Novel computational approaches to predict and reconstruct bacterial plasmids

Sergio Arredondo Alonso

Novel computational approaches to predict and reconstruct bacterial plasmids

Sergio Arredondo Alonso

Novel computational approaches to predict and reconstruct bacterial plasmids
PhD thesis, Utrecht University, the Netherlands

Author: Sergio Arredondo Alonso
Cover design: Maria Tió Coma. Photo by Alina Grubnyak on Unsplash
Lay-out: Patrique Praest and Sergio Arredondo Alonso
Printed by: Print Amsterdam BV

The research presented in this thesis was supported by the Joint Programming Initiative in Antimicrobial Resistance (third call) under the grant agreement identifier JPIA-MR2016-AC16/00039

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)

Novel computational approaches to predict and reconstruct bacterial plasmids

Focus on the nosocomial pathogen *Enterococcus faecium*

Nieuwe computer gebaseerde methoden om bacteriële plasmiden te voorspellen en te reconstrueren
(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 maandag 21 september 2020 des middags te 2.30 uur

door

Sergio Arredondo Alonso

geboren op 1 april 1993
te Sant Joan Despí (Barcelona), Spanje

Promotor:
Prof. dr. R.J.L. Willems

Copromotor:
Dr. A.C. Schürch

Commissie:
Prof. dr. A.F.J.M van den Ackerveken
Prof. dr. M.E.E Kretzschmar
Prof. dr. C. Schultsz
Prof. dr. B. Snel
Prof. dr. J.A.G.M de Visser

Paranimfen:
Jesse Kerkvliet
Julian Paganini

Table of Contents

1

General Introduction

S. Arredondo-Alonso

Parts of this chapter have been published in: A.C. Schürch, S. Arredondo-Alonso , R.J.L Willems, R.V. Goering Clin. Microbiol. Infect. (2018)
doi: 10.1016/j.cmi.2017.12.016

The emergence of bacterial strains which are resistant to antibiotics is an important threat for human health (1). Due to the rise in antimicrobial resistance (AMR), treatment of hospital acquired have become increasingly challenging (2). The capacity of a subset of bacterial pathogens to escape and evade common antimicrobial therapy in clinical settings has led to the term 'ESKAPE' pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species) which cause the majority of difficult to treat bacterial infections (3).

The surveillance of AMR is essential to guide infection control policies and monitor the effectiveness of therapy guidelines and antibiotic stewardships (4). Surveillance of antimicrobial resistance by whole-genome sequencing (WGS) has recently been adopted by public health microbiology laboratories due to the decrease in sequencing costs and the possibility of sequencing microbial genomes in a few hours (5). The implementation of WGS permits: i) a reliable microbial identification, ii) investigation in the relatedness of isolates, supporting the detection of outbreaks and phylogenetic analysis with a higher level of resolution than current molecular typing methods (6) and iii) prediction of drug susceptibility, especially relevant in slow-growing pathogens for which phenotypic data cannot be rapidly obtained (7, 8).

Whole genome sequencing: assembly and typing-based methods

Short-read sequencing technologies are most frequently used for WGS of bacteria. These sequencing technologies include Illumina NextSeq/MiSeq, 454 GS FLX or Ion Torrent, among others, that generate reads up to 300 bp and with a high read-quality (phred score 20-30, read-error rate 0.01-0.001%). Illumina NextSeq/MiSeq can generate up to 7.5–20 Gb exceedingly covering the small size of bacterial genomes which opens the possibility of multiplexing up to 96 bacterial samples in the same Illumina NextSeq sequencing run and reducing the cost per sample (9). The *de Bruijn* graph representing the sequencing data (10, 11), is commonly used by fragment assembly algorithms which make use of short substrings of the reads termed 'k-mers' (10, 11). In de Bruijn graphs, edges represent k-mers, with a fixed 'k' length, and vertices (nodes) represent (k-1)-mers. Two vertices, e.g. {v, w}, are connected if an overlap exists between the prefix of v and the suffix of w, thus creating an edge between these two vertices (12, 13). The resulting output of most assemblers is a collection of stretches of longer continuous sequences named contigs obtained by traversing the assembly graph unambiguously. Recently, assemblers also provide additional information inferred from the assembly process in a graphical fragment assembly (gfa) format (11).

Short-read WGS data enables single-nucleotide polymorphisms (SNP) analysis in which sequencing reads are mapped against a closely related reference-genome (reference-based approach) or by performing a core-genome alignment of conserved sequences (*de novo* approach) using popular bioinformatic tools such as Snippy (https://github.com/tseemann/snippy) or Roary (14), respectively. The core genome of a species is defined as the pool of genes common to all the studied genomes of a given species (15). Other approaches such as core-genome multilocus sequencing typing (cgMLST) confer a higher level of reproducibility and comparability of results between microbiology laboratories by comparison of allelic variants against a predefined and curated set of genes (MLST scheme) but they have, in general, a lower level of genomic resolution than SNP analysis approaches (15). These WGS typing techniques permit the analysis of core genome relatedness and the population structure of bacterial species to an unprecedented level.

The accessory genome and antibiotic resistance

The high flexibility and plasticity of bacteria to evolve and adapt to environmental pressures (e.g. antibiotic treatment) is importantly driven by the acquisition of large insertions or deletions (indels), genome recombination and rearrangements, or by the acquisition of foreign DNA sequences such as plasmid or phage sequences. These genetic elements belong to the 'accessory' genome, that fraction of the genome, which is only present in a subset of isolates belonging to a bacterial species.

An important part of the accessory genome linked to antimicrobial resistance is represented by plasmids. Through plasmids, a large pool of extrachromosomal genes can be gained by bacteria, including AMR genes. This pool of accessory genes can be disseminated horizontally to other isolates of the same species or other bacterial species. Despite the importance of plasmid sequences in AMR dissemination and evolution, most of the WGS epidemiological analyses are focused on the tracing of clonal strains. However, in a scenario where AMR genes are contained on plasmids and where plasmid conjugation occurs frequently (16, 17), a sole focus on the dissemination of clones will not paint the full epidemiological picture of AMR transmission. To fully understand antimicrobial resistance introduction and transmission in hospital environments, plasmid epidemiology must be taken into account (18, 19).

Challenges in the reconstruction of plasmids from short-read WGS

When analysing the molecular epidemiology of plasmids on a population level, the low-level resolution obtained by PCR-based plasmid-typing techniques and the laborious work associated with plasmid purification and subsequent sequencing are limiting factors. Accordingly, WGS has been adopted as the reference standard to analyse plasmid sequences (20, 21). However, short-read sequencing technologies cannot span plasmid repeat sequences, leading to an accurate but fragmented assembly graph resulting in hundreds of contig sequences. Several tools have been proposed to improve *de novo* plasmid assembly, but manual expert pruning is required to obtain correct plasmid boundaries, which limits the high-throughput analysis of WGS data (22).

The introduction of long-read sequencing technologies such as PacBio single-molecule real-time (SMRT) sequencing platform and Oxford Nanopore Technologies (ONT) facilitated the reconstruction of complete microbial genomes, including fully assembled plasmids (23, 24). The generation of reads with a median read length around 8-10 kbp (25) allows to span most of the repetitive structures present in bacterial genomes, including plasmid sequences, and thus obtaining genome assemblies consisting of a single sequence per replicon (26). However, these long reads are considered noisy because error rates can range from ~10% to ~1% after base-calling and consensus correction, respectively (27) which challenges the inference of SNP and cgMLST types based uniquely on long-read sequencing data. To bypass this limitation, short- and long-read sequencing can be combined in a process called hybrid assembly. Short-reads with an inherent read-error rate around 0.01-0.001% are used to perform a first short-read assembly. This high-quality but fragmented assembly is scaffolded in a second stage with long-reads by solving ambiguous paths present in the *de Bruijn* graph using assemblers such as Unicycler (28). The possibility of obtaining high-quality completed microbial genomes including fully assembled plasmids has led to the development of multiplexing strategies to reduce the long-read

sequencing cost per isolate (29). However, the costs associated with long-read WGS and the amount of short-read WGS generated and publicly available makes a reliable plasmid reconstruction using uniquely short-read WGS desirable.

Enterococcus faecium: an important multi-drug resistant nosocomial pathogen

In this thesis, I will focus on the role of plasmids on adaptation and resistance gene transfer in *Enterococcus faecium*, a nosocomial pathogen frequently associated with hospital-infections. This bacterium has a dual behaviour acting as a 'friend' by harmlessly residing in the gastrointestinal tract of mammals but also behaving like a 'foe' causing urinary tract infections and endocarditis (30, 31). The acquisition of multi-drug resistance against aminoglycosides, fluoroquinolones and most importantly against glycopeptides (*i.e.*, vancomycin), motivated the WHO to include *E. faecium* in its global priority list (32). The rise in vancomycin resistance was dramatically high in the United States increasing from 0% in the 1980s to more than 80% in late 2000s (33, 34) while in Australia, the current percentage of VREfm isolates ranges from 49% to 57% (35, 36). A significant increase from 10.5% (2012) to 17.3% (2015) in the percentage of vancomycin-resistant *E. faecium* isolates (VREfm) was also observed in Europe as indicated by the European Antimicrobial Resistance Surveillance Network (37). Although, European-wide vancomycin resistance percentages seem to be lower than in the USA or Australia, in some European countries, such as Cyprus, Romania, Ireland, Hungary, Poland, Latvia, Slovakia and Lithuania, vancomycin-resistance levels seem to be comparable to those found in Australia, with vancomycin-resistance percentages over 30%. In the Netherlands the VRE numbers in the Netherlands are still low with 1.1% of *E. faecium* isolates from blood being vancomycin-resistant.

In *E. faecium*, vancomycin-resistance is encoded by five different gene clusters of which *vanA* and *vanB* are responsible for most of the vancomycin resistance observed in clinical cases. The gene clusters consist of genes encoding a two-component regulatory system and enzymes involved in the metabolic recycle of D-Ala-D-Ala peptidoglycan precursors (38). Furthermore, both gene clusters are encoded within transposons (e.g. Tn*1546* and Tn*1549*). Tn*1546*, containing the *vanA* gene cluster has frequently been associated with plasmids (39) and horizontal transmission of large plasmids, bearing *vanA,* has frequently been reported (40). Other resistance traits, such as aminoglycoside resistance (*aac(6')-Ie-aph(2')* gene present in the transposon Tn*5281*, linezolid resistance mediated by the *cfr(B)* gene, tetracycline resistance mediated by *tet(M)*, or quinupristin-dalfopristin resistance mediated by *vat(D)* and *vat(E)* genes, are also frequently present on plasmids. Furthermore, plasmids carrying these antibiotic resistance genes are not only found in humans but also in farm animals. This underpins the importance of a One-Health perspective when studying the dissemination of antibiotic resistance in *E. faecium* (41).

Initial studies on *E. faecium* population structure based on MLST data analysed with eBurst revealed that hospitalized patient isolates formed a distinct group termed clonal complex (CC) 17 which was distributed worldwide (42). The introduction of novel algorithms to predict the population structure revealed that the assignment into CC derived from MLST data can be incorrect due to the recombination occurring to the loci involved in MLST designation (43).

To circumvent this, MLST data in combination with an algorithm based on Bayesian analysis population structure (BAPS) was later conducted on a set of more than 1,700 *E. faecium* strains (44) which confirmed the existence of separate lineages splitting hospitalized patients with BAPS 3-3 significantly associated, and farm animals isolates represented by BAPS 2-1 and BAPS 2-4 (44). Admixture analysis showed scarce recombination events within hospitalized patient isolates suggesting that once *E. faecium* isolates adapted to the hospital environment, most likely by horizontal-gene transfer (HGT) events, they became isolated and thus imposed a restriction on the gene flow to other subpopulations or lineages.

WGS-based phylogenetic studies conducted in the US and Europe split the *E. faecium* population into two distinct lineages corresponding to i) clade A, a hospital-associated clade; and ii) clade B, a community-associated clade (45). Further studies suggested the subdivision of clade A into clade A1 and clade A2 with grouping the majority of hospital-associated and animal isolates in these two distinct clades (46). However, this subdivision in only two distinct clades was not supported by recent population structure studies that suggested that isolates specifically from farm animals were genomically more heterogeneous and clustered in multiple distinct lineages (38, 47). Genomic analysis also revealed that isolates belonging to clade A and B differ in genome size (46) and that genomic variation is importantly driven by recombination events (48). This indicates that genome plasticity among *E. faecium* has contributed importantly to adaptation and colonization to new environments and hosts (38). The recombination observed in *E. faecium* has also questioned the validity of MLST-based types and remarked the importance of WGS studies to monitor VREfm outbreaks (49). The role of plasmid sequences driving *E. faecium* clade A and B population structure and the dissemination of vancomycin resistance mediated by *vanA* and *vanB* gene clusters was largely unexplored and motivated the studies conducted in this thesis.

Outline of this thesis

The aim of the research described in this thesis is to reconstruct plasmid sequences from a large collection of *E. faecium* isolates from different isolation sources, including human and non-human samples. Detection and reconstruction of plasmids from WGS sequencing data are challenged by the presence of repeat units that cannot be spanned by the

length offered by short reads and hinder the prediction of plasmid sequences in the studied *E. faecium* population. For this purpose, novel bioinformatic tools and approaches were developed to improve the prediction of plasmids from short-read WGS data. This unravelled the contribution of plasmids to source-specificity and allowed to detect several *vanA* plasmid configurations present in Dutch hospitals in different *E. faecium* clonal groups. In **chapter 2**, we benchmarked several bioinformatic tools to detect and reconstruct plasmid sequences from short-read WGS data. We determined PlasmidSPAdes as the best plasmid prediction tool to predict the origin of contigs (plasmid- or chromosome- derived) in terms of precision and completeness but observed that the plasmid boundaries were mostly incorrect. Plasmid-predicted contigs were frequently merged together and this made the tracking of a single plasmid in a large collection of bacterial isolates unfeasible. **In chapter 3**, we developed a new tool termed 'mlplasmids' for which we used a machine-learning based approach to binarily predict the origin of contigs derived from short-read WGS data based on pentamer frequencies for three different bacterial species: *E. faecium*, *Klebsiella pneumoniae* and *Escherichia coli*. The resulting support-vector machine classifiers were made available as open-source software (R package) and a web-server interface (Shiny app). Mlplasmids was fundamental to confidently predict the origin of contigs from *E. faecium* and to avoid the presence of chromosomal contamination. In **chapter 4,** we used a combination of short- and long-read sequencing data of 62 *E. faecium* isolates to maximize and increase the number and diversity of complete plasmid sequences available for *E. faecium*. This resulted in 305 plasmids that were indispensable to train and test the mlplasmids tool presented in chapter 3. Furthermore, we used mlplasmids to predict the plasmidomes, defined as the set of plasmid sequences present in a single sample, in a set of 1,582 *E. faecium* isolates for which only short-read WGS data were available. A k-mer analysis of the pan-plasmidome allowed to discern the existence of several plasmidome populations in which hospitalized-patients were clearly distinct. We also concluded that the plasmidome rather than the chromosome was most informative for source-specificity and thus indicative for restrictions in the flow of plasmid genes between isolation sources. **In chapter 5,** we developed a novel tool termed 'gplas' in which we tackled the main limitation derived from mlplasmids corresponding to the lack of plasmid boundaries in the prediction. Gplas bins plasmid-predicted sequences into different bins using k-mer coverage and composition from the nodes and edges present in short-read graphs. The approach generated a plasmidome network, based on likely plasmid walks along the graph, that was further queried to observe highly-connected components that are finally predicted as bins. Gplas was also released as open-source software and integrated a Snakemake pipeline to fully automate the analysis. **In chapter 6,** we provided a comprehensive picture of the population genomics and molecular epidemiology of *vanA* positive VRE isolated from 32 Dutch hospitals between 2012 to 2015.

In this chapter, we used gplas to predict plasmids containing the *vanA* gene cluster and compared the distribution of vancomycin resistance from both clonal (BAPS and PopPUNK analysis) and plasmid (gplas and k-mer analysis) perspectives. Using a network approach, we observed that *vanA* resistance was driven by at least four different plasmid types carried by different clonal groups and present at distinct European countries. Finally, the new methods and analyses proposed in this thesis, as well as some of the limitations and future directions, are further discussed in **chapter 7**.

References

1. WHO. 2015. Global action plan on antimicrobial resistance. http://www.who.int/antimicrobial-resistance/global-action-plan/en.

2. Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, Vlieghe E, Hara GL, Gould IM, Goossens H, Greko C, So AD, Bigdeli M, Tomson G, Woodhouse W, Ombaka E, Peralta AQ, Qamar FN, Mir F, Kariuki S, Bhutta ZA, Coates A, Bergstrom R, Wright GD, Brown ED, Cars O. 2013. Antibiotic resistance—the need for global solutions. Lancet Infect Dis 13:1057–1098.

3. Boucher HW, Talbot GH, Bradley JS, Edwards JE, Gilbert D, Rice LB, Scheld M, Spellberg B, Bartlett J. 2009. Bad Bugs, No Drugs: No ESKAPE! An Update from the Infectious Diseases Society of America. Clinical Infectious Diseases 148:1-12.

4. Tacconelli E, Sifakis F, Harbarth S, Schrijver R, van Mourik M, Voss A, Sharland M, Rajendran NB, Rodríguez-Baño J, EPI-Net COMBACTE-MAGNET Group. 2018. Surveillance for control of antimicrobial resistance. Lancet Infect Dis 18:e99–e106.

5. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8:e1002824.

6. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ. 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366:2267–2275.

7. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. 2019. Using Genomics to Track Global Antimicrobial Resistance. Front Public Health 7:242.

8. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V, Supply P, Suresh A, Utpatel C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, de Jong BC, de Vos M, Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC, Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM, Niemann S, Comas I, Van Rie A. 2019. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. Nat Rev Microbiol 17:533–545.

9. Schürch AC, van Schaik W. 2017. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. Annals of the New York Academy of Sciences 1388:108-120.

10. Pevzner PA, Tang H, Tesler G. 2004. *De novo* repeat classification and fragment assembly. Genome Res 14:1786–1796.

11. Gonnella G, Kurtz S. 2016. RGFA: powerful and convenient handling of assembly graphs. PeerJ 4:e2681.

12. Zerbino DR, Birney E. 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18:821–829.

13. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev M a., Pevzner P a. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

14. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.

15. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin Microbiol Infect 24:350–354.

16. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, Giess A, Pankhurst LJ, Vaughan A, Grim CJ, Cox HL, Yeh AJ, Modernising Medical Microbiology (MMM) Informatics Group, Sifri CD, Walker AS, Peto TE, Crook DW, Mathers AJ. 2016. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene bla_{KPC}. Antimicrob Agents Chemother 60:3767–3778.

17. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X, Yu L-F, Gu D, Ren H, Chen X, Lv L, He D, Zhou H, Liang Z, Liu J-H, Shen J. 2016. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis 16:161–168.

18. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y, Tsai Y-C, Boitano M, Dayal J, Brooks SY, Schmidt B, Young AC, Thomas JW, Bouffard GG, Blakesley RW, NISC Comparative Sequencing Program, Mullikin JC, Korlach J, Henderson DK, Frank KM, Palmore TN, Segre JA. 2014. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med 6:254ra126.

19. Conlan S, Park M, Deming C, Thomas PJ, Young AC, Coleman H, Sison C, NISC Comparative Sequencing Program, Weingarten RA, Lau AF, Dekker JP, Palmore TN, Frank KM, Segre JA. 2016. Plasmid Dynamics in KPC-Positive *Klebsiella pneumoniae* during Long-Term Patient Colonization. MBio 7: e00742-16.

20. Brolund A, Franzen O, Melefors O, Tegmark-Wisell K, Sandegren L. 2013. Plasmidome-Analysis of ESBL-Producing *Escherichia coli* Using Conventional Typing and High-Throughput Sequencing. PLoS One 8:e65793.

21. De Toro M, Pilar Garcillán-Barcia M, De F, Cruz L. 2014. Plasmid Diversity and Adaptation Analyzed by Massive Sequencing of *Escherichia coli* Plasmids. Microbiol Spectrum 2:PLAS–0031–2014.

22. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, Peto T, Crook D, Woodford N, Walker AS, Phan H, Sheppard AE. 2017. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. Front Microbiol 8:182.

23. Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM. 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14:R101.

24. Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733–735.

25. Maio ND, De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M, Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Sarah Walker A, Stoesser N, on behalf of the REHAB consortium. 2019. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microbial Genomics 5:e000294.

26. George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N, Sheppard AE, Mathers A, Norris R, Navickaite I, Eaton C, Iqbal Z, Crook DW, Phan HTT. 2017. Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. Microbial Genomics 3: e000118.

27. Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology 28:129.

28. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.

29. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics 3:e000132.

30. Bertics PJ, Wiepz GJ. 2009. New Developments with Vancomycin-Resistant Enterococci: *E. faecium*—Friend or Foe? J Infect Dis 200:679–681.

31. Van Tyne D, Gilmore MS. 2014. Friend turned foe: evolution of enterococcal virulence and antibiotic resistance. Annu Rev Microbiol 68:337–356.

32. Tacconelli E, Magrini N, Kahlmeter G, Singh N. 2017. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. World Health Organization 27.

33. Arias CA, Murray BE. 2012. The rise of the Enterococcus: beyond vancomycin resistance. Nat Rev Microbiol 10:266–278.

34. Miller WR, Murray BE, Arias CA. 2013. Emergence and management of drug-resistant enterococcal infections. Microbial Drug Resistance 6:637-55.

35. Leong KWC, Cooley LA, Anderson TL, Gautam SS, McEwan B, Wells A, Wilson F, Hughson L, O'Toole RF. 2018. Emergence of Vancomycin-Resistant *Enterococcus faecium* at an Australian Hospital: A Whole Genome Sequencing Analysis. Sci Rep 8:6274.

36. Turnidge J, Binns P, Cruickshank M, Firman J, Heaney A, McKenzie D, Others. 2017. AURA 2017: second Australian report on antimicrobial use and resistance in human health.

37. ECDC. 2019. European Centre for Disease Prevention and Control. Surveillance of antimicrobial resistance in Europe 2018.

38. Guzman Prieto AM, van Schaik W, Rogers MRC, Coque TM, Baquero F, Corander J, Willems RJL. 2016. Global Emergence and Dissemination of Enterococci as Nosocomial Pathogens: Attack of the Clones? Front Microbiol 7:788.

39. Werner G, Freitas AR, Coque TM, Sollid JE, Lester C, Hammerum AM, Garcia-Migura

L, Jensen LB, Francia MV, Witte W, Willems RJ, Sundsfjord A. 2011. Host range of enterococcal *vanA* plasmids among Gram-positive intestinal bacteria. J Antimicrob Chemother 66:273–282.

40. Freitas AR, Coque TM, Novais C, Hammerum AM, Lester CH, Zervos MJ, Donabedian S, Jensen LB, Francia MV, Baquero F, Peixe L. 2011. Human and swine hosts share vancomycin-resistant *Enterococcus faecium* CC17 and CC5 and *Enterococcus faecalis* CC2 clonal clusters harboring Tn1546 on indistinguishable plasmids. J Clin Microbiol 49:925–931.

41. McEwen SA, Collignon PJ. 2018. Antimicrobial Resistance: a One Health Perspective. Microbiol Spectr 6.

42. Top J, Willems R, Bonten M. 2008. Emergence of CC17 *Enterococcus faecium*: from commensal to hospital-adapted pathogen. FEMS Immunol Med Microbiol 52:297–308.

43. Turner KME, Hanage WP, Fraser C, Connor TR, Spratt BG. 2007. Assessing the reliability of eBURST using simulated populations with known ancestry. BMC Microbiol 7:30.

44. Willems RJL, Top J, van Schaik W, Leavis H, Bonten M, Sirén J, Hanage WP, Corander J. 2012. Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. MBio 3:e00151–12.

45. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, Desjardins C, Cerqueira G, Gevers D, Walker S, Wortman J, Feldgarden M, Haas B, Birren B, Gilmore MS. 2012. Comparative Genomics of Enterococci: Variation in *Enterococcus faecalis*, Clade Structure in *E. faecium*, and Defining Characteristics of *E. gallinarum* and *E. casseliflavus*. mBio 3: e00318-11.

46. Lebreton F, van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L, Saif S, Young S, Zeng Q, Wortman J, Birren B, Willems RJL, Earl AM, Gilmore MS. 2013. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. MBio 4:e00534–13.

47. Raven KE, Reuter S, Reynolds R, Brodrick HJ, Russell JE, Török ME, Parkhill J, Peacock SJ. 2016. A decade of genomic history for healthcare-associated *Enterococcus faecium* in the United Kingdom and Ireland. Genome Res 26:1388–1396.

48. de Been M, van Schaik W, Cheng L, Corander J, Willems RJ. 2013. Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium*. Genome Biol Evol 5:1524–1535.

49. van Hal SJ, Ip CLC, Ansari MA, Wilson DJ, Espedido BA, Jensen SO, Bowden R. 2016. Evolutionary dynamics of *Enterococcus faecium* reveals complex genomic relationships between isolates with independent emergence of vancomycin resistance. Microb Genom 2:e000048.

2

On the (Im)possibility of Reconstructing Plasmids From Whole-Genome Short-Read Sequencing Data

Sergio Arredondo-Alonso, Rob J. Willems, Willem van Schaik, Anita C. Schürch

Published in: Microb. Genom. (2017) doi: 10.1099/mgen.0.000128

Abstract

To benchmark algorithms for automated plasmid sequence reconstruction from short-read sequencing data, we selected 42 publicly available complete bacterial genome sequences spanning 12 genera, containing 148 plasmids. We predicted plasmids from short-read data with four programs (PlasmidSPAdes, Recycler, cBar and PlasmidFinder) and compared the outcome to the reference sequences. PlasmidSPAdes reconstructs plasmids based on coverage differences in the assembly graph. It reconstructed most of the reference plasmids (recall=0.82), but approximately a quarter of the predicted plasmid contigs were false positives (precision=0.75). PlasmidSPAdes merged 84% of the predictions from genomes with multiple plasmids into a single bin. Recycler searches the assembly graph for sub-graphs corresponding to circular sequences and correctly predicted small plasmids, but failed with long plasmids (recall=0.12, precision=0.30). cBar, which applies pentamer frequency analysis to detect plasmid-derived contigs, showed a recall and precision of 0.76 and 0.62, respectively. However, cBar categorizes contigs as plasmid-derived and does not bin the different plasmids. PlasmidFinder, which searches for replicons, had the highest precision (1.0), but was restricted by the contents of its database and the contig length obtained from *de novo* assembly (recall=0.36). PlasmidSPAdes and Recycler detected putative small plasmids (<10 kbp), which were also predicted as plasmids by cBar, but were absent in the original assembly. This study shows that it is possible to automatically predict small plasmids. Prediction of large plasmids (>50 kbp) containing repeated sequences remains challenging and limits the high-throughput analysis of plasmids from short-read whole-genome sequencing data.

Introduction

A bacterial cell can hold zero, one or multiple plasmids with varying sizes and copy numbers. Traditionally, plasmid sequencing involved methods to purify plasmid DNA, followed by shot-gun sequencing, which frequently necessitated closing of gaps by primer-walking (1). Plasmid DNA purification is exceedingly difficult if it involves plasmids longer than 50 kbp (1, 2). Alternatively, plasmid sequences can be assembled from whole-genome-sequencing (WGS) data generated by high-throughput short-read sequencing platforms. However, plasmids often contain repeat sequences that are shared between the different physical DNA units of the genome, which prohibits complete assembly from short-read data. Assembly often results in many fragmented contigs per genome of unclear origin (plasmid or chromosome) (3). Currently available plasmid prediction programs either aim to determine whether a previously assembled contig is from a plasmid (PlasmidFinder, cBar), or try to reconstruct whole plasmid sequences from the sequencing reads or the assembly graph (Recycler, PlasmidSPAdes, PLACNET) (Table 1).

PlasmidFinder is a web-based tool that was developed to detect replicon sequences in assemblies and is optimized for use in enterobacterial genomes (4). Since two plasmids sharing the same replication mechanism cannot coexist in the long term within the same cell, replicon sequences are used to classify plasmids into different incompatibility groups (4). cBar was specifically designed to predict plasmid-derived sequences based on differences in k-mer composition (5). It relies on differences in pentamer frequencies from 881 complete prokaryotic sequences and gives a binary classification of chromosome- or plasmid-derived contig. PLACNET (plasmid constellation network) reconstructs plasmids from WGS data by integrating three lines of evidence: (i) scaffold linking and coverage information, (ii) presence of replication initiator proteins (Rip) and relaxase proteins (Rel), (iii) similarity of the sequences with a custom database containing non-redundant plasmid sequences from the National Center for Biotechnology Information (6). Manual pruning in Cytoscape is necessary to obtain disjoint components (7, 8). Prediction reproducibility rates are thus highly dependent on the expertise of the researcher. As we aimed to test only fully automated methods for plasmid prediction, we excluded PLACNET from the comparison.

More recently, two algorithms that predict plasmids on the basis of the information contained in the *de Bruijn* graph were published: Recycler (9) and PlasmidSPAdes (10). Recycler extracts the information from the *de Bruijn* graph searching for sub-graphs (cycles) corresponding to plasmids. Selection of the cycles is based on the following assumptions: (i) nodes forming a plasmid have a uniform coverage, (ii) a minimal path must be selected between edges because of repetitive sequences, (iii) contigs belonging to the same cycle have concordant paired-end information, and (iv) plasmid cycles exceed

Program	Input	Paired-end information	Coverage	k-mer composition	*De Bruijn* graph	Similarity to replicons	Similarity to relaxases	Similarity to plasmids	Web tool	Command-line interface	Included in the study
PlasmidFinder [4]	Contigs					✓			✓		✓
cBar [5]	Contigs			✓						✓	✓
Recycler [9]	BAM+assembly graph	✓	✓		✓					✓	✓
PlasmidSPAdes [10]	Reads	✓	✓		✓					✓	✓
PLACNET [6]	BAM/SAM+contigs	✓	✓			✓	✓	✓		✓	

a minimum length. PlasmidSPAdes assumes a highly uniform contig coverage within the chromosome. It calculates the median coverage from the SPAdes assembly graph (11) to estimate a chromosome coverage. PlasmidSPAdes then builds a second assembly graph (referred to as the plasmid graph) only considering contigs with a read contig coverage differing from the chromosome coverage. After repeat resolution using ExSPAnder (12), connected components in the plasmid graph are reported as putative plasmids.

Here, we benchmarked currently available programs starting either from the reads or from assembled contigs. The aim of this study was to determine whether it was possible to obtain complete plasmid sequences in an automated fashion.

Results

Prediction per single plasmids

We defined a minimum recall value of 0.9 to classify a plasmid as correctly predicted. Out of 148 reference plasmids included in this study, 133 (89.9%) were correctly predicted by either PlasmidFinder, cBar, Recycler or PlasmidSPAdes (Figs 1 and 2). PlasmidSPAdes correctly predicted 125 plasmids, cBar 84 plasmids, Recycler 21 plasmids and PlasmidFinder 13 plasmids at a recall of 0.9 or more (Figs 1 and 2a, b).

Of all 148 plasmids, 5 plasmids were consistently correctly predicted by all of the programs (Fig. 2a). These included two large plasmids belonging to two *Klebsiella pneumoniae* strains (CAV1741 and PMK1). These plasmids were fully assembled and did not share any similarity within the bacterial genome. In contrast, 15 plasmids consistently had a recall value less than 0.9 in all predictions. Four of these fifteen plasmids were not fully covered by SPAdes contigs, precluding complete prediction of the plasmids. The definition of recall per plasmid operated here does not take into account whether plasmids were accurately predicted in unique and independent bins. Programs with a high mean recall (PlasmidSPAdes and cBar,

Table 1. Overview of the programs to predict plasmids from short-read sequencing data

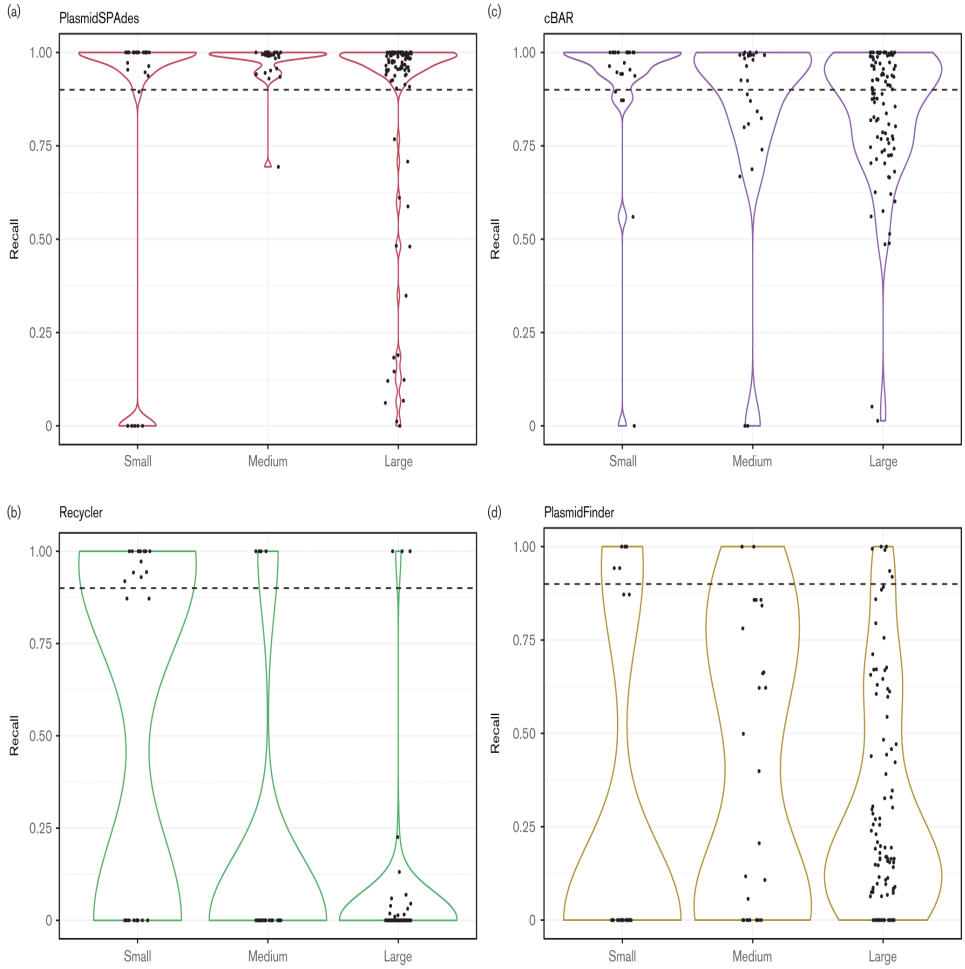


Figure 1. Performance of the programs on a single plasmid level. Recall values of small (less than 10 kbp), medium (from 10 to 50 kpb) and large (greater than 50 kbp) plasmids by PlasmidSPAdes, cBar, Recycler and PlasmidFinder. Recall was calculated by aligning the reference plasmid sequences against the plasmid predictions of each genome and disregarded plasmid binning (if any).

0.87 and 0.86, respectively) did not predict, or often incorrectly predicted, plasmid binning. cBar performs a binary classification predicting contigs as either 'plasmid' or 'chromosome', but did not sort the sequences into different plasmids from the same bacterial isolate. PlasmidSPAdes correctly predicted 120 reference plasmids (recall >0.9) present in genomes with more than one reference plasmid (n=35). From these 120 correctly predicted plasmids, 19 plasmids were accurately predicted in a single unique bin and 101 plasmids were merged in a bin with other predicted plasmids from the same genome (Supplementary Results 2). Therefore, plasmid binning was not correctly predicted in 84% of the cases and plasmid structural information was not readily retrievable.

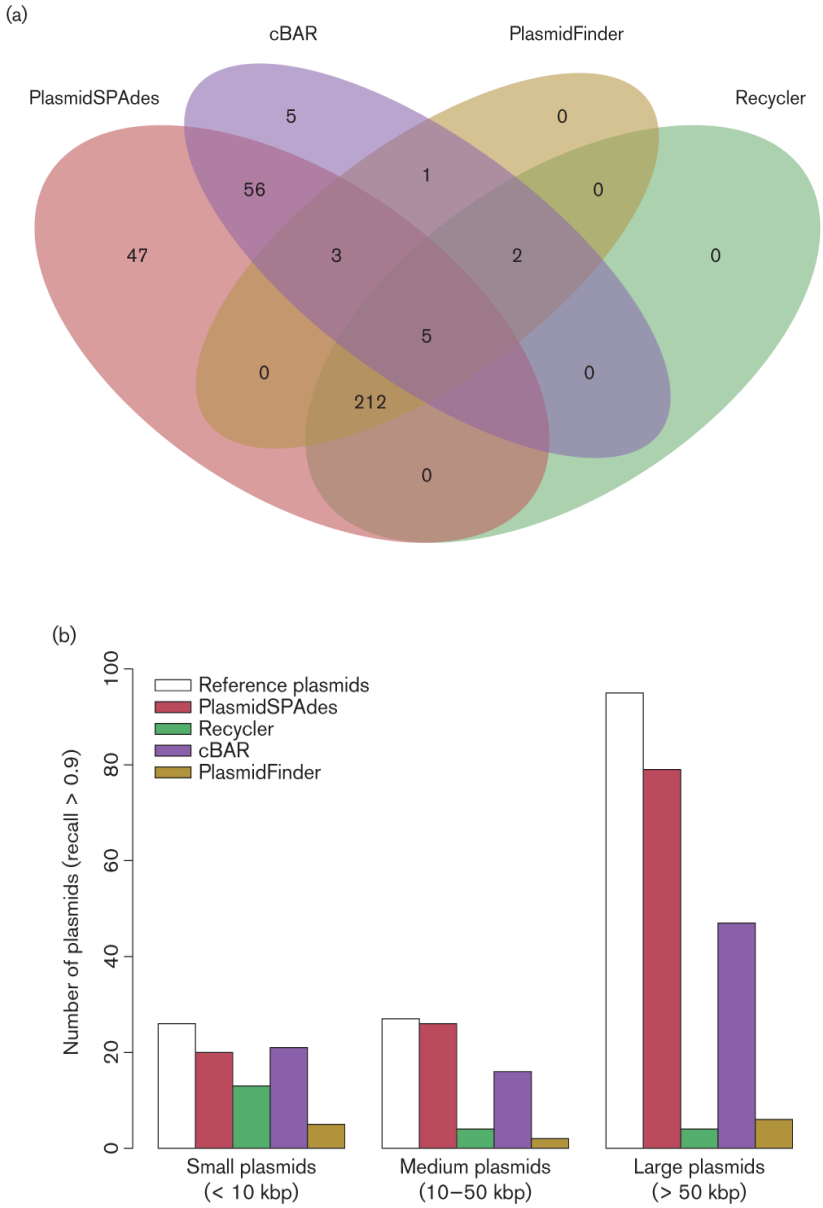


Figure 2. Comparison of program performance on a single plasmid level. (a) A minimum recall value of 0.9 in the program prediction was selected to consider a plasmid as correctly predicted. Venn diagram showing the overlap in prediction between PlasmidSPAdes (red), cBar (purple), PlasmidFinder (orange) and Recycler (green). The intersection of the ellipses showed five plasmids present in all the predictions. (b) Reference plasmids were classified into small (less than 10 kbp), medium (from 10 to 50 kbp) and large (greater than 50 kbp) plasmids depending on their size. The number of reference plasmids correctly predicted (minimum recall value of 0.9) by the programs is represented in the three categories.

Prediction per genome

Next, performance was evaluated on the genome level; thus, comparing the entirety of all predicted plasmid sequences of each genome against all plasmids of each genome. PlasmidSPAdes analysis resulted in a mean plasmid fraction of 0.72 and a mean chromosome fraction of 0.22 (Fig. 3a). Furthermore, an overall precision of 0.75 and an overall recall of 0.82 were reported. The completeness of the prediction was high even in the bacterial isolates with a high number of reference plasmids. However, PlasmidSPAdes merged plasmid contigs into a single bin if they shared repeated sequences as shown in Fig. S4.

Surprisingly, a mean fraction of 0.06 corresponding to contigs absent from the reference genomes was detected (Fig. 3a and Table S3). Most of the contigs present in the fraction of novel sequences were detected as isolated components by PlasmidSPAdes, with the exception of novel sequences in *Escherichia coli* strains JJ1886 and JJ1887. Predicted plasmid contigs that were absent in the reference genomes of *E. coli* JJ1886 and JJ1887 had high similarity with *Staphylococcus aureus* chromosome and plasmids. This potential contamination was not filtered by PlasmidSPAdes, because its coverage differed from the host chromosome. Further discussion on the identification of potential novel small cryptic plasmids is available in Supplementary Results 3 and 4.

Recycler analysis resulted in a mean plasmid fraction of 0.24, a mean chromosome fraction of 0.62 and a mean fraction of novel sequences of 0.14 (Fig. 3a). We reported an overall precision of 0.30, indicating that a high number of sequences predicted as plasmid originated from the chromosome. Recycler obtained a low overall recall of 0.12 (Fig. 3b). This low value can partly be explained by the fact that the algorithm only reports unique circular sequences. The recall value obtained by Recycler was 1.0 in samples with small or medium size plasmids (e.g. *Bacillus subtilis* BEST195 or *Enterobacter aerogenes* CAV1320). Furthermore, large plasmids not sharing any repeated sequence with other replicons were also correctly predicted by Recycler (Fig. 3b and Table S4). The Recycler chromosome fraction was further analysed to observe whether non-plasmid mobile genetic elements were predicted. A total of 14% of the contigs considered as false-positive results and mapping to their respective chromosomes were identified as prophage sequences (Supplementary Results 3, Fig. S5). Recycler more robustly detected plasmid sequences in contaminated samples than PlasmidSPAdes. This feature is reflected in *E. coli* JJ1886 and JJ1887, where the fraction of novel sequences was not higher compared to the rest of the genomes (Fig. 3a). Most of the novel contigs predicted by Recycler were also predicted by PlasmidSPAdes (Table S3). Common features of these novel contigs are a length less than 10 kbp and an intermediate copy number (Supplementary Results 4).

cBar predicted every contig as either plasmid derived or chromosome derived. cBar

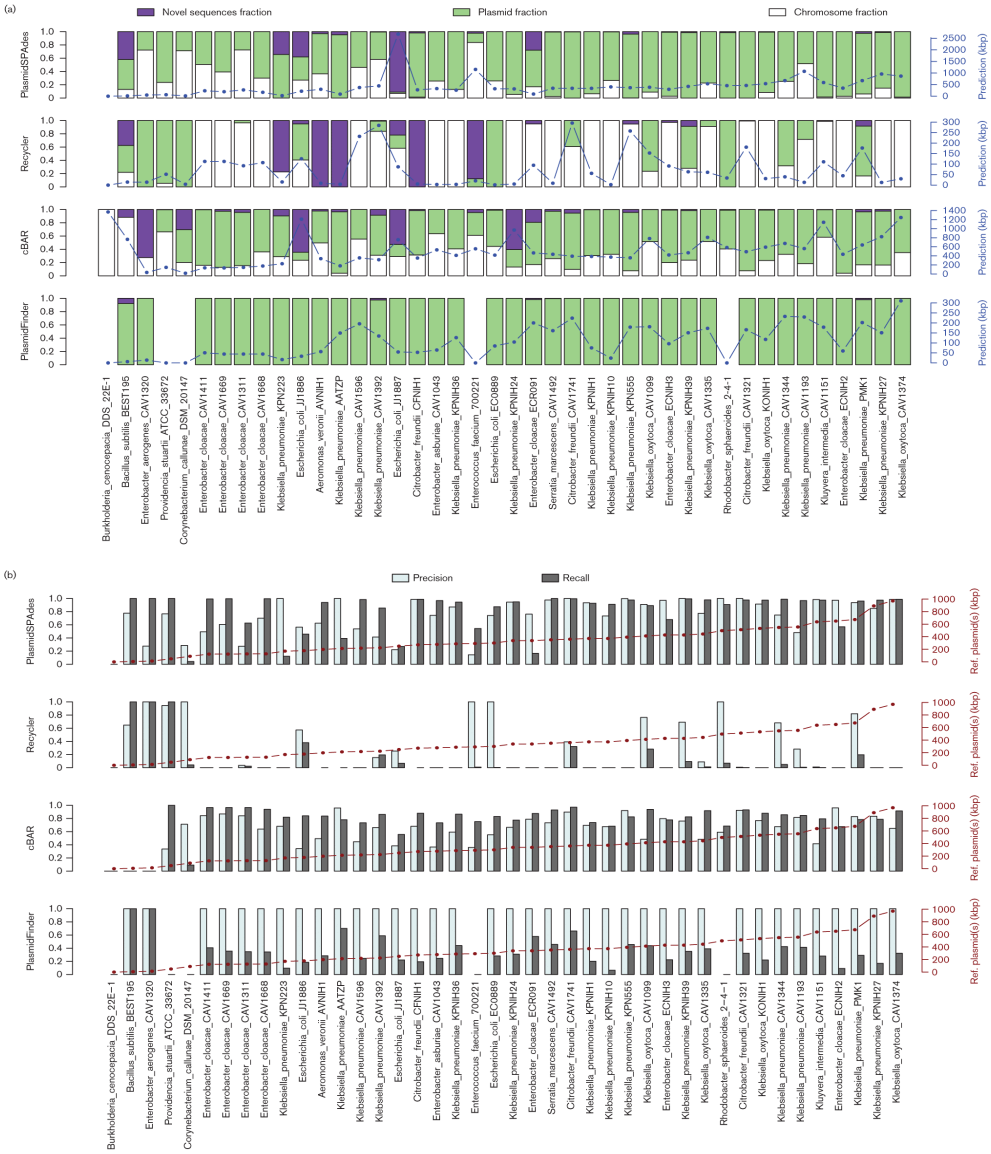


Figure 3. Performance of the programs on a genome level. (a) The prediction of each program was mapped against the reference genomes of each bacterial isolate. Contigs mapping to the reference plasmids were depicted as plasmid fraction (green bars), to the reference chromosome as chromosome fraction (white bars) or to neither as novel sequences fraction (purple bars). On the right-hand-side y-axis, the total length (in kbp) of plasmid prediction is indicated. cBar was the only program predicting contigs as plasmids in the genome that was used as a negative control (*Burkholderia cenocepacia* DDS 22E-1). (b) Precision and recall values are represented with white and grey bars, respectively. A precision of 1 indicates the absence of contigs mapping to the reference chromosome in the prediction. Recall of 1 indicates the full sequences of all the reference plasmids were present in the prediction. On the right-hand-side y-axis, the total plasmid length (in kbp) of a particular bacterial genome is indicated.

analysis resulted in a mean plasmid fraction of 0.58, a mean chromosome fraction of 0.33 and a mean fraction of novel sequences of 0.09. We reported an overall precision and recall of 0.62 and 0.76, respectively. However, the precision varied largely across genomes, as reflected in *Providencia stuartii* ATCC 33762 (recall=1.0, precision=0.34). This strain has a single plasmid of 48.87 kbp (Fig. S1), which was correctly detected by cBar, but also 19 other contigs (>500 bp) mapping to the chromosome that were wrongly predicted as plasmids (Fig. 3). In *B. subtilis* subsp. natto BEST195 and *E. aerogenes* CAV1320, which carry single plasmids, precision and recall value were 0 (Fig. 3b). Notably, in the negative control *B. cenocepacia* DDS 22E-1, which does not have any plasmids, cBar predicted a total size of 1369 kbp wrongly as plasmid-derived contigs (Fig. 3a). All previously unidentified putative plasmids (Table S3) were also classified as plasmids by cBar, with the exception of two fragments in *Aeromonas veronii* AVNIH1 and Klebsiella oxytoca KONIH1.

PlasmidFinder analysis resulted in a mean plasmid fraction of 0.99 and a mean fraction of novel sequences of 0.01. PlasmidFinder was able to detect at least one plasmid replicon sequence in 37 bacterial strains, but failed to detect any replicon sequence in *Rhodobacter sphaeroides* 2-4-1 and in the Gram-positive bacteria *Corynebacterium callunae* DSM 20147, *Enterococcus faecium* ATCC 700221 and *P. stuartii* ATCC 33672. Surprisingly, in *B. subtilis* BEST195, one of the four Gram-positive strains, a recall of 1.0 was obtained. This single plasmid of *B. subtilis* BEST195 had an identity of 88% and covered 82% of a replicon sequence (NC_015392) from *Salmonella enterica* indexed in the PlasmidFinder database. The database of PlasmidFinder was designed to detect replicon sequences of plasmids from the Enterobacteriaceae. Therefore, we excluded all Gram-positive genomes to calculate the overall precision and recall of PlasmidFinder. This resulted in an overall precision of 1.0, indicating that no false-positive sequences were predicted as plasmids. However, the low overall recall of 0.36 was due to the low connectivity of the assemblies that were generated using only short-read sequencing data (Fig. 3b).

Discussion

The large majority of plasmids (89.9%) could be correctly predicted by one of the tested programs. However, in many cases, the predictions were fragmented (all programs), contaminated by chromosome sequences (cBar, Recycler, PlasmidSPAdes), the binning of the plasmids were unclear (cBar, PlasmidSPAdes) and the plasmids were incomplete (all programs). In absence of reference plasmid sequences, disentangling or binning the sequences into separate plasmids is a challenging step.

PlasmidSPAdes fully or partially predicted most plasmids (recall=0.82). The major drawback of PlasmidSPAdes was the merging of predicted plasmids into a single bin. This limitation can partially be overcome by a similar methodology as previously applied in PLACNET (6). By visualizing the plasmid graph and connecting contigs with a similar coverage

and scaffolding linkage, plasmid sub-graphs can be separated manually, but only if the different plasmids sufficiently differ in their copy number (10) (Fig. S3). Repeat sequences, such as transposases, that merge different components in the assembly graph, can be spotted by their high number of scaffolding links and coverage. However, this process is highly dependent on the expertise of the individual analysing the data, may be difficult to reproduce independently, and can only be performed if coverage of plasmids differs. Consequently, this approach limits the high-throughput analysis of short-read WGS data to correctly predict plasmid sequences.

Recycler applies an innovative and general approach to plasmid prediction, and successfully extracted complete plasmid sequences if they had circular features present in the assembly graph. Most large plasmids, however, tend to be assembled into several contigs due to the presence of repeated sequences with high coverage. Recycler failed to extract these types of plasmids and in many cases only extracted non-plasmid mobile elements. cBar was originally designed to categorize chromosome and plasmids in metagenomic sequences. Its accuracy is known to be lower for long plasmids because the nucleotide composition of these plasmids is similar to the host chromosome (14). However, the overall recall of cBar is high (0.78) and it might be well-suited to confirm if a sequence was predicted to be plasmid-derived by another method.

The results of PlasmidFinder indicated a high reliability of the prediction. If applied to PlasmidSPAdes predictions, the detection of different incompatibility groups by PlasmidFinder could either indicate the presence of two or more plasmids merged together into a single component or the presence of a multireplicon plasmid.

To obtain the full sequences of plasmids, long-read sequencing data can be a solution (15). However, to our surprise, PlasmidSPAdes and Recycler predicted a substantial number of contigs (fraction of novel sequences: 0.06 and 0.14, respectively) that were not present in the complete reference genomes sequenced with long reads. These sequences could originate from sequences filtered as contaminants, but could also represent small replicons (Supplementary Results 2 and 4). As described elsewhere, the hierarchical genome assembly process (HGAP) can lead to missing small plasmids in the main assembly when using a seed read length cut-off greater than actual plasmid size (16, 17). Library preparation withDNA size selection prior to PacBio sequencing can also obviate small plasmids when the cut-off selected is higher than actual replicon size (2).

We showed that it is possible to automatically predict the sequences of small and circular plasmids. Nonetheless, the correct prediction of large plasmids (>50 kbp) containing repeated sequences remains challenging using short-read sequencing data only.

Material and Methods

At the time the study was conceived (July 2016), we selected all bacterial genomes with complete plasmid sequences and Illumina Miseq or Hiseq paired-end data publicly available. Complete genome sequences and reads were downloaded from GenBank and the National Center for Biotechnology Information Sequence Read Archive, respectively (Table S1). All the strains were previously sequenced by Pacific Biosystems PacBio RS II.

The above criteria resulted in a set of 42 genomes that spanned twelve different genera (Fig. S1). The test data contained 148 plasmid sequences ranging from 1.55 to 338.85 kbp (Fig. S1, Table S1) and 45 chromosomal sequences ranging from 0.93 to 6.26 Mbp. This set included five genomes used in PlasmidSPAdes and Recycler publications to ensure consistency between present and previously published results (Supplementary Results 1 and Table S2). We used QUAST 4.1 (13) to map the predicted plasmid contigs against (i) each reference plasmid separately or (ii) the reference genome (containing chromosomes and plasmids, Fig. S2). Nucmer alignments were used to assign each of the predicted plasmid contigs to one of the following three categories: 'plasmid fraction' (true positive), 'chromosome fraction' (false positive) and 'fraction of novel sequences' (absent from the reference genomes). A minimum alignment of 500 bp and 95% nucleotide identity was required to assign a contig to a certain fraction. We considered the whole contig length to evaluate the performance of the programs using recall and precision.

- Recall was defined as the percentage of the reference plasmid(s) covered by the prediction. On the individual plasmid level, a recall of 1 indicates that the full reference plasmid sequence was present among the predicted plasmids. On the whole genome level, a recall of 1 indicates all the reference plasmids were fully present among the predicted plasmids.
- Precision was defined as the fraction of true positives (plasmid fraction) divided by the sum of true and false positive results (plasmid and chromosome fraction).

For each genome (n=42), we calculated precision and recall values. To calculate the overall precision and recall of PlasmidSPAdes, Recycler and PlasmidFinder, we excluded the negative control *Burkholderia cenocepacia* strain 22E-1 as no false-positive results were obtained. Additionally, the overall precision and recall of PlasmidFinder was calculated filtering out genomes corresponding to Gram-positive bacteria. A detailed explanation of the metrics reported in the paper is available in Supplementary Methods.

References

1. Smalla K, Jechalke S, Top EM. Plasmid detection, characterization, and ecology. Microbiol Spectr 2015;3:PLAS-0038-2014.

2 Conlan S, Thomas PJ, Deming C, Park M, Lau AF et al. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae.

Sci Transl Med 2014;6:254ra126.

3. De Toro M, Garcillaon-Barcia MP, De La Cruz F. Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. Microbiol Spectr 2014;2:PLAS–0031.

4. Carattoli A, Zankari E, García-Fernandez A, Voldby Larsen M, Lund O et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 2014;58:3895–3903.

5. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 2010;26:2051–2052.

6. Lanza VF, de Toro M, Garcillan-Barcia MP, Mora A, Blanco J et al. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Genet 2014;10:e1004766.

7. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504.

8. de Been M, Lanza VF, de Toro M, Scharringa J, Dohmen W et al. Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. PLoS Genet 2014;10:e1004776.

9. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E et al. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics 2017;33:475–482.

10. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A et al. PlasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 2016;32:3380–3387.

11. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–477.

12. Prjibelski AD, Vasilinetc I, Bankevich A, Gurevich A, Krivosheeva T et al. ExSPAnder: a universal repeat resolver for DNA fragment assembly. Bioinformatics 2014;30:i293–i301.

13. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29:1072–1075.

14. Harrison PW, Lower RP, Kim NK, Young JP. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends Microbiol 2010;18:141–148.

15. 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.

16. Forde BM, Ben Zakour NL, Stanton-Cook M, Phan MD, Totsika M et al. The complete genome sequence of *Escherichia coli* EC958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b:H4-ST131 clone. PLoS One 2014;9: e104400.

17. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 2015;23:110–120.

Supplementary Results

Detailed description of genomes considered as positive controls (S1)

The following genomes were previously analyzed by the authors of PlasmidSPAdes and Recycler to validate their algorithms (9, 10).

Escherichia coli JJ1886

Recycler predicted the sequence of seven possible plasmids from *E. coli* JJ1886. Four of them corresponded to the reference plasmids whereas three sequences did not map either the plasmid references nor the chromosome of *E. coli* JJ1886. These three sequences were confirmed as plasmids by nucleotide BLAST although no other evidence of plasmid-related genes were found in the annotation by Prokka. In addition, we found two sequences of 42.6 kbp and 8.2 kbp corresponding to the chromosome of *E. coli* JJ1886. A best blast hit of the sequence with 42.6 kbp mapping to the chromosome suggested a phage origin. PlasmidSPAdes was able to recover a fraction of the plasmid pJJ1886 5, but plasmids pJJ1886 1 and pJJ1886 3 were not detected. Five components did not map to either the chromosome or the reference plasmids which suggested that they were putative unidentified plasmids. The components corresponding to *S. aureus* plasmids were present in a copy number of less than 1. Frequently, short-read length plasmids are in high copy number to ensure their prevalence in the next generations (11). These findings suggested these novel sequences identified in *E. coli* JJ1886 could constitute contamination during the library preparation. PlasmidSPAdes did not remove some parts of the chromosome from *S. aureus* because its coverage differed from *E. coli* JJ1886. cBar identified several plasmid sequences as chromosomal resulting in a low precision (Table S2). However, it was the program with the best recall value (0. 83) because it recovered 12 contigs (>500 bp) belonging to pJJ1886 5. PlasmidFinder detected the presence of two plasmid replicon initiator sequences corresponding to the incompatibility group IncF. Both replicons are located in the plasmid pJJ1886 5 of the contigs with a size of 8.0 kbp and 12.9 kbp.

Citrobacter freundii CFNIH1

PlasmidSPAdes detected a component with a length of 275.6 kbp, composed by 19 contigs (>1 kbp) that matched the reference plasmid pKEC-a3c. In addition, a second component composed by a single contig of 5.4 kbp and an inferred copy number of 14.1 was identified. Recycler was not able to recover the plasmid pKEC-a3c (Table S2). However, it also extracted the same novel component of 5.4 kbp with a coverage ratio of 14.1. We performed a dot-plot of the sequence against itself to observe the presence of circularization signatures at the ends. The sequence had a best blast hit corresponding to "Klebsiella oxytoca strain CAV1335 plasmid pCAV-1335-5410, complete sequence" with a length of 5.4 kbp. Annotation made by Prokka identified the presence of mobilization protein MbeC and relaxase MbeA. The sequence of 5.4 kbp predicted by PlasmidSPAdes

and Recycler is the same with a slight difference. Recycler extracted one of the repeat sequences present at the end of the contigs obtaining a final plasmid sequence of 5410 bp. However, PlasmidSPAdes extracted the plasmid sequences without removing one of the repeats. The previous findings suggested the presence of a complete plasmid sequence of 5.4 kbp which was not previously reported in *C. freundii* CFNIH1. PlasmidFinder detected the presence of two replication initiator proteins present in pKEC-a3c. The replicon sequences corresponded to the incompatibility groups IncN and IncA.

***Corynebacterium callunae* DSM 20147**

PlasmidSPAdes detected two components of 10.4 kbp and 4.2 kbp. A low precision value was obtained because the component of 10.3 kbp was composed by a single contig mapping to the chromosome (Table S2). The component of 4.2 kbp corresponded to the reference plasmid pCC1. Recycler detected exclusively the reference plasmid pCC1 whereas no false positive results were obtained. cBar obtained a low recall value because only one contig corresponding to pCC2 was correctly identified as plasmid (Table S2). PlasmidFinder was not able to locate any replication initiator sequence in the two reference plasmids present in *C. callunae* DSM 20147. The database of PlasmidFinder was constructed using replicon sequences from the family Enterobacteriaceae. Replicon sequences from Gram positive bacteria may differ and may explain the lack of true positive results for this genome.

***Rhodobacter sphaeroides* 2.4.1**

PlasmidSPAdes was able to detect a large component of 458 kbp including the five reference plasmids. However, the program was not able to separate the plasmids in different components. PlasmidSPAdes merged them in a single component due to the presence of repeated sequences frustrating the detection of each plasmid as different sub graphs. Visualization of the plasmid graph using Bandage spotted one contig containing a transposase shared in the different physical DNA units (Figure S3). Recycler was only able to detect small fractions from plasmid Ax and plasmid D whereas lack of false positive results were reported (Table S2). PlasmidFinder did not detect any plasmid replicon sequences.

***Burkholderia cenocepacia* DDS 22E-1**

This genome does not contain any reference plasmid but it is composed by three chromosomes with a size of 1.16 Mbp, 3.20 Mbp and 3.66 Mbp. PlasmidSPAdes and Recycler did not detect any plasmid sequence thus the outcomes of both programs corresponded to empty files. Additionally, PlasmidFinder did not find any replicon sequence within the chromosomes of *B. cenocepacia* DDS 22E-1. cBar predicted 1481 contigs (>500 bp) wrongly as plasmid-derived sequences.

PlasmidSPAdes structural report (S2)

The major pitfall of PlasmidSPAdes was the erroneous assignment of predicted plasmid contigs belonging to different reference plasmids into the same bin. To visualize and explain this issue we selected only genome projects with more than one reference plasmid (n=35). From these genomes, only plasmids correctly predicted by PlasmidSPAdes (n=120, recall >0.9), have been considered. For each reference plasmid we observed whether it was merged with another predicted plasmid from the same genomes. This resulted in 19 plasmids predicted in a single unique bin and 101 plasmids merged together with other predicted plasmids. Therefore, 84.2% of the well predicted plasmids were merged together with other plasmids from the same genome project.

Recycler chromosome fraction analysis (S3)

Recycler was designed to extract circular sequences from the assembly graph. Therefore, Recycler predictions also contained non-plasmid mobile genetic elements with a potential circularization signature. We further elab-orated on this extracting all the contigs (n=94) which were part of the chromosome fraction and annotated them using Prokka. First we checked the presence of genes annotated with the following keywords to assess if there were contigs with potential phage-related genes.

*grep -i -E "—capsid—head—integrase—plate—tail—fiber—coat—phage—transposase—portal—terminase—protease—lysin" *.tbl*

This allowed us to identify 22 contigs with potential phage-related genes. Furthermore, we used Phaster (PHAge Search Tool Enhanced Release), a program specially focused on the identification and annotation of (pro)phage sequences (12). From 22 contigs with potential-phage related genes, only 13 were identified as prophages. Therefore, 14% (13/94) of the contigs classified as false positives and assigned to the chromosme fraction were identified as prophages sequences. Furthermore, we observed the same phage sequence was extracted in genomes belonging to the same species. For example, we highlight the phage sequence of 41.9 kbp predicted by Recycler in *E. cloacae* strain CAV1311, *E. cloacae* strain CAV1411, *E. cloacae* strain CAV1668 and *E. cloacae* strain CAV1669 (Figure S5).

Components not mapping to the reference genomes (S4)

In this section we describe potential novel plasmids detected by PlasmidSPAdes and Recycler as plasmid com-ponents in the graph. Only components with a single contig and exceeding a minimum length of 1000 bp were analyzed. Novel contigs were further analyzed and annotated using Prokka (version 1.12-beta)(13). To identify poten-tial novel plasmids we compared these sequences to the non-redundant nucleotide database of the NCBI using BLAST. The best blast hit was extracted selecting

minimum e-value and highest bit-score as previously de-scribed (10). The completeness of the potential novel mobile genetic elements was corroborated by generating a dot-plot aligning the sequence to itself (14). The presence of the same repeated sequence at the ends of the contig suggested a potential circularization signature. We also considered the k-mer coverage ratio as a feature to identify potential small plasmids. Each contig reported in SPAdes 3.8.2 has an associated k-mer coverage, defined as $c_\kappa = c\,(l - k + 1)/l$, where c_κ = k-mer coverage, c = nucleotide coverage, l = read length and k = k-mer length. We considered the median coverage reported by PlasmidSPAdes as an estimation of the chromosome coverage. Finally, the reported k-mer coverage ratio corresponds to the k-mer coverage of a novel contig divided by its respective median coverage. This analysis is summarized in Table S3.

***Bacillus subtilis* subsp. natto BEST195**

PlasmidSPAdes and Recycler identified a single component not mapping to the reference assembly with a length of 5386 bp, present in a k-mer coverage ratio of 10 and with circularization signatures.

***Klebsiella pneumoniae* strain Kpn223**

PlasmidSPAdes identified two isolated components formed by a single contig with a length of 4.29 and 4.14 kbp. Recycler identified the same sequences but excluded one of the adjoining regions and detected another component not mapping to the reference of 3478 bp. The sequence of 4167 bp had as best blast hit "*Klebsiella pneumoniae* strain 0773 plasmid pKpn114, complete sequence" with a length of 4.21 kbp. The sequence of 4014 bp did not have any significant blast hit even though Prokka detected the presence of a Mobilization protein A and circularization signatures were present. In addition, the sequence identified by Recycler with a length of 3478 bp had a best blast hit corresponding to "*Enterobacter* sp. FY-07 plasmid pAKI40B, complete sequence". However, the k-mer coverage ratio suggested it is necessary to validate experimentally these novel plasmids to confirm them as stable residents in the host.

***Escherichia coli* JJ1886**

PlasmidSPAdes identified two components composed by a single contig not mapping to the reference assembly. The component with a length of 11.10 kbp had as best blast hit "*Staphylococcus aureus* subsp. aureus strain LA-MRSA ST398 isolate E154, complete genome". As explained in Supplementary Text 1, PlasmidSPAdes recovered part of the chromosome of *S. aureus* because its coverage differs from the host chromosome. The second isolated component identified by PlasmidSPAdes was also present in Recycler prediction with a length of 1.6 kbp. This component had as best blast hit "*Staphylococcus aureus* strain C2355 plasmid pUR2355, complete sequence" with a length of 7.6 kbp. Additionally, Recycler predicted two other isolated components with a length of 2361

and 2216 bp. Both components had best blast hits corresponding to "*Staphylococcus aureus* subsp. aureus strain LA-MRSA ST398 plasmid plinE154, complete sequence and "*Staphylococcus* sp. plasmid pST94 qacG and rep94 genes". The presence of *S. aureus* chromosome in PlasmidSPAdes prediction and a k-mer coverage ratio below 1.0 suggests this sample was contaminated with *S. aureus* DNA (9).

***Aeromonas veronii* strain AVNIH1**

PlasmidSPAdes identified two components formed by a single contig not mapping to the reference assembly of *A. veronii* strain AVNIH1. In addition, Recycler also identified the same components and it extracted one of the repeated sequences present at both ends of the contigs. The largest sequence had as best blast hit "*Aeromonas salmonicida* subsp. salmonicida strain JF2507 plasmid pAsal1D, complete sequence" with a length of 9.1 kbp. The other sequence corresponded to "Uncultured prokaryote from Rat gut metagenome metamobilome, isolate RGRH0694" with a length of 1.8 kbp. The presence of circularization signatures, a similar best blast hit length corresponding to previous report plasmids and the inferred plasmid copy numbers suggested the presence of two small complete plasmids not reported with a length of 7114 bp and 1736 bp.

***Klebsiella pneumoniae* strain AATZP**

PlasmidSPAdes identified an isolated component formed by a single contig of 4.29 kbp. Additionally, Recycler did not identify any reference plasmid present in *K. pneumoniae* strain AATZP but it also detected the same putative unidentified plasmid. Best blast hit corresponded to: "*Klebsiella pneumoniae* subsp. pneumoniae Kp13 plasmid pKP13c, complete sequence" with a length of 5.06 kbp. A similar blast hit length and the presence of circularization signatures at both ends of the sequence suggested the identification of a small cryptic plasmid with a length of 4167 bp not present in the reference assembly of *K. pneumoniae* strain AATZP.

***Klebsiella pneumoniae* strain CAV1392**

PlasmidSPAdes identified an isolated component composed by a single contig of 2.57 kbp with a k-mer coverage ratio of 0.1. Additionally, Recycler identified the same component and reporting the correct size (2495 bp). Best blast hit corresponded to "*Enterobacter* sp. W001 plasmid pR23, complete sequence" with a length of 10.49 kbp. The presence of circularization signatures at the end suggested the completeness of the plasmid.

***Citrobacter freundii* CFNIH1**

PlasmidSPAdes and Recycler identified an isolated component with a length of 5.4 kbp and with a k-mer coverage ratio of 14.1. Dot-plot of the contig to itself confirmed the presence of circularization. Furthermore, best blast hit corresponded to "*Klebsiella oxytoca* strain CAV1335 plasmid pCAV-1335-5410, complete sequence" with a length of 5.4 kbp.

Annotation made by Prokka identified the presence of relaxase MbeA. This novel plasmid had already been detected in PlasmidSPAdes original publication (10).

***Enterococcus faecium* strain ATCC 700221**

Both programs identified the presence of two components not mapping to the reference assembly with a length of 12462 bp and 5386 bp. The largest contig had as best blast hit "*Enterococcus faecium* plasmid p200B" with a length of 12.5 kbp. The sequence of 5386 had as best blast hit "Enterobacteria phage phiX174, complete genome" with the same length. Additionally, in both cases there was presence of circularization signatures. The presence of phage sequences was common in the report given by Recycler and PlasmidSPAdes due to the presence of circularity signatures and differences in the coverage with the host chromosome. The above findings suggest the identification of a novel plasmid with a length of 12462 bp.

***Klebsiella pneumoniae* strain Kpn555**

PlasmidSPAdes and Recycler identified the same components not mapping to the reference assembly. The sequence of 4048 bp had as best blast hit "*Escherichia coli* strain EC19 plasmid pEC19-1 hypothetical proteins, MobA, MobB, and MobC genes, complete cds" with a length of 4.86 kbp. The sequence with a length of 3478 bp had as best blast hit "*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578 plasmid pKPN7, complete sequence" with a length of 3.78 kbp. Finally, two sequences with a length of 2874 bp and 2798 bp were also reported. In both cases, the best blast hit corresponded to "Uncultured bacterium extrachromosomal DNA RGI00802" with a length of 2.80 kbp. Sequence annotation and circularization signatures indicated the presence of four small novel plasmids but their k-mer coverage ratio suggests experimental validation is required to discard these components as possible contamination.

***Klebsiella pneumoniae* strain PMK1**

PlasmidSPades and Recycler identified the same three components not present in the reference assembly. Contig annotation spotted the presence of plasmid-genes related in the sequence of 5640 bp. Best blast hit corresponded to "Uncultured prokaryote from Rat gut metagenome metamobilome, plasmid pRGRH1815" with a length of 7.10 kbp. The presence of circularization signatures and a high k-mer coverage ratio suggested the characterization of a plasmid with a length of 5640 bp. In addition, the sequence with a length of 3770 bp presented circularization signatures and a best blast hit corresponding to "*Escherichia coli* strain NCTC 9034 plasmid pEC34A, complete sequence". Additionally, cir-cularization signatures were present indicating the completeness of the plasmid. The previous findings suggested the presence of a plasmid with a length of 3770 bp. Finally the sequence with a length of 5386 bp had as best blast hit "*Echinostoma caproni* genome assembly *E caproni* Egypt, scaffold ECPE contig0001929". Several blast hit results with

a similar bit-score indicated the presence of a phage. This may explain the presence of circularization signatures at the ends of the sequence.

***Enterobacter cloacae* ECR091**

PlasmidSPAdes and Recycler identified the same isolated component formed by a single contig with a length of 4.6 kbp. Best blast hit corresponded "*Salmonella enterica* subsp. enterica serovar Typhimurium str. U288 plasmid pSTU288-3, complete sequence" with the same length. Contig annotation spotted the presence of a mobilization protein (MbeC) and circularization signatures confirmed the completeness of the plasmid. A k-mer coverage ratio of 11.8 suggests the presence of a novel plasmid with a length of 4667 bp not present in the reference assembly. Additionally, PlasmidSPAdes identified another isolated component with a contig length of 2572 bp. Best blast hit corresponded to "*Enterobacter agglomerans* ColE1-like plasmid RNA one modulator (rom) gene, complete cds" with a length of 2.49 kbp. A k-mer coverage ratio of 22.0 and the presence of circularization signatures suggests the identification of a novel plasmid.

***Enterobacter cloacae* ECNIH3**

PlasmidSPAdes and Recycler identified the same component with a contig length of 2.49 kbp. This component is the same identified by PlasmidSPAdes in the isolate *E. cloacae* ECR091. In this case, the k-mer coverage ratio is even higher 30.9 and Recycler reported the correct size of the plasmid (2495 bp).

***Klebsiella oxytoca* KONIH1**

PlasmidSPAdes identified an isolated component with a contig length of 3.71 kbp. Best blast hit corresponded to "*Enterobacter asburiae* strain CAV1043 plasmid pCAV1043-10, complete sequence" with a length of 10.40 kbp. The presence of circularization signatures and a high k-mer coverage ratio suggests the presence of a novel plasmid not present in the reference assembly.

***Klebsiella pneumoniae* strain KPNIH39**

PlasmidSPAdes and Recycler identified the same sequence with a length of 5521 bp. Best blast hit corresponded to "*Enterobacter cloacae* plasmid pNE1280, complete sequence". Presence of circularization signatures and a k-mer coverage ratio of 9.1 indicated the presence of a small cryptic plasmid with a length of 5521 bp.

Supplementary Methods

Evaluation metrics

We predicted plasmids from short reads with four different programs: PlasmidFinder, cBar, Recycler and PlasmidSPAdes. Reads were downloaded from the SRA database using the sra-toolkit and subsequently trimmed using seqtk with the command 'trimfq'. This trimmed low-quality bases from both ends using the Phred algorithm, which uses

base error probabilities calculated from the phred quality values. We selected an error probability cutoff value of 0.05. *De novo* assembly was performed using SPAdes 3.8.2 on a high performance computer running CentOS7. For each sample, the assembly graph and resulting contigs corresponding to the maximum k-mer used by SPAdes 3.8.2 were selected (1). Contigs with a size less than 500 bp were filtered out.

- PlasmidFinder. To replicate results that would be obtained through the use of the PlasmidFinder web interface, we downloaded the PlasmidFinder database containing 121 replicon sequences (updated on 16 March 2016) from the Center for Genomic Epidemiology (https//cge.cbs.dtu.dk//services/data.php). We then performed nucleotide BLAST (NCBI-BLAST version 2.2.28+) searches against this database (2). Contigs were identified as plasmids if they had a minimum identity of 80% and covered at least 60% of the replicon sequence, consistent with the parameters used to identify plasmids in bacterial whole-genome data by the authors of PlasmidFinder (3). Contigs in which a replicon sequence was identified were considered as PlasmidFinder prediction.
- cBar. We downloaded cBar version 1.2 from http://csbl.bmb.uga.edu/ ffzhou/cBar/cBar.1.2.tar.gz and used it to categorize contigs derived by SPAdes 3.8.2. Contigs categorized as plasmid-derived were con-sidered as cBar prediction.
- Recycler. We downloaded Recycler (single version, date: 07-03-2016) from https://github.com/Shamir-Lab/Recycler. The BAM file required as input by Recycler was created by alignment of the trimmed reads against the resulting contigs using Bwa 0.7.12 (4) and samtools 1.3.1 (5). Cycles reported in the assembly graph were considered as Recycler prediction.
- PlasmidSPAdes. We run PlasmidSPAdes (packaged in SPAdes 3.8.2) with standard parameters. The components reported in contigs.fasta were considered as PlasmidSPAdes prediction.

Measures for the evaluation

We evaluated the performance of each program regarding accuracy and completeness. Quast (version 4.1) (6) was used to assign each of the predicted contigs to one of the following three categories: Plasmid, chromosome or novel sequences fraction. Quast was run with the following command to map the prediction against the reference genome (chromosome and plasmid(s)) for each genome project:

python quast.py contigs.fasta -R all references.fasta -o quast analysis genome –min-alignment 500 –ambiguity-usage all

Quast takes as default a minimum identity alignment (IDY%) of 95%, all the alignments with less than this threshold are thus discarded. A minimum alignment of 500 bp was

considered to assign a contig to a specific genome. We defined the argument "all" in the flag –ambiguity-usage to report all equally good alignments of a contig (e.g. a transposase present in the chromosome and in a reference plasmid). Equally good alignments are defined in Quast using the following definition:

'all alignments are sorted by decreasing LENxIDY% value. All alignments with LENxIDY% less than Sxbest(LENxIDY%) are discarded. S should be between 0.8 and 1.0. The default value is 0.99.'

Accordingly, we encountered different scenarios:

1. Contigs with a single alignment to a reference sequence (most frequent). These contigs could be directly assigned to: Plasmid fraction (true positive result) or Chromosome fraction (false positive result).
2. Contigs without any significant alignment against the reference genome. These contigs were assigned to the fraction of novel sequences. These contigs can be the result of contamination during a sequencing project as it is apparent in *Escherichia coli* JJ1886 and JJ1887. But, some contigs corresponded to small plasmids not present in the PacBio assemblies as further elaborated in Supplementary Results and Table S3.
3. Contigs mapping to the chromosome and to a reference plasmid(s). This scenario could be further divided in two cases:

- Contigs with repeated sequences e.g. transposases which are present in the chromosome but also in a reference plasmid. In the Quast output, we observed two alignments (same length and score mapping) to two different sequences. In these cases, we assigned these contigs only to the Plasmid fraction. We considered these contigs only as true positive results because all the base pairs from the contig were present in a reference plasmid.
- Contigs containing missamblies. A missassembly corresponds to cases where different regions of the same contig are mapping to different locations of a reference genome(s). Quast defines different kind of missamblies: local missassemblies, relocations and translocations. Local missassemblies and relocations correspond to contigs where the left and right flank are mapping to different parts of the same reference sequence. Therefore, contigs can be assigned either to Plasmid fraction or Chromosome fraction because the missassembly occurs in the same reference sequence. For precision calculation purposes, we only decided to filter out contigs with a translocation event between the chromosome and a reference plasmid. This means, a single contig had two parts (left and rigth flanks; with a minimum alignment of 500 bp and passing Quast score) mapping e.g. left flank to the chromosome and right flank to a reference plasmid. In these cases, we could not assigne the whole contig to either the Plasmid fraction or the Chromosome fraction and we removed

these contigs to calculate precision.

Icarus (packaged in Quast 4.1) (7) was used to visualize the alignments between the reference genomes and the predicted sequences. We defined the previously introduced terms as:

- Plasmid fraction. Fraction of the prediction that matched the reference plasmids (true positive predic-tion).
- Chromosome fraction. Fraction of the prediction that matched the reference chromosome (false positive prediction). This fraction can include non-plasmid mobile genetic elements from the chromosome such as phages or transposable elements.
- Fraction of novel sequences. Fraction of the prediction not mapping to either the reference plasmid or the chromosome, thus corresponding to contigs absent from the reference assembly.

The programs were further evaluated using the following metrics.

- Recall was defined as the fraction of the reference plasmid(s) covered by the prediction. On the individual plasmid level, a recall of 1 indicates that the full sequence of the reference plasmid was present among the predicted plasmids. On the whole genome level, a recall of 1 indicates all reference plasmids were fully present among the predicted plasmids. However, recall does not consider whether predicted plasmid contigs were correctly binned.
- Precision. We defined precision as: Plasmid Fraction / (Plasmid Fraction + Chromosome Fraction)

The fraction of novel sequences was ignored when calculating precision. Total contig length was considered to estimate recall and precision. For each genome project (n=42) we calculated precision and recall values. To calculate the overall precision and recall of PlasmidSPAdes, Recycler and PlasmidFinder we excluded the negative control *B. cenocepacia* strain 22E-1. The overall precision and recall of cBar was calculated considering *B. cenocepacia* strain 22E-1 as the program detected a high number of contigs corresponding to false positive results (Supplementary Text 1). Finally, the overall precision and recall of PlasmidFinder was calculated filtering out all the genome projects from gram-positive bacteria. Scaffold linkage of specific contigs in the PlasmidSPAdes assembly graph of a selection of genomes was visualized with Bandage (version 0.7.1) (8). The workflow (Figure S2) was written in bash and python (version 2.7) and subsequent analysis done in R (version 0.99.982).

Supplementary Figures

Figure S1: Overview of reference plasmids. The size of the reference plasmids is shown for each bacterial isolate. Strains were sorted based on their total plasmid length. *K. oxytoca* strain CAV1374 was the most complex isolate with eleven plasmids ranging from 1.9 to 332.9 kbp.

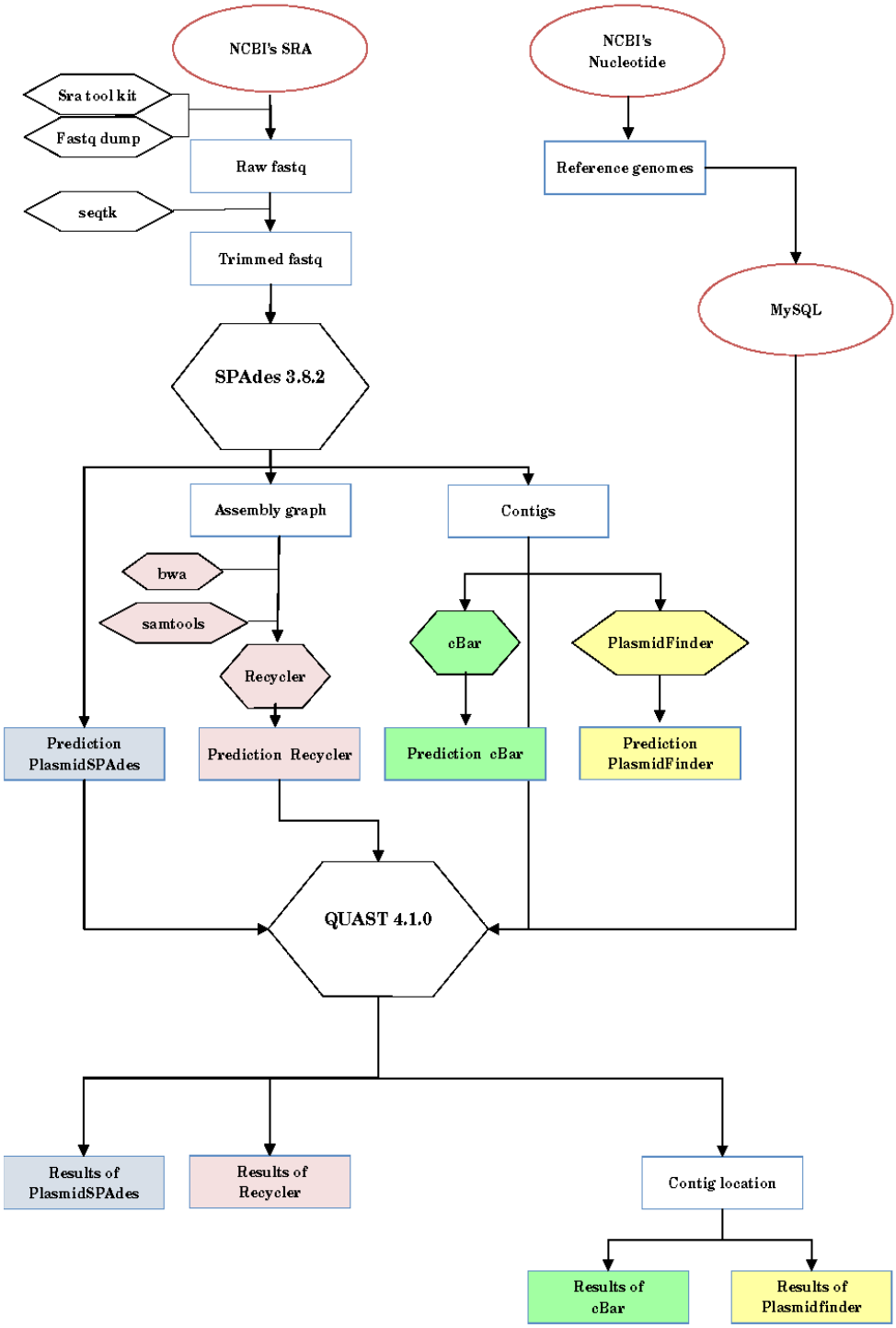


Figure S2: Analysis Workow. Diagram representing the analysis reported in the manuscript.

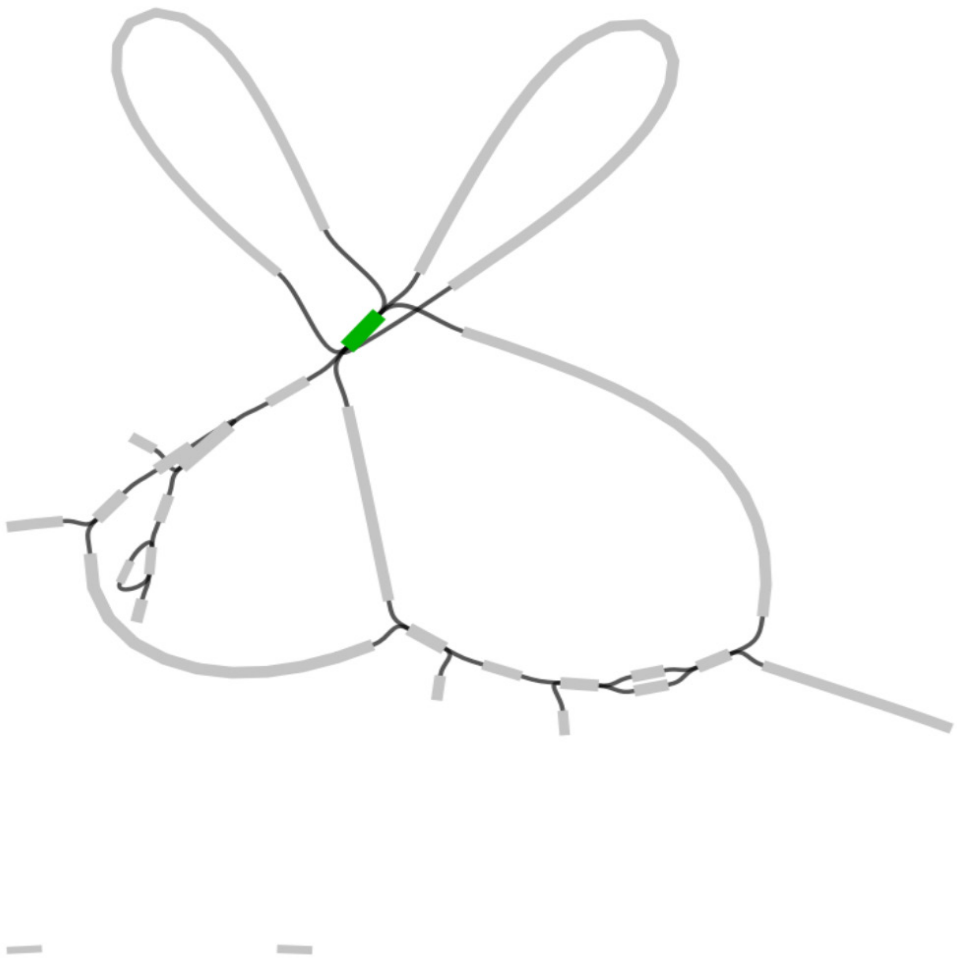


Figure S3: Bandage representation of the assembly graph generated by PlasmidSPAdes in *R. sphaeroides* 2.4.1. Green contig with a length of 1.6 kbp and coverage of 536 was identied as a transposase by BLASTx.

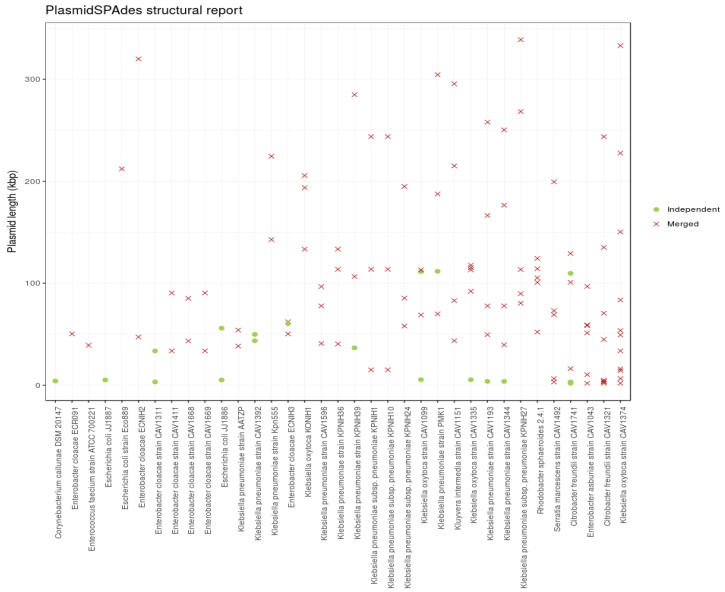


Figure S4: PlasmidSPAdes structural report. Plasmids correctly predicted by PlasmidSPAdes (recall >0.9) were further evaluated to assess if they were merged in a single unique bin with other reference plasmids. Only genome projects with more than one reference plasmid (n=35) have been considered.

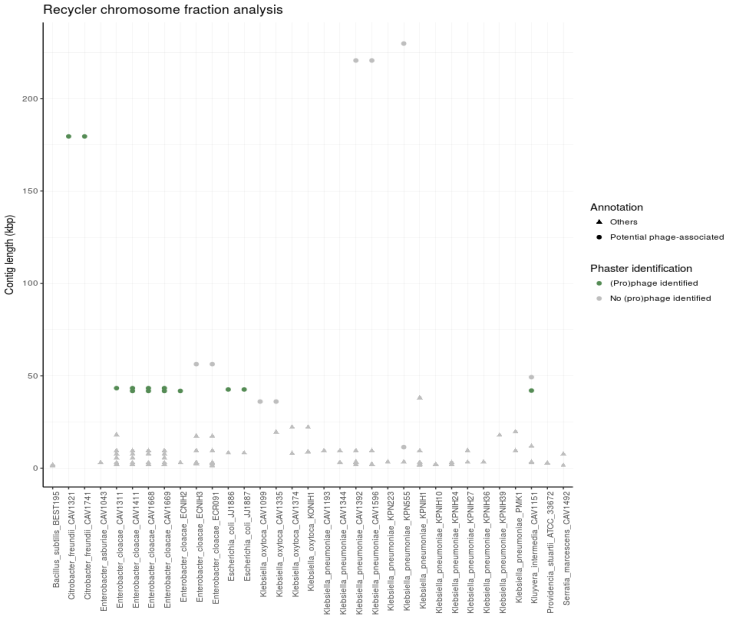


Figure S5: Recycler chromosome fraction analysis. Contigs assigned to the chromosome fraction were annotated by Prokka and queried against Phaster to observe phage sequences. Strains were sorted alphabetically (x-axis) to group genome projects belonging to the same species.

Table S1: Bacterial genomes included in this study.

Genome	SRA	Genome accession	Number of plasmids	Range size	Total
Burkholderia cenocepacia strain DDS 22E-1	SRR1618480	GCA_000755725.1	0	-	0
Bacillus subtilis subsp. natto BEST195	DRR016448	GCA_000209795.2	1	5.8	5.8
Enterobacter aerogenes strain CAV1320	SRR2965748	GCA_001021995.1	1	13.9	13.9
Providencia stuartii strain ATCC 33672	SRR1558174	GCA_000754345.1	1	48.8	48.8
Corynebacterium callunae DSM 20147	SRR892039	GCA_000420585.1	2	4.1-85.0	89.132
Enterobacter cloacae strain CAV1411	SRR2965820	GCA_001022075.1	2	33.6-90.4	124.0
Enterobacter cloacae strain CAV1669	SRR2965616	GCA_001022255.1	2	33.6-90.4	124.0
Enterobacter cloacae strain CAV1311	SRR2965815	GCA_001022015.1	3	3.2-90.4	127.2
Enterobacter cloacae strain CAV1668	SRR2965612	GCA_001022055.1	2	43.4-85.1	128.6
Klebsiella pneumoniae strain Kpn223	SRR3465557	GCA_001663435.1	1	170.9	170.9
Escherichia coli JJ1886	SRR933487	GCA_000493755.1	5	1.5-110.0	178.3
Aeromonas veronii strain AVNIH1	SRR3465535	GCA_001634325.1	1	198.3	198.3
Klebsiella pneumoniae strain AATZP	SRR3228444	GCA_001648215.1	3	38.3-121.0	213.4
Klebsiella pneumoniae strain CAV1596	SRR1582868	GCA_001022235.1	4	2.9-96.7	218.3
Klebsiella pneumoniae strain CAV1392	SRR1582895	GCA_001022035.1	3	43.6-130.7	224.1
Escherichia coli JJ1887	SRR933489	GCA_001593565.1	5	1.5-130.6	250.4
Citrobacter freundii CFNIH1	SRR1284629	GCA_000648515.1	1	272.2	272.2
Enterobacter asburiae strain CAV1043	SRR2965752	GCA_001022095.1	6	1.9-96.8	278.0
Klebsiella pneumoniae strain KPNIH36	SRR3222156	GCA_001675125.1	3	40.44-133.4	287.5
Enterococcus faecium strain ATCC 700221	SRR3176159	GCA_001594345.1	3	39.1-189.4	292.2
Escherichia coli strain ECO889	SRR3465539	GCA_001663475.1	2	88.0-212.1	300.2
Klebsiella pneumoniae subsp. pneumoniae KPNIH24	SRR1501128	GCA_000714675.1	3	58.0-194.8	338.4
Enterobacter cloacae ECR091	SRR1576808	GCA_000750275.1	3	50.3-176.9	338.5
Serratia marcescens strain CAV1492	SRR2965730	GCA_001022215.1	5	3.2-199.4	351.3
Citrobacter freundii strain CAV1741	SRR2965739	GCA_001022275.1	6	1.9-129.1	361.1
Klebsiella pneumoniae subsp. pneumoniae KPNIH1	SRR1505904	GCA_000281535.2	3	15.0-243.8	372.5
Klebsiella pneumoniae subsp. pneumoniae KPNIH10	SRR1427234	GCA_000281435.2	3	15.0-243.8	372.5
Klebsiella pneumoniae strain Kpn555	SRR3465562	GCA_001663455.1	3	26.4-224.4	393.7
Klebsiella oxytoca strain CAV1099	SRR2965639	GCA_001022295.1	5	5.4-113.9	412.8
Enterobacter cloacae ECNIH3	SRR1576778	GCA_000750225.1	4	50.3-255.0	427.9
Klebsiella pneumoniae strain KPNIH39	SRR3217430	GCA_001663295.1	3	36.7-284.8	428.1
Klebsiella oxytoca strain CAV1335	SRR2965660	GCA_001022115.1	5	5.4-117.6	443.5
Rhodobacter sphaeroides 2.4.1	SRR522246	GCA_000273405.1	5	52.1-124.3	496.7
Citrobacter freundii strain CAV1321	SRR2965690	GCA_001022155.1	9	1.9-234.7	512.4
Klebsiella oxytoca KONIH1	SRR1501122	GCA_000714655.1	3	133.3-205.5	532.7
Klebsiella pneumoniae strain CAV1344	SRR1582875	GCA_001022175.1	5	3.7-250.3	547.9
Klebsiella pneumoniae strain CAV1193	SRR2965672	GCA_001456135.1	5	3.7-257.9	555.5
Kluyvera intermedia strain CAV1151	SRR2965721	GCA_001022135.1	4	43.6-295.6	637.7
Enterobacter cloacae ECNIH2	SRR1515967	GCA_000724505.1	3	47.2-319.9	649.7

Table S2: Precision and recall of each program in the genome projects considered as positive controls

Strain	Program	Precision (%)	Recall (%)
E. coli JJ1886	pSPAdes	0.56	0.46
	Recycler	0.57	0.38
	PlasmidFinder	1.00	0.18
	cBar	0.34	0.84
C. freundii CFNIH1	pSPAdes	0.99	0.99
	Recycler	0.00	0.00
	PlasmidFinder	1.00	0.19
	cBar	0.68	0.88
C. callunae DSM 20147	pSPAdes	0.28	0.04
	Recycler	1.00	0.04
	PlasmidFinder	0.00	0.00
	cBar	0.71	0.09
R. sphaeroides 2-4-1	pSPAdes	1.00	0.91
	Recycler	1.00	0.07
	PlasmidFinder	0.00	0.00
	cBar	0.59	0.69

Table S3: Novel sequences not present in the reference genome predicted by PlasmidSPAdes and Recycler.

	pSPAdes	Recycler	cBar	k-mer coverage ratio	Blast hit	Annotation	Circularity
B. subtilis BEST195	5513	5386	Plasmid	10.3	Plasmid (CP003995)	-	✓
K. pneumoniae KPN223	4294	4167	Plasmid	0.9	Plasmid (EU932690)	-	✓
	4141	4014	Plasmid	1.5	Non significant	Mob. protein MobA	✓
	-	3478	Plasmid	1.3	Plasmid(NZ CP012489)	-	✓
E. coli JJ1886	11105	-	Chromosome	0.2	Chromosome (CP013218)	-	X
	-	2361	Plasmid	1.0	Plasmid (CP014694)	-	✓
	-	2216	Plasmid	0.8	Plasmid (Y16944)	-	✓
	1689	1634	Plasmid	0.2	Plasmid (JQ312422)	-	✓
A. veronii AVNIH1	7241	7114	Plasmid	6.3	Plasmid (KT781681)	Antitoxin RelB	✓
	1863	1736	Chromosome	15.7	Plasmid (LN853312)	-	✓
K. pneumoniae AATZP	4294	4167	Plasmid	2.4	Plasmid (CP003995)	-	✓
K. pneumoniae CAV1392	2572	2495	Plasmid	0.1	Plasmid (NC 015515)	-	✓
C. freundii CFNIH1	5487	5410	Plasmid	14.1	Plasmid (NZ_CP011613)	Relaxase MbeA	✓
E. faecium ATCC 700221	12589	12462	Plasmid	2.7	Plasmid (AB158402)	-	✓
	5513	5386	Plasmid	26.6	Phage (CP004084)	-	✓
K. pneumoniae KPN555	4175	4048	Plasmid	0.4	Plasmid (JX238446)	Relaxase MbeA	✓
	3605	3478	Plasmid	0.9	Plasmid (CP000652)	Antitoxin MazE	✓
	3001	2874	Plasmid	1.7	Plasmid (HG796369)	Plasmid recombination enzyme	✓
	2925	2798	Plasmid	2.0	Plasmid (HG796369)	-	✓
K. pneumoniae PMK1	5695	5640	Plasmid	26.0	Plasmid (LN854314)	Antitoxin IgA-2, Mob. protein MbeC	✓
	5441	5386	Plasmid	2.0	Scaffold (LL266921)	-	✓
	3825	3770	Plasmid	35.0	Plasmid (NC 019077)	-	✓
E. cloacae ECR091	4744	4667	Plasmid	11.8	Plasmid (CP004060)	Mob. protein MbeC	✓
	2572	-	Plasmid	22.0	Plasmid (AF014880)	-	✓
E. cloacae ECNIH3	2572	2495	Plasmid	30.9	Plasmid (AF014880)	-	✓
K. oxytoca KONIH1	3713	-	Chromosome	40.7	Plasmid (CP011586)	-	✓
K. pneumoniae KPNIH39	5550	5521	Plasmid	9.1	Plasmid (NC 019346)	-	✓

Supplementary References

1. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev M a., Pevzner P a. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

3. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903.

4. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

6. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

7. Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A. 2016. Icarus: visualizer for de novo assembly evaluation. Bioinformatics.

8. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352.

9. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, Shamir R. 2016. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics btw651.

10. Antipov D, Hartwick N, Shen M, Raiko M, Pevzner PA. 2016. plasmidSPAdes : Assembling Plasmids from Whole Genome Sequencing Data. Bioinformatics 32:3380–3387.

11. San Millan A, Heilbron K, MacLean RC. 2014. Positive epistasis between co-infecting plasmids promotes plasmid survival in bacterial populations. ISME J 8:601–612.

12. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21.

13. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

14. Krumsiek J, Arnold R, Rattei T. 2007. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 23:1026-1028.

3

mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species

Sergio Arredondo-Alonso, Malbert R. C. Rogers, Johanna C. Braat, Tess D. Verschuuren, Janetta Top, Jukka Corander, Rob J. L. Willems, and Anita C. Schürch

Published in: Microb. Genom . (2018) doi: 10.1099/mgen.0.000224

Abstract

Assembly of bacterial short-read whole-genome sequencing data frequently results in hundreds of contigs for which the origin, plasmid or chromosome, is unclear. Complete genomes resolved by long-read sequencing can be used to generate and label short-read contigs. These were used to train several popular machine learning methods to classify the origin of contigs from *Enterococcus faecium*, Klebsiella pneumoniae and *Escherichia coli* using pentamer frequencies. We selected support-vector machine (SVM) models as the best classifier for all three bacterial species (F1-score *E. faecium*=0.92, F1-score *K. pneumoniae*=0.90, F1-score *E. coli*=0.76), which outperformed other existing plasmid prediction tools using a benchmarking set of isolates. We demonstrated the scalability of our models by accurately predicting the plasmidome of a large collection of 1644 *E. faecium* isolates and illustrate its applicability by predicting the location of antibiotic-resistance genes in all three species. The SVM classifiers are publicly available as an R package and graphical-user interface called 'mlplasmids'. We anticipate that this tool may significantly facilitate research on the dissemination of plasmids encoding antibiotic resistance and/or contributing to host adaptation.

Introduction

Plasmids are autonomous extra-chromosomal elements that can act as major drivers of variation and adaptation in bacterial populations (1, 2). Plasmids can also facilitate the dissemination of antimicrobial resistance via horizontal transfer of resistance genes, such as plasmid-derived vancomycin resistance in *E. faecium* or extended-spectrum β-lactamase in *Enterobacteriaceae* isolates (3–6). This means that understanding plasmid epidemiology is pivotal to fully understand the introduction and transmission of antimicrobial resistance in bacterial populations (7, 8).

Analysing the plasmid content of large collections of isolates by PCR-based techniques is laborious and has low resolution. Illumina sequencing platforms, which provide short reads (ranging from 150 to 300 bp) with low error rates, have been massively used to assemble bacterial draft genomes (9). However, the frequent presence of insertion-sequences (IS) and transposable elements in bacterial genomes prohibit their full assembly, because these regions cannot be spanned by short-reads (7, 10). This results in a fragmented assembly typically consisting of hundreds of chromosomal and plasmid contigs that challenge the inference of the origin of these contigs.

Different tools (PlasmidFinder, cBAR, Recycler, PlasmidSPAdes, PlasFlow) have been proposed to automate the reconstruction of plasmids using short-read whole-genome sequencing (WGS) data (11–15). However, plasmid predictions are usually incomplete and chromosome-derived contigs are frequently present among the predicted plasmids (16). This may be partially overcome using tools such as PlacnetW, which allows users to define and solve plasmid boundaries, but limits the high-throughput analysis of short-read WGS data (17, 18).

Long-read WGS has emerged as a solution to obtain complete and error-free plasmid sequences (19, 20). Read lengths generated by these platforms allow the complete spanning of repeat sequences and obtaining a single contig per replicon (21, 22). Due to the increasing number of complete genomes available in RefSeq/National Center for Biotechnology Information (NCBI) databases, we explored the possibility of training several popular machine learning algorithms using genome signatures from single-species assemblies. These features have been previously used in cBAR and recently in PlasFlow to distinguish plasmid- and chromosome-derived sequences in metagenomic samples.

Here, we present mlplasmids, a new tool to predict plasmidand chromosome-derived sequences for a selection of Gram-positive and Gram-negative bacterial species (*E. faecium*, *K. pneumoniae* and *E. coli*) with species-specific classifiers, and we show that mlplasmids outperforms other plasmid prediction tools for these three species. We have made the plasmid models available as an R package and a web-server.

Results

Diversity of complete genome sequences

To ensure that the new classifiers were built using genome sequences from a large and diverse set of isolates belonging to each species, we first assessed the diversity present in our collections of *E. faecium*, *K. pneumoniae* and *E. coli*. We used Mash to sketch and cluster all retrieved isolates from *E. faecium*, *K. pneumoniae* and *E. coli*. For *E. faecium*, we defined three main clusters and observed that our set of *E. faecium* (n=62) extended the diversity present in complete genomes in the Assembly Entrez NCBI database (n=24) (Fig. S1). Seven isolates were part of a cluster in which we did not find NCBI complete genomes. Strikingly, we observed a single unique NCBI complete genome forming an independent cluster (GCF_000737555), corresponding to *E. faecium* T110, a probiotic strain (Fig. S1). For *K. pneumoniae*, we also observed and defined three main clusters from all complete genomes available in the Assembly Entrez NCBI database (n=156). One of the three clusters was only composed of three *K. pneumoniae* isolates (GCA_000714635, GCF_000019565 and GCF_002156765) and showed a Mash distance higher than 0.05 versus isolates present in the other two major clusters (Fig. S2). For *E. coli*, we observed three major clusters of isolates present in the *E. coli* NCBI collection. All defined *E. coli* clusters presented a high diversity versus each other in terms of Mash distances (Fig. S3).

Pentamer frequencies differentiate between plasmid- and chromosome-derived sequences in single species

We investigated the applicability of genomic signatures to distinguish between plasmid- and chromosome-derived sequences by calculating the pentamer frequencies from complete chromosomal and plasmid sequences of *E. faecium*, *K. pneumoniae* and *E. coli* available in the NCBI database. We then transformed pentamer frequencies into a distance matrix and clustered complete sequences based on their pentamer profile. We observed that pentamer frequencies provided a clear separation between plasmid and chromosome sequences (Figs 2 and S4). However, we observed that chromosome sequences for each species were clustering independently, which suggested that pentamer frequencies differed between bacterial species. Additionally, we observed that plasmid sequences from *E. coli* and *K. pneumoniae* were clustering together, which indicates that plasmids from these two species share a high fraction of k-mers that might be a result of potential plasmid transmission between both species (Fig. 2). We concluded that pentamer frequencies could be used as classifier features to distinguish chromosome and plasmid sequences for single species. In addition, we decided to use exclusively pentamer frequencies for several reasons: (i) the optimal ratio between the number of objects and features (~10) to avoid overfitting problems of the plasmid models due to increase of model complexity, (ii) fast and robust plasmid prediction allowing the

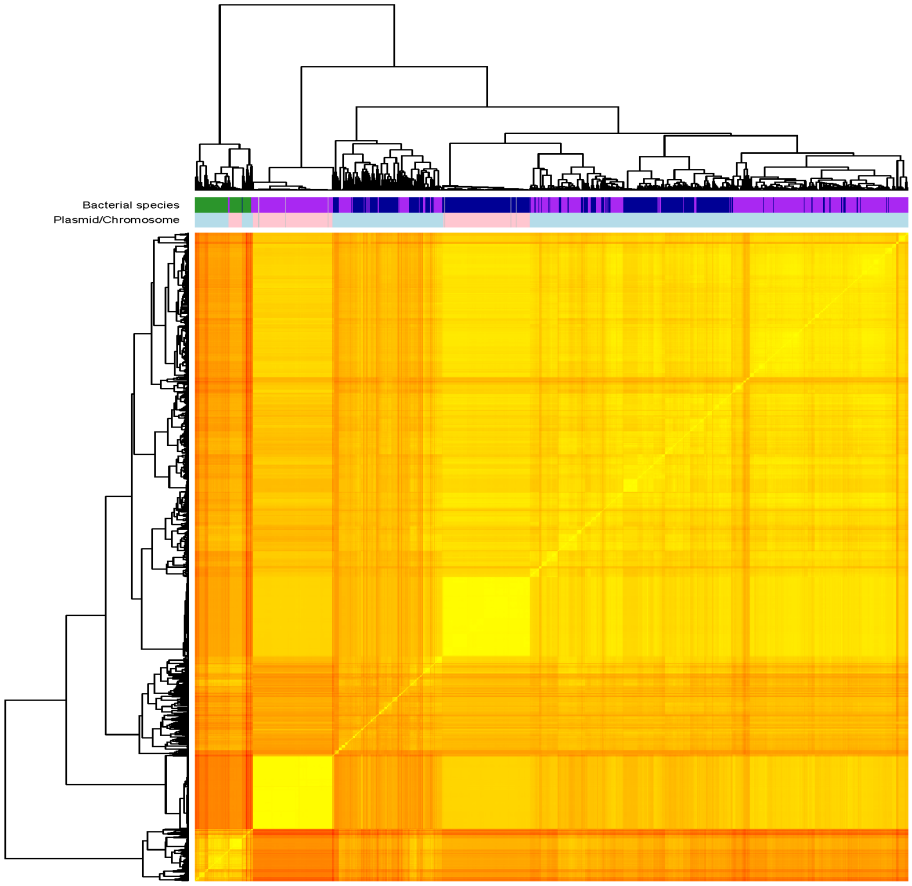


Figure 2. Ward hierarchical clustering of all chromosome and plasmid sequences from the Assembly Entrez NCBI database corresponding to *E. coli*, *K. pneumoniae* and *E. faecium* based on pentamer frequencies. Each node on the dendrogram corresponds to a either a plasmid (light blue) or chromosome (pink) sequence from *E. coli* (dark blue), *K. pneumoniae* (purple) or *E. faecium* (green).

possibility of distributing mlplasmids as a graphical-user interface, and (iii) they have been used before to distinguish plasmid sequences in metagenomic samples (12, 15).

Performance of several popular machine-learning classifiers on single species

SVM was the machine-learning algorithm selected as best classifier for predicting plasmid-derived contigs in the three bacterial species. SVM performance in *E. faecium* (accuracy=0.94; F1-score=0.92) and in *K. pneumoniae* (accuracy=0.92; F1-score=0.90) was better than the other tested machine-learning models and their F1-score, and AUC reflected that prediction of the model was balanced for both classes (Fig. 3). In the case of *E. coli*, SVM performance (accuracy=0.95; F1-score=0.76) reflected that prediction for the plasmid-class was less accurate compared to the chromosome class (sensitivity=0.71)

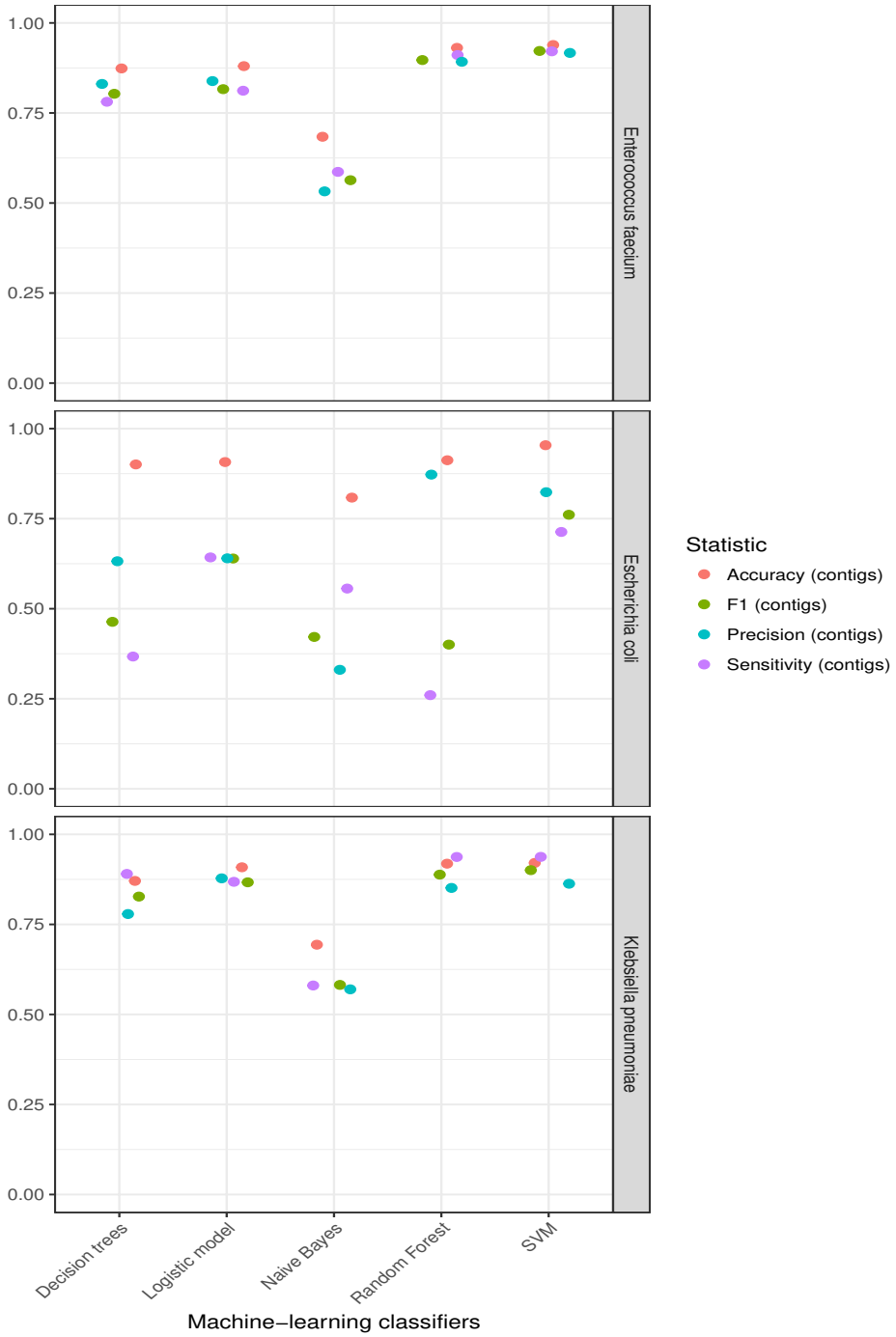


Figure 3. Performance of the optimized machine-learning classifiers. Decision trees, logistic model, Bayesian classifier (naive Bayes), RF and SVM using our test sets for *E. faecium*, *E. coli* and *K. pneumoniae*. The statistics reported are accuracy (red), F1-score (green), precision (blue) and sensitivity (purple), and are indicated using contigs as a performance measure.

(Fig. 3, Table S8). This can be explained by a lower frequency of the plasmid class (Table S4) present in the training set of the machine-learning classifiers compared to the training sets of *E. faecium* and *K. pneumoniae* or a higher diversity from isolates categorized as belonging to this species (Fig. S4). For the three selected SVM models, we observed that metrics reported were higher when considering base pairs as the unit (Table S8). This indicated that misclassification mostly occurred on short length contigs (<1 kbp) as shown for the *E. faecium* SVM model (Fig. S5).

For the *E. faecium* training and test sets, we checked the presence of chromosome-labelled contigs corresponding to putative integrated plasmids. We observed a low frequency of these contigs (n=10 contigs, frequency=0.1). We did not remove them to avoid overfitting problems. After predictions, we observed that the *E. faecium* model predicted two of these contigs as plasmid derived. These two contigs had a small contig length (1.47 and 2.3 kbp). Longer contigs (n=8, mean contig length=11.16 kbp) were predicted as chromosome derived. We implemented *E. faecium*, *K. pneumoniae*, *E. coli* SVM models in a new R package called mlplasmids.

Benchmarking mlplasmids against existing plasmid prediction tools

We benchmarked mlplasmids against other fully automated plasmid prediction tools using the isolates described in Methods in the section 'Selection of isolates for benchmarking' (*E. faecium*=7, *K. pneumoniae*=11, *E. coli*=3). Performance of mlplasmids in *E. faecium* (F1-score=0.94, precision=0.95) was higher than cBAR (F1-score=0.53, precision=0. 46), PlasFlow (F1-score=0.71, precision=0.61) and PlasmidSPAdes (precision=0.61) (Figs 4 and 5). For *E. coli*, mlplasmids performance was superior (F1-score=0.84, precision=0.88) compared to cBAR (F1-score=0.50, precision=0.4), PlasFlow (F1-score=0.58, precision=0.42) and PlasmidSPAdes (precision=0.6). In the case of *K. pneumoniae*, mlplasmids metrics were overall better (F1-score=0.88, precision=0.86) even though performance of PlasFlow (F1-score=0.82, precision=0.72) and PlasmidSPAdes (precision=0.79) was also good, and in the case of *K. pneumoniae* strain KPN555 performance was better compared to mlplasmids (Fig. 5).

The mean genome fraction values of mlplasmids for *E. faecium* (81.5%), *K. pneumoniae* (82.3) and *E. coli* (83.7%) indicated that most of the bases from the reference plasmids were covered in the prediction by mlplasmids, even though contigs with a contig length smaller than 1000 bp were filtered out (Fig. 5b). For *K. pneumoniae*, the overall genome fraction of PlasFlow (83.1) was higher than for mlplasmids, but precision (0.72) indicated that a fraction of chromosomal contigs was wrongly predicted as plasmid (Fig. 5a). We further compared mlplasmids and PlasFlow predictions showing the potential of mlplasmids unravelling the origin of contigs unclassified by PlasFlow (Supplementary Results S1, Figs. S6 and S7).

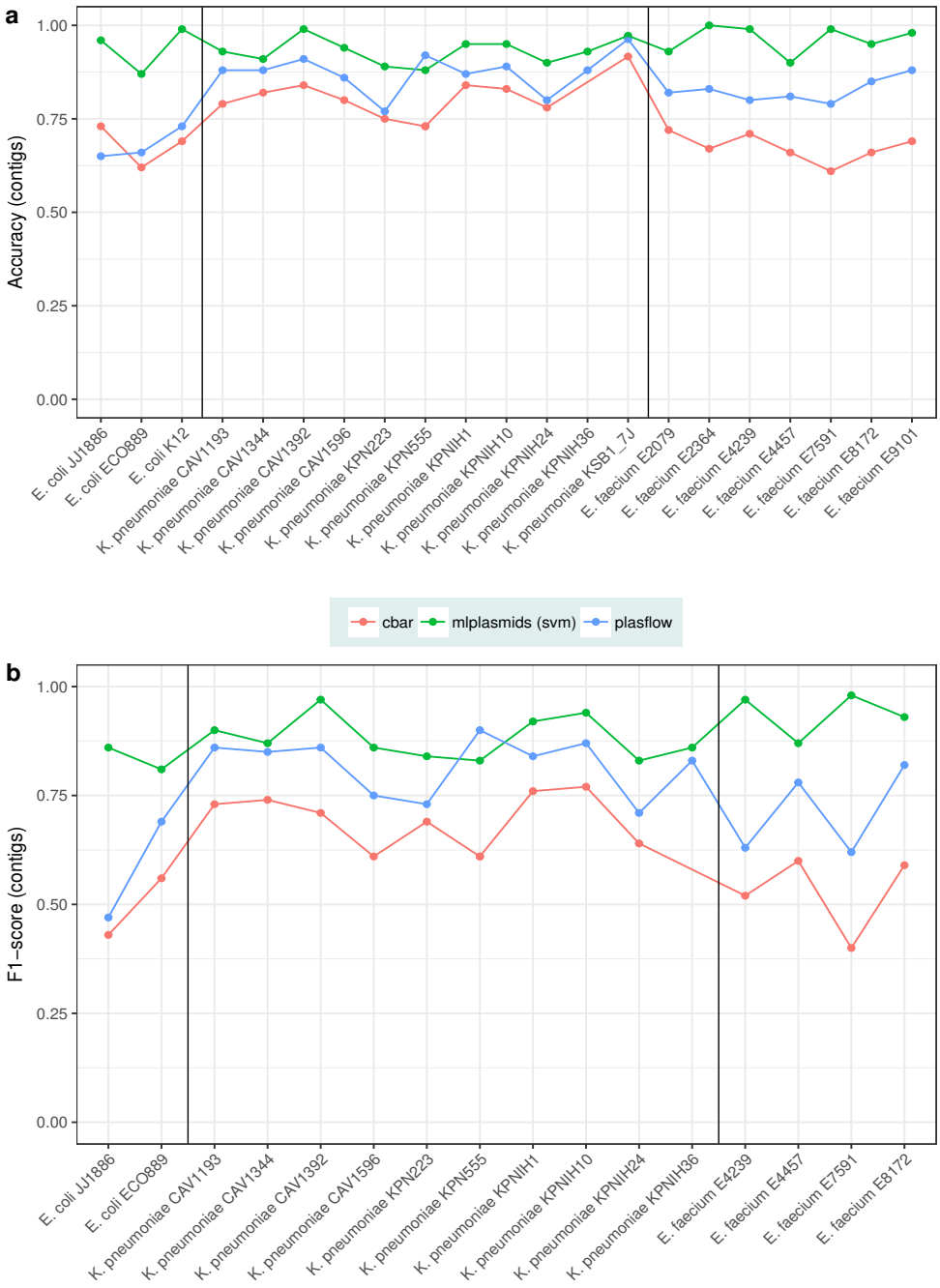


Figure 4. Benchmarking of cBAR (red), mlplasmids (green) and PlasFlow (blue) using an independent set of isolates (n=20). (a) Accuracy was measured in contigs and reported for all isolates including samples considered as negative controls (*E. coli* K. 12, *K. pneumoniae* KSB1_7J, *E. faecium* E2079, *E. faecium* E2364 and *E. faecium* E9101). (b) The F1-score was measured in contigs and only reported for isolates bearing plasmids (n=16).

Our approach of training the classifiers on datasets from single species was fundamental to obtain a good precision. This was also reflected in mlplasmids prediction for isolates corresponding to negative controls. For *E. coli* strain K-12 and *K. pneumoniae* KSB1_7J, only a single contig (>1000 bp) was erroneously predicted as plasmid derived. We also observed similar very low numbers of false-positive plasmid assigned contigs for *E. faecium* E2079 (n=6) and *E. faecium* E9101 (n=1), and for *E. faecium* E2364 (n=0) all chromosome-derived contigs were correctly predicted (Fig. 4a).

Predicting plasmids acquired by horizontal gene transfer

To assess the applicability of mlplasmids detecting plasmids acquired from related species, we considered all the plasmid-derived contigs described in Methods in the section 'Selection of isolates for benchmarking'. For each dataset of *E. coli*, *K. pneumoniae* and *E. faecium* contigs, we predicted the origin of contigs using all three models available in mlplasmids. As expected, for each dataset the best model to predict chromosome- and plasmid-derived contigs corresponded to the mlplasmids model from the same species (Fig. S8). However, we recovered most of the *E. coli* plasmid contigs (96%) when using the *K. pneumoniae* model and with an associated high probability of belonging to the plasmid class (mean=0.80) (Fig. S8c). We also observed a similar situation when predicting *K. pneumoniae* plasmid contigs with our *E. coli* model, in which plasmid-derived contigs were detected with a high probability of belonging to that class (mean=0.82) but only 57% of plasmid-derived contigs were assigned to this category, which could be explained by a lower prevalence of plasmid contigs present during the training of the *E. coli* model (Fig. S8b). This analysis suggested that mlplasmids can correctly predict plasmid sequences transferred to *E. coli* or *K. pneumoniae* coming from a related bacterial species as a result of a horizontal gene transfer event.

However, when using the *E. faecium* model against the *K. pneumoniae* and *E. coli* dataset, we obtained a high number of false-negative contigs and plasmid-derived contigs had a low probability (*K. pneumoniae* mean=0.59; *E. coli* mean=0.62) of belonging to the assigned class (Fig. S8a). This highlighted that pentamer frequencies between chromosome- and plasmid-derived contigs differ between nonrelated species. Additionally, we refuted the possibility that all sequences predicted with a particular model, but coming from another bacterial species, would have been exclusively assigned to the plasmid class (Fig. S8).

Applicability for predicting sequences derived from incomplete long-read assemblies

To rule out misclassification of complete plasmid sequences as chromosomal due to a possible correlation of pentamer frequencies and contig length, we evaluated the performance of mlplasmids with chromosomal and plasmid sequences with a sequence length higher than the mean contig size used during the training of

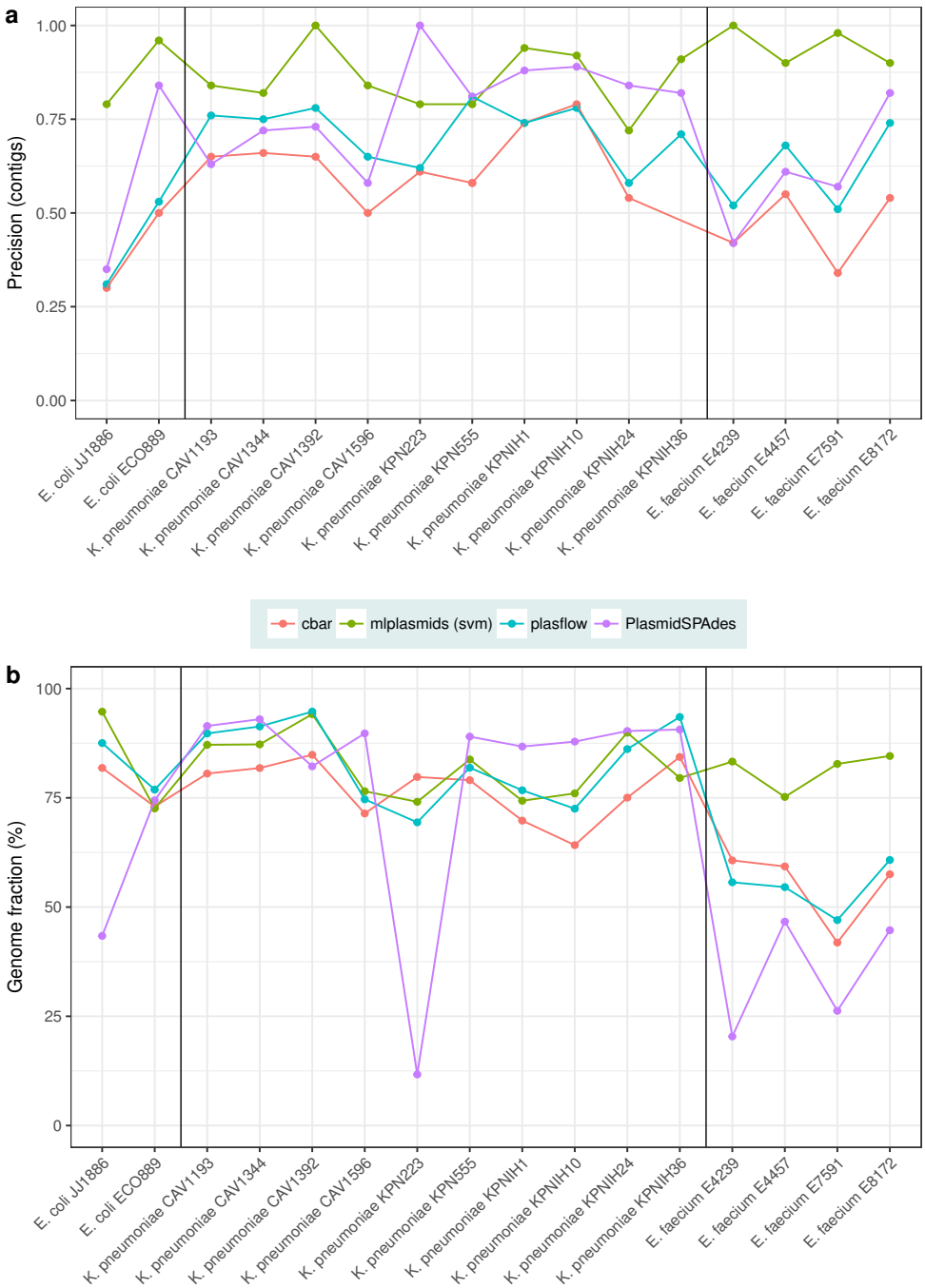


Figure 5. Comparison of cBAR (red), mlplasmids (green), PlasFlow (blue) and PlasmidSPAdes (purple) using an independent set of isolates. (a) Precision was measured in contigs and reported only for isolates bearing plasmids (n=16). (b) Genome fraction (measured as percentage of base pairs) was extracted from Quast analysis for isolates bearing plasmids (n=16).

mlplasmids classifiers. We predicted complete genome sequences from *E. faecium* (chromosomes=24; plasmids=82), *K. pneumoniae* (chromosomes=11; plasmids=33) and *E. coli* (chromosomes=3; plasmids=7). The observed mlplasmids performance for *E. faecium* (F1-score=0.99), *K. pneumoniae* (F1-score=0.98) and *E. coli* (F1-score=0.92) suggested that mlplasmids can also be used to predict these large contigs correctly. This demonstrates the flexibility of mlplasmids to predict sequences with different lengths compared to the mean contig length used to train the mlplasmids models. Consequently, mlplasmids may facilitate the classification of contigs generated from incomplete hybrid or long-read assemblies as exemplified for isolate *E. faecium* E7070 (Supplementary Results S2, Fig. S9).

Applicability for predicting the location of antibiotic-resistance genes

To show the potential of mlplasmids in determining whether a particular gene of interest is plasmid or chromosome encoded, we predicted the location of antibiotic-resistance genes in *E. faecium*, *K. pneumoniae* and *E. coli*. Firstly, we determined resistance genes in NCBI draft assemblies by using Abricate to screen contigs against the ResFinder database. Secondly, we used *E. faecium*, *K. pneumoniae* and *E. coli* SVM models in mlplasmids to determine whether these resistance genes were located in plasmid- or chromosome-derived contigs. For each identified resistance gene, we calculated the frequency of finding that particular gene on a predicted plasmid- or chromosome-derived contig.

For *E. faecium*, we assigned a total of 1058 and 1836 genes as chromosome and plasmid located, respectively. We observed that most aminoglycoside-resistance genes (e.g. ant(6)-Ia_2) were mainly present in a plasmid context (Fig. S10). Erythromycin-resistance genes were preferentially present in one genomic context depending on the gene variant as exemplified by *erm(A)_1* and *erm(B)_18* (Fig. S10). As previously described, the vanA operons were only present in plasmid-predicted contigs (37). Furthermore, vanB operons were present in both plasmid and chromosomal contexts, but the frequency of chromosome-derived contigs was higher (0.73) (Fig. S10) (38). Validation of the prediction on *E. faecium* isolates excluded from the model training (n=7) revealed that all resistance genes (n=43) predicted by Abricate were correctly predicted either as plasmid or chromosome derived (F1- score=1.0).

For *K. pneumoniae*, we assigned a total of 5107 and 10432 ResFinder hits as chromosome and plasmid located, respectively. Most of the antibiotic-resistance genes showed a clear tendency of being present in either a plasmid or chromosomal genomic context (Fig. 6). As described before (39), we observed some notable exceptions, such as *armA* or $bla_{\text{CTX-M-14\_1}}$, in which these particular resistance genes were also present in predicted chromosome-derived contigs (Fig. S11a). We performed the same analysis on *K. pneumoniae* isolates belonging to the independent set (n=10) resulting in a total of 41 and 75 genes predicted as plasmid and chromosome encoded, respectively. Mlplasmids evaluation revealed that

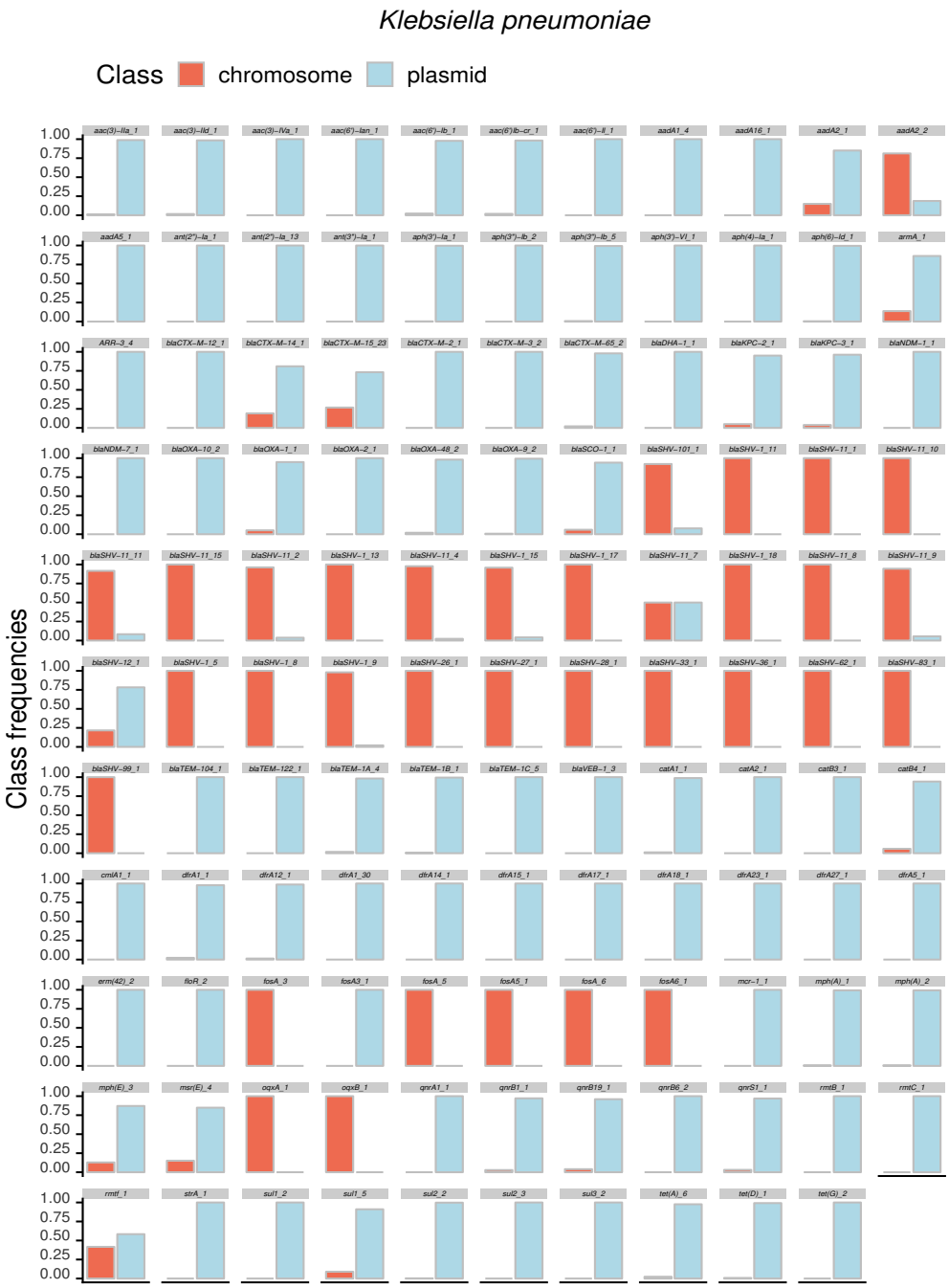


Figure 6. *K. pneumoniae* resistome. Draft genomes available in NCBI Genomes FTP (n=1346) were downloaded and screened using Abricate and ResFinder for the presence of antibiotic-resistance genes. Each contig containing a resistance gene was predicted with mlplasmids to have plasmid or chromosome origin. For visualization purposes, only antibiotic-resistance genes present more than five times are shown.

all predicted plasmid-encoded genes were correctly assigned (precision=1.0) and only five genes were misclassified as chromosome encoded (F1-score=0.96, sensitivity=0.93).

For *E. coli*, we assigned a total 4517 and 8085 ResFinder hits as chromosome and plasmid located, respectively. In contrast to *K. pneumoniae*, we observed that resistance genes were frequently identified in both plasmid and chromosomal contexts (Fig. S12). We also observed differences in gene location between resistance gene variants as exemplified for *qnrS2_1*, which was frequently encoded in predicted plasmid-derived contigs in contrast to *qnrS1_1*, which can be found in both genomic contexts (Fig. S11b). Interestingly, *mcr-1_1* was found in both plasmid and chromosomal contexts in *E. coli*, whereas for K. pneumoniae this resistance gene was only identified in plasmid-derived contigs (Fig. S11). Chromosomal locations of *mcr-1_1* for *E. coli* have been described before (40). We predicted a total of 15 resistance genes from *E. coli* isolates that belonged to the independent set (n=3). Mlplasmids performance revealed that four genes that were encoded in a single contig from *E. coli* ECO889 were wrongly predicted as chromosome-encoded, whereas gene assignment was flawless for *Escherichia coli* JJ1886 (F1-score=0.88, sensitivity=0.80). As observed for *E. faecium* and *K. pneumoniae*, all genes predicted as plasmid encoded were correctly assigned (precision=1.0).

Applicability for predicting the plasmid content of a single species

Finally, we demonstrate the utility of mlplasmids by predicting the plasmidome content of *E. faecium*. We predicted plasmid-derived sequences from a collection of 1644 Illumina-sequenced *E. faecium* isolates (Table S6). Mlplasmids prediction using our R package

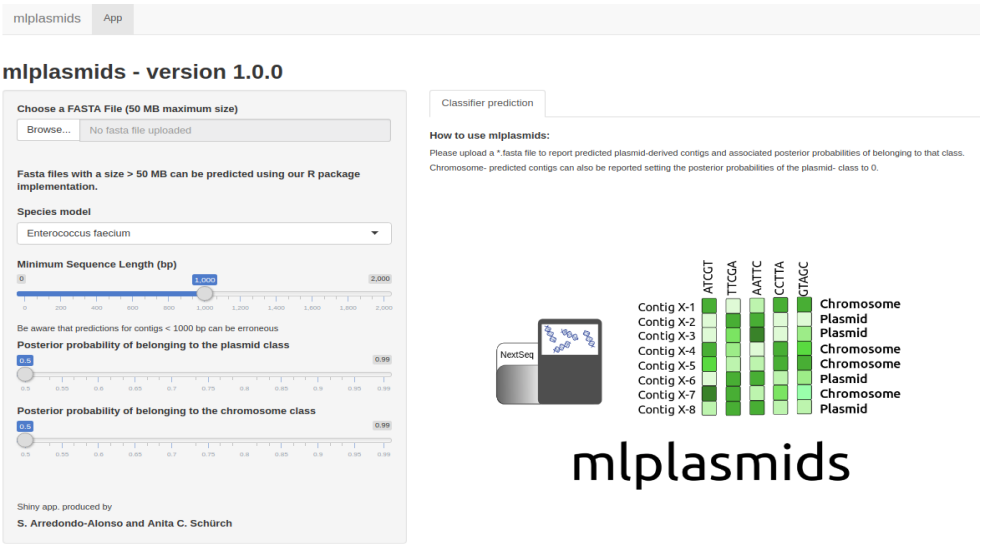


Figure 7. The mlplasmids web-server interface. To facilitate the usability of mlplasmids, we developed a Shiny app, in which users can easily upload single genome assemblies and retrieve mlplasmids prediction.

took 1624509 s (~27 min) on a Linux laptop (Ubuntu 14.04) using a single core. Classifier prediction resulted in 194884 contigs originating from the chromosome and 94485 contigs with a predicted plasmid origin in 1640 isolates. Mlplasmids did not predict any plasmid-derived contig in four strains, including one of our negative controls (*E. faecium* isolate E2364). The mean posterior probability of the predicted chromosome-derived contigs corresponded to 0.95 versus a mean posterior probability of 0.91 for plasmid-predicted contigs (Fig. S13). This suggested a high likelihood that contigs were correctly assigned to each class. We filtered out contigs with a minimum posterior probability of 0.7 of belonging either to the plasmid or chromosome class to estimate the number of plasmid- and chromosome-derived contigs per isolate. This resulted in mean numbers of ~113 chromosome- and ~52 plasmid-derived contigs per isolate. The mean cumulative length of chromosome- and plasmid- predicted contigs was 2619359 and 240324 bp, respectively, which matched with the expected *E. faecium* genome size.

To facilitate the usability of mlplasmids, we have developed a graphical-user interface in which users can upload and retrieve mlplasmids prediction of genome assemblies online (Fig. 7). As mlplasmids models use pentamer frequencies to predict plasmid-derived contigs, users can collect genome assemblies from several isolates of a single species in a single file, which can facilitate the analysis of a large collection. Assemblies can be uploaded to the web-server as tar.gz files. Users must select the species model (*E. faecium*, *K. pneumoniae* or *E. coli*) for the plasmid prediction. After uploading a genome assembly, results appear as tabular data in which each row corresponds to a sequence present in the fasta file. Additionally, results can be filtered using three options: (i) minimum sequence length to report prediction; (ii) minimum posterior probability for assignment of plasmid class; (iii) minimum posterior probability for assignment of chromosome class. Results can be downloaded in csv/xslx format.

Discussion

We present a set of species-specific machine-learning classifiers to classify plasmid-derived contigs for three clinically relevant species: the Gram-positive bacterium *E. faecium*, and the Gram-negative bacteria *K. pneumoniae* and *E. coli*. We used genomic structure information from complete genomes to label short-read contigs as plasmid- or chromosome-derived, and used them to train and test five different popular machine-learning algorithms.

Genome signatures were previously used in cBAR and more recently for PlasFlow to predict plasmid sequences from primarily metagenomes (12, 15). In contrast to cBAR and PlasFlow, we trained and benchmarked our SVM classifiers using contigs with a minimum length of 1 kbp. This is important to accurately predict contigs derived from small plasmids (length <5 kbp) or from plasmids with a high number of repeat sequences

(e.g. transposons), since this leads to a fragmented assembly with a lower mean contig length. We showed mlplasmids potential to obtain an accurate and reliable plasmidome prediction compared to cBAR and PlasFlow. Furthermore, mlplasmids' precision when predicting contigs from isolates considered as negative controls was remarkable (Fig. 4a). We have highlighted the potential of mlplasmids to classify the origin of contigs unclassified by PlasFlow. Mlplasmids also outperformed PlasmidSPAdes (Fig. 4), which relies on differences in coverage between plasmids and chromosome in the prediction of plasmid-derived contigs for single genome assemblies. Mlplasmids allows accurate prediction of contigs derived from large plasmids or linear plasmids without differences in sequencing coverage between replicons.

Mlplasmids can predict whether a particular contig is plasmid or chromosome derived. However, it is not possible to cluster plasmid contigs into different bins to observe whether predicted plasmid contigs are derived from the same replicon. Nevertheless, mlplasmids can be used as a basis for plasmid classification by other tools such as PlacnetW (18), facilitating the reconstruction of plasmid sequences in a network graph, or by PlasmidSPAdes filtering of chromosome-derived contigs regardless of contig coverage. Additionally, PlasmidFinder can be used in combination with mlplasmids to find replication genes present in predicted plasmid-derived contigs.

In contrast to cBAR or PlasFlow, mlplasmids is only suitable for genome assemblies from single species. However, we anticipate that a similar methodology can be implemented to create new models for predicting plasmid- and chromosome-derived contigs for other bacterial species with a sufficient number of diverse and complete genomes.

Material and Methods

Retrieving complete genome sequences from the NCBI database

We downloaded complete genomes for *E. faecium* (chromosomes=24; plasmids=82), *K. pneumoniae* (chromosomes=156; plasmids=561) and *E. coli* (chromosomes=168; plasmids=415) from the Assembly Entrez NCBI database (https://www.ncbi.nlm.nih.gov/assembly/) with the following criteria: (i) a status level of 'complete genome' and (ii) one or more plasmid entries in its respective genome assembly. Retrieved genomes and their corresponding accession numbers are available in Table S1.

Extending the number of complete genome sequences for *Enterococcus faecium*

Of 1644 *E. faecium* Illumina-sequenced (MiSeq/NextSeq) isolates, 62 isolates were selected based on their preliminary plasmid content using PlasmidSPAdes (version 3.8.2) and presence of known plasmid replication genes (1) (Supplementary Methods S1). We used Oxford Nanopore Technologies (ONT) MinION and hybrid assembly using Unicycler (version 0.4.1) in 'bold' mode to obtain complete genome sequences (23)

Estimating strain diversity in our collection of complete genomes

To ensure that our training and test sets contained chromosome- and plasmid-derived contigs from a diverse set of isolates belonging to each species, we estimated the diversity present in our collection of *K. pneumoniae*, *E. coli* and *E. faecium* genomes with Mash (version 1.1) (sketch size=1000; k-mer=21) (24). Mash distances were calculated using the total genome content of an isolate (chromosome plus associated plasmids). Computed pairwise Mash distances were transformed into a distance matrix and clustered using the hclust function (method='ward.D2') available in R package stats (version 3.3.3). Hierarchical clustering was visualized using the heatmap.2 function available in R package gplots (version 3.0.1) (25).

Simulating Illumina sequence reads

To calculate the number of paired reads required to simulate sequence read files, we used wgsim (version 0.3.2, https://github.com/lh3/wgsim) with 50× coverage and no error rate. We retrieved the genome size using bioawk (version 20110810, https://github.com/lh3/bioawk) for each selected complete genome of *K. pneumoniae* and *E. coli*.

Assembling Illumina sequence reads

Simulated sequence reads of *K. pneumoniae* and *E. coli* were trimmed using seqtk (version 1.2-r94, https://github.com/lh3/seqtk) with the command '–trimfq'. We used SPAdes (version 3.6.2) to perform *de novo* assembly (26). Contigs with a length smaller than 500 bp were excluded.

E. faecium Illumina NextSeq reads were trimmed using nesoni clip, part of the nesoni toolkit (version 0.132), with the following settings: '–adaptor-clip yes –match 10 –max-errors 1 –clip-ambiguous yes –quality 10 –length 90 –trim-start 0 –trim-end 0 –gzip no –out-separate yes pairs'. Trimmed reads were then assembled into contigs using SPAdes (version 3.5.0) with default settings. Contigs with a mean coverage lower than 10× and/or a length smaller than 500 bp were removed from the assemblies.

Labelling short-read contigs as chromosome or plasmid derived

To label contigs as either plasmid or chromosome derived, SPAdes contigs were mapped using bwa-mem (version 0.7.15-r1140) against complete chromosomal and plasmid sequences (27). Contig alignments were parsed using samtools (version 1.4). This approach allowed to label each SPAdes contig either as plasmid or chromosome derived. SPAdes contigs mapping both to complete chromosomal and plasmid sequences or with a length shorter than 1000 bp were discarded.

Genomic signatures as features to distinguish plasmid and chromosome sequences

To investigate the role of pentamer frequencies as classifier features to differentiate between plasmid and chromosomal sequences, we retrieved the Assembly Entrez NCBI

complete genomes available for *E. faecium*, *K. pneumoniae* and *E. coli* (as previously described in Methods in the section 'Retrieving complete genome sequences from the NCBI database'). We calculated their pentamer frequencies using the R package biostrings (version 2.42.1) (28) and transformed them into a distance matrix (Euclidean distance). We clustered the resulting matrix using the hclust function (method='ward.D2') from R package stats (version 3.3.3). Hierarchical clustering was visualized using the heatmap. plus function available in R package heatmap.plus (version 1.3). Additionally, we used the t-distributed stochastic neighbour embedding (t-SNE) (theta=0.5, iterations=1000, dims=2, is_distance=TRUE) using the implementation provided in the R package Rtsne (version 0.13) (29).

Selection of isolates for benchmarking

We excluded a set of isolates from the training set consisting of contigs derived from isolates of *K. pneumoniae* (chromosomes=11; plasmids=33), *E. coli* (chromosomes=3; plasmids=7) and *E. faecium* (chromosomes=7; plasmids=31) for which original Illumina sequencing data and complete genomes were available. Twelve of these isolates were also used in a recent benchmarking publication of plasmid prediction tools (16) (Table S2). From the benchmarking set of isolates, *E. coli* strain K-12 substrain MG1655, *K. pneumoniae* KSB1_7 and *E. faecium* E2079, E2364 and E9101 did not contain any plasmids and were considered as negative controls. None of these data was used to train *E. faecium*, *K. pneumoniae* and *E. coli* mlplasmids models.

Building a machine-learning model

For each bacterial species, we tuned and compared five different supervised algorithms provided in mlr R package (version 2.11): logistic regression, Bayesian classifier, decision trees, random forest (RF) and support-vector machine (SVM) (30–32). We defined a two-class classification problem using the category 'plasmid' as positive-class. To train and test the resulting classifiers, we considered pentamer frequencies (n=1024) that were calculated using the oligonucleotideFrequency function available in R package biostrings (version 2.42.1). The mlr package was used to split SPAdes-labelled contigs into training (80%) and test sets (20%), preserving the frequencies of each class in both sets (Supplementary Methods S2 and Table S6).

For *E. faecium* training and test sets, we checked the presence of chromosome-labelled contigs corresponding to plasmid sequences and integrated into the chromosome using blastp (version 2.6.0+) (>60% coverage, >80% identity, E-value=1×10−5) against a curated database of known enterococcal plasmid replication sequences (33). Decision trees, RF and SVMs hyperparameters were optimized using random search in a predefined search space (Table S5). We performed 10-fold cross-validation to assess the quality of hyperparameters combination, using error rate as a performance measure, except for *E.*

coli models in which the true-positive rate was considered to overcome a lower plasmid frequency. For each object, posterior probabilities were generated and the class with a highest posterior probability was assigned.

Optimized classifiers were compared for the test set through receiver operating characteristic (ROC) curves. For each classifier, area under the curve (AUC) and precision-recall curves were calculated to compare resulting classifiers based on different true-positive and false-positive thresholds (from 0 to 1). Metrics were assessed using two different units: number of contigs and sequence size in base pairs. The F1-score was reported to obtain a harmonic mean between specificity and sensitivity. Definitions of the statistics reported in this study are reported below.

$$Sensitivity = \frac{True\ positive\ (contigs/bp)}{True\ positive\ (contigs/bp) + False\ negative\ (contigs/bp)}$$

$$Specificity = \frac{True\ negative\ (contigs/bp)}{True\ negative\ (contigs/bp) + False\ positive\ (contigs/bp)}$$

$$Precision = \frac{True\ positive\ (contigs/bp)}{True\ positive\ (contigs/bp) + False\ positive\ (contigs/bp)}$$

$$Accuracy = \frac{True\ positive\ (contigs/bp) + True\ negative\ (contigs/bp)}{Total\ (contigs/bp)}$$

$$F1-score = \frac{2 \times True\ positive\ (contigs/bp)}{2 \times True\ positive\ (contigs/bp) + False\ positive\ (contigs/bp) + False\ negative\ (contigs/bp)}$$

An overview of the method followed to build the resulting classifiers is shown in Fig. 1.

For each bacterial species, we selected the best model based on the resulting F1-score to predict plasmid- and chromosome-derived sequences. We implemented them in a new R package called mlplasmids available at https://gitlab.com/sirarredondo/mlplasmids under GNU General Public License v3.0. We also developed a Shiny app, available at https://sarredondo.shinyapps.io/mlplasmids/ to enable plasmid prediction with a graphical user interface (34).

Comparison of mlplasmids against other plasmid prediction tools

We evaluated the performance of mlplasmids against PlasFlow (version 1.0), PlasmidSPAdes (version 3.8.2) and cBar (version 1.2). We considered contigs derived from the isolates described in Methods in the section 'Selection of isolates for benchmarking' for which short-read sequencing data and complete genomes were available to validate the presented plasmid prediction tools. cBAR was run using default parameters. PlasFlow was

run using standard and recommended parameters corresponding to a minimum posterior probability of 0.7 and minimum contig length of 1000 bp. Contigs with a lower probability were catalogued as 'unclassified' by PlasFlow and were excluded from this comparison. PlasmidSPAdes (version 3.8.2) generates its own assembly and the resulting contigs were labelled as true- or false-positive results following the methodology described in Methods in the section 'Labelling short-read contigs as chromosome or plasmid derived'. For all the tools, we filtered out contigs with a length shorter than 1000 bp.

We benchmarked these plasmid prediction tools using: accuracy, F1-score, and precision. PlasmidSPAdes does not predict chromosome-derived contigs; thus, we could not directly calculate its accuracy and F1-score. To overcome this, we used Quast (version 4.1) to map plasmid-predicted contigs against their respective complete plasmid sequences (35). We then retrieved the reported 'genome fraction' in Quast, which is defined as the percentage of aligned bases from the reference genome covered by contigs predicted as plasmid derived. This allowed us to obtain an estimation of PlasmidSPAdes' sensitivity (35).

Validating mlplasmids against complete plasmid sequences

To observe the performance of the resulting classifiers in sequences larger than the mean contig length present in our training and test sets, we used *K. pneumoniae* (n=11) and *E. coli* (n=3) complete genomes described in Methods in the section 'Selection of isolates for benchmarking' to observe mlplasmids performance predicting complete chromosomal and plasmid sequences. In addition, we downloaded complete genomes of *E. faecium* from the Assembly Entrez NCBI database (n=24) that were not included in the training set (Table S1).

Predicting the location of antibiotic-resistance genes

All assemblies of *E. faecium* (n=369), *K. pneumoniae* (n=1346) and *E. coli* (n=5234) with an assembly level corresponding to 'contig' were downloaded from NCBI Genomes FTP (ftp.ncbi.nlm.nih.gov/genomes/). For each downloaded draft assembly, we used Abricate (version 0.8.2) (https://github.com/tseemann/abricate) to screen contigs against the ResFinder database (release from 18th May 2016) (36) to determine the presence of antimicrobial-resistance genes. Abricate was run using a minimum DNA identity of 95% and a minimum coverage of 80%. To assign a particular contig as plasmid- or chromosome-derived, we used *E. faecium*, *K. pneumoniae* and *E. coli* SVM models in mlplasmids specifying a minimum posterior probability of 0.7 and a minimum contig length of 1000 bp (Table S3).

To validate mlplasmids' potential to predict the genomic context of a particular antibiotic-resistance gene, we used the isolates described in Methods in the section 'Selection of isolates for benchmarking'. We used mlplasmids on a contig level to assign whether a particular resistance gene was present on a plasmid or chromosome context. We used

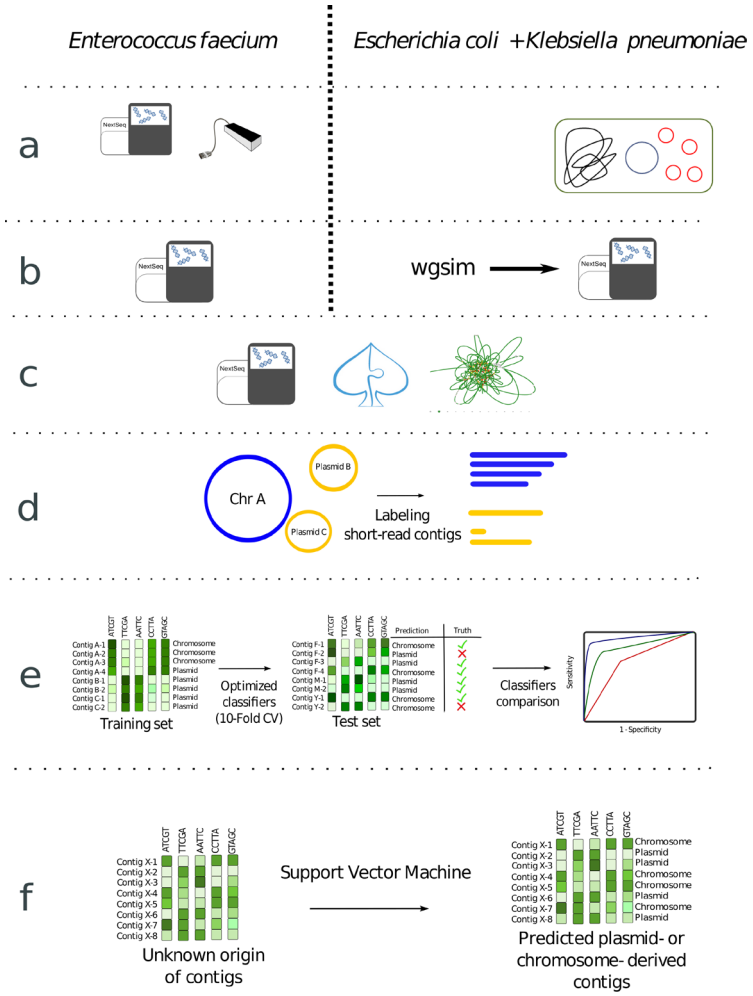


Figure 1. Workflow to create the plasmid models for *Enterococcus faecium*, Klebsiella pneumoniae and *Escherichia coli*. (a) For *E. faecium*, 62 Illumina-sequenced strains were selected for ONT sequencing and Unicycler was used to extend the number of complete genomes available for this species. For *E. coli* and *K. pneumoniae*, we downloaded complete genomes with plasmids associated from the Assembly Entrez NCBI database. (b) For *E. coli* and *K. pneumoniae*, we simulated reads with 50 coverage and no error rate using wgsim. (c) Illumina simulated and non-simulated reads were *de novo* assembled using SPAdes. (d) We mapped short-read contigs against complete genome sequences to define a reliable dataset of short-read contigs as plasmid or chromosome derived. (e) For each bacterial species, five machine-learning classifiers were trained (10-fold cross-validation) and compared using a specific bacterial species training and test set. (f) SVM models were implemented in mlplasmids and used to predict plasmid- and chromosome-derived sequences in isolates with only short-read WGS data available. The complete workflow is available from https://gitlab.com/sirarredondo/analysis_mlplasmids.

identical metrics, introduced in Methods in the section 'Building a machine-learning model', to determine performance metrics but considering genes as units.

Predicting the plasmidome content of *Enterococcus faecium*

We used the *E. faecium* optimized model to predict plasmid- and chromosome-derived contigs from the collection of 1644 *E. faecium* Illumina-sequenced (MiSeq/NextSeq) isolates (Table S6). We filtered out contigs with a length shorter than 500 bp and a minimum posterior probability of 0.7 to assign contigs either as plasmid or chromosome derived using the class with a highest posterior probability. SPAdes assembly statistics from this collection are shown in Table S7.

Data overview

To facilitate the comprehension and reproducibility of the analysis, we summarized in Supplementary Methods S3 the different sequencing and assembly files used in each of the sections previously described in Methods.

Acknowledgments

We thank the Utrecht Sequencing Facility for providing a NextSeq and ONT sequencing service. The Utrecht Sequencing Facility is subsidized by the University Medical Center Utrecht, Hubrecht Institute and Utrecht University.

References

1. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV et al. Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology. Boston, MA: Massachusetts Eye and Ear Infirmary; 2014.

2. Smalla K, Jechalke S, Top EM. Plasmid detection, characterization, and ecology. Microbiol Spectr 2015;3:PLAS-0038-2014.

3. Carattoli A. Plasmids and the spread of resistance. Int J Med Microbiol 2013;303:298–304.

4. de Been M, Lanza VF, de Toro M, Scharringa J, Dohmen W et al. Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. PLoS Genet 2014;10:e1004776.

5. Doumith M, Godbole G, Ashton P, Larkin L, Dallman T et al. Detection of the plasmid-mediated *mcr-1* gene conferring colistin resistance in human and food isolates of *Salmonella enterica* and *Escherichia coli* in England and Wales. J Antimicrob Chemother 2016;71:2300–2305.

6. Freitas AR, Tedim AP, Francia MV, Jensen LB, Novais C et al. Multilevel population genetic analysis of vanA and vanB *Enterococcus faecium* causing nosocomial outbreaks in 27 countries (1986-2012). J Antimicrob Chemother 2016;71:3351–3366.

7. Conlan S, Thomas PJ, Deming C, Park M, Lau AF et al. Single molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing *Enterobacteriaceae*. Sci Transl Med 2014;6:254ra126.

8. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ et al. Plasmid classification in

an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. Front Microbiol 2017;8:182. 2012;19:455–477.

9. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE et al. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 2012;30:434–439.

10. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A et al. Nested Russian doll-like genetic mobility drives rapid dissemina- tion of the carbapenem resistance gene $bla_{KPC.}$ Antimicrob Agents Chemother 2016;60:3767–3778.

11. Carattoli A, Zankari E, García-Ferna ndez A, Voldby Larsen M, Lund O et al. *In silico* detection and typing of plasmids using plas- midfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 2014;58:3895–3903.

12. Zhou F, Xu Y. cBar: a computer program to distinguish plasmid- derived from chromosome-derived sequence fragments in meta- genomics data. Bioinformatics 2010;26:2051–2052.

13. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E et al. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics 2017;33:475–482.

14. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A et al. plasmid- SPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 2016;32:3380–3387.

15. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signa- tures. Nucleic Acids Res 2018;46:e35.

16. 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.

17. de Toro M, Garcilla on-Barcia MP, de La Cruz F. Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. Microbiol Spectr 2014;2:PLAS–0031–2014.

18. Vielva L, de Toro M, Lanza VF, de La Cruz F. PLACNETw: a web- based tool for plasmid reconstruction from bacterial genomes. Bioinformatics 2017;33:3796–3798.

19. George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N et al. Resolving plasmid structures in *Enterobacteriaceae* using the MinION nanopore sequencer: assessment of MinION and MinION/ Illumina hybrid data assembly approaches. Microb Genom 2017;3: e000118.

20. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat Methods 2015;12:733–735.

21. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 2015;23:110–120.

22. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G et al. A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience 2015;4: 60.

23. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacte- rial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13:e1005595.

24. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 2016;17:132.

25. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W et al. gplots: Various R Programming Tools for Plotting Data, R package Version 2; 2009.

26. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–477.

27. Li H. 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. http://arxiv.org/abs/1303.3997.

28. Page s H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient Manipulation of Biological Strings, R Package Version 2.42.1; 2016.

29. der M, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–2605.

30. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J et al. mlr: machine learning in R. J Mach Learn Res 2016;17:1–5.

31. Kuhn M. caret: classification and regression training. Astrophysics Source Code Library; 2015. https://ui.adsabs.harvard.edu/#abs/ 2015ascl.soft05003K.

32. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A et al. The e1071 Package. Misc Functions of Department of Statistics. Vienna: TU Wien; 2006. www.cs.upc.edu/~belanche/Docencia/ mineria/Practiques/R/e1071.pdf.

33. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV et al. Extrachromosomal and mobile elements in enterococci: transmis- sion, maintenance, and epidemiology. In: Gilmore MS, Clewell DB, Ike Y and Shankar N (editors). Enterococci: from Commensals to Leading Causes of Drug Resistant Infection. Boston: Massachusetts Eye and Ear Infirmary; 2014.

34. Rstudio. Shiny: Easy Web Applications in R. 2014.

35. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29: 1072–1075.

36. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012;67:2640–2644.

37. Wardal E, Kuch A, Gawryszewska I, Żabicka D, Hryniewicz W et al. Diversity of plasmids and Tn1546-type transposons among VanA *Enterococcus faecium* in Poland. Eur J Clin Microbiol Infect Dis 2017;36:313–328.

38. van Hal SJ, Ip CL, Ansari MA, Wilson DJ, Espedido BA et al. Evolutionary dynamics of *Enterococcus faecium* reveals complex genomic relationships between isolates with independent emer- gence of vancomycin resistance. Microb Genom 2016;2: 000048.

39. Navon-Venezia S, Kondratyeva K, Carattoli A. Klebsiella pneumo- niae: a major worldwide source and shuttle for antibiotic resis- tance. FEMS Microbiol Rev 2017;41:252–275.

40. Falgenhauer L, Waezsada SE, Gwozdzinski K, Ghosh H, Doijad S et al. Chromosomal locations of *mcr-1* and $bla_{\text{CTX-M-15}}$ in fluoroquinolone- resistant *Escherichia coli* ST410. Emerg Infect Dis 2016;22:1689–1691.

Supplementary Results

Comparison of mlplasmids against other plasmid prediction tools (S1)

Only with the purpose of comparing mlplasmids and plasflow prediction, we created an artificial and third category for mlplasmids named 'unclassified' in which we included all contigs first assigned as plasmid- or chromosome-derived but with a posterior probability lower than 0.7. We only defined a mlplasmids 'unclassified' category for this particular analysis, since mlplasmids prediction only consists of two classes: plasmid or chromosome, and users can decide whether filter out predicted contigs based on their associated posterior probabilities.

For the three single species datasets, the frequency of this category was lower for mlplasmids (*E. faecium* = 0.03 ; *K. pneumoniae* = 0.10 ; *E. coli* = 0.05) compared to PlasFlow (*E. faecium* = 0.16; *K. pneumoniae* = 0.17 ; *E. coli* = 0.21) (Fig. S6). These results showed that most of predicted contigs had an associated high posterior probability of belonging to either plasmid- or chromosome-class with mlplasmids compared to PlasFlow. To show the potential of mlplasmids predicting unclassified contigs from PlasFlow, we considered unclassified contigs by plasflow and observed the posterior probabilities given by mlplasmids. For *E. faecium* and *K. pneumoniae* datasets, we observed that unclassified contigs from PlasFlow derived from the chromosome-class and plasmid-class were mostly correctly predicted by mlplasmids (Fig. S7a and S7b). For *E. coli*, unclassified contigs from the plasmid-class showed a non-uniform distribution whereas contigs from the chromosome-class were in general correctly predicted (Fig. S7c).

Applicability for predicting sequences derived from incomplete long-read assemblies (S2)

For all bacterial species, mlplasmids did not recover any false positive sequences (Specificity = 1). For *E. coli*, only a single plasmid sequence (NC_022662.1) was wrongly predicted as chromosome- derived but with a low posterior probability associated to that class (0.53). In the case of *K. pneumoniae*, mlplasmids misclassified a plasmid sequence with a length of 26.45 kbp (NZ_CP015133.1) from *K. pneumoniae* strain KPN555. For *E. faecium* two sequences were misclassified as chromosomal (NZ_LT598665.1 and NZ_CP019991.1) and the last sequence (NZ_CP019991.1) could correspond to a phage since its NCBI annotation showed two phage-related genes. This demonstrates the flexibility of mlplasmids to predict sequences with different lengths compared to average contig length used to train and test resulting classifiers and discarded misclassifications due to a correlation between pentamer frequencies and contig length. This may facilitate the classification of contigs generated from incomplete hybrid or long-read assemblies as exemplified for isolate *E. faecium* E7070. This isolate was selected for ONT sequencing and after hybrid assembly, 16 contigs were reported. Contigs predicted as plasmid by mlplasmids (n = 6)

contained circularization signatures whereas the rest of the contigs (n = 10) were predicted as chromosome-derived (Fig. S9). This facilitated the design of appropriate PCR reactions to complete the genome sequence for E7070.

Supplementary Methods

Extending the number of complete genome sequences for *E. faecium* (S1)

Illumina sequencing

Bacterial isolates were grown overnight (O/N) at 37°C on blood agar plates. Single colonies were picked up and grown O/N at 37°C with Brain Heart Infusion (BHI). Bacterial cell pellets were pretreated and incubated 1-4 hours with 180 µL of enzymatic lysis buffer. Subsequently, 0.75 mg proteinase K were added and incubated at 56°C until lysis completion. 20 µL of RNAse A (10mg/mL) were added and incubated for 5' at room-temperature (RT). Total DNA purification was performed using and following the protocol from NucleoSpin 96 Tissue Core Kit (Machery-Nagel), vacuum processing. DNA concentration was measured using Quant-it Picogreen (Thermo Fisher Scientific). Library preparation was carried out following Nextera DNA Library Prep Reference Guide. Finally, Nextera libraries were sequenced using Illumina NextSeq at USEQ, Utrecht, The Netherlands (http://www.useq.nl).

WGS short-read assemblies

Illumina reads were trimmed using nesoni clip, part of the nesoni toolkit (version 0.132), with the following settings: '--adaptor-clip yes --match 10 --max-errors 1 --clip-ambiguous yes --quality 10 -- length 90 --trim-start 0 --trim-end 0 --gzip no --out-separate yes pairs:'. Trimmed reads were then assembled into scaffolds using SPAdes (version 3.5.0) with default settings. Scaffolds with an average coverage lower than 10 and/or a length smaller than 500bp were removed from the assemblies.

Isolate selection for ONT sequencing

A fraction (n=60) of the total number of isolates (n=1,644) was selected for long-read sequencing with ONT. The plasmid content of the isolates *in silico* were estimated using PlasmidSPAdes (version 3.8.2) (1). Prokka (version 1.12) was used to annotate the putative plasmid contigs using the Enterococcus database included in Prokka (2). Orthologous clustered genes were estimated using Roary (version 3.8), splitting paralogues and defining a threshold of 95% amino-acid level similarity to cluster protein sequences (3). This multi-dimensionality matrix was then reduced and visualized to two dimensions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) (theta = 0.5, iterations = 1000, dims = 2) using the implementation provided in the R (version 3.3.3) package Rtsne (version 0.13) (4, 5). k-means (iter.max = 1000) provided in the R package stats was used to allocate 50 centroids into the dimensionality reduced distribution given by tSNE. Euclidean

distance of each isolate was calculated to extract the 50 isolates closest to each centroid.

To cover all plasmid replication genes not present in the first selection, 12 additional isolates were selected for ONT sequencing. This second selection was based on a reciprocal blast of the predicted plasmid orthologous genes against 76 previously described plasmid replication amino-acid sequences from the genus Enterococcus (6). Reciprocal blast allowed to identify miss-annotated genes corresponding to plasmid replication sequences. Isolates bearing plasmid replication genes not present in the first selection were sorted and selected based on highest number of orthologous genes.

ONT sequencing

E. faecium selected isolates were grown O/N at 37°C on blood agar plates, then single colonies were picked up and grown with BHI at 37°C. Genomic DNA was extracted using the Wizard Genomic DNA purification kit (Promega) following manufacturer's instructions. Isolated DNA was sheared (4000 rpm, 2x120 seconds) using G-tubes (Covaris). Library preparation was performed using Ligation Sequencing Kit 1D (SQK-LSK108) with the Native Barcoding Kit 1D (EXP-NBD103). Genomic libraries were loaded onto R9.4 (FLO-MIN106) flowcells using the MinION device (Mk2). Libraries were basecalled using Metrichor workflows (Run 1 ,2, 3), Albacore 1.01 (Run 4, 5) and Albacore 1.1.0 (Run 6). ONT Sequencing and basecalling were conducted at USEQ, Utrecht, The Netherlands (http://www.useq.nl)

ONT reads and hybrid assembly

Fastq files were obtained from base-called data using Poretools (version 0.6.0) except for Run6 in which fastq files were retrieved using Albacore (version 1.1.0). Distribution of read length and total number of reads were calculated using Bioawk (version 20110810, https://github.com/lh3/bioawk). We used Porechop (version 0.2.1, https://github.com/rrwick/Porechop) to trim reads and filter out chimeras from different bins specifying the flag "--discard_middle". Illumina reads were trimmed using seqtk (version 1.2-r94, https://github.com/lh3/seqtk) with the command "--trimfq" prior to assembly.

Hybrid assembly was performed using Unicycler (version 0.4.1), specifying "bold" mode (7). Briefly, Unicycler uses SPAdes (version 3.6.2) to create different assembly graphs based on different k-mer size only considering Illumina reads (8). The best assembly graph was selected by Unicycler based on number of dead-ends and contiguity. Next, all ONT reads were used to scaffold and solve the assembly graph. Additionally, we specified the same file as described above (Isolate selection for ONT sequencing) containing 76 known plasmid replication sequences to rotate and change the 0- coordinate of replicons resulting from hybrid assembly (6). Finally, Unicycler conducted several rounds of Pilon (version 1.22) to polish genome sequences using Illumina reads (9).

Categorization of Unicycler contigs

Unicycler contigs were labeled either as chromosome or plasmids based on size and circularity. Contigs were categorized as chromosome if they were larger than 350 kbp, regardless of circularity.

However, only contigs were categorized as plasmids if they were circular and smaller than 350 kbp. Putative plasmids smaller than 350 kbp and lacking circularization signatures were not categorized. Draft annotation (Prokka - version 1.12) of plasmid sequences allowed to identify and discard four putative complete phage sequences present as circular contigs.

Building a machine-learning model (S2)

For each bacterial species, we tuned and compared five different supervised algorithms provided in mlr R package (version 2.11): logistic regression, Bayesian classifier, decision trees, random forest (RF) and support-vector machine (SVM) (10). We defined a two-class classification problem using the category 'plasmid' as positive-class. To train and test the resulting classifiers we considered pentamer frequencies (n=1024) which were calculated using oligonucleotideFrequency function available in R package biostrings (version 2.42.1). Mlr package was used to split SPAdes labeled contigs into training (80%) and test set (20%), preserving the frequencies of each class in both sets (Supplementary Table S4).

Decision trees, random forest and support-vector machines hyperparameters were optimized using random search in a predefined search space (Supplementary Table S5). We performed 10-fold cross- validation to assess the quality of hyperparameters combination, using error rate as performance measure. For each object, posterior probabilities were generated and the class with a highest posterior probability was assigned.

Datasets (S3)

In this study, we used Illumina NextSeq/MiSeq data for 1,644 *E. faecium* isolates that are available under the ENA project PRJEB28495. A fraction (n = 62) of these 1,644 *E. faecium* isolates was completed using ONT MinION reads which are publicly available under the figshare projects: 10.6084/m9.figshare.7046804 ; 10.6084/m9.figshare.7047686

From these 62 ONT isolates, 5 were not used to label short-read contigs to train and test mlplasmids models. These 5 isolates (E2079, E2364, E4457, E7591, E8172 and E9101) were used to benchmark *E. faecium* mlplasmids models against other plasmid tools. A complete overview of the different datasets used in this study is available at Supplementary Table S6.

Supplementary Figures

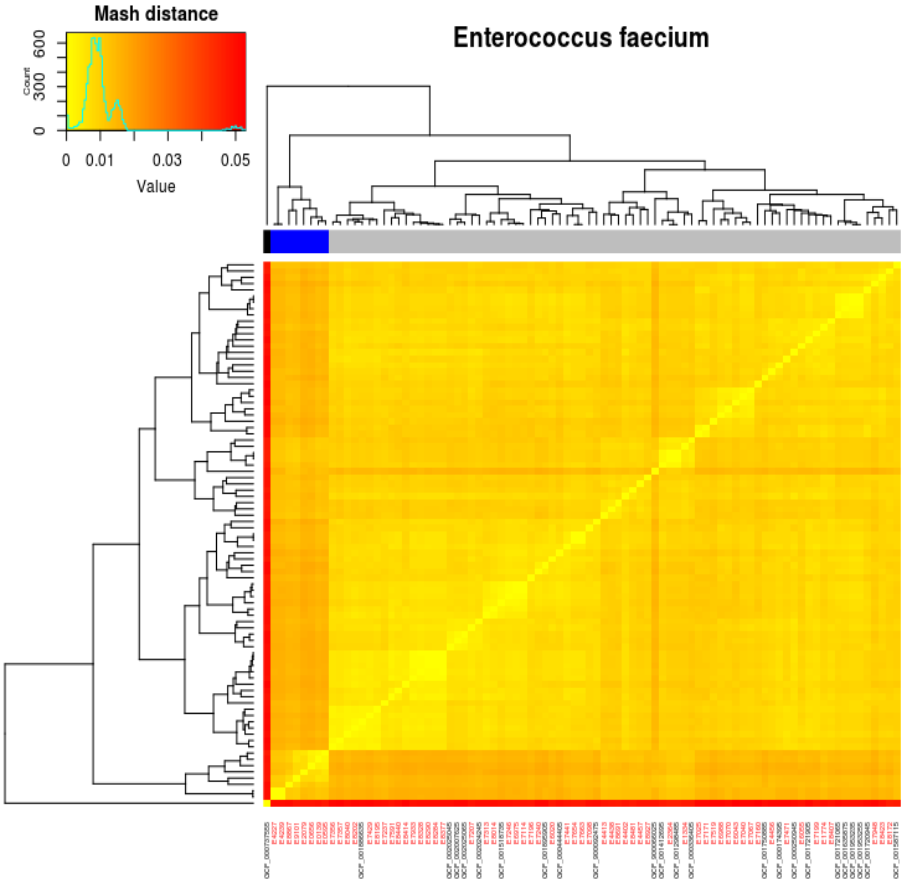


Figure S1. Ward hierarchical clustering of computed pairwise mash distances (k = 21 ; s = 1,000) from *E. faecium* isolates. Based on dendrogram branch lengths, we defined three clusters (black, blue and grey) and visualized mash distances using heatmap based on their genome content similarity. At the bottom y-axis, we coloured in red *E. faecium* isolates (n = 60) that were selected and completed using ONT sequencing and Illumina sequencing. Rest of the isolates corresponded to publicly available NCBI complete genomes from *E.faecium* (n = 24).

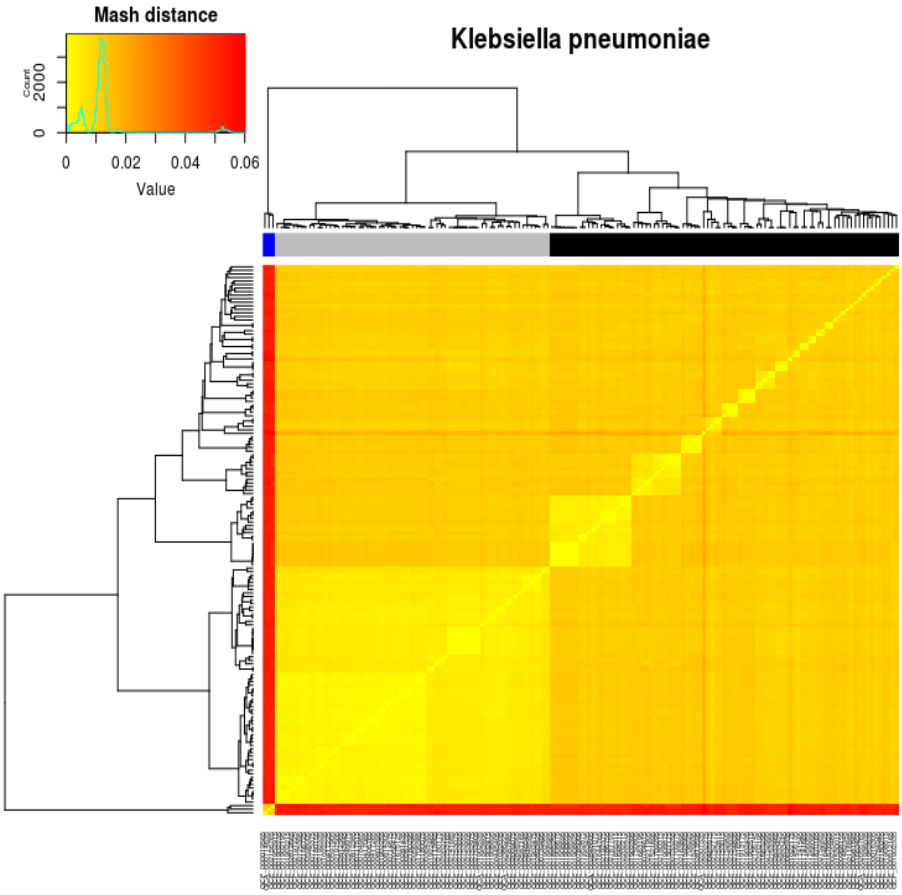


Figure S2. Ward hierarchical clustering of computed pairwise mash distances (k = 21 ; s = 1,000) from *K. pneumoniae* isolates retrieved from Assembly Entrez NCBI database (n = 156). Based on dendrogram branch lengths, we defined three clusters of isolates (blue, grey and black) and visualized mash distances using heatmap to group isolates based on their genome content similarity.

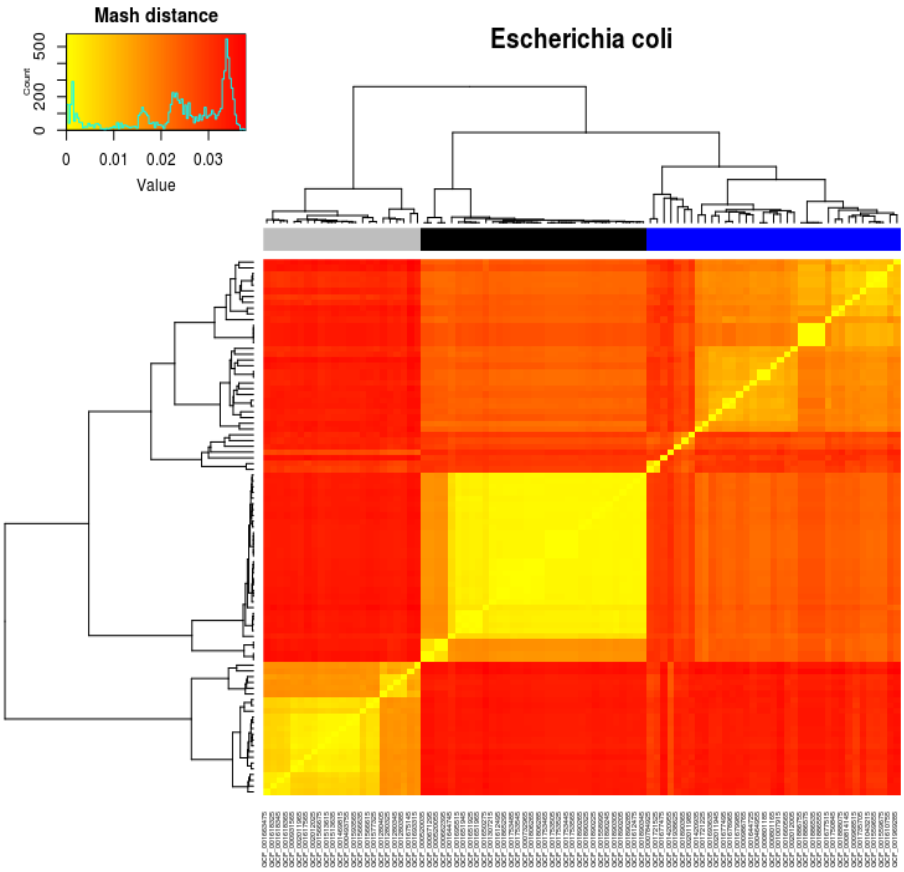


Figure S3. Ward hierarchical clustering of computed pairwise mash distances (k = 21 ; s = 1,000) from *E. coli* isolates retrieved from Assembly Entrez NCBI database (n = 168). Based on dendrogram branch lengths, we defined three clusters of isolates (grey, black and blue) and visualized mash distances using heatmap to group isolates based on their genome content similarity.

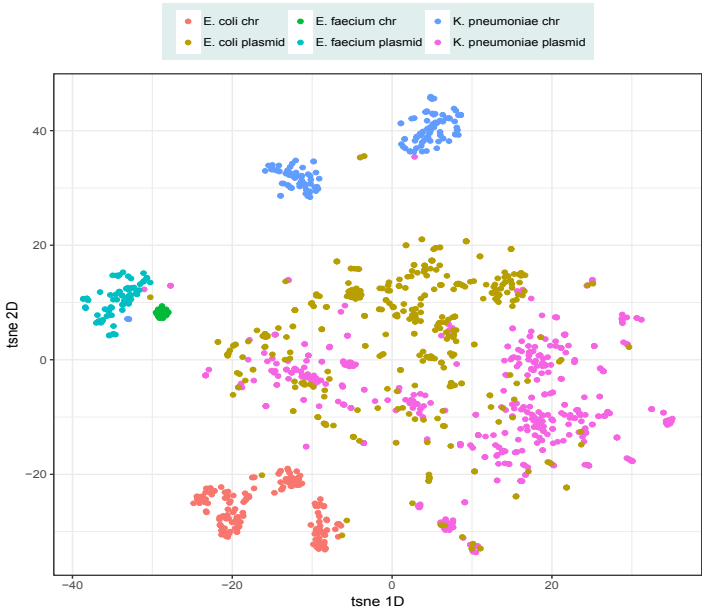


Figure S4. t-sne clustering of all chromosome and plasmid sequences from Assembly Entrez NCBI database corresponding to *E. coli*, *K. pneumoniae* and *E. faecium* based on pentamer frequencies. Each point in the graph corresponds to a different type replicon: *E. coli* chromosome (red), *E. coli* plasmid (yellow), *K. pneumoniae* chromosome (dark blue), *K. pneumoniae* plasmid (pink), *E. faecium* chromosome (green) and *E. faecium* plasmid (light blue).

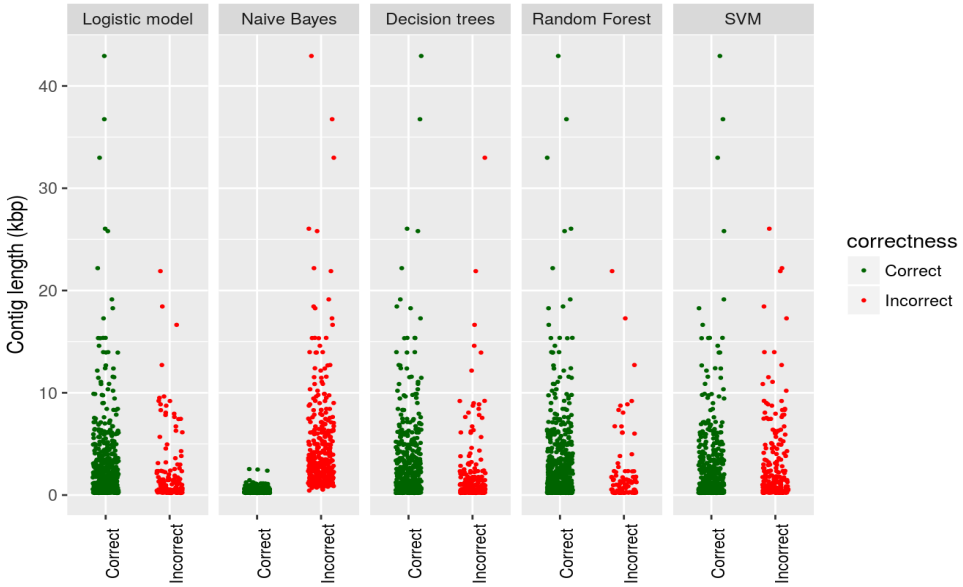


Figure S5. Distribution of correct- and miss- classified short-reads contigs for: Logistic Model, Bayesian Classifier (Naive Bayes), Decision trees, Random Forest, and Support-Vector Machine (SVM). Except for the Bayesian classifier, misclassification most notably occured in contigs with short length.

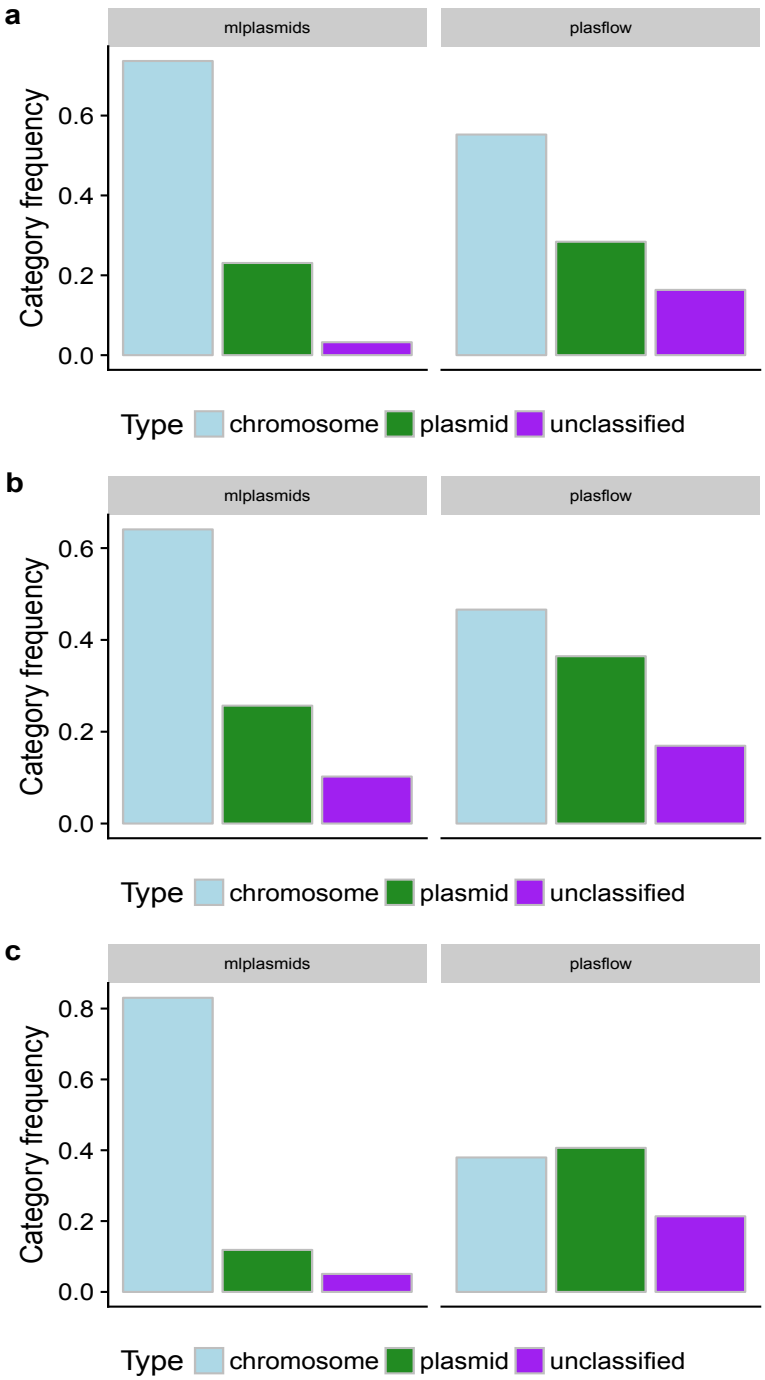


Figure S6. Categorizing the prediction of mlplasmids and plasflow for *E. faecium* (a), *K. pneumoniae* (b) and *E. coli* (c) contigs belonging to our validation sets. We used a minimum posterior probability of 0.7 to assign a contig either to the chromosome- or plasmid-class and with a minimum length of 1,000 bp. Rest of the contigs were included in the category 'unclassified'.

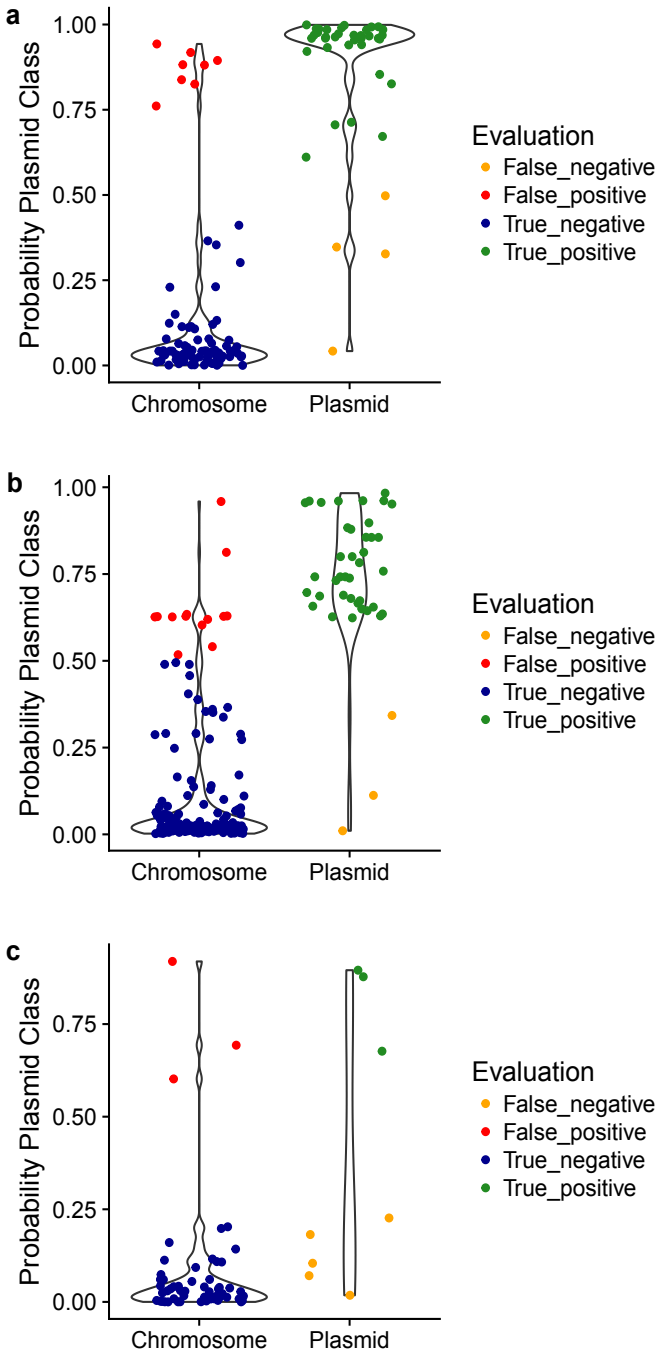


Figure S7. Unraveling the origin of contigs unclassified by plasflow using mlplasmids. *E. faecium* contigs (a), *K. pneumoniae* (b) and *E. coli* (c) which were predicted as 'unclassified' by plasflow were interrogated using mlplasmids. Each predicted contig was grouped into chromosome- or and plasmid-derived (x-axis), coloured based on prediction evaluation and associated probability plasmid-class (y-axis) represented.

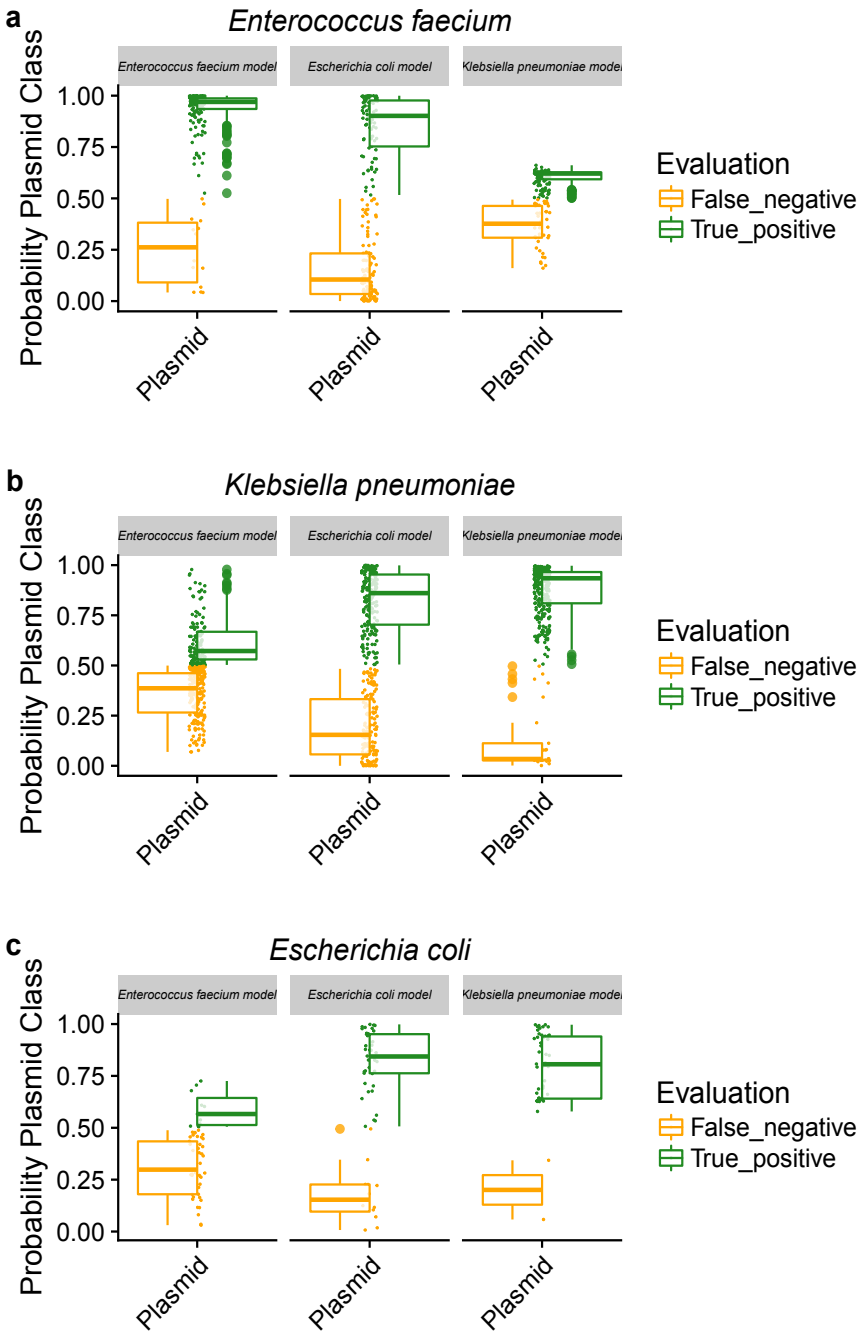


Figure S8. Estimating mlplasmids potential to predict plasmid sequences transferred by HGT events. We used all the three species models available in mlplasmids to predict contigs belonging to *E. faecium* (a), *K. pneumoniae* (b) and *E. coli* (c) validation sets. Each plasmid-derived contig was coloured as false-negative (orange) or true-positive (green) based on evaluation of mlplasmids prediction.

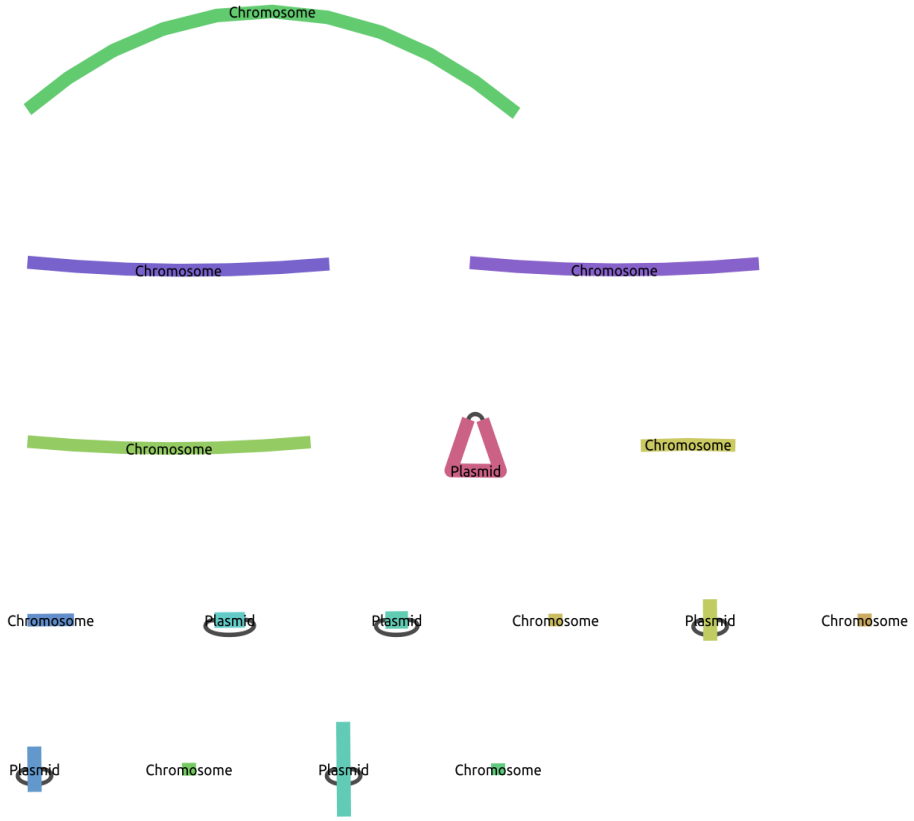


Figure S9. mlplasmids applicability to predict contigs derived from incomplete hybrid or long-read assemblies. Bandage visualization of the hybrid assembly obtained for the *E. faecium* isolate E7070. For this isolate, hybrid assembly using Unicycler did not result in a complete assembly (chromosome and plasmids in single and circular components). Resulting contigs were labeled based on mlplasmids prediction.

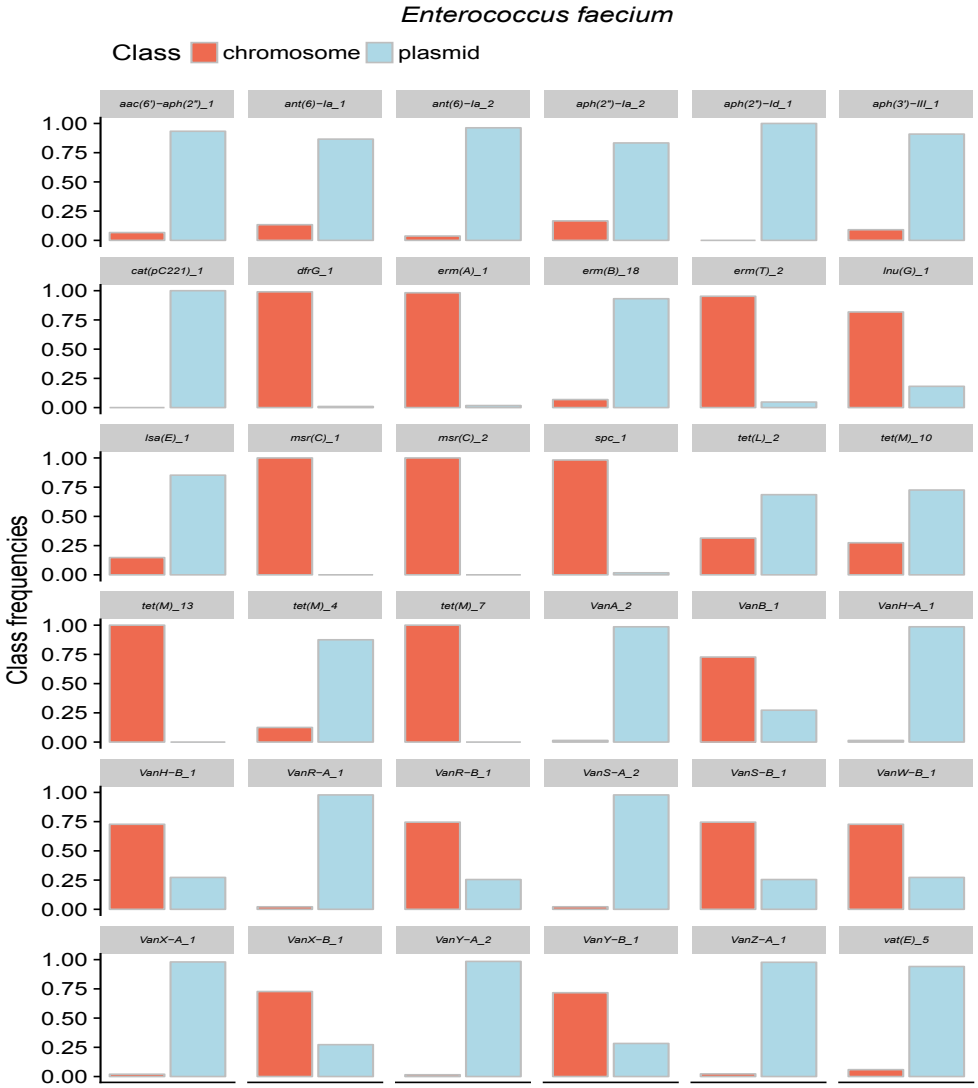


Figure S10. *Enterococcus faecium* resistome. Draft genomes available in NCBI Genomes FTP (n = 369) were downloaded and screened using Abricate and ResFinder for the presence of antibiotic resistance genes. Each contig containing a resistance gene was predicted with mlplasmids to predict plasmid- or chromosome-origin. For visualization purposes, only antibiotic resistance genes present more than five times are shown.

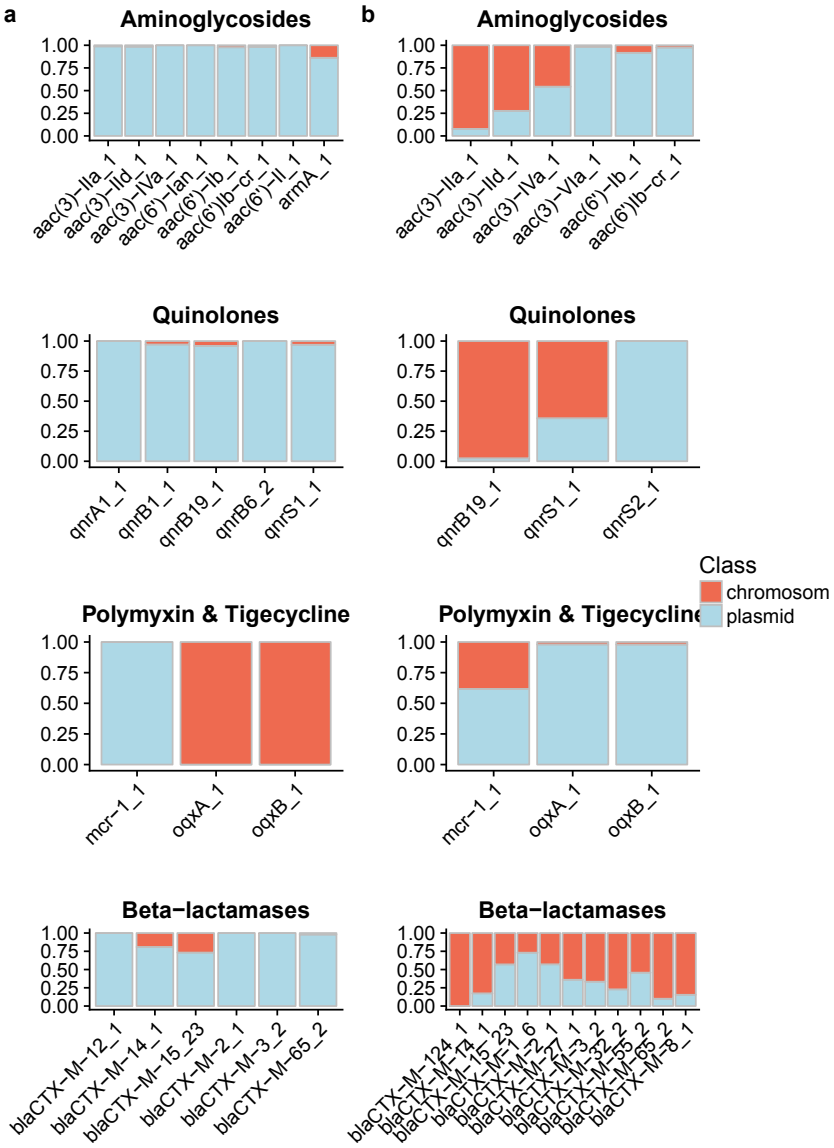


Figure S11. Highlighted genes for Klebsiella pneumoniae (panel A) and *Escherichia coli* (panel B).

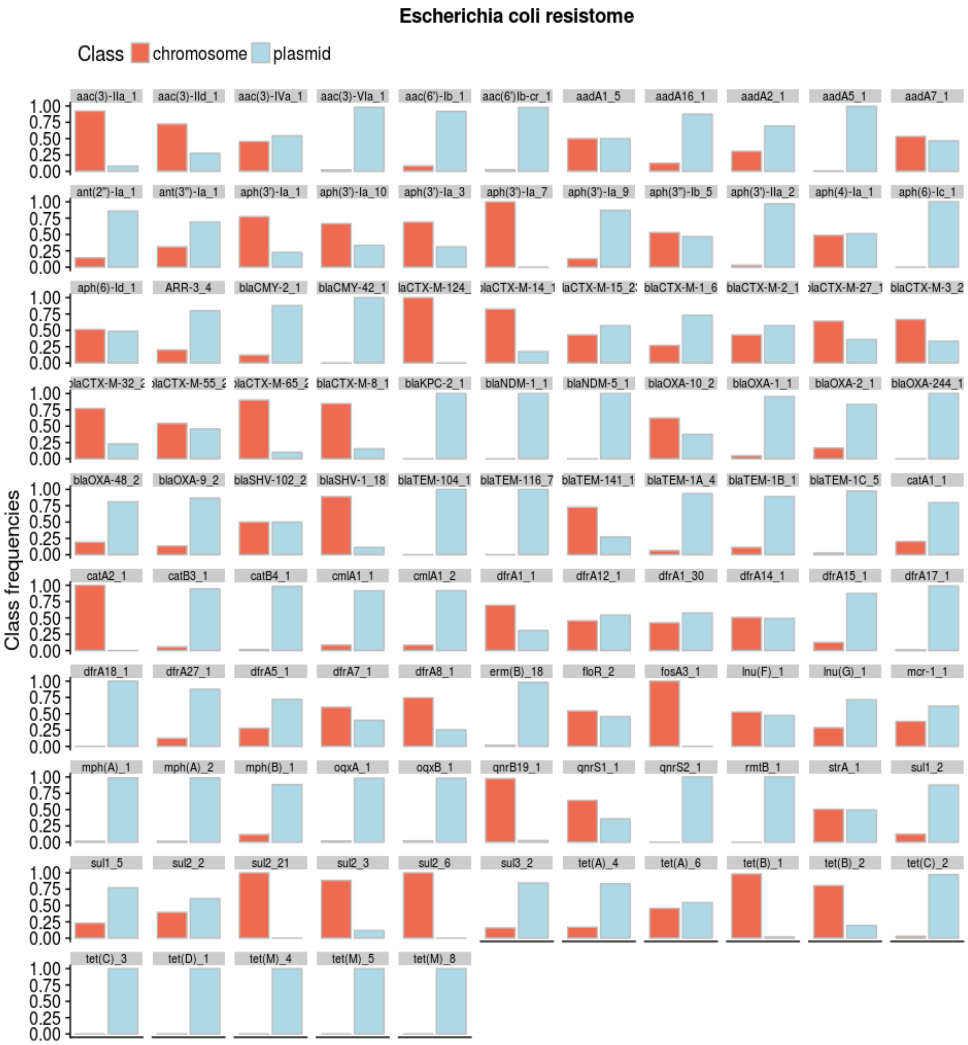


Figure S12. *Escherichia coli* resistome. Draft genomes available in NCBI Genomes FTP (n = 5,234) were downloaded and screened using Abricate and ResFinder for the presence of antibiotic resistance genes. Each contig containing a resistance gene was predicted with mlplasmids to predict plasmid- or chromosome-origin. For visualization purposes, only antibiotic resistance genes present more than five times are shown.

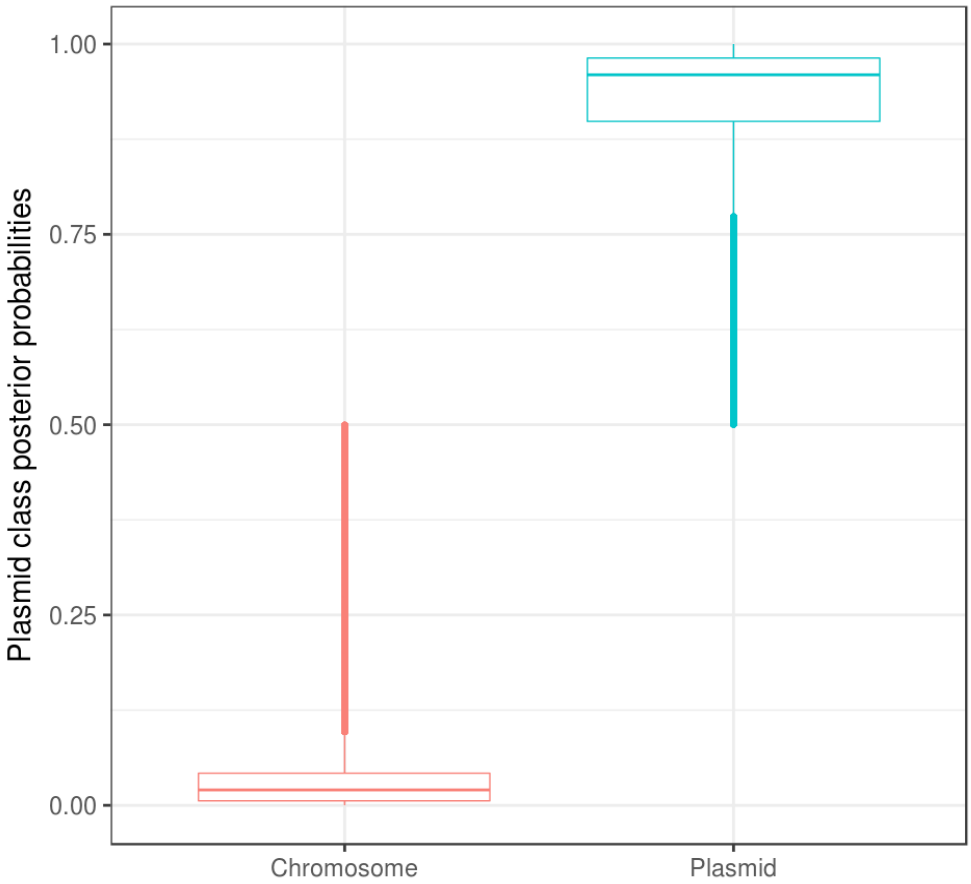


Figure S13. Predicting the plasmidome content of *E. faecium* isolates (n = 1,644). Posterior probabilities of short-read contigs (n= 289,369) of belonging to chromosome- or plasmid-class using our optimized mlplasmids *E. faecium* model for our collection of 1,644 Illumina sequenced *E. faecium* isolates.

Supplementary Tables

Supplementary Tables S1, S2 and S3 are available at
https://doi.org/10.1099/mgen.0.000224

Supplementary Table S4. Description of the training and test sets used for each bacterial species. For each dataset (training or test), the number of objects (SPAdes contigs) and number of features (5-mer combinations) are indicated.

Bacterial species	Set	Number of objects	Number of features	Prevalence plasmid-class	Prevalence chrm-class
E. faecium	Training set	8336	1024	0.33	0.67
E. faecium	Test set	2085	1024	0.34	0.66
K. pneumoniae	Training set	10051	1024	0.38	0.62
K. pneumoniae	Test set	2513	1024	0.37	0.67
E. coli	Training set	10061	1024	0.12	0.88
E. coli	Test set	2651	1024	0.14	0.86

Supplementary Table S5. Hyperparameters optimized for decision trees, random forest, and support vector machine.

Classifier	Hyperparameter	Search space (min-max value)
Decision trees	minsplit	10-50
Decision trees	minbucket	5-50
Decision trees	cp	0.001-0.2
Random Forest	ntree	50-1000
Random Forest	mtry	3-10
Random Forest	nodesize	10-50
Support-vector machine	C	(-10)/10
Support-vector machine	sigma	(-10)/10

Supplementary Table S6.Sequencing/Assembly data used in this study.

Bacterial species	Analysis	Dataset	Availability
E. faecium	Labeling short-read contigs as chromosome- or plasmid- derived	Newly generated *E. faecium* genomes (n = 55)	Illumina NextSeq/Miseq reads: ENA Project : PRJEB28495 ONT MinION reads: figshare projects: 10.6084/m9.figshare.704680410 .6084/m9.figshare.7047686
E. faecium	Benchmarking against other plasmid tools	Newly generated *E. faecium* genomes (n = 7)	Illumina NextSeq/Miseq reads; ENA Project : PRJEB28495 ONT MinION reads; figshare projects: 10.6084/m9.figshare.7046804 10.6084/m9.figshare.7047686
E. faecium	Prediction of the plasmidome content	Newly generated 1,644 *E. faecium* genomes	Illumina NextSeq/Miseq reads; ENA Project : PRJEB28495
E. faecium	Validating mlplasmids against complete genome sequences	Suppl. Table S1	Publicly available NCBI genomes
E. faecium	Predicting the location of AMR genes	Suppl. Table S3	Publicly available NCBI genomes
K. pneumoniae and *E. coli*	Labeling short-read contigs as chromosome- or plasmid- derived	Suppl. Table S1	Publicly available NCBI genomes
K. pneumoniae and *E. coli*	Benchmarking against other plasmid tools	Suppl. Table S2	Publicly available NCBI genomes
K. pneumoniae and *E. coli*	Predicting the location of AMR genes	Suppl. Table S3	Publicly available NCBI genomes

Supplementary Table S7. SPAdes assembly statistics using Illumina MiSeq/NextSeq.

Technology	No. of isolates	Mean coverage	Mean N50	Mean contig length	Median contig length	Avg. no. of contigs
Illumina MiSeq	63	98X	54616 bp	21531 bp	6898 bp	169.1
Illumina NextSeq	1581	113X	52256 bp	17989 bp	5356 bp	176.3

Supplementary References

1. Antipov D, Hartwick N, Shen M, Raiko M, Pevzner PA. plasmidSPAdes : Assembling Plasmids from Whole Genome Sequencing Data. Bioinformatics 2016;32:3380–3387.

2. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–2069.

3. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3693.

4. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–2605.

5. Krijthe J. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation (R package version 0.10). Computer Software.

6. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV, et al. Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology. 2014.

7. 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.

8. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012;19:455–477.

9. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.

10. Bischl B, Lang M, Kotthoff L, Schiffner J. mlr: Machine learning in R. J Mach Learn Res 2016;17:1–5.

4

Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus faecium*

S. Arredondo-Alonso, J. Top, A. McNally, S. Puranen, M. Pesonen, J. Pensar, P. Marttinen, J. C. Braat, M. R. C. Rogers, W. van Schaik, S. Kaski, R. J. L. Willems, J. Corander, A. C. Schürch

Published in: mBio (2020) doi: 10.1128/mBio.03284-19

Abstract

Enterococcus faecium is a gut commensal of humans and animals but is also listed on the WHO global priority list of multidrug-resistant pathogens. Many of its antibiotic resistance traits reside on plasmids and have the potential to be disseminated by horizontal gene transfer. Here, we present the first comprehensive population-wide analysis of the pan-plasmidome of a clinically important bacterium, by whole-genome sequence analysis of 1,644 isolates from hospital, commensal, and animal sources of *E. faecium*. Long-read sequencing on a selection of isolates resulted in the completion of 305 plasmids that exhibited high levels of sequence modularity. We further investigated the entirety of all plasmids of each isolate (plasmidome) using a combination of short-read sequencing and machine-learning classifiers. Clustering of the plasmid sequences unraveled different *E. faecium* populations with a clear association with hospitalized patient isolates, suggesting different optimal configurations of plasmids in the hospital environment. The characterization of these populations allowed us to identify common mechanisms of plasmid stabilization such as toxin-antitoxin systems and genes exclusively present in particular plasmidome populations exemplified by copper resistance, phosphotransferase systems, or bacteriocin genes potentially involved in niche adaptation. Based on the distribution of k-mer distances between isolates, we concluded that plasmidomes rather than chromosomes are most informative for source specificity of *E. faecium*.

Introduction

Enterococcus faecium ranks among the most frequent causative agents of hospital- acquired infections, specifically, central-line associated bloodstream infections (1). The burden of disease due to *E. faecium* is augmented by the fact that *E. faecium* has acquired resistance against almost all available antibiotics, most notably, against ampicillin, gentamicin, and vancomycin and less frequently against the more recently introduced antibiotics linezolid, daptomycin, and tigecycline (2). Antibiotic resistance, including vancomycin resistance, is not a feature exclusively found among hospitalized patient isolates, as *E. faecium* isolates from farm animals also contain these resistance traits (3).

Previous whole-genome sequencing (WGS)-based studies split the *E. faecium* population into two lineages corresponding to a hospital-associated clade (clade A) and a community-associated clade (clade B) (4, 5). Subsequently, clade A was first subdivided into clade A1, mainly represented by clinical isolates, and clade A2, with a majority of animal isolates (6). Recent reports indicated that animal isolates do not form a monophyletic subclade and no longer support the split of clade A isolates into two single subclades (2, 7).

Plasmids can act as vehicles for the transmission of virulence and antimicrobial resistance genes (8). Several mechanisms of plasmid-mediated resistance have been described in *E. faecium* (9, 10), including glycopeptide resistance caused by the presence of *vanA* and *vanB* gene clusters (Tn*1546* and Tn*1549*, respectively), aminoglycoside resistance caused by the presence of *aac(6=)-Ie-aph(2)* gene (Tn*5281*), tetracycline resistance mediated by *tet(M)*, linezolid resistance due to the presence of *cfr*, *cfr(B)*, *optrA*, and *poxtA*, or quinupristin-dalfopristin resistance due to plasmids harboring *vat(D)* and *vat(E)*.

Enterococcal plasmids have been conventionally grouped in four main family groups (RepA_N, Inc18, RCR, and Rep_3) based on their sequence homology against known replication initiator proteins (RIP) (11). The presence of conjugation systems and mobilization systems in *enterococcal* plasmids suggests that horizontal gene transfer (HGT) may act as a major source of DNA mobility between *E. faecium* hosts (11). Previous attempts to investigate the mobilome and HGT in *E. faecium* have been restricted to microarray-based studies using custom-designed probes (12).

In this study, we sequenced the genomes of 1,644 clade A isolates from human (hospitalized patients and nonhospitalized persons) and animal (pet, farm, and wild animals) sources using short-read sequencing technology. We elucidated complete plasmid sequences from a representative subset of 62 isolates by long-read sequencing, resulting in 305 complete plasmids. Furthermore, we used a recently developed machine-learning classifier (mlplasmids) to predict the plasmidome of *E. faecium* isolates with only short-read sequencing data (13). Using this novel genomic tool, we accurately predicted and defined the plasmidome of all isolates that were sequenced as part of this study, which allowed

the study of the population pan-plasmidome of *E. faecium* in terms of plasmid k-mers and gene diversity in the clade A isolates. Our analysis shows that the plasmidome rather than the chromosome of *E. faecium* is most informative for understanding niche adaptation.

Results

Core gene phylogeny confirms distinct clustering of hospitalized patient isolates

To determine the core genome variability of clade A *E. faecium* isolates, we constructed a core gene alignment for 1,644 isolates of *E. faecium* clade A. This alignment was filtered for recombination, and the remaining variable sites were analyzed to classify the 1,644 isolates into (85) sequence clusters (SCs) using hierBAPS (postBNGBAPS.2 group) (see Data Set S1 in the supplemental material). In total, 955 genes (orthologous groups) were used to reconstruct the population phylogeny of our *E. faecium* collection (Fig. 1A) (https://microreact.org/project/BJKGTJPTQ).

In accordance with previous *E. faecium* population studies, we split the 1,644 *E. faecium* isolates into clade A1 and non-clade A1 isolates (Fig. 1). Hospitalized patient isolates (1,142) were mostly designated clade A1 (1,098; 96%), representing the most frequent source in this clade (1,098/1,227 [89%]). We also identified clade A1 isolates in nonhospitalized persons (18) and pets (102) (Fig. 1B). Furthermore, pet isolates represented the biggest nonhospital source (78%) present in clade A1 (Fig. 1B). These pet isolates were mainly from dogs from the Netherlands, randomly selected in an unbiased nationwide survey of healthy pet owners with no recent antibiotic usage history. In this survey, cocarriage of vancomycin-resistant *E. faecium* between owners and dogs was not observed (14).

Human community isolates from nonhospitalized patients were widely dispersed over the phylogenetic tree outside clade A1 (Fig. 1A). Farm animal isolates, represented in this study mostly by isolates from poultry and pigs, clustered in clade A distinct from the hospital clade A1 in polyphyletic groups, confirming that there is no distinct clade A2 representing isolates from farm animals (2, 7), in contrast to what was reported previously (6). Pig and poultry isolates were grouped in a limited number of distinct SCs, with 88% of pig isolates grouping in SCs 29 and 30 and 93% of poultry isolates grouping in SCs 24, 25, and 35 (Data Set S1).

Completed plasmid sequences show extensive modularity

To elucidate whether plasmids have shaped the observed *E. faecium* population structure, we first fully resolved the plasmids of *E. faecium* by performing Oxford Nanopore Technologies sequencing (ONT) and subsequently constructed a hybrid assembly of 62 *E. faecium* isolates. These isolates were selected to capture the highest plasmidome variability present in our 1,644 clade A *E. faecium* isolates based on PlasmidSPAdes predictions (15) and a homology search against a curated database of replication initiator proteins in en-

terococci (11), as previously described (13) (see Text S1). Hybrid assemblies resulted in 48 completed (finished) chromosome sequences (and 14 chromosomes distributed among two contigs or more), 305 plasmids, and 6 phage sequences present in single circular contigs (Data Set S1). The 48 complete chromosomes ranged in size from 2.42 to 3.01 Mbp. Hospitalized patient isolates (n=32) had the largest chromosomes (mean, 2.82 Mbp), whereas poultry isolates (n=2) carried the smallest chromosomes (mean, 2.42 Mbp). Notably, hospitalized patient isolates had up to 20% larger chromosomes than *E. faecium* from other sources, which highlights the considerable genomic flexibility of this organism.

The set of 305 completed plasmid sequences ranged in length from 1.93 to 293.85 kbp (median, 15.15 kbp; mean, 53.48 kbp) (Fig. 2A, S1, and S2). Hospitalized patient isolates (n=43) with complete plasmid sequences (n=247) contained the highest number of plasmids (mean, 5.70), and their cumulative plasmid length was substantially larger than those from other isolation sources (mean, 308.01 kbp).

We characterized these plasmids using a standard classification (11) based on (i) presence of replication initiator proteins (RIP) (Data Set S1) and (ii) presence of relaxases (MOB) (Data Set S1). A considerable proportion of plasmids (48/294 [16%]) were multireplicon plasmids, with plasmids encoding up to four different RIP gene families, indicating a high degree of plasmid modularity (see Fig. S1). This was most prominent in Rep_1 and Inc18 family plasmids, which contained at least one other RIP with a frequency of 1.0 (8/8) and 0.53 (30/57) (Fig. 2B), respectively. The predominant RIP family RepA_N (n=82) was mainly encoded by large plasmids (mean plasmid length, 155.3 kbp) and was less frequently associated with other RIP sequences (n=15, 18%) (Fig. 2B). Plasmids encoding the Rep_3 family (n=56; mean plasmid length, 12.4 kbp) and Rep_trans (n=24; mean plasmid length, 25.7 kbp) were less frequently present in multireplicon plasmids (n=6, 11%) (Fig. 2B). No RIP family was characterized for 11 plasmids (mean plasmid length, 9.6 kbp).

The observed modularity of *E. faecium* plasmids became even more apparent when relaxase gene families were linked to the fully sequenced plasmids. All identified relaxases cooccurred in plasmids with different RIP genes and even in multireplicon plasmids (Fig. 2B). In total, we observed 46 different Rep-relaxase combinations (Fig. 2B). A more extensive characterization of mosaicism of plasmid sequences is available in Text S1.

Hospitalized patient isolates have the largest predicted plasmidome sizes

To predict the plasmidome content present in the other 1,582 *E. faecium* isolates that were only sequenced with short-read technology, we previously used the information derived from the completed plasmid sequences to develop and validate a machine- learning classifier called mlplasmids (13). The classifier achieved an accuracy of 0.95 and an F1 score (harmonic mean between precision and recall) of 0.92 on a test set of *E. faecium* sequences generated by short-read sequencing. A more extensive description of the classifier

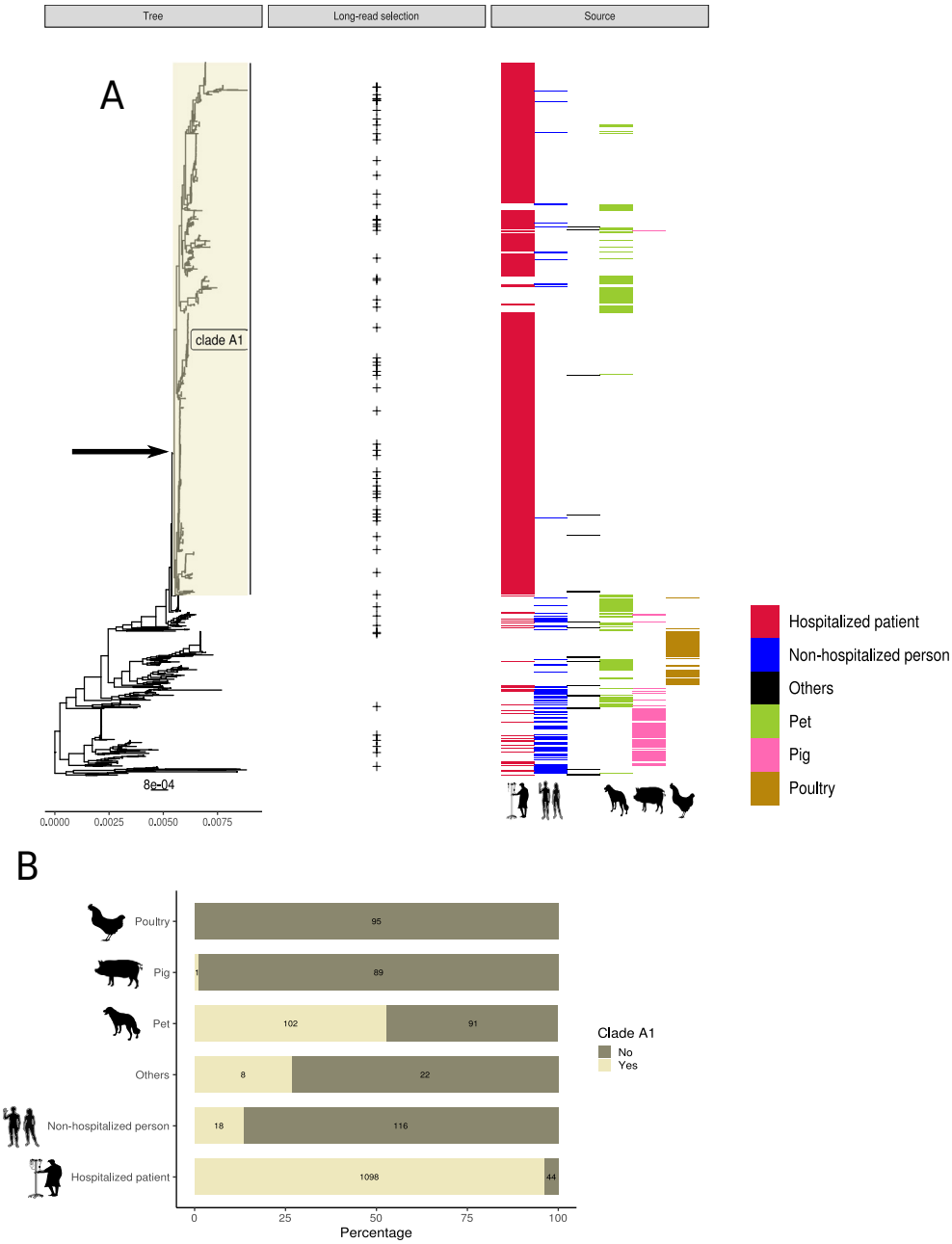


Figure 1. (A) RAxML tree based on 955 E. faecium core genes in 1,644 clade A strains. Isolates selected for long-read sequencing are indicated with under long-read selection. Isolates were colored based on their isolation source: hospitalized patients (red), nonhospitalized persons (blue), pet (green), pig (pink), poultry (brown), and other sources (black). Arrow in the RAxML tree indicates the internal node 1227 used to split the clade A1 and non-clade A1 isolates. (B) For each isolation source (x axis), we specified the count and percentage (y axis) of isolates belonging or not to clade A1.

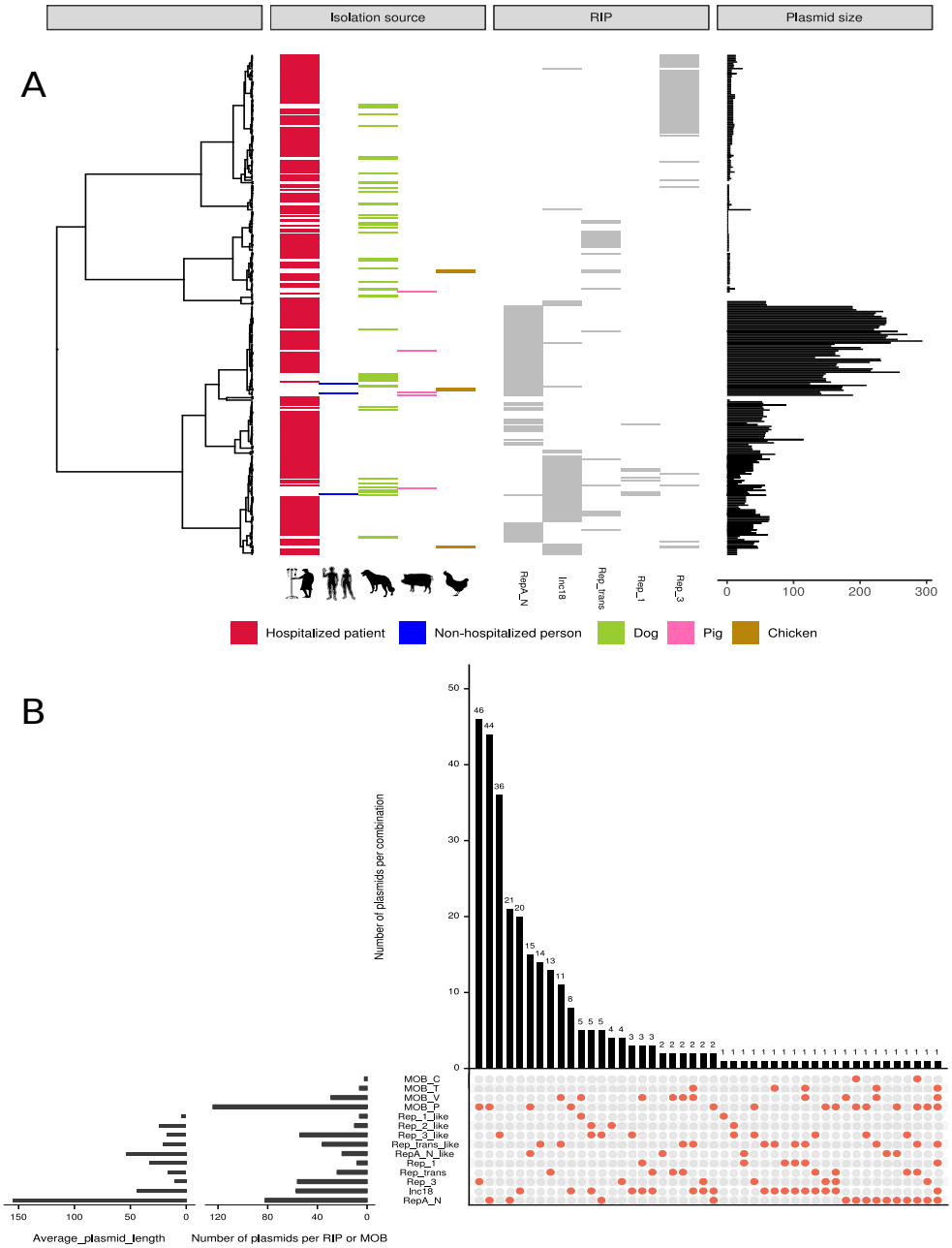


Figure 2. Overview of completed plasmid sequences (n=305). (A) Pairwise Mash distances (k=21, s= 1,000) of the completed plasmid sequences (n=305) were transformed into a distance matrix and clustered using hierarchical clustering (ward.D2). Node positions in the dendrogram were used to sort and represent in different panels: (i) isolation source, (ii) replication initiator gene (RIP), and (iii) plasmid size (kbp) of the completed plasmid sequences. (B) Intersection plot of the combination of RIP and relaxases found in the set of completed plasmid sequences with associated RIP sequences (n 294).

validation and its performance compared to that of existing plasmid prediction tools can be found in the study by Arredondo-Alonso et al. (13).

mlplasmids was used on the present collection of *E. faecium* isolates, resulting in an average number of base pairs predicted as plasmid derived of 240,324 bp (52 contigs), while the average number of chromosome-derived base pairs was 2,619,359 bp (113 contigs) per isolate. mlplasmids did not predict plasmid-derived contigs in four isolates, including one isolate that was previously described as plasmid-free (64/3, in this study named E2364) (16).

We observed significant differences in the number of base pairs predicted as plasmid derived depending on the source of the *E. faecium* isolates ($P<0.05$) (Fig. 3A). Predicted plasmidome size of isolates from hospitalized patients was considerably larger (mean, 276.16 kbp; $P<0.05$) than that from other isolation sources (Fig. 3). This finding is in line with previous reports which showed that isolates from clade A1 are enriched for mobile genetic elements (6, 17).

Plasmidome populations are strongly associated with isolation source

To structure the pan-plasmidome of *E. faecium*, we determined pairwise distances of isolates based on the k-mer content of their predicted plasmidome. We computed a neighbor-joining tree (bioNJ) to cluster *E. faecium* isolates exclusively on the basis of gain and loss of plasmid sequences (Fig. 4A). During this analysis, 37 isolates were excluded, as they showed no signs of plasmid carriage signatures based on their distribution of pairwise distances (see Fig. S3).

To evaluate the core genome clonality of isolates clustering in the same plasmidome population, we incorporated information regarding isolation source and SCs into the plasmidome tree (Fig. 4A) and core genome phylogeny (Fig. 4B). Isolates with a similar plasmidome contents but different SCs were positioned in different parts of the core genome phylogeny (Fig. 4B), which could be indicative of horizontal transmission of plasmid sequences.

To quantify and formalize these observations of horizontal or vertical transfer of plasmid sequences, we estimated clusters of isolates with similar plasmidomes. The k-mer distances of the plasmidomes were clustered using hierarchical clustering (ward.D2), and we estimated an optimal number of 26 clusters (average silhouette width, 0.45) (Fig. S4A). To enable meaningful statistical inferences, we only considered clusters that contained more than 50 isolates and had an average silhouette width, as a measure of goodness of fit, higher than 0.3 (Fig. S4B). This resulted in 9 clusters that are referred to as plasmidome populations 1 to 9 (Fig. 3B, S4, and S5). We then calculated the SC diversity of all isolates of each plasmidome population (Simpson index) and tested for enrichment of particular isolation sources (Fig. S4B). However, these plasmidome populations may be driven by

the k-mer content of large plasmid sequences and could obscure the potential transfer of small plasmid sequences between isolates. An extensive evaluation of the plasmidome populations and potential transfer of the complete plasmid sequences obtained in our study is described in Text S1.

To evaluate the influence of other factors than source category to explain the plasmidome clustering, we modeled the observed plasmid k-mer distances using three linear regression models with three different covariates: source, isolation time, and geographical distance between pairs of isolates. We observed that modeling k-mer distances using exclusively source explained 39% of the variance present in the plasmid k-mer distances, whereas using time (difference in years between the isolates) as covariate explained 29% of the variance. Geographical distance between isolates explained less than 1% of the variance. Finally, we incorporated the three predictors into a multiple linear regression model, which increased the variance explained up to 43%. This elucidated that isolation source was the most important predictor to explain plasmidome clustering, but a difference in time between strains must be considered: isolates which are circulating during the same period of time are more likely to share plasmid sequences. Geographical distance between isolates seems not relevant to explain the observed clustering, which suggests a high mobility and spread of *E. faecium* plasmid sequences.

Restriction modification systems, but not CRISPR-Cas, could act as barriers of horizontal gene transfer

The absence of CRISPR-Cas systems in clade A1 isolates was previously postulated as a plausible explanation for a nondiscriminatory accumulation of plasmid sequences in clade A1 isolates (6, 18). However, we only observed a CRISPR-Cas system in a single non-clade A1 isolate and no occurrence of the recently described Jet system in any of the isolates (19). The absence of a CRISPR-Cas system is therefore unlikely to result in a higher and different plasmidome content of clade A1 isolates from hospitalized patients.

Recently, a novel defense mechanism consisting of a restriction modification (RM) system was postulated as contributing to the subspeciation of *E. faecium* (20). The specificity of the RM system resides in the S subunit, which binds to different DNA sequences by two target recognition domains. In our collection, we also identified the S subunit (WP_002287733) as present and enriched in clade A1 isolates ($P<0.05$), whereas the subunits M and R were identical in both clade and non-clade A1 isolates and always present together with the S subunit. Furthermore, we identified 8 novel S subunit variants in our set of 62 isolates with complete genome sequences. Of these, four variants (E1774_00555, E7313_02981, E4413_00571, and E4438_00276) were significantly enriched in clade A1, while two other variants (E0139_00520 and E4227_ 02943) were enriched in non-clade A1 isolates, which reinforces the hypothesis that different RM systems contribute to the differentiation of the

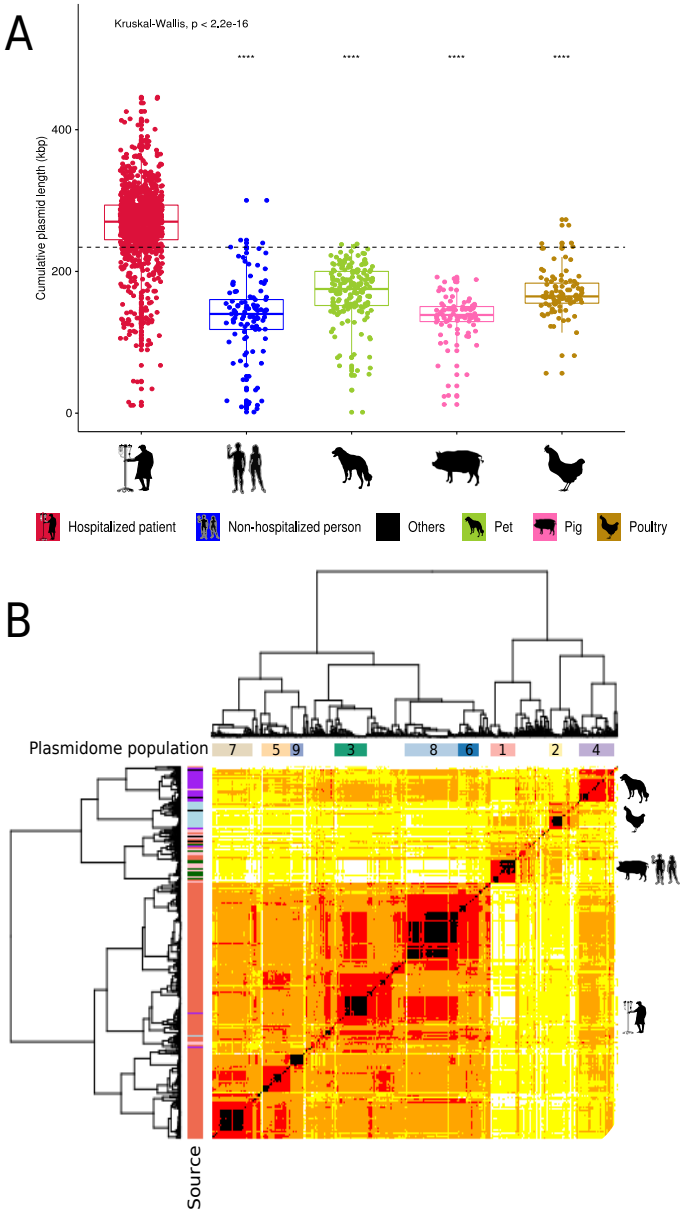


Figure 3. Predicted pan-plasmidome of 1,644 *E. faecium* isolates. (A) Boxplot of the numbers of base pairs (kbp) predicted as plasmid derived per isolation source. Horizontal dashed line indicates the mean cumulative plasmid length across all the groups. (B) Pairwise Mash distances (k =21, s=1,000) of plasmid-predicted contigs in 1,607 isolates were transformed into a distance matrix and clustered using hierarchical clustering (ward.D2). Based on the quantile function of our defined gamma distribution, we grouped isolates in five blocks: black (0 to 0.01), red (0.01 to 0.25), orange (0.25 to 0.5), yellow (0.5 to 0.75), and white (0.75 to 1.0). Dissimilarity matrix of the isolates was visualized as a heat map colored based on the previous blocks. We incorporated the defined plasmid populations (n=9) and isolation source information on top and left dendrograms, respectively.

plasmidome content between isolation sources (Text S1).

Characterization of genes driving the plasmidome populations

To identify which genes were potentially driving the observed plasmidome populations (n=9), we determined, for each plasmidome population, which genes were present in more than 95% of the isolates and defined those as plasmidome population core genes. We further characterized these genes using eggNOG to retrieve the cluster of orthologous genes (COG) and associated KEGG pathways. These plasmidome population core genes were then searched in our set of complete plasmid sequences to identify the type of replicon sequences bearing these genes, such as large RepA_N or Inc18 plasmids.

Most of the plasmidome population core genes belonged to COG S (unknown function) and COG L (DNA replication, recombination, and repair) (Fig. S6; Data Set S1). Within these

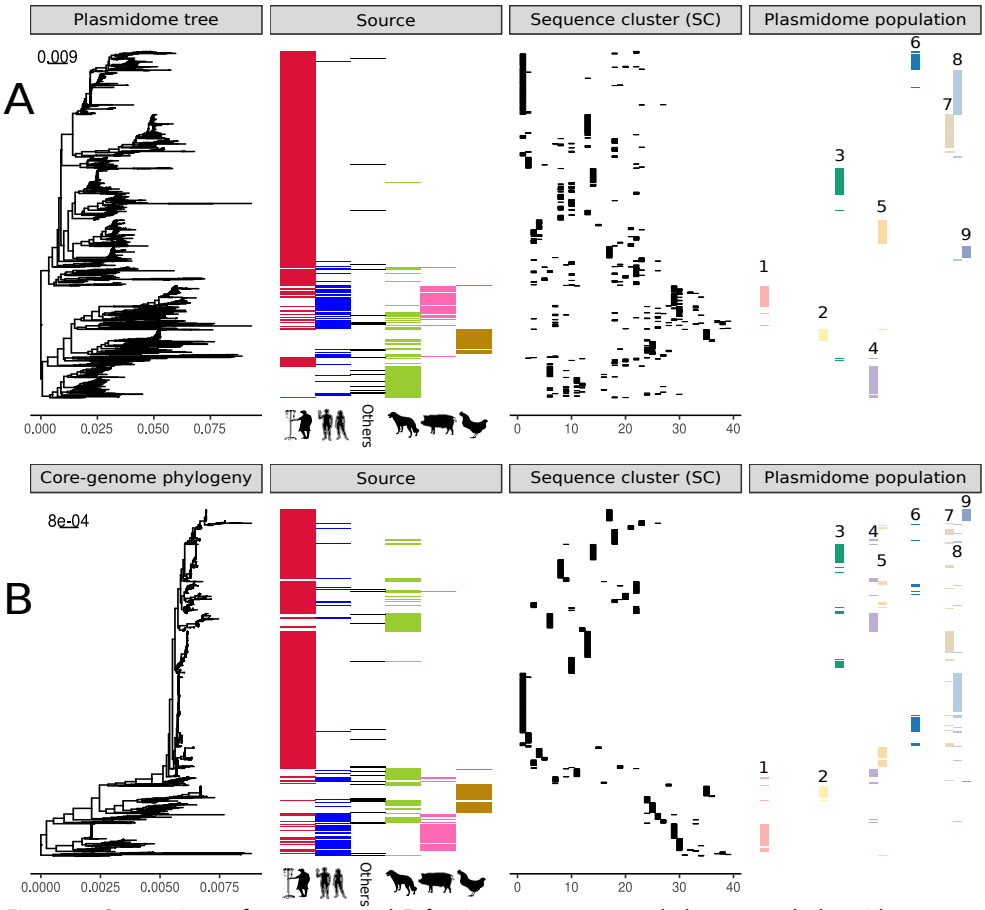


Figure 4. Comparison of reconstructed *E. faecium* core genome phylogeny and plasmidome trees. The figure includes three different panels: isolation source, sequence cluster (SC), and plasmidome population. (A) bioNJ tree based on the dissimilarity matrix of Mash distances (k=21, n=1,000) of 1,607 isolates uniquely considering plasmid-predicted contigs. (B) RAxML core genome tree based on 955 *E. faecium* core genes in 1,644 clade A strains.

two COG groups, we identified functions such as toxin-antitoxin (TA) systems, involved in the stabilization of large plasmid sequences (e.g., RelE/AbrB, MazEF, and HicAC systems), and a type IV TA "innate immunity" bacterial abortive infection (Abi) system that protects bacteria from the spread of a phage infection (AbiEi/AbiEii). This TA system interferes with phage RNA synthesis, enables stabilization of mobile genetic elements (21), and was extensively described in lactococcal plasmids (22).

Interestingly, we identified some plasmidome population core genes only present in particular populations. For plasmidome population 1 (pig and nonhospitalized isolates), we identified a copper resistance operon (*tcrYAZB*) that provides a mechanism to tolerate high concentrations of this heavy metal as plasmidome population core genes. Copper was commonly used as a growth-promoting agent for pigs (23). However, high levels of copper result in toxicity for the bacterial cells. The *tcrYAZB* operon provides a plasmid survival mechanism to tolerate high concentrations of this heavy metal. In addition, we identified the glycopeptide resistance-encoding *vanA* gene cluster as a plasmidome population core in the population. These genes were harbored on a RepA_N conjugative plasmid of 140 kbp (LR132068.1 and LR135180.1) and colocalized with genes encoding plasmid stabilization systems (RelE/AbrB and AbiEi/AbiEii), which may explain the persistence of this large plasmid in the population.

Plasmidome population 2 (poultry associated) also showed plasmidome population core genes which were exclusively present as core in this population. This included the bile salt hydrolase (BSH) choloylglycine hydrolase and putatively a tetronasin resistance-encoding permease gene. BSH is involved in the deconjugation (hydrolysis) of bile acids, which have antimicrobial activity, especially against Gram-positive bacteria (24). Therefore, acquisition of BSH could serve as a selective advantage for *E. faecium* for gut colonization. In a recent review, BSHs have been described as the gatekeepers of bile acid metabolism and host-microbe cross talk in the gastrointestinal tract (25). In addition, as mentioned, homologous searches revealed only hits for *E. faecium* strains isolated from chicken, but we also obtained hits for *Enterococcus cecorum* (100% similarity in amino acids [AA]), which is a species mainly found in birds. In both strains, BSH was located downstream of the same site-specific recombinase, which highly suggests HGT between these 2 species. We also observed a tetronasin resistance gene as a plasmidome population core gene. The presence of this gene on a mobile element among *E. faecium* poultry isolates was previously described and may be the result of selective pressure due to the wide use of ionophores, e.g., tetronasin for coccidiosis prophylaxis in poultry (26). Interestingly, this gene is often colocated on a plasmid with Tn*1546* encoding vancomycin resistance and TA systems.

In the case of the hospital-associated plasmidome populations (3, 5, 6, 7, 8, and 9), we characterized some genes present in all these populations. Of these, a locus of three ge-

nes putatively encodes an ABC transport system, while one gene encodes an ATP-binding protein and the other two genes encode permeases. These genes were assigned to COG V (defense mechanisms) and were similar to the previously described *vex* locus of *Streptococcus pneumoniae*. In *S. pneumoniae*, this gene cluster was initially linked to vancomycin tolerance (27), but Moscoso and coauthors disproved these results (27, 28). Protein analysis of the ATP binding protein Vex2 revealed the presence of domains with similarity to lipoprotein/bacteriocin/macrolide export systems, which may suggest that this system is involved in antibiotic resistance. We also observed antimicrobial resistance genes such as aminoglycoside resistance (*aacA-aphD*) and erythromycin resistance (*erm*) present in the plasmidome population core of all the hospitalized patient populations.

In line with the hypothesis of different routes of hospital adaptation, we observed some plasmidome population core genes that are only present as core in some plasmidome populations associated with hospitalized patients. We observed the presence of a bacteriocin with homology to *BacA* in populations 5, 7, and 8 and previously described as a plasmid-borne bacteriocin in *E. faecalis* (29). *BacA* can act as a more evolved toxin-antitoxin system in which not only daughter cells but also cells from the same generation not bearing the *BacA* plasmid are excluded. Furthermore, it was demonstrated that plasmid dissemination was more prominent under conditions of fluctuations in the population of *E. faecium*, since *BacA* activity exclusively affects dividing cells (29). We also observed a complete phosphotransferase system putatively involved in mannose/fructose/sorbose utilization present in the plasmidome cores of populations 6 and 7. This may provide novel pathways for the utilization of complex carbohydrates in these hospital-associated populations.

A complete characterization of the plasmidome population core genes and the complete plasmid sequences in which these genes are located can be found in Text S1.

Plasmidome content is the major genomic component driving niche specificity

To assess which of the genomic components (chromosome or plasmidome) contributed most to source specificity, we compared the distributions of k-mer pairwise distances using three different inputs: (i) whole-genome contigs, (ii) chromosome- derived contigs, and (iii) plasmid-derived contigs. We hypothesized, for source-specific components, that k-mer distances between pairs of isolates belonging to the same source were lower than pairs of isolates from different or random sources. This difference can reflect the association strength between niche and genomic component (whole-genome, chromosome-derived, and plasmid-derived contigs). We followed a bootstrap approach to compare and average k-mer pairwise distances of (i) pairs of isolates from the same isolation source (within-source group), (ii) pairs of isolates belonging to different isolation sources (between-source group), and (iii) pairs of isolates randomly selected (random group).

Whole-genome contigs explained most of the source specificity of all the isolation sources except for nonhospitalized person isolates, based on the highest k-mer pairwise distance differences between isolates from the same source (within source) and randomly selected isolates (Fig. 5 and S7).

However, with the exception of nonhospitalized person isolates, the plasmidome contribution was higher than the chromosome contribution to explain source specificity. This was based on the highest difference in k-mer pairwise distances between isolates from the same (within-source group) and different (between-source group) sources when comparing the plasmidome versus the chromosome (Fig. 5 and S7).

Most notably, we observed significant similarities of the whole genome and chromosome of dog and hospitalized patient isolates (positive difference, 0.20; $P < 0.05$) but a significant dissimilarity between these two sources when considering their plasmidomes (negative difference, 0.13; $P<0.05$) (Fig. 5 and S7). In addition, pig and nonhospitalized person isolates had significantly similar plasmidomes as observed by a small difference in k-mer distances (positive difference, 0.15; $P<0.05$), corroborating the postulated exchange of plasmid sequences between these two groups (Fig. S7).

Discussion

We used a combination of ONT long-read and Illumina short-read technologies to perform a comprehensive analysis of the pan-plasmidome of the nosocomial pathogen

E. faecium which has evolved in different niches. The high number of multireplicon plasmids consisting of several combinations of RIP families confirmed the high levels of mosaicism previously observed for *E. faecium* plasmids, which challenges the classification of *Enterococcal* plasmids based on RIP schemes (30).

We observed that the total plasmidome size of isolates from hospitalized patients was substantially larger than that from animal isolates and isolates from nonhospitalized persons. Moreover, clustering of k-mer pairwise distances from our set of predicted plasmid sequences revealed a high level of diversity in *E. faecium* plasmidomes. We estimated the potential contribution of different genomic components (whole genome, chromosome, and plasmid) to source specificity and observed that the plasmidome explains source specificity in dogs and hospitalized patients, while their corresponding core genomes share an evolutionary history. This finding suggests that either the hospital-adapted population was founded by a host jump from the canine population or, alternatively, the host jump happened in the other direction. In line with previous reports (31, 32), we observed that nonhospitalized person isolates in our collection shared their plasmidomes with pig isolates, which indicates an exchange of plasmids or strains between both sources.

Source specificity of plasmid sequences was highest in pigs and poultry isolates and sig-

nificantly differed from the other sources, but also, the plasmidomes of clinical isolates were highly dissimilar to isolates from other sources. This suggests that the pan-plasmidome of *E. faecium* plays a role in the emergence of this organism as a nosocomial pathogen of major importance. There was not, however, a single preferred plasmidome configuration for hospital patient isolates, but rather, these isolates were associated with six different plasmidome populations, indicating different possible routes of plasmid acquisition within the hospital environment.

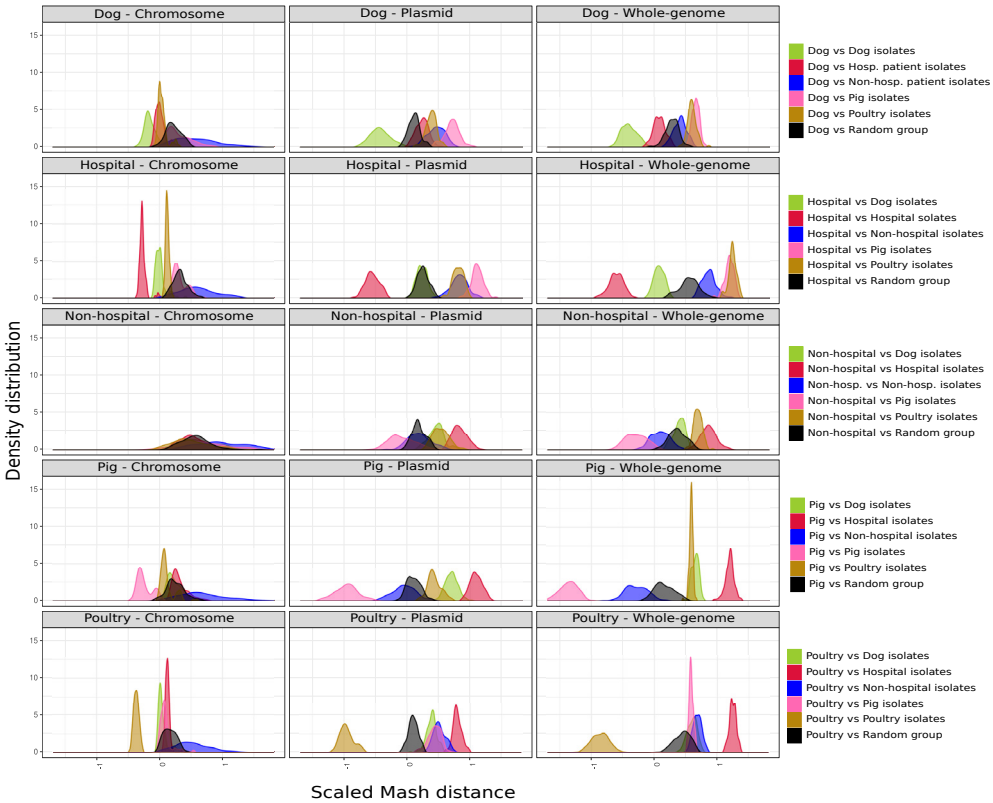


Figure 5. Evaluation of the source specificity from each genomic component. Mash distances computed from chromosome-predicted (first column), plasmid- predicted (second column), and whole-genome (third column) contigs were scaled and compared between all the isolation sources. Each row corresponds to a particular isolation source (e.g., first row refers to dog isolates) and the distribution of pairwise distances against other sources (dog in green, hospitalized patient in red, nonhospitalized person in blue, pig in pink, poultry in brown, and random isolates in black) for each genomic component. These average distances were computed using a bootstrap approach (100 iterations). The distribution of pairs of isolates from the same source type with respect to the distribution of pairs from random isolates (black group) reflects the specificity of the genome component in each source. If pairs from the same source deviate to the left, it indicates a higher specificity of that particular genomic component, whereas a deviation to the right with respect to the pairs of random isolates (black group) indicates a lower specificity than expected by chance.

The existence of distinct host-associated plasmidome populations indicates that the dissemination of plasmids within the *E. faecium* population is restricted. The presence of particular S subunit variants belonging to a type I RM system enriched either in clade A1 isolates or non-clade A1 isolates in the *E. faecium* population suggests that they play an active role as HGT barriers between isolates from different sources (20). Restriction modification systems potentially limit the exchange of plasmid sequences and might contribute to source specificity. In a few cases, we observed the presence of single isolates from a specific source in plasmidome populations dominated by a different source, as exemplified in the case of plasmidome population 4 (dog enriched) and the hospitalized patient isolate E8172. In this case, we identified a similar RepA_N conjugative plasmid potentially transmitted from or to dogs to or from that particular hospitalized patient's isolate. The presence of identical S subunit variants between hospitalized patient and dog isolates (clade A1 enriched) could enable an occasional exchange of plasmid sequences between different sources.

Exploration of the core genes of the predicted plasmidome populations revealed that most plasmid genes are poorly characterized. We further characterized some of the plasmid genes with an unknown function as toxin-antitoxin systems. The widespread occurrence of these selfish systems is indicative of their importance in plasmid maintenance and stabilization. Previous reports have shown a high prevalence of particular toxin-antitoxin systems, such as *mazEF*, in *E. faecium* clinical isolates (33). They could contribute to the stabilization of plasmid-mediated antibiotic resistance by the maintenance of a single plasmid structure and might thus provide an interesting alternative target for antibiotic therapy.

We also identified a set of copper resistance genes (*tcrYAZB* operon) in the core plasmidome of population 1 (pig and nonhospitalized associated). Copper was used as a growth-promoting agent in piglets (34), and high levels of copper are toxic for most bacterial species. The acquisition of copper resistance genes may have contributed to the adaptation of *E. faecium* to environmental constraints imposed by pig farming. Recently, Gouliouris et al. also described the same copper resistance operon as over- represented in pig isolates, thus confirming that this set of plasmid-borne genes has played an important role in *E. faecium* survival in farms (35). Those plasmid genes were identified in our set of complete plasmid sequences and were present in a RepA_N conjugative plasmid (140 kbp) identified in pig and nonhospitalized isolates. Furthermore, we identified a BSH gene widely present in the poultry-associated plasmidome population. *E. faecium* was previously characterized as one of the microorganisms with the highest level of BSH activity in the intestines of chickens (36) and capable of developing new mechanisms to tolerate a high concentration of bile salts (37). The BSH gene described here could be functionally

responsible for the bile tolerance of poultry isolates.

The presence of several plasmid genes involved in carbohydrate metabolism and utilization in plasmidome populations associated with hospitalized patients may indicate the acquisition of novel pathways to process complex carbohydrates. This observation is in line with previous reports (6, 38) in which phosphotransferase systems enriched in clade A1 isolates and encoded by mobile genetic elements were fundamental for *E. faecium* during gastrointestinal (GI) tract colonization. The high frequency of plasmid genes with an unknown function or corresponding to hypothetical proteins could mask the presence of other plasmid-mediated mechanisms contributing to niche adaptation. This highlights the importance of further functional studies to elucidate the roles of these plasmid genes.

The observations that plasmid sequences are highly informative for source specificity and that particular genes may have a clear benefit for *E. faecium* in particular niches suggest that the distribution of plasmid genes among *E. faecium* isolates is regulated by complex ecological constraints, and thus contributes to niche adaptation, rather than by opportunities arising from physical interactions between different sources. Of note, this approach does not calculate the contribution of a single genomic sequence but of the whole genomic component (plasmid or chromosome) to the niche specificity. Small chromosomal alterations or rearrangements could also be involved and play an important role in niche specificity.

Based on our findings, we elucidated that isolation source was the most important predictor to explain the observed plasmidome clustering and indicated that isolates from the same niche can exchange plasmid sequences during the same time frame. Combining extensive short- and long-read sequencing of a large collection of isolates from a diverse set of sources, as reported here for *E. faecium*, may serve as a broadly applicable approach to study the pan-plasmidome of evolutionary and ecologically diverse populations.

Material and Methods

Genomic DNA sequencing, assembly, and characterization of plasmids

Detailed description of Illumina and ONT sequencing is available in Text S1 in the supplemental material and in the study by Arredondo-Alonso et al. (13), which includes a full description of ONT selection of E. faecium isolates (n = 62) and consecutive hybrid assembly using Unicycler (39). Characterization of fully assembled plasmids is also described in Text S1.

Population genomic analysis

Pangenomes for the entire genome data set (1,684 strains) and the clade A data set (1,644 strains) were created using Roary (40) with default settings. A core gene alignment was generated using the –mafft option in Roary, resulting in a core gene alignment of 859 genes

for the entire data set and of 978 genes for the clade A data set. To estimate recombination events and to remove them from the core genome alignment, we used BratNextGen with default settings, including 20 hidden Markov model (HMM) iterations, 100 permutations run in parallel on a cluster, and 5% significance level, similar to those in earlier publications (41, 42). To determine sequence clusters (SCs) in the core genome alignment where significant recombinations had been removed, we used 5 estimation runs of the hierBAPS method (43) with 3 levels of hierarchy and the prior upper bound for the number of clusters ranging in the interval 50 to 200. All runs converged to the same estimate of the posterior mode clustering. We considered the second level of hierarchy (postBNGBAPS.2) to determine SCs in our collection. To estimate a phylogenetic tree, we used RAxML (44) with the GTR Gamma model on a core gene alignment stripped of recombination. The bootstrap option was disabled in RAxML due to an extremely long runtime.

CRISPR-Cas and restriction modification system detection

To detect CRISPR-Cas arrays present in our set of 1,644 *E. faecium* isolates, we first used CRISPRDetect (version 2.2) (45), and detected hits were further validated using CRISPR-CasFinder (version 1.1.1) (46).

To observe the presence of the restriction modification system described by Huo et al. (20), we retrieved the nucleotide sequences of the S subunit (WP_002287733.1), M subunit (WP_002287732.1), and R subunit (WP_002287735.1) from the *E. faecium* genome sequence (NZ_GG688488). We screened for the presence of these subunits in our entire collection of isolates (1,644) using Abricate and defined a 95% minimum identity and 90% coverage as thresholds (version 0.8.2). Later, we focused our analysis on the set of complete genome isolates (62) and performed a multiple-sequence alignment on the protein level of all the S subunits identified using Clustal Omega (version 1.2.4) (47). Based on the multiple-sequence alignment, we defined 8 novel S subunit variants that were tested for enrichment in either clade A1 or non-clade A1 isolates using a Fisher exact test with the function fisher.test from R stats package (version 3.4.4).

Predicting the plasmidome content of short-read sequenced *E. faecium* isolates

To determine the plasmidome content of the remaining 1,582 isolates, we used mlplasmids (13). mlplasmids (version 1.0.0) was run, specifying "*Enterococcus faecium*" model and a minimum contig length of 1,000 bp. For further analysis, we discarded predicted contigs with a posterior probability lower than 0.7 of belonging to the assigned class (chromosome/plasmid; https://gitlab.com/sirarredondo/efaecium_population/raw/ master/Files/mlplasmids_prediction/prediction_svm.tsv). Differences in the numbers of chromosome- and plasmid-derived base pairs predicted by mlplasmids between hospitalized patient isolates and other isolation sources were assessed using the Kruskal-Wallis test (significance threshold, 0.05) available in ggpubr package (version 0.1.7) (48).

We calculated pairwise Mash distances (k=21, s=1,000; version 1.1) between isolates (n 1,640), only considering plasmid-predicted contigs. We reconstructed a plasmidome tree with the bioNJ algorithm implemented in the R ape package (version 5.1) using computed Mash distances (49, 50). The resulting phylogenetic tree was midrooted using the midpoint function in the R phangorn package (version 2.4.0) (51). To improve the resolution of the bioNJ tree, we observed the distribution of the computed Mash distances and fitted a gamma distribution using the fitdist function (distr="gamma" ,and method="mle") available in the R fitdistrplus package (52). We discarded isolates with an average pairwise mash distance superior to 0.12, which was calculated using the qgamma function (P= 0.9, shape = 2.344073, rate = 35.870449, lower.tail = TRUE) in the R stats package (version 3.4.4). All remaining isolates (n=1,607) were used to reconstruct the plasmidome tree.

We used the function NbClust (method "ward.D2" and index "silhouette") available in the R NbClust package (version 3.0) (53) to evaluate an optimal number of clusters derived from pairwise Mash distances. We computed hierarchical clustering using the hcut function (method ="ward.D2", isdiss=TRUE, k=26) and cut the resulting dendrogram specifying 26 clusters. For each resulting cluster, we uniquely defined plasmidome populations (n=9) based on two criteria: (i) clusters with more than 50 isolates and (ii) an average silhouette width greater than 0.3.

Correlation of plasmidome populations and isolation sources was determined using a one-sided Fisher exact test (alternative "greater") from the fisher.test function (R stats package version 3.4.4) and naive P values were adjusted using the Benjamini-Hochberg (BH) method implemented in p.adjust function (R stats package, version 3.4.4). We considered an adjusted P value threshold of 0.05 to determine enrichment of isolation sources for specific plasmidome populations. We incorporated metadata and plasmid population information into plasmid bioNJ and the *E. faecium* core genome tree using the R ggtree package (version 1.13.3). Simpson index based on SC diversity (postBNGBAPS.2 group) (Data Set S1) and its associated 95% confidence interval from 1,000 bootstrap replications was computed using the R package iNEXT (version 2.0.19) (54).

We evaluated the influence of two other covariate (time and distance) in the clustering derived from Mash distances. For each pair of isolates, we determined (i) if they belonged to the same or different isolation source, (ii) time difference (in years) between their isolation times, and (iii) geographical distance. To calculate the geographical distance, we considered the latitude and longitude of each isolate and used the distm function (R geosphere package, version 1.5-7). We fitted three linear regression models (function lm in R stats package, version 3.4.4) considering as response the pairwise Mash distances and the previous defined covariates. For each model, we retrieved its adjusted R2 to explain the percentage of variance explained by each covariate. We combined all three covariates

in a multiple linear regression model using the function lm (R stats package, version 3.4.4) and further evaluated the observed correlations by performing a permutation test with the function lmp from the package lmPerm (version 2.1.0) (55).

Contribution of genomic components to source specificity

To evaluate the contribution of genomic components on source specificity, we considered three different inputs: (i) Mash pairwise distances from whole-genome contigs, (ii) Mash pairwise distances from chromosome-derived contigs, and (iii) Mash pairwise distances from plasmid-derived contigs. Pairwise distances were scaled using the scale function (scale=TRUE, center=TRUE) from the R stats package (version 3.4.4). For each isolation source (hospitalized patient, dog, poultry, pig, and nonhospitalized person), we used a bootstrap approach (100 iterations) to calculate the average pairwise distances of 50 random isolates belonging to the following combinations: (i) pairs of isolates belonging to the same niche (within-source group), (ii) pairs of isolates belonging to different niches (between-source group), and (iii) pairs of isolates belonging to random isolation sources (random group). This random group consisted of an artificial group in which we merged 50 random isolates belonging to any of the five isolation sources after sampling 100 isolates from each of the sources to avoid overrepresentation of hospitalized patient isolates. This random group was used to statistically assess whether the distribution of pairwise distances belonging to within-source and between-source groups differed from that of random pairwise distances. We used a one-way analysis of variance (ANOVA) test (aov function, R stats package version 3.4.4) and computed differences in the observed means using Tukey's honestly significant difference (HSD) function available in the R stats package (version 3.4.4). Significant (adjusted $P<0.05$) positive and negative observed differences of the means were considered indications of niche adaptation similarity and dissimilarity, respectively.

Estimating the core plasmidome of the defined populations

We used Roary (version 3.8) (40) to define orthologous groups present in each plasmidome population by defining a threshold of 95% amino-acid-level similarity and nonsplitting paralogues. We defined the core plasmidome of each population as the total number of core genes (OGs present in more than 99% isolates) and soft-core genes (OGs present in more than 95% of the isolates but less than 99% of the isolates). To group these core plasmidome genes into different COG categories, we used eggNOG (version 1.0.3-5-g6972f60) with the translate option and the bacterial database (4.5.1) provided.

Data availability

The complete code used to generate the analysis reported in the manuscript is publicly available at the following GitLab repository:

https://gitlab.com/sirarredondo/efaecium_population.

Illumina NextSeq 500/MiSeq reads of the 1,644 *E. faecium* isolates used in this study have been deposited in the following European Nucleotide Archive (ENA) public project: PRJEB28495. Oxford Nanopore Technologies MinION reads used to complete the 62 *E. faecium* genomes are available under the following figshare projects: 10.6084/m9.figshare.7046804 and 10.6084/m9.figshare.7047686.

Hybrid assemblies generated by Unicycler (v.0.4.1) are available under the ENA and NCBI project PRJEB28495 and also retrievable at the following GitLab repository: https://gitlab.com/sirarredondo/ efaecium_population/tree/master/Files/Unicycler_assemblies. Annotation of the complete genome se- quences generated in this study are available on NCBI under BioProject PRJEB28495.

Pangenomes of the observed plasmidome populations and eggNOG annotation are available at https://gitlab.com/sirarredondo/efaecium_population/tree/master/Files/Plasmid_populations.

Exploratory analysis of our data and metadata set is available at the following microreact project: https://microreact.org/project/BJKGTJPTQ.

Acknowledgments

This study was supported by the Joint Programming Initiative in Antimicrobial Resistance (JPIAMR Third call, STARCS, JPIAMR2016-AC16/00039 to S.A.-A., W.V.S., and R.J.L.W.) and by the Academy of Finland (grant no. 286607 and 294015 to P.M.). J.C. was funded by the European Research Council (grant no. 742158). W.V.S. was supported by a Royal Society Wolfson Research merit award (grant no. WM160092).

References

1. Weiner LM, Webb AK, Limbago B, Dudeck MA, Patel J, Kallen AJ, Edwards JR, Sievert DM. 2016. Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the Na- tional Healthcare Safety Network at the Centers for Disease Control and Prevention, 2011–2014. Infect Control Hosp Epidemiol 37:1288 –1301. https://doi.org/10.1017/ice.2016.174.

2. Guzman Prieto AM, van Schaik W, Rogers MRC, Coque TM, Baquero F, Corander J, Willems R. 2016. Global emergence and dissemination of enterococci as nosocomial pathogens: attack of the clones? Front Microbiol 7:788. https://doi.org/10.3389/fmicb.2016.00788.

3. Bonten MJ, Willems R, Weinstein RA. 2001. Vancomycin-resistant enterococci: why are they here, and where do they come from? Lancet Infect Dis 1:314 –325. https://doi.org/10.1016/S1473-3099(01)00145-1.

4. Galloway-Peña J, Roh JH, Latorre M, Qin X, Murray BE. 2012. Genomic and SNP analyses demonstrate a distant separation of the hospital and community-associated clades of *Enterococcus faecium*. PLoS One 7:e30187. https://doi.org/10.1371/journal.pone.0030187.

5. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, Desjardins C, Cerqueira G, Gevers D, Walker S, Wortman J, Feldgarden M, Haas B, Birren B, Gilmore MS. 2012. Comparative ge-

nomics of enterococci: variation in *Enterococcus faecalis*, clade structure in *E. faecium*, and defining characteristics of *E. gallinarum* and *E. casseliflavus*. mBio 3:e00318-11. https:// doi.org/10.1128/mBio.00318-11.

6. Lebreton F, van Schaik W, McGuire AM, Godfrey P, Griggs A, Mazumdar V, Corander J, Cheng L, Saif S, Young S, Zeng Q, Wortman J, Birren B, Willems RJL, Earl AM, Gilmore MS. 2013. Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. mBio 4:e00534-13. https://doi.org/10.1128/mBio.00534-13.

7. Raven KE, Reuter S, Reynolds R, Brodrick HJ, Russell JE, Török ME, Parkhill J, Peacock SJ. 2016. A decade of genomic history for healthcare- associated *Enterococcus faecium* in the United Kingdom and Ireland. Genome Res 26:1388 –1396. https://doi.org/10.1101/gr.204024.116.

8. Palmer KL, Kos VN, Gilmore MS. 2010. Horizontal gene transfer and the genomics of *enterococcal* antibiotic resistance. Curr Opin Microbiol 13:632– 639. https://doi.org/10.1016/j.mib.2010.08.004.

9. Hegstad K, Mikalsen T, Coque TM, Werner G, Sundsfjord A. 2010. Mobile genetic elements and their contribution to the emergence of antimicrobial resistant *Enterococcus faecalis* and *Enterococcus faecium*. Clin Microbiol Infect 16:541–554. https://doi.org/10.1111/j.1469-0691.2010.03226.x.

10. Sadowy E. 2018. Linezolid resistance genes and genetic elements enhancing their dissemination in enterococci and streptococci. Plasmid 99:89 –98. https://doi.org/10.1016/j.plasmid.2018.09.011.

11. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV, Hayes F. 2014. Extrachromosomal and mobile elements in enterococci: transmission, maintenance, and epidemiology, In Gilmore MS, Clewell DB, Ike Y, Shankar N (ed), Enterococci: from commensals to leading causes of drug resistant infection. Massachusetts Eye and Ear Infirmary, Boston, MA.

12. Mikalsen T, Pedersen T, Willems R, Coque TM, Werner G, Sadowy E, van Schaik W, Jensen LB, Sundsfjord A, Hegstad K. 2015. Investigating the mobilome in clinically important lineages of *Enterococcus faecium* and *Enterococcus faecalis*. BMC Genomics 16:282. https://doi.org/10.1186/ s12864-015-1407-6.

13. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, Willems RJL, Schürch AC. 2018. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genom 4:e000224. https://doi.org/10.1099/mgen.0.000224.

14. van den Bunt G, Top J, Hordijk J, de Greeff SC, Mughini-Gras L, Corander J, van Pelt W, Bonten MJM, Fluit AC, Willems R. 2017. Intestinal carriage of ampicillin- and vancomycin-resistant *Enterococcus faecium* in humans, dogs and cats in the Netherlands. J Antimicrob Chemother 73:607– 614. https://doi.org/10.1093/jac/dkx455.

15. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 32:3380 –3387. https://doi.org/10.1093/bioinformatics/ btw493.

16. Bender JK, Fiedler S, Klare I, Werner G. 2015. Complete genome sequence of the gut commensal and laboratory strain *Enterococcus faecium* 64/3. Genome Announc 3:e01275-15. https://doi.org/10.1128/genomeA.01275-15.

17. Buultjens AH, Lam MMC, Ballard S, Monk IR, Mahony AA, Grabsch EA, Grayson ML, Pang S, Coombs GW, Robinson JO, Seemann T, Johnson PDR, Howden BP, Stinear TP. 2017. Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*. PeerJ 5:e2916. https://doi.org/10.7717/peerj.2916.

18. Palmer KL, Gilmore MS. 2010. Multidrug-resistant enterococci lack CRISPR-cas. mBio 1:e00227-10. https://doi.org/10.1128/mBio.00227-10.

19. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R. 2018. Systematic discovery of antiphage defense systems in the microbial pangenome. Science 359:eaar4120. https://doi.org/10.1126/ science.aar4120.

20. Huo W, Adams HM, Trejo C, Badia R, Palmer KL. 2019. A type I restriction- modification system associated with *Enterococcus faecium* subspecies separation. Appl Environ Microbiol 85:e02174-18. https://doi.org/10.1128/AEM.02174-18.

21. Dy RL, Przybilski R, Semeijn K. 2014. A widespread bacteriophage abortive infection system functions through a type IV toxin-antitoxin mechanism. Nucleic Acids 42:4590 –4605. https://doi.org/10.1093/nar/gkt1419.

22. O'Connor L, Tangney M, Fitzgerald GF. 1999. Expression, regulation, and mode of action of the AbiG abortive infection system of lactococcus lactis subsp. cremoris UC653. Appl Environ Microbiol 65:330 –335. https://doi.org/10.1128/AEM.65.1.330-335.1999.

23. Hasman H. 2005. The tcrB gene is part of the *tcrYAZB* operon conferring copper resistance in *Enterococcus faecium* and *Enterococcus faecalis*. Microbiology 151:3019 –3025. https://doi.org/10.1099/mic.0.28109-0.

24. Bustos AY, Font de Valdez G, Fadda S, Taranto MP. 2018. New insights into bacterial bile resistance mechanisms: the role of bile salt hydrolase and its impact on human health. Food Res Int 112:250 –262. https://doi

.org/10.1016/j.foodres.2018.06.035.

25. Foley MH, O'Flaherty S, Barrangou R, Theriot CM. 2019. Bile salt hydrolases: gatekeepers of bile acid metabolism and host-microbiome crosstalk in the gastrointestinal tract. PLoS Pathog 15:e1007581. https:// doi.org/10.1371/journal.ppat.1007581.

26. Nilsson O, Myrenås M, Ågren J. 2016. Transferable genes putatively conferring elevated minimum inhibitory concentrations of narasin in *Enterococcus faecium* from Swedish broilers. Vet Microbiol 184:80 – 83. https://doi.org/10.1016/j.vetmic.2016.01.012.

27. Novak R, Henriques B, Charpentier E, Normark S, Tuomanen E. 1999. Emergence of vancomycin tolerance in *Streptococcus pneumoniae*. Nature 399:590 –593. https://doi.org/10.1038/21202.

28. Moscoso M, Domenech M, García E. 2010. Vancomycin tolerance in clinical and laboratory *Streptococcus pneumoniae* isolates depends on reduced enzyme activity of the major LytA autolysin or cooperation between CiaH histidine kinase and capsular polysaccharide. Mol Microbiol 77:1052–1064. https://doi.org/10.1111/j.1365-2958.2010.07271.x.

29. Kurushima J, Ike Y, Tomita H. 2016. Partial diversity generates effector immunity specificity of the Bac41-like bacteriocins of *Enterococcus faecalis* clinical strains. J Bacteriol 198:2379 –2390. https://doi.org/10.1128/JB.00348-16.

30. Freitas AR, Tedim AP, Francia MV, Jensen LB, Novais C, Peixe L, Sánchez- Valenzuela A, Sundsfjord A, Hegstad K, Werner G, Sadowy E, Hammerum AM, Garcia-Migura L, Willems RJ, Baquero F, Coque TM. 2016. Multilevel population genetic analysis of *vanA* and *vanB*

Enterococcus faecium causing nosocomial outbreaks in 27 countries (1986 –2012). J Antimicrob Chemother 71:3351–3366. https://doi.org/10.1093/jac/dkw312.

31. Willems RJ, Top J, van den Braak N, van Belkum A, Mevius DJ, Hendriks G, van Santen-Verheuvel M, van Embden JD. 1999. Molecular diversity and evolutionary relationships of Tn1546-like elements in enterococci from humans and animals. Antimicrob Agents Chemother 43:483– 491. https://doi.org/10.1128/AAC.43.3.483.

32. Freitas AR, Coque TM, Novais C, Hammerum AM, Lester CH, Zervos MJ, Donabedian S, Jensen LB, Francia MV, Baquero F, Peixe L. 2011. Human and swine hosts share vancomycin-resistant *Enterococcus faecium* CC17 and CC5 and *Enterococcus faecalis* CC2 clonal clusters harboring Tn1546 on indistinguishable plasmids. J Clin Microbiol 49:925–931. https://doi.org/10.1128/JCM.01750-10.

33. Soheili S, Ghafourian S, Sekawi Z, Neela VK, Sadeghifard N, Taherikalani M, Khosravi A, Ramli R, Hamat RA. 2015. The *mazEF* toxin-antitoxin system as an attractive target in clinical isolates of *Enterococcus faecium* and *Enterococcus faecalis*. Drug Des Devel Ther 9:2553–2561. https://doi.org/10.2147/DDDT.S77263.

34. Poulsen HD. 1998. Zinc and copper as feed additives, growth factors or unwanted environmental factors. J Anim Feed Sci 7:135–142. https://doi.org/10.22358/jafs/69961/1998.

35. Gouliouris T, Raven KE, Ludden C, Blane B, Corander J, Horner CS, Hernandez-Garcia J, Wood P, Hadjirin NF, Radakovic M, Holmes MA, de Goffau M, Brown NM, Parkhill J, Peacock SJ. 2018. Genomic surveillance of *Enterococcus faecium* reveals limited sharing of strains and resistance genes between livestock and humans in the United Kingdom. mBio 9:e01780-18. https://doi.org/10.1128/mBio.01780-18.

36. Knarreborg A, Engberg RM, Jensen SK, Jensen BB. 2002. Quantitative determination of bile salt hydrolase activity in bacteria isolated from the small intestine of chickens. Appl Environ Microbiol 68:6425– 6428. https://doi.org/10.1128/aem.68.12.6425-6428.2002.

37. Zhang X, Bierschenk D, Top J, Anastasiou I, Bonten MJM, Willems RJL, van Schaik W. 2013. Functional genomic analysis of bile salt resistance in *Enterococcus faecium*. BMC Genomics 14:299. https://doi.org/10.1186/ 1471-2164-14-299.

38. Zhang X, Top J, de Been M, Bierschenk D, Rogers M, Leendertse M, Bonten MJM, van der Poll T, Willems RJL, van Schaik W. 2013. Identification of a genetic determinant in clinical *Enterococcus faecium* strains that contributes to intestinal colonization during antibiotic treatment. J Infect Dis 207:1780 –1786. https://doi.org/10.1093/infdis/jit076.

39. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

40. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

41. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 40:e6. https:// doi.org/10.1093/nar/gkr928.

42. de Been M, van Schaik W, Cheng L, Corander J, Willems RJ. 2013. Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium*. Genome Biol Evol 5:1524 –1535. https://doi.org/10.1093/gbe/evt111.

43. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30:1224 –1228. https://doi.org/10.1093/molbev/mst028.

44. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

45. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. 2016. CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC Genomics 17:356. https://doi.org/10.1186/s12864-016-2627-0.

46. Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57. https://doi.org/10.1093/nar/gkm360.

47. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.

48. Kassambara A. 2017. ggpubr: ggplot2 based publication ready plots. R package version 0 1 6. https://CRAN.R-project.org/package ggpubr.

49. Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685– 695. https:// doi.org/10.1093/oxfordjournals.molbev.a025808.

50. Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20:289 –290. https://doi.org/ 10.1093/bioinformatics/btg412.

51. Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593. https://doi.org/10.1093/bioinformatics/btq706.

52. Delignette-Muller ML, Dutang C. 2015. fitdistrplus: an R package for fitting distributions. J Stat Softw 64:1–34.

53. Charrad M, Ghazzali N, Boiteau V, Niknafs A. 2014. NbClust: an examination of indices for determining the number of clusters. R package version 1. J Stat Softw 61:1–36.

54. Hsieh T, Ma K, Chao A. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods Ecol Evol 7:1451–1456. https://doi.org/10.1111/2041-210X.12613.

55. Wheeler B, Torchiano M. 2010. lmPerm: permutation tests for linear models. R package version 1. https://CRAN.R-project.org/package lmPerm.

Supplementary Results

Characterization of completed plasmid sequences obtained by ONT sequencing

The completely sequenced plasmids were first characterized based on their replication initiator proteins (RIP). Most of the completed plasmid sequences contained a RepA_N initiator sequence (n = 82) with similarity versus RIP sequences described in pLG1 me-

gaplasmid (n = 55, accession number ADO66907) and the non-conjugative pRUM plasmid (n = 27, accession number NP_863172). RepA_N family was found in large plasmids (mean = 155.2 kbp) (Figure 2B), occasionally associated with other RIP sequences (n = 15, freq = 0.18) (Figure 2B) and present in hospitalized patients (n = 63), dog (n = 12), pig (n = 3), non-hospitalized persons (n = 2) and chicken isolates (n = 2) (Suppl. Figure S1). We also identified plasmids containing RepA_N-like (<80% identity) initiators (n = 20) in medium plasmids (mean 53.9 kbp).

Also the Inc18 family was ubiquitous in the collection of plasmid sequences (n = 57) and present in plasmids with a medium size (mean = 44.7 kbp) (Figure 2B). Plasmids bearing Inc18 family showed higher levels of mosaicism than RepA_N plasmids and were frequently present in multireplicon plasmids (n = 30 ; freq = 0.53). In this study, we mainly found Inc18 sequences with similarity to the initiator sequence from pRE25 plasmid (n = 44, accession number Q9AL28) which was originally identified in *E. faecalis* from a raw-fermented sausage and associated with multiple antibiotic resistance genes (1). We identified Inc18 plasmids in isolates from hospitalized patients (n = 49), dog (n = 7), chicken (n = 2), pig (n = 1) and non-hospitalized persons (n = 1) (Suppl. Figure S1).

The Rep_3 family was mostly found on small plasmids (n = 56, mean plasmid length = 10.8 kbp) (Figure 2B) and rarely present in multireplicon plasmids (n = 6, freq = 0.11.) (Figure 2B). We found Rep_3 sequences with similarity to other small theta-replicating plasmids such as *Enterococcus durans* pGL (n = 22, accession number ADW93773) or *E. faecium* p200B (n = 8, accession number BAF44066). In contrast to RepA_N and Inc18 families, we only found Rep_3 sequences in isolates from hospitalized patients (n = 49), dogs (n = 5) and chickens (n = 2). We detected a high number of plasmids containing Rep_3-like (<80% identity) initiator sequences (n = 54, mean pl. length = 17.1 kbp) and were occasionally associated with other rip families (n = 19, freq = 0.35).

Rep_trans family (n = 24) was mainly identified in plasmids with a medium size (mean = 16.36 kbp) (Figure 2B) and occasionally present in multireplicon plasmids (n = 9, freq = 0.38) (Figure 2B). We mainly found similarity with Rep_trans sequence from pRI1 (n = 16, accession number YP_001672021), a small cryptic mobilizable *E. faecium* plasmid from human and animal origin (2). Plasmids bearing Rep_trans family were present in hospitalized patients (n = 19), dog (n = 3) and chicken (n = 2) isolates (Suppl. Figure S1). We also characterised Rep_trans-like (<80% identity) sequences (n = 36, mean pl. length = 20.8 kbp) with a similar frequency of being associated to other rip sequences (n = 11, freq = 0.31).

Rep_2-like (<80% identity) sequences (n = 10) were present in small-medium plasmids (24.0 kbp), frequently present in multireplicon plasmids (n = 6, freq = 0.6), similar to pJB01 (n = 9, accession number YP_138502) (3) and only present in hospitalized (n = 8) and dog isolates (n = 2) (Suppl. Figure S1).

Finally, we described another known Enterococcus Rip family corresponding to Rep_1 (n = 8, mean = 33.4 kbp). This Rip group was only present in multireplicon plasmids in association with RepA_N and Inc18 families (n = 8, freq = 1.0). All Rep_1 sequences had similarity to a previously described non-functional Rep of the non-conjugative pAMα1 (accession number NP_863351) plasmid from *E. faecalis* (4). Our findings were in accordance with previous reports suggesting Rep_1 family may not be functional and other rip initiators are required for plasmid replication (5). Plasmids carrying Rep_1 initiators were found in patients (n = 5), dogs (n = 2) and non-hospitalized person isolates (n = 1). We additionally found Rep_1-like (<80% identity) sequences (n = 6) present in small plasmids (mean = 4.0 kbp), not associated with other replication initiator families and with similarity to Rep_1 sequences from *E. faecalis* pTEF1 (n = 6, accession number NP_816941) and *E. faecium* pNJAKD (n = 2, accession number YP_004747351).

MOB_P family was the most predominant relaxase family (n = 124) (Figure 2B) and present in plasmids with a single RepA_N (n = 44) or Rep_3 (n = 46) initiator sequence (Figure 2A). This relaxase family was also found in RepA_N-like (n = 15), Inc18 (n = 8), and other multireplicon plasmids (n = 11). MOB_V family was mainly found in plasmids carrying a single Rep_trans-like (n = 11), Rep_1-like (n = 5), Rep_trans (n = 2) families and multireplicon plasmids (n = 11) containing different combinations of Inc18, Rep_1, RepA_N and Rep_1 families (Figure 2A). MOB_T family was identified in multireplicon plasmids containing a Rep_trans-like group and several combinations of Inc18, RepA_N an Rep_1 sequences (n = 6). MOB_C was only found in RepA_N plasmids (n = 2) including a multireplicon plasmid (RepA_N, Rep_trans and Rep_trans-like) (Figure 2A).

Restriction-modification systems, but not CRISPR-cas, could act as barriers of horizontal gene transfer

Recently, a new type I restriction modification (RM) system has been described in *E. faecium* (6). This RM system is composed of three subunits (M, R, and S). The latter subunit is responsible for the recognition and binding to foreign DNA sequences through the presence of two target recognition domains and was enriched in a set of clade A1 isolates (6). A different set of RM systems could act as barrier in the gene exchange of *E. faecium* and contribute to the subspecies separation (6). In our collection, we also identified this S-subunit (WP_002287733) as present and enriched in clade A1 isolates (Fisher's exact test, $P < 0.05$). As previously reported (6), the subunits M and R were present in both clade A1 and non-clade A1 isolates, which suggests that the specificity of the system resides in the subunit S which binds to different DNA sequences. Based on this, we further explored and identified 8 novel S-subunits variants which were present in our set of 62 isolates with complete genomes. The multiple sequence alignment of the unique S-subunit variants (8 novel variants and reference S-subunit variant (accession number WP_002287733)) found

in our set of completed genomes is available at: https://gitlab.com/sirarredondo/efaecium_population/tree/master/Files/rmsystems.

Several S-subunits were enriched in clade A1 isolates (E1774_00555, E7313_02981 , E4413_00571, E4438_00276) and mainly present in hospitalized patients and dog isolates. In contrast, S-subunit variants E0139_00520 and E4227_02943 were enriched in non-clade A1 isolates. These observations may indicate that different sets of RM systems in *E. faecium* population have contributed to the differentiation of the plasmidome content depending on the isolation source.

Plasmidome populations are strongly associated with isolation source

In the following section, we unravel and provide a detailed characterization of the genes present in each plasmidome population. We first assess whether there is a particular *E. faecium* isolation source significantly overrepresented in the population and whether it exists a high SC diversity of the isolates which could indicate horizontal transmission of plasmid sequences (Figure 3B, Suppl. Figure 4B and Figure S5). To identify which genes are driving these populations, we defined the plasmidome-population core genes (present > 95% isolates) and characterized their COG and manually curated and searched some of their functions in literature. We finally compared the sequences to our set of complete plasmid sequences to match the type of plasmid replicons bearing these plasmidome-population core genes.

Some of the plasmidome-population core genes defined below were present in several populations. In this analysis, we tried to define which genes are commonly present in the isolates from a particular population. The presence of a plasmidome-population core gene for a particular population does not imply the absence of that gene in another population, as not a single, but the pool of plasmid genes are the ones defining the population. Plasmidome-population core genes shared between populations were mainly involved in plasmid replication, recombination, repair, mobilization or stabilization. These mechanisms are used by plasmid sequences and belong to the backbone of plasmids and thus can be found in populations with overrepresentation of different hosts (e.g. poultry and dog population).

The pangenome of each population and the COG annotation generated by eggNOG are available at: https://gitlab.com/sirarredondo/efaecium_population/tree/master/Files/Plasmid_populations In the sections below we refer to the plasmidome-population genes using a locus tag present each pangenome fasta file provided at: https://gitlab.com/sirarredondo/efaecium_population/tree/master/Files/Plasmid_populations.

The complete annotation of each plasmidome core gene with each associated locus tag is also available in Supplementary Dataset S1.

Population 1

Plasmidome population 1 was enriched for pig and non-hospitalized person isolates (Bonferroni-corrected P < 0.05) suggesting transmission of plasmid sequences between these two sources. We confirmed this by inspecting our set of completed plasmid sequences and chromosomes. Isolate E0139 (from a non-hospitalized person) and isolate E0595 (derived from a pig) shared a near-identical (99.4% identity, 96% coverage) RepA_N conjugative plasmid of 140 kbp (accession numbers LR132068.1 and LR135180.1) which suggested an exchange of this plasmid between these different source types. However, the corresponding chromosome of isolates E0595 and E0139 exhibited different SC's (30 and 29), indicating that the presence of this identical plasmid in these two isolates is the result of horizontal transfer of plasmids, rather than vertical transfer. In general, plasmidome population 1 had a large diversity of SC's (Simpson index = 0.53, CI = 0.524-0.630) suggesting horizontal spread of at least a part of the plasmid sequences defining this population. We identified a total of 111 genes present as part of the core-plasmidome this population. From these 111 genes, only 68 had an associated COG. The most predominant COG was the category S (unannotated function) with 18 genes. We manually inspected these genes and predicted their potential function which included a i) toxin-antitoxin component (TA) system corresponding to the toxin belonging to RelE and AbrB transcriptional regulator (IIAENCCH_00121, IIAENCCH_00120, and a toxin component of the Fic family (IIAENCCH_00140), ii) Abi system formed by the AbiEi (IIAENCCH_00016) and AbiEii (IIAENCCH_00017), iii) a starvation protective gene against oxidative damage (IIAENCCH_00129) and iv) an ABC transporter permease FetB, exporting iron (IIAENCCH_00146).

These groups include genes involved in mechanisms of plasmid stabilization such as TAs that may explain the persistence of large plasmids in the population in the absence of a selective pressure. Furthermore, TA systems have been postulated as attractive targets to stop the dissemination of vancomycin resistance in *E. faecium* (7, 8). The AbiEi/AbiEii system described corresponds to an innate immune mechanism that provides viral protection against phage dissemination and its mechanism of action interferes with phage RNA synthesis and also enable stabilization of mobile genetic elements (9). Interestingly, this system has been extensively described in lactococcal plasmids (10). The following most predominant COG groups corresponded to COG L (14 genes) and COG M (10 genes). COG L genes belonged mainly to genes involved in plasmid replication, recombination and repair such as ISEfa7 transposase (IIAENCCH_00090), IS1476 transposases (IIAENCCH_00105, IIAENCCH_00106) ISEnfa3 transposase (IIAENCCH_00109) or DNA topoisomerase III (IIAENCCH_00019) among other examples (Suppl. Dataset S1). In the COG M group, we detected a copper resistance gene operon (*tcrYAZB* operon) (IIAENCCH_00096, IIAENCCH_00107,

IIAENCCH_00137, IIAENCCH_00139, IIAENCCH_00095, IIAENCCH_00136) that mediates resistance against this heavy-metal and was previously described in *E. faecium* as plasmid-borne (11). Copper was commonly used as a growth-promoting agent to increase pig production (11). However, high-levels of copper result is toxic for the cells. The *tcrYA-ZB* operon provides a plasmid-survival mechanism to tolerate high concentrations of this heavy-metal. We found that the glycopeptide resistance gene *vanA* (IIAENCCH_00115) was also part of the set of plasmidome-population core genes of this population. Furthermore, we also identified an ATPase from a type IV secretion system (IIAENCCH_00032) and the TraG family conjugation protein (IIAENCCH_00037) as a plasmidome-population core gene which suggests the presence of an active and widespread system to enhance plasmid mobilization. The plasmidome-population core genes described above were present in the complete plasmid sequences previously mentioned corresponding to a RepA_N conjugative plasmid (accession numbers LR132068.1 and LR135180.1). The introduction of this plasmid in the population together with the pool of genes described could be explained due to selective pressures such as high concentrations of copper or the usage of glycopeptides during pig breeding. The presence of TA systems or Abi systems such as AbiEi/AbiEii may play a role in stabilizing the plasmid structure after removing the initial selection pressure.

Population 2

Plasmidome population 2 was significantly overrepresented by poultry isolates (Bonferroni $P < 0.05$) and exhibited a high homogeneity of SC's (Simpson index = 0.21, CI = 0.210-0.359) suggesting that plasmid sequences within this population were mainly vertically inherited.

We identified a total of 93 genes as belonging to the core-plasmidome of population 2 and 58 genes had an associated COG category. The most predominant COG group corresponded to COG L with a total of 14 genes which included plasmid replication, recombination and repair genes such as DNA topoisomerases III (GDKHCPLE_00029, GDKHCPLE_00091), IS66 Orf2-like proteins (GDKHCPLE_00142), HNH endonuclease (GDKHCPLE_00092) among other examples (Suppl. Dataset S1). The second most predominant group was COG G which included mainly genes associated to carbohydrate utilisation such as PTS systems involved in: i) trehalose/maltose utilisation (GDKHCPLE_00038, GDKHCPLE_00109), ii) PTS system involved in N-acetylglucosamine (GDKHCPLE_00037), iii) glycosyl hydrolase (GDKHCPLE_00105) or iv) fructokinases (GDKHCPLE_00104, GDKHCPLE_00110) among other examples. This may confer novel pathways for carbohydrate usage in this poultry-associated population. We detected a BSH choloylglycine hydrolase member (GDKHCPLE_00140, COG M) as being part of the core of this plasmidome populations. The role of BSH activity in the intestine of poultry is unclear but it has been hypo-

thesized that can confer tolerance to the bile (12). Again, we find glycopeptide resistance mediated by *vanA* (GDKHCPLE_00050) within the set of core-genes in this population. We could also identify an additional resistance gene corresponding to a streptomycin adenyltransferase (*aadE*, GDKHCPLE_00054). Also TA systems mediated by RelE toxin (NGNCBCCO_00115) and an ATPase from a type IV secretion system (GDKHCPLE_00017) which may enhance the mobilization of plasmid sequences, were part of the core genes in this plasmid population. We identified as plasmidome-population core a tetronasin resistance gene (GDKHCPLE_00084). The presence of this tetronasin resistance gene on mobile element among *E. faecium* poultry isolates has been previously described and may be related to the widely use of ionophores, e.g. tetronasin for coccidiosis prophylaxis in poultry (13).

In our pool of complete plasmid sequences, we observed the presence of two near-identical plasmids present in the poultry isolates E4227 and E4239 which were sequenced to completion: i) a RepA_N conjugative plasmid (175 kbp) corresponding to accession numbers LR135171 and LR135783 (100% identity and 100% coverage) and ii) multireplicon Inc18 and Rep3 plasmid (46 kbp) corresponding to accession numbers LR135172 and LR135784 (100% identity and 99% coverage).

Population 3

This population was significantly overrepresented with hospitalized patient isolates. SC diversity measured by the Simpson index (0.04, CI = 0.037-0.110) indicated that the isolates belonging to this population shared plasmid sequences which were mainly transmitted due to vertical inheritance.

In this population, we identified a total of 73 plasmidome-population core genes from which 37 had an associated COG. The most predominant COG corresponded to the category S (unknown function) with 13 genes. Within this group, i) TA system formed by the Txe/YoeB module (GCEJCPEH_00117, GCEJCPEH_00118) and ii) a duplicated Abi system formed by AbiEi (GCEJCPEH_00086, OLDHMAJC_00207) and AbiEii (GCEJCPEH_00085, OLDHMAJC_00206) were identified. The second most predominant COG corresponded to COG L with a total of 7 genes and mainly corresponding to genes involved in plasmid replication, recombination and repair such as DNA topoisomerase III (GCEJCPEH_00088), helix-destabilizing (GCEJCPEH_00084) or resolvase proteins (OLDHMAJC_00341) among other examples (Suppl. Dataset S1). We identified three genes categorized as COG V (defense mechanism) corresponding to an ABC transport system formed by an ATP-binding protein (OLDHMAJC_00277) and two permeases (OLDHMAJC_00276, OLDHMAJC_00278). These three genes are similar to the previously described *vex* locus in *Streptococcus pneumoniae* (14). As plasmidome-population core genes, we identified an ATPase involved in a type IV secretion system (GCEJCPEH_00068) and the TraG family conjugation protein (GCEJCPEH_00063) which may contribute to the mobilization and spread of

plasmid sequences in the population, and the *erm* gene (GCEJCPEH_00113) conferring resistance to macrolide, lincosamide and streptogramin B. We inspected our set of complete plasmid sequences derived from long-read isolates (E7196, E7654 and E7663) belonging to the population 3. We found three plasmid structures bearing the described plasmidome-population core genes: i) a RepA_N plasmid with a length higher than 160 kbp (LR135271, LR135325, LR135318), ii) a different RepA_N plasmid with a length around 62 kbp (LR135272, LR135326, LR135319) and iii) a multireplicon plasmid Inc18 & Rep_1 only present in the isolates E7654 (LR135327) and E7663 (LR135320) with a length around 38 kbp.

Population 4

Plasmidome population 4 was enriched among dog isolates (Bonferroni corrected $P < 0.05$) with a high SC diversity (Simpson index = 0.73, CI = 0.728-0.781) suggesting horizontal spread of plasmid sequences between isolates from this population.

We found a total of 27 genes as part of the core-plasmidome of population 4 from which 17 genes had an associated COG (Supplementary Dataset S1). The most predominant COG corresponded to the unknown function S with 8 genes; i) TA systems formed by the toxin RelE (HBJAKGIG_00118) and a consecutive antitoxin component from the AbrB family (HBJAKGIG_00119), plus another the toxin component of the Fic family (HBJAKGIG_00176) ii) a starvation protective gene against oxidative damage, DPS protein (HBJAKGIG_00170), iii) Abi system formed by AbiEi (HBJAKGIG_00149) and AbiEii (HBJAKGIG_00148). Two plasmidome-population core genes belonging to the category COG G, represented a predicted PTS systems, a sucrose-6-phosphate hydrolase (HBJAKGIG_00172) and a subunit of a beta-glucoside transporter (HBJAKGIG_00173). We found an ATPase part of a type IV secretion system (HBJAKGIG_00052) which could enhance the mobilization of plasmid sequences in the population. The lower number of genes (n = 27) present in the core-plasmidome respect to other populations may indicate that isolates belonging to this population present a higher heterogeneity of plasmid content. Two hospitalized patient isolates (E8040 and E8172) clustered in this dog plasmidome population. E8172 was long-read sequenced and contained a conjugative RepA_N plasmid (156 kbp, accession number LR135373.1) with high levels of similarity but some structural rearrangements (99.9% identity, 77% coverage) when compared to another conjugative RepA_N plasmid (148 kbp, accession number LR135259.1) present in the completely sequenced dog isolate E4457 from the same plasmidome population. Similar RepA_N plasmids are found in other long-read isolates present in the population such as E4402 (accession number LR135175, 149 kbp), E4413 (accession number LR135186, 172 kbp), E8481 (accession number LR536671, 149 kbp) and E4438 (accession number LR135192, 145 kbp).

Population 5

This population was significantly overrepresented by hospitalized patient isolates. SC diversity measured by the Simpson index (0.72, CI = 0.720-0.777) suggested horizontal transmission of plasmid sequences in isolates belonging to this population.

We found a total of 152 plasmidome-population core genes from which 106 had an associated COG. There two most predominant COG groups with 25 genes respectively, COG S (unknown function) and COG L (replication, recombination, repair). COG S included the following TA systems: i) a RelE (AAEHJEFK_00145, AAEHJEFK_00224) and AbrB (AAEHJEFK_00144, AAEHJEFK_00223) system, ii) a Txe/YoeB module (AAEHJEFK_00221,AAEHJEFK_00222), iii) a HicA/HicB module (AAEHJEFK_00072, AAEHJEFK_00073), iv) toxin component of the Fic family (AAEHJEFK_00035) and v) MazE/MazF system (AAEHJEFK_00121,AAEHJEFK_00122). MazE had not COG associated but was defined as plasmidome-population core gene in the population. From COG L, included a variety of genes such as IS1216 transposase (AAEHJEFK_00150), IS256 transposase (AAEHJEFK_00252), ISEf1 transposase (AAEHJEFK_00199), DNA topoisomerase III (AAEHJEFK_00243), resolvases (AAEHJEFK_00079) among other examples (Suppl. Dataset S1). Three genes constituted the *panBCD* (AAEHJEFK_00201, AAEHJEFK_00202, AAEHJEFK_00203) locus, previously described in *Streptococcus gallolyticus* and that encodes for the complete biosynthetic pathways of panthotenate. This locus may provide a selective advantage of *E. faecium* isolates containing this locus to outcompete and grow in an environment with a variety of carbohydrates and poor amino acid source (15). Other plasmidome-population core genes putatively encode a peptidoglycan binding domain protein (COG M, AAEHJEFK_00157) with two domains implicated in: i) peptidoglycan binding and ii) glycoside hydrolase superfamily (GH25_*BacA*-like). We searched the protein sequence against *BacA* homologues that were previously described as a plasmid-encoding bacteriocin in *E. faecalis* (16). *BacA* homologues are splitted into five different clades (16). In our study, we observed a perfect match (blastp, e-value = 0.0, identity = 99%) between AAEHJEFK_00157 and EOK45589 which belongs to clade IV variant. We could not identify other Bac41-like genes in the adjacent areas of *BacA*, which is in accordance with the findings of *BacA* clade IV described by Kurushima et al 2016. Furthermore the authors argue about the functionality of this *BacA* homologue gene since they showed that the presence of *BacL1* (another Bac41-like gene) is required for bacteriolysin activity. Kurushima et al. 2016 described that *BacA* gene can act as a more evolved toxin-antitoxin system in which not only daughter cells but also cells from the same generation not bearing the plasmid gene are excluded. Furthermore, the authors showed that plasmid dissemination was more prominent under conditions of *E. faecium* populations fluctuations since the gene activity exclusively affects dividing cells.

We again find the locus of three genes corresponding to an ABC transport system formed by an ATP-binding protein (AAEHJEFK_00219) and two efflux ABC transport systems (AAEHJEFK_00218, AAEHJEFK_00220) acting as permeases which was previously described in *S. pneumoniae* as *vex* locus. We also detected the TraG gene (AAEHJEFK_00015, COG U) encoding for a conjugation protein and an ATPase from a type IV secretion system (AAEHJEFK_00020), which may indicate that the horizontal transmission of plasmid sequences within this population is mediated by this conjugation system. Several antimicrobial resistance genes were part of the core in plasmidome population 5 including: i) aminoglycoside resistance (*aacA-aphD* gene) (AAEHJEFK_00194), ii) macrolide, lincosamide and streptogramin B resistance (*erm* gene) (AAEHJEFK_00107), iii) glycopeptide resistance (*vanA* gene) (AAEHJEFK_00234) and iv) teicoplanin resistance (AAEHJEFK_00142). We detected a RepA_N plasmid with a length around 165 kbp shared between the isolates E6043 (accession number LR134106), E7040 (accession number LR135220) and E7067 (accession number LR135236). Secondly, we observed a RepA_N like plasmid present also in these E6043 (accession number LR134108), E7040 (accession number LR135222), E7067 (accession number LR135238) and E7207 with a length around 55 kbp and containing the TraG gene described before. And lastly, we identified a multireplicon Inc18 & Rep_3-like plasmid shared between the isolate E6043 (accession number LR134110), E7067 (accession number LR135239) and E7040 (accession number LR135223) with structural rearrangements and length ranging from 38 kbp to 52 kbp. The aminoglycoside resistance gene (AAEHJEFK_00194) was located in the RepA_N plasmid of 165 kbp whereas the other resistance genes (*erm*, *vanA* and teicoplanin resistance gene) were carried by the multireplicon Inc18 & Rep_3-like plasmid. The existence of different plasmid replicons present in the population underpins the importance of analysing the entire pool of plasmid genes rather than focusing on an individual plasmid replicons.

Population 6

Population 6 was significantly overrepresented by hospitalized patient isolates. Based on the heterogeneity of SC's (Simpson index = 0.30, CI = 0.299-0.422) , we concluded that the plasmid sequences present in the population were mainly vertically inherited since most of the isolates of the population belonged to a narrow range of SC's.

In total, we identified 128 plasmidome-population core genes from which 86 had an associated COG. We observed again that the most predominant COG group corresponded to the category S (unknown function) (Suppl. Dataset S1). Core genes in this population represent i) toxin RelE (LDCOMLJG_00062) and antitoxin system (LDCOMLJG_00063) from AbrB family, TA system MazE/MazF (LDCOMLJG_00194, LDCOMLJG_00195) and a toxin component from Fic family (LDCOMLJG_00190), ii) Abi system formed by AbiEi (LDCOMLJG_00151) and AbiEii (LDCOMLJG_00152). The following most predominant

group was COG L which included ISL3 transposase (LDCOMLJG_00028), IS256 transposase (LDCOMLJG_00214), IS200 transposase (LDCOMLJG_00269) or a DNA topoisomerase III (LDCOMLJG_00149) among other examples. Eight genes grouped within the category COG G were encoding among other functions for a complete PTS system involved in mannose/fructose/sorbose utilisation: i) IIA component (LDCOMLJG_00079), ii) IIB (LDCOMLJG_00080), iii) IIC (LDCOMLJG_00081), and iv) IID (LDCOMLJG_00082). We also found the TraG conjugation protein (LDCOMLJG_00222, COG U) present as plasmidome-population core gene suggesting the presence of a conjugative plasmid widely spread in this population. Also the *erm* gene (macrolide, lincosamide and streptogramin B resistance) (LDCOMLJG_00008) belonged to the plasmidome-population core. The locus of three genes encoding for an ABC transporter system including two efflux ABC transporters acting as permeases (LDCOMLJG_00181, LDCOMLJG_00183) and an ATP-binding protein (LDCOMLJG_00182) also belonged to the core in this population. We confirmed the presence of the plasmidome-population core genes in our set of complete plasmid sequences derived from long-read sequenced isolates belonging to population 6 (accession numbers for E8202; LR135345, E7429; LR135298, E6055; LR135198, E7356; LR135340 and E8195; LR135365). We observed a highly similar RepA_N plasmid but showing structural rearrangements depending on the isolates and carrying most of the previously described plasmidome-population core genes.

Population 7

This population was also hospital-associated and its SC diversity (Simpson index = 0.56, CI = 0.561-0.649) suggested horizontal transmission of the plasmid sequences in the population. We identified a total of 138 plasmidome-population core genes from which 86 had an associated COG. In this case, the most predominant COG group corresponded to the category G (carbohydrate transport and metabolism). This group included a complete set of PTS systems involved in mannose/fructose/sorbose-specific utilisation consisting of: i) IIA components (FIBMOKAC_00130, FIBMOKAC_00193), ii) IIB components (FIBMOKAC_00129, FIBMOKAC_00194), iii) IC components (FIBMOKAC_00128, FIBMOKAC_00195), iv) IID components (FIBMOKAC_00127, FIBMOKAC_00196), and v) glycosyl hydrolases (FIBMOKAC_00134), sugar kinases (FIBMOKAC_00257) (Suppl. Dataset S1). This highlights that most of the defined plasmidome-population core genes in this population are responsible for the utilisation of complex carbohydrates. In this population, we also identified several TA systems catalogued as plasmidome-population core genes: i) MazE/MazF system (FIBMOKAC_00110, FIBMOKAC_00109), ii) toxin component of the Fic family (FIBMOKAC_00062) and iii) toxin RelE (FIBMOKAC_00122). We observed a plasmidome-population core gene encoding a peptidoglycan binding domain protein (FIBMOKAC_00168) and with similarity to *BacA* belonged to the plasmidome-population core

in population 7. Furthermore, we identified the set of three genes forming the *vex* locus previously mentioned, formed by two efflux ABC transporters acting as permeases (EIAGHGLI_00196, EIAGHGLI_00200) and an ATP-binding protein (EIAGHGLI_00196). We also observed the presence of an ATPase from a type IV secretion system (FIBMOKAC_00012) that could be involved in the mobilization of plasmid sequences. We could confirm the widespread of these plasmidome-population core genes by inspecting our set of completed genome isolates from long-read sequenced isolates (n = 4) corresponding to hospitalized patients (E7313, E8014 and E8423) and belonging to plasmidome population 7. These isolates shared an identical RepA_N plasmid (> 200 kbp) with some structural rearrangements between their sequences and bearing the same set of PTS system present in two different parts of the plasmid replicon. To highlight this observation, we focused on the isolates E8014 and E8423 belonging to the SC groups 13 and 18 respectively. Both isolates carried a similar large RepA_N plasmid sequence (accession numbers LR135352 and LR135476, 99.7% identity and 86% coverage, length > 200kbp) despite being non-clonally related.

Population 8

This population was also overrepresented among hospitalized patients but its associated Simpson index (0.15, CI = 0.151-0.216) indicated that the plasmid sequences present in this population were mainly clonally inherited. In total, we found 138 plasmidome-population core genes from which 88 had an associated COG. The most predominant COG group was the category L (replication, recombination, repair) with a total of 21 genes. Within this group, we found DNA topoisomerases III (EIAGHGLI_00076, EIAGHGLI_00227), ISEnfa3 transposases (HGPANJKB_00188, HGPANJKB_00238) or an IS256 transposase (EIAGHGLI_00195) among other examples (Suppl. Dataset S1). The second most predominant group (20 genes) represented category S (unknown function). Here, we could group genes in two classes: i) TA systems, including RelE/AbrB (EIAGHGLI_00212,EIAGHGLI_00211), MazEF (EIAGHGLI_00255, EIAGHGLI_00256) and Xre antitoxin component (EIAGHGLI_00223) and ii) Abi system formed by AbiEi and AbiEii (HGPANJKB_00033, HGPANJKB_00034) . We also found 12 plasmidome-population core genes belonging to COG category G which were mainly predicted to encode PTS systems: i) N-acetylglucosamine-specific IIABC component (EIAGHGLI_00204), ii) trehalose/maltose-specific IIBCA component (EIAGHGLI_00205) and iii) lactose/cellobiose-specific IIC component (EIAGHGLI_00221), among other examples of carbohydrate degradation and transport (Suppl. Dataset S1). We identified the same two efflux ABC transporters acting as permeases (EIAGHGLI_00196, EIAGHGLI_00200) and an ATP-binding protein (EIAGHGLI_00197) as well as the peptidoglycan binding domain protein (EIAGHGLI_00233) with highly similarity to BacA as plasmidome core genes. Also an ATPase gene from a type IV secretion system (EIAGHGLI_00030, COG

U) and the antimicrobial resistance gene, *erm* gene (EIAGHGLI_00163) conferring macrolide, lincosamide and streptogramin B resistance were assigned as plasmidome core genes. We confirmed the presence of these plasmidome-population core genes in the complete plasmid sequences derived from long-read isolates belonging to population 8. The genes were carried into two different complete plasmid structures: i) a RepA_N plasmid with a length around 240 kbp present in the long-read isolates E8290 (accession number LR135395), E8414 (accession number LR135489), E7933 (accession number LR135385), E8328 (accession number LR135415), E8284 (accession number LR135409) and E8377 (accession number LR135402) and ii) an Inc18 plasmid with a length around 63 kbp present in the long-read isolates E8290 (accession number LR135396), E8414 (accession number LR135491), E7933 (accession number LR135387), E8284 (accession number LR135410) and E8377 (accession number LR135403). Most of the plasmidome-population core

genes resided in the RepA_N plasmid (240 kbp) whereas the *erm* gene and Xre toxin/antitoxin system was present in the Inc18 plasmid.

Population 9

This population was significantly associated to hospitalized patients and its SC diversity (Simpson index = 0.04, CI = 0.037-0.110) indicated that the plasmid sequences shared in the isolates belonging to the population were mainly vertically inherited. This population had the largest number of plasmidome-population core genes with a total of 205 from which 134 had an associated COG. There were two predominant COG groups L (replication, recombination and repair) and S (unknown function) which 34 genes respectively. Within COG S group, we identified a similar set of genes and included the following categories: i) TA systems, RelE/AbrB system (OCOMIPFD_00142, OCOMIPFD_00141), MazEF system (OCOMIPFD_00214, OCOMIPFD_00215), HicAB system (OCOMIPFD_00079, OCOMIPFD_00080), toxin component of the Fic family (OCOMIPFD_00039), Txe/YoeB module (OCOMIPFD_00114, OCOMIPFD_00113) and ii) Abi system formed by AbiEi (OCOMIPFD_00180) and AbiEii (OCOMIPFD_00179). In COG group L, we identified several transposases such as ISEfa8 (OCOMIPFD_00144), ISEf1 (OCOMIPFD_00151), ISEfa7 (OCOMIPFD_00197) transposases, DNA topoisomerases III (OCOMIPFD_00168), among other examples (Suppl. Dataset S1). We again observed the presence of an ABC transporter system formed by two permeases (OCOMIPFD_00186, OCOMIPFD_00188) and an ATP-binding protein (OCOMIPFD_00187) as plasmidome core gene. In this population, we also observed a large set of antimicrobial resistance core genes including: i) aminoglycoside resistance genes, *aacA-aphD* (OCOMIPFD_00249), *aadE* (OCOMIPFD_00201) and *aphA* (OCOMIPFD_00271), ii) chloramphenicol resistance, *cat* gene (OCOMIPFD_00281), iii) macrolide, lincosamide and streptogramin B resistance, *erm* gene (OCOMIPFD_00086) and iv) glycopeptide resistance, *vanA* gene (OCOMIPFD_00136).

Plasmidome-population core genes present in hospitalized patient isolates

In the previous section, we described the plasmidome-population core genes present in each population. Next, we analyze which plasmidome core genes were found among all the hospitalized patient isolates. This revealed only 10 plasmidome-population core genes present, which had an associated COG annotation. These genes corresponded to: i) antitoxin component from the AbrB family (OMLCIILE_00043) , ii) replication-associated protein (OMLCIILE_00041), iii) replication initiator protein (OMLCIILE_00049), iv) single-strand binding protein, ssb (OMLCIILE_00299), v) tyrosine recombinase, XerS (OMLCIILE_00437) and vi) putative transposon Tn*552* DNA-invertase bin3 (OMLCIILE_00343). These genes are mainly involved in plasmid replication, recombination, repair or stabilization and thus are present within different plasmid structures. The low number of plasmidome core genes among hospitalized patient isolates may be explained by the heterogeneity (differences in isolation time, year and countries) of isolates falling under this category and emphasizing that different plasmid configurations have evolved within this source group. If we lower the threshold to define a gene as plasmidome-population core (from 95% to > 90% isolates) we identified 39 plasmidome-population core genes from which 22 had an associated COG annotation. This set of 39 plasmidome-population core genes (present in > 90% of hospitalized patient isolates), included several genes previously highlighted for some of the populations, like antimicrobial resistance genes such as aminoglycoside resistance (*aacA-aphD*, OMLCIILE_00488) and ii) erythromycin resistance (*erm*, OMLCIILE_00367) and the RelE toxin (OMLCIILE_00042). RelE is frequently coupled with the antitoxin component from the AbrB family described above, which makes this TA system an attractive target to combat antimicrobial resistance in hospitalized patients by plasmid clearance. Furthermore, we observed the locus of three genes formed by two permeases (HKLEHDKC_00083, OMLCIILE_00480) and an ATP-binding protein (OCOMIPFD_00187). This locus of three genes was not present in populations not related to hospitalized patients (population 1, 2 and 4) which suggests that these three genes may have contributed to the adaptation of *E. faecium* to the hospital environment. We also identified two genes described several times in the previous populations related to mobilization of plasmid sequences formed by a type IV secretion system (OMLCIILE_00070) and TraG involved in conjugation machinery (OMLCIILE_00070).

Supplementary Methods

Illumina sequencing

Bacterial isolates were grown overnight (O/N) at 37°C on blood agar plates. Single colonies were picked up and grown O/N at 37°C with Brain Heart Infusion (BHI). Bacterial cell pellets were pretreated and incubated 1-4 hours with 180 µL of enzymatic lysis buffer. Subsequently, 0.75 mg proteinase K were added and incubated at 56°C until lysis comple-

tion. 20 µL of RNAse A (10mg/mL) were added and incubated for 5' at room-temperature (RT). Total DNA purification was performed using and following the protocol from NucleoSpin 96 Tissue Core Kit (Machery-Nagel), vacuum processing. DNA concentration was measured using Quant-it Picogreen (Thermo Fisher Scientific). Library preparation was carried out following Nextera DNA Library Prep Reference Guide. Finally, Nextera libraries were sequenced using Illumina NextSeq at USEQ, Utrecht, The Netherlands (http://www.useq.nl).

WGS short-read assemblies

Illumina reads were trimmed using nesoni clip, part of the nesoni toolkit (version 0.132), with the following settings: '--adaptor-clip yes --match 10 --max-errors 1 --clip-ambiguous yes --quality 10 --length 90 --trim-start 0 --trim-end 0 --gzip no --out-separate yes pairs:'. Trimmed reads were then assembled into scaffolds using SPAdes (version 3.5.0) with default settings. Scaffolds with an average coverage lower than 10 and/or a length smaller than 500bp were removed from the assemblies.

Selection of isolates to sequence by ONT

A fraction (n=62) of the total number of isolates was selected for long-read sequencing using Nanopore technology. We initially predicted the plasmid content of the isolates *in silico* using PlasmidSPAdes (version 3.8.2) which performs *de novo* assembly filtering out contigs with a coverage similar to the host chromosome coverage (17). Prokka (version 1.12) was used to annotate the putative remaining plasmid contigs specifying the custom Enterococcus database provided (18). Orthologous clustered genes were estimated using Roary (version 3.8), splitting paralogues and defining a threshold of 95% amino-acid level similarity to cluster protein sequences (19). This multi-dimensionality matrix was then reduced and visualized to two dimensions using the t-Distributed Stochastic Neighbor Embedding (t-SNE) (theta = 0.5, iterations = 1000, dims = 2) using the implementation provided in the R package Rtsne (version 0.13) (20, 21). To avoid manual selection of the isolates, k-means function (iter.max = 1000) provided in the R package stats (version 3.4.4) was used to and allocated 50 centroids into the dimensionality reduced distribution given by tSNE. Euclidean distance of each isolate was calculated to extract the 50 isolates closest to each centroid. To cover all plasmid replication genes not present in the first selection, 12 additional isolates were selected for Nanopore sequencing. This second selection was based on a reciprocal blast (blastx and tblastn, -evalue 1e-10) of the predicted plasmid orthologous genes against 76 previously described plasmid replication amino-acid sequences from the genus Enterococcus (22). Isolates bearing plasmid replication genes not present in the first selection were sorted and selected based on the highest number of orthologous genes.

ONT sequencing

E. faecium selected isolates (n = 62) were grown O/N at 37°C on blood agar plates, then single colonies were picked up and grown with BHI at 37°C. Genomic DNA was extracted using the Wizard Genomic DNA purification kit (Promega) following manufacturer's instructions. Isolated DNA was sheared (4000 rpm, 2x120 seconds) using G-tubes (Covaris). Library preparation was performed using Ligation Sequencing Kit 1D (SQK-LSK108) with the Native Barcoding Kit 1D (EXP-NBD103). Genomic libraries were loaded onto R9.4 (FLO-MIN106) flowcells using the MinION device (Mk2). Libraries were basecalled using Metrichor workflows (Run 1 ,2, 3), Albacore 1.01 (Run 4, 5) and Albacore 1.1.0 (Run 6). ONT Sequencing and basecalling were conducted at USEQ, Utrecht, The Netherlands (http://www.useq.nl)

Assembly of ONT sequenced isolates

Fastq files were obtained from base-called data using Poretools (version 0.6.0) except for Run6 in which fastq files were retrieved using Albacore (version 1.1.0). Distribution of read length and total number of reads were calculated using Bioawk (version 20110810, https://github.com/lh3/bioawk). We used Porechop (version 0.2.1, https://github.com/rrwick/Porechop) to trim reads and filter out chimeras from different bins specifying the flag "--discard_middle". Illumina reads were trimmed using seqtk (version 1.2-r94, https://github.com/lh3/seqtk) with the command "--trimfq" prior to assembly. Hybrid assembly was performed using Unicycler (version 0.4.1), specifying "bold" mode (23). Briefly, Unicycler uses SPAdes (version 3.6.2) to create different assembly graphs based on different k-mer size only considering Illumina reads (24). The best assembly graph was selected by Unicycler based on number of dead-ends and contiguity. Next, all ONT reads were used to scaffold and solve the assembly graph. Additionally, we specified the same file as described above (section 'Selection of isolates to sequence by ONT') containing 76 known plasmid replication sequences to rotate and change the 0-coordinate of circular replicons resulting from hybrid assembly (22). Finally, Unicycler conducted several rounds of Pilon (version 1.22) to polish genome sequences using Illumina reads (25).

Characterization of fully assembled plasmids

Contigs derived from hybrid assembly were labeled either as chromosome or plasmid based on sequence length and circularization signatures. Contigs were categorized as plasmid if they presented circularization signatures and a sequence length smaller than 350 kbp. Putative plasmids smaller than 350 kbp and lacking circularization signatures were not considered for further analysis. Rapid annotation by Prokka (version 1.12) (18) allowed us to discard four putative circular phage sequences. We used Abricate (version 0.8.2) to query (> 80% identity & > 60% coverage) our set of completed plasmid sequences (n = 305) versus a curated database of known replication initiator and relaxases proteins from

Enterococcus (5, 26). Replicon sequences with a lower identity (< 80% ID) were classified as Rip-like. Completed plasmid sequences were clustered using Mash (k = 21, s = 1000 ; version 1.1) (27) and the resulting distance matrix was clustered using the hclust function (method = 'ward.D2') provided in R package stats (version 3.4.4). Dendrogram visualization and metadata associated to plasmid isolates was displayed using ggtree (version 1.13.3) (28).

Supplementary Figures

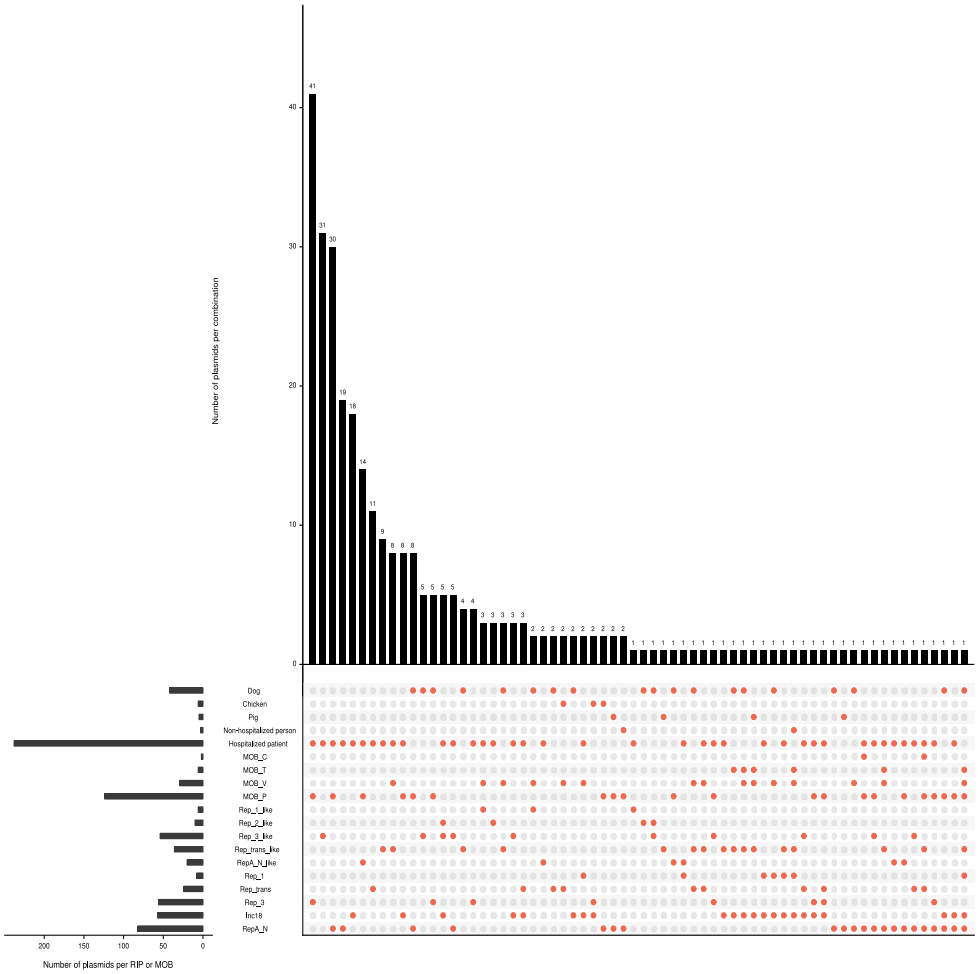


Figure S1. Intersection plot of the combination of RIP and relaxases found in the set of complete plasmid sequences present in long-read sequenced isolates indicating the isolation source in which they were identified.

Figure S2. Boxplot of the distribution in length from the plasmids identified in our set of long-read sequences isolates (n = 59) with complete plasmid sequences. Each isolate (y axis) was colored based on isolation source (brown, poultry; green, pe; red, hospitalized patient; blue, non-hospitalized person; pink, pig). Isolates are displayed in ascending order based on the total number of plasmids identified.

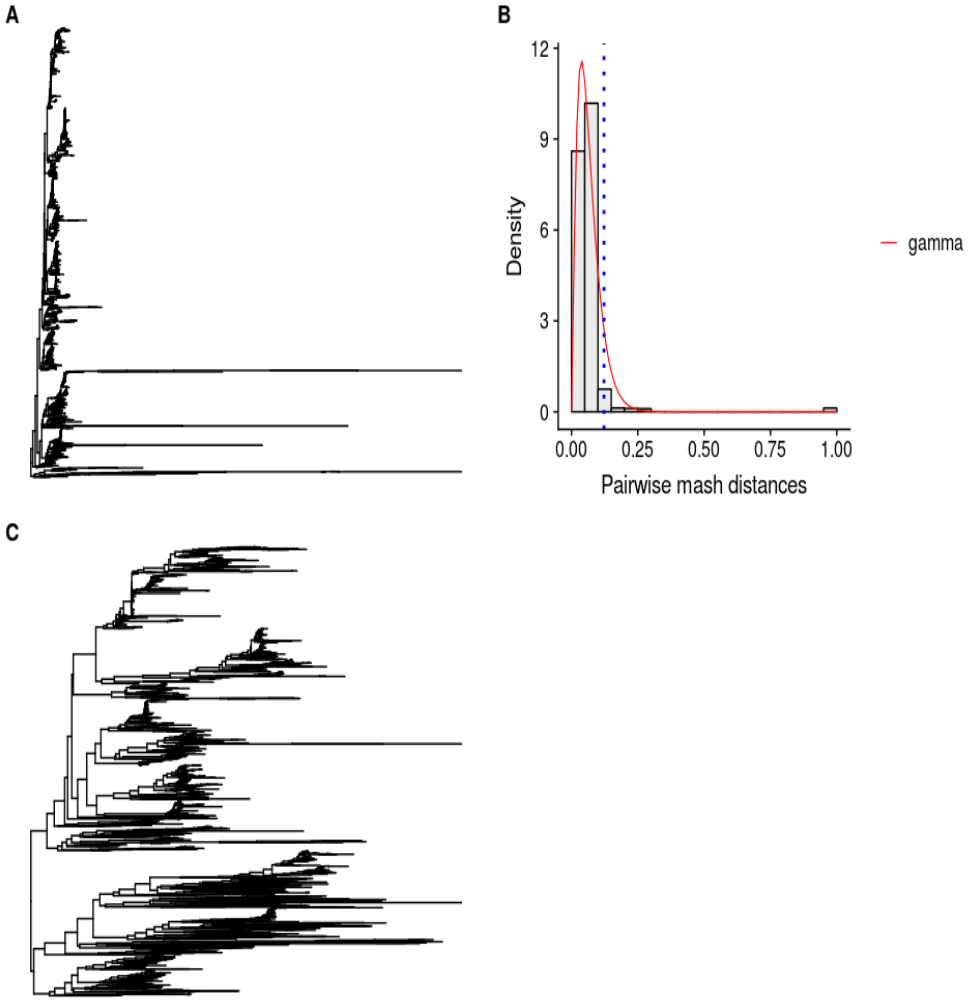


Figure S3. Maximizing resolution of the bioNJ plasmid-based tree. (A) bioNJ tree of 1,639 isolates considering plasmid-predicted contigs by mlplasmids. (B). Histogram against fitted density functions of pairwise Mash distances obtained by denscomp function (fitdistrplus R package). Vertical dashed line indicates the Mash distance (0.1224967) used to filter out isolates (n = 37) with a higher average pairwise Mash distance. (C) bioNJ plasmid-based tree of 1,607 isolates after exclusion of 32 isolates.

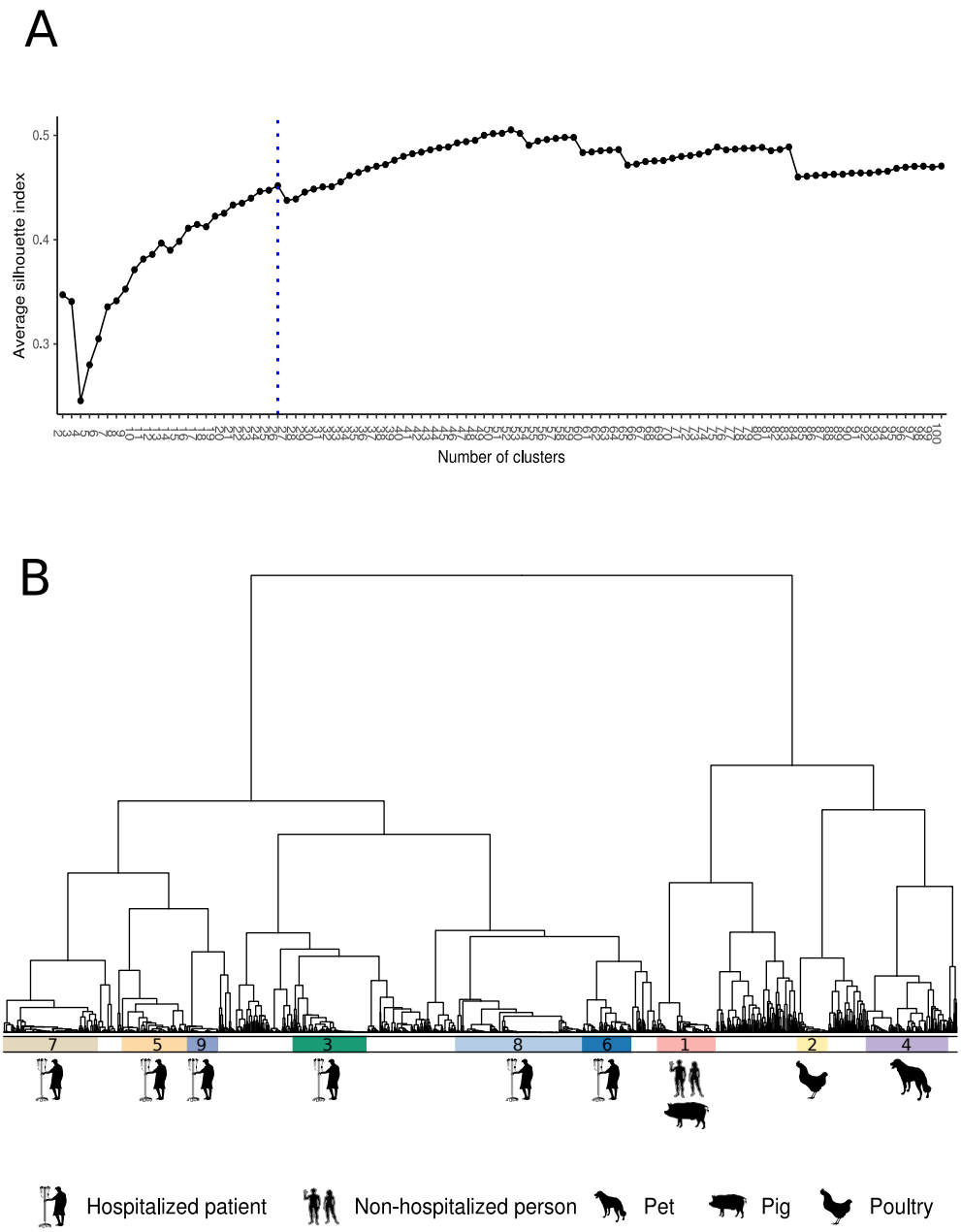


Figure S4. Definition of plasmidome populations. (A). Average silhouette index for plasmid dissimilarity matrix of pairwise Mash distances clustered using hierarchical clustering (ward.D2), computed-for different clustering solutions (2 to 100). We selected 26 as the optimal number of clusters present in the data, which corresponded to an average silhouette index of 0.42. (B) Dendrogram with 26 clusters. From these 26 clusters, we only selected plasmidome populations (n = 9) if a particular cluster had a size larger than 50 isolates and an average silhouette index higher than 0.3. Each plasmid population showed overrepresentation of at least one isolation source.

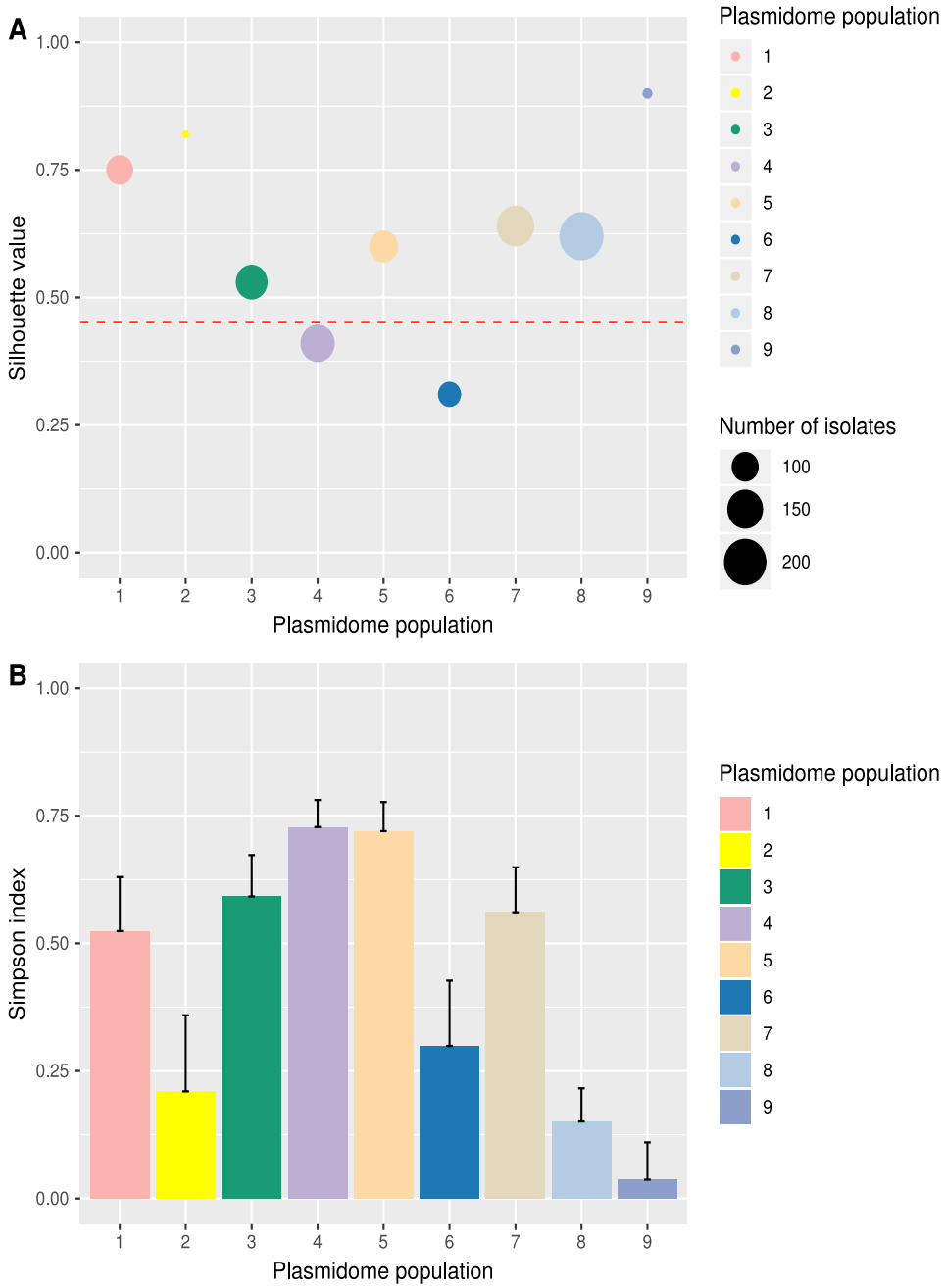


Figure S5. Clustering quality and diversity of the defined plasmidome populations (n = 9). (A) Average silhouette index of each plasmidome population. Size of the point indicates the number of isolates belonging to that particular population. Horizontal dashed line indicates the average silhouette index of the selected clustering solution (k = 26, average silhouette index, 0.42). (B) Simpson indexes and their associated confidence intervals (95%, 1,000 bootstrap replications), based on SC diversity.

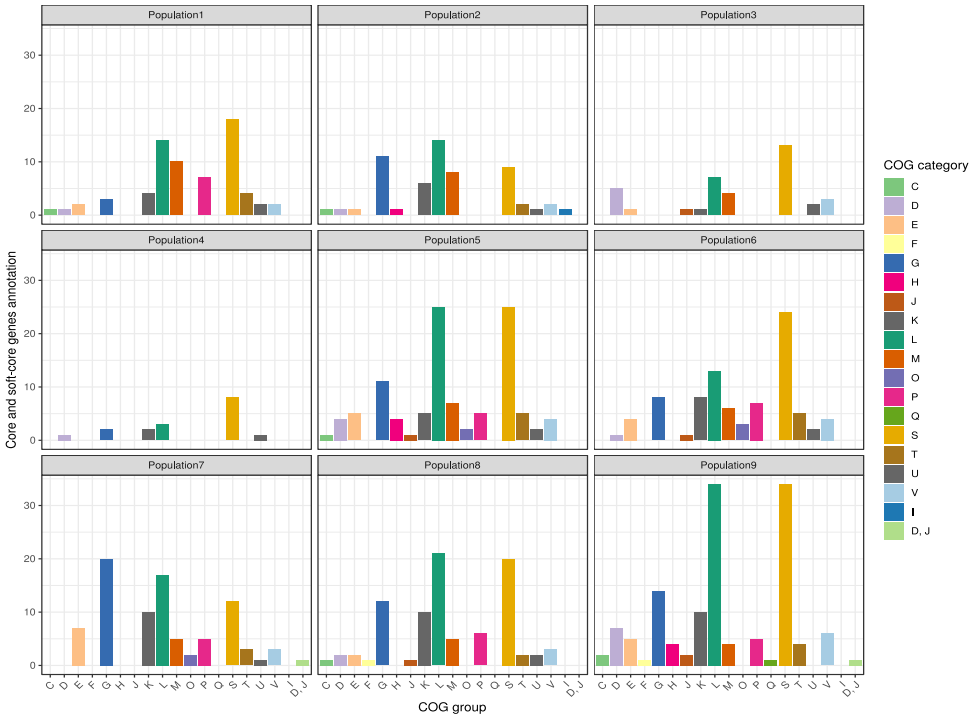


Figure S6 Number of core plasmidome genes (y axis) grouped into COG categories (x axis) from each plasmidome population

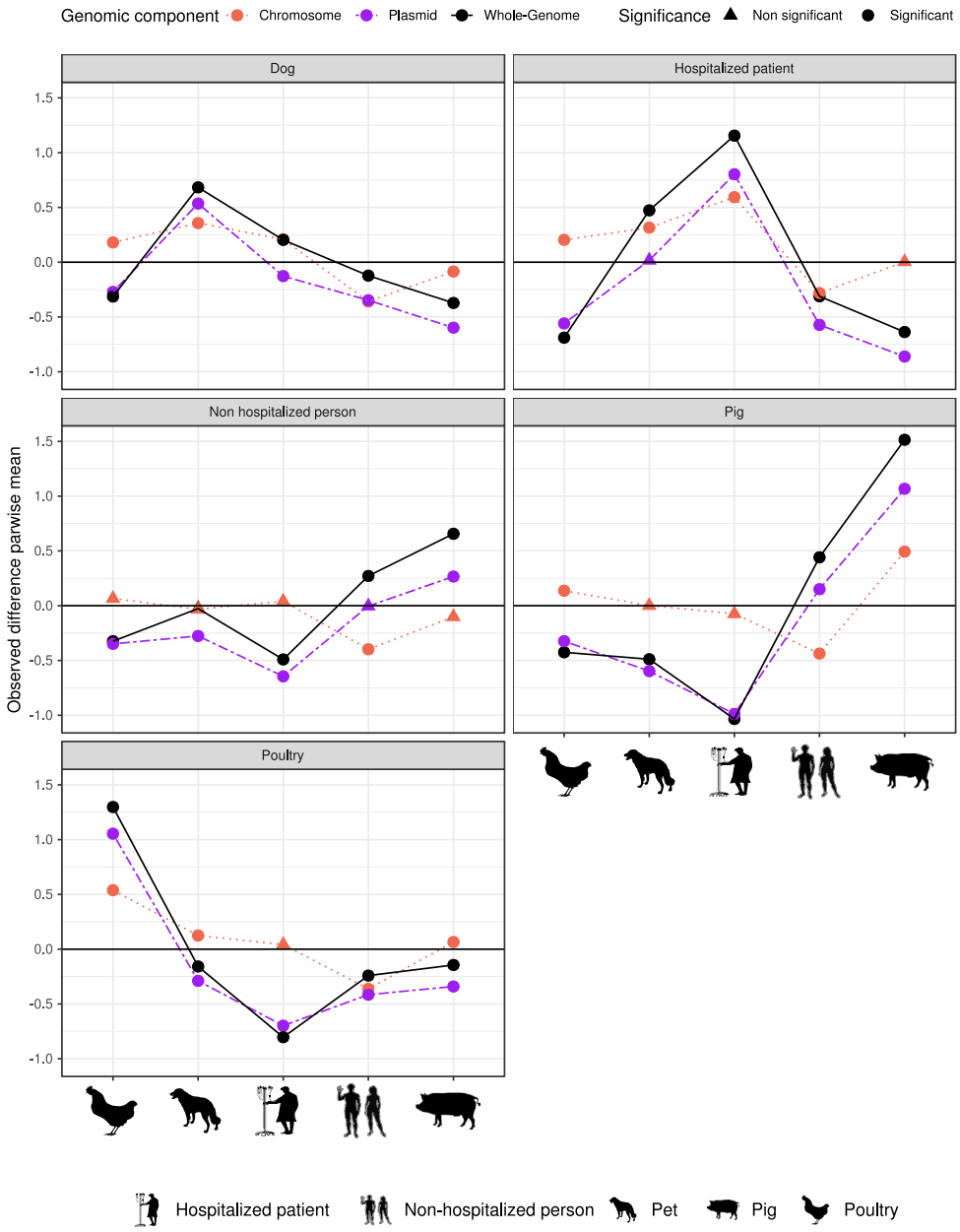


Figure S7 Differences in the observed means of average pairwise distances when comparing within-host and between-host groups against our defined random group of isolates. Each line corresponds to a different genomic component (whole genome, black solid; chromosome, dotted red; plasmid, dashed purple), and test significance is indicated based on shape (triangle, nonsignificant; circle, significant).

Supplementary Tables

Dataset S1 is available at https://mbio.asm.org/content/11/1/e03284-19.

Supplementary References

1. Teuber M, Schwarz F, Perreten V. 2003. Molecular structure and evolution of the conjugative multiresistance plasmid pRE25 of *Enterococcus faecalis* isolated from a raw-fermented sausage. Int J Food Microbiol 88:325–329.

2. Garcia-Migura L, Hasman H, Jensen LB. 2009. Presence of pRI1: a small cryptic mobilizable plasmid isolated from *Enterococcus faecium* of human and animal origin. Curr Microbiol 58:95–100.

3. Kim SW, Jeong EJ, Kang HS, Tak JI, Bang WY, Heo JB, Jeong JY, Yoon GM, Kang HY, Bahk JD. 2006. Role of RepB in the replication of plasmid pJB01 isolated from *Enterococcus faecium* JC1. Plasmid 55:99–113.

4. Francia MV, Clewell DB. 2002. Amplification of the Tetracycline Resistance Determinant of pAMα1 in *Enterococcus faecalis* Requires a Site-Specific Recombination Event Involving Relaxase. J Bacteriol 184:5187–5193.

5. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV, Hayes F. 2014. Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology. Massachusetts Eye and Ear Infirmary.

6. Huo W, Adams HM, Trejo C, Badia R, Palmer KL. 2019. A Type I Restriction-Modification System Associated with *Enterococcus faecium* Subspecies Separation. Appl Environ Microbiol 85.

7. Fernández-García L, Blasco L, Lopez M, Bou G, García-Contreras R, Wood T, Tomas M. 2016. Toxin-Antitoxin Systems in Clinical Pathogens. Toxins 8.

8. Soheili S, Ghafourian S, Sekawi Z, Neela VK, Sadeghifard N, Taherikalani M, Khosravi A, Ramli R, Hamat RA. 2015. The *mazEF* toxin-antitoxin system as an attractive target in clinical isolates of *Enterococcus faecium* and *Enterococcus faecalis*. Drug Des Devel Ther 9:2553–2561.

9. Dy RL, Przybilski R, Semeijn K. 2014. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. Nucleic acids.

10. O'Connor L, Tangney M, Fitzgerald GF. 1999. Expression, regulation, and mode of action of the AbiG abortive infection system of lactococcus lactis subsp. cremoris UC653. Appl Environ Microbiol 65:330–335.

11. Hasman H. 2005. The tcrB gene is part of the *tcrYAZB* operon conferring copper resistance in *Enterococcus faecium* and *Enterococcus faecalis*. Microbiology 151:3019–3025.

12. Lin J. 2014. Antibiotic growth promoters enhance animal production by targeting intestinal bile salt hydrolase and its producers. Front Microbiol 5:33.

13. Nilsson O, Myrenås M, Ågren J. 2016. Transferable genes putatively conferring elevated minimum inhibitory concentrations of narasin in *Enterococcus faecium* from Swedish broilers. Vet Microbiol 184:80–83.

14. Haas W, Sublett J, Kaushal D, Tuomanen EI. 2004. Revising the role of the pneumococcal *vex*-vncRS locus in vancomycin tolerance. J Bacteriol 186:8463–8471.

15. Rusniok C, Couvé E, Da Cunha V, El Gana R, Zidane N, Bouchier C, Poyart C, Leclercq

R, Trieu-Cuot P, Glaser P. 2010. Genome sequence of *Streptococcus gallolyticus*: insights into its adaptation to the bovine rumen and its ability to cause endocarditis. J Bacteriol 192:2266–2276.

16. Kurushima J, Ike Y, Tomita H. 2016. Partial Diversity Generates Effector Immunity Specificity of the Bac41-Like Bacteriocins of *Enterococcus faecalis* Clinical Strains. J Bacteriol 198:2379–2390.

17. Antipov D, Hartwick N, Shen M, Raiko M, Pevzner PA. 2016. plasmidSPAdes : Assembling Plasmids from Whole Genome Sequencing Data. Bioinformatics 32:3380–3387.

18. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

19. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.

20. Maaten L van der, Hinton G. 2008. Visualizing Data using t-SNE. J Mach Learn Res 9:2579–2605.

21. Krijthe J. 2015. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation (R package version 0.10). Computer Software.

22. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV, Hayes F. 2014. Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology.

23. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.

24. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev M a., Pevzner P a. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

25. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963.

26. Jensen LB, Garcia-Migura L, Valenzuela AJS, Løhr M, Hasman H, Aarestrup FM. 2010. A classification system for plasmids from enterococci and other Gram-positive bacteria. J Microbiol Methods 80:25–43.

27. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132.

28. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36.

5

gplas: a comprehensive tool for plasmid analysis using short-read graphs

Sergio Arredondo-Alonso, Martin Bootsma, Yaïr Hein, Malbert R. C. Rogers, Jukka Corander, Rob J. L. Willems and Anita C. Schürch

Published in: Bioinformatics (2020) doi: 10.1093/bioinformatics/btaa233

Abstract

Plasmids can horizontally transmit genetic traits, enabling rapid bacterial adaptation to new environments and hosts. Short-read whole-genome sequencing data are often applied to large-scale bacterial comparative genomics projects but the reconstruction of plasmids from these data is facing severe limitations, such as the inability to distinguish plasmids from each other in a bacterial genome. We developed gplas, a new approach to reliably separate plasmid contigs into discrete components using sequence composition, coverage, assembly graph information and network partitioning based on a pruned network of plasmid unitigs. Gplas facilitates the analysis of large numbers of bacterial isolates and allows a detailed analysis of plasmid epidemiology based solely on short-read sequence data.

Gplas is written in R, Bash and uses a Snakemake pipeline as a workflow management system. Gplas is available under the GNU General Public License v3.0 at https://gitlab.com/sirarredondo/gplas.git.

Introduction

A single bacterial cell can harbor several distinct plasmids; however, current plasmid prediction tools from short-read WGS often have a binary outcome (plasmid or chromosome). To bin predicted plasmids into discrete entities, we built a new method based on the following concepts: (i) contigs of the same plasmid have a uniform sequence coverage (1,2), (ii) plasmid paths in the assembly graph can be searched for using a greedy approach (3) and (iii) removal of repeat units from the plasmid graphs disconnects the graph into independent components (4).

Here, we refined these ideas and introduce the concept of unitigs co-occurrence to create a pruned plasmidome network. Using an unsupervised approach, the network is queried to find highly connected nodes corresponding to sequences belonging to the same discrete plasmid unit, representing a single plasmid. We show that our approach outperforms other *de novo* and reference-based tools and fully automates the reconstruction of plasmids from short reads.

Results

Gplas in combination with mlplasmids obtained the highest average precision (0.88) indicating that the predicted components were mostly formed by nodes belonging to the same discrete plasmid unit (Table 1 and Supplementary Fig. S1). The reported average completeness value (0.79) showed that most of the nodes from a single plasmid were recovered as a discrete plasmid bin by gplas (Table 1 and Supplementary Fig. S2). We observed a decline in the performance of gplas in combination with mlplasmids (precision = 0.82, completeness = 0.72) when considering uniquely bins with a size larger than one which indicated merging problems of large plasmids with a similar k-mer coverage (Supplementary Fig. S3 and Results S2). However, in all cases, the performance of gplas in combination with mlplasmids performed better than other *de novo* and reference-based tools tested here (Table 1). To show the potential of gplas in combination with mlplasmids, we showcase the performance of our approach in two distinct bacterial isolates (Supplementary Results S1 and S2).

Mlplasmids only contains a limited range of species models (Supplementary Methods). For other bacterial species, we observed that plasflow probabilities in combination with gplas performed similar than the other *de novo* approaches but also introduced bias when wrongly predicting chromosome contigs as plasmid nodes (Table 1 and Supplementary Fig. S1), thereby creating bins corresponding to chromosome and plasmid chimeras (precision = 0.62).

Table 1. Gplas benchmarking

Tool	Precision	Completeness	Bin size
gplas - mlplasmids	0.88/0.82[a]	0.79/0.72[a]	6.02/10.9[a]
gplas - plasflow	0.62/0.45[a]	0.52/0.32[a]	7.17/11.1[a]
hyasp	0.64/0.56[a]	0.36/0.30[a]	3.84/5.65[a]
mob-recon	0.79/0.71[a]	0.56/0.51[a]	3.4/7.22[a]
plasmidSPAdes	0.52/0.27[a]	0.56/0.38[a]	6.99/13.7[a]

[a]Components > 1 node

Discussion

We present a new tool called gplas, which enables the binning and a detailed analysis workflow of binary classified plasmid contigs into discrete plasmid units by relying on the structure of the assembly graph, k-mer information and partitioning of a pruned plasmidome network. A limitation of the presented approach is the generation of chimeras resulting from plasmids with similar k-mer profiles, k-mer coverage and sharing repeat unit(s), such as a transposase or an IS element. These cases cannot be unambiguously solved. Here, we integrated and extended upon features to predict plasmid sequences and exploit the information present in short-read graphs to automate the reconstruction of plasmids.

Material and Methods

Gplas algorithm

Given a short-read assembly graph (gfa format), segments (nodes) and edges (links) are extracted from the graph. Gplas uses mlplasmids (version 1.0.0, prediction threshold 0.5) or plasflow (version 1.1, prediction threshold 0.7) to classify segments as plasmid- or chromosome-derived and selects segments with an in- and out- degree of 1 (unitigs) (5,6). The k-mer coverage SD of the chromosome-derived unitigs is computed to quantify the fluctuation in the coverage of segments belonging to the same replicon unit. Plasmid-derived unitigs are considered to search for plasmid walks with a similar coverage and composition using a greedy approach (Supplementary Methods S1). Gplas creates a plasmidome network (undirected graph) in which nodes correspond to plasmid unitigs and edges are created and weighted based on the co-existence of the nodes in the solution space of the computed walks. Modularity values computed using a selection of partitioning algorithms (7,8,9) are considered to perform a voting decision regarding the split of the components into different bins (subcomponents) in the undirected network (Supplementary Methods S1). These bins represent the set of plasmids present in the bacterial isolate and are plotted in the plasmidome network using igraph R package (10). The pseudocode and formalization of the algorithm are available in Algorithm 1 and Supplementary Methods

```
Data: Graph G from SPAdes or Unicycler
Result: Plasmidome network G_P. Assignment of plasmid nodes
        N_P into different bins
Initialization;
Extract nodes N and links L from G;
Divide N as collection of plasmid-derived nodes P and
 chromosome-derived nodes C using mlplasmids or plasflow;
Discard P and C with an d^i(v) and d^o(v) != 1 and length < 1 kbp;
Determine the s_C^2 of C based on the k-mer coverage;
for each v_0 ∈ P do
    Search through all the possible plasmid-like walks W starting
     from v_0;
    for W in number of walks do
        while ∃ eligible extension E(W) do
            Consider the last v in W
            Retrieve all candidate extensions E(W)
            Compute gplas scores g(W,v) of E(W)
            Filter E(W) with a g(W,v) < ξ (default = 0.1, tunable by
             the user)
            Sample a E(W) based on the vector g(W,v)
            Extension of W using the selected v
        end
        Create a new set of links L_P connecting N_P in W;
        Reinitialize W considering again v_0 as first element;
    end
end
Compute the weights H_P of L_P based on their frequency in W;
Create a novel plasmidome network G_P(N_P, L_P,H_P);
Consider components (subgraphs) G_P^i from G_P;
for each G_P^i with N_P^i > 1 do
    Compute modularity values Q from G_P^i using three
     partitioning algorithms ;
    Consider all Q > 0.2 (tunable by the user) to split G_P^i and
     perform a voting decision ;
    Predict N_P^i as a single bin or classify N_P^i into bins based on
     the partitioning algorithm with a highest Q;
end
Classification of N_P^i in G_P^i with N_P^i = 1 as singletons;
Plot G_P with colours according to bin classification;
```

Algorithm 1. Gplas pseudocode

S1, respectively.

Benchmarking dataset

Gplas was benchmarked against current existing tools to bin plasmid contigs from short-read WGS: (i) plasmidSPAdes (*de novo*- based approach, version 3.12) (1), (ii) mob-recon (reference-based approach, version 1.4.9.1) (11) and (iii) hyasp (hybrid approach, version 1.0.0) (3). To evaluate the binning tools, we selected a set of 28 genomes with short- and long-read WGS available including 106 plasmids from 9 different bacterial species, which were not present in the databases or training sets of the tools (Supplementary Methods S3 and Table S1) (12,13,14,15).

Let n_{bin} be the total number of nodes present in the predicted bin and define *ref* as the reference replicon sequence with a highest number of nodes in each bin. Let n_{ref} be the total number of nodes comprised in *ref*. We then define two metrics commonly used in metagenomics for binning evaluation: (i) precision and (ii) completeness (Supplementary Methods S4).

$$\text{precision} = \frac{n_{\text{bin}} \in n_{\text{ref}}}{n_{\text{bin}}}$$

$$\text{completeness} = \frac{n_{\text{bin}} \in n_{\text{ref}}}{n_{\text{ref}}}$$

Acknowledgements

We would like to thank Dr Bryan Wee for testing and contributing to the development of gplas.

References

1. Antipov,D. et al. (2016) plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics, 32, 3380–3387.

2. Rozov,R. et al. (2016) Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics, 33, 475–482.

3. Müller,R. and Chauve,C. (2019) HyAsP, a greedy tool for plasmids identification. Bioinformatics, 35, 4436–4439.

4. Vielva,L. et al. (2017) PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics, 33, 3796–3798.

5. Arredondo-Alonso,S. et al. (2018) mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb. Genom., 4, e000224.

6. Krawczyk,P.S. et al. (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res., 46, e35.

7. Blondel,V.D. et al. (2008) Fast unfolding of communities in large networks. J. Stat. Mech., 2008, P10008.

8. Newman,M.E.J. (2006) Finding community structure in networks using the eigenvectors

of matrices. Phys. Rev. E, 74, 036104.

9. Pons,P. and Latapy,M. (2005) Computing communities in large networks using random walks. Computer and Information Sciences – ISCIS 3733, 284–293.

10. Csardi,G. et al. (2006) The igraph software package for complex network research. InterJ. Complex Syst., 1695, 1–9.

11. Robertson,J. and Nash,J.H.E. (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb. Genom., 4, e000206.

12. Arredondo-Alonso,S. et al. (2020) Plasmids shaped the recent emergence of the major nosocomial pathogen enterococcus faecium. mBio 11, e03284-19.

13. De Maio,N. et al.; on behalf of the REHAB Consortium. (2019) Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb. Genom., 5, e000294.

14. Decano,A.G. et al. (2019) Complete assembly of *Escherichia coli* sequence type 131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. mSphere, 4, e00130.

15. Wick,R.R. et al. (2017) Completing bacterial genome assemblies with multiplex MinION sequencing. Microb. Genom., 3, e000132

Supplementary Results

Gplas showcase: *K. pneumoniae* KSB1_7G (S1)

To show the potential of gplas to bin contigs into discrete plasmid units, we showcase the isolate Klebsiella pneumoniae KSB1_7G (Supplementary Table S1). This genome has several features that allows to showcase the features used by gplas (Table S2): i) presence of two discrete components showing a similar k-mer coverage and containing plasmid starting nodes, ii) a low number of nodes that reduces the total number of walks, indispensable for visualization purposes and iii) no dead-ends present in the short-read graph.

Component B (Figure S5, left component) and C (Figure S5, right component) correspond to two discrete plasmid sequences with an approximated length of 161 kbp and 112 kbp respectively (Figure S5). Component C has a unique edge corresponding to a self-loop since that plasmid sequence has circularization signatures present in the boundaries of the node 15.

Gplas uses mlplasmids (or plasflow, in case a bacterial species not listed in mlplasmids is chosen) to predict plasmid- derived and chromosome-derived nodes. Predicted plasmid nodes with an in- out-degree equal to 1 (unitigs) are considered as plasmid starting nodes. Chromosome-derived nodes corresponding to unitigs are used to calculate the k-mer coverage standard deviation (0.034) present in nodes belonging to the same replicon. Plasmid starting- nodes (15, 18, 20, 24, 25, 26, 27) are considered to search for plasmid-like walks. For the purpose of visualization procedures, we create a space search of 5 solutions per each plasmid starting node. We use the plasmid starting node 18+ to show the work-

flow of gplas:

- 18+ is the first node of the walk.

- The coverage of the walk corresponds to 1.35.

- The unique outcoming edge (18+,33-) corresponds to a connection with a repeat unit since 33- has two incoming and two outcoming edges (in-degree and out-degree of 2) resulting in a coverage of 2.70. This outcoming edge (18+,33-) is assigned with a default gplas score of 0.25 corresponding to a default k-mer composition and coverage score of 0.5 respectively.

- We sample the vector of gplas scores to select for an outcoming edge to elongate the walk. In this case, only a single edge can be selected to elongate the walk.

- We include the node (33-) from the outcoming edge (18+,33-). The first elongation results in the walk: 18+,33-

-We update the k-mer coverage of the walk. In this case, the coverage of the walk remains 1.35 since the k- mer coverage of 33- is not considered. Only unitig nodes with a length larger than 1 kbp are considered in this step.

- There are two outcoming edges from the node 33- : i) 33-, 20- and ii) 33-, 24+.

- We calculate the probcoverage and retrieve the probplasmid of the outcoming edges. In this case, the edges connect to nodes showing a similar probcoverage (Figure S6 and S7) since both are part of the same replicon. We retrieve the probplasmid corresponding to the probabilities of being plasmid-derived: 0.86 (20-) and 0.94 (24+). This results in a gplas score of 0.31 (33-, 20-) and 0.48 (33-,24+).

- We sample one of the outcoming two edges and elongate the walk with the selected connection.

- In this case, we update the k-mer coverage of the walk using 18- and 20- or 24+, depending on which outcoming node was selected.

We follow this procedure until reaching the following four scenarios: i) all outcoming edges have a final gplas score lower than the filtering threshold (default = 0.10), ii) no outcoming edges are available to elongate the walk (e.g. dead-end), iii) the last node incorporated in the walk corresponds to the initial plasmid starting node (circularization signature) and iv) the length of the path exceeds 100 nodes. We repeat this procedure 5 times (only for visualization purposes) and end up with the solutions illustrated in Figure S8. We find the solution: '18+,33-,20-,31-,25-,31+,18+' repeated 4 times and the solution: '18+, 33-, 24+, 38-, 27-, 38+, 26-, 33-, 20-, 31-, 25-, 21+, 18+' (third panel in Figure S8).

Interestingly, we find repeated 5 times the solution '18-,31-,25+,31+,20+,33+,18-' if the

plasmid search is initialized from the opposite direction 18- (Figure S9). We repeat the same approach with the other plasmid starting nodes (15, 20, 24, 25, 26 and 27) to obtain our space of solutions.

Using this space of solutions, we create a novel set of edges connecting all the plasmid unitigs present within the same solution. Furthermore, we weight these edges based on the frequency in which they appear in the set of solutions. This is considered to create a new plasmidome network corresponding to an undirected graph in which nodes are plasmid unitigs and edges correspond to the co-occurrence of these plasmid unitigs in the solutions with different weights based on the frequency of the connections.

The bins with a size larger than 1 node and present in the plasmidome network are queried to decide whether to split the component into different bins (subcomponents) or predict a single bin and retain the original component. For this purpose, we compare three different partitioning algorithms: i) walktrap (cluster_walktrap function in igraph R package (4, 11)), based on finding communities in the graph via random walks, ii) leading eigen (cluster_leading_eigen function in igraph R package (3, 4)), based on the calculation of the leading non-negative eigen vector from the graph modularity matrix, iii) Louvain method (cluster_louvain function in igraph R package (4, 10)), based on a greedy way implementation that tries to maximize the modularity by reassigning vertices from the graph.

The modularity of the graph after partitioning the networks with each algorithm is computed using the formula described in the function modularity from the igraph R package(4).

We consider a modularity value larger than 0.2 as a support vote to split the component into different bins (sub components). We perform a voting decision between the four listed algorithms and in case there is a majority of algorithms supporting the split of the component, we consider the algorithm achieving a highest modularity to split the component into the proposed number of bins.

For the components present in the plasmidome network with a size equal to 1, we classified them as singletons which are bins with only one node. These usually correspond to small plasmids in which there is a self-edge already present in the original graph given to gplas.

We finally incorporate the bin assignment of the nodes (Table S3) into the plasmidome network by using different colours. Thus, nodes belonging to the same bin are identically coloured in the final plasmidome network (Figure S10).

In this simplified example in which we only requested for the search of 5 plasmid-like walks per starting node, two different bins were obtained. Bin 2 contains a single node,

15, with a self-loop indicating the presence of circularization signatures (precision 1.0, completeness = 1.0). Interestingly, even though node 15 has a k-mer coverage similar to other plasmid starting nodes, the structure of the original assembly graph indicated that it corresponded to an independent plasmid sequence and thus no edges can cross these sequences. We encounter this scenario only if the plasmid sequences share no repeat sequences. This highlights the importance of searching for walks using the original graph rather than simply binning nodes based on similar k-mer coverage or composition. Bin 1 (precision= 1.0, completeness = 1.0) belongs to plasmid 2 (Table S3).

Gplas showcase on plasmids with a similar k-mer coverage (S2)

To highlight the main limitation of gplas, we show the results obtained for the *K. pneumoniae* isolate SAMN10819819 (Supplementary Table S1). This isolate contained five plasmids (Figure S11), and two of them corresponded to large plasmids with a very similar k-mer coverage (1.69x and 1.66x) and length (107.577 kbp and 88.581 kbp).

The short-read graph associated to this isolate was complex with a total of 289 nodes, 398 edges and 12 dead-ends (Figure S12). Gplas in combination with mlplasmids predicted a total of 4 bins in the plasmidome network shown in Figure S13. Two of these four bins (light and dark green) had only one node and were formed by the small plasmids, node 84 and node 81 (Figure S13). For these cases, gplas obtained a precision and completeness of 1.0.

The other two bins corresponded to: i) orange bin (reported as bin number 2 in Supplementary Table S2) with a precision and completeness of 1.0, formed by the nodes of the 175.881 kbp plasmid (Figure S11) and ii) light blue bin (reported as bin number 1 in Supplementary Table S2) with a precision of 0.58 and completeness of 1.0 with a mixture of nodes from the 107.577 kbp and 88.581 kbp plasmid (Figure S11).

In the case of light blue bin (Figure S13), gplas in combination with mlplasmids created paths corresponding to chimeras between these two plasmids since they share repeat units and have a similar k-mer coverage that resulted in the acceptance of connections belonging to different plasmid units.

Supplementary Methods

Gplas formalization (S1)

Gplas requires a single input corresponding to a graph in gfa format (version 1.0) (`https://github.com/GFA-spec/GFA-spec`). We can define the nodes N present in a graph G by:

$$N = \{N_1, N_2, ..., N_n\},$$

i.e., the number of nodes in G, $|N|$ equals n. Nodes can also be referred to as segments, contigs or vertices. Furthermore, we can define the set of links L as the set of directed connections between two elements of N:

$$L = \{(i, j) : |1 \leq i, j \leq n \quad \text{with a link from } N_i \text{ to } N_j\}.$$

Links can also be referred as edges. We also denote the edge from vertex v to vertex w by $e(v, w)$. We can define the graph G given by the user as:

$$G = (N, L).$$

For each $v \in N$, we define the in-degree $d^i(v)$ and the out-degree $d^o(v)$ as the number of edges in L towards and from vertex v respectively. We define $|v|$ as the length of segment v, i.e., the number of nucleotides of segment v.

Mlplasmids (version 1.0.0) provides for each $v \in N$ a probability whether v is plasmid-derived if the given G corresponds to a bacterial species included in the tool (*Enterococcus faecium*, *Klebsiella pneumoniae* or *Escherichia coli*). To generalize the prediction to other bacterial species, the probabilily that the segment is plasmid-derived can also be derived using plasflow (version 1.1) which provides a metagenomics classifier to retrieve plasmid-derived sequences. For either method, we define $\pi(v)$ as the probability that segment v is plasmid-derived.

For a graph G generated by SPAdes, the k-mer count information of node N_i is extracted from the "kc" tag present in the header line of N_i for each node $1 \leq i \leq n$. The k-mer count is divided by $|N_i|$ and normalised against the median k-mer count. If G was generated from Unicycler, the normalised depth of N_i (k-mer coverage) is retrieved from the tag "dp" which is already present in the header line of N_i. In both cases, this value is further considered and defined as the *coverage* of a particular node $v \in N$, and denoted by $c(v)$.

Let t be the threshold for posterior probability of the plasmid class π reported by mlplasmids or plasflow. We then define $\mathcal{P}$ as the set of plasmid-predicted contigs by:

$$\mathcal{P} = \left\{v \in N : \pi(v) \geq t, d^i(v) = d^o(v) = 1, |v| \geq 1\,kbp\right\}.$$

And $\mathcal{C}$ as the set of chromosome-predicted contigs by:

$$\mathcal{C} = \left\{v \in N : \pi(v) < t, d^i(v) = d^o(v) = 1, |v| \geq 1\,kbp\right\}.$$

$|\mathcal{P}|$ and $|\mathcal{C}|$ are the number of contigs that are plasmid-predicted and chromosome-predicted, respectively. For the set of chromosome-predicted contigs, we define the average coverage $\mu_{\mathcal{C}}$ and the variance of the coverage $s^2_{\mathcal{C}}$ by:

$$\mu_{\mathcal{C}} = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}} c(v)$$

$$s^2_{\mathcal{C}} = \frac{1}{|\mathcal{C}|-1}\sum_{v\in\mathcal{C}}(c(v)-\mu_{\mathcal{C}})^2.$$

For a graph $G = (N, L)$, a walk of length $k \in \mathbb{N}$ is defined as a finite sequence of alternating vertices and edges $(v_0, e(v_0, v_1), v_1, e(v_1, v_2), \ldots, e(v_{k-1}v_k), v_k)$ with $v_i \in N$ for $0 \leq i \leq k$ and edge $e(v_{i-1}v_i)$ is the edge in L from v_{i-1} to v_i for all $1 \leq i \leq k$. We denote the set of vertices $\{v_0, v_1, \ldots, v_k\}$ of a walk W by $V(W)$.

Here we consider a special class of walks with the following properties:

1. $v_0 \in \mathcal{P}$, i.e., the walk starts from a segment which is predicted to be plasmid-derived.
2. $k \leq M = 100$, i.e, we only consider of walks of length less or equal to $M = 100$.
3. $v_i \neq v_0$ for $1 \leq i < k$ i.e., the walks either does not return to the starting vertex or ends when it returns to the starting vertex.

If we have a walk $W = (v_0, e(v_0, v_1), v_1, e(v_1, v_2), \ldots, e(v_{k-1}, v_k), v_k)$ which can be extended, i.e., $k < M$ and $v_k \neq v_0$, we want to quantify whether a node v for which the edge $e(v_k, v)$ exists, is a likely extension of the walk or not. We call such a node v a candidate extension node. Slightly abusing notation, we define the average coverage of walk W, $c(W)$, as:

$$c(W) = \frac{1}{|V(W)\bigcap(\mathcal{C}\cup\mathcal{P})|}\sum_{v\in(V(W)\bigcap(\mathcal{C}\cup\mathcal{P}))} c(v),$$

i.e., we base the average coverage of W only on the coverage of the nodes of W which are at least 1 *kbp* long and have an in-degree and out-degree equal to one.

Alternatively, we can also determine the average coverage by weighting each contig by its length. We then obtain an alternative average coverage of walk W, denoted by $\tilde{c}(W)$, defined by:

$$\tilde{c}(W) = \frac{\sum_{v\in(V(W)\bigcap(\mathcal{C}\cup\mathcal{P}))} c(v)|v|}{\sum_{v\in(V(W)\bigcap(\mathcal{C}\cup\mathcal{P}))} |v|}.$$

If the candidate extension node v belongs to either $\mathcal{C}$ or $\mathcal{P}$, we want to determine how similar $c(v)$ and $c(W)$ are. For simplicity, we assume that the variance in the coverage of a node due to chance events related to the sequencing process is $s^2_{\mathcal{C}}$, i.e., the variance in the coverage of the chromosome-predicted contigs, more precisely, we assume that the coverage of the next contig of the walk is distributed according to a normal distribution with mean $c(W)$ and variance $s^2_{\mathcal{C}}$ if the walk and the candidate extension node belong to the same plasmid. We define the similarity in coverage between the walk W and the candidate extension node v, $S(W, v)$ as:

$$S(W, v) := \Phi(\frac{c(v)-c(w)}{s_{\mathcal{C}}}+1) - \Phi(\frac{c(v)-c(w)}{s_{\mathcal{C}}}-1)$$

with Φ the cumulative distribution function of the standard normal distribution. This means that the closer $c(v)$ is to $c(W)$, the higher $S(W, v)$.

The similarity $S(W, v)$ is based only on the *coverage*. In order to avoid chimeras between chromosomal and large plasmid replicons with a similar *coverage*, we define a score $g(W, v)$ of each candidate extension node by:

$$g(W, v) = \begin{cases} 0.25 & \text{if } v \notin (\mathcal{P} \cup \mathcal{C}) \\ S(W, v) \cdot \pi(v) & \text{if } v \in (\mathcal{P} \cup \mathcal{C}) \end{cases}$$

Candidate extension nodes v assigned with a score g of 0.25 may belong to repeat units such as transposases or IS elements and thus $\pi(v)$ and $S(W, v)$ cannot be confidently estimated. Additionally, we apply a filtering threshold ξ (default $\xi = 0.10$) to avoid the selection of edges potentially leading to the creation of replicon chimeras. This threshold can be tuned by the user to accept or reject a higher number of connections.

Let $E(W)$ be the set of candidate extension nodes of walk W. We define the function $h_W(v)$ by:

$$h_W(v) = \begin{cases} 1 & \text{if } g(W, v) \geq \xi \\ 0 & \text{if } g(W, v) < \xi \end{cases},$$

i.e., $h_W(v)$ equals one, if the candidate extension node has a score higher or equal to the threshold ξ.

If $h_W(v) = 0$ for all $v \in E(W)$, we have reached a dead-end, otherwise we select the extension $v \in E(W)$ with probability

$$\frac{g(W, v) h_W(v)}{\sum_{v' \in E} g(W, v') h_W(v')}$$

This results in the walk W' which is the walk W extended with the node v, i.e.,

$$W' = (v_0, e(v_0, v_1), v_1, e(v_1, v_2), \ldots, e(v_{k-1}, v_k), v_k, e(v_k, v), v).$$

Starting from a node which is plasmid-predicted, we repeat this extension procedure of the walk until we reach one of the following scenarios:

1. There are no outgoing links from the last element of W i.e., we have reached a dead-end.
2. The last node incorporated in W corresponds to the starting node of the walk.
3. The length of the walk $|W|$ exceeds $M = 100$.

For each $v \in \mathcal{P}$, we generate K walks (default $K = 20$, but tunable by the user).

We use these $|\mathcal{P}|K$ walks, denoted by $\{W_1, W_2, \ldots W_{|\mathcal{P}|K}\}$ to generate a new undirected pruned plasmidome graph, which we denote by $G_{\mathcal{P}}(\mathcal{P}, L_{\mathcal{P}}, H_{\mathcal{P}})$. The nodes of $G_{\mathcal{P}}$ are the plasmid-predicted contigs. There is an edge between two contigs v and w if the contigs co-occur in a walk, and the weight of an edge $e(v, w) \in L_{\mathcal{P}}$, denoted by $H_{\mathcal{P}}(e(v, w))$ is the number of times the two contigs co-occur in a walk.

More formally, to define the edges of G_P, denoted by $L_{\mathcal{P}}$, we first define for two nodes $v, w \in \mathcal{P}$, a set $J(v, w)$ which describes the walks in which both v and w are present, i.e.,

$$J(v, w) := \{i \in \mathbb{N} : 1 \leq i \leq |\mathcal{P}|K, v \in V(W_i) \text{ and } w \in V(W_i)\}.$$

Let $|J(v,w)|$ be the number of walks in which both v and w are present.

With this definition, we can define the set of edges of the plasmidome graph:

$$L_{\mathcal{P}} = \{e(v,w) : v,w \in \mathcal{P} \text{ and } |J(v,w)| > 0\}.$$

The function $H_{\mathcal{P}} : L_{\mathcal{P}} \to \mathbb{N}$ denotes the weight of each edge, i.e., $H_{\mathcal{P}}(e(v,w)) = |J(v,w)|$.

With this approach we create non-directed links between plasmid unitigs avoiding intermediary links connecting to nodes not present in $\mathcal{P}$. These intermediary nodes can correspond to either unitgs which are classified as chromosomal, or to transposases or repetitive elements that are shared between replicon sequences in G.

Next, the plasmidome network $G_{\mathcal{P}}$ is split into its connected subgraphs, i.e., let $\{N^1_{\mathcal{P}}, N^2_{\mathcal{P}}, \ldots, N^m_{\mathcal{P}}\}$ be the unique partition of $\mathcal{P}$ with the following properties:

- $N^i_{\mathcal{P}} \neq \emptyset \quad \forall 1 \leq i \leq m$.
- For $v,w \in N^i_{\mathcal{P}}$ with $v \neq w$, there exists a path from v to w in $G_{\mathcal{P}}$, i.e., there is a finite sequence of edges from $L_{\mathcal{P}}$ which connects v to w.
- For $v \in N^i_{\mathcal{P}}$ and $w \in N^j_{\mathcal{P}}$ and $i \neq j$, there is no path from v to w.

For each element of the partition the corresponding graph keeps the original edges and weight, i.e., the subgraph $G^i_{\mathcal{P}}(N^i_{\mathcal{P}}, L^i_{\mathcal{P}}, H^i_{\mathcal{P}})$ corresponding to the set $N^i_{\mathcal{P}}$ of the partition, is the graph with as nodes $N^i_{\mathcal{P}}$, as edges the set $L^i_{\mathcal{P}} := \{e(v,w) \in L_{\mathcal{P}} : v,w \in N^i_{\mathcal{P}}\}$ and as weight the function $H^i_{\mathcal{P}} : L^i_{\mathcal{P}} \to \mathbb{N}$ which is the restriction of $H_{\mathcal{P}}$ to $L^i_{\mathcal{P}}$.

The subgraphs $G^i_{\mathcal{P}}$ consisting of more than 1 node are queried against three different partitioning algorithms available in the igraph R package[4]: i) walktrap (cluster_walktrap function in igraph R package[4, 11]), based on finding communities in the graph via random walks, ii) leading eigen (cluster_leading_eigen function in igraph R package[3, 4]), based on the calculation of the leading non-negative eigen vector from the graph modularity matrix, iii) Louvain method (cluster_louvain function in igraph R package[4, 10]), based on a greedy way implementation that tries to maximize the modularity by reassigning vertices from the graph.

To decide whether to split $G^i_{\mathcal{P}}$ into further subcomponents (later referred as bins) or predict a single bin for $G^i_{\mathcal{P}}$, we compute the modularity of $G^i_{\mathcal{P}}$ with the three algorithms listed above. For this purpose, we use the modularity function implemented in the igraph R package[4] in which modularity is explicitly defined as:

$Q = \frac{1}{2m} \sum_{i,j} \frac{A_{ij} - k_i k_j}{2m} \delta(c_i, c_j)$

here m is the number of edges, A_{ij} is the element of the adjacency matrix A in row i and column j, k_i is the degree of i, k_j is the degree of j, c_i is the type (or component) of i, c_j that of j, the sum goes over all i and j pairs of vertices, and $\delta(x,y)$ is 1 if $x = y$ and 0 otherwise.

We consider a modularity value Q larger than 0.2 (tunable by the user) as a support value to split the component $G^i_{\mathcal{P}}$ into different bins (subcomponents). We perform a voting decision between the four listed algorithms and in case there is a majority of algorithms supporting the split of the $G^i_{\mathcal{P}}$, we consider the algorithm achieving a highest Q to split $G^i_{\mathcal{P}}$ and classify $N^i_{\mathcal{P}}$ into different bins according to the partitioning solution found.

Disconnected subgraphs from $G_{\mathcal{P}}$ with $|N^i_{\mathcal{P}}| = 1$ are classified into singletons (bins with a single node).

Gplas finally reports how the elements of $\mathcal{P}$ are distributed over the bins according to the above classification. Ele-

ments of $\mathcal{P}$ belonging to the same bin are considered as coming from the same plasmid replicon unit. $G_{\mathcal{P}}$ is represented using the igraph R package[4] and the nodes, i.e., $\mathcal{P}$, are coloured according to the bin classification.

Gplas pseudocode (S2)

Data: Graph G from SPAdes or Unicycler
Result: Plasmidome network $G_{\mathcal{P}}$. Assignment of plasmid nodes $N_{\mathcal{P}}$ into different bins
Initialization;
Extract nodes N and links L from G;
Divide N as collection of plasmid-derived nodes $\mathcal{P}$ and chromosome-derived nodes $\mathcal{C}$ using mlplasmids or plasflow;
Discard $\mathcal{P}$ and $\mathcal{C}$ with an $d^i(v)$ and $d^o(v)$!= 1 and length $<$ 1 kbp;
Determine the $s^2_{\mathcal{C}}$ of $\mathcal{C}$ based on the k-mer coverage;
for *each* $v_0 \in \mathcal{P}$ **do**
  Search through all the possible plasmid-like walks W starting from v_0;
  for *W in number of walks* **do**
    while $\exists$ *eligible extension E(W)* **do**
      Consider the last v in W
      Retrieve all candidate extensions $E(W)$
      Compute gplas scores g(W,v) of $E(W)$
      Filter $E(W)$ *with a g(W,v)* $< \xi$ *(default = 0.1, tunable by the user)*
      Sample a $E(W)$ *based on the vector g(W,v)*
      Extension of W using the selected v
    end
    Create a new set of links $L_{\mathcal{P}}$ *connecting* $N_{\mathcal{P}}$ *in W;*
    Reinitialize W considering again v_0 *as first element;*
  end
end
Compute the weights $H_{\mathcal{P}}$ of $L_{\mathcal{P}}$ based on their frequency in W;
Create a novel plasmidome network $G_{\mathcal{P}}(\mathcal{P}, L_{\mathcal{P}}, H_{\mathcal{P}})$;
Consider components (subgraphs) $G^i_{\mathcal{P}}$ from $G_{\mathcal{P}}$;
for *each* $G^i_{\mathcal{P}}$ *with* $|N^i_{\mathcal{P}}| > 1$ **do**
  Compute modularity values Q from $G^i_{\mathcal{P}}$ using three partitioning algorithms ;
  Consider all $Q > 0.2$ (tunable by the user) to split $G^i_{\mathcal{P}}$ and perform a voting decision ;
  Predict $N^i_{\mathcal{P}}$ as a single bin or classify $N^i_{\mathcal{P}}$ into bins based on the partitioning algorithm with a highest Q;
end
Classification of $N^i_{\mathcal{P}}$ in $G^i_{\mathcal{P}}$ with $|N^i_{\mathcal{P}}| = 1$ as singletons;
Plot $G_{\mathcal{P}}$ with colours according to bin classification;

Benchmarking dataset (S3)

To evaluate the performance of gplas against existing plasmid binning tools, we selected a set of 28 genomes with short- and long-read WGS available including 106 plasmids from 9 different bacterial species (Supplementary Table S1) (2,5,6,14). These genomes were selected due to their release date (after June 2017) to avoid any bias in favour of reference-based approaches as a result of including plasmids present in the databases of the tools. Importantly, these genomes were not part of the training sets of mlplasmids or plasflow.

We trimmed short-reads using trim galore (version 0.6.1) and determined a minimum quality phred score of 20 (8). Long-reads were filtered out using filtlong (v0.2.0) (https://github.com/rrwick/Filtlong.git) specifying a minimum length of 1 kbp, removing 10% of the worst base reads, filtering out reads with a mean quality weight inferior to 20 using short-reads as references and finally keeping a number of kbp corresponding to a genome coverage of 20x. We subsequently used Unicycler (version v0.4.7) using SPAdes (version 3.12.0) to perform a hybrid assembly and obtain complete genomes (15). If the assembly resulted in an non-complete genome, we retrieved the uncompleted path using Bandage (version 0.8.1) (13) and used filtlong (v0.2.0) indicating the non-completed path as external ref- erence and selecting reads from that path until reaching a path coverage of 10x. Previous filtered long-reads and reads corresponding to the non-completed path were merged and passed to Unicycler to rerun the hybrid assembly and obtain a complete genome.

Benchmarking tools (S4)

PlasmidSPAdes (from SPAdes 3.12) was run using default parameters and specifying the trimmed short-reads (1). Hyasp (version 1.0.0) was run using the flag "–bin" and specifying as input the graph created by Unicycler using only short-reads and after removing overlaps (002_overlaps_removed.gfa) (9). We created the database proposed by hyasp authors to identify plasmid seeds using the file "ncbi_database_genes.fasta" provided in the hyasp github repo (https://github.com/cchauve/HyAsP). Mob-recon (version 1.4.9.1) was run using default values with the proposed database (https://github.com/phac-nml/mob-suite) after first-installation of mob-suite (12). Nodes from the short-read graph (002_overlaps_removed.gfa) were used as input for mob-recon. Gplas (version 0.6.1) was run with the following values: f = 0.1, x = 50, q = 0.2, mlplasmids threshold = 0.5, plasflow threshold = 0.7 and indicating the short-read graph derived from Unicycler (15) as input '002_overlaps_removed.gfa' (Supplementary Table S1).

We used Quast (version 4.6.3) to map nodes predicted as belonging to the same bin against the complete genomes from the same bacterial isolate (7). To determine the precision and completeness (see section below) of the predictions, we only considered nodes

with a length larger than 1 kbp, mapping unambiguously to a replicon sequence and thus excluding transposases and other repetitive sequences mapping totally or partially to more than one genome sequence.

Benchmarking metrics (S5)

For each bin predicted by the tools, we determined which reference replicon (either plasmid or chromosome) had a larger representation in terms of number of nodes. The purity of the predicted bins (precision) was defined as the number of nodes belonging to the reference replicon and present in the bin divided by the total size (nodes) of the bin. Completeness was defined as the number of nodes belonging to the reference replicon in the bin divided by the total size (nodes) of the reference replicon.

Bins in which the most predominant reference replicon was the chromosome unit were assigned with a precision and completeness of 0. We strongly penalized these bin predictions since they are mostly formed and contaminated by chromosome-derived sequences.

To evaluate the tools, we filtered out bins with a single node (size = 1) if the associated reference replicon had a size larger than 1. In this way, we could still consider bins corresponding to small plasmids in which the reference replicon is formed by a single node. However, these bins can mask the problems of binning present in medium and large plasmids formed by several plasmid unitigs and with a similar k-mer coverage. To elucidate the performance of the tools in these cases, we also reported the average precision and completeness of the tools in predicted bins with a size (nodes) larger than 1.

Gplas output files (S6)

- results/*results.tab: Tab delimited file containing the full output information retrieved by gplas. The file contains the following columns: contig number, probability of being chromosome-derived, probability of being plasmid-derived, class prediction, contig name, k-mer coverage, length, bin assigned.

- results/*bins.tab: Tab delimited file containing the bin prediction reported by gplas with the following columns: contig number, bin assignment.

- results/*bin*.fasta: Fasta file generated per each bin containing the nodes assigned to them.

- results/*plasmidome_network.png: Png file of the plasmidome network generated by gplas after creating an undirected graph considering as nodes the plasmid unitigs and as edges the co-occurrence of these

nodes in the space of solutions.

- walks/*solutions.csv: Comma-separated file containing the plasmid-like walks generated by gplas.

- walks/*connections.tab: Tab delimited file containing the following information: Variance factor, number of tries selected by the user, try number, elongation number, plasmid starting node, last node of the walk, possible outcoming node, probplasmid, probcoverage, gplas score, frequency of the score, selection or discardance of the connection. This file may facilitate the visualization of the walks generated by gplas as exemplified in Figures S8 and S9.

Supplementary Tables

Supplementary Table S1 is available online at:

https://doi.org/10.1093/bioinformatics/btaa233

Table S2. KSB1_7G summary short-read graph stats

KSB1_7G short-read graph	Class	Nodes	Edges	Dead-ends
Component A	Chromosome	74	100	0
Component B	Plasmid 1	9	12	0
Component C	Plasmid 2	1	1	0

Table S3. Assignment of the plasmid unitigs into different bins for the isolate *K. pneumoniae* K B1_7G. Colour of the bins corre- spond to the assignment given in Figure S10.

Plasmid unitig	Bin
15	2 (orange)
18	1 (light blue)
20	1 (light blue)
24	1 (light blue)
25	1 (light blue)
26	1 (light blue)
27	1 (light blue)

Supplementary Figures

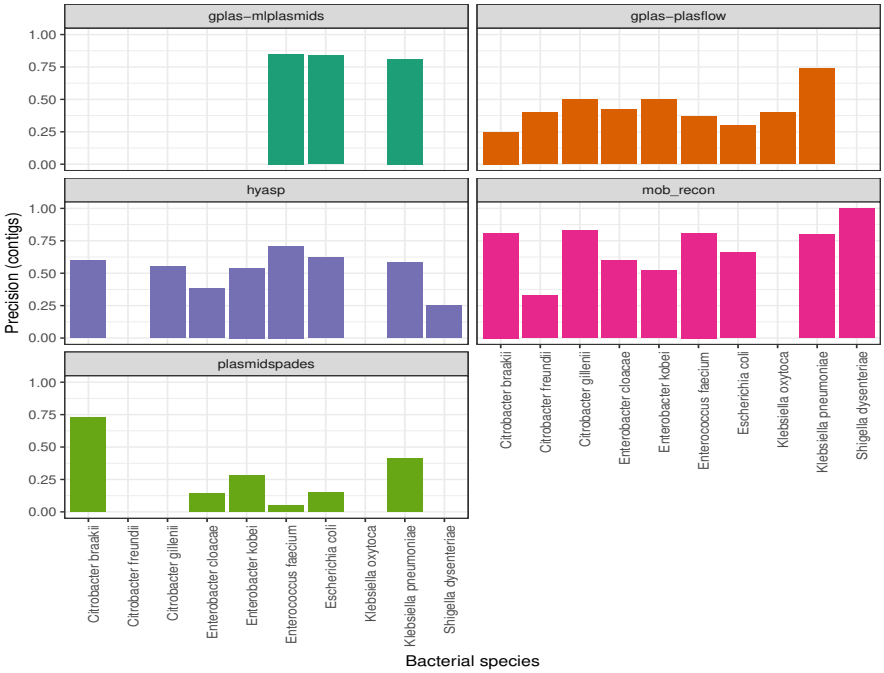


Figure S1. Barplot with the precision (y-axis) achieved by the different tools depending on the bacterial species analysed (x-axis).

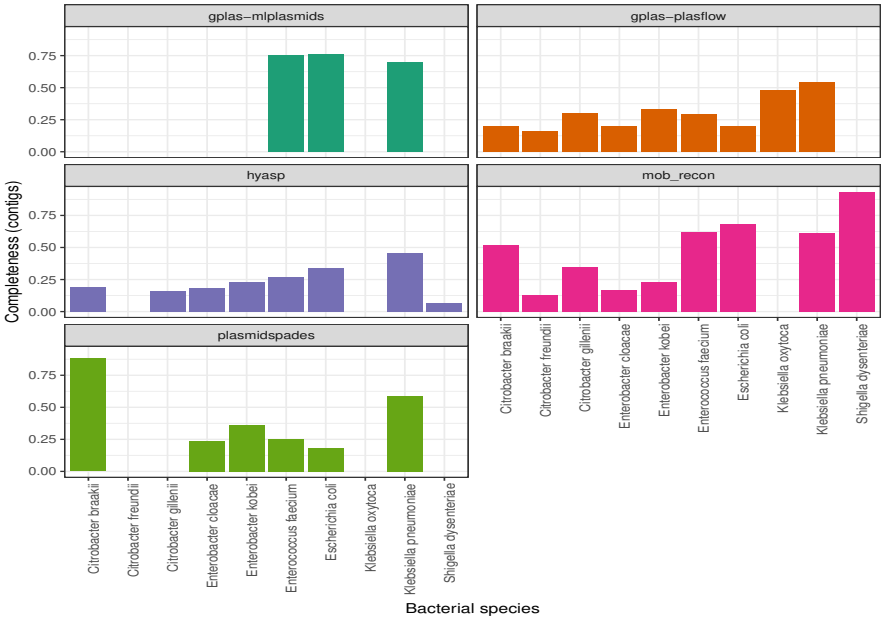


Figure S2. Barplot of the completeness (y-axis) achieved by the different tools depending on the bacterial species analysed (x-axis).

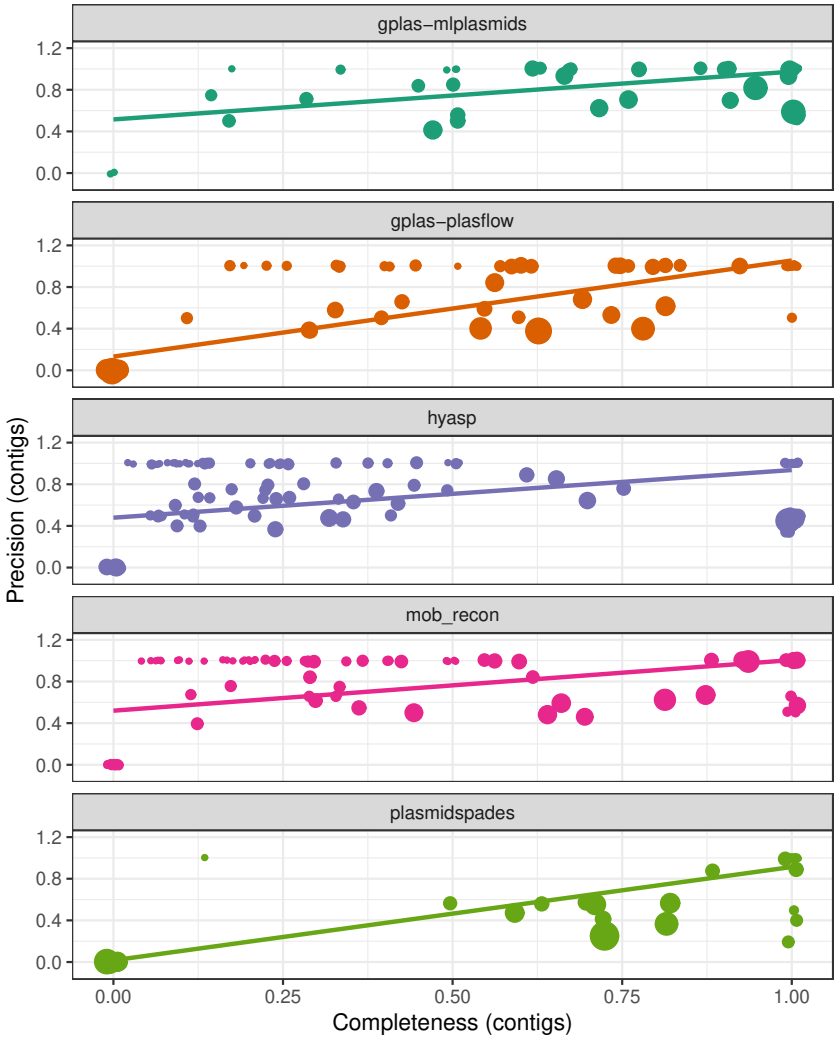


Figure S3. Scatterplot of the completeness (x-axis) and precision (y-axis) obtained by each bin predicted by the tools included in the benchmarking. For each classifier, we fitted a linear regression model and indicated the standard error with a shadow area to observethe correlation between precision and completeness in each of the predictions.

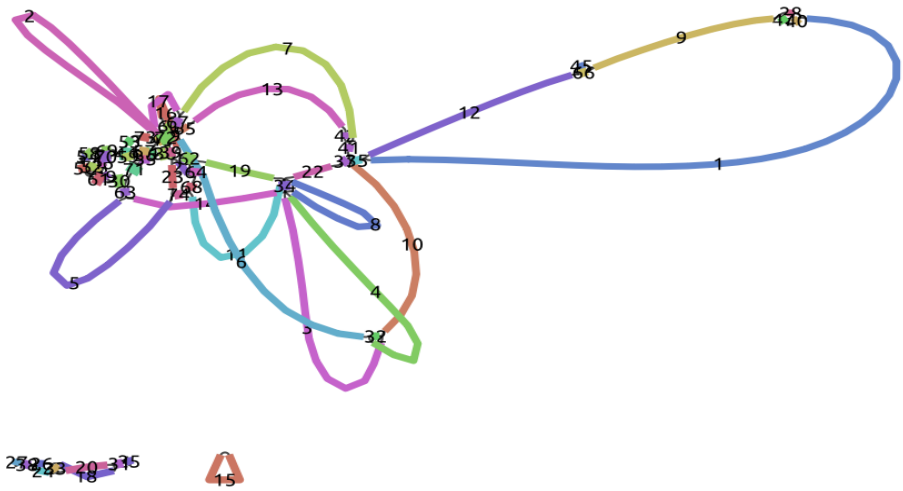


Figure S4. Visualization of the entire KSB1_7G isolate short-read WGS graph using Bandage.

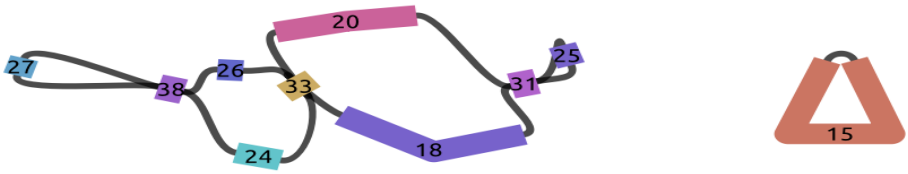


Figure S5. Zoom-in of the two components corresponding to plasmid sequences from KSB1_7G isolate short-read WGS graph .

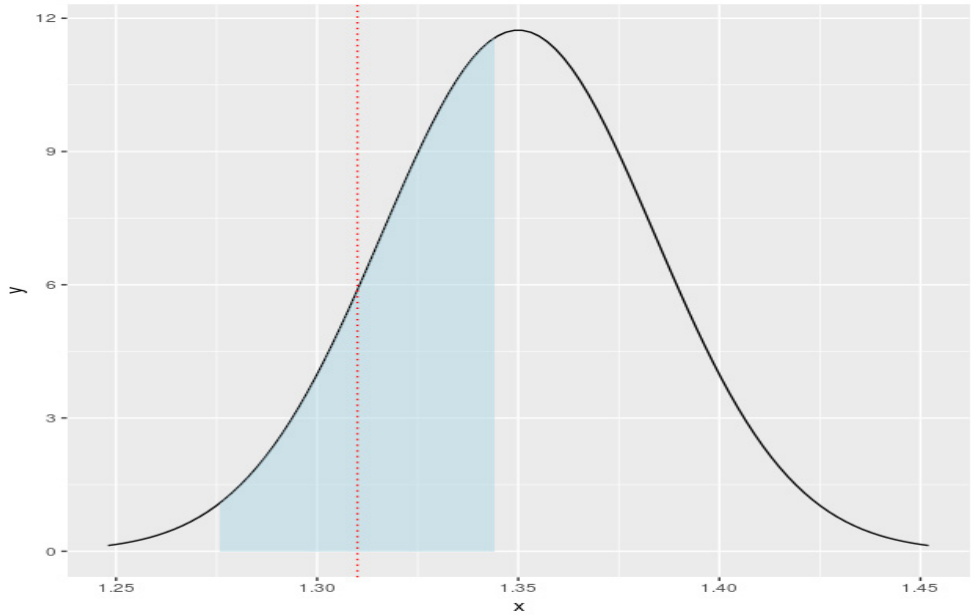


Figure S6. Probcoverage (0.36) of the outcoming node 20-. This score corresponds to the area under the curve between the lower limit (1.32 - 0.034) and upper limit (1.32 + 0.034) for a normal distribution with mean 1.35 and sd of 0.034. Dashed vertical red line indicates the coverage of 20-

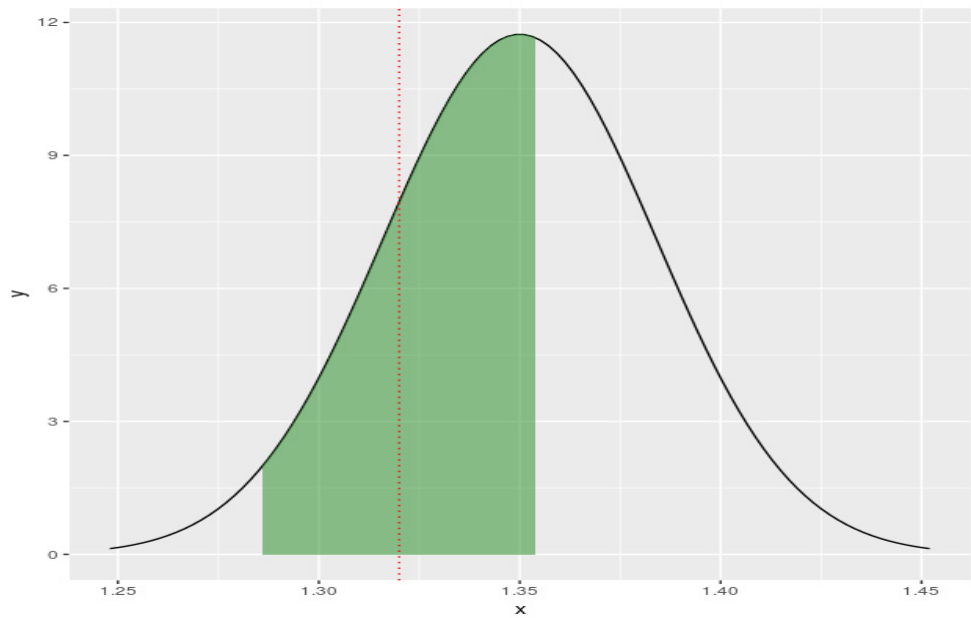


Figure S7. Probcoverage (0.51) of the outcoming node 24+. This score corresponds to the area under the curve between the lower limit (1.32 - 0.034) and upper limit (1.32 + 0.034) for a normal distribution with mean 1.35 and sd of 0.034. Dashed vertical red line indicates the coverage of 24+.

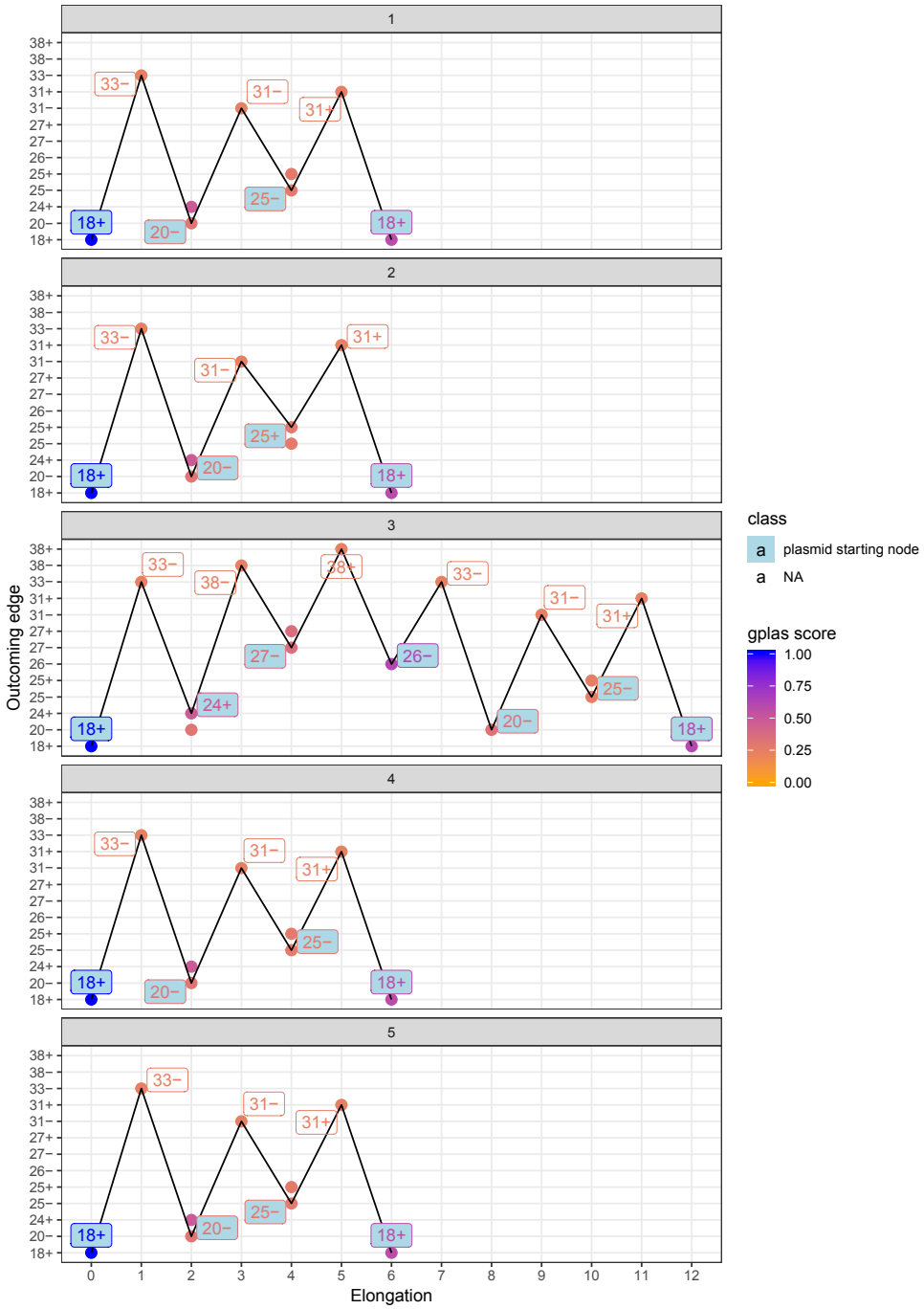


Figure S8. Space of solutions (n =5) starting from the plasmid starting node 18+. We observe the presence of the same solution '18+,33-,20-,31-,25-,31+,18+' in 4 panels and the solution '18+,33-,24+,38-,27-,38+,26-,33-,20-,31-,25-,21+,18+' in the third panel. Nodes in the solutions corresponding to plasmid unitigs are filled with light blue.

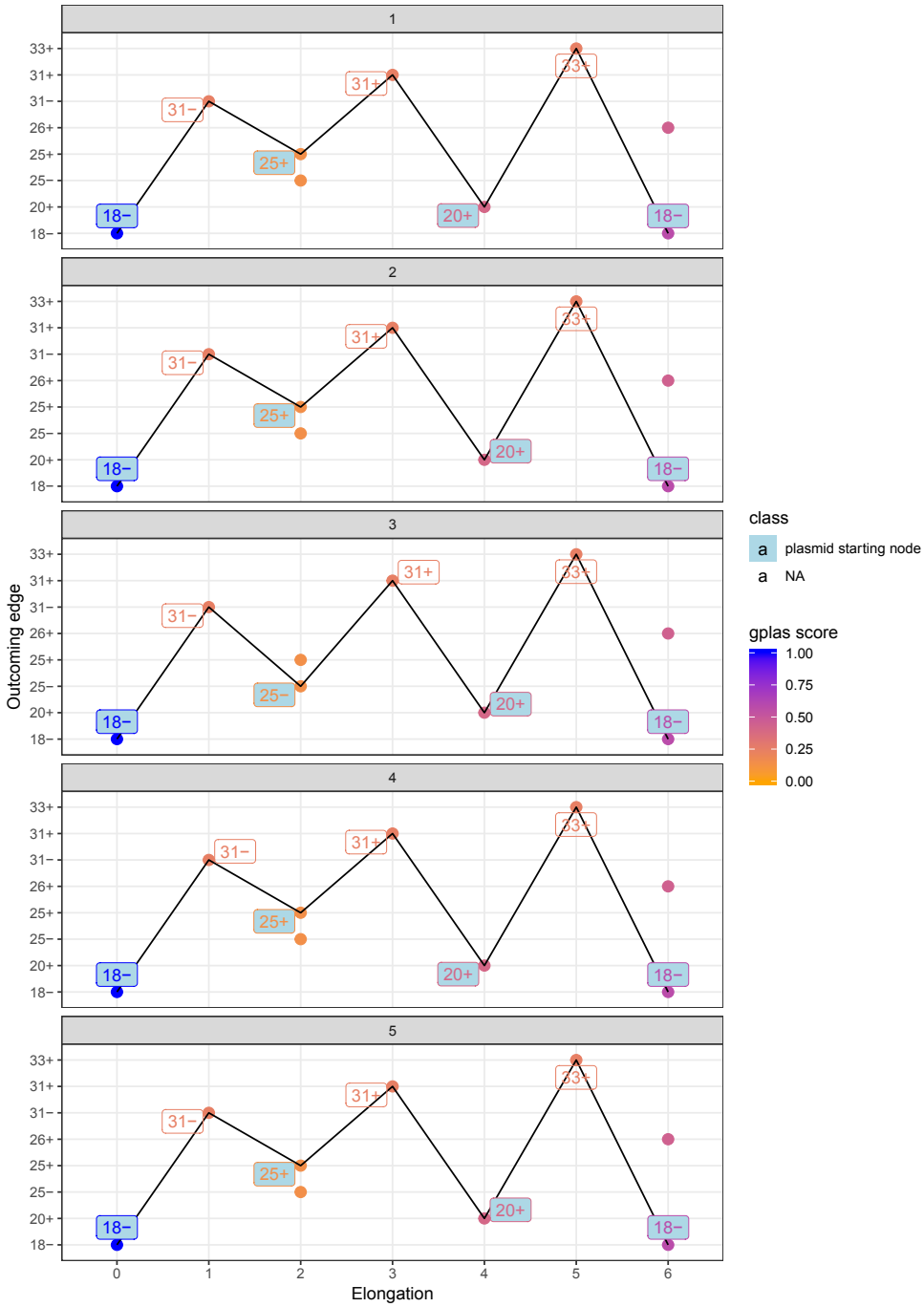


Figure S9. Space of solutions (n =5) starting from the plasmid starting node 18-. We observe the presence of the same solution in all the panels: 18-,31-,25+,31+,20+,33+,18-. Nodes in the solutions corresponding to plasmid unitigs are filled with light blue.

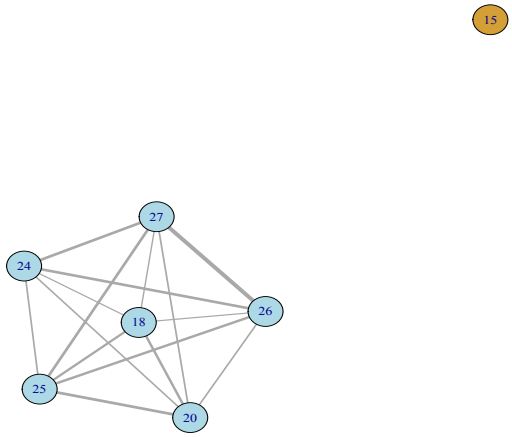


Figure S10. Plasmidome network from the isolate *K. pneumoniae* KSB1_7G. The network corresponds to an undirected graph in which nodes (circles) are plasmid unitigs and edges (lines) with associated weights (line width) are drawn based on the co-existence of the nodes in the solutions found by gplas. Colour of the nodes (circles) correspond to the bin assignment given by gplas.

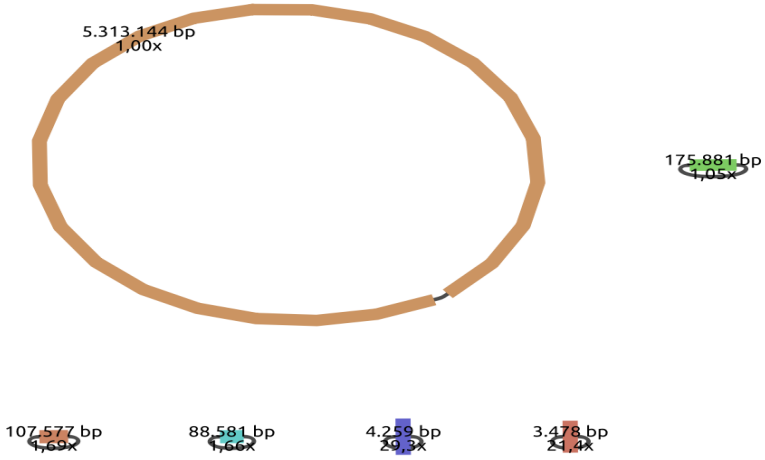


Figure S11. Bandage visualization of the complete genome from the isolate *K. pneumoniae* SAMN10819819 with sequence length (bp) and normalised coverage displayed.

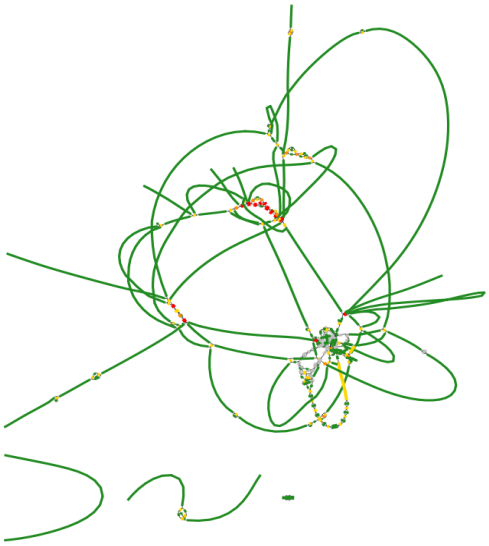


Figure S12. Bandage visualization of the short-read graph from the isolate *K. pneumoniae* SAMN10819819.

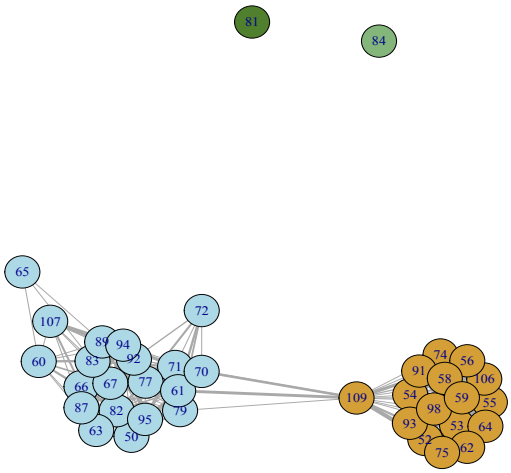


Figure S13. Plasmidome network created by gplas from the isolate *K. pneumoniae* SAMN10819819. Plasmid unitigs are represented by nodes (circles) and edges (lines) with associated weights (line width) represent the connections found by gplas. Colour of the nodes (circles) correspond to the bin assignment given by gplas.

Supplementary References

1. Antipov, D. et al. (2016). plasmidSPAdes : Assembling plasmids from whole genome sequencing data. Bioinformatics, 32, 3380–3387.

2. Arredondo-Alonso, S. et al. (2019). Genomes of a major nosocomial pathogen enterococcus faecium are shaped by adaptive evolution of the chromosome and plasmidome. bioRxiv.

3. Blondel, V. D. et al. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment , 2008(10), P10008.

4. Csardi, G. et al. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5), 1–9.

5. De Maio, N. et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microbial Genomics, 5(9).

6. Decano, A. G. et al. (2019). Complete assembly of *Escherichia coli* sequence type 131 genomes using long reads demonstrates antibiotic resistance gene variation within diverse plasmid and chromosomal contexts. mSphere, 4(3).

7. Gurevich, A. et al. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075.

8. Krueger, F. (2012). Trim galore: a wrapper tool around cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (reduced representation Bisufite-Seq).

9. Müller, R. and Chauve, C. (2019). HyAsP, a greedy tool for plasmids identification. Bioinformatics, 35(21), 4436–4439.

10. Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical Review E , 74(3).

11. Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. Computer and Information Sciences - ISCIS 2005. 3733.

12. Robertson, J. and Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom, 4(8).

13. Wick, R. R. et al. (2015). Bandage: Interactive visualization of *de novo* genome assemblies. Bioinformatics,31(20), 3350–3352.

14. Wick, R. R. et al. (2017a). Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics, 3(10).

15. Wick, R. R. et al. (2017b). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol., 13(6), e1005595.

6

Mode and dynamics of *vanA*-type vancomycin-resistance dissemination in Dutch hospitals

Sergio Arredondo-Alonso, Janetta Top, Jukka Corander, Rob J.L. Willems, Anita C. Schürch

Manuscript in preparation

Abstract

Enterococcus faecium is a commensal of the gastrointestinal tract of animals and humans but also a causative agent of hospital acquired infections. Resistance against glycopeptides and especially to vancomycin, a first-line antibiotic treatment to treat infections with multidrug resistant Gram positive pathogens, has motivated the inclusion of *E. faecium* in the WHO global priority list. Vancomycin resistance can be conferred by the *vanA* gene cluster on the transposon Tn*1546*, which is frequently present in plasmids. Here, we studied the dissemination of plasmids carrying the *vanA* gene cluster in 309 *E. faecium* isolates from 32 Dutch hospitals between 2012 and 2015. Plasmid prediction was conducted using gplas, a graph-based approach to predict distinct plasmid sequences using short-read WGS graphs. Subsequently, a network analysis based on shared k-mer content was undertaken to cluster genetically similar types of plasmids. Based on completed *vanA* plasmid sequences (n = 26), we differentiated six *vanA* plasmid types (A-F) (84% alignment coverage, 99% alignment identity) present in *E. faecium* isolates from different clonal backgrounds, isolation years, hospitals, and also detected in other European countries. The integration of these plasmid types into the network of predicted Dutch *vanA* plasmid sequences (n = 309) unravelled the presence of several predicted plasmid types (1-8). This allowed to compare clonal background, plasmid type and Tn*1546* variant of isolates within Dutch regions. Overall, clonal dissemination contributed most (~45%) to the spread of vancomycin-resistance. However, we also identified potential transposon-mediated (e.g. Overijssel 2012-2013) and plasmid-mediated (Zuid-Holland 2014-2015) dissemination. Here, we provide a comprehensive picture of the interplay between nested genomic elements to explain the dynamics of plasmid-mediated *vanA* resistance dissemination occurring in the Netherlands between 2012-2015.

Introduction

Enterococcus faecium is commonly inhabiting the gut of animals and humans but has also emerged as a nosocomial pathogen causing a sizable fraction of health-care associated infections, specifically device-associated infections like central line associated bloodstream infections and surgical site infections (1, 2). The intrinsic and acquired multidrug resistance against fluoroquinolones, aminoglycosides and more importantly against glycopeptides motivated the inclusion of *E. faecium* in the WHO global priority list (3). The number of strains resistant against vancomycin, a first-line glycopeptide antibiotic to treat infections with multidrug resistant Gram-positive pathogens, dramatically increased first in the US in the 1990s, followed by other parts of the world (4). Resistance against vancomycin can be acquired through eight different gene clusters (*vanA*, *vanB*, *vanD*, *vanE*, *vanG*, *vanL*, *vanM*, and *vanN*) (5, 6) of which *vanA* and *vanB*, associated to transposon sequences Tn*1546* and Tn*1549* respectively, are the predominant vancomycin-resistance gene clusters (7).

Clonal spread of vancomycin-resistant *Enterococcus faecium* (VRE) has been extensively described using a plethora of molecular typing schemes. They range from fingerprint-based methods like pulsed-field gel electrophoresis (8), to PCR-based methods such as multiple-locus variable number tandem repeat analysis (9), multilocus sequence typing (10)and whole genome sequencing (11). However, due to the fact that vancomycin resistance genes are encoded by mobile genetic elements, vancomycin resistance can also be transferred horizontally. In fact, mobilization of the *vanA* gene cluster via insertion in different plasmid backbones has already been reported (12, 13). To identify the dissemination of *vanA* plasmids within hospital settings, whole-genome sequencing (WGS) based on short-read technologies has been recently applied to collections of hundred hospitalized patient isolates in Denmark and Australia (14, 15). These studies undertook a reference-based approach to map short-reads against complete plasmids from a selection of isolates. However, this approach can overestimate the presence of a reference plasmid by neglecting the mosaicism observed in these types of sequences as previously observed for *Enterobacteriaceae* isolates (16) and *Enterococcus* populations (17).

In this study, we first analyzed previously completed *vanA* plasmid sequences (n = 26) (18) and observed six distinct plasmid types (A-F) that were present in different clonal complexes and isolates from distinct European countries. Secondly, we predicted *vanA* plasmid sequences in our collection consisting of short-read WGS (n = 309) using a *de novo* approach (19). This reference-free approach predicts plasmid boundaries in *E. faecium* by inspecting plasmid-like walks in the *de-Bruijn* graph produced by short-read assemblers (19). Using a network approach, we observed that plasmids from the included VRE isolates clustered in eight different predicted *vanA* plasmid types (1-8). Furthermore, these

predicted *vanA* plasmid types (n = 8) were co-occurring in the network with four of the plasmid types defined using complete genome sequences. Furthermore, core genome (hierBAPS SC, PopPUNK core-genome tree), accessory genome (PopPUNK clustering) and *vanA* transposon relatedness (Tn*1546* variant scheme) of *vanA*-VRE were related to the eight predicted *vanA* plasmid types. This allowed to fully reconstruct the nested genomic elements, consisting of clonal background (hierBAPS SC), *vanA* plasmid type (*de novo* prediction and network assignment) and Tn*1546* variants in which the *vanA* resistance gene cluster is present. We observed and quantified that clonal dissemination, defined by vertical inheritance of the same SC, *vanA* plasmid type and Tn*1546* variant, was the most frequent scenario of vancomycin-resistance dissemination occurring in the Netherlands between 2012-2015. However, we also detected transposon-mediated and plasmid-mediated dissemination of the *vanA* resistance gene cluster.

Results

The population structure of VRE from Dutch hospitals

We used short-read WGS data of 309 VRE carrying the *vanA*-type cluster from hospitalized patients from 32 Dutch hospitals, isolated between March 2012 and November 2015 (18). The clonality of these VRE samples was determined using hierBAPS (20), after filtering for recombination events as previously described (18). HierBAPS defined 18 different sequence clusters (SCs) of which SC13 (n = 102, 33%), SC17 (n = 52, 16.8%) and SC18 (n = 42, 13.6%) represented the most predominant clones present in the dataset (Figure 1A). The distribution of these SC across time and geographical position showed that SC13 was widespread in Dutch hospitals for the entire collection period (2012-2015) (Figure 1B) compared to SC17 which was observed in distinct regions (Amsterdam, Lelystad, Zwolle) from August-September 2012 (Figure 1B). SC18 was detected around 2014 in several Dutch regions (Figure 1A). To facilitate the visualization of genomic and metadata information, we provide the following Microreact project https://microreact.org/project/mfEYIjcuu.

To obtain a higher genomic resolution, we used PopPUNK to define genomic clusters by calculating core and accessory distances (21). In general, we observed a high concordance between PopPUNK clusters and hierBAPS-based SC, where most PopPUNK clusters consisted of one predominant SC as exemplified by the linked PopPUNK cluster 2 and hierBAPS SC17 (n = 52, 98.1%) or PopPUNK cluster 3 and hierBAPS SC1 (n = 17, 100%) (Suppl. Table S1). However, we encountered two cases in which a single PopPUNK cluster was composed by several hierBAPS SCs : i) PopPUNK cluster 1 consisting of hierBAPS SC13 (n = 98, 89.1%) and SC10 (n = 10, 8.2%) and ii) PopPUNK cluster 6 consisting of hierBAPS SC29 (n = 5, 41.7%), SC30 (n = 4, 33.3%) and SC32 (n = 3, 25%) (Suppl. Table S1). This suggested that isolates belonging to these clusters have a similar accessory genome while their core

genomes were distinct as they belonged to differentSCs.

Defining a network of complete *vanA* plasmid sequences

To establish a partitioning scheme similar to hierBAPS SCs or PopPUNK cluster assignments but uniquely based on similarity between plasmids carrying the *vanA*-type gene cluster, we retrieved 26 complete *vanA* plasmids from which metadata information regarding isolation, country and source was known (Table 1). These complete sequences were pairwise compared using Mash (k = 21, s = 1,000) and integrated into a network. This network consisted of nodes which corresponded to complete *vanA* plasmid sequences.

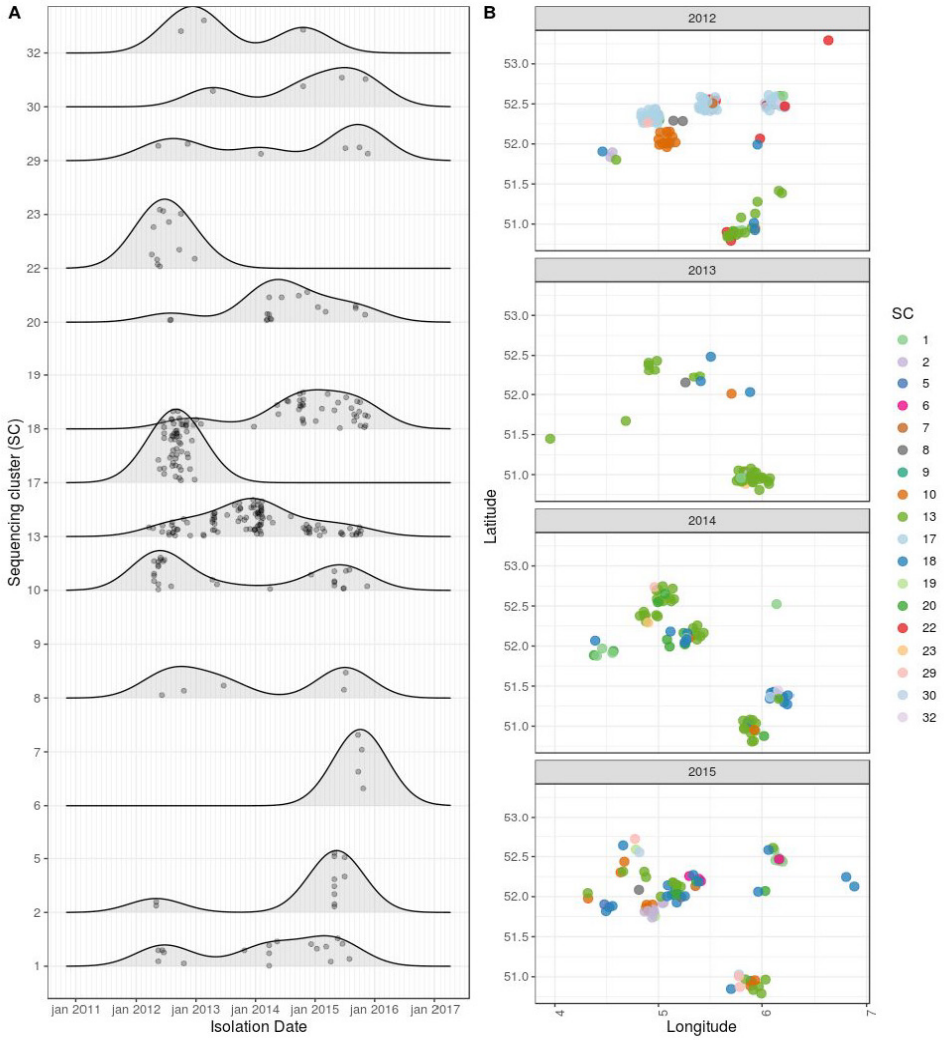


Figure 1. A) Distribution of VREfm sequencing clusters (SC) based on their isolation date . B) Geographical position of the VREFm, longitude (x-axis) and latitude (y-axis) were fixed based on the coordinates from the Netherlands.

To determine a similarity cutoff to draw an edge between two nodes, the k-mer distance distribution was analyzed (Suppl. Figure S1). A Mash distance of 0.025 was an optimal cutoff to draw an edge in the network and to consider two sequences as similarly related. This resulted in a network with 24 nodes (complete plasmid sequences) consisting of 6 independent components. The components, represented by independent subgraphs in the network, signify *vanA* plasmid types (A-F) with a similar content and structure (Figure 2). Two *vanA* plasmid sequences (from isolates E8202, E8172) were not present in this network (Figure 2) because they represented singletons without connections to other plasmid sequences.

We inferred that *vanA* plasmid types (A-F) were shared between isolates from: i) different hierBAPS SC types, observed in 83% of the *vanA* types; ii) different countries, observed in 50% of the *vanA* plasmid types and iii) different isolation years, observed in all *vanA* plasmid types (100%). Furthermore, all these three characteristics were found to be true for plasmid types B, D and F (Table1).

To better understand the modularity and similarity of these plasmid types, we performed pairwise comparisons using pyani (22). This allowed to retrieve the coverage and identity values of the aligned regions between two complete plasmid sequences. We observed that the average coverage between plasmid alignments belonging to the same complete plasmid type was 84% compared to a coverage of 35% when comparing alignments from different plasmid types. This indicated that the edges present in the network of plasmid types and defined by a minimum k-mer distance of 0.025 represented an average alignment coverage of 84%. The coverage between plasmids belonging to different plasmid types (35%) indicated the expected shared plasmid fraction between non-related *vanA* plasmids due to the presence of common elements, i.e. Tn*1546*. The heatmap and single-linkage clustering of the alignment coverage reported by hyasp suggested the same plasmid types as inferred in our network approach (Figure 3). The coverage values obtained for each plasmid type are reported in Table 1. We did not observe differences in the average identity values between the aligned regions within (98.5%) and between (99.7%) plasmid types (Suppl. Figure S2).

To illustrate the genome synteny between plasmids of a single plasmid type, we considered the complete plasmid type B because of the large number of isolates (n = 7) and high-diversity of hierBAPS SC types associated. The alignment in Figure 4 showed the modules present in all plasmids (core modules) belonging to *vanA* plasmid type B, and thus omitting further regions of homology between pairs of plasmids (non-core modules).

Predicting the *vanA* plasmid network of VRE

To reconstruct the *vanA* plasmid sequences present in the *vanA*-VRE from hospitalized patients from 32 Dutch hospitals (n = 309), we predicted and binned plasmid sequenc-

es with gplas (19), by using a combination of machine-learning and a graph-based approach, in 303 (98.1%) VRE isolates. The presence of a fragmented graph consisting of thousands of contigs prohibited the prediction of plasmid boundaries in 5 VRE isolates (1.9%). This approach resulted in plasmids bins with a collection of contigs that belong to distinct plasmid sequences. We then focused on bins bearing the *vanA* gene cluster. Apart from the contigs encoding for the *vanA* gene cluster, these plasmid bins contained other contigs which were connected through walks in the assembly graph by gplas (19). Secondly, we used Mash (k = 21, s = 1,000) to compare the *vanA* plasmid bins present in

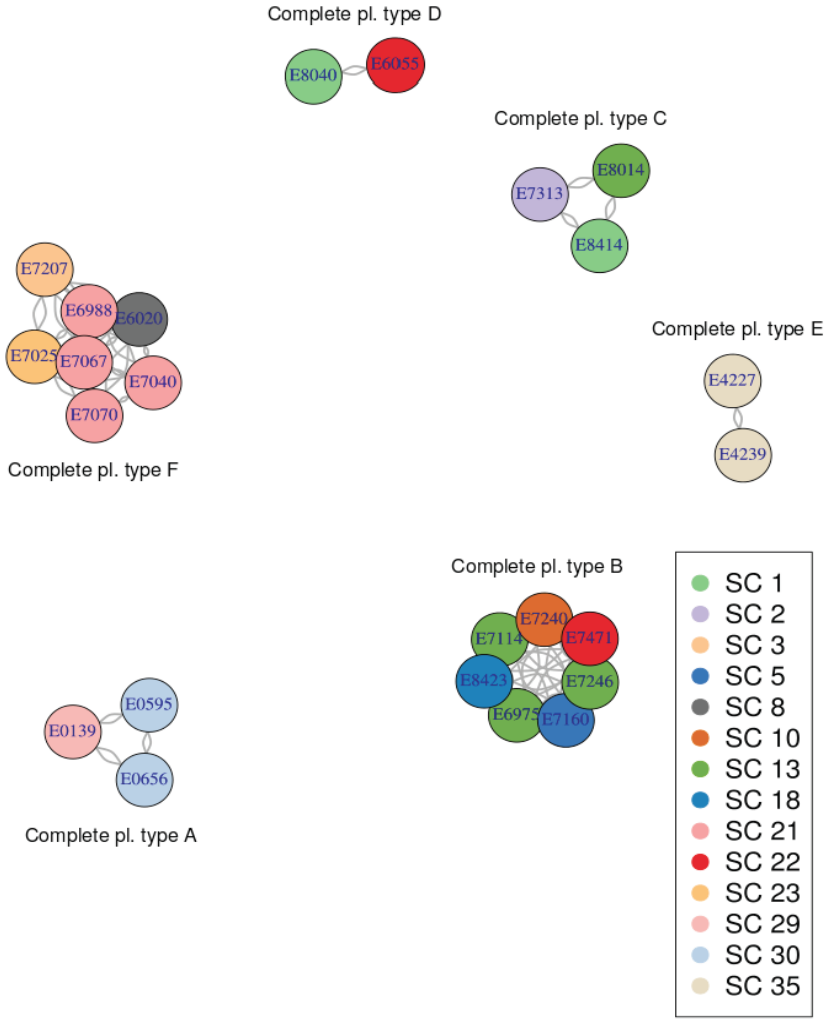


Figure 2. Network of *vanA* complete plasmid sequences, nodes in the graph correspond to isolates and edges to connections between similar (Mash distance = 0.025, k =21, s =1,000) plasmid sequences. The independent components were designated as plasmid types (n = 6) and nodes were coloured according to SC.

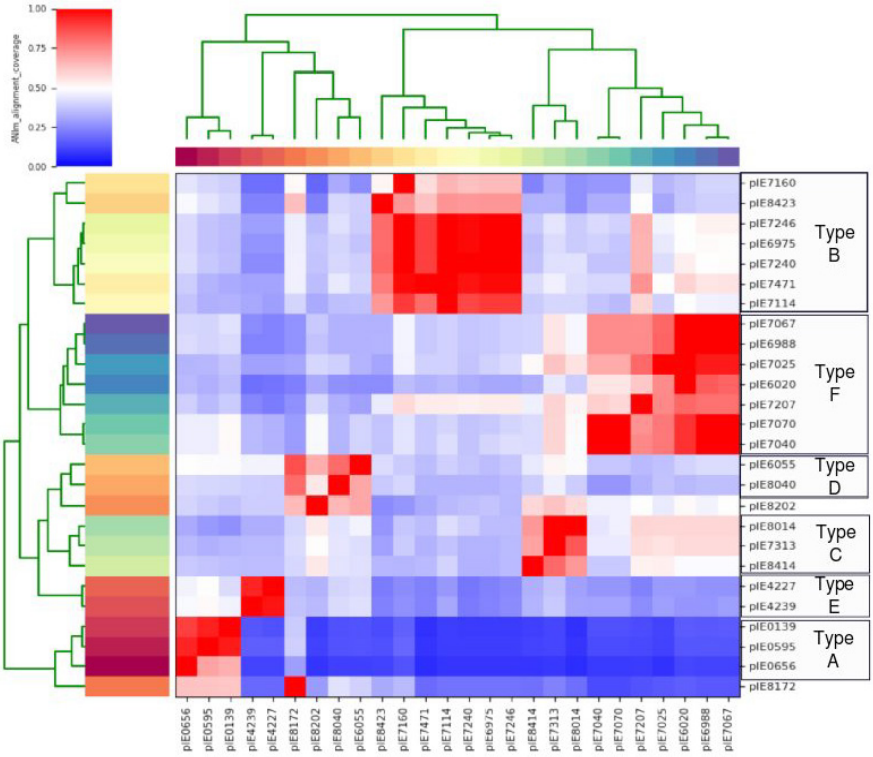


Figure 3. Heatmap and single-linkage clustering of the pyani pairwise alignment coverage obtained from our set of *vanA* complete plasmid sequences (n = 26). The rectangles present on the right side indicate the grouping of the complete plasmid sequences into the previously defined plasmid types (A-F).

each isolate (n = 303). The k-mer distances between these bins were also integrated into a network approach following the same procedure as previously explained for the complete *vanA* plasmids.

The distribution of k-mer distances between bins (Suppl. Figure S3) followed the same distribution as observed with the *vanA* complete plasmids (Suppl. Figure S1). This showed that the k-mer content of the predicted bins resembled the *vanA* complete plasmid sequences and supported the prediction given by gplas. We then applied the same threshold (Mash distance) of 0.025 to create edges between two nodes. This resulted in a network with 270 nodes and 16 components (subgraphs with > 1 node) (Figure 5). Component 3 (236 nodes) was partitioned into 3 different graph bins based on its modularity value (0.42) (Figure 5, central component). In our next analysis, we focused on components/graph bins with > 10 isolates representing eight different predicted *vanA* plasmid types (1-8) (Table3).

To discern the association between strain and *vanA* plasmid relatedness we combined the PopPUNK core-genome inferred neighbour-joining (nj) tree in Figure 6 together with: i) hierBAPS SC, ii) PopPUNK clusters and iii) predicted *vanA* plasmid types. Some closely related isolates in the core-genome tree had distinct predicted *vanA* plasmid types as exemplified by SC13 or SC17 containing plasmid types 2, 3, 5 and 4, 5 respectively (Figure 6). The observation of divergence in the plasmid content within the same SC indicates that particular VRE strains acquired different *vanA* plasmids (Figure 6).

The predicted *vanA* plasmid types (n = 8) were further analysed by the characterization of Tn*1546* variants (Suppl. Figure S4). We considered single-nucleotide polymorphisms (SNPs) and large deletion events (e.g. orf1 and orf2) to define 14 Tn*1546* variants in our collection (Table 2). In the next section, we describe whether isolates from the same predicted *vanA* plasmid types harboured identical or similar Tn*1546* variants. On average, predicted *vanA* plasmid types had the same predominant Tn*1546* variant in 86.2% of the cases but this differed depending on each predicted *vanA* plasmid type. The scheme used to characterize the Tn*1546* variants is described in Material and Methods and full results are given in Suppl. Table S2.

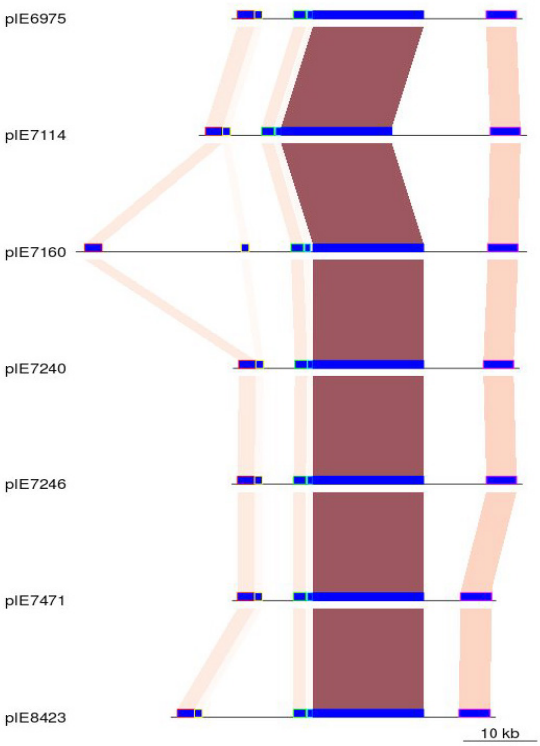


Figure 4. Visualization of the multiple genome alignment of complete sequences (n = 7) belonging to the plasmid type B.

Description of the predicted *vanA* plasmid types

To facilitate the exploration of the data presented in this section which included: i) *vanA* plasmid types, ii) hierBAPS SC, iii) PopPUNK cluster assignments, iv) geographical position and iv) isolation time, we combined and integrated all information in the Microreact project https://microreact.org/project/mfEYIjcuu. This information is also available in Suppl. Table S1.

The predicted plasmid type 1 was formed entirely by samples (n = 11) belonging to SC10 (Figure 4). The majority of the isolates were from Utrecht and isolated in April-May 2012 (n = 9). However, we also found two isolates from May 2013 (E7837) and March 2014 (E8046)

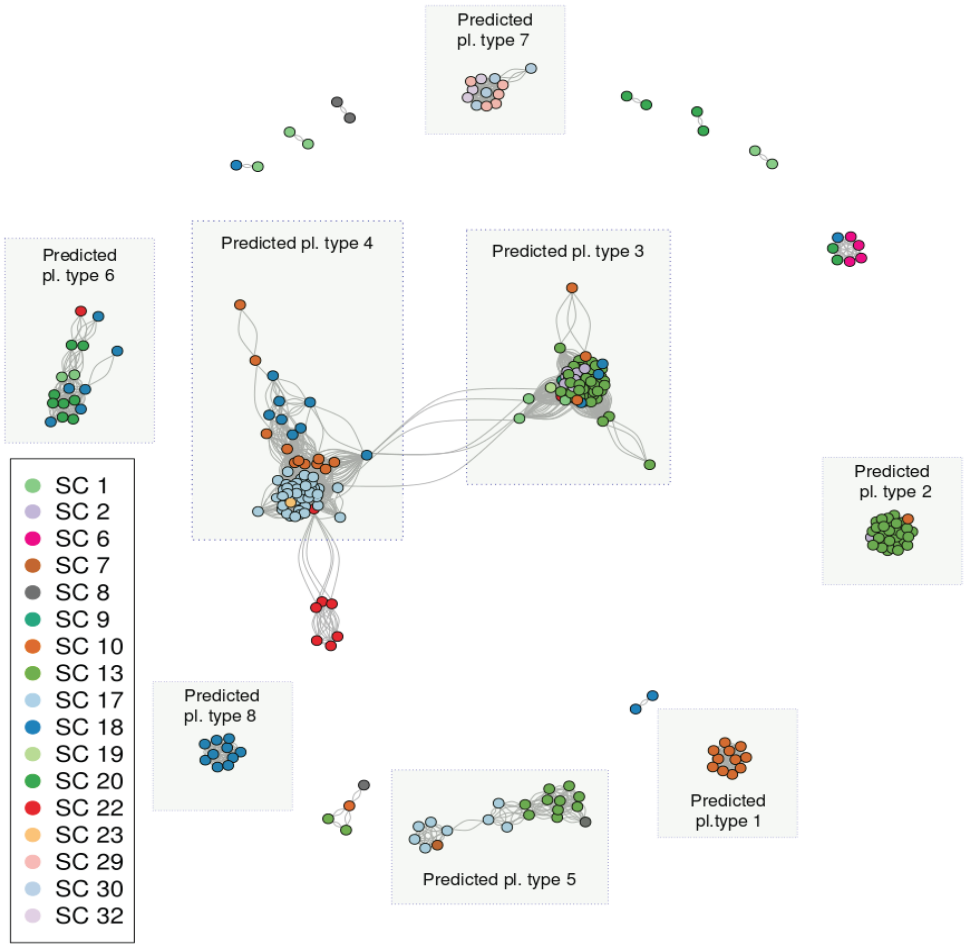


Figure 5. Network of predicted *vanA* plasmid sequences. Nodes in the graph correspond to *vanA* plasmid bins predicted by gplas and edges to connections between similar (Mash distance = 0.025, k = 21, s = 1,000) bins. The components highlighted with a rectangle were considered as *vanA* predicted plasmid types (n = 8). Nodes in the network are coloured according to SC.

which were present in nearby Dutch cities (Ede and Amersfoort). All isolates (n = 11, 100%) shared the Tn*1546* variant 46 corresponding to SNP positions T7658C and G8234T (Suppl. Figure S4). This predicted plasmid type was vertically transmitted as the same *vanA* plasmid sequence was present in SC10 isolates from 2012 and circulated during at least two consecutive years.

The predicted *vanA* plasmid type 2 was mainly formed by isolates belonging to SC13 (Table 3) but geographical widespread in the Netherlands between 2012 and 2015. In this predicted plasmid type, the Tn*1546* variant MNI dominated (n = 23, 92%) corresponding to deletions in *orf1*, *orf2* and intergenic regions (Suppl. Figure S4). The final coordinates of the deletion present in *orf2* varied between Tn coordinates 1-3417 (n = 10, 43.5%) and 1-3676 (n = 10, 43.5%) whereas the deletion observed in the intergenic regions mainly corresponded to the coordinates 8650-8827 (n = 22, 96%). This highlighted that the exact deletion size between the defined Tn*1546* variants (n = 14) may differ.

The same Tn*1546* MNI variant was also found in the predicted *vanA* plasmid type 3 (n = 64, 90.1%) (Suppl. Figure S4). The deletion in orf2 differed in coordinates between 1-3417 (n = 34, 53.1%) and 1-3676 (n = 21, 32.8%), which is different from the MNI variant previously

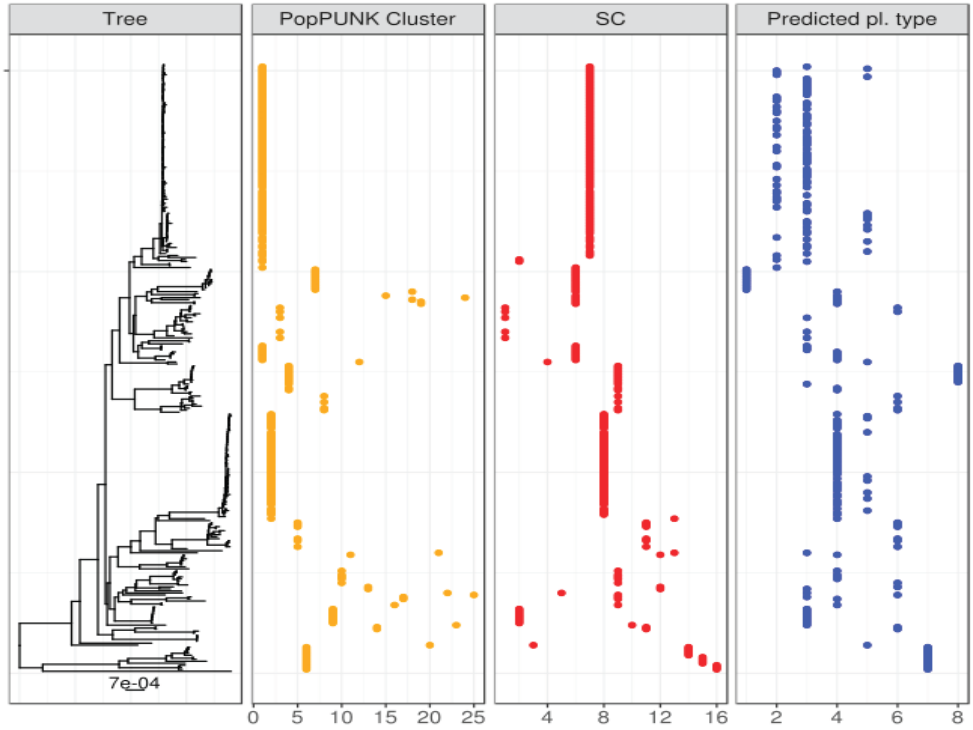


Figure 6. PopPUNK neighbour-joining tree, based on the core genome, in combination with PopPUNK cluster, hierBAPS Sequencing Cluster (SC) and *vanA* predicted plasmid type assignments.

observed in *vanA* plasmid type 2. Isolates from predicted *vanA* plasmid type 3 belonged to SC13, SC2 and SC1 (Table 3) indicating horizontal transmission of the *vanA* plasmid between these hierBAPS SCs. This predicted plasmid type 3 was widely observed in isolates in the Netherlands during the entire collection period (2012-2015).

The isolates carrying *vanA* plasmid type 4 mainly belonged to three different SCs (10,17,18) (Table 3) and were recovered from different Dutch cities (Amsterdam, Lelystad, Zwolle, Utrecht, Nieuwegein and Rotterdam) in 2012. The predominant Tn*1546* variants were 46MN (n = 45, 72.6%) and 468MN (n =6, 9.7%) corresponding to deletions in orf1 and orf2 (1-3343 in 88.2% of the cases), two SNP positions (T7658C and G8234T) and the additional SNP (C9692T) present in the 468MN variant (Suppl. Figure S4).

Isolates containing *vanA* plasmid type 5 belonged to two predominant SCs, SC13 and SC17. (Table 3). However, the network topology of this plasmid type indicated that plasmids contained in both SCs were not highly interconnected and thus can indicate that these belong to two different plasmid subtypes (Figure 5). A close inspection of the Tn*1546* variants revealed two variants MNI (n = 10, 52.6%) and 46MN (n = 9, 47.4%) (Suppl. Figure S4) supporting the split of this *vanA* plasmid type into two subtypes. The predicted plasmid type 5 was mainly represented by isolates from 2012 (80%) and present in distinct Dutch provinces (Limburg, Utrecht, Flevoland and Overijssel).

Isolates with the *vanA* plasmid type 6 belonged to four different SCs, SC20, SC18, SC1, and SC22 and were located widespread around the Netherlands, and were retrieved between 2012-2015. The dominant Tn*1546* variant was 5II (n = 14, 82.4%) corresponding to the G7747T SNP and two deletions in intergenic regions (5896-5931, 10706-10851). In addition, the variants 25II (n= 1, 5.9%), 6M (n = 1, 5.9%) and the original Tn*1546* sequence (n = 1, 5.9%) were observed in this plasmid type (Suppl. Figure S4).

The *vanA* plasmid type 7 was found in isolates belonging to three distinct SCs, SC29, SC30, and SC32 (Table 3). They all shared the same Tn*1546* 6M variant corresponding to the SNP position G8234T and deletion of orf1 (1-119) (Suppl. Figure S4). Plasmid type 7 was represented by isolates from 2012 to 2015 but only from a single Dutch province (North Holland).

The *vanA* plasmid type 8 provides a clear example of clonal dissemination related to a hospital outbreak (Viecuri MC Venlo, Limburg province). All isolates (n = 10) carrying this *vanA* plasmid type belonged to SC18 and were isolated in Venlo (Limburg province) during September-October 2014. In addition, these isolates all shared the same Tn*1546* structure (n = 10, 100%), original variant (Suppl. Figure S4).

Elucidating the content of predicted *vanA* plasmid types by the integration of the complete plasmids

The genetic characterization involving gene content and gene synteny analysis of the eight predicted *vanA* plasmid types described above is challenging as these plasmid types represent bins of short-read contigs from which the plasmid structure and contig order was unknown. To elucidate gene content and synteny, we integrated the six plasmid types (A-F) derived from complete plasmid sequences (Table 1, Figure 2) into the network of the eight predicted *vanA* plasmid types (1-8) (Figure 7). We calculated Mash distances (k = 21, s = 1,000) between the complete *vanA* plasmid sequences (n = 26) and the predicted *vanA* plasmid bins (n = 303). The presence of edges connecting complete plasmids and predicted *vanA* plasmid bins revealed that the predictions had a similar k-mer content and thus further validated the predicted eight *vanA* plasmid types (Figure 7). Furthermore, this approach also allowed to elucidate the content and structure of the plasmids within the predicted *vanA* plasmid types.

The plasmid type B was exclusively embedded in, and only connected to, predicted *vanA* plasmid type 4 (Figure 7). This revealed that the predicted plasmid type 4 could be characterized as a multireplicon RepA_N & Rep_3-like plasmids with a length ranging from 35.7 kbp to 61.1 kbp (mean = 43.3 kbp, median = 39.4 kbp) (Figure 7). The isolates carrying this plasmid type (B-4) were from: i) four different European countries (Greece, Latvia, Slovenia, the Netherlands), ii) belonged to eight different hierBAPS SCs (1,2,4,10,17,18,22,23) and iii) from four distinct isolation years (2009, 2010, 2012, 2014, 2015).

The plasmid type C was uniquely embedded in the predicted *vanA* plasmid type 3 (Figure 7). Type C-3 plasmids are multireplicon Inc18 & Rep_trans plasmids ranging in size from 42.6 kbp to 49.9 kbp (mean = 45.1 kbp, median = 42.7 kbp). This plasmid type (C-3) was carried by Dutch isolates belonging to nine different SCs (1,2,9,10,13,18,19,22,23) which suggested a high-level mobility of this plasmid type between different *E. faecium* clonal complexes. Here, we also spotted a limitation of the predicted *vanA* plasmid network. The predicted plasmid types of the isolates E7313, E8014, E8414, with associated complete plasmid sequences, were linked to two differentially predicted *vanA* plasmid clusters (2 and 3). However, these complete plasmid sequences belonged to a single plasmid type C. Furthermore, both predicted *vanA* types (2,3) shared the same Tn*1546* variant (MNI). These observations suggested that the *vanA* plasmid types 2 and 3 could be merged based on plasmid configuration, SC type distribution, and Tn*1546* variants. Based on these observations, in the next section we reassigned plasmid type 2 as plasmid type 3.

The plasmid type D linked to the predicted *vanA* plasmid type 6 (Figure 7). Type D-6 plasmids are Inc18 plasmids with a size ranging from 41.1 kbp to 47.9 kbp (mean = 44.5 kbp, median = 44.5 kbp). The isolates carrying this plasmid type belonged to different coun-

tries (Portugal, the Netherlands), and six SCs (1,7,8,13,17,22).

Plasmid type A, found in isolates from non-hospitalized persons and pigs from 1996-1999 (Table 2) were uniquely embedded in the predicted *vanA* plasmid type 7 which was found in Dutch hospitalized patients belonging to SC29 and SC30 (Figure 7). Type A-7 plasmids are large RepA_N plasmids with a size ranging from 141.9 kbp to 189.4 kbp (mean = 157.1 kbp, median= 141.9 kbp). In this case, the presence of a similar *vanA* plasmid sequence (coverage ~ 85%) in the predicted network could be explained by vertical transmission of isolates belonging to these SCs.

Finally, we observed that the plasmid type F was weakly connected to predicted plasmid type 4 (Figure 7) whereas the type E formed an independent component (subgraph). This suggests that these *vanA* plasmid types were not present in our Dutch VRE collection.

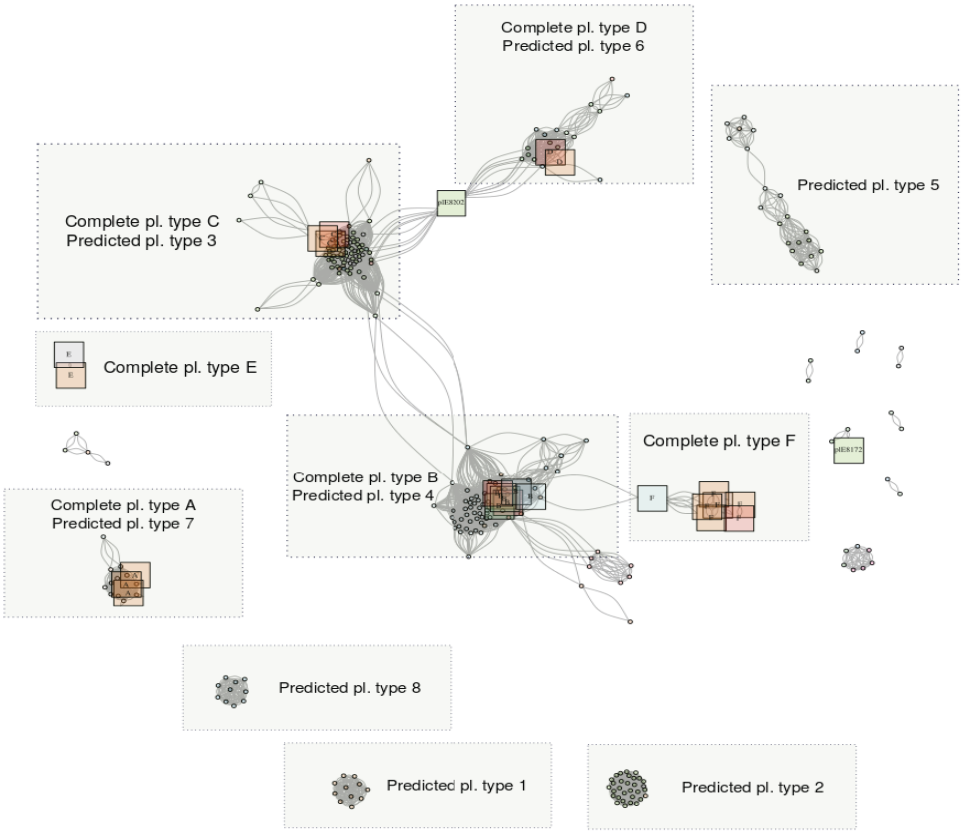


Figure 7. Network of predicted and complete *vanA* plasmid sequences. Nodes corresponding to complete plasmid sequences are highlighted with squared shapes while nodes from predicted plasmid sequences are represented with circles. Edges are connections between nodes with similar *vanA* sequences (Mash distance = 0.025, k = 21, s = 1,000). Rectangles indicate the grouping of nodes based on the previously defined complete (A-F) and predicted plasmid types (1-8).

In contrast, the predicted *vanA* plasmid types (1, 2, 5, 8) did not have any connections to complete *vanA* plasmid types. Most likely, this is due to the fact that these *vanA* plasmid types were not represented by the collection of completed plasmids used in this study. Alternatively, absence of a link with the completed plasmids could also be the result of incorrect *vanA* plasmid type predictions.

Dynamics of *vanA*-type resistance dissemination

The characterization of the different *vanA* plasmid types present in our VRE collection allowed us to include the missing element into the nested genomic complexes (SC, plasmid type, Tn*1546* variant) involved in the dissemination of the *vanA* gene cluster. Next, we aimed to quantify the importance of SC, *vanA* plasmid-type and Tn*1546* mobilisation, in the dissemination of vancomycin-resistance. For this, we first grouped VRE isolates with a potential epidemiological link which corresponded to isolates from the same Dutch region and recovered within a period of 12 months. To estimate the importance of inter-regional spread of vancomycin-resistance, the same approach was taken without taking into account the origin of the isolates. We then performed a pairwise "all vs. all" comparison of hierBAPS SC, predicted *vanA* plasmid type and/or Tn*1546* variants of VRE isolates. Based on this, we defined three scenarios: i) clonal dissemination, ii) plasmid-mediated dissemination, and iii) transposon-mediated dissemination of *vanA*-type vancomycin resistance. We refer to the Methods section for a complete description of the scenarios explored in this section.

We observed that, on average, clonal dissemination was the predominant mode of vancomycin-resistance spread (~45%), followed by Tn*1546* mobilisation/transposition into another plasmid type (~12%) and plasmid-mediated dissemination (~6%). However, the dynamics of vancomycin-resistance spread were clearly distinct between regions (Figure 8).

Clonal dissemination driven by the hierBAPS SC 17, *vanA* plasmid type B-4 and Tn*1546* 46MN variant, had the strongest contribution *vanA*-type vancomycin resistance spread in Flevoland (2012-2013, frequency avg. = 70%) and North-Holland (2012-2013, frequency avg. = 75%) (Figure 8). Interestingly, in North-Holland (2013-2014), clonal-dissemination was still the predominant mode of spread (frequency = 70%) but was driven by a different clone defined by the hierBAPS SC 13, plasmid type C-3, and Tn*1546* MNI variant. In contrast, transposition of Tn*1546* variants MNI and 46MN variant followed by horizontal gene transfer was predominant in Flevoland (2012-2013, freq. avg = 29%), Limburg (2012-2013, freq. avg = 21%) and Overijssel (2012-2013, freq = 47%), respectively.

The highest contribution of plasmid mediated vancomycin-resistance was identified in South Holland (2014-2015, freq. avg = 42%). There, we found plasmid type C-3 together with MNI variant was present in four distinct clonal backgrounds (SCs: 2, 10, 13, 19). This plasmid-mediated outbreak mainly occurred in the Albert Schweitzer hospital between

2014 and 2015 (Suppl. Figure S5). Finally, we also observed complex scenarios of mixtures of genomic units exemplified by Limburg in 2012-2013, in which clonal (freq = 38%), plasmid-mediated (freq = 7%) and Tn*1546* transposition (freq = 21%) all contributed to the spread of vancomycin-resistance.

Finally, we analysed the spread of vancomycin-resistant on a country-wide perspective. Pairwise "all vs. all '' comparisons revealed that in most cases VRE strains, plasmid and Tn*1546* variants were unrelated (~59%), while clonal-dissemination was detected in ~27% of the comparisons, while plasmid spread and transposition of Tn*1546* accountanted for ~7% of the comparisons (Suppl. Figure S6). However, during the period between 2014-2015 plasmid-mediated dissemination increased up to ~29% which could be linked to the South Holland plasmid outbreak (2014-2015) described above (Suppl. Figure S6).

Discussion

In this study, we proposed a novel approach for studying the molecular epidemiology of *vanA*-type vancomycin resistance based on whole genome sequence data. In order to fully reconstruct the mode of *vanA* dissemination, we applied a machine learning and graph-based prediction combined with a network analysis based on the shared plasmid k-mer content which allowed us to distinguish and quantify clonal, plasmid- and transposon-mediated dissemination.

Combining existing short-read WGS with complete *vanA* plasmids allowed us to define and characterize several plasmid types present in the collection. The integration of previously completed *vanA* plasmids was essential to elucidate the genetic content of the predicted *vanA* plasmid types present in our predicted network. These *vanA* plasmid types were defined by a similarity of ~99% identity and ~84% coverage and were present in different clonal backgrounds (SCs), and carried a predominant Tn*1546* variant (86.2% on average) that accumulated additional SNPs or deletions (e.g 46MN and 468MN). The genomic relatedness of strains, using hierBAPS and PopPUNK, plasmid types and Tn*1546* variants calling was combined to sketch a comprehensive picture of the molecular epidemiology of *vanA*-type vancomycin resistance in Dutch hospitals.

In our set of Dutch VRE collected between 2012 and 2015, clonal dissemination was the predominant (~45%) mode of spread of *vanA*-type of vancomycin resistance, followed by transposon-mediated (~12%) and plasmid-mediated (~6%) dissemination. While our data set allowed us to define modes of dissemination, it was not possible to analyse transmission events between patients. Raven et al. (11) showed that transmission routes of VRE can potentially span numerous wards and years and happen between hospitals. To achieve this level of epidemiological resolution, not only WGS data but also patient admission and ward movement data is necessary which was lacking in the Dutch dataset.

The dissemination of vancomycin resistance was previously investigated using WGS in several recent studies (11, 23–26)but seldom with a focus on distinguishing clonal and plasmid outbreaks. One exception is the study by Pinholt et al (14) that used a combination of short-read and long-read sequencing to describe the clonal expansion of VRE in the Capital Region of Denmark between 2012 and 2015. Here, the clonal group ST80 was defined as responsible for the first observed local outbreaks. This clonal group subsequently spread to other hospitals in the same region. The plasmid bearing the *vanA* gene cluster was disseminated to other, non-clonally related vancomycin-susceptible isolates. In our data set covering the same timeframe, ST80 represented by SC18 was also a predominant clonal complex (Figure 1A). However, the absence of data on vancomy-

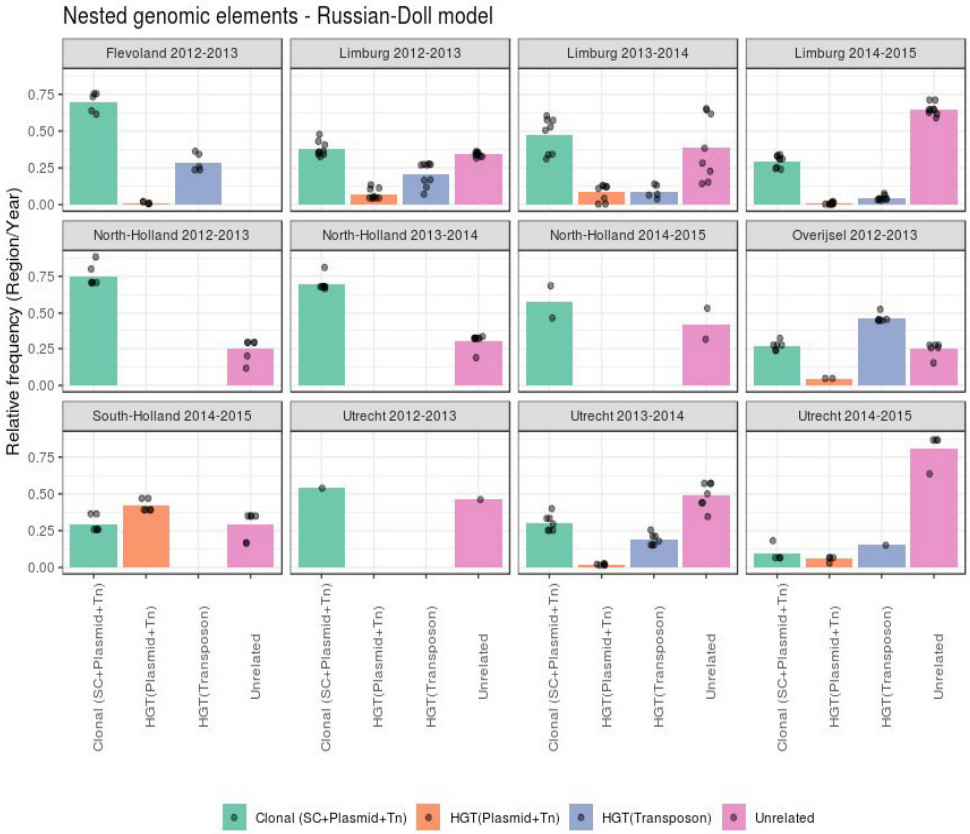


Figure 8. Contribution of nested genomic elements in regional dissemination of vancomycin resistance in the Netherlands. VRE samples isolated from 12 consecutive months and from the same Dutch region were considered. Clonal dissemination (green bar) corresponded to pairs of isolates sharing the same hierBAPS SC, *vanA* plasmid type and Tn*1546* variant. Plasmid-mediated dissemination (orange bar) corresponded to pairs of isolates sharing the same *vanA* plasmid type and Tn*1546* variant but different hierBAPS SC. Transposon-dissemination (blue bar) corresponded to pairs of isolates sharing the same Tn*1546* variant, different *vanA* plasmid type and same or different the hierBAPS SC. Unrelated (purple bar) isolates corresponded to pairs of isolates with a different Tn *1546* variant.

cin-susceptible isolates in our study does not allow us to deduce if the introduction of the plasmid-types in the different clonal complexes occurred before or during the collection period (2012-2015).

The emergence of *vanA*-type resistance was also investigated in Australia during 2015 using a combination of short and long-read sequencing (15). The study showed the presence of several *vanA* plasmid types which were dominant in each ST group with distinct Tn*1546* variants. This unravelled that, in Australia, the emergence of the *vanA*-type resistance most likely occurred by multiple introductions of different clones which suggested that HGT is not solely responsible for the spread of the *vanA* gene cluster, which is in line with our own results.

Both studies (14, 15) however, followed a reference-based approach to deduce the presence of a particular *vanA* plasmid type. This could mask the presence of plasmid-types which are distinct from the selected reference plasmid(s). In our study, prediction of plasmids has been integrated into a network that avoids the arbitrary usage of a reference-plasmid and takes plasmid modularity into account. Bipartite networks were previously postulated to explore the pangenome of bacterial species with a particular emphasis on the accessory genome (27). A network approach also allows to classify plasmids in the absence of a common evolutionary history as it can integrate both horizontal and vertical inheritance, in contrast to phylogenetic trees (28, 29). The classification of plasmids based on k-mer similarity networks has also recently been proposed by Acman et al. (30). Consequently, our network-based analysis could be expanded to include other *vanA* plasmids from plasmid databases such as Plasmid Atlas or PLSDB (31, 32), and could provide a global picture of the dissemination of vancomycin-resistance. Freitas et al., using an experimental approach showed that *vanA* resistance was located not only on Inc18-pIP186 (Europe) but also on pRUM/Axe-Txeplasmids (33) and also concluded that both clonal and plasmid dissemination contributed to the vancomycin resistance spread.

To get full resolution of plasmid architecture, completed genomes are a useful resource to complement the short-read WGS data. Long-read completed plasmids were also used to investigate plasmid-mediated dissemination of carbapenem resistance (34). In a recent study, the dissemination of carbapenem resistance in *Enterobacterales* was observed in eight different species involving more than 100 sequence types. The plasmids bearing the carbapenem resistance showed a high genome plasticity by frequently exchanging modules to other non-carbapenamese resistant plasmids (35). Several antibiotic resistance-related plasmid outbreaks in a clinical context have been described in Enterobacteriaceae (36–40) but often it is not straight-forward to elucidate exact routes of transmission because of the nestedness of genetic elements (16).

A focus on the core-genome can overestimate the number of isolates that are consid-

ered as non-related and thus missing potential epidemiological links. In line with Harris et. al. 2020 (41), we encourage the shift from a traditional core genome view on outbreak investigations to a new perspective that also includes the analyses of transposon-mediated and plasmid-mediated mobilisation of AMR genes to effectively confirm potential epidemiological links and correctly evaluate the effectiveness of infection control policies. We showed that highly similar plasmids can be transferred between different SC which challenges the interpretation of AMR outbreak studies that are solely focused on core-genome analysis. A factor contributing to this clonality perspective is driven by the limitations inherent from short-read WGS from which the assembly of plasmids is difficult and error-prone due to the high number of repeated sequences (42).

In summary, we propose a novel approach to infer the dynamics of *vanA* resistance dissemination. This allowed to obtain a comprehensive picture of the interplay between the nested genomic elements involved in the spread of the *vanA* gene cluster occurring in the Netherlands (2012-2015). Specifically, it revealed that in most Dutch regions, clonal spread, defined as isolates that were indistinguishable with respect to strain, plasmid and transposon type, contributed most to the spread of *vanA* type vancomycin-resistance. However, we also detected episodes of transposon-mediated and plasmid-mediated outbreaks in which the dissemination of the *vanA* gene cluster did not occur vertically.

Material and Methods

Dutch VREfm collection, short-read WGS and genome assembly

The isolates from this collection represent a subset of isolates belonging to a previous study we conducted and that consisted of 1,644 *E. faecium* isolates including hospitalized patients but also samples from other sources (18). *VanA*-type vancomycin-resistance isolates (n=309) from 32 Dutch hospitals recovered between 2012 and 2015 were further analyzed. DNA extraction, and whole-genome sequencing using Illumina NextSeq were conducted as previously described (43). Short-reads were trimmed using Trim Galore (version 0.6.4_dev) using the flag '--paired' and specifying a phred score of 20 with the flag '--quality' (44). We used Unicycler (version 0.4.7) passing the short paired-end trimmed reads from Trim Galore and specifying the normal mode (--mode) (45). Unicycler was used to compute the assembly graph provided in the file 'assembly.gfa' which selects for the k-mer size that optimises the ratio between number of dead-ends and contig size in the graph given by SPAdes (version 3.14.0) (46). In-silico prediction using Abricate (https://github.com/tseemann/abricate, version 0.8), with the ResFinder database (indexed on 16th of July 2018) (47)was conducted to search and select for Dutch isolates bearing the *vanA* resistance gene.

Population structure of VREfm isolates

Recombination events and estimation of sequence clusters using BratNextGen and hierBAPS were performed as previously described (18). PopPUNK (version 2.0.1) was run specifying the flag '--easy-run' with a minimum k-mer size of 13 (flag --min-k) and creating the files required to generate a microreact project (flag --microreact)(21).

Prediction and characterisation of *vanA* plasmids

Gplas (version 0.6.1) was used to predict the plasmids present in the assembly graph of each VREfm isolate (19). Gplas was run using mlplasmids (43)as classifier to predict plasmid sequences (flag '-c'), specifying the species model 'Enterococcus faecium' (flag -s), a modularity threshold of 0.1 to partition the resulting bins (flag -q), and 50 walks per plasmid seed (flag -x). TETtyper (unique version) was used (48) to detect SNPs and deletions present in the Tn*1546* sequences of each *vanA* bin against the reference (--ref) corresponding original transposon structure (Accession M97297) described by Arthur et al. (49), and passing the trimmed reads to perform the analysis with other default parameters. We used the *ad-hoc* scheme described in Table 2 to term the observed Tn*1546* variants.

Analysis of predicted *vanA* bins and complete plasmids

Mash (version 2.2.2) (50) was used to perform k-mer pairwise comparisons of the *vanA* plasmids bins and complete plasmid sequences specifying a k-mer size (-k) of 21 and sketch size (-s) of 1,000. The igraph R package (version 1.2.4) (51) was used to represent in a network the *vanA* plasmid bins or complete plasmid sequences as nodes, and edges as connections between sequences with a minimum Mash distance of 0.025. The function 'cluster_louvain' from the igraph R package was used to split components into further subgraphs based on the reported modularity value.

The starting coordinate position of complete *vanA* plasmids was adjusted using the function fixstart from circlator (version 1.5.5) using a customized database of known plasmid replication initiator sequences (52). Pairwise coverage and identity of aligned regions between complete plasmid sequences was performed using the script 'average_nucleotide_identity.py' from pyani (version 0.2.10) using default parameters (22). The function 'progressiveMauve' from mauve (53) with default parameters was used to produce a multiple genome alignment of *vanA* complete plasmids belonging to the same configuration. The backbone file created by mauve was visualized with the R package genoplotR (version 0.8.9) (54).

Contribution of nested genomic elements in the dissemination of *vanA*-type gene cluster

Pairwise comparisons were computed between VRE isolates sampled within 12 consecutive months and isolated at: i) same hospital, ii) same Dutch region and iii) country-wide.

We observed which genomic elements were shared between pairs of VRE isolates and defined the following scenarios: i) clonal dissemination, characterized by identical SC, *vanA* plasmid type and Tn*1546* structure; ii) HGT plasmid dissemination, characterized by identical *vanA* plasmid type and Tn*1546* structure but distinct SC type, iii) HGT transposon dissemination, characterized by identical Tn*1546* structure but distinct SC and *vanA* plasmid types and iv) no linkage, distinct Tn*1546* structure.

Visualization of genomic elements

The R package ggtree (version 1.14.6) (55) was used to integrate the neighbour-joining tree based on the core-genome given by PopPUNK together with SC assignment, PopPUNK clusters and predicted *vanA* plasmid types. A Microreact project (version 15.0.0) (56) was also created with metadata regarding geographical position and isolation date.

Data availability

The complete code used to generate the results present in this manuscript is provided in a RMarkdown document available through

https://gitlab.com/sirarredondo/vancomycin_dissemination.

Supplementary Tables S1 and S2 are also available in the previous gitlab repository.

The raw paired-end reads of the isolates are available through the European Nucleotide Archive project PRJEB28495. The complete *vanA* plasmid sequences (n = 26) can also be retrieved through the NCBI Bioproject PRJEB28495 and the gitlab project

https://gitlab.com/sirarredondo/vancomycin_dissemination

Table 1. Metadata information of the *vanA* complete plasmid sequences (n = 26) and average pairwise alignment coverage between and within plasmid types.

Isolate	Complete pl. type	Country	Year	Source	SC type	Between coverage [a]	Within coverage [b]
E0139	A	the Netherlands	1996	Non-hospital	29	40%	85%
E0595	A	the Netherlands	1996	Pig	30		
E0656	A	the Netherlands	1999	Non-hospital	30		
E6975	B	Greece	2010	Hospital	13	32%	85%
E7114	B	Latvia	2010	Hospital	13		
E7160	B	Slovenia	2009	Hospital	5		
E7240	B	Greece	2010	Hospital	10		
E7246	B	Greece	2009	Hospital	13		
E7471	B	the Netherlands	2012	Hospital	22		
E8423	B	the Netherlands	2015	Hospital	18		
E7313	C	the Netherlands	2012	Hospital	2	38%	80%
E8014	C	the Netherlands	2014	Hospital	13		
E8414	C	the Netherlands	2014	Hospital	1		
E6055	D	Portugal	2010	Hospital	22	35%	75%
E8040	D	the Netherlands	2014	Hospital	1		
E4227	E	Sweden	2005	Chicken	35	28%	96%
E4239	E	Sweden	2007	Chicken	35		

E6020	F	Latvia	2010	Hosp.	8	38%	82%
E6988	F	Latvia	2010	Hosp.	21		
E7025	F	Latvia	2010	Hosp.	23		
E7040	F	Latvia	2010	Hosp.	21		
E7067	F	Latvia	2010	Hosp.	21		
E7070	F	Latvia	2010	Hosp.	21		
E7207	F	Greece	2008	Hosp.	3		

[a] Between coverage refers to the average coverage resulting from pairwise comparisons of complete plasmid sequences belonging to different plasmid types (A-F).
[b] Within coverage refers to the average coverage resulting from pairwise comparisons of complete plasmid sequences belonging to the same plasmid type.

Table 2. *Ad-hoc* scheme to name the Tn*1546* variants present in the predicted *vanA* plasmid sequences.

Type	Position	Nomenclature
SNP	C806T	1
SNP	G3966T	2
SNP	G4351T	3
SNP	T7658C	4
SNP	G7747T	5
SNP	G8234T	6
SNP	C8833T	7
SNP	C9692T	8
Deletion	orf1	M
Deletion	orf2	N
Deletion	*vanR*	R
Deletion	*vanS*	S
Deletion	*vanH*	H
Deletion	*vanA*	A
Deletion	*vanX*	X
Deletion	*vanY*	Y
Deletion	*vanZ*	Z
Deletion	Intergenic	I

Table 3. Description of the *vanA* predicted plasmid types (> 10 isolates). For each predicted plasmid type, we indicated the graph component, partitioning of the component (see footnote), number of isolates forming the predicted plasmid type, percentage of SCs (relative to the predicted plasmid type size) and percentage of PopPUNK clusters (relative to the predicted plasmid type size).

Predicted pl. type	Graph Compo-nent	Partitioning of Graph Compo-nent [a]	Size	hierBAPS SC	PopPUNK cluster
1	1	-	11	SC10 (100%)	PUNK7(100%)
2	2	-	31	SC13(93.1%) SC10(3.4%) SC1(3.4%)	PUNK1(100.0%)
3	3	2	76	SC13(72.0%) SC2(12.0%) SC1(4.0%) SC10(4.0%) SC18(2.7%) SC9(1.3%) SC19(1.3%) SC22(1.3%) SC23(1.3%)	PUNK1(77.3%) PUNK9 (10.7%) PUNK3 (4.0%) PUNK4 (1.3%) PUNK14 (1.3%) PUNK19 (1.3%) PUNK28 (1.3%) PUNK30 (1.3%) PUNK32 (1.3%)
4	3	3	62	SC17(66.1%) SC10(17.7%) SC18(12.9%) SC22(1.6%) SC23(1.6%)	PUNK2(67.7%) PUNK1(8.5%) PUNK10(6.5%) PUNK4(3.2%) PUNK20(3.2%) PUNK11 (1.6%) PUNK16 (1.6%) PUNK17 (1.7%) PUNK19 (1.7%) PUNK 33(1.7%
5	6	-	20	SC13 (50.0%) SC17(40.0%) SC7(5.0%) SC8(5.0%)	PUNK1(50.0%) PUNK2(40.0%) PUNK13(5.0%) PUNK25(5.0%)
6	7	-	17	SC20 (47.1%) SC18 (35.3%) SC1(11.8%) SC2(5.9%)	PUNK5(35.3%) PUNK8(23.5%) PUNK3(11.8%) PUNK15(11.8%) PUNK10(5.9%) PUNK14(5.9%) PUNK35(5.9%)
7	8	-	12	SC29(41.7%) SC30(33.3%) SC32(25.0%)	PUNK6(100.0%)

8	13	-	10	SC18(100%)	PUNK4(100.0%)

[a] Partitioning of the graph component refers to the split obtained after applying the Louvain method, for finding community structure, in component 3.

References

1. Weiner-Lastinger LM, Abner S, Edwards JR, Kallen AJ, Karlsson M, Magill SS, Pollock D, See I, Soe MM, Walters MS, Dudeck MA. 2020. Antimicrobial-resistant pathogens associated with adult healthcare-associated infections: Summary of data reported to the National Healthcare Safety Network, 2015-2017. Infect Control Hosp Epidemiol 41:1–18.

2. Novosad SA, Fike L, Dudeck MA, Allen-Bridson K, Edwards JR, Edens C, Sinkowitz-Cochran R, Powell K, Kuhar D. 2020. Pathogens causing central-line-associated bloodstream infections in acute-care hospitals-United States, 2011-2017. Infect Control Hosp Epidemiol 41:313–319.

3. Tacconelli E, Magrini N, Kahlmeter G, Singh N. 2017. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. World Health Organization 27.

4. Cattoir V, Leclercq R. 2013. Twenty-five years of shared life with vancomycin-resistant enterococci: is it time to divorce? J Antimicrob Chemother 68:731–742.

5. Lebreton F, Depardieu F, Bourdon N, Fines-Guyon M, Berger P, Camiade S, Leclercq R, Courvalin P, Cattoir V. 2011. D-Ala-d-Ser *VanN*-type transferable vancomycin resistance in *Enterococcus faecium* . Antimicrob Agents Chemother 55:4606–4612.

6. Xu X, Lin D, Yan G, Ye X, Wu S, Guo Y, Zhu D, Hu F, Zhang Y, Wang F, Jacoby GA, Wang M.

2010. *vanM*, a new glycopeptide resistance gene cluster found in *Enterococcus faecium*. Antimicrob Agents Chemother 54:4643–4647.

7. Guzman Prieto AM, van Schaik W, Rogers MRC, Coque TM, Baquero F, Corander J, Willems RJL. 2016. Global Emergence and Dissemination of Enterococci as Nosocomial Pathogens: Attack of the Clones? Front Microbiol 7:788.

8. Saeedi B, Hallgren A, Jonasson J, Nilsson LE, Hanberger H, Isaksson B. 2002. Modified pulsed-field gel electrophoresis protocol for typing of enterococci. APMIS.

9. Top J, Schouls LM, Bonten MJM, Willems RJL. 2004. Multiple-locus variable-number tandem repeat analysis, a novel typing scheme to study the genetic relatedness and epidemiology of *Enterococcus faecium* isolates. J Clin Microbiol 42:4503–4511.

10. Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E, Van Embden JDA, Willems RJL. 2002. Multilocus sequence typing scheme for *Enterococcus faecium*. J Clin Microbiol 40:1963–1971.

11. Raven KE, Gouliouris T, Brodrick H, Coll F, Brown NM, Reynolds R, Reuter S, Török ME, Parkhill J, Peacock SJ. 2017. Complex Routes of Nosocomial Vancomycin-Resistant *Entero-*

coccus faecium Transmission Revealed by Genome Sequencing. Clin Infect Dis 64:886–893.

12. Werner G, Freitas AR, Coque TM, Sollid JE, Lester C, Hammerum AM, Garcia-Migura L, Jensen LB, Francia MV, Witte W, Willems RJ, Sundsfjord A. 2011. Host range of enterococcal *vanA* plasmids among Gram-positive intestinal bacteria. J Antimicrob Chemother 66:273–282.

13. Freitas AR, Coque TM, Novais C, Hammerum AM, Lester CH, Zervos MJ, Donabedian S, Jensen LB, Francia MV, Baquero F, Peixe L. 2011. Human and swine hosts share vancomycin-resistant *Enterococcus faecium* CC17 and CC5 and *Enterococcus faecalis* CC2 clonal clusters harboring Tn*1546* on indistinguishable plasmids. J Clin Microbiol 49:925–931.

14. Pinholt M, Bayliss SC, Gumpert H, Worning P, Jensen VVS, Pedersen M, Feil EJ, Westh H. 2019. WGS of 1058 *Enterococcus faecium* from Copenhagen, Denmark, reveals rapid clonal expansion of vancomycin-resistant clone ST80 combined with widespread dissemination of a *vanA*-containing plasmid and acquisition of a heterogeneous accessory genome. J Antimicrob Chemother 74:1776–1785.

15. Lee RS, Gonçalves da Silva A, Baines SL, Strachan J, Ballard S, Carter GP, Kwong JC, Schultz MB, Bulach DM, Seemann T, Stinear TP, Howden BP. 2018. The changing landscape of vancomycin-resistant *Enterococcus faecium* in Australia: a population-level genomic study. J Antimicrob Chemother 73:3268–3278.

16. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, Giess A, Pankhurst LJ, Vaughan A, Grim CJ, Cox HL, Yeh AJ, Modernising Medical Microbiology (MMM) Informatics Group, Sifri CD, Walker AS, Peto TE, Crook DW, Mathers AJ. 2016. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene bla_{KPC}. Antimicrob Agents Chemother 60:3767–3778.

17. Freitas AR, Novais C, Tedim AP, Francia MV, Baquero F, Peixe L, Coque TM. 2013.Microevolutionary events involving narrow host plasmids influences local fixation of vancomycin-resistance in Enterococcus populations. PLoS One 8:e60589.

18. Arredondo-Alonso S, Top J, McNally A, Puranen S, Pesonen M, Pensar J, Marttinen P, Braat JC, Rogers MRC, van Schaik W, Kaski S, Willems RJL, Corander J, Schürch AC. 2020. Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus faecium*. MBio 11.

19. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJL, Schürch AC. 2020. gplas: a comprehensive tool for plasmid analysis using short-read graphs. Bioinformatics.

20. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30:1224–1228.

21. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316.

22. Pritchard L, Cock P, Esen Ö. 2019. pyani v0. 2.8: average nucleotide identity (ANI) and related measures for whole genome comparisons.

23. Brodrick HJ, Raven KE, Harrison EM, Blane B, Reuter S, Török ME, Parkhill J, Peacock SJ. 2016. Whole-genome sequencing reveals transmission of vancomycin-resistant *Enterococcus faecium* in a healthcare network. Genome Med 8:4.

24. Carter GP, Buultjens AH, Ballard SA, Baines SL, Tomita T, Strachan J, Johnson PDR, Fer-

guson JK, Seemann T, Stinear TP, Howden BP. 2016. Emergence of endemic MLST non-typeable vancomycin-resistant *Enterococcus faecium*. J Antimicrob Chemother 71:3367–3371.

25. Howden BP, Holt KE, Lam MMC, Seemann T, Ballard S, Coombs GW, Tong SYC, Grayson ML, Johnson PDR, Stinear TP. 2013. Genomic insights to control the emergence of vancomycin-resistant enterococci. MBio 4.

26. Abdelbary MHH, Senn L, Greub G, Chaillou G, Moulin E, Blanc DS. 2019. Whole-genome sequencing revealed independent emergence of vancomycin-resistant *Enterococcus faecium* causing sequential outbreaks over 3 years in a tertiary care hospital. Eur J Clin Microbiol Infect Dis 38:1163–1170.

27. Lanza VF, Baquero F, de la Cruz F, Coque TM. 2017. AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics 33:283–285.

28. Corel E, Lopez P, Méheust R, Bapteste E. 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol 24:224–237.

29. Bernard G, Greenfield P, Ragan MA, Chan CX. 2018. k-mer Similarity, Networks of Microbial Genomes, and Taxonomic Rank. mSystems 3.

30. Acman M, van Dorp L, Santini JM, Balloux F. 2020. Large-scale network analysis captures biological features of bacterial plasmids. Nat Commun 11:2452.

31. Jesus TF, Ribeiro-Gonçalves B, Silva DN, Bortolaia V, Ramirez M, Carriço JA. 2019. Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. Nucleic Acids Res 47:D188–D194.

32. Galata V, Fehlmann T, Backes C, Keller A. 2019. PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Res 47:D195–D202.

33. Freitas AR, Tedim AP, Francia MV, Jensen LB, Novais C, Peixe L, Sánchez-Valenzuela A, Sundsfjord A, Hegstad K, Werner G, Sadowy E, Hammerum AM, Garcia-Migura L, Willems RJ, Baquero F, Coque TM. 2016. Multilevel population genetic analysis of *vanA* and *vanB Enterococcus faecium* causing nosocomial outbreaks in 27 countries (1986-2012). J Antimicrob Chemother 71:3351–3366.

34. Weingarten RA, Johnson RC, Conlan S, Ramsburg AM, Dekker JP, Lau AF, Khil P, Odom RT, Deming C, Park M, Thomas PJ, Henderson DK, Palmore TN, Segre JA, Frank KM. 2018. Genomic Analysis of Hospital Plumbing Reveals Diverse Reservoir of Bacterial Plasmids

Conferring Carbapenem Resistance. mBio.

35. Stoesser N, Phan HTT, Seale AC, Aiken Z, Thomas S, Smith M, Wyllie D, George R, Sebra R, Mathers AJ, Vaughan A, Peto TEA, Ellington MJ, Hopkins KL, Crook DW, Orlek A, Welfare W, Cawthorne J, Lenney C, Dodgson A, Woodford N, Sarah Walker A, the TRACE Investigators' Group. Genomic epidemiology of a complex, multi-species plasmid-borne bla_{KPC} carbapenemase outbreak in Enterobacterales in the UK, 2009-2014.

36. Torres-González P, Bobadilla-Del Valle M, Tovar-Calderón E, Leal-Vega F, Hernández-Cruz A, Martínez-Gamboa A, Niembro-Ortega MD, Sifuentes-Osornio J, Ponce-de-León A. 2015. Outbreak caused by Enterobacteriaceae harboring NDM-1 metallo-β-lactamase carried in an IncFII plasmid in a tertiary care hospital in Mexico City. Antimicrob Agents Chemother 59:7080–7083.

37. Bosch T, Lutgens SPM, Hermans MHA, Wever PC, Schneeberger PM, Renders NHM, Leenders AC, Kluytmans JA, Schoffelen A, Notermans D, Others. 2017. 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 55:2380–2390.

38. Martin J, Phan HTT, Findlay J, Stoesser N, Pankhurst L, Navickaite I, De Maio N, Eyre DW, Toogood G, Orsi NM, Others. 2017. Covert dissemination of carbapenemase-producing *Klebsiella pneumoniae* (KPC) in a successfully controlled outbreak: long-and short-read whole-genome sequencing demonstrate multiple genetic modes of transmission. J Antimicrob Chemother 72:3025–3034.

39. Hidalgo L, de Been M, Rogers MRC, Schürch AC, Scharringa J, van der Zee A, Bonten MJM, Fluit AC. 2019. Sequence-based epidemiology of an OXA-48 plasmid during a hospital outbreak. Antimicrob Agents Chemother.

40. Conlan S, Lau AF, Deming C, Spalding CD, Lee-Lin S, Thomas PJ, Park M, Dekker JP, Frank KM, Palmore TN, Segre JA. 2019. Plasmid Dissemination and Selection of a Multidrug-Resistant *Klebsiella pneumoniae* Strain during Transplant-Associated Antibiotic Therapy. MBio 10.

41. Harris PNA, M WA. 2020. Beyond the Core Genome: Tracking Plasmids in Outbreaks of Multidrug-resistant Bacteria. Clinical Infectious Diseases.

42. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. 2017. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genom 3.

43. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, Willems RJL, Schürch AC. 2018. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genom 4.

44. Krueger F. 2012. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. URL http://www bioinformatics babraham ac uk/projects/trim_galore.

45. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.

46. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev M a., Pevzner P a. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455–477.

47. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644.

48. Sheppard AE, Stoesser N, German-Mesner I, Vegesana K, Sarah Walker A, Crook DW, Mathers AJ. 2018. TETyper: a bioinformatic pipeline for classifying variation and genetic contexts of transposable elements from short-read whole-genome sequencing data. Microb Genom 4.

49. Arthur M, Molinas C, Depardieu F. 1993. Characterization of Tn*1546*, a Tn3-related transposon conferring glycopeptide resistance by synthesis of depsipeptide peptidoglycan precursors in *Enterococcus faecium* BM4147. Journal of Bacteriology 175:117–127.

50. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM.

2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132.

51. Csardi G, Nepusz T, Others. 2006. The igraph software package for complex network research. InterJournal, Complex Systems 1695:1–9.

52. Clewell DB, Weaver KE, Dunny GM, Coque TM, Francia MV, Hayes F. 2014. Extrachromosomal and Mobile Elements in Enterococci: Transmission, Maintenance, and Epidemiology, p. . In Gilmore, MS, Clewell, DB, Ike, Y, Shankar, N (eds.), Enterococci: From Commensals to Leading Causes of Drug Resistant Infection. Massachusetts Eye and Ear Infirmary, Boston.

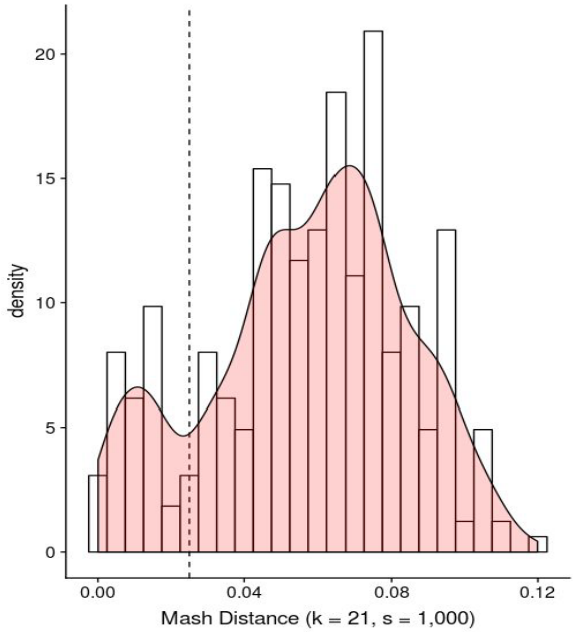
53. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403.

54. Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334–2335.

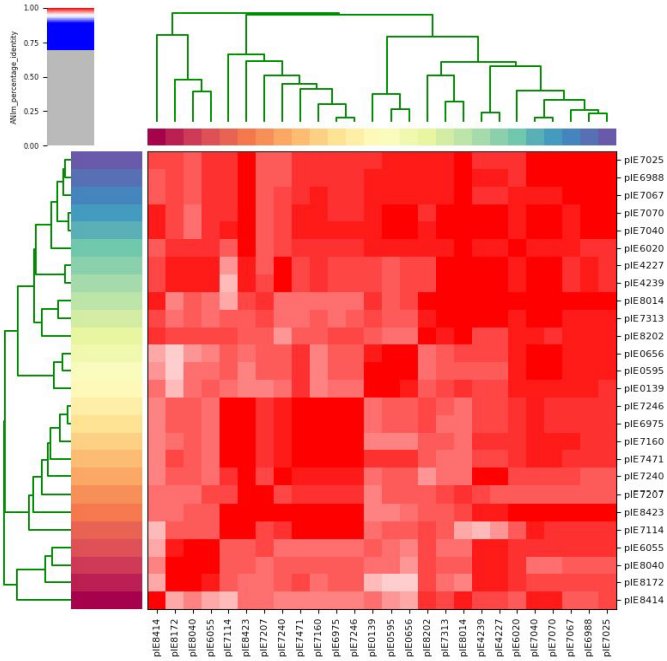
55. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36.

56. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2.

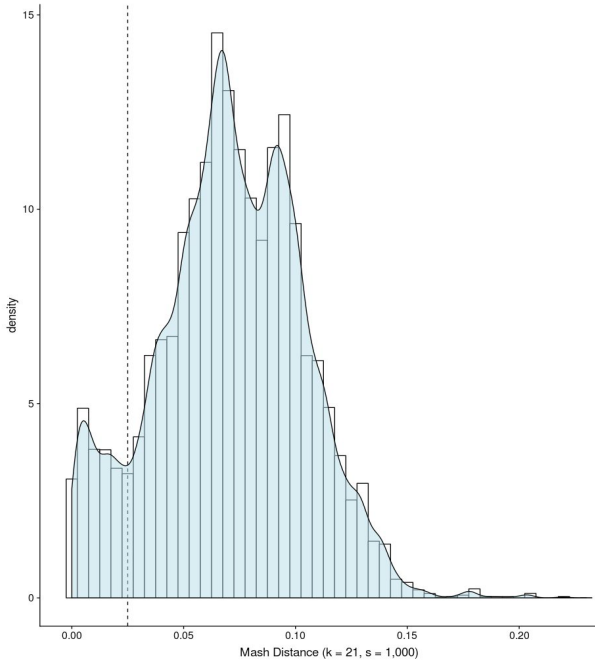
Supplementary Figures



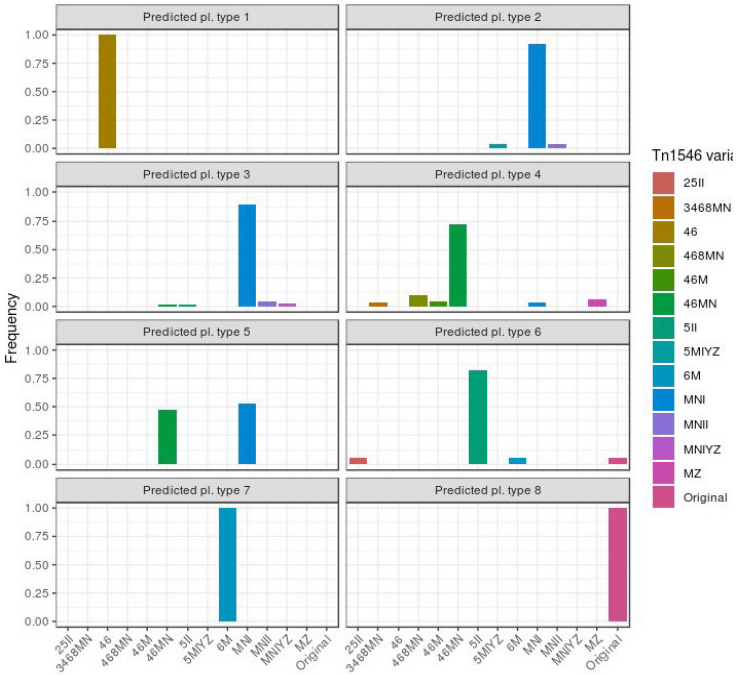
Suppl. Figure S1 Distribution of pairwise Mash distances (k = 21, s = 1,000) between *vanA* complete plasmid sequences (n = 26).



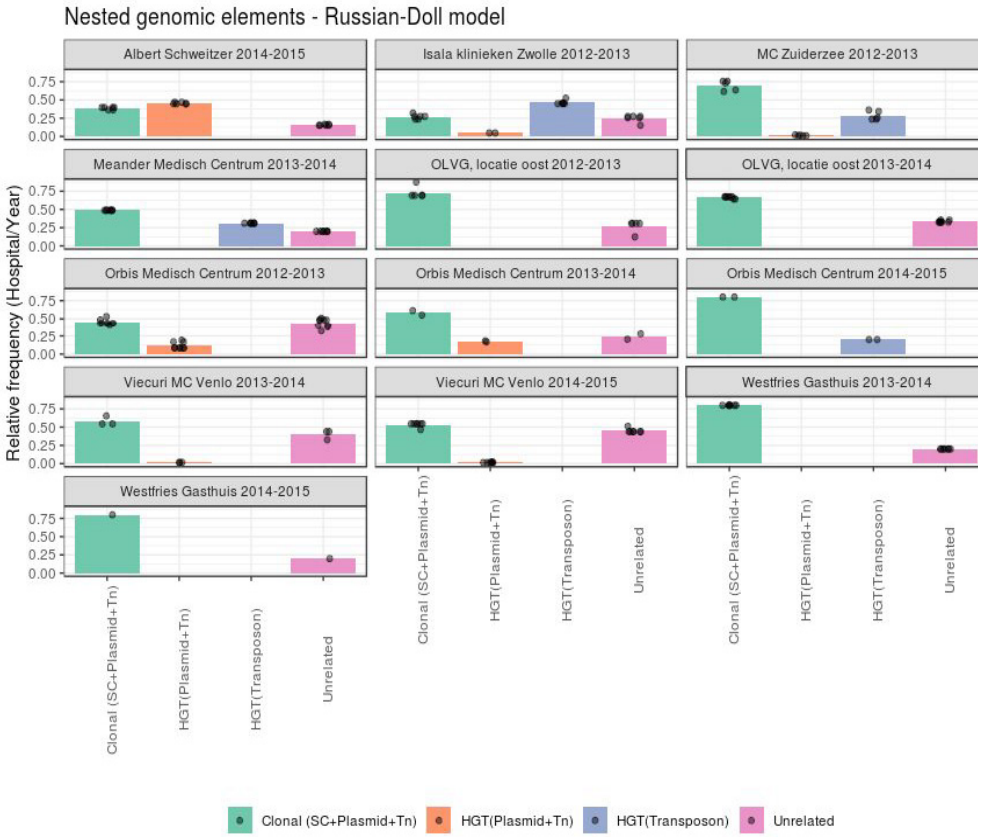
Suppl. Figure S2. Heatmap and single-linkage clustering of the pyani pairwise alignment identity in the set of *vanA* complete plasmid sequences (n = 26).



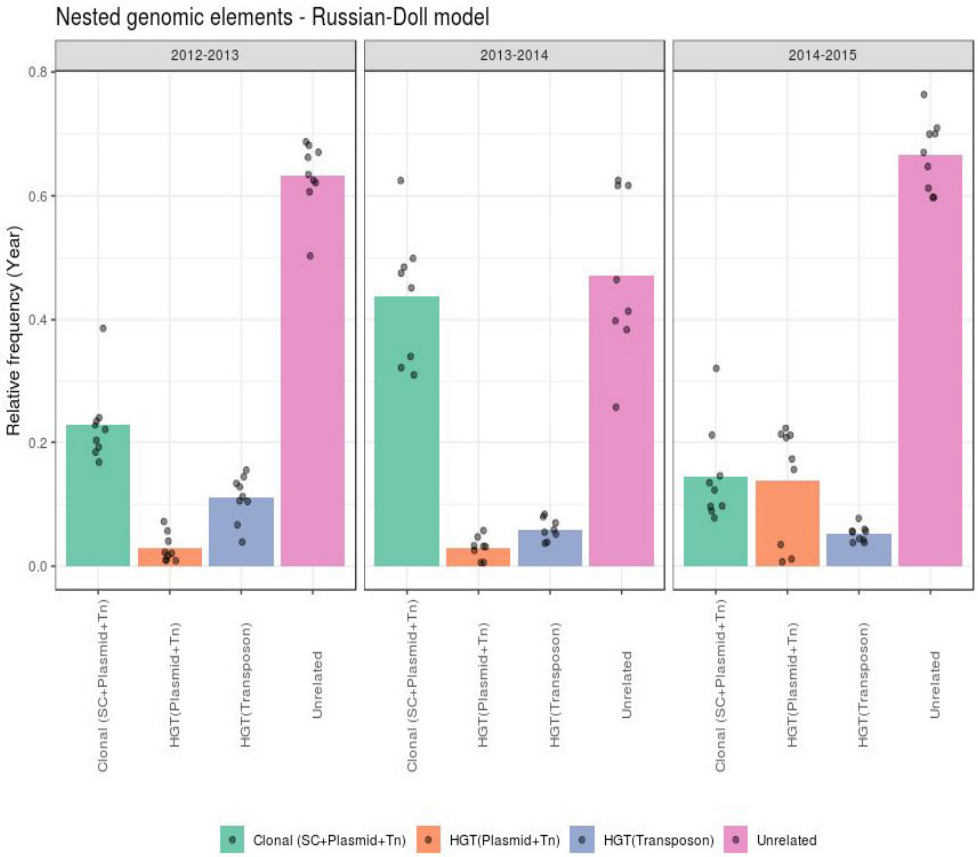
Suppl. Figure S3. Distribution of pairwise Mash distances (k = 21, s = 1,000) between predicted *vanA* plasmid sequences (n = 303).



Suppl. Figure S4. Frequency of the 14 Tn *1546* variants present in our *vanA* predicted plasmid types (n = 8) .



Suppl. Figure S5. Contribution of nested genetic elements in the dissemination of vancomycin resistance in specific Dutch hospitals. VRE samples isolated in 12 consecutive months and from the same hospital were considered. Clonal dissemination (green bar) corresponded to pairs of isolates sharing the same hierBAPS SC, *vanA* plasmid type and Tn*1546* variant. Plasmid-mediated dissemination (orange bar) corresponded to pairs of isolates sharing the same *vanA* plasmid type and Tn*1546* variant but different hierBAPS SC. Transposon-dissemination (blue bar) corresponded to pairs of isolates sharing the same Tn*1546* variant, different *vanA* plasmid type and same or different the hierBAPS SC. Unrelated (purple bar) isolates corresponded to pairs of isolates with a different Tn*1546* variant.



Suppl. Figure S6. C ontribution of nested genetic elements in country-wide dissemination of vancomycin resistance. VRE samples isolated in 12 consecutive months in the Netherlands were considered. Clonal dissemination (green bar) corresponded to pairs of isolates sharing the same hierBAPS SC, *vanA* plasmid type and Tn*1546* variant. Plasmid-mediated dissemination (orange bar) corresponded to pairs of isolates sharing the same *vanA* plasmid type and Tn *1546* variant but different hierBAPS SC. Transposon-dissemination (blue bar) corresponded to pairs of isolates sharing the same Tn *1546* variant, different *vanA* plasmid type and same or different the hierBAPS SC. Unrelated (purple bar) isolates corresponded to pairs of isolates with a different Tn*1546* variant.

7

General discussion

Sergio Arredondo-Alonso

Plasmids are extrachromosomal elements that can disseminate between strains or even between different bacterial species. Plasmids can also acquire novel genetic traits by acquisition of transposons or cointegration with other plasmids thereby providing host bacterial strains novel adaptive traits. These inherent characteristics make plasmids optimal vectors for disseminating antimicrobial resistance (AMR) (1, 2). Plasmid-mediated AMR dissemination often follows a Russian-Doll model in which nested genomic elements intervene in resistance propagation by vertical and horizontal transmission (3). However, current epidemiological studies on AMR dissemination often solely focus on clonal outbreaks (4). Traditional typing methods such as MLST, but also current whole genome sequence (WGS)-based epidemiological studies fail to resolve the dynamics of plasmids and thus challenge the monitoring of plasmid-mediated AMR (3). A main reason for this focus on clonal dissemination is the challenge of reconstructing plasmid sequences from short-read WGS as explained in **chapter 2.** The scope of this thesis was to develop a new set of tools in order to overcome this limitation and accurately detect and trace plasmid sequences, from short-read WGS data, in the nosocomial pathogen *Enterococcus* faecium. This was especially relevant to investigate the dissemination of antimicrobial resistance (AMR), in particular vancomycin resistance, as well as other disseminated genes. Below I will review the overall results of **chapters 2 to 6** and discuss the limitations and further directions of the developed tools presented **in chapter 3** (mlplasmids) and **chapter 5** (gplas).

In **chapter 2**, we faced the challenge of detecting plasmid sequences from WGS with a limited set of bioinformatic tools. At the start of this thesis, the reconstruction of plasmids seemed to be apparently solved by the introduction of short-read WGS and tools such as PlasmidFinder that permitted the detection of *Enterobacteriaceae* plasmid replicons, or by Placnet that required manual expert analysis of plasmid networks (5, 6). To shed some light on whether the reconstruction of plasmids was feasible from short-read WGS, we performed a benchmark of different tools to automate the reconstruction of plasmids. In this benchmark, we used a selection of 42 bacterial genomes spanning 12 different genera, which corresponded to all complete genomes in the NCBI public databases and with short-read data available at that time (July 2016). This selection was clearly affected by the content of the NCBI database which is skewed for clinically relevant species belonging to the *Enterobacteriaceae* family (7). Furthermore, we did not consider the complexity of the genomes in terms of number of contigs, number of components or dead-ends present in the *de Bruijn* graph that could affect the performance of *de-novo* assembly tools such as Recycler (8) or PlasmidSPAdes (9).

Despite the limitations of the benchmarking set, this comparison was well received by the community because it provided one of the first independent studies which reflected some of the bottlenecks and limitations encountered by other researchers trying to predict plasmid sequences from WGS. We determined PlasmidSPAdes as the best tool to reconstruct plasmid sequences. This tool uses a *de-novo* approach that could be applied to any bacterial species and assumes that plasmids have a different copy number than the chromosome. However, we clearly showed the two limitations of this tool: i) large plasmids with a similar copy number as the chromosome were wrongly absent from the prediction and ii) 84% of the predictions reported by PlasmidSPAdes merged one or more plasmids into the same component. The benchmark presented **in chapter 2** motivated the development of new plasmid prediction tools such as PlasmidTron, MOB-Suite, PlasFlow, PlaScope, metaplasmidSPAdes (10–13) that mentioned our comparison as a rational to develop their tools and plasmid databases (14–16). Furthermore, Orlek et al. also cited the study presented **in chapter 2** to emphasize that the binary origin of contigs (plasmid or chromosome) was a more feasible problem than predicting plasmid boundaries (17). Today, the emergence of these and other new plasmid prediction tools (11, 12, 15, 18–20) and the increase of the availability of complete genomes would require a new independent benchmark study in which the current status of plasmid reconstruction using short-read WGS should be updated.

A novel methodology to predict the repertoire of plasmid sequences within *E. faecium* was developed using a large collection of isolates from different sources including human and non-human samples. For this we first performed short-read WGS on 1,644 isolates.

Furthermore, we sequenced a subset of 62 isolates with long-read WGS (Oxford Nanopore Technologies). These 62 isolates were fundamental to develop a novel machine-learning classifier to predict the origin of contigs derived from *E. faecium*. The selection of these isolates had to ensure enough plasmid diversity and variability. According to the results of **chapter 2**, we performed an initial plasmid prediction of the 1,644 isolates using Plasmid-SPAdes. Orthologous genes (OG) were searched to estimate the presence and absence of plasmid genes and t-SNE clustering (unsupervised approach) was considered to reduce the dimensionality problem. Finally, the selection of the long-read isolates was achieved by fixing the number of centroids, using a supervised approach (k-means), corresponding to the number (62) of long-read isolates. This selection performed **in chapter 3** was fundamental to achieve a machine-learning classifier that confidently predicted plasmid sequences in *E. faecium* and to avoid underrepresentation of sequences derived from non-human sources. The classifier (support-vector machine, SVM) developed for *E. faecium* achieved an accuracy of 0.94 and F1-score of 0.92 on an independent set of isolates. To facilitate the usage of the tool to a wider non-bioinformatics community, we made the classifier available as an R package called 'mlplasmids' but also as a web-server graphical interface and provided two other SVM classifiers to predict the origin of contigs in *Klebsiella pneumoniae* and *Escherichia coli*.

The accuracy (0.95) and F1-score (0.76) obtained in the *E. coli* model highlighted one limitation of training a machine-learning classifier using sequences from NCBI databases. The model was trained on only 168 completely assembled genomes available at the time (March 2018). We observed that the *E. coli* model suffered from false-negatives, plasmids which were not predicted as belonging to that class, which was related to a lack of plasmid variability and problems with the ratio between plasmid and chromosomal contigs present in the training set and used in the construction of the model. This confirms the importance of training the algorithms with a diverse strain set of completely assembled genomes covering the plasmid diversity present in the bacterial species. Today, with more than 1000 completely assembled *E. coli* genomes available in public databases, an update and re-training of the *E. coli* model is necessary to effectively apply mlplasmids and gplas to *E. coli* plasmid population analyses.

In *E.faecium*, a strain set was selected that captured much of the plasmid diversity of this species to achieve a classifier with a good balance between specificity and sensitivity. Furthermore, the models use pentamer frequencies (1,024) to differentiate between plasmids and chromosomal sequences, as previously used by cBar in a metagenomics approach (21). In our case, we decided to train the models for individual species and thus limited the range of usability of these classifiers. However, this decision was fundamental to increase the accuracy of prediction and to avoid the incorrect prediction of chromosomal

sequences (false positives) in the models which can contaminate all downstream analyses as shown in **chapter 5** when combining gplas together with PlasFlow.

The models in mlplasmids have been used in the community not only to predict short-read contigs but also to predict contigs coming from incomplete long-read assemblies (22). There are two other challenging cases in which the length of the contigs being predicted plays a fundamental role: i) prediction of contigs bearing AMR genes flanked by IS elements or transposon sequences and ii) prediction of plasmids integrated into the chromosome. In the first case, the presence of left or right flanks that surround the transposon or IS element around the AMR gene, is fundamental to predict the origin of that contig. Otherwise, the model may report a posterior probability around the threshold (0.5) which makes assignment to both classes equally likely. If there are no right- or left-flanks with a chromosome k-mer signature, the contig would be predicted as plasmid-derived, despite being integrated into the chromosome. The assignment of plasmid contigs for these two cases can be improved by looking at the neighbouring contigs in the plasmid network as shown **in chapter 5**.

In chapter 4, we used mlplasmids to perform a plasmid prediction of the 1,644 *E. faecium* isolates present in our collection. This prediction allowed us to observe the presence of several host-associated plasmidome populations. Most notably, the plasmidome from hospitalized patient isolates were clearly distinct and these isolates also had an overall larger plasmidome content. These plasmidome populations were compared using a k-mer analysis with Mash (23) based on a previous benchmarking study that evaluated and provided guidance on different phylogenetic strategies from WGS (24). K-mer analysis and comparison is becoming a preferred method to compare plasmids (15, 25) which are unsuited to be analysed by traditional phylogenetic methods because of a lack of core elements and high ratio of recombination.

We observed that plasmid k-mer distances were the key factor in explaining distinct host populations compared to other factors such as geographical distance or time of isolation between pairs of isolates. These plasmidome populations are shaped by the flow of plasmid sequences but with restrictions imposed by ecological constraints and active HGT barriers, e.g. type I RM systems, and less by physical interaction between hosts. Most of the plasmid genes enriched in these populations were poorly characterized and their function was unknown. This attests to the importance of performing functional genomic-based approaches like Tn-seq (26) and Crispr-Cas technology (27) to functionally characterize the role of plasmid genes. Finally, a limitation of the approach followed in chapter 3 is that the proposed k-mer analysis allows studying the plasmidome but not individual plasmids. It is therefore important to realize that the plasmidome populations defined in the study may be driven by the k-mer content of large plasmid sequences, which could mask the

role of medium or smaller plasmids in e.g. host adaptation or AMR dissemination.

The plasmidome prediction given by mlplasmids only allowed to confidently predict the binary origin of contigs (plasmid- or chromosome-derived) and thus lacked plasmid boundaries. These plasmid boundaries, however, are required to predict whether AMR genes, such as the *vanA* gene cluster, are found in a single conserved plasmid sequence or whether there are different plasmid configurations bearing the *vanA* gene. This motivated the development of a novel bioinformatic tool ('gplas', **chapter 5**) that opened the possibility of binning plasmid-predicted sequences into different clusters or plasmid-types. These clusters correspond to the plasmid sequence from which they originated and thus allowed to track and trace individual plasmids in the population.

Gplas built on different concepts of some of the tools benchmarked **in chapter 2**: i) contigs originating from the same plasmid should have a uniform sequence coverage (8, 9), ii) a greedy-approach can be used to find plasmid walks in the assembly graph (19) and iii) in the generated plasmidome network, highly-connected components correspond to independent plasmid sequences (18). In order to retrieve these features, gplas uses the assembly graph (in gfa format) and not merely the collection of contigs reported by the assembler. Of note here is that we built on plasmid features from tools such as Recycler (8) that, by themselves, showed a poor performance in the benchmark study of **chapter 2**.

In **chapter 5,** we describe the combination and adaptation of previously postulated plasmid features and the introduction of novel features to interrogate the resulting plasmid network. This dramatically increased the prediction of plasmid boundaries. We observed that the initial binary prediction was fundamental to achieve a reliable prediction as shown with the precision (0.82) and completeness (0.72) obtained when using gplas in combination with mlplasmids. The performance of gplas, especially in the combination with mlplasmids, outperformed other *de-novo* or reference-based tools. However, when plasmids had a similar k-mer coverage and shared repeat sequences in the underlying *de Bruijn* graph, this resulted in the presence of bins in which two or more plasmids were merged. Despite this limitation, the rest of the plasmid network gave us detailed insights into the topology and composition of plasmids between two or more isolates. Two isolates from different plasmidome populations sharing a bin while the rest of the plasmid network is different can be indicative for the transmission of an individual plasmid between these isolates. This was fundamental to detect AMR transmission between different clonal but also plasmidome backgrounds. Furthermore, the prediction given by gplas was able to help to corroborate the initial assignment given by mlplasmids or PlasFlow in challenging cases. An example are AMR genes or plasmid sequences integrated in the chromosome. In these cases, additional information was obtained by looking into the neighbouring contigs surrounding the contig of interest in the plasmidome network. The

presence of a contig in a bin with multiple other plasmid-predicted contigs can help to confirm the plasmid origin of that sequence. In the case of a plasmid integrated into the chromosome, the node is unbinned as the contig is not linked to other plasmid contigs in the network.

PlasFlow (12) was developed using a machine-learning approach but trained on a large number of different bacterial species and genera. This general approach allowed to predict plasmid sequences from any bacterial species and was mainly developed to be used in metagenomic samples. However, as shown **in chapter 5**, the usage of such a general approach comprises a risk of wrongly predicting chromosome-sequences which can hinder downstream analysis such as the binning of plasmid-predicted contigs into individual sequences. For this reason, **in chapter 3**, we focused our efforts on developing several machine-learning classifiers for distinct bacterial species, such as pathogens included in the ESKAPE list. The combination of gplas with PlasFlow allowed to obtain plasmid boundaries for any bacterial species and thus increased the usage and applicability of the tool.

A current limitation of gplas is that it can only be applied to single genome isolates and is thus not suitable for metagenomics samples, unlike plasmid prediction tools as metaPlasmidSpades (9), Recycler (8), Plasflow (12) and SCAPP (28). However, gplas could be repurposed to be used with metagenomics samples. Gplas assumes that there is a single chromosomal sequence in the sample and all chromosomal predicted contigs are considered to estimate the expected variation of the chromosomal k-mer coverage introduced during the library or sequencing preparation. This estimation needs to be adapted in the case of metagenomic samples in which several genomes are present in the same sample. The rest of the assumptions applied in gplas could be implemented in a metagenomics approach without further modifications.

In **chapter 6**, we used gplas in combination with mlplasmids, presented in **chapters 5 and 3**, respectively, to predict the plasmids carrying the *vanA* gene cluster in 309 vancomycin resistant *E. faecium* (VRE) isolates from 32 Dutch hospitals and isolated between 2012 and 2015. For these isolates, short-read WGS was generated in **chapter 4.** We exclusively focused on bins predicted by gplas containing the contig encoding for the *vanA* gene cluster. Next, we considered a network approach to integrate all k-mer distances between *vanA* bins. This avoided the usage of an arbitrary reference plasmid to compare plasmid bins and integrated the modularity observed in bacterial plasmids (7). Lanza et al. (29) already postulated the use of a network approach to investigate the accessory genome of bacterial species. In this **chapter 6,** we considered a network in which nodes represented gplas bins bearing the *vanA* gene cluster, and edges connections between bins with a high similarity based on their pairwise k-mer distance. The topology and structure of the network suggested the presence of 8 distinct clusters that were considered

later as predicted plasmid-types. These clusters contained plasmids carried by isolates belonging to different sequence clusters (SC) which already suggested the dissemination of highly similar *vanA* plasmid-types between distinct VRE clonal complexes. We then integrated into the network the complete plasmids bearing the *vanA* gene cluster which were assembled in **chapter 4.** These complete plasmids elucidated the content of four of the predicted plasmid-types and allowed to partially validate the predicted network. The predicted plasmid-types lacking a complete *vanA* plasmid could represent novel *vanA* plasmid sequences not assembled in **chapter 4** or could reflect difficulties in the prediction given by gplas. In this case, the integration of all previously completed *vanA* resistant plasmids present in public databases using tools such as Plasmid Atlas or PLSDB (14, 15) would elucidate the genomic content of the remaining predicted plasmid-types, and harmonize the plasmid annotation using previously defined plasmid-type schemes (30).

The reconstruction of the plasmid-types bearing the *vanA*-type gene cluster facilitated the nestedness analysis of clonal complex, plasmid-type and Tn*1546* variant. This elucidated that, overall, clonal dissemination, in which the same clone, plasmid-type and Tn*1546* variant were vertically inherited, contributed most (~45%) to the spread of *vanA*-type of vancomycin resistance in Dutch hospitals between 2012-2015. However, we also observed scenarios of transposon-mediated outbreaks in which Tn*1546* transposition and mobilisation using distinct plasmid sequences was responsible (> 70%) for most of the observed vancomycin resistance dissemination (e.g. Flevoland 2012-2013 or North-Holland 2012-2013). Furthermore, we also observed a potential plasmid-mediated outbreak occurring in South Holland in 2014 to 2015 in which a particular *vanA* plasmid-type contributed most (~42%) to the dissemination of vancomycin resistance in that particular Dutch region. Based on this, we suggested that, to confirm potential epidemiological links and correctly assess the effectiveness of infection control policies or antimicrobial stewardships in clinical settings, dissemination of all the different layers of the Russian-Doll model of nested genomic complexes (3, 17) such as clone, plasmid, transposon and other mobile genetic elements, need to be taken into account.

In this thesis, we propose several approaches to improve plasmid prediction and reconstruction of short-read sequence data. I demonstrated that gplas in combination with mlplasmids or PlasFlow can provide detailed insights on shared similar plasmidome networks among isolates and shared plasmid bins containing AMR genes. Without doubt the decrease in costs of long-read sequencing and increase in the read-accuracy and output of long-read sequencing platforms such as the GridION or PromethION is causing a shift in the current genomics paradigm by facilitating not only plasmid assembly but also detection of structural variants among a wide range of applications in bacterial and eukaryotic genomes (31–34). However, the enormous amount of short-read WGS data already gen-

erated and publicly available in databases (35) makes the tools developed in this thesis to predict and reconstruct plasmid sequences based on existing short-read sequencing data extremely valuable for studying plasmid-based host-adaptation and the epidemiology of plasmid-encoded AMR, which is important for the effective implementation of infection control policies and antibiotic stewardships.

References

1. Courvalin P. 1994. Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. Antimicrob Agents Chemother 38:1447–1451.

2. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, Mevius DJ, Hordijk J. 2018. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. J Antimicrob Chemother 73:1121–1137.

3. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, Giess A, Pankhurst LJ, Vaughan A, Grim CJ, Cox HL, Yeh AJ, Modernising Medical Microbiology (MMM) Informatics Group, Sifri CD, Walker AS, Peto TE, Crook DW, Mathers AJ. 2016. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene bla_{KPC}. Antimicrob Agents Chemother 60:3767–3778.

4. Harris PNA, M WA. 2020. Beyond the Core Genome: Tracking Plasmids in Outbreaks of Multidrug-resistant Bacteria. Clinical Infectious Diseases ciaa052.

5. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903.

6. Lanza VF, de Toro M, Garcillan-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz F. 2014. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. PLoS Genet 10:e1004766.

7. Pesesky MW, Tilley R, Beck DAC. 2019. Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. Plasmid 102:10–18.

8. Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E, Mizrahi I, Shamir R. 2016. Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. Bioinformatics 33:475-482.

9. Antipov D, Hartwick N, Shen M, Raiko M, Pevzner PA. 2016. plasmidSPAdes : Assembling Plasmids from Whole Genome Sequencing Data. Bioinformatics 32:3380–3387.

10. Page AJ, Wailan A, Shao Y, Judge K, Dougan G, Klemm EJ, Thomson NR, Keane JA. 2018. PlasmidTron: assembling the cause of phenotypes and genotypes from NGS data. Microb Genom 4:e000164.

11. Robertson J, Nash JHE. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom 4:e000206.

12. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res 46:e35.

13. Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, Vallenet D. 2018. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. Microb Genom 4:e000211.

14. Galata V, Fehlmann T, Backes C, Keller A. 2019. PLSDB: a resource of complete bacterial plasmids. Nucleic Acids Res 47:D195–D202.

15. Jesus TF, Ribeiro-Gonçalves B, Silva DN, Bortolaia V, Ramirez M, Carriço JA. 2019. Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. Nucleic Acids Res 47:D188–D194.

16. Douarre P-E, Mallet L, Radomski N, Felten A, Mistou M-Y. 2020. Analysis of COMPASS, a New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive Diversity of IncF Plasmids. Front Microbiol 11:483.

17. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, Peto T, Crook D, Woodford N, Walker AS, Phan H, Sheppard AE. 2017. Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. Front Microbiol 8:182.

18. Vielva L, de Toro M, Lanza VF, de la Cruz F. 2017. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics 33:3796-3798

19. Müller R, Chauve C. 2019. HyAsP, a greedy tool for plasmids identification. Bioinformatics 35:4436-4439.

20. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. 2020. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein-sequence-based replicon distribution scores. bioRxiv.

21. Zhou F, Xu Y. 2010. cBar: A computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 26:2051–2052.

22. Decano AG, Ludden C, Feltwell T, Judge K, Parkhill J, Downing T. 2019. Complete Assembly of *Escherichia coli* Sequence Type 131 Genomes Using Long Reads Demonstrates Antibiotic Resistance Gene Variation within Diverse Plasmid and Chromosomal Contexts. mSphere 4:e00130-19.

23. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132.

24. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, Harris SR. 2018. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Res 3:33.

25. Acman M, van Dorp L, Santini JM, Balloux F. 2020. Large-scale network analysis captures biological features of bacterial plasmids. Nat Commun 11:2452.

26. Zhang X, de Maat V, Guzmán Prieto AM, Prajsnar TK, Bayjanov JR, de Been M, Rogers MRC, Bonten MJM, Mesnage S, Willems RJL, van Schaik W. 2017. RNA-seq and Tn-seq reveal fitness determinants of vancomycin-resistant *Enterococcus faecium* during growth in human serum. BMC Genomics 18:893.

27. de Maat V, Stege PB, Dedden M, Hamer M, van Pijkeren J-P, Willems RJL, van Schaik W. 2019. CRISPR-Cas9-mediated genome editing in vancomycin-resistant *Enterococcus faecium*. FEMS Microbiol Lett 366.

28. Pellow D, Probst M, Furman O, Zorea A, Segal A. 2020. SCAPP: An algorithm for improved plasmid assembly in metagenomes. bioRxiv.

29. Lanza VF, Baquero F, de la Cruz F, Coque TM. 2017. AcCNET (Accessory Genome Con-

stellation Network): comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics 33:283–285.

30. Freitas AR, Tedim AP, Francia MV, Jensen LB, Novais C, Peixe L, Sánchez-Valenzuela A, Sundsfjord A, Hegstad K, Werner G, Sadowy E, Hammerum AM, Garcia-Migura L, Willems RJ, Baquero F, Coque TM. 2016. Multilevel population genetic analysis of *vanA* and *vanB Enterococcus faecium* causing nosocomial outbreaks in 27 countries (1986-2012). J Antimicrob Chemother 71:3351–3366.

31. Rang FJ, Kloosterman WP, de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol 19:90.

32. Wick RR, Judd LM, Holt KE. 2019. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology 20:129.

33. Nicholls SM, Quick JC, Tang S, Loman NJ. 2019. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. Gigascience 8:giz043.

34. Coster WD, De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, Sleegers K, Van Broeckhoven C. 2019. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Research 29:1178-1187.

35. Schürch AC, van Schaik W. 2017. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. Annals of the New York Academy of Sciences 1:108-120.

English summary

Bacteria can hold autonomous replicating elements called plasmids which can act as major drivers of genetic variation and adaptation. Plasmids can facilitate the dissemination of antimicrobial resistance genes (AMR) within and between bacterial species by horizontal gene transfer (HGT). Thus, the reconstruction plasmid sequences is essential to fully understand how the introduction and dissemination of AMR genes takes place into bacterial populations.

In this thesis, I have focused on the nosocomial pathogen *Enterococcus faecium* which commonly inhabits the intestinal tract of animals and humans. However, *E. faecium* is also an important causative agent of hospital-acquired infections. In the last three decades, *E. faecium* has acquired resistance against multiple antibiotics, most notably ampicillin, gentamicin and vancomycin. The presence of many antibiotic resistance traits contained on plasmids, such as vancomycin resistance, has facilitated the rapid dissemination of AMR genes into the *E. faecium* population.

To study the genetic determinants of antimicrobial resistance in bacterial species and its spread, whole-genome sequencing (WGS) of bacteria by short-read technologies became the gold-standard. However, plasmid sequences frequently contain many repetitive sequences that cannot be spanned by short-read sequences. The resulting assembly of bacterial genomes often consists of hundreds of longer contiguous sequences, termed contigs, from which the chromosomal or plasmid origin remains largely unknown.

At the start of my thesis research, only a limited set of bioinformatics tools was available to predict and reconstruct bacterial plasmids from short-read WGS technologies. In **chapter 2,** we provided one of the first independent benchmarks that highlighted the limitations and challenges behind plasmid reconstruction using short-read WGS data. We concluded that none of the programs was able to accurately reconstruct plasmid sequences in an automated fashion. These results motivated the development of novel tools to improve the prediction and reconstruction of bacterial plasmids.

In **chapter 3,** we developed a user-friendly machine-learning tool called mlplasmids. This tool used pentamer frequencies to predict the total of plasmid sequences, the plasmidome, of a selection of bacterial species. Mlplasmids could confidently predict the replicon origin, plasmid- or chromosome-derived, of a contig sequence.

In **chapter 4,** we performed WGS of a large collection of 1,644 *E. faecium* isolates coming from different origins including hospital, commensal and animal sources. In this study, we used mlplasmids to predict the *E. faecium* plasmidome content which unravelled that hospital isolates carried a larger and distinct plasmid content. We showed that not a single but several plasmid configurations were present in hospital isolates which suggested several routes of plasmid adaptation into the hospital environment. Furthermore, we observed that the content of plasmid sequences was stronger linked to source specificity than the

chromosome content.

The main limitation of mlplasmids is that all sequences predicted to be of plasmid origin remained unbinned, precluding the reconstruction of individual plasmids. This limits its applicability to study the dissemination of individual plasmid sequences into bacterial populations. Therefore, we developed in **chapter 5,** a novel tool termed gplas that separated contigs with a plasmid origin into different bins corresponding to individual plasmid sequences. This tool uses pentamer frequencies, k-mer coverage, and assembly graph information to perform its prediction. We showed that the combination of mlplasmids together with gplas outperformed other binning tools and allowed the reconstruction of single plasmid sequences from short-read WGS data.

In **chapter 6,** we used the newly developed tool gplas to study the dissemination of vancomycin resistance from a clone, plasmid and transposon perspective in a set of 309 vancomycin-resistant *E. faecium* isolates, from patients from 32 Dutch hospitals, carrying the *vanA* gene cluster isolated between 2012-2015. We observed that, on average, clonal dissemination was the major driver of vancomycin resistance spread in Dutch hospitals in this strain set. However, we also identified some hospital outbreaks where HGT, either mediated by plasmid or transposon transfer dominated the spread of the *vanA* gene cluster.

In this thesis I describe the development of two novel tools, mlplasmids and gplas, to predict and reconstruct bacterial plasmid sequences from short-read WGS data. I applied these tools to large collections of mainly *E. faecium* isolates and showed how they were instrumental in disclosing the role of plasmids in adaptation of *E. faecium* to different ecological habitats and in dissemination of vancomycin resistance. Mlplasmids and gplas are not *E. faecium* specific thus can also be applied to study the role of plasmids in bacterial adaptation and resistance dissemination in a wide variety of bacterial species, Gram-positive as well as Gram-negative.

Nederlandse samenvatting

Bacteriën kunnen autonome elementen bevatten, plasmiden genaamd, die als belangrijke aanjagers van variatie en aanpassing kunnen fungeren. Deze plasmide sequenties kunnen de verspreiding van antimicrobiële resistentie genen (AMR) naar andere bacteriesoorten door horizontale genoverdracht (HGT) vergemakkelijken. De reconstructie-plasmide sequenties zijn dus essentieel om volledig te begrijpen hoe de introductie en verspreiding van AMR-genen plaatsvindt in bacteriële populaties.

In dit proefschrift hebben we ons gericht op de nosocomiale pathogeen *Enterococcus faecium* die gewoonlijk in het darmkanaal van zoogdieren voorkomt. *E. faecium* is echter ook een belangrijke veroorzaker van ziekenhuisinfecties. In de afgelopen drie decennia is *E. faecium* resistent geworden tegen meerdere antibiotica, met name ampicilline, gentamicine en vancomycine. De aanwezigheid van vele kenmerken van antibioticaresistentie in plasmide sequenties, zoals vancomycine resistente, zou de snelle verspreiding van AMR-genen in de *E. faecium*-populatie kunnen vergemakkelijken.

Om plasmide sequenties te bestuderen, werd sequentiebepaling van het hele genoom (WGS) van bacteriën door kort gelezen technologieën de gouden standaard. Plasmide-sequenties bevatten echter vaak veel repetitieve sequenties die niet kunnen worden overbrugd door korte-read-sequenties. De resulterende assemblage bestaat vaak uit honderden langere aaneengesloten sequenties, contigs genoemd, waarvan de chromosomale of plasmide-oorsprong onbekend blijft.

Aan het begin van dit proefschrift was er slechts een beperkte set van bio-informatica-instrumenten beschikbaar om bacteriële plasmiden te voorspellen en te reconstrueren op basis van korte-read WGS-technologieën. In **hoofdstuk 2** bieden we een van de eerste onafhankelijke benchmarks die de beperkingen en uitdagingen benadrukken achter plasmide-reconstructie met behulp van short-read WGS-gegevens. We concludeerden dat geen van de programma's in staat was om plasmide-sequenties op een geautomatiseerde manier nauwkeurig te reconstrueren. Deze resultaten motiveerden de ontwikkeling van nieuwe tools om de voorspelling en reconstructie van bacteriële plasmiden te verbeteren.

In **hoofdstuk 3** hebben we een gebruiksvriendelijke machine learning tool ontwikkeld, genaamd mlplasmids. Deze tool gebruikte pentameer frequenties om het plasmidoom gehalte van een selectie van bacteriesoorten te voorspellen. Mlplasmids kon met vertrouwen de oorsprong voorspellen, afkomstig van plasmiden of chromosomen, van een opeenvolgende sequentie. De belangrijkste beperking achter mlplasmids bestond er echter in dat alle sequenties met een plasmide oorsprong losgemaakt bleven en dus de toepasbaarheid ervan om de verspreiding van individuele plasmide sequenties in bacteriële populaties te bestuderen, beperkte.

In **hoofdstuk 4** bestudeerden we een grote collectie van 1.644 *E. faecium*-isolaten van verschillende oorsprong, waaronder ziekenhuis-, commensale en dierlijke bronnen. In deze

studie gebruikten we mlplasmids om het *E. faecium* plasmidoom gehalte te voorspellen, wat ontrafelde dat ziekenhuisisolaten een groter en duidelijker plasmidegehalte droegen. We toonden aan dat er niet één maar meerdere plasmide-configuraties aanwezig waren in ziekenhuisisolaten, wat verschillende routes voor plasmide-aanpassing in de ziekenhuisomgeving suggereerde. Verder hebben we waargenomen dat het gehalte aan plasmide sequenties sterker was gekoppeld aan de bron specificiteit dan het chromosoom gehalte.

In **hoofdstuk 5** hebben we een nieuwe tool ontwikkeld, gplas genaamd, die contigs met een plasmide-oorsprong scheidt in verschillende bakken die overeenkomen met individuele plasmide-sequenties. Deze tool maakt gebruik van pentameer frequenties, k-mer-dekking en informatie over assemblage grafieken om de voorspelling uit te voeren. We toonden aan dat de combinatie van mlplasmiden samen met gplas beter presteerde dan andere binning-tools en maakte de reconstructie mogelijk van enkele plasmide-sequenties op basis van korte WGS-gegevens.

In **hoofdstuk 6** hebben we ons gericht op 309 *E. faecium*-isolaten, afkomstig uit 32 Nederlandse ziekenhuizen, die het *vanA*-gencluster dragen dat resistentie tegen vancomycine verleent. De vanA-gencluster wordt vaak gedragen door plasmide sequenties en wordt gecodeerd binnen een transposon-element. Met gplas konden we de plasmide sequenties reconstrueren die dit resistentie gencluster dragen. Dit maakt het mogelijk om de verspreiding van vancomycine resistente te bestuderen vanuit een kloon-, plasmide- en transposon perspectief. We constateerden dat klonale verspreiding gemiddeld de belangrijkste oorzaak was van de verspreiding van vancomycine resistente in Nederlandse ziekenhuizen (2012-2015). We hebben echter enkele uitbraak instellingen geïdentificeerd waarvoor HGT, ofwel gemedieerd door plasmide- of transposons overdracht. domineerde de verspreiding van het *vanA*-gencluster.

Dit proefschrift leverde twee nieuwe tools, mlplasmids en gplas, om bacteriële plasmide sequenties te voorspellen en te reconstrueren op basis van kort gelezen WGS-gegevens. We hebben deze tools toegepast op grote verzamelingen isolaten en laten zien hoe de bijdrage van plasmide sequenties aan bacteriële variatie en adaptatie kan worden bestudeerd.

Acknowledgments

Anita, I don't know where to start! Thanks for absolutely everything, you have been my supervisor since I came to do my internship and I have always felt that you were always there to support me in all the different stages of these 4 years. I am extremely glad to be your first PhD student, and I also feel very happy to see how your group started growing in the last couple of years. You were always supportive and probably you are one of the biggest reasons why I decided to continue my career in academia. You are a brilliant scientist, supervisor and teacher, I hope to continue my journey following your steps, once again thanks for everything!

Rob, I am also very grateful that you were my promoter during this journey. I loved all the meetings that we got together, you always had a positive energy which surrounded you that makes work a much easier task! Thanks for always keeping an eye on me and give me the time and freedom to develop some of the tools that I present in this dissertation. The perspective that you always brought into the projects have been priceless and I have learnt a lot from you. I couldn't ask for a better supervisor!

Alex, amigo!!, you are one of the best persons that I met during this time. You have always been there, especially at the beginning of my stay in the Netherlands, and I am always happy when I remember all the times that we spent together. Thanks for all the fun nights and dinners together with Maria! I hope we can always keep in touch, either short trips to London or you come to visit whenever I live in the future! You are always welcome in our home :)

Gosia and **Elena!** What a duet of ragazzas! I am super glad that our paths joined in the Netherlands. I only have good memories of all the things that we have lived together, thanks for all the time that we spent at the UMC, either having coffee, lunch or just taking a break from work. Thanks for all the dinners and nights, and for the food and drinks (you know what drink I mean Gosia…). I am looking forward to seeing you soon, I hope we always stay in touch. Na zdrowie!!

Angelino, I loved sitting together with you in the office and enjoyed all the football discussions that we got in the office. I can only smile when I remember all the things that happened in the little office inside the lab, I hope you still keep around that drone. Gracias por todo amigo :)

Yuxi and **Vincent,** my favourite couple of the department! I only have positive words for you, I loved being part of your wedding, and loved every minute of the trip that we did to China! I hope we can see each other soon.

Only a few people will understand the following order. **Paul.** You are one of the craziest but nicest guys I have ever met! Thanks for always bringing such a positive mood wherever you are, you are probably one of the best laughing therapies that I could anyone to

recommend! Tot sinas amigo! :) **Leire,** me alegro mucho de haberte conocido! Al igual que Paul estás un poco loca, pero eres una de las personas más geniales que he conocido. Espero que todo te vaya genial en tu futuro y que puedas volver pronto a Bilbao! Mucha suerte en todo, espero que siempre estemos en contacto! **Jelle,** my Dutch teacher, probably the person that has taught me more random and funny words. I loved all the coffee breaks that we spent together after lunch! Schapen scheren scheveningen and aju paraplu! The other words can't be written here!

Mark D, una de las primeras personas que conocí cuando llegué al departamento. Me alegro mucho haberte podido ayudar en tu carrera yendo a Barcelona y más tarde a México. Eres una gran persona!

Jesse, I am very glad that you are now part of Anita's group. You are a great guy and one of the best team mates that you can always dream about! Thanks for being my paranymph and all the help in this last part of the journey. I hope we always stay in touch :) **Julian!** Que quilombo más bonito habernos cruzado en esta vida! Eres una persona magnífica y me alegro mucho de considerarte mi amigo, espero que siempre estemos en contacto estos años. También muchas gracias por ayudarme en esta recta final! Eres genial michimiau :) **Alessia,** I am very confident that you are going to be an excellent researcher or anything that you propose in your career, I am glad that I could supervise you during your internship and will always make time for coffee if you are around.

Janetta, thanks for all your help and supervising some of the projects from this dissertation. You are an excellent scientist and I have learnt a lot from you, some of the work presented here would not have been possible without you. **Jukka,** I have always admired you as a scientist, and it is been a pleasure collaborating with you in most of the chapters of this thesis. I appreciate a lot all the confidence and support that you are giving me for my next career step. It will be a blast working with you as a postdoc! **Willem,** thanks for being one my first supervisors when I arrived to Utrecht. You are a brilliant scientist and now a full professor in England, thanks also for showing me how useful Twitter can be for sharing science and learning from others in the bioinformatics community.

Malbert, it is always been great working with you, you are an excellent bioinformatician and also a great person, thanks for all the work along these years! **Tess,** I had a lot of fun collaborating with you in mlplasmids, thanks also for all the meetings and chats that we got during these years. I wish you the best in your promising career. **Fernanda,** I always loved chatting with you, thanks for all the tips and talks during these years. I wish you the best to you and your family! **Iris,** thanks for all the brilliant lab work that you did and also for being always extremely helpful whenever I encounter some problems with metadata. I wish you the best in your successful career! **Rodrigo,** you are also one of the persons that

you would always have in your team, I hope we can have some drinks together at some point, and thanks for all your brilliant work.

Rita, I am not sure if I should write this in Spanish because probably you would get most of it! Thanks for all the fun times together and I hope we always keep in touch. **Stephanie!** My favourite singer, it's been great spending time together! **Hendrik,** que pasa cab--- !!! I loved our chats where you tried to show off all your Spanish in one go, I have been always impressed! I hope everything goes 'xaxi', you need to remember that word! Un abrazo :), **Patrique,** thanks a million for all the help with the layout!!! I hope now you understand better the borders of the Mediterranean countries despite we all share basically same the same schedules, food and our languages sound quite similar. **Julia,** espero que todo te vaya genial en el doctorado, mucha suerte con tu proyecto y espero que nos veamos pronto.

Bart, I loved sharing the office with you, and thanks for all the talks and drinks that we got together along these years, I hope you and **Danni** are doing great, and wish you the best with your family :) **Vincent**, **Jery, Axel** and **Roos.** It is been great having you around at work, and also sharing some fun activities outside. I hope you are all doing great and would love seeing each other again at some point.

Sjors, Astrid, Leonardo, Shu, Jiannan, Dennis, Lisanne, Janneke, Priscilla and the rest and former **PhDs** and **postdocs** in our department. It's been a pleasure working in the same department with you, thanks for all the time that we spent together!

Silvia y **Alberto,** muchas gracias por ser las mejores personas que uno puede tener a su lado. Se me hace muy duro a veces no poder estar más con vosotros, espero que en unos años podamos volver de forma definitiva a Barcelona!

Mama y **papa,** gracias por todo el esfuerzo que habéis hecho y la ayuda que siempre nos habéis dado sobretodo al principio cuando vinimos a Holanda! Estoy muy orgulloso de vosotros :)

Maria, crec que res de lo que estic vivint hagués sigut possible sense tu. Gràcies per haver decidit venir a viure a Utrecht amb mi, gràcies per no donar-te mai per vençuda fins que vas trobar feina a Holanda, gràcies per haver trobat un pis en el que podem construir el nostre futur, gràcies per acompanyar-me un cop més a un altre país en els propers anys, gràcies per donar-me tant suport en aquesta recta final. Sóc molt feliç al teu costat i estic enormement feliç de saber que la persona que més estimo en aquest món sempre estarà al meu costat. Espero que en els propers anys puguem tornar a Barcelona i construir definitament el nostre futur a la nostra terra, t'estimo molt! :)

About the author
List of publications

About the author

Sergio Arredondo Alonso was born on the 1st of April 1993, in Sant Joan Despí, Barcelona, Spain. In 2011, he started his Bachelor studies in Microbiology at the Universitat Autònoma de Barcelona. During his bachelor's degree, he joined the laboratory of Dr. Isidre Gibert and Dr. Daniel Yero at the Institut de Biotecnologia and Biomedicina, and worked on the construction of knock-out strains for studying virulence regulation in *Pseudomonas aeruginosa*. During his undergraduate studies, he developed a strong interest in microbial genomics and, in 2016, he started his Master studies in Bioinformatics at the Universitat Autònoma de Barcelona. During his master studies, he obtained an Erasmus+ grant to perform an internship for his master's thesis in the Department of Medical Microbiology at the UMC Utrecht, under the supervision of Dr. Anita C. Schürch and Prof. Willem van Schaik. The topic of the master's thesis was to perform a comprehensive comparison of bioinformatic approaches to reconstruct plasmid sequences. In September 2016, he obtained his master's degree, and started in October 2016 his PhD studies in the group of Prof. Rob Willems, under the supervision of Dr. Anita C. Schürch. In the last 4 years, he has performed the work presented in this thesis and most parts have been published in international scientific journals. After his PhD promotion, he will continue as a Postdoctoral fellow at the University of Oslo under the supervision of Prof. Jukka Corander.

List of publications

Arredondo-Alonso, Sergio, Rob J. Willems, Willem van Schaik, and Anita C. Schürch. 2017. "On the (im)possibility of Reconstructing Plasmids from Whole-Genome Short-Read Sequencing Data." Microbial Genomics 3 (10). doi: 10.1099/mgen.0.000128.

Schürch, A. C., **S. Arredondo-Alonso**, R. J. L. Willems, and R. V. Goering. 2018. "Whole Genome Sequencing Options for Bacterial Strain Typing and Epidemiologic Analysis Based on Single Nucleotide Polymorphism versus Gene-by-Gene–based Approaches." Clinical Microbiology and Infection 24 (4). doi: 10.1016/j.cmi.2017.12.016.

Arredondo-Alonso, Sergio, Malbert R. C. Rogers, Johanna C. Braat, Tess D. Verschuuren, Janetta Top, Jukka Corander, Rob J. L. Willems, and Anita C. Schürch. 2018. "Mlplasmids: A User-Friendly Tool to Predict Plasmid- and Chromosome-Derived Sequences for Single Species." Microbial Genomics 4 (11). doi: 10.1099/mgen.0.000224.

Arredondo-Alonso, S., J. Top, A. McNally, S. Puranen, M. Pesonen, J. Pensar, P. Marttinen, et al. 2020. "Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus Faecium*." mBio 11 (1). doi: 10.1128/mBio.03284-19.

Arredondo-Alonso, Sergio, Martin Bootsma, Yaïr Hein, Malbert R. C. Rogers, Jukka Corander, Rob J. L. Willems, and Anita C. Schürch. 2020. "Gplas: A Comprehensive Tool for Plasmid Analysis Using Short-Read Graphs." Bioinformatics , 36 (12). doi: 10.1093/bioinformatics/btaa233

Arredondo-Alonso, Sergio, Janetta Top, Jukka Corander, Rob J. L. Willems, and Anita C. Schürch. 2020. "Mode and Dynamics of *vanA*-Type Vancomycin-Resistance Dissemination in Dutch Hospitals." Submitted for publication. Available as a preprint in medRxiv. doi: 10.1101/2020.07.21.20158808