Using bacterial DNA sequencing data to investigate the epidemiology of plasmid-mediated antibiotic resistance

Alex Orlek

Wolfson College

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy in Clinical Medicine

Trinity Term 2020

Acknowledgements

I'm grateful to my supervisors and to my funding body for giving me the opportunity to carry out research in such a dynamic and exciting field. Sarah Walker has been perhaps my closest scientific mentor and has supported me with insightful feedback throughout the project; with Sarah's statistical expertise, complex statistical challenges became surmountable. Thanks to Anna Sheppard for her steadfast support; she has contributed her knowledge of bioinformatics and bacterial genetics, and provided insightful feedback on written work. Muna Anjum (Animal and Plant Health Agency) has contributed her knowledge of antibiotic resistance, particularly in the context of antibiotic usage/resistance in livestock; I'm grateful for her enthusiastic support, including helpful feedback on written work. I've been lucky to have the support of Nicole Stoesser who, throughout the project, generously provided me with helpful advice, and feedback on written work, even though she is not officially listed as a supervisor. Thanks also to Hang Phan who helped during the earlier stages of the project while she was based in Oxford, and to Alison Mather (Quadram Institute) for advice and encouragement during the later stages of the project.

Outside of PhD work, I'm grateful to Ade Oladeinde (USDA, Georgia, USA) who collaborated with me on a paper (Oladeinde et al., 2018; see below) and made me feel at home while I was in Atlanta for the ASM Microbe 2018 conference. Thanks to my colleagues from the Modernising Medical Microbiology group, for cake, lunchtime chats, and journal club discussions. Last but certainly not least, I'm grateful to my family and friends for their unstinting support.

Funding

The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) [grant HPRU-2012-10041].

Declaration and attributions

I, Alex Orlek, designed and conducted all the analyses presented in this thesis with the support of my supervisors. I undertook the literature searching and writing for all chapters; I implemented amendments to draft chapters following feedback from supervisors. More detailed attributions for each research chapter are given below.

Chapter 1

The idea to write a review article on WGS-based plasmid analysis was suggested by Sarah. This led me to successfully publish a review article in *Frontiers in Microbiology* (see below). I wrote the manuscript for this article which was modified in line with comments from my supervisors and Nicole Stoesser. This published article formed the basis for Chapter 1.

Chapter 2

The idea to retrieve NCBI plasmids was part of an original project plan devised by my supervisors before I joined the group. However, it was through my work and scrutiny that the importance of curating plasmid sequences became apparent. I designed and conducted subsequent analyses to investigate the performance of plasmid typing schemes, with the support of my supervisors. Anna

suggested that I submit this work as an article to the *Plasmid* journal, and this was successfully published as the "Ordering the mob…" article (see below). I wrote the manuscript for this article which was modified in line with comments from my supervisors and Nicole Stoesser. This published article formed the basis for Chapter 2. I conceived and conducted all work relating to the development of the bacterialBercow software.

Chapter 3

I developed the reference-based and replicon-based pipelines for plasmid analysis. With respect to the Manchester outbreak samples, as well as the REHAB and University of Virginia outbreak samples (used for benchmarking analysis), all sample collection and wet lab sequencing work was conducted by collaborators.

Chapter 4

All Manchester outbreak sample collection and wet lab sequencing work was conducted by collaborators. Assembled PacBio-sequenced samples were provided by Hang Phan. Hang Phan also conducted *in silico* species identification using Kraken, and ran BLAST-based methods for *in silico* multi-locus sequence typing (MLST), and for typing the bla_{KPC} gene and the Tn*4401* transposon. All other sequence data processing and analysis work is my own. The ATCG software development was entirely my own work. Nicole Stoesser encouraged me to explore methods for assessing plasmid structural similarity, which led me to implement breakpoint distance calculation as part of the ATCG software. The "Comparative Genomics for Prokaryotes" book was helpful in giving me a comprehensive introduction to comparative genomics methods. The "Bioinformatics Data Skills" book introduced me to the GenomicRanges R package which was central to ATCG software development (book references are listed below).

Chapter 5

Sarah guided me towards conducting non-linear multivariate logistic regression and provided expert statistical guidance throughout. I undertook to learn about GAM modelling in R, and all modelling work and biological interpretation was my own. I realised the importance of addressing issues of pseudoreplication, and also realised the importance of metadata curation.

Finally, I am indebted to countless learning resources; a selection is listed below.

Buffalo, V. (2015). *Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools*. Sebastopol, California, USA: O'Reilly Media.

Chang, W. (2013). *R Graphics Cookbook: Practical Recipes for Visualizing Data*. Sebastopol, California, USA: O'Reilly Media.

Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. London, UK: SAGE Publications Ltd.

Ross, N. (2019). GAMs in R: A Free, Interactive Course using mgcv. Available at: https://noamross.github.io/gams-in-r-course/.

Setubal, J. C., Almeida, N. F., and Wattam, A. R. (2018). "Comparative Genomics for Prokaryotes," in *Comparative genomics: Methods and Protocols*, eds. J. C. Setubal, J. Stoye, and P. F. Stadler (New York, NY: Humana Press).

Papers arising directly from work presented in this thesis

The published articles listed below relate to work presented in this thesis. Orlek et al. 2017c was used as a basis for the thesis Introduction (Chapter 1). Orlek et al. 2017a was used as a basis for Chapter 2; Orlek et al. 2017b also relates to work presented in Chapter 2. Text from the published articles is reproduced with editing in the corresponding thesis chapters. All co-authors gave permission for me to re-use text from the articles in my thesis.

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017a). A curated dataset of complete Enterobacteriaeceae plasmids compiled from the NCBI nucleotide database. *Data Br.*

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017b). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* 91, 42–52. doi:10.1016/j.plasmid.2017.03.002.

Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., and Ellington, M. (2017c). Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front. Microbiol.* doi:doi: 10.3389/fmicb.2017.00182.

Other papers arising during the course of this thesis

Oladeinde, A., Cook, K., Orlek, A., Zock, G., Herrington, K., Cox, N., et al. (2018). Hotspot mutations and ColE1 plasmids contribute to the fitness of Salmonella Heidelberg in poultry litter. *PLoS One*. doi:10.1371/journal.pone.0202286.

Stoesser, N., Phan, H. T. T., Seale, A. C., Aiken, Z., Thomas, S., Smith, M., et al. (2020). Genomic epidemiology of complex, multispecies, plasmid-borne blaKPC carbapenemase in Enterobacterales in the United Kingdom from 2009 to 2014. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.02244-19.

Other academic achievements

I was awarded an Outstanding Abstract Award for an abstract I presented as a poster at the ASM Microbe 2018 conference. The work I presented at the conference forms the basis of Chapter 3.

Abstract

Using bacterial DNA sequencing data to investigate the epidemiology of plasmid-mediated antibiotic resistance

Alex Orlek, Wolfson College, University of Oxford, Nuffield Department of Medicine
A thesis submitted for the degree of Doctor of Philosophy in Clinical Medicine, Trinity Term 2020

Bacterial plasmids are extra-chromosomal genetic elements, which can act as efficient vectors of antibiotic resistance. Epidemiological insight into plasmids may be gained by applying plasmid typing schemes, which exploit loci involved in replication and mobility functions (replicon and MOB typing, respectively). In Chapter 2, I compiled a curated dataset of complete NCBI plasmids to assess the performance of *in silico* replicon and MOB typing in terms of concordance and 'typeability' (proportion of plasmids typed). I found a degree of non-concordance between the schemes, which was attributed to either ambiguous boundaries between MOBP/MOBQ types, or the mosaic nature of some plasmid genomes. Ultimately, I showed that the schemes fail to accommodate the diversity of plasmid genomes; of ~14000 curated bacterial plasmids, only 42% and 55% could be assigned a replicon and MOB type, respectively. Given the limitations of plasmid typing, I subsequently focused on whole genome sequencing (WGS) analysis approaches capitalising on the wider plasmid genome. High-throughput DNA sequencing has produced 1000s of bacterial WGS datasets. However, such datasets commonly comprise short sequencing reads, which yield fragmented assemblies; this makes comparative analysis of plasmid genomes challenging. In Chapter 3, I developed two methods for comparative plasmid analysis, which cluster short-read sequenced samples according to 1) plasmid replicon types; 2) sample-vs-reference plasmid distance score profiles. However, benchmarking suggested neither method is completely reliable. The rise of long-read sequencing technology has increased the availability of complete plasmid assemblies, facilitating comparative plasmid genomic analyses. Nevertheless, available alignment-based comparative genomic tools have limitations: they often do not provide metrics on structural similarity and lack flexibility in terms of input/output options. Therefore, in Chapter 4, I developed a novel alignment-based tool ('ATCG') for calculating pairwise average nucleotide identity (ANI), coverage breadth, and structural similarity, while addressing limitations of existing alignment-based tools. Benchmarking demonstrated favourable runtimes and supported the validity of calculated ANI scores. In Chapter 5, besides curating an updated plasmid dataset, I curated sample metadata (e.g. isolation source, geography). Using this metadata and plasmid biological features, I conducted multivariate statistical analyses to determine factors associated with plasmid resistance gene carriage, analysed across major resistance gene classes. The analysis yielded interesting findings, for example, demonstrating that patterns of plasmid antibiotic resistance carriage in livestock and humans reflect known antibiotic usage.

Contents

1 Introduction

1.1 Plasmids as vehicles of antibiotic resistance

Sexual reproduction in eukaryotes, involving fusion of gametes between individuals of the same species, generates genetic diversity on which natural selection can act; in contrast, bacteria reproduce through an asexual process called binary fission (Narra and Ochman, 2006). Therefore, one might expect bacteria to evolve relatively slowly, reflecting incremental acquisition of mutations, inherited vertically along asexual lineages. Yet, many bacteria show remarkable abilities to evolve adaptively in response to environmental challenges such as antibiotic resistance exposure – an apparent contradiction which puzzled early microbiologists (Snyder et al., 2013a). For example, in the 1950s, outbreaks of bacterial dysentery in Japan developed resistance to multiple antibiotics far more rapidly than could be explained solely by vertical inheritance of *de novo* resistance mutations (Watanabe, 1967). At the time, researchers were beginning to understand that genetic material can transfer intercellularly between bacteria (Avery et al., 1944; Lederberg and Tatum, 1946; Zinder and Lederberg, 1952) – a phenomenon called horizontal gene transfer (HGT) – and they realised that HGT was driving resistance dissemination during the dysentery outbreaks. We now know that the mechanism of HGT in this case was the transfer of plasmids carrying antibiotic resistance genes (Lederberg, 1998).

Plasmids are extra-chromosomal genetic elements, found ubiquitously in bacteria. They are often transmissible between host cells (as described below), but can also control their replication, to maintain a stable copy number within the host cell, enabling vertical inheritance. Stable vertical inheritance may also be promoted by specific stable inheritance mechanisms (Snyder et al., 2013b), as well as plasmid-host chromosome co-evolution, which can alleviate fitness costs of plasmid carriage (San Millan and MacLean, 2017). Plasmid genomes generally include a 'backbone' of core genetic loci, which are somewhat conserved amongst broadly related plasmids (Phan et al., 2009), and associated with key plasmid-specific functions such as replication and transfer. 'Accessory' genes may also be

present, and often confer clinically relevant traits such as virulence and antibiotic resistance (Thomas and Summers, 2008). Through conjugation, plasmids can act as efficient vehicles of HGT. Notably, conjugative plasmids encode all the molecular machinery required for self-transfer: a relaxase, a type 4 coupling protein (T4CP), and a type 4 secretion system (T4SS) comprised of mating pair formation (MPF) proteins. The relaxase recognises a plasmid sequence called the origin of transfer (*oriT*) and introduces a single-stranded break at a *nic* site within the *oriT*. The resulting complex of single-stranded DNA and relaxase, is delivered to the T4SS (facilitated by the T4CP), where it is translocated from recipient to donor cell, via a mating pore; concurrent replication of the two single-stranded plasmid molecules regenerates double-stranded plasmids within donor and recipient cells (Smillie et al., 2010; Guglielmini et al., 2014; Kohler et al., 2019). Plasmids lacking complete conjugative machinery can still transfer by conjugation, so long as functions are provided *in trans* by a co-resident plasmid (Ramsay et al., 2016).

Plasmids are the most well-studied conjugative elements. Other genetic elements called integrative and conjugative elements (ICEs) can also transfer intercellularly by conjugation; as the name suggests, they generally integrate into the host chromosome, rather than replicating extra-chromosomally, as plasmids do (Johnson and Grossman, 2015). Besides conjugation, other major mechanisms of HGT include transformation and transduction (Frost et al., 2005). Transformation involves uptake of free DNA from the extracellular environment, usually mediated by chromosomally-encoded proteins expressed by a 'competent' recipient cell (Blokesch, 2016). Transduction is mediated by phages; during infection, they may incorporate host bacterial DNA, which can then be transferred to a new host cell, by injection of incorporated donor host DNA (generalised transduction) or by integration of a hybrid lysogenic phage (specialised transduction) (Chiang et al., 2019). Transformation and transduction can mediate transfer of chromosomal and plasmid DNA, and sometimes complete plasmids can be transferred by these mechanisms (Keen et al., 2017; Hasegawa et al., 2018; Moreno-Switt et al., 2019). All described HGT mechanisms can drive the spread of resistance genes. However, numerous authors have speculated that conjugation is generally the most important driver of

resistance gene dissemination, given limitations associated with the other HGT mechanisms. (Norman et al., 2009; Von Wintersdorff et al., 2016; Lerminiaux and Cameron, 2019). For example, HGT by transformation is limited by the requirement for a competent recipient, as well as degradation of extracellular DNA, whereas HGT by transduction is limited by the tendency of phages to exhibit high host range specificity (Popa et al., 2017). In contrast, some plasmids have broad host taxonomic range – capable of transferring between different species or even different phyla (Jain and Srivastava, 2013; Klümper et al., 2014). Furthermore, plasmids can be inherited vertically with their hosts. However, stable inheritance following transformation or generalised transduction – except in the case of complete plasmid transfer – depends on integration of transferred DNA fragments into the chromosome, for example, by homologous recombination (Thomas and Nielsen, 2005). Ultimately, the relative importance of different HGT drivers remains poorly understood, and is likely to vary according to different species and habitats (Hall et al., 2017). What is clear is that plasmids can act as efficient vehicles of resistance dissemination (see below), and overall, they play a central role in facilitating horizontal gene flows, as supported by genome analysis: Halary et al. built a network comprising plasmid, phage, and chromosomal genomes, linked by gene-sharing. They found that compared to phage genomes, plasmids were more likely to act as bridges linking disparate parts of the gene-sharing network, presumably reflecting the efficiency with which many plasmids can shuttle genetic content across taxonomic boundaries (Halary et al., 2010).

Plasmid genomes often contain nested mobile genetic elements, such as transposons and integron-associated gene cassettes, which can harbour resistance genes (Partridge et al., 2018). Transposons are transposable elements which can mobilise themselves and associated cargo genes to new genetic locations within the cell by transposition, which occurs by a cut-and-paste or replicative mechanism (Snyder et al., 2013c). Transposons may be composed of smaller transposable elements called insertion sequences (ISs) which flank cargo genes (this is called a composite transposon) (Siguier et al., 2015). Integrons are genetic platforms which catalyse the insertion and excision of mobile gene cassettes, located within the integron, by site-specific recombination; excised cassettes form circular

intermediates which may reinsert at a new site within the same integron, or within a different integron (Bennett, 2008; Gillings, 2014). Therefore, not only are resistance genes spread between bacteria by virtue of being located on transmissible plasmids, they can also spread intracellularly to co-resident plasmids or the chromosome (He et al., 2016; Xie et al., 2018). The dissemination of resistance genes via resistance plasmids and their nested mobile genetic elements is a key driver of the antibiotic resistance crisis, which now threatens modern medicine (Carattoli, 2013; World Health Organization, 2014).

Following the widespread application of antibiotics in clinical and agricultural contexts, plasmids have accumulated resistance genes, which are often found clustered together in multidrug resistance regions (Stokes and Gillings, 2011). Plasmids now commonly confer resistance to major classes of clinically important antibiotics, including beta-lactams, aminoglycosides and quinolones (Bennett, 2008; Carattoli, 2013). They drive resistance dissemination in certain clinically important Gram-positive taxa such as *Enterococcus* and *Staphylococcus* (Chancey et al., 2012), as well as various Gram-negative taxa, including *Acinetobacter* and *Pseudomonas* (Livermore and Woodford, 2006). Plasmid-mediated antibiotic resistance is of particular concern among members of the Gram-negative Enterobacteriaceae family, which includes clinically important taxa such as *E. coli* and *Klebsiella* (Iredell et al., 2016). Plasmid-encoded extended-spectrum beta-lactamase resistance (ESBL) genes such as those of the CTX-M type can confer resistance to penicillins and cephalosporins, whilst carbapenemase genes such as those of the KPC type can confer resistance to carbapenems – beta-lactams of last-resort (Livermore and Woodford, 2006; Nordmann et al., 2012). The efficacy of another last resort antibiotic called colistin has been jeopardised by the recent emergence of plasmid-mediated colistin resistance in Enterobacteriaceae (Liu et al., 2016; Wang et al., 2018). Once acquired by a pathogenic bacterium, a resistance plasmid can drive the success of the host strain (Holt et al., 2013), which may lead to the emergence of a clonal population of resistant bacteria, associated with infections that may be more difficult to treat (de Kraker et al., 2011; Nathwani et al., 2014). Whilst clonal spread of a resistance plasmid-carrying strain may be important, studies indicate that plasmids

may transmit between strains frequently, even over short timescales (Conlan et al., 2014; Sheppard et al., 2016). In this case, the chain of transmission no longer simply corresponds to strain transmission; resistance plasmid dissemination across strains recruits different recipient strains into the outbreak too, resulting in a 'plasmid outbreak'. Therefore, understanding the epidemiology of bacterial resistance requires examination of plasmids as well as host strains (Adler and Carmeli, 2011).

1.2 Investigating resistance plasmid epidemiology using plasmid typing

One widely used approach to gain insight into plasmid-mediated antibiotic resistance is to classify plasmids according to a typing scheme: for example, studying the composition of plasmid types can indicate whether an antibiotic resistance epidemic is driven by diverse plasmids or one dominant plasmid type (Valverde et al., 2009). In addition, typing plasmids harboured by strains from a resistance outbreak can inform hypotheses about resistance transmission (Pecora et al., 2015); although evidence from typing alone is weak and inconclusive (see below). The principal plasmid typing schemes are replicon and MOB typing, which can be carried out using PCR-based methods or through *in silico* analysis of DNA sequence data. For a detailed outline of plasmid typing schemes and associated methods, see Table 1. Briefly, replicon typing exploits genetic elements of the replicon region (encoding replication machinery) to classify plasmids (Carattoli et al., 2005, 2014). MOB typing exploits the conserved N-terminal sequence of the relaxase proteins encoded by transmissible plasmids (Garcillán-Barcia et al., 2009). Whilst these single-locus typing schemes have been widely and successfully applied, they provide limited resolution (Fricke et al., 2009), restricting epidemiological inference: in an outbreak context, if two patients are infected by unrelated strains harbouring resistance plasmids of the same type, this raises the possibility of plasmid transmission, but plasmid transmission cannot be conclusively ruled-in using single-locus plasmid typing alone; further higher-resolution investigation would be required (Foxman et al., 2005). If resistance plasmids

are unrelated, a direct plasmid transmission link can be ruled-out, though an indirect transmission link (e.g. via resistance gene transposition) is possible.

Plasmid typing may provide a stepping-stone to higher resolution analyses; identifying shared ('core') genes amongst related plasmids can inform development of plasmid multi-locus sequence typing (pMLST) schemes (Carattoli and Hasman, 2020). However, even pMLST schemes analyse only a small fraction of total plasmid DNA sequence. Whole-genome sequencing (WGS) data can now be obtained for many bacterial isolates, at relatively low cost, within short timescales (Metzker, 2010). WGS data potentially enables higher-resolution plasmid relationships to be determined, through analysis of the wider plasmid genome. High-throughput bacterial WGS can be achieved using next-generation sequencing (NGS) technologies, conducted using either second- or third-generation sequencing platforms (Goodwin et al., 2016). Second generation sequencers lack the sensitivity required to analyse single molecules, so instead rely on clonal amplification of template DNA; this approach can only generate short reads (50–500 bp) due to progressive signal degradation. On the other hand, third-generation sequencers use a single molecule sequencing approach, which does not depend on clonal amplification, and can therefore generate long reads (~1–100 kb) (Ameur et al., 2019). The implications of short- and long-reads for downstream bioinformatic analysis are discussed in detail later in the chapter.

Using bacterial WGS data, researchers can conduct chromosomal core-genome SNP analysis, which has proven instrumental for high-resolution analysis at the strain-level (Croucher and Didelot, 2015). Unfortunately, determining high-resolution plasmid relationships using WGS is more challenging. Plasmid genomes often exhibit high plasticity, since their nested mobile genetic elements promote rearrangements through replicative transposition and homologous recombination (Stokes and Gillings, 2011; Conlan et al., 2016; He et al., 2016). The tendency of plasmids to gain, lose and rearrange genetic content means sets of plasmids – even if of the same type – will tend to share few phylogenetically concordant core genes (Fondi et al., 2010; Tazzyman and Bonhoeffer, 2014),

impeding subtyping and phylogenetic analysis (Maiden, 2006). Even backbone genes may not be well conserved across all plasmids of the same type (Lanza et al., 2014), and sometimes show mosaic phylogenetic origins (Sen et al., 2013). Nevertheless, a recent study indicates that in certain cases, plasmid phylogenetics is possible: Wang et al. performed the first large-scale phylogenetic analysis of plasmids, focusing on the plasmid-mediated spread of the *mcr-1* gene which confers colistin resistance (Wang et al., 2018). The *mcr-1* gene emerged in plasmids recently (estimated dates of origin in two major plasmid types are 2006 and 2011), and the transposon containing the *mcr-1* gene tends to stabilise over time through IS deletion. As a result, conserved sequence content surrounding the *mcr-1* gene could be identified; however, plasmid collections with longer and more complex evolutionary histories are less amenable to phylogenetic analysis. The challenge of plasmid phylogenetics, means that plasmid typing remains a widely used approach, even in an era of WGS.

In the rest of this chapter I examine WGS-based plasmid analysis and its application in studies of antibiotic resistance epidemiology in more depth. I focus on analysis of cultured rather than metagenomic samples, since this reflects the focus on my thesis as a whole. For a discussion of metagenomic WGS analysis see recent reviews (Martínez et al., 2017; Browne et al., 2020). WGS datasets offer exciting opportunities for large-scale plasmid analysis, but popular high-throughput sequencers produce short reads, presenting the additional challenge of assembling reads to resolve individual plasmid structures. Firstly, I assess current approaches for plasmid analysis using short-read WGS data and highlight novel algorithms for improving the reconstruction of plasmids from short-reads. Recently, researchers have been turning to long-read sequencers – the long reads they produce make it easier to generate accurate complete plasmid sequences, because the long reads can span repetitive regions. Towards the end of the chapter, I explore how analysis of complete plasmid genomes can provide improved insights into plasmid epidemiology. For example, comparing complete plasmid structures from a given study site can elucidate the shorter-term dynamics of plasmid evolution and transmission, even though longer-term phylogenetic relationships remain obscured.

Table 1. Summary of plasmid typing and subtyping schemes

Replicon typing schemes	
Inc grouping	Plasmids with similar replication machinery are often unable to stably co-exist within the same host cell (Snyder et al., 2013b); this phenomenon was traditionally used to classify plasmids into incompatibility (Inc) groups. Inc grouping has been applied to plasmids from Enterobacteriaceae (Hedges and Datta, 1973), *Pseudomonas aeruginosa*, and *Staphylococcus aureus* (Taylor et al., 2004).
Replicon probe hybridization	Couturier et al. (1988) cloned replicons representing Enterobacteriaceae Inc groups; plasmids were classified according to Southern blot hybridizations using the replicons as probes. Probe hybridization lacks specificity when closely related replicons are present, and is no longer widely used except for its application subsequent to PCR-based replicon typing (PBRT); here, amplicons derived from PCR can be used as probes to type plasmids isolated on a gel (EFSA, 2011).
PCR-based replicon typing (PBRT)	PBRT for plasmids of the well-studied Enterobacteriaceae family detects replicons based on various genetic loci, notably replicase (*rep*) genes, and replication regulatory sequences. PBRT types roughly correspond to traditional Inc groups, so Inc nomenclature is still used (Carattoli et al., 2005). A commercial Enterobacteriaceae PBRT kit is available, and currently detects 30 replicons using multiplex PCRs (Diatheva, 2020). PBRT has also been devised for *Acinetobacter baumannii* plasmids (Bertini et al., 2010); multiplex PCRs targeting 27 replicons are used to classify plasmids into 19 'GR' types. A PBRT scheme has also been applied to plasmids of gram-positive taxa, focusing on enterococcal (Jensen et al., 2010) and staphylococcal (Lozano et al., 2012) plasmids. A closely related scheme focuses on plasmids of *Enterococcus faecium* (Rosvoll et al., 2010, 2012).
Replicon subtyping	Allelic profiles are assessed at 2–6 core loci (depending on the specific scheme). Plasmids are assigned a pMLST subtype nesting within the broader replicon type. pMLST schemes are available for six common replicon types of Enterobacteriaceae plasmids (IncF, HI1, HI2, I1, N, A/C). PCR-based and *in silico* methods are available (Carattoli et al., 2014).
In silico replicon typing/subtyping	Replicon and pMLST allele databases can be downloaded for local use, but user-friendly web-tools (PlasmidFinder/pMLST for replicon typing/subtyping) can run the analysis pipeline, including read assembly. The PlasmidFinder replicon database currently (Jan 2020) contains 139 reference replicons for Enterobacteriaceae plasmids; a dataset of replicons for gram-positive plasmids based on the schemes devised by Jensen et al. (2010) and Lozano et al. (2012) is also available (Carattoli and Hasman, 2020). Instead of relying on read assembly and BLAST, unassembled reads can be mapped to the PlasmidFinder database or pMLST database using SRST2 (Inouye et al., 2014).
MOB typing schemes	
PCR-based MOB typing	PCR-based 'degenerate primer MOB typing' (DPMT) is used to type γ-Proteobacterial plasmids; 19 degenerate primer pairs target relaxase sequences to partition plasmids into five of the main MOB types identified by *in silico* MOB typing (Rose et al., 1998; Alvarado et al., 2012). PCR-based MOB typing has also been demonstrated for enterococcal plasmids (Goicoechea et al., 2008; Freitas et al., 2016).
In silico MOB typing	The relatively conserved N-terminal relaxase sequences are used as PSI-BLAST or profile Hidden Markov Model (HMM) queries to detect relaxase sequences of transmissible plasmids, and MOB types (Garcillán-Barcia et al., 2009). Currently available software tools partition plasmids into up to 9 MOB families (Cury et al., 2020; Garcillan-Barcia et al., 2020)
Other plasmid typing schemes	

Schemes targeting loci other than replicon/relaxase	Other locus-based methods for plasmid typing and subtyping exist, but tend to be applicable to a more restricted set of plasmids (Guglielmini et al., 2011; Freitas et al., 2013; Compain et al., 2014; Dealtry et al., 2014a, 2014b; Bousquet et al., 2015).
Plasmid 'fingerprinting' (RFLP typing)	Restriction fragment length polymorphism (RFLP) is sometimes used to subtype plasmids, especially when pMLST is unavailable. However, band patterns can be difficult to interpret, and do not provide a reliable phylogenetic marker (Laguerre et al., 1992). Shearer et al. (2011) used RFLP to assign a subset of conserved staphylococcal plasmids to three major RFLP types.

1.3 Short-read WGS data for plasmid analysis: opportunities and challenges

When analysing whole genomic DNA, limited information can be derived from plasmid typing alone: bacterial cells may contain multiple different plasmids, and a single plasmid may contain multiple replicons, obscuring correspondence between detected replicons and the set of plasmid types within a host cell (Johnson et al., 2007). Therefore, in PCR-based studies – where genomic context of an amplicon remains unknown – plasmids are commonly isolated first, before being individually characterised by replicon and resistance typing (see Table 11 in EFSA, 2011). This is time-consuming, restricts the number of isolates that can be analysed, and has inherent limitations: if plasmids from the same isolate are of similar size they cannot be separated by pulsed-field gel electrophoresis, and isolation by transfer to recipient cells is not always achieved (Dib et al., 2015).

Potentially, WGS enables *in silico* analyses in which plasmid typing and analysis of loci of interest, such as resistance genes (and their plasmid or chromosomal genetic context), are performed in a unified way. Consequently, much larger isolate collections can be analysed – a key advantage given that plasmid studies have indicated a need for including more strains in analyses (e.g. environmental strains, or isolates exhibiting only low-level resistance) to uncover more complex transmission routes (Carrër et al., 2010; Stoesser et al., 2014, 2015a). Unfortunately, there are significant obstacles to *in silico* analysis of WGS data. Popular high-throughput sequencing technologies (e.g. Illumina) produce short (50–500 bp) reads, for which assembly is inherently challenging (Nagarajan and Pop, 2013). Isolating individual plasmids prior to sequencing simplifies assembly, potentially enabling complete

plasmid reconstruction (Mathers et al., 2015), but is laborious. Long-read sequencing vastly simplifies assembly and is becoming increasingly cost-effective (Goldstein et al., 2019; Taylor et al., 2019). However, high costs have – at least until very recently – restricted its use, so the bulk of currently available bacterial WGS datasets comprise short-reads (Koren and Phillippy, 2015). Hence, a major challenge lies in extracting useful information from short sequencing reads derived from different sources (different co-resident plasmids, the chromosome).

1.3.1 Reference-based read mapping versus *de novo* read assembly

Having obtained short reads from WGS of isolate DNA, reads are generally mapped to a reference and/or assembled *de novo* using a de Bruijn graph assembler (Zerbino and Birney, 2008; Compeau et al., 2011) (for detailed workflows see Edwards and Holt, 2013; Lynch et al., 2016). Reference-based read mapping is a fast and accurate method to characterise SNPs and detect loci of interest (Li and Durbin, 2009). Identified core genome SNPs can be used to construct strain phylogenies. In addition, rapid epidemiological surveillance of replicons and resistance genes can potentially be achieved with read mapping tools such as SRST2 (Inouye et al., 2014). However, the read mapping approach is limited when structural information is of interest; for example, when assessing whether a detected resistance gene is located on chromosome or plasmid, and if the latter, assessing with which plasmid type it is associated. One approach to overcome this is *de novo* read assembly, which can be less sensitive for SNP or locus detection (Inouye et al., 2014), but provides reference-free structural information, and can identify loci not represented on available references.

1.3.1.1 Determining the genetic context of resistance genes from *de novo* assemblies

Sometimes, short reads can be assembled into complete plasmid structures, and plasmid-localized resistance genes can then be identified and linked to plasmid types through *in silico* analysis

(Oladeinde et al., 2018); however, this is frequently not the case (Arredondo-Alonso et al., 2016). Notably, presence of multiple copies of the same repeat structure – a common situation in plasmid genomes – introduces assembly ambiguity, which can fragment assemblies (Pevzner et al., 2001; Treangen and Salzberg, 2011). Paired-end sequencing data can resolve repeat location, but only if the paired reads span the length of the repeat. Resistance genes are frequently flanked by repetitive mobile elements, and are therefore prone to poor assembly, obscuring their genetic context. There are several tools which can help resolve the location of specific loci of interest (Holt, 2015). Bandage allows visualization and annotation of the assembly graph; for example, users can zoom to unresolved repeat regions and BLAST search connecting contigs (Wick et al., 2015). If all connecting contigs match either plasmid or chromosomal references, then an ambiguously-linked region can be assigned accordingly. For example, Bandage helped to reveal diverse plasmid contexts for the *mcr-1* colistin resistance gene in UK clinical isolates (Doumith et al., 2016). However, this manual approach is unfeasible when analysing large datasets.

1.3.1.2 Reference-based mapping to track plasmid transmission during short-term outbreaks

As well as detecting loci of interest, reference-based mapping has been used to track plasmids during short-term outbreaks. Specifically, as part of a standard bioinformatic workflow, plasmid(s) from an index case isolate are fully assembled, often through long-read sequencing. Short reads or contigs from subsequent isolates are then mapped to the index plasmid, which is deemed present if sufficient homology is demonstrated (Mathers et al., 2015; Pecora et al., 2015; Stoesser et al., 2015b). This approach implicitly assumes that the index plasmid is important throughout the study period, and that plasmid structures are relatively conserved in the short-term. In some cases, these assumptions may hold (Stoesser et al., 2014), but other studies show major structural changes can occur, including recombination of large segments (Conlan et al., 2016) as well as mobilization of resistance genes (Sheppard et al., 2016). Crucially, Sheppard et al. demonstrated that reference-based mapping can be

misleading if plasmid plasticity is high. Specifically, a reference *bla*$_{KPC}$ resistance plasmid from the index isolate was detected across diverse strains by a contig-alignment approach, leading to the initial interpretation that resistance had spread via the original *bla*$_{KPC}$ plasmid. Instead, long-read sequencing showed that often the *bla*$_{KPC}$ gene was actually present on a co-resident plasmid, and the original plasmid lacked the *bla*$_{KPC}$ gene, suggesting that mobilization of *bla*$_{KPC}$ had recruited diverse plasmids to the outbreak. It was usually not possible to determine the genetic context of *bla*$_{KPC}$ with short reads due to long repetitive flanking sequences (Sheppard et al., 2016).

1.3.2 Algorithms to improve plasmid reconstruction from fragmented assemblies

Approaches described so far aim to address specific questions, such as the location of a resistance gene, or the short-term transmission of particular plasmids. It is yet more challenging to generate complete structures of diverse plasmids from fragmented assemblies generated from short-read WGS data, although several algorithms attempt to do this. The Plasmid Constellation Network (PLACNET) method (Lanza et al., 2014; Vielva et al., 2017) takes the assembly graph and adds reference genomes as nodes; references are linked to homologous contigs, and the assembly graph is reconfigured according to the additional links. Finally, manual reconfiguration is required to retrieve disjoint connected components that should represent distinct genetic units (chromosome/plasmid), with plasmids identified by presence of replication or relaxase proteins. However, poor assembly of repetitive sequences remains a challenge for resistance gene localization, and reliance on reference sequences to order the network can lead to large-scale errors in structure. For example, de Been et al. used long-read sequencing to validate PLACNET reconstructions; in one case, plasmid contigs had been incorrectly reconstructed as two distinct plasmids rather than one, probably because the plasmid was a fusion of two previously-observed reference plasmids (de Been et al., 2014). The requirement for manual reconfiguration means PLACNET cannot easily be applied to large datasets, and also makes it difficult to benchmark its performance against competitor software tools.

Alternative algorithms for identifying plasmid sequences are implemented by entirely automated software tools; some of these tools, like PLACNET, aim to reconstruct distinct plasmid sequences; namely, plasmidSPAdes, MOB-recon, HyAsP, and gplas (Antipov et al., 2016; Robertson and Nash, 2018; Arredondo-Alonso et al., 2019a; Müller and Chauve, 2019). Other automated tools aim only to separate plasmid from chromosomal sequence, without resolving plasmid units; namely, PlaScope, PlasFlow, and mlplasmids (Arredondo-Alonso et al., 2018; Krawczyk et al., 2018; Royer et al., 2018). Among these tools, two principal methodologies are represented. First, using coverage depth to distinguish plasmid and chromosomal contigs (plasmidSPAdes), and to resolve individual plasmids, with the assumption that different genetic units exhibit distinct copy numbers. Second, using plasmid and chromosomal reference sequences to classify contigs, either based on reference matching (MOB-recon, PlaScope) or by using reference sequence genomic signatures to train machine learning classifiers (PlasFlow, mlplasmids). Information on contig length and graph connectivity may also be incorporated. Recent tools combine reference-based and coverage depth approaches (HyAsP, gplas). Benchmarking indicates that the goal of accurately resolving plasmid structures remains unmet, although distinguishing plasmid from chromosomal contigs is more achievable (Arredondo-Alonso et al., 2016; Müller and Chauve, 2019).

1.4 Long-read WGS for plasmid analysis

Plasmid sequences reconstructed by algorithms described above enable analyses of the wider plasmid genome using short-read WGS data (Arredondo-Alonso et al., 2019b). However, results need to be interpreted cautiously, given that plasmid structures may be inaccurate. Long-read sequencing technologies, notably single molecule real-time sequencing (SMRT, Pacific Biosciences) and nanopore sequencing (Oxford Nanopore) promise to revolutionise plasmid analysis (Chin et al., 2013; Loman et al., 2015). Repeat regions that would convolute short-read assembly can be spanned by the long reads, enabling assembly of complete bacterial genomes including resolution of plasmid structures (George

et al., 2017; De Maio et al., 2019). Compared with Illumina short reads, long reads currently have lower base-calling accuracy. Therefore, while long-read only assembly is possible (Wick and Holt, 2019), researchers often rely on hybrid sequencing approaches to combine base-calling accuracy (from short-reads) with structural accuracy (from long-reads). Current hospital outbreak studies commonly sequence the complete isolate collection using Illumina sequencing, and only apply long-read sequencing to a subset of isolates, for which hybrid assemblies are generated; inferences about plasmid diversity from short-read data can guide selection of a representative subset of isolates (Conlan et al., 2016; Evans et al., 2019). In future, further technological improvements and cost-reductions are likely to encourage long-read only assembly, as well as long-read sequencing of larger isolate collections (Goldstein et al., 2019).

Complete and accurate plasmid structures open up new avenues of research. Resistance genes can be trivially localised to plasmids; in turn, the short-term evolutionary trajectories of resistance plasmids during an outbreak (e.g. structural rearrangements and transmission events) can be inferred using standard comparative genomic tools such as the Artemis Comparison Tool (ACT) (Carver et al., 2005; Peter et al., 2019). In addition, analysis of complete plasmids can improve understanding of plasmid biology. For example, tracking the transfer of plasmids during *in vitro* experiments using long-read sequencing was used to gain insight into plasmid-specific genetic factors associated with plasmid transmission (Benz et al., 2019). Studies of single hospital outbreaks overlook a broader 'One Health' perspective – this perspective acknowledges that antibiotic resistance is found in human, livestock, and environmental bacteria, and to understand and control the spread of resistance, we need to study bacteria across all three contexts (Hernando-Amado et al., 2019). Most existing studies of plasmid transmission are based on plasmid typing, but reliable inferences can only be made by comparison of complete plasmids (Muloi et al., 2018). Ludden et al. recently investigated whether resistance spreads between humans and livestock using short- and limited long-read sequencing, applied to isolates from multiple sites in Cambridgeshire, UK. Recent transmission was generally not supported, but a possible case of recent plasmid transfer was identified based on high sequence similarity (99% identity, 98%

coverage breadth) (Ludden et al., 2019). However, the study was limited by sparse sampling of human isolates, and the regional focus, which limits the generalisability of conclusions (Hanage, 2019).

Analysis of larger datasets of complete plasmids from multiple sites potentially allows for more robust and generalisable conclusions. However, descriptive assessments, based on aligning plasmids to one or a handful of reference sequences and inferring possible short-term evolutionary events – a common approach for single-site outbreak studies – is usually not a viable approach for large datasets; nor is phylogenetic analysis (see above). Instead, plasmids can be compared using all-v-all sequence similarity metrics such that plasmid sequence diversity can be assessed, while underlying evolutionary processes may remain obscure. Sequence similarity metrics can be based on gene-content similarity (Brilli et al., 2008; de Been et al., 2014; Corel et al., 2015; Suzuki et al., 2020); alternatively, average nucleotide identity (ANI) metrics can be calculated, using either alignment-based or alignment-free methods (Ondov et al., 2016; Fernandez-Lopez et al., 2017; Jain et al., 2018). Plasmid relationships can then be visualised using a dendrogram or network; the latter representation is popular since it reflects the fact that plasmid evolution does not conform to a tree-like pattern (Bapteste et al., 2009; Suzuki et al., 2020).

Currently, analysis of large datasets of complete plasmids often entails retrieving sequences from public sequence databases, namely the NCBI nucleotide database (Orlek et al., 2017a), after which plasmid diversity can be visualised (Galata et al., 2019; Jesus et al., 2019), and the structure of plasmid diversity can be analysed statistically, e.g. assessing correlation of plasmid groupings with traditional typing schemes (Acman et al., 2020). Unfortunately, retrieving plasmid sequences from NCBI nucleotide is not straightforward, since curation is required to filter chromosomal sequences which have been misannotated as plasmids (Orlek et al., 2017b). NCBI nucleotide database sequences are linked to corresponding sample metadata such as geographic location and isolation source (Barrett et al., 2012), but this is of variable quality and completeness, obstructing research (Card et al., 2019; Gonçalves and Musen, 2019). Hence, sample metadata remains a relatively untapped data source,

which – if curated properly – could be used to address important questions in plasmid-epidemiology; for example, understanding how plasmid-mediated antibiotic resistance epidemiology is influenced by metadata characteristics such as isolation source (e.g. human/livestock).

1.5 Thesis outline

During my PhD, there has been massive growth in availability of bacterial WGS data. Alongside sustained use of short-read sequencing, long-read sequencing has emerged, facilitating resolution of (near) complete plasmid structures. I ultimately aimed to capitalise on the growing quantities of plasmid sequence data – including large numbers of publicly available complete plasmid sequences archived in NCBI – in order to address biological/epidemiological research questions. For example, one of the advantages of analysing complete plasmids is that co-associations between plasmid genetic traits can be investigated more easily, opening up interesting research avenues, as demonstrated in Chapter 5 (see below). However, I found I needed to develop bioinformatic methods to support biological research. Therefore, this thesis presents a combination of method development work as well as research addressing biological questions relating to plasmid-mediated antibiotic resistance. Method development work includes curating complete plasmid sequences and associated metadata from NCBI (Chapters 2 and 5); and developing tools for comparative genomics of plasmids assembled from short and long sequencing reads (Chapters 3 and 4). Key biological research avenues included a comparative genomic analysis of plasmids from a hospital outbreak (Chapter 4), and an investigation into the biological and abiotic factors associated with plasmid resistance gene carriage (Chapter 5). A more detailed outline of each chapter is given below.

Plasmid typing, especially replicon typing, has been used widely to investigate plasmid epidemiology and distinguish plasmid from chromosomal sequences. Through my early explorations of the literature, I came across 'untyped' plasmids that did not match any loci of a given typing scheme. However, recent systematic analysis of the performance of *in silico* plasmid typing (replicon and MOB

typing) and subtyping (pMLST) schemes was lacking. Therefore, in **Chapter 2**, I began by assessing the performance of *in silico* plasmid typing schemes in terms of 'typeability' (the proportion of plasmids that are typeable), as well as the concordance between replicon and MOB typing – if they show the same phylogenetic signal, then replicon types should nest within MOB types, since MOB typing is a broader resolution scheme. As part of this analysis I developed methods for curating NCBI plasmid sequences, which were ultimately packaged into a software tool. Overall, I showed that the schemes fail to accommodate the diversity of plasmid genomes within NCBI; the typeability of the replicon typing scheme was especially poor for plasmids that fell outside relatively well-studied (clinically relevant) taxa, such as Enterobacteriaceae.

Results from Chapter 2 underscored the limitations of plasmid typing schemes for studying plasmid epidemiology. Therefore, in subsequent chapters, I focused on WGS analysis approaches which capitalise on the wider plasmid genome sequence. It is difficult to accurately reconstruct individual plasmids from short-read data, so downstream comparative genomics of individual plasmids may not be possible. Hence, in **Chapter 3**, I aimed for the more attainable goal of comparing samples in terms of their plasmidomes (i.e. total plasmid genetic content), without needing to reconstruct individual plasmids. Specifically, I developed two methods for assessing sample plasmidome diversity: 1) hierarchical clustering according to per-sample plasmid replicon types; 2) BLAST aligning sample contigs to a plasmid reference database, then clustering samples according sample-vs-reference plasmid distance score profiles. Unfortunately, benchmarking suggested that neither method is completely reliable.

Complete plasmid sequences are becoming increasingly available, thanks in large part to the emergence of long-read sequencing technology. However, available tools for plasmid comparative genomics have limitations: they often do not provide metrics on structural similarity and in the case of current alignment-free tools, the output metric conflates sequence identity and breadth of coverage. Therefore, in **Chapter 4**, I proceeded to develop and validate a tool for comparative

genomics of complete genomes, which quantifies all-v-all pairwise nucleotide identity, coverage breadth, and structural similarity. I applied this tool to plasmid contigs, assembled from 38 long-read sequenced samples from a study focusing on a hospital outbreak in Manchester, UK. This enabled the construction of a network displaying sharing of plasmid genetic content between samples. Inter-sample relationships were further investigated using comparative genomic visualisation. I found evidence for persistence of plasmid genetic content across a period of several years, as well as examples of inferred horizontal dissemination of plasmid genetic content between different species and strains.

In **Chapter 5**, I aimed to investigate whether there are particular plasmid genetic traits or abiotic characteristics that are associated with plasmid antibiotic resistance gene carriage. I first curated a dataset of bacterial plasmids and linked sample metadata. Using the curated metadata (geographic location, isolation source, collection date) as well as plasmid biological features, I conducted multivariate statistical analyses to determine correlates of resistance gene carriage across different resistance gene classes. Importantly, I attempted to reduce 'pseudoreplication' (e.g. clonal plasmids re-sequenced during the course of an outbreak), which could otherwise have undermined the robustness of the analysis. Overall, the analysis yielded numerous interesting findings. For example, resistance patterns in different isolation sources (human/livestock), across time, and across different host bacterial taxa reflected known antibiotic usage patterns. Hence, the modelling results were consistent with antibiotic usage as an important driver of plasmid resistance carriage.

Bibliography

Acman, M., van Dorp, L., Santini, J. M., and Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* doi:10.1038/s41467-020-16282-w.

Adler, A., and Carmeli, Y. (2011). Dissemination of the Klebsiella pneumoniae carbapenemase in the health care settings: Tracking the trails of an elusive offender. *MBio* 2. doi:10.1128/mBio.00280-11.

Alvarado, A., Garcillán-Barcia, M. P., and de la Cruz, F. (2012). A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. *PLoS One* 7. doi:10.1371/journal.pone.0040438.

Ameur, A., Kloosterman, W. P., and Hestand, M. S. (2019). Single-Molecule Sequencing: Towards Clinical Applications. *Trends Biotechnol.* doi:10.1016/j.tibtech.2018.07.013.

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., and Pevzner, P. A. (2016). Genome analysis plasmidSPAdes : assembling plasmids from whole genome sequencing data. *Bioinformatics* 33, 3380–3387.

Arredondo-Alonso, S., Bootsma, M., Hein, Y., Rogers, M. R. C., Corander, J., Willems, R. J., et al. (2019a). gplas: a comprehensive tool for plasmid analysis using short-read graphs. *bioRxiv*. doi:10.1101/835900.

Arredondo-Alonso, S., Rogers, M. R. C., Braat, J. C., Verschuuren, T. D., Top, J., Corander, J., et al. (2018). mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. genomics*. doi:10.1099/mgen.0.000224.

Arredondo-Alonso, S., Top, J., Schürch, A. C., McNally, A., Puranen, S., Pesonen, M., et al. (2019b). Genomes of a major nosocomial pathogen Enterococcus faecium are shaped by adaptive evolution of the chromosome and plasmidome. *bioRxiv*. doi:10.1101/530725.

Arredondo-Alonso, S., van Schaik, W., Willems, R. J., and Schurch, A. C. (2016). On the (im)possibility to reconstruct plasmids from whole genome short-read sequencing data. *bioRxiv*, 1–18. doi:http://dx.doi.org/10.1101/086744.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *J. Exp. Med.* doi:10.1084/jem.79.2.137.

Bapteste, E., O'Malley, M. a, Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., et al. (2009). Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* 4, 34. doi:10.1186/1745-6150-4-34.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., et al. (2012). BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Res.* 40. doi:10.1093/nar/gkr1163.

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br. J. Pharmacol.* 153 Suppl, S347–S357. doi:10.1038/sj.bjp.0707607.

Benz, F., Huisman, J. S., Bakkeren, E., Herter, J. A., Stadler, T., Ackermann, M., et al. (2019). Extended-spectrum beta-lactamase antibiotic resistance plasmids have diverse transfer rates and can be spread in the absence of selection. *bioRxiv*. doi:10.1101/796243.

Bertini, A., Poirel, L., Mugnier, P. D., Villa, L., Nordmann, P., and Carattoli, A. (2010). Characterization and PCR-based replicon typing of resistance plasmids in Acinetobacter baumannii. *Antimicrob. Agents Chemother.* 54, 4168–4177. doi:10.1128/AAC.00542-10.

Blokesch, M. (2016). Natural competence for transformation. *Curr. Biol.* doi:10.1016/j.cub.2016.08.058.

Bousquet, A., Henquet, S., Compain, F., Genel, N., Arlet, G., and Decré, D. (2015). Partition locus-based classification of selected plasmids in Klebsiella pneumoniae, Escherichia coli and Salmonella enterica spp.: An additional tool. *J. Microbiol. Methods* 110, 85–91. doi:10.1016/j.mimet.2015.01.019.

Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Liò, P., and Fani, R. (2008). Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics* 9, 551. doi:10.1186/1471-2105-9-551.

Browne, P. D., Kot, W., Jørgensen, T. S., and Hansen, L. H. (2020). "The Mobilome: Metagenomic Analysis of Circular Plasmids, Viruses, and Other Extrachromosomal Elements," in *Methods in Molecular Biology* doi:10.1007/978-1-4939-9877-7_18.

Carattoli, A. (2013). Plasmids and the spread of resistance. *Int. J. Med. Microbiol.* 303, 298–304. doi:10.1016/j.ijmm.2013.02.001.

Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K. L., and Threlfall, E. J. (2005). Identification of plasmids by PCR-based replicon typing. *J. Microbiol. Methods* 63, 219–228. doi:10.1016/j.mimet.2005.03.018.

Carattoli, A., and Hasman, H. (2020). "PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS)," in *Methods in Molecular Biology* doi:10.1007/978-1-4939-9877-7_20.

Carattoli, A., Zankari, E., Garcia-Fernández, A., Larsen, M. V., Lund, O., Villa, L., et al. (2014). In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi:10.1128/AAC.02412-14.

Card, G. E., Pickett, B. D., Ridge, P. G., and Robison, R. A. (2019). Molecular epidemiology of carbapenem-resistance plasmids using publicly available sequences. *Genome*. doi:10.1139/gen-2019-0100.

Carrër, A. A., Poirel, L., Yilmaz, M., Akan, Ö. A., Feriha, C., Cuzon, G., et al. (2010). Spread of OXA-48-encoding plasmid in Turkey and beyond. *Antimicrob. Agents Chemother.* 54, 1369–1373. doi:10.1128/AAC.01312-09.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., and Parkhill, J. (2005). ACT: The Artemis comparison tool. *Bioinformatics* 21, 3422–3423. doi:10.1093/bioinformatics/bti553.

Chancey, S. T., Zähner, D., and Stephens, D. S. (2012). Acquired inducible antimicrobial resistance in Gram-positive bacteria. *Future Microbiol.* doi:10.2217/fmb.12.63.

Chiang, Y. N., Penadés, J. R., and Chen, J. (2019). Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLoS Pathog.* doi:10.1371/journal.ppat.1007878.

Chin, C.-S. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi:10.1038/nmeth.2474.

Compain, F., Poisson, A., Le Hello, S., Branger, C., Weill, F. X., Arlet, G., et al. (2014). Targeting

relaxase genes for classification of the predominant plasmids in Enterobacteriaceae. *Int. J. Med. Microbiol.* 304, 236–242. doi:10.1016/j.ijmm.2013.09.009.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29, 987–991. doi:10.1038/nbt.2023.

Conlan, S., Park, M., Deming, C., and Thomas, P. J. (2016). Plasmid Dynamics in KPC-Positive Klebsiella pneumoniae during Long-Term Patient Colonization. *MBio* 7, 1–9.

Conlan, S., Thomas, P. J., Deming, C., Park, M., Lau, A. F., Dekker, J. P., et al. (2014). Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* 6, 254ra126. doi:10.1126/scitranslmed.3009845.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2015). Network-Thinking : Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* xx, 1–14. doi:10.1016/j.tim.2015.12.003.

Couturier, M., Bex, F., Bergquist, P. L., and Maas, W. K. (1988). Identification and classification of bacterial plasmids. *Microbiol. Rev.* 52, 375–395. doi:10.1016/0042-6822(62)90018-1.

Croucher, N. J., and Didelot, X. (2015). The application of genomics to tracing bacterial pathogen transmission. *Curr. Opin. Microbiol.* 23, 62–67. doi:10.1016/j.mib.2014.11.004.

Cury, J., Abby, S. S., Doppelt-Azeroual, O., Neron, B., and Rocha, E. P. C. (2020). "Chapter 19: Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan," in *Horizontal Gene Transfer: Methods and Protocols*, ed. F. de la Cruz (New York, NY: Humana Press), 265–283.

de Been, M., Lanza, V. F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., et al. (2014). Dissemination of Cephalosporin Resistance Genes between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages. *PLoS Genet.* 10. doi:10.1371/journal.pgen.1004776.

de Kraker, M. E. A., Wolkewitz, M., Davey, P. G., Koller, W., Berger, J., Nagler, J., et al. (2011). Burden of antimicrobial resistance in European hospitals: Excess mortality and length of hospital stay associated with bloodstream infections due to Escherichia coli resistant to third-generation cephalosporins. *J. Antimicrob. Chemother.* 66, 398–407. doi:10.1093/jac/dkq412.

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., et al. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics*. doi:10.1099/mgen.0.000294.

Dealtry, S., Ding, G. C., Weichelt, V., Dunon, V., Schlüter, A., Martini, M. C., et al. (2014a). Cultivation-independent screening revealed hot spots of IncP-1, IncP-7 and IncP-9 plasmid occurrence in different environmental habitats. *PLoS One* 9. doi:10.1371/journal.pone.0089922.

Dealtry, S., Holmsgaard, P. N., Dunon, V., Jechalke, S., Ding, G. C., Krögerrecklenfort, E., et al. (2014b). Shifts in abundance and diversity of mobile genetic elements after the introduction of diverse pesticides into an on-farm biopurification system over the course of a year. *Appl. Environ. Microbiol.* 80, 4012–4020. doi:10.1128/AEM.04016-13.

Decraene, V., Phan, H. T. T., George, R., Wyllie, D. H., Akinremi, O., Aiken, Z., et al. (2018). A large, refractory nosocomial outbreak of klebsiella pneumoniae carbapenemase-producing Escherichia coli demonstrates carbapenemase gene outbreaks involving sink sites require novel approaches to infection control. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01689-18.

Diatheva (2020). PBRT KIT- PCR-based replicon typing: PBRT 2.0 kit. Available at: https://www.diatheva.com/molecular-biology/pcr-based-replicon-typing/pbrt-2-0-kit-details [Accessed January 27, 2020].

Dib, J. R., Wagenknecht, M., Farías, M. E., and Meinhardt, F. (2015). Strategies and approaches in plasmidome studies-uncovering plasmid diversity disregarding of linear elements? *Front. Microbiol.* 6, 1–12. doi:10.3389/fmicb.2015.00463.

Doumith, M., Godbole, G., Ashton, P., Larkin, L., Dallman, T., Day, M., et al. (2016). 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.*, dkw093-. doi:10.1093/jac/dkw093.

Edwards, D. J., and Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb. Inform. Exp.* 3, 2. doi:10.1186/2042-5783-3-2.

EFSA (2011). Scientific Opinion on the public health risks of bacterial strains producing extended-spectrum β-lactamases and/or AmpC β-lactamases in food and food-producing animals. *EFSA J.* 9, 1–95. doi:10.2903/j.efsa.2011.2322.

Evans, D. R., Griffith, M. P., Mustapha, M. M., Marsh, J. W., Sundermann, A. J., Cooper, V. S., et al. (2019). Comprehensive analysis of horizontal gene transfer among multidrug-resistant bacterial pathogens in a single hospital. *bioRxiv*. doi:10.1101/844449.

Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M. P., and de la Cruz, F. (2017). Towards a taxonomy of conjugative plasmids. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2017.05.005.

Fondi, M., Bacci, G., Brilli, M., Papaleo, M. C., Mengoni, A., Vaneechoutte, M., et al. (2010). Exploring the evolutionary dynamics of plasmids: the Acinetobacter pan-plasmidome. *Bmc Evol. Biol.* 10. doi:10.1186/1471-2148-10-59.

Foxman, B., Zhang, L., Koopman, J. S., Manning, S. D., and Marrs, C. F. (2005). Choosing an appropriate bacterial typing technique for epidemiologic studies. *Epidemiol. Perspect. Innov.* 2, 10. doi:10.1186/1742-5573-2-10.

Freitas, A. R., Novais, C., Tedim, A. P., Francia, M. V., Baquero, F., Peixe, L., et al. (2013). Microevolutionary Events Involving Narrow Host Plasmids Influences Local Fixation of Vancomycin-Resistance in Enterococcus Populations. *PLoS One* 8. doi:10.1371/journal.pone.0060589.

Freitas, A. R., Tedim, A. P., Francia, M. V., Jensen, L. B., Novais, C., Peixe, L., et al. (2016). Multilevel population genetic analysis of vanA and vanB Enterococcus faecium causing nosocomial outbreaks in 27 countries (1986-2012). *J. Antimicrob. Chemother.*, dkw312. doi:10.1093/jac/dkw312.

Fricke, W. F., Welch, T. J., McDermott, P. F., Mammel, M. K., LeClerc, J. E., White, D. G., et al. (2009). Comparative genomics of the IncA/C multidrug resistance plasmid family. *J. Bacteriol.* 191, 4750–4757. doi:10.1128/JB.00189-09.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat.Rev.Microbiol.* 3, 722–732. doi:10.1038/nrmicro1235.

Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019). PLSDB: A resource of complete bacterial plasmids. *Nucleic Acids Res.* doi:10.1093/nar/gky1050.

Garcillán-Barcia, M. P., Francia, M. V., and De La Cruz, F. (2009). The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* 33, 657–687. doi:10.1111/j.1574-6976.2009.00168.x.

Garcillan-Barcia, M. P., Redondo-Salvo, S., Vielva, L., and de la Cruz, F. (2020). "Chapter 21: MOBscan: Automated Annotation of MOB Relaxases," in *Horizontal Gene Transfer: Methods*

and Protocols, ed. F. de la Cruz (New York, NY), 295–308.

George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., et al. (2017). Resolving plasmid structures in enterobacteriaceae using the MinION nanopore sequencer: Assessment of MinION and MinION/illumina hybrid data assembly approaches. *Microb. Genomics*. doi:10.1099/mgen.0.000118.

Gillings, M. R. (2014). Integrons: past, present, and future. *Microbiol. Mol. Biol. Rev.* 78, 257–77. doi:10.1128/mmbr.00056-13.

Goicoechea, P., Romo, M., and Coque, T. (2008). Identification of enterococcal plasmids by multiplex-PCR-based relaxase typing. in *18th European Congress of Clinical Microbiology and Infectious Diseases. Barcelona, Spain*, Abstract P1669.

Goldstein, S., Beka, L., Graf, J., and Klassen, J. L. (2019). Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics*. doi:10.1186/s12864-018-5381-7.

Gonçalves, R. S., and Musen, M. A. (2019). Analysis: The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data*. doi:10.1038/sdata.2019.21.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49.

Guglielmini, J., Néron, B., Abby, S. S., Garcillán-Barcia, M. P., La Cruz, D. F., and Rocha, E. P. C. (2014). Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* 42, 5715–5727. doi:10.1093/nar/gku194.

Guglielmini, J., Quintais, L., Garcillán-Barcia, M. P., de la Cruz, F., and Rocha, E. P. C. (2011). The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7. doi:10.1371/journal.pgen.1002222.

Halary, S., Leigh, J. W., Cheaib, B., Lopez, P., and Bapteste, E. (2010). Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci.* 107, 127–132. doi:10.1073/pnas.0908978107.

Hall, J. P. J., Brockhurst, M. A., and Harrison, E. (2017). Sampling the mobile gene pool: Innovation via horizontal gene transfer in bacteria. *Philos. Trans. R. Soc. B Biol. Sci.* doi:10.1098/rstb.2016.0424.

Hanage, W. P. (2019). Two health or not two health? That is the question. *MBio*. doi:10.1128/mBio.00550-19.

Hasegawa, H., Suzuki, E., and Maeda, S. (2018). Horizontal plasmid transfer by transformation in Escherichia coli: Environmental factors and possible mechanisms. *Front. Microbiol.* doi:10.3389/fmicb.2018.02365.

He, S., Chandler, M., Varani, A. M., Hickman, A. B., Dekker, J. P., and Dyda, F. (2016). Mechanisms of Evolution in High-Consequence Drug Resistance Plasmids. *MBio* 7, e01987-16. doi:10.1128/mBio.01987-16.

Hedges, R. W., and Datta, N. (1973). Plasmids Determining I Pili Constitute a Compatibility Complex. *J. Gen. Microbiol.* 77, 19–25. doi:10.1099/00221287-77-1-19.

Hernando-Amado, S., Coque, T. M., Baquero, F., and Martínez, J. L. (2019). Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat. Microbiol.* 4, 1432–1442. doi:10.1038/s41564-019-0503-9.

Holt, K. E. (2015). Locating drug resistance regions in short reads, using Bandage and ISMapper. Available at: https://holtlab.net/2015/07/20/locating-drug-resistance-regions-in-short-reads-using-bandage-and-ismapper/ [Accessed July 11, 2016].

Holt, K. E., Thieu Nga, T. V., Thanh, D. P., Vinh, H., Kim, D. W., Vu Tra, M. P., et al. (2013). Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* 110, 17522–7. doi:10.1073/pnas.1308632110.

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., et al. (2014). SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 6, 90. doi:10.1186/s13073-014-0090-6.

Iredell, J., Brown, J., and Tagg, K. (2016). Antibiotic resistance in Enterobacteriaceae: mechanisms and clinical implications. *Bmj*, h6420. doi:10.1136/bmj.h6420.

Jain, A., and Srivastava, P. (2013). Broad host range plasmids. *FEMS Microbiol. Lett.* 348, 87–96. doi:10.1111/1574-6968.12241.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 1–8. doi:10.1038/s41467-018-07641-9.

Jensen, L. B., Garcia-Migura, L., Valenzuela, A. J. S., Løhr, M., Hasman, H., and Aarestrup, F. M. (2010). A classification system for plasmids from enterococci and other Gram-positive bacteria. *J. Microbiol. Methods* 80, 25–43. doi:10.1016/j.mimet.2009.10.012.

Jesus, T. F., Ribeiro-Gonçalves, B., Silva, D. N., Bortolaia, V., Ramirez, M., and Carriço, J. A. (2019). Plasmid ATLAS: Plasmid visual analytics and identification in high-Throughput sequencing data. *Nucleic Acids Res.* doi:10.1093/nar/gky1073.

Johnson, C. M., and Grossman, A. D. (2015). Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu. Rev. Genet.* 49, annurev-genet-112414-055018. doi:10.1146/annurev-genet-112414-055018.

Johnson, T. J., Wannemuehler, Y. M., Johnson, S. J., Logue, C. M., White, D. G., Doetkott, C., et al. (2007). Plasmid replicon typing of commensal and pathogenic Escherichia coli isolates. *Appl. Environ. Microbiol.* 73, 1976–1983. doi:10.1128/AEM.02171-06.

Keen, E. C., Bliskovsky, V. V., Malagon, F., Baker, J. D., Prince, J. S., Klaus, J. S., et al. (2017). Novel "superspreader" bacteriophages promote horizontal gene transfer by transformation. *MBio*. doi:10.1128/mBio.02115-16.

Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L. H., Sørensen, S. J., et al. (2014). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* 9, 934–945. doi:10.1038/ismej.2014.191.

Kohler, V., Keller, W., and Grohmann, E. (2019). Regulation of gram-positive conjugation. *Front. Microbiol.* doi:10.3389/fmicb.2019.01134.

Koren, S., and Phillippy, A. M. (2015). One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* 23, 110–120. doi:10.1016/j.mib.2014.11.014.

Krawczyk, P. S., Lipinski, L., and Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* doi:10.1093/nar/gkx1321.

Laguerre, G., Mazurier, S. I., and Amarger, N. (1992). Plasmid profiles and restriction fragment length polymorphism of Rhizobium leguminosarum bv. viciae in field populations. *FEMS Microbiol.*

Lett. 101, 17–26. doi:Doi: 10.1016/0378-1097(92)90693-i.

Lanza, V. F., de Toro, M., Garcillán-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., et al. (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. doi:10.1371/journal.pgen.1004766.

Lederberg, J. (1998). Personal perspective: Plasmid (1952-1997). *Plasmid*. doi:10.1006/plas.1997.1320.

Lederberg, J., and Tatum, E. L. (1946). Gene recombination in Escherichia coli [23]. *Nature*. doi:10.1038/158558a0.

Lerminiaux, N. A., and Cameron, A. D. S. (2019). Horizontal transfer of antibiotic resistance genes in clinical environments. *Can. J. Microbiol.* doi:10.1139/cjm-2018-0275.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Liu, Y. Y., Wang, Y., Walsh, T. R., Yi, L. X., Zhang, R., Spencer, J., et al. (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. doi:10.1016/S1473-3099(15)00424-7.

Livermore, D. M., and Woodford, N. (2006). The β-lactamase threat in Enterobacteriaceae, Pseudomonas and Acinetobacter. *Trends Microbiol.* 14, 413–420. doi:10.1016/j.tim.2006.07.008.

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi:10.1038/nmeth.3444.

Lozano, C., García-Migura, L., Aspiroz, C., Zarazaga, M., Torres, C., and Aarestrup, F. M. (2012). Expansion of a plasmid classification system for gram-positive bacteria and determination of the diversity of plasmids in Staphylococcus aureus strains of human, animal, and food origins. *Appl. Environ. Microbiol.* 78, 5948–5955. doi:10.1128/AEM.00870-12.

Ludden, C., Raven, K. E., Jamrozy, D., Gouliouris, T., Blane, B., Coll, F., et al. (2019). One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *MBio*. doi:10.1128/mBio.02693-18.

Lynch, T., Petkau, A., Knox, N., Graham, M., and Van Domselaar, G. (2016). A Primer on Infectious Disease Bacterial Genomics. *Clin. Microbiol. Rev.* 29, 881–913. doi:10.1128/CMR.00001-16.

Maiden, M. C. (2006). Multilocus sequence typing of bacteria. *Annu.Rev.Microbiol.* 60, 561–588. doi:10.1146/annurev.micro.59.030804.121325.

Martínez, J. L., Coque, T. M., Lanza, V. F., de la Cruz, F., and Baquero, F. (2017). Genomic and metagenomic technologies to explore the antibiotic resistance mobilome. *Ann. N. Y. Acad. Sci.*, 1–16. doi:10.1111/nyas.13282.

Mathers, A. J., Stoesser, N., Sheppard, A. E., Pankhurst, L., Giess, A., Yeh, A. J., et al. (2015). Klebsiella pneumoniae carbapenemase (KPC)-producing K. pneumoniae at a single institution: Insights into endemicity from Whole-genome sequencing. *Antimicrob. Agents Chemother.* 59, 1656–1663. doi:10.1128/AAC.04292-14.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46. doi:10.1038/nrg2626.

Moreno-Switt, A. I., Pezoa, D., Sepúlveda, V., González, I., Rivera, D., Retamal, P., et al. (2019). Transduction as a Potential Dissemination Mechanism of a Clonal qnrB19-Carrying Plasmid Isolated From Salmonella of Multiple Serotypes and Isolation Sources. *Front. Microbiol.* doi:10.3389/fmicb.2019.02503.

Müller, R., and Chauve, C. (2019). HyAsP, a greedy tool for plasmids identification. *Bioinformatics*. doi:10.1093/bioinformatics/btz413.

Muloi, D., Ward, M. J., Pedersen, A. B., Fèvre, E. M., Woolhouse, M. E. J., and Van Bunnik, B. A. D. (2018). Are Food Animals Responsible for Transfer of Antimicrobial-Resistant Escherichia coli or Their Resistance Determinants to Human Populations? A Systematic Review. *Foodborne Pathog. Dis.* 15, 467–474. doi:10.1089/fpd.2017.2411.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. *Nat. Rev. Genet.* 14, 157–67. doi:10.1038/nrg3367.

Narra, H. P., and Ochman, H. (2006). Of What Use Is Sex to Bacteria? *Curr. Biol.* doi:10.1016/j.cub.2006.08.024.

Nathwani, D., Raman, G., Sulham, K., Gavaghan, M., and Menon, V. (2014). Clinical and economic consequences of hospital-acquired resistant and multidrug-resistant Pseudomonas aeruginosa infections: a systematic review and meta-analysis. *Antimicrob. Resist. Infect. Control* 3, 1–16. doi:10.1186/2047-2994-3-32.

Nordmann, P., Dortet, L., and Poirel, L. (2012). Carbapenem resistance in Enterobacteriaceae: Here is the storm! *Trends Mol. Med.* 18, 263–272. doi:10.1016/j.molmed.2012.03.003.

Norman, A., Hansen, L. H., and Sørensen, S. J. S. J. (2009). Conjugative plasmids: Vessels of the communal gene pool. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 2275–2289. doi:10.1098/rstb.2009.0037.

Oladeinde, A., Cook, K., Orlek, A., Zock, G., Herrington, K., Cox, N., et al. (2018). Hotspot mutations and ColE1 plasmids contribute to the fitness of Salmonella Heidelberg in poultry litter. *PLoS One*. doi:10.1371/journal.pone.0202286.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* doi:10.1186/s13059-016-0997-x.

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017a). A curated dataset of complete Enterobacteriaeceae plasmids compiled from the NCBI nucleotide database. *Data Br.*

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017b). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* 91, 42–52. doi:10.1016/j.plasmid.2017.03.002.

Partridge, S. R., Kwong, S. M., Firth, N., and Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* doi:10.1128/CMR.00088-17.

Pecora, N. D., Li, N., Allard, M., Li, C., Albano, E., Delaney, M., et al. (2015). Genomically informed surveillance for carbapenem-resistant enterobacteriaceae in a health care system. *MBio* 6, 1–11. doi:10.1128/mBio.01030-15.

Peter, S., Bosio, M., Gross, C., Bezdan, D., Gutierrez, J., Oberhettinger, P., et al. (2019). Tracking of antibiotic resistance transfer and rapid plasmid evolution in a hospital setting by Nanopore sequencing. *bioRxiv*. doi:10.1101/639609.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.* 98, 9748–53. doi:10.1073/pnas.171285098.

Phan, M. D., Kidgell, C., Nair, S., Holt, K. E., Turner, A. K., Hinds, J., et al. (2009). Variation in Salmonella enterica serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever. *Antimicrob. Agents Chemother.* 53, 716–727. doi:10.1128/AAC.00645-08.

Popa, O., Landan, G., and Dagan, T. (2017). Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *ISME J.* doi:10.1038/ismej.2016.116.

Ramsay, J. P., Kwong, S. M., Murphy, R. J. T., Yui Eto, K., Price, K. J., Nguyen, Q. T., et al. (2016). An updated view of plasmid conjugation and mobilization in *Staphylococcus*. *Mob. Genet. Elements* 6, e1208317. doi:10.1080/2159256X.2016.1208317.

Robertson, J., and Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. genomics*. doi:10.1099/mgen.0.000206.

Rose, T. M., Schultz, E. R., Henikoff, J. G., Pietrokovski, S., McCallum, C. M., and Henikoff, S. (1998). Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res.* 26, 1628–1635. doi:10.1093/nar/26.7.1628.

Rosvoll, T. C. S., Lindstad, B. L., Lunde, T. M., Hegstad, K., Aasnæs, B., Hammerum, A. M., et al. (2012). Increased high-level gentamicin resistance in invasive Enterococcus faecium is associated with aac(6')Ie-aph(2'')Ia-encoding transferable megaplasmids hosted by major hospital-adapted lineages. *FEMS Immunol. Med. Microbiol.* 66, 166–176. doi:10.1111/j.1574-695X.2012.00997.x.

Rosvoll, T. C. S., Pedersen, T., Sletvold, H., Johnsen, P. J., Sollid, J. E., Simonsen, G. S., et al. (2010). PCR-based plasmid typing in Enterococcus faecium strains reveals widely distributed pRE25-, pRUM-, pIP501- and pHTβ-related replicons associated with glycopeptide resistance and stabilizing toxin-antitoxin systems. *FEMS Immunol. Med. Microbiol.* 58, 254–268. doi:10.1111/j.1574-695X.2009.00633.x.

Royer, G., Decousser, J. W., Branger, C., Dubois, M., Médigue, C., Denamur, E., et al. (2018). PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb. genomics*. doi:10.1099/mgen.0.000211.

San Millan, A., and MacLean, R. C. (2017). Fitness Costs of Plasmids: A Limit to Plasmid Transmission. *Microb. Transm.*, 65–79. doi:10.1128/microbiolspec.mtbp-0016-2017.

Sen, D., Brown, C. J., Top, E. M., and Sullivan, J. (2013). Inferring the evolutionary history of INCP-1 plasmids despite incongruence among backbone gene trees. *Mol. Biol. Evol.* 30, 154–166. doi:10.1093/molbev/mss210.

Shearer, J. E. S., Wireman, J., Hostetler, J., Forberger, H., Borman, J., Gill, J., et al. (2011). Major families of multiresistant plasmids from geographically and epidemiologically diverse staphylococci. *G3 (Bethesda).* 1, 581–91. doi:10.1534/g3.111.000760.

Sheppard, A. E., Stoesser, N., Wilson, D. J., Sebra, R., Kasarskis, A., Anson, L. W., et al. (2016). Nested Russian Doll-like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.00464-16.

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., and Chandler, M. (2015). "Everyman's Guide to Bacterial Insertion Sequences," in *Mobile DNA III* doi:10.1128/9781555819217.ch26.

Smillie, C., Garcillan-Barcia, M. P., Francia, M. V, Rocha, E. P., and de la Cruz, F. (2010). Mobility of plasmids. *Microbiol Mol Biol Rev* 74, 434–452. doi:10.1128/MMBR.00020-10.

Snyder, L., Peters, J. E., Henkin, T. M., and Champness, W. (2013a). “Chapter 1 Introduction,” in *Molecular Genetics of Bacteria* (Washington, DC: ASM Press), 1–12.

Snyder, L., Peters, J. E., Henkin, T. M., and Champness, W. (2013b). “Chapter 4 Plasmids,” in *Molecular Genetics of Bacteria* (Washington, DC: ASM Press), 183–218.

Snyder, L., Peters, J. E., Henkin, T. M., and Champness, W. (2013c). “Chapter 9 Transposition, site-specific recombination, and families of recombinases,” in *Molecular Genetics of Bacteria* (Washington, DC: ASM Press), 361–402.

Stoesser, N., Giess, A., Batty, E. M., Sheppard, A. E., Walker, A. S., Wilson, D. J., et al. (2014). Genome sequencing of an extended series of NDM-producing Klebsiella pneumoniae isolates from neonatal infections in a Nepali hospital characterizes the extent of community- Versus hospital-associated transmission in an endemic setting. *Antimicrob. Agents Chemother.* 58, 7347–7357. doi:10.1128/AAC.03900-14.

Stoesser, N., Sheppard, A. E., Moore, C. E., Golubchik, T., Parry, C. M., Nget, P., et al. (2015a). Extensive within-host diversity in fecally carried extended-spectrum-beta-lactamase-producing Escherichia coli Isolates: Implications for transmission analyses. *J. Clin. Microbiol.* 53, 2122–2131. doi:10.1128/JCM.00378-15.

Stoesser, N., Sheppard, A. E., Shakya, M., Sthapit, B., Thorson, S., Giess, A., et al. (2015b). Dynamics of MDR Enterobacter cloacae outbreaks in a neonatal unit in Nepal: Insights using wider sampling frames and next-generation sequencing. *J. Antimicrob. Chemother.* 70, 1008–1015. doi:10.1093/jac/dku521.

Stokes, H. W., and Gillings, M. R. (2011). Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* 35, 790–819. doi:10.1111/j.1574-6976.2011.00273.x.

Suzuki, M., Doi, Y., and Arakawa, Y. (2020). Plasmid ORF-based binarized structure network analysis of plasmids (OSNAp), a novel approach to core gene-independent plasmid phylogeny. *Plasmid* 108, 102477. doi:10.1016/j.plasmid.2019.102477.

Taylor, D., Gibreel, A., Lawley, T., and Tracz, D. (2004). “Chapter 23: Antibiotic resistance plasmids,” in *Plasmid Biology*, eds. B. Funnell and G. Phillips (Washington, DC: ASM Press), 473–485.

Taylor, T. L., Volkening, J. D., DeJesus, E., Simmons, M., Dimitrov, K. M., Tillman, G. E., et al. (2019). Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Sci. Rep.* doi:10.1038/s41598-019-52424-x.

Tazzyman, S. J., and Bonhoeffer, S. (2014). Why There Are No Essential Genes on Plasmids. *Mol. Biol. Evol.* 32, 3079–3088. doi:10.1093/molbev/msu293.

Thomas, C. M., and Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat.Rev.Microbiol.* 3, 711–721. doi:10.1038/nrmicro1234.

Thomas, C., and Summers, D. (2008). “Bacterial Plasmids,” in *eLS*, John Wiley & Sons, Ltd: Chichester. Available at: http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0000468.pub2/full.

Treangen, T. J., and Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36–46. doi:10.1038/nrg3117.

Valverde, A., Cantón, R., Garcillán-Barcia, M. P., Novais, Â., Galán, J. C., Alvarado, A., et al. (2009). Spread of blaCTX-M-14 is driven mainly by IncK plasmids disseminated among Escherichia coli phylogroups A, B1, and D in Spain. *Antimicrob. Agents Chemother.* 53, 5204–5212. doi:10.1128/AAC.01706-08.

Vielva, L., De Toro, M., Lanza, V. F., and De La Cruz, F. (2017). PLACNETw: A web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics*. doi:10.1093/bioinformatics/btx462.

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., et al. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* 7, 1–10. doi:10.3389/fmicb.2016.00173.

Wang, R., Van Dorp, L., Shaw, L. P., Bradley, P., Wang, Q., Wang, X., et al. (2018). The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* 9, 1–9. doi:10.1038/s41467-018-03205-z.

Watanabe, T. (1967). Infectious drug resistance. *Sci. Am.*

Wick, R. R., and Holt, K. E. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 1; peer review: 2 approved]. *F1000Research* 8. doi:https://doi.org/10.12688/f1000research.21782.1.

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* 31, 3350–3352. doi:10.1093/bioinformatics/btv383.

World Health Organization (2014). Antimicrobial Resistance: Global Report on Surveillance. WHO Press.

Xie, M., Li, R., Liu, Z., Chan, E. W. C., and Chen, S. (2018). Recombination of plasmids in a carbapenem-resistant NDM-5-producing clinical Escherichia coli isolate. *J. Antimicrob. Chemother.* doi:10.1093/jac/dkx540.

Zerbino, D. R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi:10.1101/gr.074492.107.

Zinder, N. D., and Lederberg, J. (1952). Genetic exchange in Salmonella. *J. Bacteriol.*

2 Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids

2.1 Summary

This chapter is based on the following article: Orlek et al. (2017). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. Plasmid. 91:42–52.

In the article, I assessed the performance of plasmid typing schemes, using a curated database of "Enterobacteriaceae" plasmids. Additional analysis conducted at a later stage is also incorporated into the chapter (specifically, analysis of a curated dataset of all bacterial plasmids retrieved in 2019). Note that since publication of the "Ordering the mob" article, the taxonomic nomenclature of the Enterobacteriaceae group has undergone revision, and taxonomic annotations in NCBI have been updated accordingly. Briefly, the family previously known as Enterobacteriaceae has essentially been re-assigned as a new order (Enterobacterales ord. nov.), which has in turn been split into seven distinct families. Of these families, the largest is a revised Enterobacteriaceae family which still retains many clinically-relevant genera such as Escherichia, Klebsiella, and Salmonella (Adeolu et al., 2016; McAdam, 2020). In this chapter, I use the contemporaneous nomenclature (i.e. revised nomenclature for the 2019 dataset analysis, otherwise, original nomenclature).

Plasmid typing, principally, replicon and MOB typing, are used to detect plasmid sequences and study the epidemiology of plasmid-mediated antibiotic resistance. Generally, *in silico* replicon typing involves BLASTN querying the PlasmidFinder database of plasmid replication loci, whereas MOB typing involves remote homology searches using a small set of representative relaxase proteins. Consequently, replicon typing classifies plasmids at finer resolution than MOB typing, and is therefore potentially more informative for studying plasmid epidemiology. The usefulness of plasmid typing schemes also depends on the proportion of plasmids which can be assigned a type ('typeability'). In addition, plasmid epidemiology studies implicitly assume that plasmid types are coherent, such that related plasmids are assigned the same type. If this is the case, replicon types should nest within the broader-resolution MOB types. Thousands of plasmid sequences have been added to NCBI in recent years, without consistent annotation to indicate which sequences represent complete plasmids. Here, curated datasets of complete plasmids were retrieved from NCBI, and used to assess the typeability

and concordance of *in silico* replicon and MOB typing schemes. Concordance was assessed at hierarchical replicon type resolutions, from replicon family-level to plasmid multi-locus sequence type (pMLST)-level. Concordance analyses, and initial typeability analyses, were conducted on a dataset of Enterobacteriaceae plasmids; additional typeability analyses were conducted on a dataset of all bacterial plasmids retrieved in 2019. I found that 85% and 65% of curated Enterobacteriaceae plasmids could be replicon and MOB typed. Enterobacteriaceae plasmids for which a replicon or MOB type could be assigned were generally larger and/or encoded more resistance genes. Assessed across all bacterial taxa, typeability of replicon typing varied markedly between taxa, reflecting the taxonomic composition of the PlasmidFinder database. I found some degree of non-concordance between replicon families and MOB types, which was only partly resolved when partitioning plasmids into finer-resolution groups (replicon and pMLST types). In some cases, non-concordance was attributed to ambiguous boundaries between MOBP and MOBQ types; in other cases, backbone mosaicism was considered a more plausible explanation. Overall, replicon and MOB typing schemes are likely to continue playing an important role in plasmid analysis, but their performance is constrained by the diverse and dynamic nature of plasmid genomes.

2.2 Introduction

Plasmid typing may provide insights into resistance plasmid epidemiology (Leverstein-van Hall et al., 2011; de Been et al., 2014; Pecora et al., 2015; Orlek et al., 2017c), such as whether resistance dissemination involves diverse plasmids or one dominant 'epidemic' type (Valverde et al., 2009). Other applications of plasmid typing include plasmid detection; notably, distinguishing plasmid from chromosomal contigs in short-read *de novo* assemblies (Lanza et al., 2014; Robertson and Nash, 2018).

The most widely used plasmid classification schemes are replicon and MOB typing, which exploit loci encoding plasmid replication (replicons) and mobility functions (relaxases), respectively (Carattoli et al., 2005, 2014; Garcillán-Barcia et al., 2009; Alvarado et al., 2012). Replicons include various different

loci, none of which are universal across plasmids (del Solar et al., 1998), whereas relaxases are thought to be universally present amongst plasmids that mobilise via the relaxase-*in-cis* mechanism (Garcillán-Barcia et al., 2011; Ramsay et al., 2016). Replicon types can be assigned by querying plasmids against various replicon sequences using BLASTN (Carattoli et al., 2014), whilst MOB types can be assigned using profile-based searches such as PSI-BLAST. A PSI-BLAST approach involves using *in silico* probes representing each relaxase family to generate search "profiles" which are in turn used to iteratively query a dataset of plasmids, and thereby assign MOB types to plasmids within the dataset (Francia et al., 2004; Garcillán-Barcia et al., 2009; Guglielmini et al., 2011). Six probes have been used to detect relaxases of Gammaproteobacterial plasmids, whilst a further two probes have been used to detect relaxases in other taxa (Guglielmini et al., 2011). PSI-BLAST can uncover distant homology, so MOB typing provides a lower resolution but potentially more inclusive classification (Altschul et al., 1997; Garcillán-Barcia and de la Cruz, 2013). Individual plasmids can contain multiple replicons, complicating classification by replicon typing, whereas usually just one relaxase per plasmid is encoded (Garcillán-Barcia and de la Cruz, 2013).

Within the replicon typing framework, plasmids can be classified at hierarchical resolutions: for common replicon types, plasmid multi-locus sequence typing (pMLST) schemes have been devised for sub-typing (Brolund and Sandegren, 2015; Hancock et al., 2016), whilst replicon types also belong to broader replicon families. Most replicon families were originally defined according to plasmid incompatibility, which is a manifestation of the genetic relatedness of replicons (Taylor et al., 2004). However, the Col 'family' of plasmids was defined according to the more superficial phenotype of colicin production, and it has long been known that this so-called 'family' includes phylogenetically diverse plasmids (Riley and Gordon, 1992). Nevertheless, generally, owing to its higher resolution, replicon typing can provide more detailed information on plasmid relatedness and epidemiology, particularly if a pMLST scheme is available (Pallecchi et al., 2007; Muloi et al., 2018).

'Typeability' is the proportion of plasmids that can be assigned a type by a given typing scheme. MOB typing only types relaxase-encoding plasmids, which constitute varying proportions of total plasmids across different taxa (Smillie et al., 2010; Guglielmini et al., 2011). This variation in typeability across taxa is biologically interesting. For example, Smillie et al. found that *Borrelia* plasmids were not MOB-typeable; *Borrelia* are obligate intracellular parasites so plasmid mobility may not be adaptive given limited opportunities for plasmid transfer (Smillie et al., 2010). As autonomously replicating elements, all plasmids are expected to encode replicon loci. However, replicon loci from less well-studied plasmids (e.g. from non-clinically relevant taxa) may be uncharacterised and hence missing from replicon loci databases used for *in silico* replicon typing (namely, the widely used PlasmidFinder database of replicon loci) (Carattoli et al., 2014). As a result, it may not be possible to assign a replicon type to certain plasmids. A full understanding of how typeability varies (e.g. across source taxa), could be useful for highlighting, and ultimately addressing, gaps in the PlasmidFinder database. More generally, it is important to assess the typeability of *in silico* replicon typing in order to better understand its limitations as a research tool. As discussed previously, replicon typing is often used for detecting plasmid sequences and studying plasmid epidemiology, but 'untyped' plasmids undermine the usefulness of replicon typing in these contexts (Arredondo-Alonso et al., 2017).

Besides typeability, it is also important to understand whether replicon and MOB types group plasmids in a phylogenetically coherent way, such that related plasmids are grouped together. Assuming plasmid types are coherent, replicon and MOB types should be concordant – that is, replicon families should nest within the broader-resolution MOB type families. An assumption that plasmid types can be used to identify related plasmids underlies many plasmid epidemiology studies. However, if plasmid backbones undergo frequent recombination, plasmid types are unlikely to be coherent, and this would undermine the usefulness of plasmid typing as a tool for plasmid epidemiology. Previous studies have supported concordance, with several exceptions (see Figure 5 in Garcillán-Barcia et al., 2011). As alluded to above, non-concordance could result from backbone mosaicism (Hiraga et al., 1994; Kulinska et al., 2008), frequently due to recombination (Bianco and Kowalczykowski, 1997;

Osborn et al., 2000). Alternatively, non-concordance may reflect fuzzy boundaries between the MOB families. Indeed, there is known to be some level of overlap between the MOBP and MOBQ families such that PSI-BLAST searches with MOBP and MOBQ probes sometimes uncover the same homologs (Garcillán-Barcia et al., 2009). However, non-concordance has not been the focus of previous studies, so it remains unclear which explanation is most likely.

Re-investigating the performance of replicon and MOB typing is timely given the increasing numbers of publicly available complete plasmid sequences (Conlan et al., 2014). In this study, I curated a dataset of all complete publicly available Enterobacteriaceae plasmids to allow an up-to-date, comprehensive assessment of replicon and MOB typing schemes in terms of typeability and concordance. Concordance was assessed at hierarchical resolutions from replicon family-level to pMLST-level, where a pMLST scheme was available. When I began this research, the PlasmidFinder database for *in silico* replicon typing only covered Enterobacteriaceae plasmids, but later, a Gram-positive component became available. Hence, to investigate how typeability varies across taxa, I later downloaded a dataset of all bacterial plasmids, and re-ran plasmid typing with the updated PlasmidFinder database.

2.3 Methods

2.3.1 Retrieving complete Enterobacteriaceae plasmid sequences and associated metadata from NCBI

Putative complete Enterobacteriaceae plasmid accessions were retrieved from the NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/) on 26th August 2016, using an Entrez query with filters to exclude some incomplete or non-plasmid accessions at this stage (Supplementary Methods S1). Duplicate sequences (those sharing 100% sequence identity with another retrieved sequence) were removed, preferentially retaining RefSeq over GenBank accessions (Pruitt et al., 2007), and more richly annotated over less richly annotated GenBank accessions. Biopython scripts (Cock et al., 2009)

were used to retrieve information and filter accessions, including filtering-out non-coding sequences, and eliminating incomplete plasmid sequences using a regular expression search of accession title descriptions (Supplementary Methods S1). Multi-locus sequence typing (MLST) using all available Enterobacteriaceae schemes (http://pubmlst.org/data/) was also conducted to identify and remove chromosomal sequences mis-annotated as plasmids, using BLAST as described (Larsen et al., 2012) (Appendix A: Methods 1). Additional manual filtering involved examining putative plasmids at the tails of the sequence length distribution, to remove accessions thought to represent chromosomal sequences or partial plasmid sequences.

EDirect (Kans, 2013) was used to retrieve accession metadata, including the 'completeness' annotation which was used to guide filtering. Edirect was also used to retrieve additional metadata not used for filtering, including the sequencing technology, and the 'create date' of the accession in NCBI (Nahin, 2008). Plasmids in the quality-filtered dataset were assigned as clinical or non-clinical according to the source taxa. Specifically, genus and, where necessary, species-level information on host organism and human pathogenicity derived from the PATRIC database was used to guide assignments (Wattam et al., 2014).

2.3.2 Retrieving an updated dataset of complete plasmids including all bacterial taxa

When work for this chapter began in 2016, I retrieved and analysed only Enterobacteriaceae plasmids because the Gram-positive component of the PlasmidFinder database remained under construction. However, since then, the Gram-positive component of PlasmidFinder has been developed (and contains over 328 replicon sequences, as of Dec 2019). Consequently, I later decided to retrieve complete plasmids across all bacterial taxa (see below). This allowed me to comprehensively investigate how typeability varies across bacterial taxa, and how this relates to the taxonomic composition of replicons included in the PlasmidFinder database (see Section 2.3.3).

A dataset of complete bacterial plasmids was retrieved on 1st May 2019 and curated using updated curation methods. This dataset is the same as the dataset analysed in Chapter 5 (see therein for a detailed description of plasmid curation methods). Briefly, initial curation methods were similar to those described in Section 2.3.1, but rMLST rather than MLST loci were used to exclude chromosomal sequences. This updated methodology was packaged into a software tool called bacterialBercow (https://github.com/AlexOrlek/bacterialBercow; for details see Appendix A: Methods 3). Next, additional curation steps were applied, aiming to exclude plasmids isolated from laboratory or commercial strains (see Figure 1 in Chapter 5). I was concerned about remaining biases in the dataset, resulting from re-sequencing of similar plasmids, sampled from the same location (e.g. during the course of a hospital outbreak). To address concerns about the impact of dataset bias on results, typeability analyses were conducted with a smaller dataset compiled after steps aiming to reduce bias had been implemented. Specifically, I identified clusters of similar plasmids based on sequence similarity (>95% identity, >50% coverage breadth) and metadata sharing (taxonomic/spatiotemporal metadata). Plasmids belonging to a given cluster were considered to be duplicates, and a single representative plasmid was selected (See Chapter 5, Section 5.3.5 for details).

Note that since retrieving the "Enterobacteriaceae" plasmid dataset in 2016 (Section 2.3.1) taxonomic nomenclature underwent revision, such that the former Enterobacteriaceae family now corresponds to the Enterobacterales order. Taxonomic metadata linked to the 2019 bacterial plasmid dataset was assessed, confirming adherence to the updated taxonomic groupings in the Enterobacterales order (with reference to Adeolu et al., 2016). Throughout this chapter, contemporaneous nomenclature is used (i.e. revised nomenclature for the 2019 dataset analysis, otherwise original nomenclature).

2.3.3 Replicon typing and pMLST, MOB typing, resistance gene detection

The Enterobacteriaceae dataset was analysed using the following methods. For replicon typing, the complete Enterobacteriaceae plasmid sequences were queried against a locally downloaded version

of the PlasmidFinder database (retrieved on 20th April 2016 from http://www.genomicepidemiology.org/), using recommended percentage identity and coverage thresholds of 80% and 60% respectively (Carattoli et al., 2014). Where multiple hits aligned to the same locus (defined by an overlap spanning 50% of the length of the shorter sequence), best hits were selected according to percentage identity and coverage. *In silico* pMLST was conducted for IncF, IncN, IncA/C, IncHI1, IncHI2 and IncI1 Enterobacteriaceae plasmids (García-Fernández et al., 2008, 2011; Phan et al., 2009; García-Fernández and Carattoli, 2010; Villa et al., 2010; Hancock et al., 2016). pMLST was conducted using locally downloaded allele databases (http://pubmlst.org/plasmid/) with recommended identity and coverage thresholds of 85% and 66% respectively (Carattoli et al., 2014). For each allelic locus, the best hit was again selected according to percentage identity and coverage.

To perform MOB typing of Enterobacteriaceae plasmids, PSI-BLAST searches (Altschul et al., 1997) were conducted using N-terminal relaxase protein sequences as queries against a database of complete plasmid sequences, translated in all six frames. The code used to implement MOB typing has been packaged as a freely available software tool (https://github.com/AlexOrlek/MOBtyping). Searches were run for ≤14 iterations – the maximum number of iterations used by previous authors (Garcillán-Barcia et al., 2009). Initially, I used E-value thresholds in accordance with those used previously. However, using this approach, I found that certain plasmids that had been assigned a MOB type in a previous study by Smillie et al., were left unassigned (Appendix A: Methods 1 and 2). As the calculation of E-values accounts for database size, the E-value associated with a given hit is inflated when BLAST searching against a larger database (Jones and Swindells, 2002). I hypothesised that this could account for the discrepancy in MOB typing, compared to the previous study. Hence, to optimise the E-value threshold, I used a set of plasmids for which MOB typing had previously been conducted and validated (Appendix A: Methods 1 and 2). Based on this, I chose E-value thresholds as follows: MOBC, 0.001; MOBF, 0.01; MOBH, 0.01; MOBP, 1; MOBQ, 0.0001; MOBV, 0.01. After hits were produced by PSI-BLAST, no identity or coverage thresholds were applied (allowing for distant homologies to be detected); however, stringent filtering was used to select best hits (Appendix A:

Methods 1). To investigate the robustness of the MOB typing results, I re-ran PSI-BLAST searches using a different set of six MOB query proteins representing each MOB family (Appendix A: Methods 1 and 2).

Resistance genes were detected by BLAST querying the Enterobacteriaceae plasmids against the ResFinder database, using recommended identity and coverage thresholds of 98% and 60% respectively (Zankari et al., 2012). Best hits were selected as described above for replicon typing.

The bacterial plasmid dataset retrieved in 2019 (Section 2.3.2) was analysed using updated typing methods. Replicon typing was conducted using the PlasmidFinder Enterobacteriaceae and Gram-positive databases (retrieved 25th September 2019). MOB typing was conducted using the CONJscan module of MacSyFinder (for details, see Chapter 5, Section 5.3.6). The taxonomic composition of PlasmidFinder replicon sequences was determined as follows: accession ids from replicon sequence FASTA headers were extracted and used as queries to retrieve source taxon metadata from the NCBI nucleotide database.

2.3.4 Visualisation and statistical analysis

All analyses were performed using R (https://www.r-project.org/). The circlize package (Gu et al., 2014) was used to create chord diagrams to visualise associations between replicon families/types/sub-types and MOB types using an edgelist as input. A given replicon type and MOB type detected on the same plasmid were represented as a single edge.

In chord diagram visualisations of plasmid typing data, if multiple replicons of the same family/type were detected on a given plasmid, the plasmid type was considered to be the set of unique families/types detected. That is, a plasmid typed as IncFIB, IncFIC, IncFIC would be represented as IncFIB, IncFIC at the replicon type level, and as IncF at the family level. Likewise, for MOB typing, a MOBF, MOBF plasmid would be represented simply as MOBF in visualisations.

Variables associated with plasmid typing success were investigated visually and using statistical analysis. Binary logistic regression models were used to investigate whether plasmid replicon and MOB typeability were independently associated with plasmid size (log10-transformed) and the number of detected resistance genes (all resistance genes included; values truncated at the 95th percentile to reduce outlier influence).

2.4 Results and discussion

2.4.1 Characteristics of the curated Enterobacteriaceae plasmid dataset

A total of 6952 sequences representing putative Enterobacteriaceae plasmids were retrieved from the NCBI nucleotide database. Deduplicating identical sequences (2815 removed) and programmatically filtering probable non-plasmid/partial plasmid sequences (1892 removed) left 2245 accessions. Of these, 2063 met inclusion criteria (Supplementary Methods S1), whilst the remaining 182 accessions were further examined to decide upon inclusion or exclusion; 148 were excluded according to methods described above (104 not annotated as 'complete', 32 found to contain MLST loci, 12 manually filtered) leaving 2097 complete Enterobacteriaceae plasmids for analysis (Appendix A: Table 1). The plasmid sequence files have been made publicly available (https://figshare.com/s/18de8bdcbba47dbaba41), and can be utilized in various areas of plasmid research (Orlek et al., 2017a).

Enterobacteriaceae plasmids ranged in size from 1.3 kb to 794 kb. As reported previously (Shintani et al., 2015), the plasmid size distribution was log-bimodal (see Orlek et al., 2017b). The source organisms were dominated by human pathogens, especially *Escherichia coli, Klebsiella pneumoniae*, and *Salmonella enterica* (see Orlek et al., 2017b). PacBio sequencing technology has clearly been a key driver of the increase in complete plasmid sequences since 2014 (Figure 1; data prior to 2014 not shown due to a lack of annotation for sequencing technology). The first complete plasmids sequenced

using Oxford Nanopore technology were added in June 2016 (Both et al., 2016) and this technology will likely drive further expansion in availability of complete plasmid sequences in future.

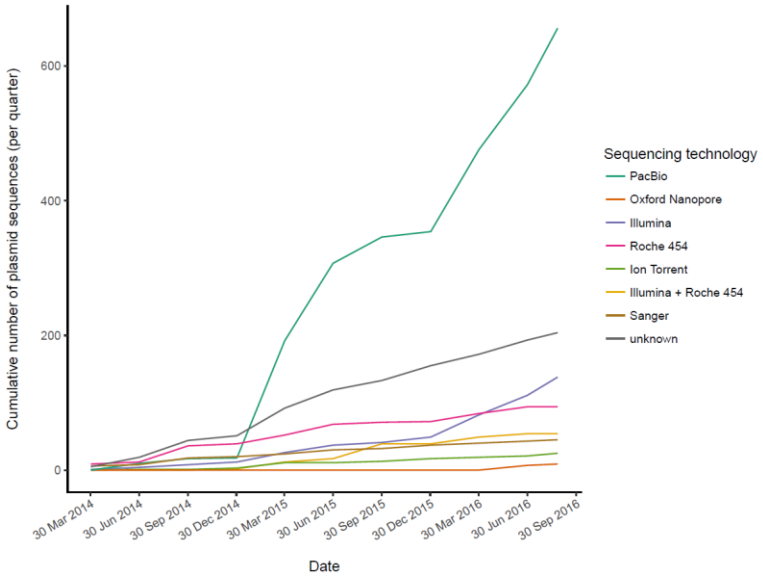


Figure 1. Complete plasmid accessions added to NCBI since 2014. Where a hybrid short-read/long-read sequencing approach was used (as indicated by the accession metadata), this is represented as long-read only (PacBio or Oxford Nanopore).

2.4.2 Assessment of Enterobacteriaceae plasmid database curation methods

Sequence topology annotation (circular/linear) was an unreliable indicator of complete plasmid sequences, with 57 of the 2097 curated plasmids annotated as 'linear'. At least one of these accessions represents a genuine linear plasmid (accession NC_011422) (Baker et al., 2007) but remaining accessions may represent mis-annotations, since 'linear' is set as the default value on submission (Fetchko and Kitts, 2011). Not counting accessions excluded as duplicates, 190 excluded accessions were annotated as 'circular'. Of these, 27 were excluded due to MLST allele detection (indicating an accession was likely chromosomal), and four were excluded following manual inspection. The majority of the other excluded 'circular' accessions were filtered due to being described as 'whole genome shotgun' sequences which should only be used to describe incomplete genomes (NCBI, 2014).

Additional manual review of these accessions might have identified some that would warrant inclusion as complete plasmids.

2.4.3 Assessment of typing performance: typeability

Over 1000 complete Enterobacteriaceae plasmids have been added to NCBI from 2014 to 2016 (Figure 2A), underscoring the need to re-assess the performance of plasmid typing schemes. Overall, a replicon type was detected in 1784 (85%) plasmids whilst a MOB type could be assigned to 1371 (65%) plasmids; together the schemes typed 1872 (89%) plasmids (Appendix A: Table 1). When typeability was assessed over time (based on the date on which an accession became available), the proportion of plasmids replicon typed remained relatively constant, whilst the proportion of plasmids MOB typed tended to increase (Figure 2B). This increase may reflect a bias in the size of available plasmid accessions over time (Supplementary Figure S1); specifically, larger plasmids which became available later were also more likely to be MOB typed (see Section 2.4.3.1). Previous authors reported in 2010 that only around half of publicly available plasmids from Gammaproteobacteria – the taxonomic class to which the Enterobacteriaceae family belongs – could be MOB typed (Smillie et al., 2010). However, the increase in proportion of MOB-typeable plasmids over time would seem to explain this discrepancy (Figure 2B); of the plasmids in my dataset that were available in 2010, roughly half could be assigned a MOB type, in agreement with previous authors.

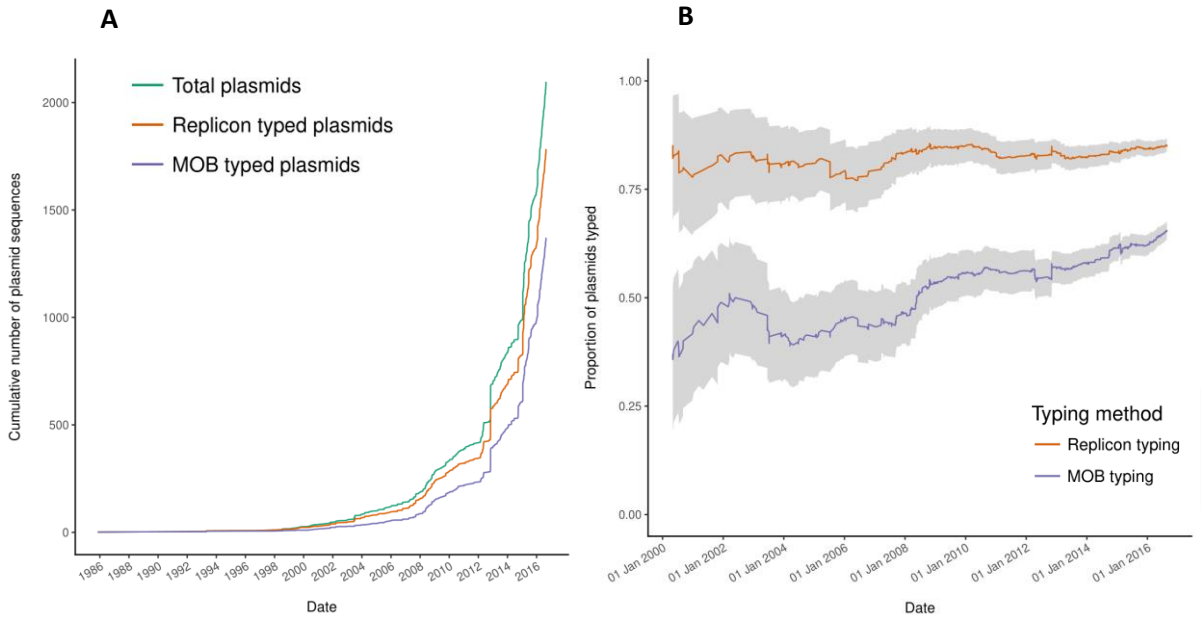


Figure 2. (A) Cumulative number of plasmids added to NCBI from initial accession (7th November 1985) to sequence retrieval date (26th August 2016). Cumulative counts reflect all plasmids (green), as well as the subsets that were replicon typed (red) and MOB typed (blue). (B) Proportion of plasmids added to NCBI prior to a given date that could be typed by replicon typing (red) and MOB typing (blue). Grey shading around the lines represents a 95% binomial confidence interval, calculated by the Agresti-Coull method.

The finding that only 85% of plasmids could be replicon typed contrasts with results of Carattoli et al., who demonstrated 100% typing success. Carattoli et al. used a selection of clinically relevant Enterobacteriaceae plasmids to design and assess *in silico* replicon typing. When plasmids from non-clinical taxa were excluded from my analyses, the typeability of replicon typing increased to 92% (1675/1818 plasmids were assigned a replicon type) (Appendix A: Table 1). This suggests that non-clinically relevant taxa are less likely to be replicon typed, because their replicon loci are not represented in PlasmidFinder. With non-clinical taxa excluded, the typeability of MOB typing also increased, from 65% to 73% (1321/1818 plasmids were assigned a MOB type). A more detailed investigation of the factors influencing typeability is provided in Sections 2.4.3.1 and 2.4.3.2.

One suggested advantage of MOB typing lies in its ability to detect divergent plasmid backbones not detected by the replicon typing scheme (Garcillán-Barcia and de la Cruz, 2013). Supporting this, 88 plasmids (4% of total) were MOB typed, but not replicon typed. Interestingly, of these 88 plasmids,

non-clinical plasmids were over-represented (47% versus 13% of plasmids in the whole dataset), which may reflect the fact that *in silico* replicon typing has only been validated for clinical plasmids (Carattoli et al., 2014).

Another aspect of typeability is the extent to which the types detected are unambiguous; more specifically, when multiple types are detected on the same plasmid, this makes it difficult to interpret the phylogenetic affiliation of the plasmid, especially when these types belong to different families of replicon/MOB type. Multi-replicon types are thought more common than multi-type MOB types, and this was also supported in my dataset. Of the 1784 plasmids replicon typed in total, 502 (28%) contained multiple replicons, of which 167 (9% of the total) represented plasmids with replicons belonging to different replicon families. In comparison, of the 1371 plasmids MOB typed in total, only 58 (4%) contained multiple relaxases, and of these, 31 (2% of the total) were plasmids with relaxases belonging to different MOB types.

Typeability of the pMLST schemes was also assessed (Supplementary Figure S2). pMLST was conducted on plasmids belonging to an unambiguous replicon type, for which a pMLST scheme was available; overall pMLST could be conducted on 868/1784 (48%) replicon typed plasmids. Of these 868 plasmids, 82% could be assigned a known pMLST type. With the exception of the IncHI1 scheme (represented by only six plasmids), the pMLST schemes did not achieve 100% typeability. Failure to assign a pMLST type was either due to no allele(s) being detected, or the detected alleles not corresponding to a known allelic profile (see Supplementary Figure S2). My findings suggest plasmid backbones are more diverse than envisaged when the pMLST schemes were originally devised.

2.4.3.1 Association between plasmid characteristics and typeability

I wanted to investigate whether plasmids that can be replicon/MOB typed tend to exhibit particular characteristics (e.g. size, resistance genes) relative to non-typed plasmids. Previous research has

indicated that relaxase-encoding (MOB typeable) plasmids tend to be larger on average than non-transmissible plasmids (Smillie et al., 2010), so plasmid size was investigated as a predictor of typing success. As described above, plasmids from clinical Enterobacterial taxa exhibited higher replicon/MOB typeability. Because only 13% plasmids came from non-clinical taxa, clinical status was not further investigated as a predictive factor in multivariate models. Size and number of resistance genes were univariably (Figure 3) and multivariably associated with typeability; logistic regression analysis showed that plasmid size and number of resistance genes were significant independent predictors of plasmid typing success (Table 1).

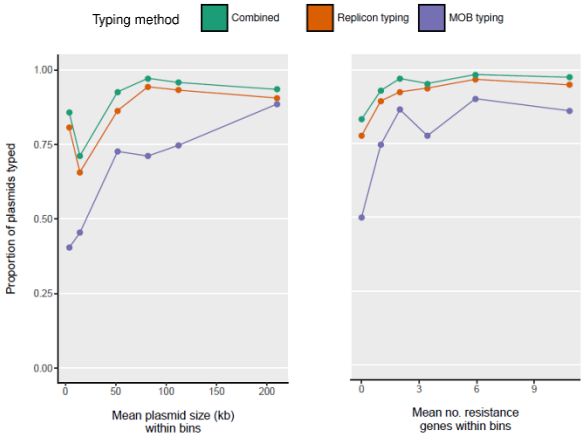


Figure 3. Relationship between typing success and different plasmid characteristics. For plasmid size, bins are equally sized (kb) 1.3–6.1; 6.1–35; 35–68; 68–95; 95–137; 137–794. For number of resistance genes (all resistance classes included), bin ranges and sizes are: 0, n=1089; 1, n=311; 2, n=134; 3–4, n=192; 5–8, n=183; 9–34, n=194.

Table 1. Logistic regression analysis of plasmid typing success

Explanatory variables	B (SE)†	Odds ratio (95% CI)††
Model of replicon typing success		
Plasmid size (log10 kb)	0.40 (0.10)***	1.49 (1.22–1.81)
Number of resistance genes	0.25 (0.04)***	1.28 (1.19–1.39)
Model of MOB typing success		
Plasmid size (log10 kb)	1.04 (0.09)***	2.82 (2.38–3.34)
Number of resistance genes	0.18 (0.02)***	1.20 (1.14–1.26)

[†]B denotes explanatory variable coefficients; SE is standard error; [††]Odds ratios indicate the change in odds resulting from a unit change in the value of the explanatory variable. Values >1 indicate that as the value of the explanatory variable increases the odds of typing success increase. ***indicates p<0.0001.

The odds of a plasmid being replicon typed were 1.49 times higher per $\log_{10}$ kb increase in plasmid size and 1.28 times higher per additional resistance gene. The odds of a plasmid being MOB typed were 2.82 times higher per $\log_{10}$ kb increase in plasmid size and 1.2 times higher per additional resistance gene (Table 1). Therefore, plasmids for which a replicon or MOB type could be assigned were generally larger and/or encoded more resistance genes. Given that antibiotic resistance often occurs in clinical settings, encoding one or more resistance genes may be a proxy for clinical Enterobacterial source taxa, which in turn is positively associated with typeability, based on univariate analysis (see above). Figure 3 indicates that relative to replicon typing, MOB typing exhibits especially poor typeability for small (less than ~50 kb) plasmids encoding no resistance genes. With respect to the influence of size on MOB typing success, one biological explanation is as follows: laboratory studies suggest transformation efficiency is higher for smaller plasmids (Chan et al., 2002); therefore, encoding a relaxase may confer small plasmids less selective advantage relative to larger plasmids (which may depend more on conjugative transfer).

Figure 6A shows that the typeability of MOB typing varied considerably between plasmids belonging to different replicon families. The majority of Col plasmids were not assigned a MOB type. This may be linked with the small average size of Col plasmids (mean size ~6.7 kb) combined with the particularly poor typeability of MOB typing relative to replicon typing for plasmids of this size (Figure 3).

2.4.3.2 Association between host bacterial taxonomy and typeability

Results so far suggest an association between typeability and (clinical/non-clinical) Enterobacterial source taxa. I wanted to further investigate the relationship between typeability and taxonomy, across

all bacterial taxa. In the case of replicon typing, I wanted to understand how typeability relates to the taxonomic composition of the PlasmidFinder database (Enterobacteriaceae and Gram-positive components). Typeability analysis was performed on an updated dataset of 14143 bacterial plasmids, representing >200 families; Enterobacteriaceae was the dominant family, making up close to a third of all accessions.

When typeability is analysed across bacterial families, replicon typing shows particularly pronounced variability (Figure 4); among Gram-negative families, Enterobacteriaceae and Yersiniaceae show high replicon typeability (90% and 84% replicon typed, respectively). The high replicon typeability of plasmids from these families reflects the taxonomic composition of replicons in the Enterobacteriaceae PlasmidFinder database (Figure 5A). Among Gram-positive families, Staphylococcaceae and Enterococcaceae plasmids showed the highest replicon typeability (88% and 78%, respectively), followed by Streptococcaceae and Lactobacillaceae (46% and 38%, respectively). Likewise, this reflects the content of the Gram-positive PlasmidFinder database (Figure 5B).

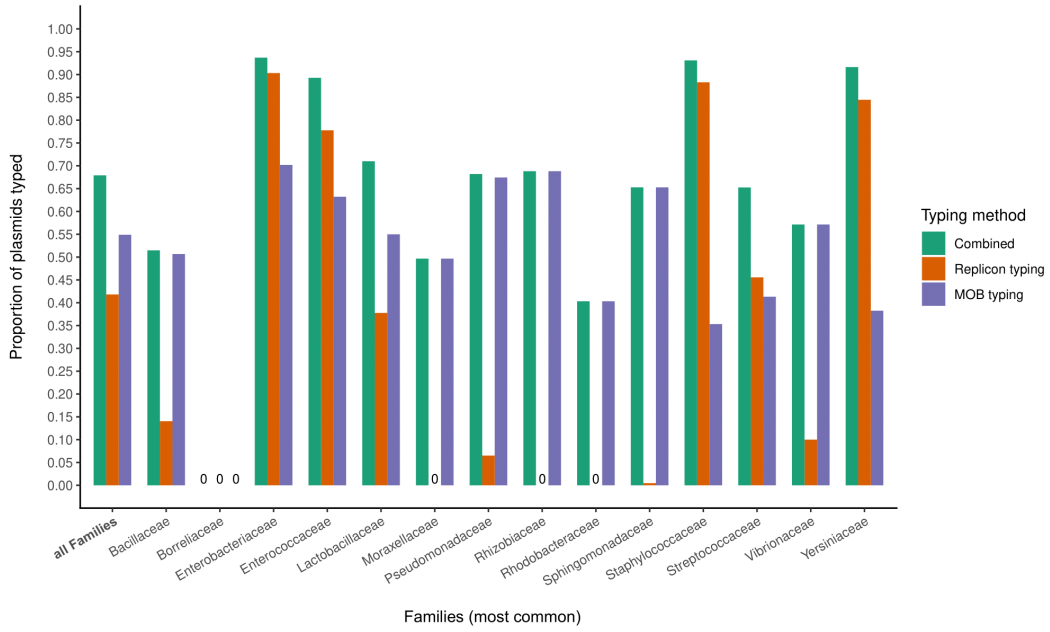


Figure 4. Typing success of replicon typing, MOB typing, and both combined. Results are shown for all plasmids, and also broken down according to the most common families (>200 plasmids).

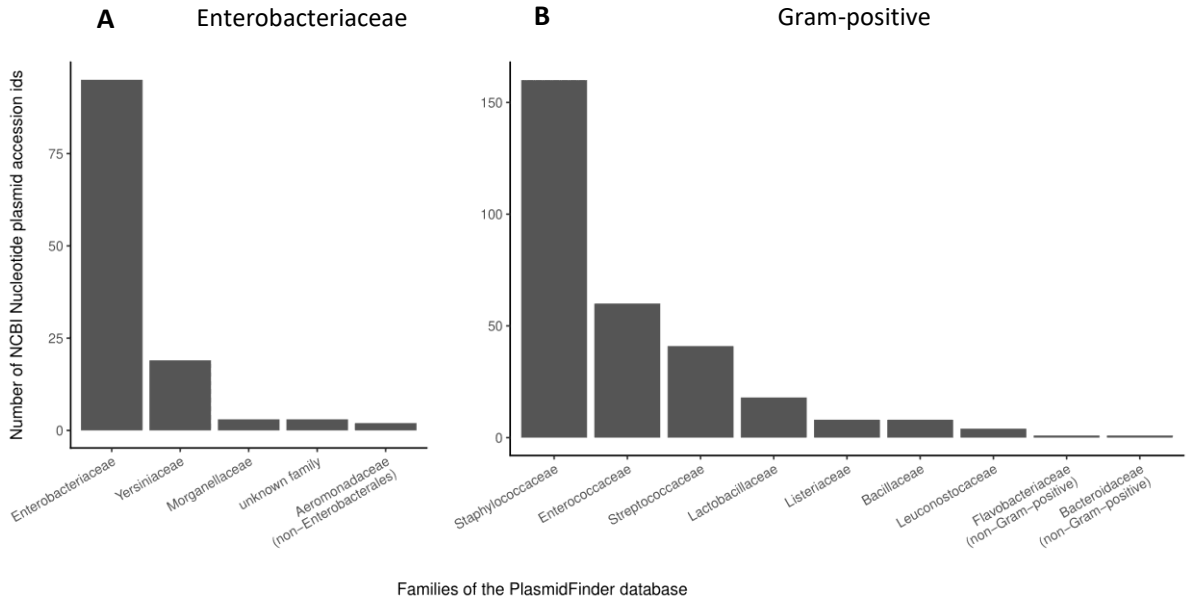


Figure 5. Taxonomic composition of the PlasmidFinder database (version retrieved 25th September 2019). The number of NCBI Nucleotide plasmid accession identifiers represented among PlasmidFinder replicons is broken down by bacterial families of the (A) Enterobacteriaceae and (B) Gram-positive database components.

The Enterobacteriaceae and Yersiniaceae plasmids which were detected by replicon typing were detected almost exclusively by Enterobacteriaceae PlasmidFinder replicons (with two exceptions for Enterobacteriaceae plasmids). Conversely, Gram-positive plasmids which were detected by replicon typing were mostly detected by replicons belonging to the Gram-positive PlasmidFinder database (Table 2). Overall, the replicon typing results suggest that, with current methods, PlasmidFinder replicons can detect plasmids from the same (family-level) taxonomic groups to those that they were isolated from. It follows that plasmids from families not covered by replicons within PlasmidFinder are generally not replicon typed, which explains why family-level typeability results mirror the family-level taxonomic composition of the PlasmidFinder database. The results likely reflect the *in silico* replicon typing methods per se (BLASTN cannot detect distant homology), combined with barriers to plasmid transfer across broad taxonomic ranges, meaning that plasmid replicon composition is structured by taxonomy, at least at higher taxonomic levels. Indeed, recent findings suggest that plasmids generally do not transfer frequently between higher-level taxonomic boundaries (Acman et al., 2020). As

described previously, MOB typing is accomplished using methods allowing for detection of more distant homology; as a result, MOB typing is thought to be a more inclusive scheme. Accordingly, when aggregating typing results across all plasmids, MOB typing shows higher typeability than replicon typing (55% vs 42%).

Table 2. The number of plasmids detected by replicons from the Enterobacteriaceae and Gram-positive PlasmidFinder databases. Results are broken down according to the most common families (>200 plasmids). Common families in which no plasmids were replicon typed are not shown.

	Number of plasmids detected by PlasmidFinder	
Family (number of plasmids)	Enterobacteriaceae database	Gram-positive database
Enterobacteriaceae (4630)	4176	2
Bacillaceae (748)	22	82
Lactobacillaceae (662)	0	250
Staphylococcaceae (521)	0	458
Enterococcaceae (261)	0	203
Pseudomonadaceae (261)	17	0
Streptococcaceae (259)	0	118
Yersiniaceae (251)	212	0
Sphingomonadaceae (216)	1	0
Vibrionaceae (210)	21	0

There are four major families for which no plasmids could be replicon typed (Borreliaceae, Morexellaceae, Rhizobiaceae, and Rhodobacteraceae) and one family with no MOB-typed plasmids (Borreliaceae). The finding that Borreliaceae plasmids are not MOB-typeable has been remarked upon before. Smillie et al. point out that the genus *Borrelia* (within the Borreliaceae family) comprises obligate intracellular parasites. Hence, there may be reduced opportunities for plasmid transfer, so encoding relaxases may not be adaptive (Smillie et al., 2010). Literature searching suggests that the lack of replicon typeable plasmids in the aforementioned taxa is not solely due to a fundamental lack of biological knowledge, since there is research on replication loci among plasmids within these taxa

(Bertini et al., 2010; Chaconas and Kobryn, 2010; Petersen et al., 2011; Mazur and Koper, 2012). If knowledge about plasmid replicons from a wider range of taxa can be curated and incorporated into the PlasmidFinder database, the usefulness of replicon typing will be enhanced. However, the current results highlight the limitation of plasmid typing, especially when dealing with taxonomically diverse sets of plasmids.

2.4.4 Assessment of typing performance: concordance

The schemes showed a degree of phylogenetic concordance, with some replicon families nested entirely within a single MOB type, consistent with MOB typing being a phylogenetically broader scheme. Col plasmids are not shown at the family level, in Figure 6B, since non-concordance is to be expected; Col plasmids include phylogenetically distinct plasmids, and do not encode a single replicon type. Indeed, Col plasmids replicate by different mechanisms: in some cases, a plasmid-encoded Rep protein initiates replication; in other cases (ColE1 and ColE1-like plasmids) replication is initiated by binding of a transcript called RNAII, and inhibited by antisense RNAI (del Solar et al., 1998). Accordingly, the *in silico* replicon typing scheme targets several different Col plasmid replicon loci, including RNAI-encoding loci as well as loci encoding RepA replication proteins; these proteins in turn belong to distinct protein families (Appendix A: File 1). Of the remaining plasmids that were investigated at the family level, IncA/C, IncF, IncH, IncQ, IncU and IncX replicon families did not show complete concordance (Figure 6B); in cases where only a few replicon families nest outside a primary MOB type (e.g. IncA/C) non-concordance is more clearly observed by inspecting Appendix A: Table 1. For plasmids with multiple different replicon families detected, patterns of replicon–MOB type association correspond with those of the constituent single family replicons (see Orlek et al., 2017b). For example, IncA/C, IncN type was associated with MOBF, MOBH type, consistent with Figure 6B (IncN associated with MOBF and IncA/C primarily associated with MOBH). Compared with previous reports of non-concordance for just IncQ and IncP families and associated MOB types (Garcillán-Barcia

et al., 2011), I found more widespread non-concordance, although I did not find non-concordance for IncP, presumably reflecting the narrower taxonomic scope of this study, in which IncP plasmids were infrequent.

There are several explanations for the patterns of non-concordance in Figure 6B, as described previously. To investigate whether non-concordance reflects fuzzy boundaries between MOB families, I examined PSI-BLAST hits to identify cases where different MOB query proteins aligned to the same locus on a given plasmid. This would indicate that the relaxase detected at that locus showed homology to different MOB queries, reflecting overlap between the corresponding MOB protein families; this could in turn result in non-concordance between typing schemes. There were 204 loci from 201 plasmids involved in such alignments and the MOB queries involved were all MOBP/MOBQ (Appendix A: Table 2). This finding reiterates previous reports of overlap between MOBP and MOBQ families and is supported by subsequent analysis in this study with a different set of MOB queries (see Section 2.4.5). Accordingly, MOBP/Q overlap could potentially explain the pattern of non-concordance observed for IncQ and IncX replicon families, which each associate with both MOBP and MOBQ. It could also explain the non-concordance of ColRNAI and Col(pWES) replicon types (Figure 7A).

To gain additional insight, I further examined alignments involving different MOB queries at the same locus. If the best two hits at a locus involve different MOB queries, the assigned MOB type is presumably less reliable, compared with a situation where top hits involve the same query; therefore, MOBP/Q overlap would seem a more plausible explanation for non-concordance in the former compared with the latter situation. Applying this reasoning to my data, I found that IncX plasmids were commonly associated with (putatively) less reliable MOB type assignments, as was one ColRNAI plasmid (accession NC_013509.1) (Appendix A: Table 2). This ColRNAI plasmid, having MOBQ and subsequently MOBP as top hits, corresponds to the single MOBQ-associated ColRNAI plasmid represented in Figure 7A. In contrast, for IncQ plasmids, although there are cases where different queries align to the same locus, the top five hits at a locus consistently involve the same MOB query

(Appendix A: Table 2); this suggests MOB type assignments for IncQ plasmids are reliable. These findings are supported by analyses in Section 2.4.5.

Given my findings, for IncQ plasmids, as well as remaining replicon families not associated with MOBP and MOBQ, backbone mosaicism provides a more plausible alternative explanation for non-concordance (Osborn et al., 2000). To investigate further, I examined the extent to which non-concordance could be resolved by partitioning plasmids into finer resolution replicon/pMLST types. If non-concordance stems from mosaicism that arose deep in plasmid evolutionary history, partitioning plasmids into more homogenous groups might mean such groups would show concordance.

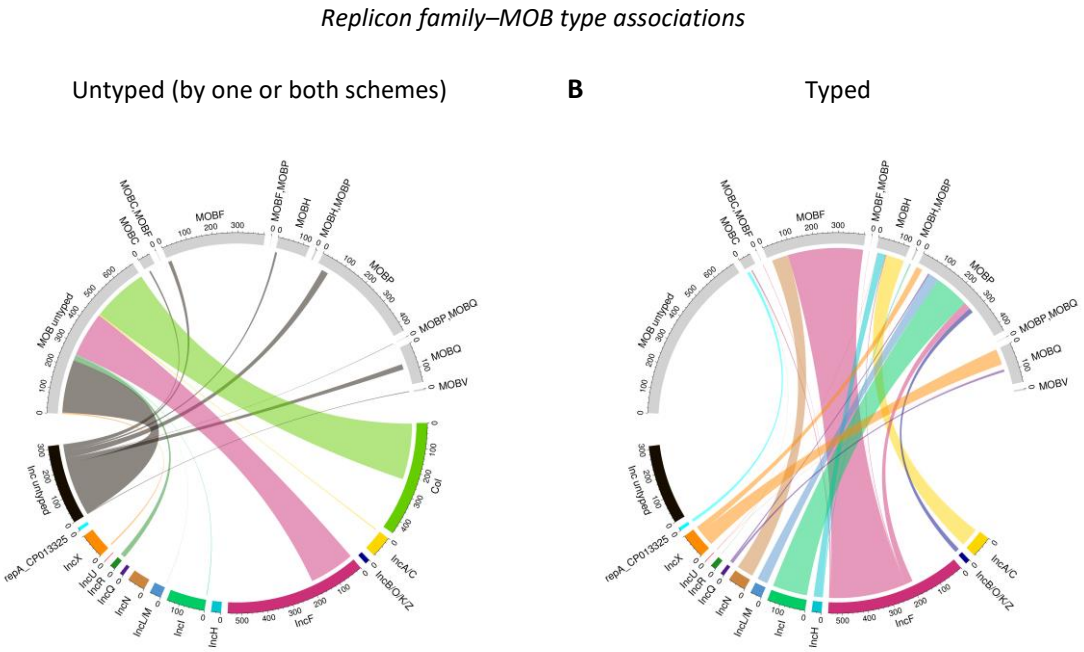


Figure 6. Chord diagrams illustrate associations between replicon families and MOB types; circularly arranged sectors represent replicon families and MOB types, and scale bars indicate their relative sizes. Associations are indicated by intersecting chords. Replicon family sectors, coloured; MOB type sectors, grey. (A) Replicon family-MOB type associations amongst plasmids un-typed by one or both schemes. (B) Replicon family-MOB type associations amongst plasmids with both replicon and MOB type detected. Data on Col plasmids are deliberately not shown (see main text). Plasmid types are represented as the unique set of families detected on a plasmid, as described in Methods (i.e. IncF, IncF = IncF). Where multiple types of different replicon families are detected,

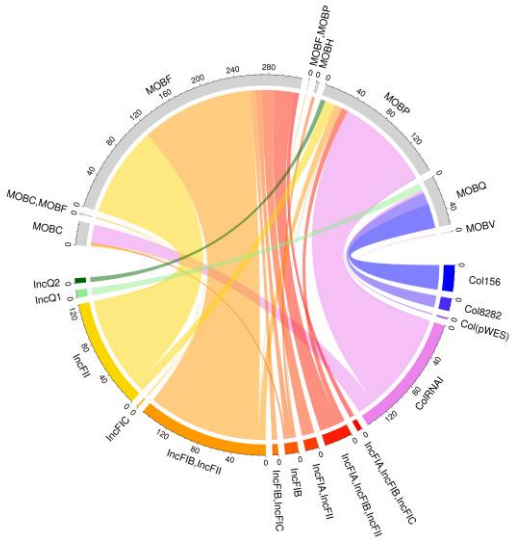
data are not shown. Replicon family types detected in fewer than 10 plasmids are not shown except where a non-concordant pattern of association is observed (IncU) or where associated with a multi-type MOB type.

Figure 7A and B show that partitioning plasmids into finer-resolution replicon types resolves non-concordance for the IncQ family, with IncQ1 and IncQ2 associated with MOBQ and MOBP respectively. This reflects a previous study (Garcillán-Barcia et al., 2011), and is in line with the known biology of IncQ plasmids. Specifically, IncQ2 plasmids are known to encode a mobilization system unrelated to IncQ1 plasmids but resembling that of IncP plasmids, indicating backbone mosaicism (Rawlings and Tietze, 2001). For other replicon families, some replicon types show concordance, but ColRNAI and Col(pWES) remain non-concordant as do IncA/C2, IncHI1, and many of the IncF replicon types. IncX replicon types are largely concordant, but this is unlikely to be attributed to resolution of mosaicism (see Section 2.4.5).

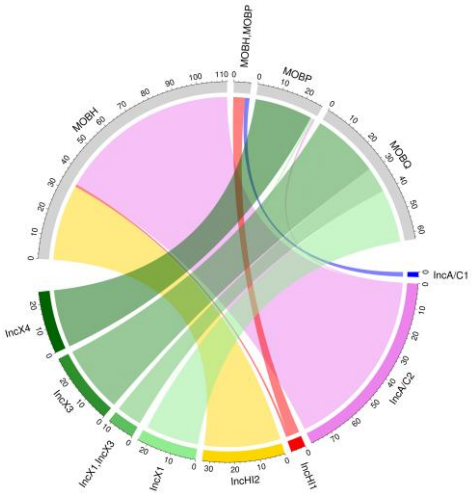
Concordance is further improved by partitioning replicon types into pMLST types (Figure 7C and D). However, some IncF pMLST types show non-concordance: F17:A-:B-, F12:A-:B-, F6:A-:B- and F4:A-:B-, F1:A-:B17. Non-concordant typing of IncF plasmids is perhaps unsurprising given the available literature on plasmid backbone mosaicism. Specifically, a study of IncFII replicons demonstrated mosaicism, with homology to the closely-related IncB, IncK and IncZ replicons (treated as IncB/O/K/Z in the Carattoli et al. *in silico* scheme), putatively due to recombination events (Praszkier et al., 1991). In my dataset, IncB/O/K/Z is associated with MOBP (Figure 6B) as are some IncFII replicon types (Figure 7A), providing an explanation for IncFII non-concordance. Overall, the finding of non-concordance across hierarchical resolutions for IncF replicons indicates that non-concordance in this family may in part reflect backbone mosaicism generated by relatively recent evolutionary events.

Replicon type–MOB type associations (non-concordant at replicon family-level)

A Col, IncF, IncQ types

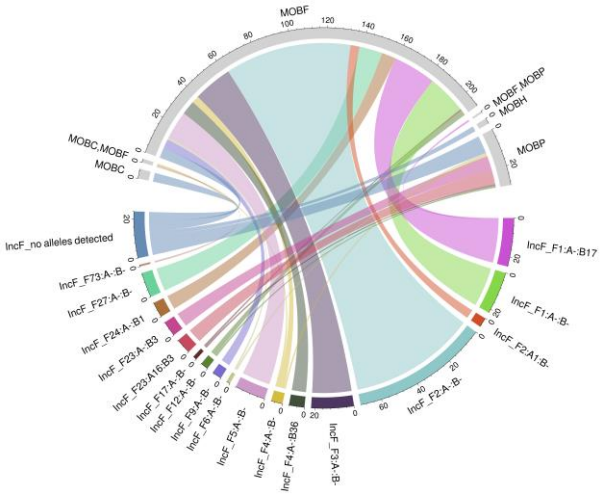


B IncA/C, IncH, IncX types



pMLST–MOB type associations (non-concordant at replicon type-level)

C IncF pMLSTs



D IncA/C pMLSTs

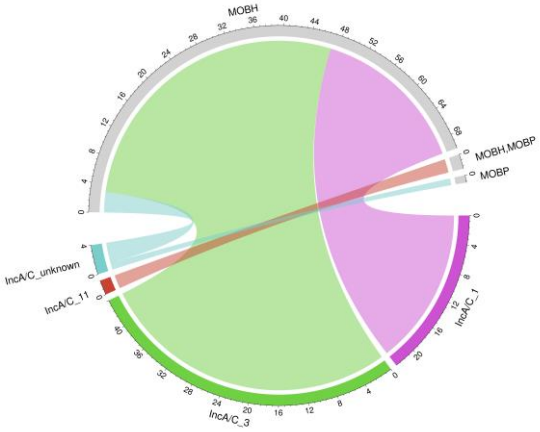


Figure 7. Chord plots A and B illustrate associations between replicon and MOB types, for replicon types belonging to non-concordant replicon families. Where replicon types are shown to be non-concordant and a pMLST scheme is available, chord plots C and D show associations between pMLST types and MOB types; IncHI1 pMLST associations are not shown due to the small number of plasmids belonging to this subtype. Plasmid types are represented as the unique set of types detected on a plasmid, as described in Methods (i.e. IncFIA, IncFII, IncFII = IncFIA, IncFII). Replicon types and pMLST types detected in fewer than 5 plasmids are not shown except where a non-concordant pattern of association is observed, or where associated with a multi-type MOB type. Associations with untyped plasmids are not shown. Col family replicon types correspond to the replicon probe names in the PlasmidFinder database. IncF pMLST is based on the so-called FAB formula where sequence types

are determined according to the allele for IncFII, IncFIA and IncFIB. As an example, the most common IncF pMLST type is F2:A-:B- which reflects detection of allele 2 for IncFII, and detection of no alleles for IncFIA and IncFIB.

2.4.5 Assessment of the robustness of MOB typing results, using alternative MOB queries

PSI-BLAST searches against the same database but using different queries may retrieve a different set of hits (Bhagwat and Aravind, 2007). Therefore, to assess the robustness of my MOB typing results, I conducted MOB typing using an alternative set of six MOB queries representing each MOB type. 54% of plasmids were MOB typed – this is lower than the 65% achieved in my original analysis, and only 6 plasmids not previously assigned a MOB type could be MOB typed using the alternative prototypes. Overall, the re-analysis supports the principal findings from my primary analyses which suggests that my conclusions are robust: for example, typing schemes were again found to be partially concordant, and concordance improved when plasmids were partitioned into higher resolution replicon types/sub-types (for details see Orlek et al., 2017b).

However, there are some key differences in the assignment of plasmid types using the original vs. alternative sets of MOB queries, as summarised in Appendix A: Table 3. Where a plasmid is assigned a different MOB type using the original vs. alternative set of queries, this indicates that the MOB type assignment may be unreliable. Crucially, such discrepancies provide a complementary way to assess the conclusions in Section 2.4.4 regarding MOBP/Q overlap as an explanation for non-concordance of the IncX replicon family and Col accession NC_013509.1. Overall, plasmids assigned a MOBQ type with original queries were more likely to be assigned a MOBP type using alternative queries (there were 94 plasmids for which this is the case). This underscores the ambiguous boundaries between MOBP and MOBQ families. Of plasmids assigned a different MOB type using alternative queries, the majority belong to IncX family. When typing with alternative queries, the IncX replicon family is entirely nested within MOBP, whereas primary analysis demonstrated non-concordance at the family-level, with the majority of IncX replicon types, except for IncX4, associated with MOBQ. One Col plasmid (NC_013509.1) was assigned a different MOB type using alternative queries; this is the same Col

plasmid identified in Section 2.4.4 as having an unreliable MOB type. Also in support of conclusions in Section 2.4.4, there were no discrepancies in the typing of IncQ plasmids. Hence, findings from my primary analyses are corroborated: MOBP/Q overlap can complicate typing, and cause non-concordance between replicon and MOB typing schemes, as appears to be the case for IncX non-concordance in Figure 6B.

2.5 Conclusions

Retrieval and curation of a dataset of complete plasmids and associated metadata from NCBI requires bioinformatics expertise. Without such a dataset, even the most basic aspects of plasmid biology, such as the size distribution, are obscured. Overall, my experience suggests that obtaining a high-quality dataset of complete plasmids requires some degree of quality-filtering, and this is likely to become more important as more sequences are added to NCBI. However, my curation methods were guided by the annotations (and mis-annotations) observed in the retrieved dataset. Building on and fine-tuning the suggested curation methods, as publicly available databases expand, would provide a valuable resource for future research. Indeed, curated plasmid datasets have a wide range of applications beyond investigation of typing schemes (Orlek et al., 2017a).

My study demonstrated that 42% and 55% bacterial plasmids available in 2019 could be typed by replicon and MOB typing schemes, respectively. Typeability was associated with plasmid characteristics including plasmid size and host taxonomy. Smaller plasmids were less likely to be MOB typed, possibly due to reduced selection pressure for conjugative transfer (transformation efficiency is higher for smaller plasmids). When typeability was assessed across bacterial families, replicon typing showed particularly pronounced variability, which mirrored the family-level taxonomic composition of replicons included in the PlasmidFinder database. The PlasmidFinder database is biased towards well-studied clinically relevant taxa, such as Enterobacteriaceae; in turn, these taxa show high typeability (90% Enterobacterial plasmids available in 2019 could be replicon typed). Conversely,

replicon typing success is especially low for plasmids falling outside well-studied taxa. As many more plasmids are sequenced, including from non-clinical strains, the proportion of plasmids that can be replicon typed may further decrease. This has clear practical implications, given the widespread use of replicon typing in plasmid epidemiology, and the use of replicon loci as plasmid-specific markers. Future work to functionally characterise plasmid genes, identify replicon loci, and incorporate novel *in silico* replicon probes into PlasmidFinder is required to detect and classify currently non-typeable plasmids.

Although previous studies have examined plasmid backbone mosaicism and the concordance of replicon/MOB typing schemes, this has not been based on comprehensive bioinformatic analysis at hierarchical replicon typing resolutions. My findings demonstrate that plasmid typing is inherently difficult and even the relatively conserved loci used for major typing schemes exhibit phylogenetic discordance, seen most notably with the IncF replicon. In this case, partitioning plasmids into more homogenous groups using pMLST improved concordance, but only partially, perhaps reflecting mosaicism resulting from relatively recent evolutionary events.

My study has also highlighted issues with reproducibility of typing results, in particular, MOB typing results. To conduct replicon typing, plasmids are searched against the PlasmidFinder database, which contains curated replicon sequences, annotated to reflect established replicon typing nomenclature (e.g. incompatibility groups). Hence, replicon typing results should be relatively stable between database versions, and reproducible if necessary (using a contemporaneous database version). On the other hand, MOB typing was achieved by PSI-BLAST querying MOB proteins (each representing a MOB type) against the curated plasmid database. Choice of MOB query proteins, as well as differences in database size and content can influence MOB typing results. PSI-BLAST search profiles may also become corrupted if non-homologous sequence is aligned (e.g. through spurious extension of a homologous alignment). Hence, to assess the robustness of MOB typing results and downstream concordance analyses, I re-ran analyses with a different set of MOB query proteins. The re-analysis

supported principal findings from primary analyses, but corroborated previous reports of overlap between MOBP and MOBQ relaxase families; hence, MOBP/MOBQ type assignments should be treated with caution.

In summary, 'Ordering the mob' denotes not only the challenging pursuit of plasmid typing, but also the process of obtaining a dataset of complete plasmids from NCBI. Using my curated plasmid dataset, I demonstrate that current typing schemes fail to classify the complete diversity of plasmids, and that there is a degree of non-concordance between replicon and MOB typing schemes. In some cases, non-concordance is likely to reflect the plasticity (and consequent mosaicism) of plasmid genomes, whilst in other cases it reflects the ambiguous boundaries between MOBP and MOBQ types.

2.6 Supplementary material

2.6.1 Supplementary Methods

Methods presented below detail the regular expressions used for the curation of the Enterobacteriaceae plasmid dataset, retrieved in 2016.

Entrez query used for initial retrieval of putative complete plasmid accessions from NCBI:

Entrez.esearch(db="nucleotide", term=str("Enterobacteriaceae[porgn] AND biomol_genomic[PROP] AND plasmid[filter] NOT complete cds[Title] NOT gene[Title] NOT genes[Title] NOT contig[Title] NOT scaffold[Title] NOT whole genome map[Title] NOT partial sequence[Title] NOT partial plasmid[Title] NOT locus[Title] NOT region[Title] NOT fragment[Title] NOT integron[Title] NOT transposon[Title] NOT insertion sequence[Title] NOT insertion element[Title] NOT phage[Title] NOT operon[Title]")

Filtering non-plasmid or non-complete plasmid accessions

The following regular expression term was used to filter non-plasmid and non-complete plasmid accessions based on the description in the accession title:

'contig|\sgene(?!tic|ral|rat|ric)|integron|transposon|scaffold|insertion sequence|insertion element|phage|operon|partial sequence|partial plasmid|region|fragment|locus|complete (?!sequence|genome|plasmid|\.|,)|(?<!complete sequence,)whole genome shotgun|artificial|synthetic|vector'

Having excluded accessions using the above term, the following term was used as an inclusion criterion for the remaining accessions:

'complete(?= sequence| genome| plasmid)'

For accessions matching neither exclusion nor inclusion criteria, MLST typing was conducted to exclude chromosomal accessions, and manual filtering was conducted, focusing on the tails of the size distribution.

A similar but distinct regular expression search was used for curation of the bacterial plasmid dataset in 2019. For details see Appendix A: Methods 3.

2.6.2 Supplementary Figures

Figures are based on analysis of the 2097 curated Enterobacteriaceae plasmids retrieved in 2016.

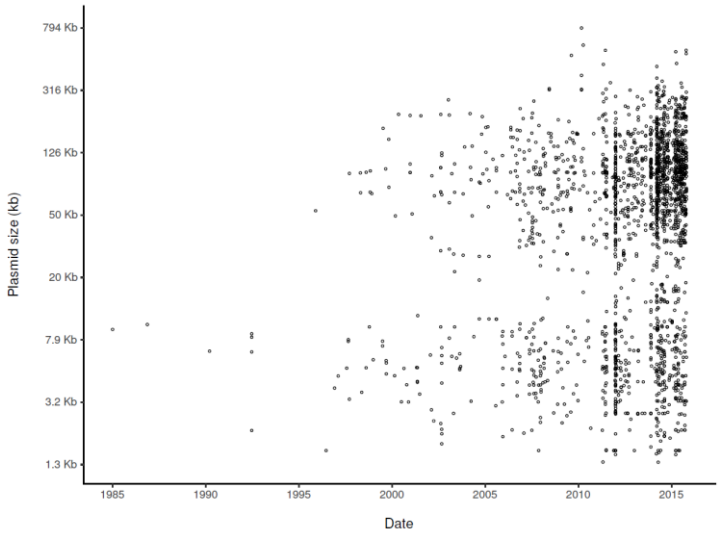


Figure S1. Plasmid size in relation to the date on which a plasmid was added to NCBI. Plasmid size is on a log10 logarithmic scale, and axis values are shown in terms of kilobase pairs.

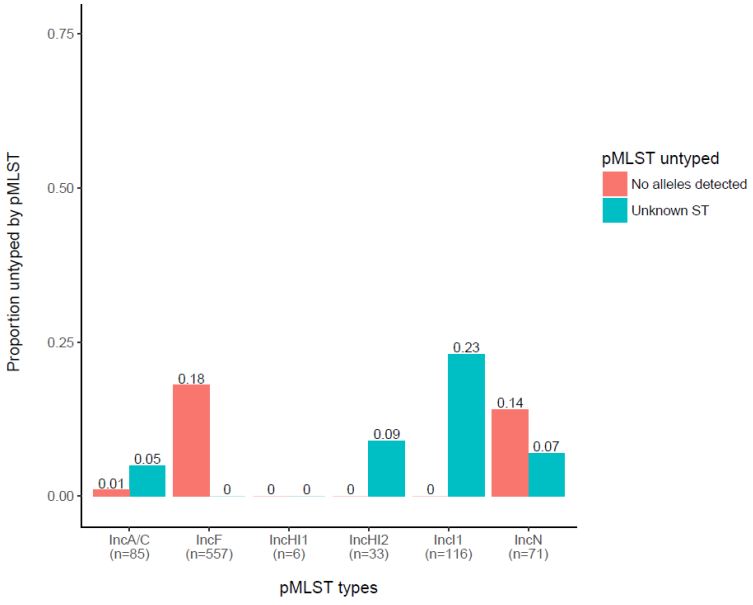


Figure S2. Proportion of plasmids that could not be assigned to a pMLST type following pMLST typing using a given scheme. Failure of pMLST typing may be due to no alleles being detected (pink) or detected alleles not matching a known allelic profile (turquoise). The number of plasmids on which pMLST typing using a given scheme was conducted is indicated on the x-axis. Values above bars indicate proportions. Note that IncF pMLST cannot produce an unknown ST since STs are assigned according to the FAB formula rather than being based on correspondence to a recognised allelic profile.

Bibliography

Acman, M., van Dorp, L., Santini, J. M., and Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* doi:10.1038/s41467-020-16282-w.

Adeolu, M., Alnajar, S., Naushad, S., and Gupta, R. S. (2016). Genome-based phylogeny and taxonomy of the '*Enterobacteriales*': proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov.,. *Int. J. Syst. Evol. Microbiol.*

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389.

Alvarado, A., Garcillán-Barcia, M. P., and de la Cruz, F. (2012). A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. *PLoS One* 7. doi:10.1371/journal.pone.0040438.

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., and Schürch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics*. doi:10.1099/mgen.0.000128.

Baker, S., Hardy, J., Sanderson, K. E., Quail, M., Goodhead, I., Kingsley, R. A., et al. (2007). A novel linear plasmid mediates flagellar variation in Salmonella Typhi. *PLoS Pathog.* 3, 0605–0610. doi:10.1371/journal.ppat.0030059.

Bertini, A., Poirel, L., Mugnier, P. D., Villa, L., Nordmann, P., and Carattoli, A. (2010). Characterization and PCR-based replicon typing of resistance plasmids in Acinetobacter baumannii. *Antimicrob. Agents Chemother.* 54, 4168–4177. doi:10.1128/AAC.00542-10.

Bhagwat, M., and Aravind, L. (2007). PSI-BLAST tutorial. *Methods Mol. Biol.* 395, 177–86. doi:1-59745-514-8:177 [pii].

Bianco, P. R., and Kowalczykowski, S. C. (1997). The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5'-GCTGGTGG-3'. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6706–11. doi:10.1073/pnas.94.13.6706.

Both, A., Huang, J., Kaase, M., Hezel, J., Wertheimer, D., Fenner, I., et al. (2016). First report of Escherichia coli co-producing NDM-1 and OXA-232. *Diagn. Microbiol. Infect. Dis.* 86, 437–438. doi:10.1016/j.diagmicrobio.2016.09.005.

Brolund, A., and Sandegren, L. (2015). Characterization of ESBL disseminating plasmids. *Infect Dis* 4235, 1–8. doi:10.3109/23744235.2015.1062536.

Campos, F. S., Cerqueira, F. B., Santos, G. R., Pereira, E. J. G., Corrêia, R. F. T., Cangussu, A. S. R., et al. (2019). Complete Sequences of Two Plasmids Found in a Brazilian Bacillus thuringiensis Serovar israelensis Strain . *Microbiol. Resour. Announc.* doi:10.1128/mra.00051-19.

Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K. L., and Threlfall, E. J. (2005). Identification of plasmids by PCR-based replicon typing. *J. Microbiol. Methods* 63, 219–228. doi:10.1016/j.mimet.2005.03.018.

Carattoli, A., Zankari, E., Garcia-Fernández, A., Larsen, M. V., Lund, O., Villa, L., et al. (2014). In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi:10.1128/AAC.02412-14.

Chaconas, G., and Kobryn, K. (2010). Structure, Function, and Evolution of Linear Replicons in Borrelia . *Annu. Rev. Microbiol.* doi:10.1146/annurev.micro.112408.134037.

Chan, V., Dreolini, L. F., Flintoff, K. A., Lloyd, S. J., and Mattenley, A. A. (2002). The Effect of Increasing Plasmid Size on Transformation Efficiency in Escherichia coli. *J. Exp. Microbiol. Immunol.* 2, 207–223.

Chen, J., Guo, M., Wang, X., and Liu, B. (2016). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.*, 1–14. doi:doi: 10.1093/bib/bbw108.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163.

Conlan, S., Thomas, P. J., Deming, C., Park, M., Lau, A. F., Dekker, J. P., et al. (2014). Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* 6, 254ra126. doi:10.1126/scitranslmed.3009845.

de Been, M., Lanza, V. F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., et al. (2014). Dissemination of Cephalosporin Resistance Genes between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages. *PLoS Genet.* 10. doi:10.1371/journal.pgen.1004776.

del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M., and Díaz-Orejas, R. (1998). Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* 62, 434–464. doi:1092-2172/98/$04.0010.

DiCenzo, G. C., and Finan, T. M. (2017). The Divided Bacterial Genome. *Microbiol. Mol. Biol. Rev.* doi:10.1128/ MMBR.00019-17.

Ferrés, I., and Iraola, G. (2018). MLSTar: Automatic multilocus sequence typing of bacterial genomes in R. *PeerJ*. doi:10.7717/peerj.5098.

Fetchko, M., and Kitts, A. (2011). The “Nucleotide” Page. *GenBank Submissions Handb.* Available at: https://www.ncbi.nlm.nih.gov/books/NBK293904/ [Accessed December 3, 2016].

Francia, M. V., Varsaki, A., Garcillán-Barcia, M. P., Latorre, A., Drainas, C., and De La Cruz, F. (2004). A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol. Rev.* 28, 79–100. doi:10.1016/j.femsre.2003.09.001.

García-Fernández, A., and Carattoli, A. (2010). Plasmid double locus sequence typing for IncHI2 plasmids, A subtyping scheme for the characterization of IncHI2 plasmids carrying extended-spectrum β-lactamase and quinolone resistance genes. *J. Antimicrob. Chemother.* 65, 1155–1161. doi:10.1093/jac/dkq101.

García-Fernández, A., Chiaretto, G., Bertini, A., Villa, L., Fortini, D., Ricci, A., et al. (2008). Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum β-lactamases in Escherichia coli and Salmonella of human and animal origin. *J. Antimicrob. Chemother.* 61, 1229–1233. doi:10.1093/jac/dkn131.

García-Fernández, A., Villa, L., Moodley, A., Hasman, H., Miriagou, V., Guardabassi, L., et al. (2011). Multilocus sequence typing of IncN plasmids. *J. Antimicrob. Chemother.* 66, 1987–1991. doi:10.1093/jac/dkr225.

Garcillán-Barcia, M. P., Alvarado, A., and De la Cruz, F. (2011). Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol. Rev.* 35, 936–956. doi:10.1111/j.1574-6976.2011.00291.x.

Garcillán-Barcia, M. P., and de la Cruz, F. (2013). Ordering the bestiary of genetic elements transmissible by conjugation. *Mob. Genet. Elements* 3, e24263. doi:10.4161/mge.24263.

Garcillán-Barcia, M. P., Francia, M. V., and De La Cruz, F. (2009). The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* 33, 657–687. doi:10.1111/j.1574-6976.2009.00168.x.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812. doi:10.1093/bioinformatics/btu393.

Guglielmini, J., Quintais, L., Garcillán-Barcia, M. P., de la Cruz, F., and Rocha, E. P. C. (2011). The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7. doi:10.1371/journal.pgen.1002222.

Hancock, S. J., Phan, M.-D., Peters, K. M., Forde, B. M., Chong, T. M., Yin, W.-F., et al. (2016). Identification of IncA/C Plasmid Replication and Maintenance Genes and Development of a Plasmid Multi-Locus Sequence-Typing Scheme. *Antimicrob. Agents Chemother.*, AAC.01740-16. doi:10.1128/AAC.01740-16.

Hiraga, S.-I., Sugiyama, T., and Itoh, T. (1994). Comparative Analysis of the Replicon Regions of Eleven ColE2-related plasmids. 176, 7233–7243.

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2017). A RESTful application programming interface for the PubMLST molecular typing and genome databases. *Database (Oxford).* doi:10.1093/database/bax060.

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Res.* doi:10.12688/wellcomeopenres.14826.1.

Jones, D. T., and Swindells, M. B. (2002). Getting the most from PSI-BLAST. *Trends Biochem. Sci.* 27, 161–164. doi:10.1016/S0968-0004(01)02039-4.

Kans, J. (2013). Entrez Direct: E-utilities on the UNIX Command Line. In: Entrez Programming Utilities Help. *Entrez Program. Util. Help*. Available at: http://www.ncbi.nlm.nih.gov/books/NBK179288/ [Accessed November 24, 2016].

Kulinska, A., Czeredys, M., Hayes, F., and Jagura-Burdzy, G. (2008). Genomic and functional characterization of the modular broad-host-range RA3 plasmid, the archetype of the IncU group. *Appl. Environ. Microbiol.* 74, 4119–4132. doi:10.1128/AEM.00229-08.

Lanza, V. F., de Toro, M., Garcillán-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., et al. (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. doi:10.1371/journal.pgen.1004766.

Larsen, M. V., Cosentino, S., Lukjancenko, O., Saputra, D., Rasmussen, S., Hasman, H., et al. (2014). Benchmarking of methods for genomic taxonomy. *J. Clin. Microbiol.* 52, 1529–1539. doi:10.1128/JCM.02981-13.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* 50, 1355–1361. doi:10.1128/JCM.06094-11.

Leverstein-van Hall, M. A., Dierikx, C. M., Cohen Stuart, J., Voets, G. M., van den Munckhof, M. P., van Essen-Zandbergen, A., et al. (2011). Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. *Clin. Microbiol. Infect.* 17, 873–880.

doi:10.1111/j.1469-0691.2011.03497.x.

Low, A. J., Koziol, A. G., Manninger, P. A., Blais, B., and Carrillo, C. D. (2019). ConFindr: Rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. *PeerJ*. doi:10.7717/peerj.6995.

Mazur, A., and Koper, P. (2012). Rhizobial plasmids - replication, structure and biological role. *Cent. Eur. J. Biol.* doi:10.2478/s11535-012-0058-8.

McAdam, A. J. (2020). Enterobacteriaceae? Enterobacterales? What should we call enteric gram-negative bacilli? A micro-comic strip. *J. Clin. Microbiol.* doi:10.1128/JCM.01888-19.

Muloi, D., Ward, M. J., Pedersen, A. B., Fèvre, E. M., Woolhouse, M. E. J., and Van Bunnik, B. A. D. (2018). Are Food Animals Responsible for Transfer of Antimicrobial-Resistant Escherichia coli or Their Resistance Determinants to Human Populations? A Systematic Review. *Foodborne Pathog. Dis.* 15, 467–474. doi:10.1089/fpd.2017.2411.

Nahin, A. (2008). Create Date — New Field Indicates When Record Added to PubMed. *NLM Tech. Bull.* Available at: https://www.nlm.nih.gov/pubs/techbull/nd08/nd08_pm_new_date_field.html [Accessed November 24, 2016].

NCBI (2014). What is Whole Genome Shotgun (WGS)? *Whole Genome Shotgun Submissions*. Available at: https://www.ncbi.nlm.nih.gov/genbank/wgs/ [Accessed December 8, 2016].

NCBI (2020). BioProject Frequently Asked Question: What is Project Type? Available at: https://www.ncbi.nlm.nih.gov/bioproject/docs/faq/#what-is-project-type [Accessed February 10, 2020].

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017a). A curated dataset of complete Enterobacteriaeceae plasmids compiled from the NCBI nucleotide database. *Data Br.*

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017b). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* 91, 42–52. doi:10.1016/j.plasmid.2017.03.002.

Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., and Ellington, M. (2017c). Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front. Microbiol.* doi:doi: 10.3389/fmicb.2017.00182.

Osborn, A. M., da Silva Tatley, F. M., Steyn, L. M., Pickup, R. W., and Saunders, J. R. (2000). Mosaic plasmids and mosaic replicons: Evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. *Microbiology* 146, 2267–2275.

Page, A., Taylor, B., and Keane, J. (2016). Multilocus sequence typing by blast from de novo assemblies against PubMLST. *J. Open Source Softw.* 1, 118. doi:10.21105/joss.00118.

Pallecchi, L., Bartoloni, A., Fiorelli, C., Mantella, A., Di Maggio, T., Gamboa, H., et al. (2007). Rapid dissemination and diversity of CTX-M extended-spectrum β-lactamase genes in commensal Escherichia coli isolates from healthy children from low-resource settings in Latin America. *Antimicrob. Agents Chemother.* 51, 2720–2725. doi:10.1128/AAC.00026-07.

Pecora, N. D., Li, N., Allard, M., Li, C., Albano, E., Delaney, M., et al. (2015). Genomically informed surveillance for carbapenem-resistant enterobacteriaceae in a health care system. *MBio* 6, 1–11. doi:10.1128/mBio.01030-15.

Petersen, J., Brinkmann, H., Berger, M., Brinkhoff, T., Päuker, O., and Pradella, S. (2011). Origin and

evolution of a novel DnaA-like plasmid replication type in rhodobacterales. *Mol. Biol. Evol.* doi:10.1093/molbev/msq310.

Phan, M. D., Kidgell, C., Nair, S., Holt, K. E., Turner, A. K., Hinds, J., et al. (2009). Variation in Salmonella enterica serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever. *Antimicrob. Agents Chemother.* 53, 716–727. doi:10.1128/AAC.00645-08.

Praszkier, J., Wei, T., Siemering, K., and Pittard, J. (1991). Comparative analysis of the replication regions of IncB, IncK, and IncZ plasmids. *J. Bacteriol.* 173, 2393–2397.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35. doi:10.1093/nar/gkl842.

Ramsay, J. P., Kwong, S. M., Murphy, R. J. T., Yui Eto, K., Price, K. J., Nguyen, Q. T., et al. (2016). An updated view of plasmid conjugation and mobilization in *Staphylococcus*. *Mob. Genet. Elements* 6, e1208317. doi:10.1080/2159256X.2016.1208317.

Rawlings, D. E., and Tietze, E. (2001). Comparative biology of IncQ and IncQ-like plasmids. *Microbiol. Mol. Biol. Rev.* 65, 481–496, table of contents. doi:10.1128/MMBR.65.4.481-496.2001.

Riley, M., and Gordon, D. (1992). A survey of Col plasmids in natural isolates of Escherichia coli and an investigation into the stability of Col-plasmid lineages. *J. Gen. Microbiol.* 138, 1345–1352.

Robertson, J., and Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. genomics*. doi:10.1099/mgen.0.000206.

Shintani, M., Sanchez, Z. K., and Kimbara, K. (2015). Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.* 6, 1–16. doi:10.3389/fmicb.2015.00242.

Smillie, C., Garcillan-Barcia, M. P., Francia, M. V, Rocha, E. P., and de la Cruz, F. (2010). Mobility of plasmids. *Microbiol Mol Biol Rev* 74, 434–452. doi:10.1128/MMBR.00020-10.

Taylor, D., Gibreel, A., Lawley, T., and Tracz, D. (2004). “Chapter 23: Antibiotic resistance plasmids,” in *Plasmid Biology*, eds. B. Funnell and G. Phillips (Washington, DC: ASM Press), 473–485.

Valverde, A., Cantón, R., Garcillán-Barcia, M. P., Novais, Â., Galán, J. C., Alvarado, A., et al. (2009). Spread of blaCTX-M-14 is driven mainly by IncK plasmids disseminated among Escherichia coli phylogroups A, B1, and D in Spain. *Antimicrob. Agents Chemother.* 53, 5204–5212. doi:10.1128/AAC.01706-08.

Villa, L., García-Fernández, A., Fortini, D., and Carattoli, A. (2010). Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. *J. Antimicrob. Chemother.* 65, 2518–2529. doi:10.1093/jac/dkq347.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42. doi:10.1093/nar/gkt1099.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics*. doi:10.1099/mgen.0.000132.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi:10.1093/jac/dks261.

3 Developing bioinformatic pipelines for comparing plasmid genetic content across bacterial isolates using short-read sequencing data

3.1 Summary

It is challenging to accurately reconstruct individual plasmids from short-read sequencing data, so downstream comparative genomics of individual plasmids is often not possible. This constrains insights into plasmid epidemiology and evolution. However, useful insights may still be gained without reconstructing individual plasmids. In this chapter, I developed two novel automated bioinformatic pipelines designed to compare short-read sequenced samples in terms of their total plasmid genetic content ('plasmidomes'); because analysis takes place at the level of sample plasmidomes, individual plasmids need not be reconstructed. The first ('replicon-based') method clusters samples according to replicon type counts (for a crude assessment of plasmidome diversity). The second ('reference-based') method involves BLAST querying sample contigs against a plasmid reference database; filtering contigs with ≤30% query coverage (deemed likely chromosomal); and finally, clustering samples according to sample contig–reference plasmid distance scores. Preliminary benchmarking of the methods indicated that neither provides a completely reliable assessment of plasmidome diversity. Given simplistic filtering of contig alignments, a presumed limitation of the reference-based method is that many chromosomal contigs may still align to plasmid references, biasing inference of plasmidome diversity.

The research conducted in this chapter began in mid-2017 before the release of reference-based contig classification tools such as mlplasmids and PlaScope; these tools are able to classify contigs in short-read assemblies as plasmid/chromosomal with a relatively good degree of accuracy for well-studied taxa such as E. coli. If I was to conduct plasmid analysis with short-read sequencing data, I would opt to use a tool such as mlplasmids or PlaScope to extract plasmid content, rather than using methods applied in this chapter. Nevertheless, methods developed in this chapter were incorporated into other parts of the thesis. Notably, code for the reference-based pipeline was used for the ATCG tool described fully in Chapter 4, and used in a preliminary state for the benchmarking analysis performed in this

chapter. In addition, the use of rMLST loci as chromosomal markers was incorporated into the bacterialBercow tool, outlined in Chapter 2 (Appendix A: Methods 3).

Note that all sample collection and wet lab sequencing work was conducted by collaborators.

3.2 Introduction

When I began this chapter, several automated bioinformatic pipelines had been published aiming to reconstruct plasmid genomes from short-read WGS data; namely, plasmidSPAdes and Recycler, which use coverage depth and assembly graph information (Rosov et al., 2015; Antipov et al., 2016). The PLACNET tool uses plasmid/chromosomal reference sequence alignments with the aim of plasmid reconstruction, but it relies partly on manual decision-making (Lanza et al., 2014). Moreover, as mentioned in Chapter 1 (Section 1.3.2), all these algorithms often fail to accurately reconstruct plasmid structures. Consequently, downstream comparative genomics of individual plasmids resolved from short-read data may not be possible. A more attainable goal is to compare short-read sequenced samples in terms of their total plasmid genetic content ('plasmidomes'), without needing to reconstruct individual plasmids. In other words, one can aim to assess the between-sample diversity ('beta diversity') of sample plasmidomes (Brolund et al., 2013; Wagner et al., 2018). Researchers have commonly used detection of plasmid replicon loci, or alignment to plasmid references as a means to gain insight into sample plasmid content and compare plasmid content across samples, as described below. However, these methods have generally been applied on an ad hoc basis; automated and benchmarked bioinformatic pipelines for replicon- or reference-based comparative plasmidome analysis are lacking – a deficit I attempt to address in this chapter.

Plasmidome diversity has traditionally been assessed descriptively using replicon typing data (Bleicher et al., 2013; Brolund et al., 2013; Song et al., 2013; Dib et al., 2015). To accommodate larger sequencing datasets, automated statistical, rather than descriptive analysis, of replicon data might be beneficial. Microbial species beta diversity has traditionally been examined by applying metrics such as Bray-Curtis dissimilarity index to 16S amplicon data (Birtel et al., 2015; Sinclair et al., 2015).

Likewise, per-sample replicon count data could be analysed statistically using the Bray-Curtis index, although to my knowledge, rigorous statistical (as opposed to descriptive) analysis of replicon count data has not been widely used for comparative plasmidome analysis.

When this research was initiated, automated tools using reference alignment to analyse plasmid content within short-read assemblies were lacking (although see Addendum for an updated perspective). As explained in Chapter 1 (Section 1.3.1.2), reference-based mapping approaches for tracking plasmids during outbreaks have been widely used, but should be interpreted cautiously since they cannot reliably uncover all plasmid structures. Nevertheless, a reference-based approach may be appropriate for more limited goals, such as assessing overall plasmid genetic content (see below) or identifying a single candidate plasmid per sample, based on the best matching plasmid reference. Indeed, mapping short-read assembled contigs to plasmid references has previously been used in this way to compare plasmid content across samples, in the context of hospital outbreak analyses (Decraene et al., 2018; Stoesser et al., 2020). Given the continued expansion of publically-available bacterial sequence data, more plasmid sequence diversity is being uncovered, so a reference matching approach for plasmid analysis is attractive. Rather than simply identifying best plasmid reference matches, I wanted to gain a broader perspective on the overall diversity of plasmid genetic content across short-read sequenced samples (see below).

In this chapter, I developed the following two automated bioinformatic pipelines for comparative analysis of plasmid content using short-read sequencing data. 1) hierarchical clustering according to per-sample plasmid replicon type counts (for a crude assessment of plasmid diversity across samples). 2) a pipeline for clustering samples according sample-vs-reference plasmid distance score profiles, determined by BLASTN querying sample contigs against a reference plasmid database. I conducted a preliminary assessment of the reliability of the methods. Specifically, using paired short- and long-read data from 60 Gammaproteobacterial samples, I benchmarked comparative analysis results from

short-read methods against a comparative analysis based on resolved plasmids from hybrid short- and long-read assemblies.

3.3 Methods

3.3.1 Developing automated tools for comparative analysis of plasmid content using short-read data

Replicon-based clustering method

First, replicon typing was conducted using a locally downloaded version of the PlasmidFinder Enterobacteriaceae database (retrieved on 29th May 2018); contigs were BLASTN searched against the PlasmidFinder database using percentage identity and coverage thresholds of 80% and 60%, respectively (Carattoli et al., 2014). Counts of replicon types were aggregated on a per-sample basis. From raw counts, pairwise dissimilarities between samples were calculated using the Bray-Curtis metric implemented with the vegan package in R (Oksanen et al., 2017). From sample–sample dissimilarity scores, a hierarchical clustering dendrogram was built using complete linkage clustering, as implemented by the hclust function in R.

Reference-based clustering method

A reference database of Gammaproteobacterial plasmids was compiled using a broadly similar approach to that described in Orlek et al. (2017) (see Chapter 2, Section 2.3.1). Firstly, putative plasmids were retrieved from NCBI RefSeq using an Edirect Entrez query, and filtering was conducted using regular expression searches of accession titles (Supplementary Methods S1). In contrast to Orlek et al. (2017), a length filter (1kb–2Mb) was applied and only circular plasmid accessions were included (with the aim of stringently excluding chromosomal accessions). In common with the bacterialBercow curation pipeline (Appendix A: Methods 3), rMLST locus detection was used to filter chromosomal (and chromid) sequences. rMLST alleles were downloaded on 1st March 2018 from the PubMLST website (https://pubmlst.org/rmlst/). Putative complete plasmid accessions were then BLASTN

searched against rMLST alleles (parameters: max_hsps=1, perc_identity=95, evalue=1e-10), and hits were further filtered using a 95% rMLST allele coverage threshold. Accessions with hits to one or more rMLST alleles were excluded. To reduce database redundancy, remaining plasmid accessions were clustered using CD-HIT-EST (Li and Godzik, 2006), with the following parameters: 95% sequence identity threshold (-c 0.95); 95% alignment coverage of the longer sequence (-aL 0.95); word size 9 (-n 9); slow but more accurate clustering (-g 1). To compile the final reference plasmid database, one representative plasmid per CD-HIT-EST cluster was selected. Plasmids from bacterial samples included in the benchmarking analysis (Section 3.3.2) were excluded from the plasmid database; in this way, the reference-based approach could be assessed in the context of analysing novel samples (containing plasmid sequences not represented exactly in the database).

To cluster samples according to reference database matches, contigs from each sample were first BLASTN searched against the Gammaproteobacterial reference plasmid database (parameters: evalue=1e-8, word_size=38, culling_limit=5). The relatively long word size of 38 was selected since this has previously been proposed for calculation of genome-genome distances (Meier-Kolthoff et al., 2014). The culling limit parameter was used to reduce the number of overlapping hits. BLAST alignments associated with contigs with ≤30% query coverage (deemed likely to be chromosomal) were excluded. For each sample, remaining alignments between sample contigs and reference plasmids were parsed to calculate sample-reference plasmid distance scores (see below for distance score calculation method). Alignment parsing was implemented using the GenomicRanges R package (Lawrence et al., 2013); the alignment parsing procedure is similar to the approach described for the ATCG pipeline developed in Chapter 4 (see Section 4.3.1), albeit less complex (for example, as described below, a distance score was derived from subject alignment intervals only, as opposed to both query and subject alignment intervals). The per-sample alignment data comprised query intervals (on sample contigs) and subject intervals (on reference plasmids). Firstly, subject intervals were divided into a “disjoint decomposition” of non-overlapping intervals using the GenomicRanges disjoin function with the arguments “ignore.strand=TRUE” and “with.revmap=TRUE”. For each disjoint

interval, an optimal intersecting alignment was selected (favouring longer alignments). Finally, contiguous disjoint intervals represented by the same optimal alignment were merged, and sample-reference plasmid statistics were calculated from merged subject (reference plasmid) alignment intervals. Specifically, a distance score was calculated from the quotient of total identical nucleotide positions across reference alignment intervals divided by reference genome length. Consequently, the distance scores take account of both percent identity and reference plasmid coverage breadth. Following calculation of distance scores, a matrix of sample-reference distance scores was generated. A sample–sample distance matrix was computed using Euclidean distance, and a hierarchical clustering dendrogram was built using complete linkage clustering.

3.3.2 Benchmarking short-read plasmid analysis methods

To assess the extent to which the replicon-based and reference-based dendrograms accurately reflect sample plasmidome diversity, a preliminary benchmarking study was conducted, as described below. The benchmarking dataset comprised bacterial samples with paired short-read (Illumina) and long-read (Nanopore/PacBio) sequencing data, from three separate sequencing projects. Firstly, the Manchester *bla*$_{KPC}$ Enterobacterales hospital outbreak study, described in Chapter 4 (see Appendix C: Table 3). Briefly, this dataset includes samples from a clonal *E. coli* ST216 strain, as well as other *bla*$_{KPC}$-encoding Enterobacterales species. Secondly, the University of Virginia *bla*$_{KPC}$ Enterobacterales hospital outbreak study, described in Sheppard et al. (2016) (NCBI BioProject 246471). Finally, the REHAB project (in-house sequencing data representing Enterobacterales samples from river/wastewater sites; for project details see http://modmedmicro.nsms.ox.ac.uk/rehab/). With the exception of the Sheppard et al. dataset, sequencing data was 'in-house' i.e. not uploaded to NCBI at the time. Short-read Illumina assemblies were generated using SPAdes (following methods described above). Long-read assemblies were either downloaded from NCBI (in the case of the University of Virginia samples); assembled using the long-read only HGAP pipeline (Chin et al., 2013) (24 PacBio-

sequenced Manchester outbreak samples, provided by Dr Hang Phan); or hybrid short-/long-read assembled using Unicycler (v0.4.1) (Wick et al., 2017) (17 Nanopore-sequenced Manchester outbreak samples, and all REHAB samples). Unicycler was initially run in "standard" mode, but samples that were not completely circularised in standard mode, were re-assembled using "bold" mode. Finally, contigs from long-read/hybrid-assembled samples were classified as plasmid, chromosomal, or unknown based on replicon and rMLST locus detection (following replicon/rMLST annotation methods described above). Contig classification methods represented a preliminary version of the bacterialBercow contig classification pipeline (see Appendix A: Methods 3). Briefly, contigs with ≥1 rMLST loci were annotated as chromosomal; contigs with a replicon and no rMLST loci were annotated as plasmids; circular contigs with no marker loci which belonged to a sample with a single circular chromosome were annotated as plasmids; all other contigs were annotated as unknown. For samples with a single chromosomal contig and no unknown contigs, plasmid contigs were extracted and used for benchmarking. This left 60 Enterobacterales samples with paired short-read assemblies and long-read assembled plasmid contigs, which were included in the benchmarking dataset (see Results for details).

Benchmarking was conducted as follows: the replicon- and reference-based plasmid analysis pipelines were applied to short-read assemblies to generate two corresponding dendrograms. These dendrograms were benchmarked against a dendrogram generated from all-vs-all comparison of per-sample plasmid contigs. Specifically, long-read assembled plasmid contigs were compared on a per-sample basis using a preliminary version of the ATCG comparative genomics pipeline, which is described in Chapter 4, and shares similarities with the genome-genome distance calculator (GGDC) (Meier-Kolthoff et al., 2013). Briefly, all-vs-all BLAST was conducted between sample contigs (BLASTN, evalue=1e-8, word_size=38); alignments were parsed to produce an optimal set of non-overlapping alignment intervals from which sample–sample distance scores were calculated. Specifically, 'distance formula d8' (as in Meier-Kolthoff et al. (2013)) was applied; this is a 'resemblance' distance metric

(*sensu* Broder, 1997) calculated based on the quotient of total identical base pairs across alignments divided by combined genome size (i.e. total length of both plasmid contig sets).

For benchmarking, dendrogram concordance was visualised and quantified using tanglegrams and the Baker's gamma statistic. Specifically, tanglegrams were produced using the dendextend package in R (Galili, 2015). To improve visualisation of non-concordant relationships, the tanglegram was 'untangled' using the "step2side" algorithm, which rotates dendrogram branches around internal nodes to achieve a more optimal layout, without changing dendrogram topology. In addition, in the case of the replicon dendrogram, where terminal branches subtended multiple leaves of the same replicon type, corresponding sample labels were shuffled to achieve further untangling. Topological concordance between the two dendrograms was assessed statistically using the Baker's gamma statistic (Baker, 1974). Its value can range from -1 (highly non-concordant) to 1 (highly concordant), with values around 0 indicating no relationship between dendrogram topologies. The index is not affected by the height of a branch, only its relative position compared with other branches. To calculate a p-value for the Baker's gamma statistic, a permutation test was used, which involved randomly shuffling labels of the reference-based dendrogram 200 times.

3.4 Results and discussion

3.4.1 Description of the curated plasmid dataset

The reference-based analysis depended on a curated plasmid database. Initially, 4913 circular Gammaproteobacterial putative plasmid accessions were retrieved from NCBI. 569 accessions were excluded after applying regular expression match inclusion/exclusion criteria, and a further 12 accessions were excluded as chromosomal based on rMLST locus detection (leaving 4332 accessions). Finally, after clustering plasmid sequences and selecting one representative plasmid per cluster, 3730 curated reference plasmids remained.

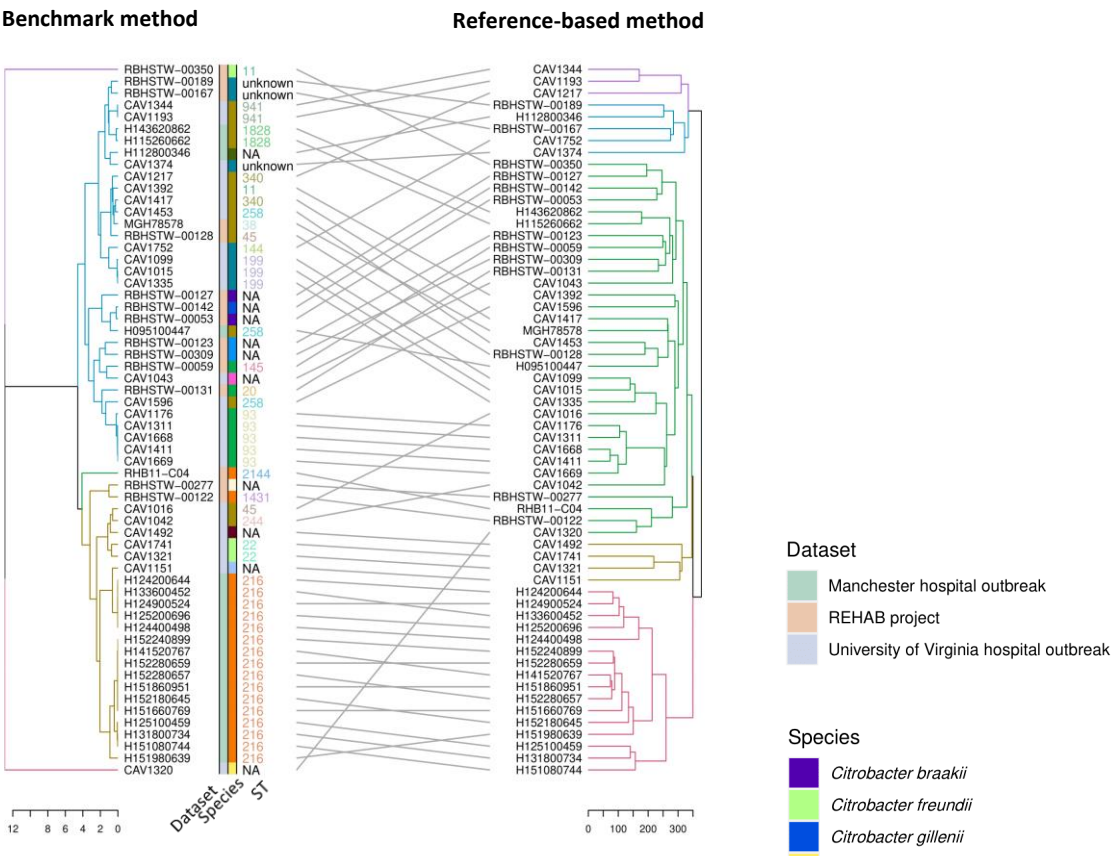
3.4.2 Benchmarking of the short-read plasmid analysis pipelines

For the benchmarking analysis, 87 bacterial samples were initially assembled; of these assemblies, 60 comprised a complete set of characterised contigs (i.e. plasmid/chromosomal contigs as opposed to "unknown" contigs) and were therefore included in the benchmarking analysis. The 60 samples represented 14 species including *E. coli* (18 samples, 3 STs), *K. pneumoniae* (14 samples, 8 STs), *K. oxytoca* (7 samples, 2 characterised STs and several unknown STs), and *Enterobacter cloacae* (7 samples, 3 STs) (Appendix B: Table 1). Figure 1 shows the relationship between the benchmarking dendrogram (based on BLAST comparison of long-read resolved sample plasmids) and the dendrograms generated using the short-read analysis pipelines (reference-based and replicon-based dendrograms shown in Figure 1A and Figure 1B, respectively). Both tanglegrams demonstrate numerous cases of non-concordance, indicating that the short-read plasmid analysis methods fail to provide a completely accurate assessment of sample plasmidome diversity. Nevertheless, concordance, assessed using Baker's gamma index, was statistically significant for both Figure 1A and 1B tanglegrams (Baker's gamma indices 0.31 and 0.17 respectively; both $p<0.001$). According to the Baker's gamma indices, the replicon-based dendrogram shows lower concordance (0.17) against the benchmark dendrogram. However, this result should be interpreted cautiously; firstly, because Baker's gamma index confidence intervals were not calculated, and secondly, because the benchmarking was only a preliminary analysis conducted on a relatively small sample set.

The sample set of Enterobacterales species should favour both methods relative to a non-Enterobacterales sample set. Enterobacterales is a relatively well-studied, clinically important taxon; hence, Enterobacterales plasmids are likely to be relatively well-represented in the reference database and – as demonstrated in Chapter 2 (Section 2.4.3.2) – Enterobacterales replicons are well-represented in PlasmidFinder. The corollary is that the pipelines would likely perform even more poorly if applied to non-Enterobacterales plasmids. Finally, the reference-based method may perform

relatively well when analysing plasmids from a clonal strain such as the *E. coli* ST216 strain from the Manchester dataset. As I explain, Figure 1 is consistent with this conjecture. Firstly, in the Figure 1A benchmark dendrogram, the *E. coli* ST216 samples form a cluster, which is matched by a concordant cluster in the reference-based dendrogram, whereas in Figure 1B, *E. coli* ST216 samples do not form a pure cluster in the replicon-based dendrogram. Furthermore, in Figure 1A, it appears that non-clonal samples tend to be more 'tangled' compared with *E. coli* ST216 samples. For example, the non-clonal *K. pneumoniae* samples CAV1016 and CAV1042 (STs 45 and 244, respectively) are closely related in the benchmark dendrogram but nest in separate clusters of the reference-based dendrogram. Once again, the caveat to the described interpretation is the small sample size of the benchmarking analysis.

A


Benchmark method
Reference-based method
Dataset
Species
ST
12 8 6 4 2 0
0 100 200 300
Dataset
Manchester hospital outbreak
REHAB project
University of Virginia hospital outbreak
Species
Citrobacter braakii
Citrobacter freundii
Citrobacter gillenii
Enterobacter aerogenes
Enterobacter asburiae
Enterobacter cloacae
Enterobacter kobei
Escherichia coli
Klebsiella oxytoca
Klebsiella pneumoniae
Kluyvera intermedia
Raoultella ornithinolytica
Serratia marcescens
Shigella dysenteriae

B

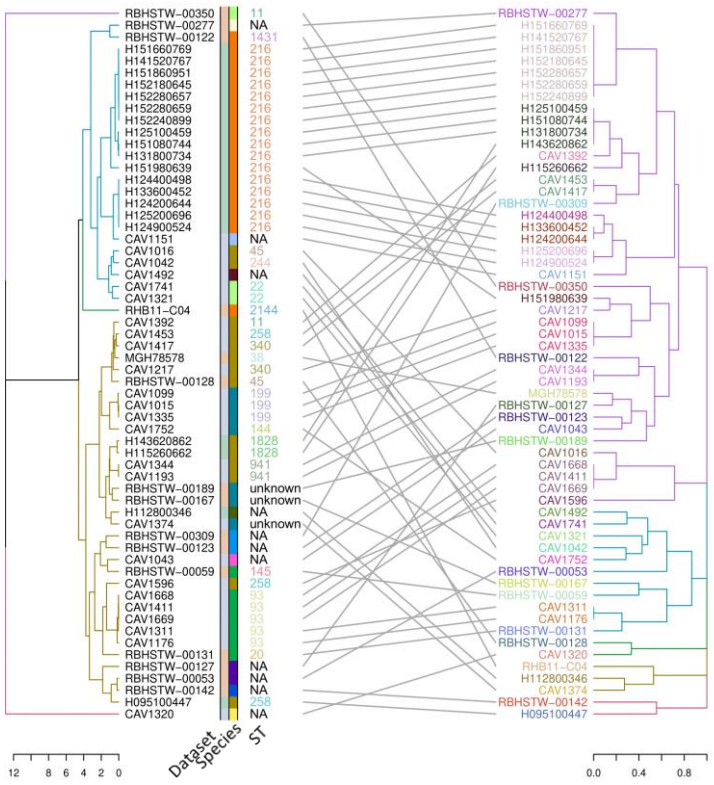

Benchmark method
Replicon-based method
Dataset
Species
ST
12 8 6 4 2 0
0.0 0.4 0.8

Figure 1. Analysis of 60 Enterobacterales samples using the short-read plasmid analysis pipelines (right dendrograms) were benchmarked against a BLAST-based comparison of long-read resolved plasmid structures (left dendrogram). (A) Shows the reference-based pipeline dendrogram (right) alongside the benchmark dendrogram (left); (B) shows the replicon-based pipeline dendrogram (right) alongside the benchmark dendrogram (left). Sample labels on the replicon-based dendrogram are coloured according to replicon type combination. Dendrogram branches are coloured arbitrarily to distinguish different clusters. An intersecting line connects a given sample in one dendrogram with the same sample in the other dendrogram; together, the lines allow concordance between dendrogram topologies to be visualised. The dataset, species, and multi-locus sequence type (ST) associated with each sample are indicated using coloured bars and text alongside the benchmark dendrogram.

3.4.3 A reflection on short-read plasmid analysis in light of the benchmarking results

It is unsurprising that the replicon-based dendrogram in Figure 1B fails to accurately describe sample plasmidome diversity. Single-locus approaches provide a crude assessment of diversity relative to whole genome analyses. The diverse and fluid nature of plasmid genomes (see Chapter 2) means that replicon-based methods are likely to give an especially unreliable picture of plasmid genetic content within samples. The reference-based dendrogram in Figure 1A arguably shows more promise (an improved Baker's gamma correlation score relative to Figure 1B) but the benchmarking is consistent with *a priori* concerns that divergent strain genetic backgrounds bias results. An obvious next step in developing the reference-based pipeline would be inclusion of chromosomal references, allowing contig classification according to best match against chromosomal/plasmid reference (see Addendum). To fully understand the shortcomings of the plasmid analysis pipelines, a more detailed benchmarking analysis would be required. In the case of the reference-based pipeline for example, this might involve investigating false positive plasmid matches (i.e. chromosomal contigs matching a plasmid reference) and false negative plasmid matches (i.e. plasmid contigs not matching any reference plasmid, indicative of a novel plasmid sequence/gaps in the reference database).

Both pipelines analyse samples at the level of the plasmidome. The advantage of this approach is that individual plasmid structures need not be reconstructed. Instead, plasmid content just needs to be distinguished from chromosomal content; this was the aim of using plasmid-specific markers (replicon-based approach) and plasmid reference matching (reference-based approach). However, it

is unclear to what extent the sample plasmidome is an informative unit of analysis for understanding plasmid epidemiology. Bacteria commonly harbour multiple plasmids, which may transfer independently between strains (Dionisio et al., 2019); when analysing sample plasmidomes, inter-sample plasmid transmission links may be obscured in the context of two samples harbouring otherwise disparate plasmid populations. More generally, in the context of comparative plasmid analysis, interpretation of plasmidome relationships is challenging: for example, does a similar plasmdome pair represent shared plasmid populations; a plasmid transmission link; or, homologous regions of plasmid sequence, perhaps transferred by homologous recombination? Comparative genomics of completely resolved plasmid sequences is required to gain a fuller understanding of plasmid epidemiology.

3.5 Conclusions

Reconstructing plasmid sequences from bacterial short-read WGS data is challenging, which in turn hinders plasmid analysis. In recent years, various algorithms for plasmid reconstruction have emerged (see Chapter 1, and Addendum below). However, in this chapter, I aimed to bypass plasmid reconstruction and instead conduct comparative analysis of sample plasmidomes, using reference-based and replicon-based bioinformatic pipelines. Unfortunately, benchmarking indicated that neither method (alone) provides a reliable assessment of plasmid diversity. Further analysis would be required to fully understand the limitations of each method, but given the development of competitor tools (see Section 3.6 below) I decided to curtail this line of research. Increased availability of long-read sequence data enables more detailed understanding of plasmid epidemiology and transmission. Indeed, in chapters 4 and 5, I decided to focus on method development and biological analysis of complete plasmid structures, rather than continuing to work with short-read data.

3.6 Addendum. An up-to-date perspective on short-read plasmid analysis

Following the preliminary benchmarking, I remained interested in further developing the reference-based pipeline. I considered the following approaches: 1) filtering chromosomal contigs using rMLST loci before aligning remaining contigs to plasmid references; and/or 2) aligning to both chromosomal and plasmid references and classifying contigs according to the best matching reference. However, these plans were overtaken by rapid developments in the field, which saw the release of relatively successful reference-based automated software tools for distinguishing plasmid and chromosomal contigs, such as mlplasmids and PlaScope (Arredondo-Alonso et al., 2018; Royer et al., 2018), which are described in Chapter 1 (Section 1.3.2). These tools use taxon-specific reference databases comprising both plasmid and chromosomal sequences to classify contigs as plasmid or chromosomal (or unassigned, in the case of PlaScope). The 'unassigned' classification alludes to fundamental difficulties of reference-based plasmid/chromosomal contig classification, presumably arising from the fact that sequence content can be exchanged between plasmids and host chromosomes. Rather than re-analyse the short-read data using one or more of the novel contig classification tools (which still fail to achieve perfect classification), I decided to focus on analysis of complete plasmids (Chapters 4 and 5) since, with more widespread adoption of long-read sequencing, this is increasingly the direction in which plasmid analysis is headed.

3.6.1 Assessment of chromosomal/plasmid marker-based methods for contig classification

This chapter was conducted before bacterialBercow (see Appendix A: Methods 3) was fully developed. Besides curating complete NCBI plasmids, bacterialBercow can potentially also be used to classify in-house bacterial assemblies, based on detection of rMLST and plasmid replicon loci (which are used as chromosomal/plasmid marker loci). Therefore, following development of bacterialBercow, I wanted to briefly investigate how well short-read derived contigs can be classified using bacterialBercow. To

my knowledge, while plasmid replicons are often used as plasmid marker loci, no current contig classification tools use rMLST loci as chromosomal markers. If sufficient chromosomal contigs encode rMLST loci, rMLST locus detection could be a promising way to help distinguish chromosomal from plasmid contigs.

The extent to which contigs can be classified as chromosomal/plasmids using rMLST/replicon loci was assessed using an in-house dataset of 125 *bla*$_{KPC}$ *E. coli* ST216 samples, previously Illumina sequenced by Decraene et al. (2018) (see therein for further details on the sample set). Adapter trimming and filtering of Illumina reads was conducted using BBduk from the BBTools package (https://jgi.doe.gov/data-and-tools/bbtools/) (v37.64), using previously described parameters (Street et al., 2017). Samples were assembled using SPAdes (v.3.10.0) (parameters: --careful --cov-cutoff auto) (Bankevich et al., 2012). Illumina reads from the 125 *E. coli* ST216 were re-assembled using Unicycler (v0.4.9b) since, unlike SPAdes, Unicycler provides contig circularity information, which could improve contig classification. The 125 samples comprised 28772 contigs in total, of which only 26 across 19 samples were circularised. Assembly N50 was assessed using QUAST (Gurevich et al., 2013). The median assembly N50 was 100.15 kb (inter-quartile range [IQR] 98.27–101.09 kb). This is relatively low compared with *E. coli* assembly metadata contained in the EnteroBase database (Zhou et al., 2020), suggesting that my *E. coli* assemblies are relatively less contiguous on average. Specifically, among 104007 *E. coli* paired-end Illumina assemblies reported in EnteroBase (retrieved 13th July 2020; filtered using inclusion and exclusion criteria described in Supplementary Methods S2), the median assembly N50 was 135.57 kb (IQR 100.90–184.50 kb).

I proceeded to run bacterialBercow (v0.1.0, https://github.com/AlexOrlek/bacterialBercow) with the aim of characterising assembly contigs. The vast majority of contigs were classified as "unknown" (Figure 2; Appendix B: Table 2), demonstrating that, at least in this case, rMLST/plasmid replicon detection is – on its own – not a particularly useful strategy for distinguishing chromosomal from

plasmid contigs. In other situations, where assemblies are more contiguous, the approach might be more fruitful.

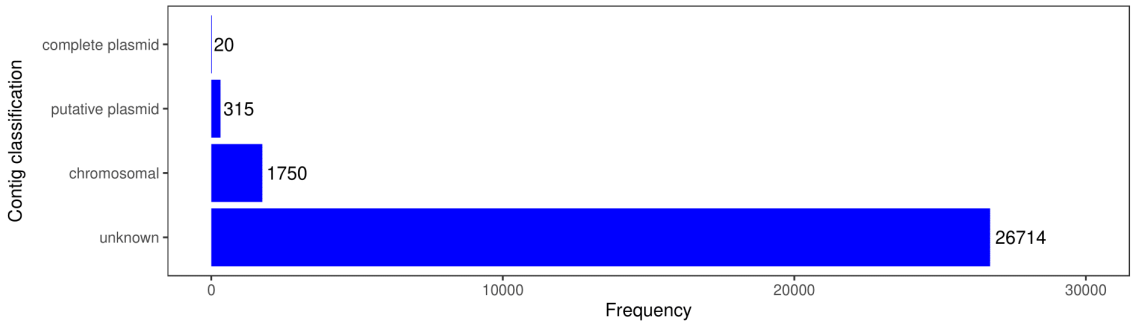


Figure 2. Frequency of contig classifications after applying bacterialBercow to 125 short-read sequenced *E. coli* samples assembled using Unicycler (comprising 28772 contigs in total).

3.7 Supplementary material

3.7.1 Supplementary Methods S1

Entrez query used for initial retrieval of putative complete plasmid accessions from NCBI:

```
esearch -db nuccore -query "g-proteobacteria[porgn] AND biomol_genomic[PROP] AND 1000:2000000[SLEN] AND plasmid[filter]" | efilter -source refseq | efetch -format docsum | xtract -pattern DocumentSummary -element AccessionVersion Topology Slen Title | awk '$2 ~ /circular/ { print $0 }'
```

Filtering plasmid accessions using regular expression searches of accession title text

The following regular expression term was applied to accession titles as an exclusion criterion to exclude non-plasmid, non-complete plasmid, and synthetic/laboratory vector plasmid accessions:

```
'contig|\sgene(?!tic|ral|rat|ric)|integron|transposon|scaffold|insertion sequence|insertion element|phage|operon|partial sequence|partial plasmid|region|fragment|locus|complete (?!sequence|genome|plasmid|\.|,)|(?<!complete sequence, )whole genome shotgun|artificial|synthetic|vector'
```

The following regular expression term was applied to accession titles as an inclusion criterion to select for complete accessions:

```
'complete(?= sequence| genome| plasmid)'
```

Accessions matching the regex exclusion criterion or not matching the regex inclusion criterion were discarded. Remaining accessions were further filtered according to rMLST locus detection, as described in the main text.

3.7.2 Supplementary Methods S2

Addendum methods

Retrieving and filtering EnteroBase accessions to obtain *E. coli* Illumina paired-end assembly metadata

Assembly metadata from the *Escherichia/Shigella* database was downloaded from EnteroBase (https://enterobase.warwick.ac.uk/), and *E. coli* accessions were selected. Then, with the aim of ensuring inclusion of Illumina paired-end assemblies only, the following inclusion and exclusion criteria (implemented in R using the grepl command) were applied to the "Data Source" field of the tabular metadata file:

```
ecolienterobaseassemblies[grepl('illumina',ecolienterobaseassemblies[,3],ignore.case = TRUE) &
grepl('paired',ecolienterobaseassemblies[,3],ignore.case = TRUE),]

ecolienterobaseassemblies[!(grepl('pacbio',ecolienterobaseassemblies[,3],ignore.case = TRUE) |
grepl('nanopore',ecolienterobaseassemblies[,3],ignore.case = TRUE)),]
```

Bibliography

Acman, M., van Dorp, L., Santini, J. M., and Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* doi:10.1038/s41467-020-16282-w.

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., and Pevzner, P. A. (2016). Genome analysis plasmidSPAdes : assembling plasmids from whole genome sequencing data. *Bioinformatics* 33, 3380–3387.

Arredondo-Alonso, S., Rogers, M. R. C., Braat, J. C., Verschuuren, T. D., Top, J., Corander, J., et al. (2018). mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. genomics*. doi:10.1099/mgen.0.000224.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques case I: Sensitivity to data errors. *J. Am. Stat. Assoc.* doi:10.1080/01621459.1974.10482971.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* doi:10.1089/cmb.2012.0021.

Birtel, J., Walser, J. C., Pichon, S., Bürgmann, H., and Matthews, B. (2015). Estimating bacterial diversity for ecological studies: Methods, metrics, and assumptions. *PLoS One*. doi:10.1371/journal.pone.0125356.

Bleicher, A., Schöfl, G., Rodicio, M. del R., and Saluz, H. P. (2013). The plasmidome of a Salmonella enterica serovar Derby isolated from pork meat. *Plasmid*. doi:10.1016/j.plasmid.2013.01.001.

Broder, A. Z. (1997). On the resemblance and containment of documents. *Proc. Int. Conf. Compression Complex. Seq.*, 21–29. doi:10.1109/sequen.1997.666900.

Brolund, A., Franzén, O., Melefors, Ö., Tegmark-Wisell, K., and Sandegren, L. (2013). Plasmidome-Analysis of ESBL-Producing Escherichia coli Using Conventional Typing and High-Throughput Sequencing. *PLoS One* 8. doi:10.1371/journal.pone.0065793.

Carattoli, A., Zankari, E., Garcia-Fernández, A., Larsen, M. V., Lund, O., Villa, L., et al. (2014). In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58, 3895–3903. doi:10.1128/AAC.02412-14.

Chin, C.-S. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi:10.1038/nmeth.2474.

Decraene, V., Phan, H. T. T., George, R., Wyllie, D. H., Akinremi, O., Aiken, Z., et al. (2018). A large, refractory nosocomial outbreak of klebsiella pneumoniae carbapenemase-producing Escherichia coli demonstrates carbapenemase gene outbreaks involving sink sites require novel approaches to infection control. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01689-18.

Dib, J. R., Wagenknecht, M., Farías, M. E., and Meinhardt, F. (2015). Strategies and approaches in plasmidome studies-uncovering plasmid diversity disregarding of linear elements? *Front. Microbiol.* 6, 1–12. doi:10.3389/fmicb.2015.00463.

Dionisio, F., Zilhão, R., and Gama, J. A. (2019). Interactions between plasmids and other mobile genetic elements affect their transmission and persistence. *Plasmid*. doi:10.1016/j.plasmid.2019.01.003.

Galili, T. (2015). dendextend: An R package for visualizing, adjusting and comparing trees of

hierarchical clustering. *Bioinformatics*. doi:10.1093/bioinformatics/btv428.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*. doi:10.1093/bioinformatics/btt086.

Lanza, V. F., de Toro, M., Garcillán-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., et al. (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. doi:10.1371/journal.pgen.1004766.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1003118.

Li, W., and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi:10.1093/bioinformatics/btl158.

Ludden, C., Raven, K. E., Jamrozy, D., Gouliouris, T., Blane, B., Coll, F., et al. (2019). One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *MBio*. doi:10.1128/mBio.02693-18.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-60.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., and Göker, M. (2014). Highly parallelized inference of large genome-based phylogenies. in *Concurrency Computation Practice and Experience* doi:10.1002/cpe.3112.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2017). Community Ecology Package "vegan". Version 2.4-3. *R Packag.* doi:10.4135/9781412971874.n145.

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* 91, 42–52. doi:10.1016/j.plasmid.2017.03.002.

Phan, M. D., Kidgell, C., Nair, S., Holt, K. E., Turner, A. K., Hinds, J., et al. (2009). Variation in Salmonella enterica serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever. *Antimicrob. Agents Chemother.* 53, 716–727. doi:10.1128/AAC.00645-08.

Rosov, R., Brown Kav, A., Bogumil, D., and Halperin, E. (2015). Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *bioRxiv*, 1–17.

Royer, G., Decousser, J. W., Branger, C., Dubois, M., Médigue, C., Denamur, E., et al. (2018). PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb. genomics*. doi:10.1099/mgen.0.000211.

Sheppard, A. E., Stoesser, N., Wilson, D. J., Sebra, R., Kasarskis, A., Anson, L. W., et al. (2016). Nested Russian Doll-like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.00464-16.

Sinclair, L., Osman, O. A., Bertilsson, S., and Eiler, A. (2015). Microbial community composition and diversity via 16S rRNA gene amplicons: Evaluating the illumina platform. *PLoS One*. doi:10.1371/journal.pone.0116955.

Song, X., Sun, J., Mikalsen, T., Roberts, A. P., and Sundsfjord, A. (2013). Characterisation of the

Plasmidome within Enterococcus faecalis Isolated from Marginal Periodontitis Patients in Norway. *PLoS One*. doi:10.1371/journal.pone.0062248.

Stoesser, N., Phan, H. T. T., Seale, A. C., Aiken, Z., Thomas, S., Smith, M., et al. (2020). Genomic epidemiology of complex, multispecies, plasmid-borne blaKPC carbapenemase in Enterobacterales in the United Kingdom from 2009 to 2014. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.02244-19.

Street, T. L., Sanderson, N. D., Atkins, B. L., Brent, A. J., Cole, K., Foster, D., et al. (2017). Molecular diagnosis of orthopedic-device-related infection directly from sonication fluid by metagenomic sequencing. *J. Clin. Microbiol.* doi:10.1128/JCM.00462-17.

Wagner, B. D., Grunwald, G. K., Zerbe, G. O., Mikulich-Gilbertson, S. K., Robertson, C. E., Zemanick, E. T., et al. (2018). On the use of diversity measures in longitudinal sequencing studies of microbial communities. *Front. Microbiol.* doi:10.3389/fmicb.2018.01037.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1005595.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* doi:10.1186/gb-2014-15-3-r46.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi:10.1093/jac/dks261.

Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y., and Achtman, M. (2020). The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* doi:10.1101/gr.251678.119.

4 ATCG: A novel alignment-based tool for comparative genomics

4.1 Summary

Whole-genome comparison metrics such as alignment-free mash distance and alignment-based average nucleotide identity (ANI) are commonly used for comparative genomic analysis of plasmid genomes, given the difficulties of plasmid phylogenetics. Although alignment-free methods are substantially faster, alignment-based comparative genomic methods remain popular, and potentially offer numerous advantages compared with current alignment-free methods. Unfortunately, existing alignment-based tools have numerous limitations. For example, existing BLAST-based methods fragment input genomes, eliminating the possibility of assessing structural similarity between genomes, or storing whole-genome alignments for downstream visualisation – this is a particular weakness in the context of analysing plasmid genomes, given that plasmids commonly undergo structural rearrangements. In this chapter, I develop a novel BLAST-based tool called ATCG which parses alignments in order to calculate ANI, genome-genome distances, and structural relatedness metrics, notably an alignment adjacency breakpoint distance. ATCG also optionally stores raw BLAST alignments and/or a set of best hit BLAST alignments, which can then be used for comparative genomic visualisation. Lastly, unlike all existing alignment-based tools, ATCG implements flexible input and comparison options (all-vs-all comparisons, query-vs-reference comparisons, pairwise comparisons specified by a list of genome pairs). Benchmarking of ATCG along with other ANI calculation tools (OrthoANIb, OrthoANIu, MUMmer) shows that with dc-megaBLAST as the aligner, ATCG computational runtime is favourable, and calculated ANI scores correlate well with OrthoANIb (an established high-sensitivity ANI calculation method). At the end of the chapter, I used ATCG to conduct all-vs-all comparative genomic analysis of 38 long-read sequenced samples from the Manchester *bla*$_{KPC}$ outbreak dataset. As well as showcasing ATCG as a useful addition to the comparative genomic

software ecosystem, the analysis also yielded interesting results. From ATCG output, I generated a network representing shared plasmid genetic content between samples. Among *E. coli* ST216 samples, there was evidence for persistence of plasmid genetic content across a period of several years, presumably supported by stable vertical inheritance with the clonal strain. However, there were also examples shared plasmid content between different species and strains represented in the network.

Note that for the Manchester outbreak samples, all sample collection and wet lab sequencing work was conducted by collaborators. Assembled PacBio sequenced samples were provided by Dr Hang Phan. All other work, including processing and assembly of Oxford Nanopore sequencing data, is my own, except were the text explicitly states that work was performed previously by Decraene et al. (e.g. bla_{KPC} resistance gene typing).

4.2 Introduction

As explained in Chapter 1, the plasticity and diversity of plasmid genomes means that collections of plasmid genomes often share few orthologous core genes. As a result, approaches which have been successfully applied for high-resolution strain-level analysis – such as core genome multi-locus sequence typing (cgMLST) (Jolley et al., 2018; Eyre et al., 2020) and core genome SNP-based phylogenetics – are generally not applicable to plasmids. Plasmid replicon and MOB loci may be used to infer low-resolution phylogenetic relationships among plasmids (Fricke et al., 2009; Zhang et al., 2019). However, the results of Chapter 2 indicate that even these 'backbone' loci are unreliable phylogenetic markers. Plasmid multi-locus sequence typing (pMLST) may provide higher resolution insight into plasmid relatedness, which could elucidate plasmid epidemiology. Yet, Chapter 2 demonstrates that shortcomings in terms of 'typeability' limit the usefulness of pMLST (and replicon typing in general), especially outside well-studied taxa such as Enterobacterales.

Ultimately, determining accurate evolutionary relationships of plasmids, using robust phylogenetic methods (i.e. involving an explicit model of molecular evolution) (Bromham, 2016), is generally not possible, especially when considering large collections of diverse plasmids, and/or long-term evolutionary dynamics. A more viable approach is to simply compare plasmid genomes in terms of

their overall similarity, such that patterns of plasmid sequence diversity can be uncovered, without aiming to achieve a full understanding of the underlying (vertical and/or reticulate) evolutionary processes (Morrison, 2014). Patterns of plasmid diversity can be intrinsically interesting. For example, Acman et al. found that the structure of plasmid diversity correlated with biological features such as plasmid replicon and MOB type, plasmid GC content, and host strain taxonomy (Acman et al., 2020). Plasmid sequence similarity can also be used to identify putative plasmid transmission events. For example, if highly similar plasmids are found among different strains, this may indicate a recent plasmid horizontal transfer event (Decraene et al., 2018; Ludden et al., 2019). In the context of a hospital outbreak study, accompanying epidemiological data, such as patient admission dates or dates spent in specific wards, can be used to help define the scope of an outbreak and gain insight into likely transmission chains (Weingarten et al., 2018). Given a collection of similar plasmids linked to an outbreak, it is reasonable to conclude that they are related through vertical descent and/or horizontal transfer; in this case, short-term evolutionary dynamics can be studied in more detail. Notably, visualising alignments of related plasmid sequences is a useful approach for inferring mutation events such as intra- and inter-genomic rearrangements (Conlan et al., 2016).

Assessing the overall similarity of plasmid genomes relies on comparative genomic methods (Setubal et al., 2018). Considering the main approaches applied to microbial genomes, some methods compare sequences in terms of their gene content or gene order ('gene-based analysis'), while others compare nucleotide sequences directly using 'whole-genome similarity' (or distance) metrics (Jain et al., 2018). Gene-based analysis begins with structural gene annotation of input genomes (i.e. identification of gene boundaries) (Lynch et al., 2016). A workflow for differential gene content analysis generally proceeds by identifying orthologs (Galperin et al., 2018; Setubal et al., 2018; Altenhoff et al., 2019). To represent plasmid diversity in terms of gene (ortholog) presence/absence, dendrograms or networks can be produced from pairwise similarity/distance metrics, calculated from binary presence/absence data (Snel et al., 2005; Dicks and Savva, 2007; Brilli et al., 2008; de Been et al., 2014; Knudsen et al., 2018; Suzuki et al., 2020). Comparative gene content analysis is a common component

of pangenome analysis, which seeks to delineate genes according to their conservation across genomes (e.g. conserved 'core' genes are distinguished from 'accessory' genes) (Page et al., 2015). An interesting approach is to restrict gene content analysis to the accessory genome, enabling insights into the functional differences which might have arisen through adaptive genome evolution (McNally et al., 2016; Uchiyama et al., 2016; Abudahab et al., 2019; Brockhurst et al., 2019).

Differential gene content results from gene gain and loss events; genome evolution also takes place through rearrangement events, such as inversions and transpositions, which alter gene order (and the relative orientation of genes, in the case of inversions). Hence, another way to obtain pairwise genome similarity metrics from gene annotations is to calculate a distance metric between two signed (stranded) permutations of a common set of genes (e.g. shared orthologs), with each permutation representing a gene order in a given genome (Dicks and Savva, 2007; Moret et al., 2013; Hartmann et al., 2018). There are two principal types of distance metric that can be derived from gene orders: edit distances and gene adjacency distances. Edit distances reflect the minimum number of rearrangement operations required to transform one gene order into another. Different types of edit distance permit different rearrangements; edit distances based on inversions (Hannenhalli and Pevzner, 1995) or double-cut-and-join (DCJ) operations (Yancopoulos et al., 2005) are widely used (Hilker et al., 2012; Moret et al., 2013). The DCJ is a general operation accounting for various rearrangements including inversions, transpositions, fissions, and fusions. In contrast to edit distances, gene adjacency distances are not based on rearrangement operations. Instead, they analyse the number of gene adjacencies shared between two gene orders. Hence, they are simple to calculate, and make no assumptions about the relative importance of different types of rearrangement event in driving gene order transformations (Bernt et al., 2013). A commonly used gene adjacency metric is the breakpoint distance. A breakpoint occurs when two genes found in adjacent order in one genome are not ordered adjacently and in the same relative orientation in another genome (Sankoff and Blanchette, 1998; Blanchette et al., 1999). A breakpoint distance metric is derived by normalising the breakpoint count using the total number of gene pairs, or the length of DNA assessed, as a denominator (Sankoff et al.,

2000; Neafsey et al., 2015). Various similar distance metrics based on gene adjacency have been devised (Keogh et al., 1998; Rocha, 2006), and an equivalent metric in the field of text string matching is called the q-gram distance (although this compares linear order only – relative orientation is not assessed given that text strings do not have strand orientation). To date, comparative analysis of plasmid structural similarity has tended to involve a descriptive assessment, drawn from visualisation of aligned plasmid genomes, rather than quantitative metrics such as gene order edit or breakpoint distances. However, as larger datasets of complete plasmid structures become increasingly available (see Chapter 2), gene order metrics may gain traction, since they are more scalable (Noll et al., 2018). As mentioned above, gene order analysis is generally conducted on shared (core) orthologs. Edit distances tend not to accommodate gene gain and loss events due to the complexity that this would entail (Moret et al., 2013); breakpoint distances could be calculated on distinct gene sets, but this would conflate gene order with gene content. Hence, gene order and gene content analyses are usually distinct endeavours.

Given the plasticity of plasmid genomes in terms of both gene exchange and structural rearrangement, comparing plasmids in terms of gene content and order is an attractive approach. However, there are several drawbacks relative to gene-independent whole-genome comparison metrics. Notably, gene-based analysis is limited by gene annotation inaccuracies, which can accrue at all stages of the gene annotation process. Gene structural annotation is imperfect, and can be hindered by sequencing errors, which can, for example, introduce spurious stop codons within coding sequences (Poptsova and Gogarten, 2010; Tripp et al., 2015). In addition, gene sequences may be split across contigs in a fragmented assembly, leading to false-negative gene calling (Gabrielaite and Marvig, 2019). Besides problems associated with annotation accuracy, an obvious limitation of gene-based analysis is that intergenic regions are excluded from comparative genome analyses. Bacterial genomes are relatively gene dense (Koonin, 2009). Nevertheless, on average, intergenic regions make up around 15% of a bacterial genome. In contrast to gene-based analyses, whole-genome comparison metrics circumvent

the complexities of gene annotation and also incorporate information contained within intergenic regions.

Methods for calculating whole-genome comparison metrics use either an alignment-based or an alignment-free approach. Alignment-based approaches use computationally-intensive dynamic programming in order to align sequences such that every position in the alignment is categorised as a match, mismatch or gap. Furthermore, the number of possible alignments grows rapidly with sequence length. Consequently, even when using heuristic sequence aligners such as BLAST and MUMmer (Altschul et al., 1997; Marçais et al., 2018) alignment-based methods can be orders of magnitude slower than alignment-free methods (Zielezinski et al., 2017; Jain et al., 2018). Rather than attempting to align sequences on a residue-by-residue basis, popular alignment-free tools for genome comparison assess similarity in terms of 'k-mer' occurrence. Genomes are first fragmented into short overlapping sub-sequences of length k, known as k-mers. Then, k-mer sets are converted to compressed representations called sketches, and pairwise similarity/distance scores are determined by comparing sketches (Ondov et al., 2016; Jain et al., 2018; Baker and Langmead, 2019). Thanks to their speed, alignment-free metrics, such as mash distance (introduced by Ondov et al.), have been useful for comparing large datasets of plasmids (Galata et al., 2019; Jesus et al., 2019; Acman et al., 2020). Alignment-free methods are under active development, with recent innovations including implementation of 'containment' distances e.g. mash screen is an asymmetric containment distance, measuring the extent to which one genome nests within another (Baker and Langmead, 2019; Ondov et al., 2019). In contrast, standard mash distance measures resemblance. Resemblance and containment are two overarching distance concepts (Broder, 1997). Resemblance is sensitive to genome size difference, so if a smaller genome matches exactly to part of a larger genome, the distance will be >0 according to resemblance, but according to containment, the distance score should in this case be 0.

Slower alignment-based methods remain an essential component of the comparative genomic toolkit because alignment-free approaches have several drawbacks. Firstly, alignment-free distances are intended only as approximate estimates of alignment-based distances such as average nucleotide identity (ANI) (discussed below). Moreover, unlike alignment-based methods, alignment-free methods conflate sequence identity with breadth of coverage (Ondov et al., 2019). This is particularly problematic for identifying homology among genomes which commonly undergo large-scale rearrangements, such as plasmids: tracts of high identity homologous sequence may be masked if overall coverage breadth is low. Therefore, in plasmid transmission studies, where accurate calculation of sequence identity is paramount, alignment-based methods are still used (Ludden et al., 2019). To handle large datasets, alignment-free distances can initially be used to reduce the size of the problem, leaving a subset of genomes of interest (e.g. genomes belonging to candidate transmission clusters) which can be aligned (Argimón and Aanensen, 2016). Another limitation of alignment-free methods is that most implementations do not assess structural similarity, since genomes are treated as sets of unordered k-mers; however, positional information can potentially be stored in order to assess structural similarity (Hosseini et al., 2019; Marçais et al., 2019).

Various tools are available for calculating alignment-based whole-genome comparison metrics, and these tools fall within two broad methodological categories. Firstly, there are various tools for calculating average nucleotide identities (ANI), by averaging percent identity scores across alignments. Secondly, there is the genome-genome distance calculator (GGDC) tool which first calculates various distance scores, from which digital (*in silico*) DNA-DNA hybridisation (dDDH) similarity scores are derived. dDDH scores are intended to be used in the context of genomic taxonomy and are calculated using statistical models which optimise for correlation with traditional wet-lab DNA-DNA hybridisation (DDH) values (Meier-Kolthoff et al., 2013; Hayashi Sant'Anna et al., 2019). Both ANI and dDDH metrics, used in combination with curated type strain reference genomes, have been widely used for genomic taxonomy of bacterial species (Richter and Rosselló-Móra, 2009; Ciufo et al., 2018; Meier-Kolthoff and Göker, 2019). In addition, large-scale comparative genomic analysis of bacterial strains and plasmids

can be achieved through clustering of ANIs or GGDC comparison metrics (Fernandez-Lopez et al., 2017; Wang et al., 2018; Weingarten et al., 2018). More detailed discussion of key methods is provided below.

The first method for ANI calculation was devised by Goris et al., and called ANIb (the b suffix reflecting use of BLASTN). To calculate ANI using the ANIb method, first, a query genome is fragmented *in silico* (~1kb long fragments) and BLASTN searched against a subject genome. A single best hit per query fragment is selected, and of these, hits showing >30% nucleotide identity and >70% fragment coverage breath are retained; from remaining hits, ANI is calculated as the mean percent identity (Goris et al., 2007). A more recent implementation, called OrthoANIb, fragments both query and subject genomes; conducts reciprocal BLASTN comparisons between query-subject fragments; and uses only bidirectional best hits (BBHs) to calculate ANI (Lee et al., 2016). Restricting analysis to BBHs means that ANI estimates should be based on one-to-one homology between (putatively) orthologous sequence regions (Wolf and Koonin, 2012). A corollary is that one-to-many alignments – involving repetitive or paralogous sequence content – which would inflate similarity estimates, are excluded. A key drawback of the described ANI implementations lies in the fragmentation step, which abolishes structural information (since information on genomic positions of fragments is not stored), precluding analysis of structural similarity. Although structural similarity was discussed above in the context of genes, other positional genetic markers, including alignment positions, can also be used to calculate structural similarity, as described below. One justification for the *in silico* fragmentation step is that it facilitates selection of BBHs, which would otherwise be difficult to determine from a complex set of overlapping genome alignments. Another justification is that *in silico* fragmentation mimics the wet-lab DDH procedure, which has traditionally been considered as a gold standard in bacterial taxonomy (Meier-Kolthoff et al., 2013). This procedure (now largely replaced by the *in silico* methods described here) involves fragmenting, mixing, and heating extracted DNA from two different species, then assessing similarity based on reassociation of heat-denatured DNA in heteroduplex vs homoduplex form (Rosselló-Mora and Amann, 2001).

Alternative methods do not involve *in silico* fragmentation. Notably, an alternative ANI metric called ANIm is calculated from whole-genome alignments generated by the NUCmer aligner from the MUMmer software package (Richter and Rosselló-Móra, 2009; Richter et al., 2016; Marçais et al., 2018). NUCmer finds maximal unique matches (MUMs) between genome pairs, and outputs non-overlapping putatively orthologous alignments (Dewey, 2019). As well as ANI, structural similarity statistics including the number of alignment adjacency breakpoints can also be obtained using the dnadiff script available in MUMmer. NUCmer is reported to achieve faster genome alignments than BLAST (Arahal, 2014), and the latest MUMmer implementation (MUMmer4), introduces parallel processing (Marçais et al., 2018). However, a remaining drawback of NUCmer is that, at least with default settings, it is less sensitive than BLAST alignment (Armstrong et al., 2019). In particular, by default, NUCmer uses relatively long exact matches to seed alignments, and therefore it performs less well when comparing more distantly related genomes, which share fewer exact subsequence matches (Marçais et al., 2018; Leimeister et al., 2019). Indeed, when ANIm is benchmarked against ANIb, correlation between the two metrics begins to deteriorate at ANIb values below 90%, and especially at ANIb values below 75% – presumably reflecting reduced sensitivity of the NUCmer aligner when comparing more distantly related genomes (Li et al., 2015; Yoon et al., 2017).

In common with ANIm, the GGDC method does not depend on *in silico* fragmentation, but unlike ANIm, GGDC calculates pairwise genome relatedness metrics (distance scores and dDDH values) from BLAST alignments rather than using a dedicated whole-genome aligner (like NUCmer). Use of BLAST potentially allows for higher sensitivity, as discussed above. However, genome-genome BLAST generally produces a rather chaotic set of overlapping alignments from which useful genome-genome statistics cannot be immediately derived (Meier-Kolthoff et al., 2013). Overlapping alignments reflect paralogous sequence and/or the BLAST algorithm per se – rather than a minimal set of optimal alignments, by default, BLAST produces all alignments fulfilling specified (E-value) thresholds. To generate distance scores from BLAST output, GGDC first parses alignments to generate a non-overlapping set of unidirectional best hit alignments, representing one-to-one homology. Then, from

the optimal alignments, distance scores are derived according to formulae shown in Table 1. Because BLAST results are asymmetric (i.e. results differ if query/subject sequence are switched), reciprocal BLAST is conducted, and distance scores are calculated by averaging calculations across the two alignment sets (one from each BLAST direction) (Auch et al., 2010a, 2010b). Different scores capture different distance concepts: scores d_0 and d_1 reflect alignment coverage breadth; score d_4 reflects percent identity; scores d_6 and d_7 capture both percent identity and coverage breadth, and reflect resemblance and containment distances, respectively.

Table 1. Principal distance scores associated with the GGDC method

Distance score	Formula	Description
d_0	$1 - \frac{L_{AB} + L_{BA}}{\lambda(A,B)}$	Distance metric reflecting overall coverage breadth
d_1	$1 - \frac{L_{AB} + L_{BA}}{\lambda_{min}(A,B)}$	Distance metric reflecting coverage breadth of the smaller genome
d_4	$1 - \frac{N_{AB} + N_{BA}}{L_{AB} + L_{BA}}$	Distance metric reflecting percent identity
d_6	$1 - \frac{N_{AB} + N_{BA}}{\lambda(A,B)}$	Distance metric reflecting total number of identical base pairs across alignments relative to combined genome size (resemblance)
d_7	$1 - \frac{N_{AB} + N_{BA}}{\lambda_{min}(A,B)}$	Distance metric reflecting total number of identical base pairs across alignments relative to twice the length of the smaller genome (containment)

Adapted from Table S3 in Meier-Kolthoff et al. (2013); distance scores d_2, d_3, d_5, d_8 and d_9 are not shown since they are simply rescaled variants. $L_{AB} + L_{BA}$ is the combined total length of alignments from pairwise BLAST between two genomes (named here genome A and genome B) run in both directions (i.e. genome A queried against genome B and vice-versa). $\lambda(A,B)$ is the combined length of the two genomes. $\lambda_{min}(A,B)$ is twice the length of the smaller genome. $N_{AB} + N_{BA}$ is the combined total number of identical base pairs across alignments from both BLAST directions.

One of the major limitations of the GGDC method is that it is only available as a web tool (at https://ggdc.dsmz.de/ggdc.php) which, at the time of writing (May 2020), can handle a maximum of 75 input genomes per run. In contrast, a range of local command-line and web tools are available for calculating the ANI metrics (Blom et al., 2016; Lee et al., 2016; Pritchard et al., 2016; Richter et al., 2016). The lack of a scalable tool for calculating genome distances/dDDH values impedes large-scale

analyses as well as benchmarking efforts by other research groups. Furthermore, because source code is not available, methodological details cannot be scrutinised. The GGDC developers state on the GGDC website that large-scale analyses can be run "as part of a scientific collaboration on our in-house compute cluster". However, other researchers may not always want to engage in collaboration, for example, due to legal concerns around sharing confidential data (UKRI, 2018). A further drawback is that the GGDC tool provides only a subset of the distance scores in Table 1 (d_0, d_4, d_6); distances d_1 and d_7 are omitted. In addition, GGDC does not implement a structural similarity/distance metric, although the possibility of calculating breakpoint distance from alignments has been mentioned in the literature (Auch et al., 2006). Finally, GGDC does not provide the option to store alignments, precluding downstream analyses, such as comparative genomic visualisation using the pairwise genome alignments.

ANI and GGDC metrics are useful for comparing relatively large numbers of genomes. If all-vs-all comparisons are conducted among genomes, similarity or distance matrices can be created, and relationships can then be visualised using dendrograms or networks. However, to investigate a selection of pairwise comparisons in more detail requires tools for comparative genomic visualisation. In particular, displaying sequence comparisons in a linear fashion helps to convey structural rearrangements. Popular visualisation tools for displaying linear comparisons include Artemis Comparison Tool (ACT) (Carver et al., 2005) and Easyfig (Sullivan et al., 2011), which can be used with GUI or command-line interfaces. Command-line only tools, namely, namely genoplotR and GenomeDiagram (Pritchard et al., 2006; Guy et al., 2011), allow more customisable plots to be produced, including visualising multiple segments in a single genome track, in order to represent comparisons between contig assemblies (not possible using ACT and Easyfig). A drawback of genoplotR and GenomeDiagram is that they require scripting expertise in R and Python, respectively.

In the context of comparative genomic visualisation, a key advantage of MUMmer is that whole-genome alignments are stored, allowing for both quantification and visualisation of pairwise genome

similarity within a unified analysis framework. MUMmer alignments can be visualised using the integrated MUMmer dotplot function or using an external tool such as ACT (Pevsner, 2015). As mentioned above, GGDC is the only BLAST-based ANI tool which runs BLAST using unfragmented genomes, meaning useful alignment data could potentially be generated for downstream comparative visualisation; however, alignments are not stored. BLAST is nevertheless commonly used for comparative genomic visualisation, in conjunction with tools such as ACT (Edwards and Holt, 2013). Since BLAST is a local aligner, linear visualisations of BLAST-based genome alignments generally appear chaotic due to overlapping alignments (Alikhan et al., 2011); it is therefore common practice to apply an arbitrary threshold to exclude low-scoring alignments (Pritchard, 2014).

In this chapter, I describe ATCG (Alignment-based Tool for Comparative Genomics), a novel tool for calculating ANI and distance metrics from pairwise BLASTN alignments; a wrapper function for genoplotR visualisation is also under development. ATCG addresses limitations of existing alignment-based tools for obtaining whole-genome comparison metrics. For example, unlike other BLAST-based methods, ATCG does not fragment input genomes, enabling structural similarity to be assessed. ATCG can also output best hit BLAST alignments, enabling downstream comparative genomic visualisation. Having developed ATCG, I benchmark it against other tools for calculating ANI: OrthoANIb, OrthoANIu, and MUMmer (specifically, a pipeline using NUCmer and the dnadiff script). Finally, I use ATCG to analyse available long-read sequenced samples from an outbreak of *bla*$_{KPC}$-encoding Enterobacterales.

4.3 Methods

4.3.1 ATCG bioinformatic workflow

ATCG is a command-line tool for BLAST-based comparisons of nucleotide sequence assemblies. It is written primarily in Python and R, and freely available at https://github.com/AlexOrlek/ATCG. To obtain ANI and distance metrics, users run the compare.py executable script (the workflow is

described below). A visualise.py executable script runs a comparative genomic visualisation workflow, which is still under active development, and hence, not described in detail. Briefly, the visualise.py workflow uses genoplotR software (Guy et al., 2011) to visualise comparisons. As input, it takes BLAST tabular output, and an optional annotation file. The visualisation workflow aims to combine the flexibility of genoplotR – notably in terms of allowing contig assembly comparisons to be visualised – with a user-friendly command-line interface. The rest of the chapter focuses on the compare.py component of ATCG. The workflow underlying the compare.py script is outlined below.

Input and specification of comparisons

As input, ATCG takes nucleotide sequence assemblies in FASTA format; optionally, FASTA files can be gzip compressed. There are three input options, which determine the specification and implementation of pairwise comparisons between input genomes, as summarised in Figure 1. The -s flag option implements all-vs-all comparisons between a set of genomes. The flag can be provided with a directory path containing FASTA files, which are recognised based on filename suffix (the suffix must be ".fa", ".fna", or ".fasta"; compressed FASTAs should be additionally suffixed with ".gz"). Each FASTA file within a directory should correspond to a genome. An incomplete genome assembly or a complete genome assembly comprising multiple sequences (chromosome, plasmids) is provided as a single multi-FASTA file. As an alternative to providing a FASTA directory path, the -s flag can be provided with a single multi-FASTA file, either as a file path or as input from the stdin stream. In this case, genomes must be delimited by the format of the FASTA header line identifiers. Specifically, vertical bars in the header identifier in the format “genome|contig” indicate the affiliation of sequences with genomes.

Another common task is to query genomes against one or more reference genomes, and this type of comparison is implemented with query and subject input flags -1 and -2 respectively. Sets of genomes provided to the two flags should be non-overlapping i.e. no genome should appear as both query and subject. Similar to when using the -s flag, the -1/-2 flags can be provided with input from two

directories containing FASTA files; from two FASTA files; or from two stdin streams. In the case of multi-FASTA file or stdin stream input, genomes must be delimited according to the FASTA header formatting described for the -s flag.

Finally, one might want to specify comparisons between a list of genome pairs of interest. For example, having identified similar genome pairs using an alignment-free tool such as mash, these pairwise relationships can be investigated in more detail using ATCG. In this case, the -p flag must be provided with a two-column tabular file containing FASTA file paths; each row of the file specifies a pair of genomes to be compared.

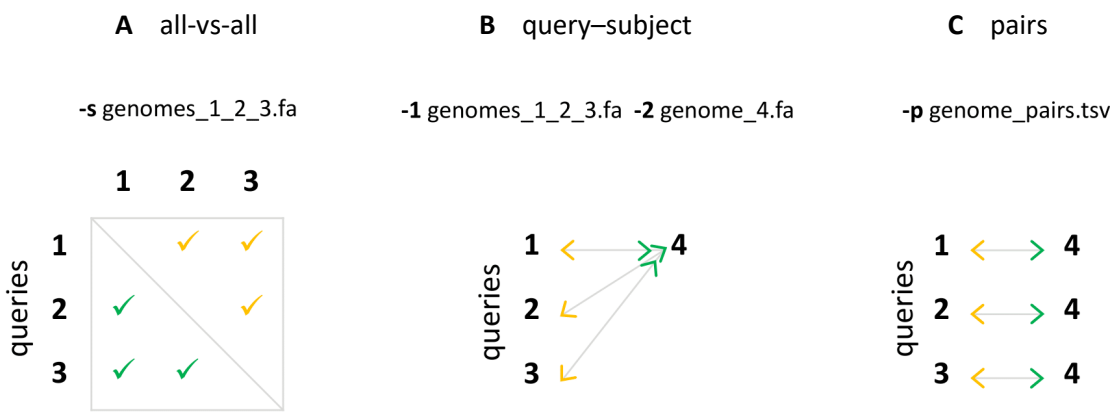


Figure 1. Genome comparisons implemented by different ATCG input options are illustrated with example diagrams in which comparisons are indicated by ticks (A) or lines and arrowheads (B, C). By default, pairwise BLAST comparisons will be implemented in one direction only, between query and subject genomes (indicated with green ticks/arrowheads). If the "--bidirectionalblast" flag is provided, BLAST searches are run in both directions (indicated by both green and gold ticks/arrowheads) (see main text for more explanation). (A) The -s flag runs all-vs-all comparisons between genomes. In the example, genomes 1, 2, and 3 are provided in a multi-FASTA file (other input format options are described in the main text). (B) The -1 and -2 flags are used to run comparisons between one or more query genomes (provided to flag -1) and one or more subject genomes (provided to flag -2). In the example, genomes 1, 2, and 3 are provided as a multi-FASTA file and queried against genome 4. (C) The -p flag is provided with a tabular file listing the file paths of pairs of genomes which will be compared. In the example, the genome_pairs.tsv tabular file contains 3 rows and specifies the same comparisons as those specified in example B.

FASTA parsing and BLAST searches

FASTA input is parsed using Biopython (Cock et al., 2009). During parsing, the sequence lengths of each FASTA record in the input (multi-)FASTA files is determined. In addition, if file or stdin input is provided, then, during the parsing step, FASTA records are partitioned by genome (according to the FASTA headers, as described above). Genome names (e.g. for labelling output files) are determined from FASTA input as follows: for directory input or -p flag input, the genome name is the file name with suffix excluded; for file or stdin input, genome names are extracted from FASTA headers during the parsing step. After parsing FASTA input, BLAST databases are created for each genome (this step will be parallelised if multiple threads are specified). Pairwise BLAST searches are run between input genomes, with comparisons determined according to input flag, as described above. BLAST parameters including “evalue”, “word_size”, “task”, and “num_threads” (Camacho et al., 2009) can be adjusted by the user.

By default, ATCG runs BLAST comparisons in one direction only – between query and subject genomes – but if the “--bidirectionalblast” flag is provided, BLAST searches are run in both directions (as illustrated in Figure 1). A motivation for running bi-directional BLAST is that BLAST results are non-symmetric (Auch et al., 2006); with unidirectional BLAST, results may differ according to which set of genomes is provided to the query (-1) flag and which to the subject (-2) flag (in the case of -s and -p flag input, genomes are assigned as queries or subjects according to the alphanumeric order of genome names). If bi-directional BLAST is specified, statistics are averaged across both directions such that results do not vary according to which genomes are provided to the -1 and -2 flags. Nevertheless, previous research demonstrates that BLAST asymmetry has only a minor influence on whole-genome comparison statistics (Auch et al., 2006); hence, to reduce computation time, unidirectional BLAST is the default option in ATCG.

Alignment parsing and calculation of ANI and distance statistics

BLAST alignment data comprises multiple components: each alignment has a strand orientation, and is also associated with similarity statistics such as percent identity and bitscore; in addition, an

alignment comprises two sets of interval data (query and subject start and end positions) (Altschul et al., 1997). Genome-genome BLAST alignment data must be parsed in order to calculate statistics such as ANI and distance scores. Existing algorithms for parsing whole-genome BLAST alignments include the BioPerl MapTiling function (Stajich et al., 2002; bioperl.org, 2019) and GGDC (mentioned in the Introduction) (Meier-Kolthoff et al., 2013). Briefly, the MapTiling function divides query and subject intervals into a "disjoint decomposition" of non-overlapping intervals, from which intergenomic statistics can be calculated. However, because query and subject interval sets are parsed independently, paralogous regions may not be completely eliminated from ANI calculations (as illustrated in Figure 2). On the other hand, using the "greedy-with trimming" algorithm (Auch et al., 2010a; Meier-Kolthoff et al., 2013), GGDC filters and truncates intervals to obtain a set of putatively orthologous alignments which are non-overlapping in both query and subject interval sets (Henz et al., 2005; Meier-Kolthoff et al., 2014b) (and see Figure 2). Unfortunately, the GGDC algorithm is not described in detail, and up-to-date source code is not available. Therefore, I developed an *ab initio* alignment parsing algorithm using the GenomicRanges R package (Lawrence et al., 2013). In common with many other ANI implementations, including GGDC, ATCG aims to eliminate paralogous alignments from the computation of intergenomic statistics (Jain et al., 2018), as described below.

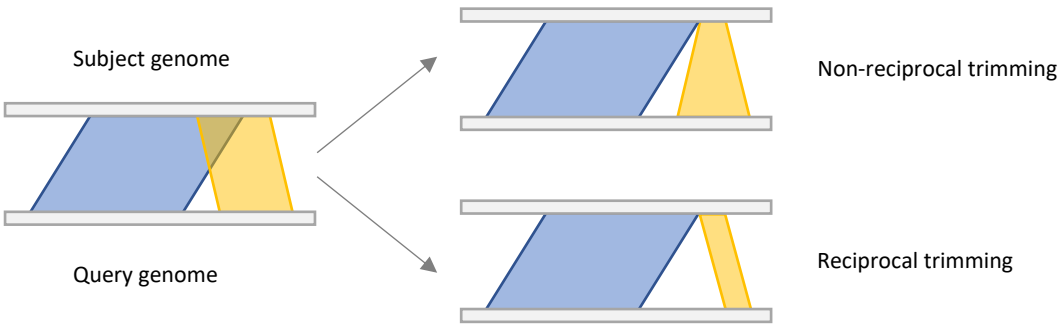


Figure 2. Approaches for trimming overlapping alignments. In this scenario, there are two alignments (coloured blue and gold) in positive strand orientation, which overlap on the subject genome. The gold alignment is suboptimal and must therefore be truncated. In non-reciprocal trimming (implemented by BioPerl MapTiling), subject and query genome alignment intervals are considered independently, so only the subject genome alignment interval is truncated. In reciprocal trimming (implemented by GGDC and ATCG), the truncation of the subject genome interval is applied reciprocally to the corresponding query interval.

Overall, ATCG alignment parsing proceeds by (parallelised) iteration through alignments associated with each subject genome, and at each iteration, a split-apply-combine strategy is adopted (Wickham, 2011b). Specifically, alignments are split by query genome; alignments representing query-subject pairs are parsed so that statistics can be calculated from non-overlapping alignments; finally, statistics are combined across splits and across subject genome iterations to produce a primary output file, recording statistics for all query-subject comparisons. If a genome comprises multiple sequences, query and subject alignment intervals for that genome are initially shifted so that intervals from different contigs of the same genome do not overlap (this is essentially equivalent to concatenating contigs prior to analysis). Figure 3 shows the alignment parsing process for a given query-subject pair. Specifically, the Figure illustrates real alignment data from the comparison of two closely related plasmids from a recent hospital outbreak study (Weingarten et al., 2018) (NCBI accessions for query and subject genomes are NZ_CP026395.1 and NZ_CP026188.1, respectively). To generate Figure 3, ATCG was run with dc-megaBLAST to generate alignments which, for simplicity, were filtered using stringent E-value and alignment length thresholds ($1e^{-200}$ and 1 kb, respectively). ATCG alignment parsing code was used in conjunction with the R packages genoplotR and ggplot2 (Wickham, 2011a) to generate plots representing the initial query-subject alignment set, and the query/subject alignment intervals at each step in the alignment parsing workflow. For full details, see the corresponding tabular data, which is given in Appendix C: Table 1.

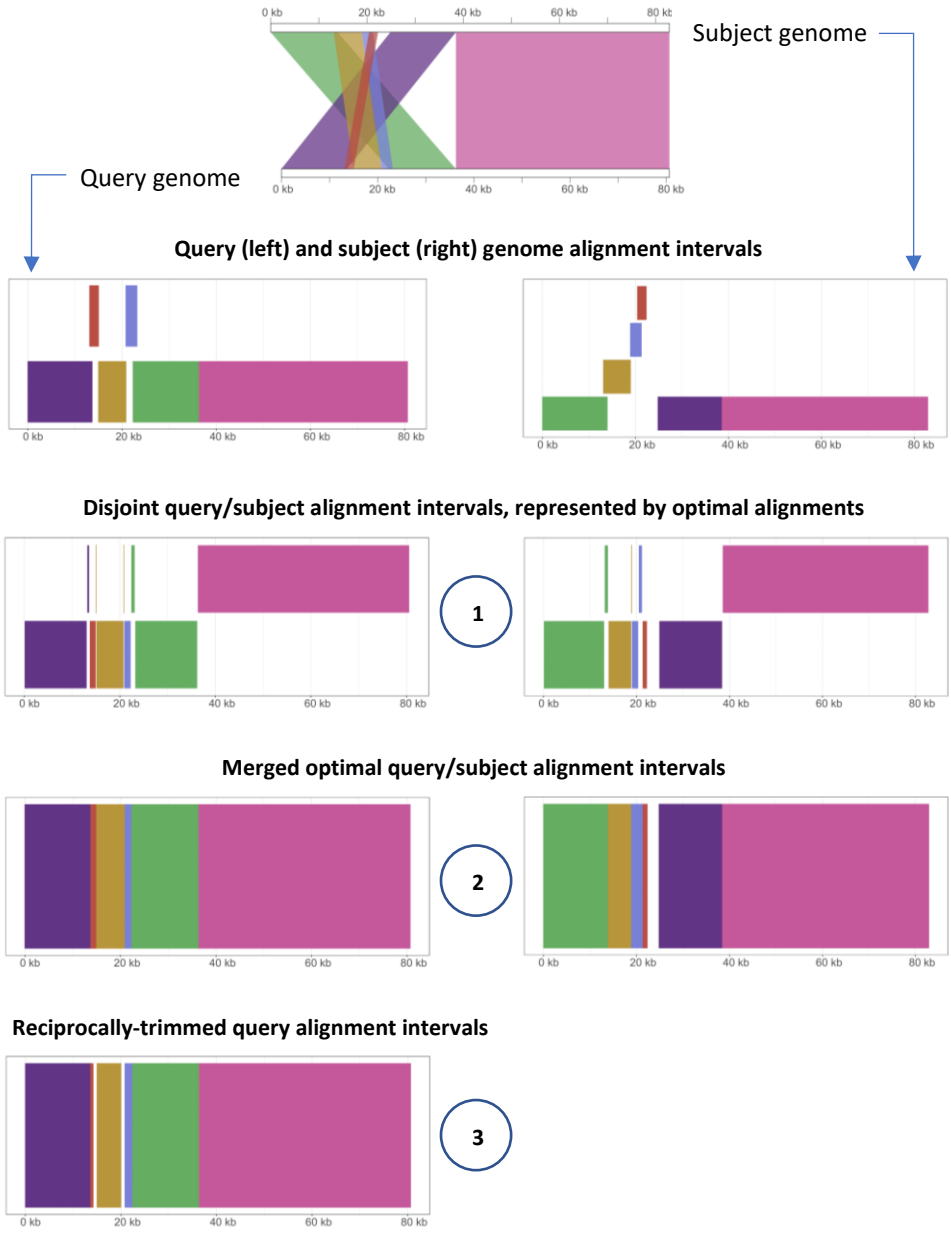


Figure 3. Steps implemented by ATCG when parsing query-subject BLAST alignments from comparison of two closely related plasmids (see main text for details). The top plot shows initial alignments between the query and subject genomes; different alignments are plotted in different colours, and transparency is used to show alignment overlaps (in other plots, the same colour coding is used except colours are opaque and therefore appear darker). All alignments except the blue alignment are in -ve strand (reverse complement) orientation. The plot second from top shows the same alignments, but with query and subject intervals shown separately (on the left and right, respectively). Overlapping alignment intervals are stacked in order of decreasing bitscore, from bottom to top. Subsequent plots show the 3-step alignment parsing process. 1) Query/subject intervals are divided into a "disjoint decomposition" of non-overlapping intervals. To aid visualisation, contiguous disjoint intervals are plotted in alternate top/bottom positions. As indicated by colouring, each disjoint interval is represented by a single optimal alignment (see main text for details). 2) Contiguous query/subject disjoint intervals represented by the same optimal alignment are merged, and some statistics are calculated at this stage (query/subject percent identity and coverage breadth). Alignments not represented in both query and subject interval sets are excluded, leaving a common set of alignments represented across query/subject intervals, from which breakpoint distance can be calculated. 3) Reciprocal trimming is conducted – query alignment intervals are truncated according to corresponding subject intervals (see main text for details). ANI is calculated from the reciprocally-trimmed query alignment intervals.

Alignment parsing proceeds in three steps, as illustrated in Figure 3, and similarity/distance metrics are calculated from the alignment data at different stages. A detailed description of the 3-step parsing procedure is given below. An outline of similarity and distance metrics, including formulae for their calculation, is provided in Supplementary Table S1.

1. Query and subject alignment intervals are divided into a "disjoint decomposition" of non-overlapping intervals (this is the same procedure implemented by the BioPerl MapTiling function, mentioned above). ATCG conducts the procedure using the GenomicRanges disjoin function with the arguments "ignore.strand=TRUE" and "with.revmap=TRUE". The latter argument means the alignment(s) intersecting each disjoint interval are recorded; where multiple alignments intersect a disjoint interval, a single optimal alignment is recorded, favouring higher bitscore.
2. Contiguous query and subject disjoint intervals represented by the same optimal alignment are merged using the GenomicRanges reduce function. After merging intervals, any alignments represented by multiple intervals are excluded, since this occurs when a suboptimal alignment is bisected by a higher-scoring (but shorter) optimal alignment. At this stage, query and subject genome percent identity and coverage breadth statistics are calculated from the merged intervals. The optimal alignments represented by the merged intervals can also be used as a basis for generating a set of best hit blast alignments (as described below). Subsequently, alignments not represented in both query and subject interval sets are excluded, such that remaining query and subject intervals represent a common set of optimal alignments. At this stage, the adjacency breakpoint distance can be calculated by normalising the number of adjacency breakpoints relative to either the total number of alignments (breakpoint distance d0) or the total length of aligned sequence (breakpoint distance d1). The contiguity of the optimal alignments can also be assessed using alignment N50 and L50 statistics, which are analogous to the widely used assembly contiguity N50/L50 statistics (Gurevich et al., 2013) (Supplementary Table S1). By default, a length filter

of 100 bp is applied to alignment intervals prior to calculating breakpoint distance and alignment contiguity statistics. The rationale for this is as follows: after steps 1 and 2, short remnants of suboptimal alignments (the flanks that did not overlap the optimal alignment) may remain. By excluding such alignments whilst retaining longer (e.g. at least gene-sized) alignments, I aim to obtain more robust breakpoint distance and alignment contiguity statistics.

3. As a result of steps 1 and 2, suboptimal query and subject alignment intervals are either excluded completely, or truncated to the endpoint of a higher-scoring alignment, with which they partially overlap. However, query and subject intervals have so far been considered independently, meaning that the alignment data may still be biased by overlapping alignments (see Figure 2). Therefore, the final stage of alignment parsing involves a reciprocal trimming process, applied to the query intervals. For example, this can be seen for the bronze coloured alignment in Figure 3. The subject alignment interval is truncated on the 5' end following alignment parsing steps 1 and 2. Since the alignment is in -ve strand orientation, the 5' subject truncation is applied reciprocally to the 3' end of the corresponding query interval. From the reciprocally-trimmed query alignment intervals, average nucleotide identity and distance metrics are calculated (Supplementary Table S1). The distance metrics calculated by ATCG closely correspond to the GGDC distance metrics, outlined by Meier-Kolthoff et al. (2013) (see Table 1 and Supplementary Table S1 for details).

In addition to generating intergenomic similarity and distance statistics, the alignment parsing procedure can also be used as a basis for outputting a set of best hit BLAST alignments, which can then be used for comparative genomic visualisation. Specifically, optimal alignments represented in query and/or subject genome merged intervals from step 2 (see above) are recorded, and corresponding original BLAST alignments are extracted. These alignments can be stored as best hits at this stage. However, by default, a further filtering step is applied: any alignments which are nested within longer alignments on either query or subject genome are filtered according to a coverage threshold (by

default, nested alignments must cover 50% of a longer alignment to be retained). This filtering parameter for nested alignments is analogous to the BLAST parameter “best_hit_overhang”, although the BLAST parameter can only be applied to query genome intervals (Camacho et al., 2009). Filtering of nested alignments by ATCG is achieved using the IRanges findOverlaps function with the argument “type=within”.

ATCG output

Table 2 shows output files produced from running ATCG with default parameters, using the -s (all-vs-all) input flag (similar output is produced using the other input flags). The primary output file is the comparisonstats.tsv tabular file; each row in this file contains similarity and distance statistics for a given pairwise genome comparison. With the -s input flag, it is also possible to specify output in matrix format, for a given similarity or distance metric of interest. In addition, with the -s flag, distance-based trees can be generated from any of the calculated distance metrics; specifically, either a hierarchical clustering dendrogram, or a tree built using the balanced minimum evolution method, implemented using the R ape package (Desper and Gascuel, 2002; Paradis et al., 2004). Trees are plotted and tree objects are also stored, allowing trees to be re-loaded for further analysis or re-plotting. Finally, with the --bestblastalignments flag, best hit BLAST alignments can be stored to file on a per subject genome basis; original unfiltered BLAST alignments can optionally be stored to file too. This allows pairwise genome alignments to be visualised as part of downstream analysis.

Table 2. ATCG output files

File name	Description
seqlengths.tsv	lengths of FASTA sequence records
filepathinfo.tsv	record of input genome file paths
blastsettings.txt	record of BLAST settings (BLAST parameters, whether unidirectional or bidirectional BLAST comparisons were run)
included.txt	genomes with BLAST alignments, which will therefore appear in comparisonstats.tsv
excluded.txt	genomes for which no BLAST alignments were detected (file may be blank)

excludedcomparisons.tsv	pairwise genome comparisons yielding no BLAST alignments (file may be blank)
output/comparisonstats.tsv	similarity and distance statistics for each pairwise combination of genomes which yielded BLAST alignments

4.3.2 Benchmarking

Benchmarking of ATCG was conducted using a subset of the benchmarking dataset previously used by Lee et al. (2016). In the Lee et al. study, bacterial genome assemblies were downloaded from the EzBioCloud genome database, and genome pairs comprising genomes belonging to the same genus were randomly selected to create the benchmarking dataset (>60,000 pairs were selected in total). Lee et al. did not provide the bacterial assemblies but NCBI Assembly database accessions were provided (see Table S1 from Lee et al. 2016). Therefore, to obtain assemblies for my benchmarking dataset, the assembly accessions from Lee et al. were used to retrieve assembly FASTA files using an EDirect command (Kans, 2013) (Supplementary Methods). To reduce computation time, I selected only 500 intra-genus pairs from each of the following most frequent genera: *Enterococcus, Vibrio, Streptococcus, Bacillus, Acinetobacter, Mycobacterium, Staphylococcus, Pseudomonas, Salmonella* (4500 genome pairs in total). The list of genome pairs and associated FASTA files are provided in Appendix C: Table 2 and File 1.

Benchmarking ANI calculations is challenging since there is no clear gold standard, arising from the fact that there are no gold standard whole genome alignments from which to compute ANI (Dewey, 2019). A common approach in this situation is to compare results from different methods in order to gauge sensitivity to different parameters. In addition, results can be interpreted according to theoretical expectations: it is reasonable to presume that alignments generated with more sensitive parameters (e.g. alignments seeded with shorter matches) may yield more reliable ANI calculations. Indeed, a previous benchmarking study considered ANIb as a pseudo-gold standard against which to benchmark other algorithms (Yoon et al., 2017), since ANIb is widely used, and based on a sensitive

BLAST algorithm (BLASTN, word size 11) (Goris et al., 2007). In this study, OrthoANIb was used as a benchmark against which to evaluate ATCG in terms of ANI calculation, as well as computation time (wall-clock and CPU time). ATCG was run with a range of different BLAST parameters (summarised in Table 3); selection of parameters was informed by parameters used in previous studies. For example, BLASTN has been used with default word size (word size 11) in the context of ANIb and OrthoANIb (Goris et al., 2007; Lee et al., 2016), while a longer (less sensitive) word size parameter (word size 38) has been proposed by developers of GGDC (Meier-Kolthoff et al., 2014a). Others have used megaBLAST rather than BLASTN for ANI calculations (Ciufo et al., 2018).

As well as benchmarking ATCG, other alignment-based tools for ANI calculation were also included in the benchmarking: OrthoANIu and MUMmer (specifically, a NUCmer/dnadiff pipeline, referred to as MUMmer for simplicity). These tools were selected because they are widely used, and freely available to run on the command-line. By default, MUMmer uses relatively long (20 bp) matches to seed alignments (compared to a default of 11 bp for BLASTN). In addition, by default, MUMmer seeds must be unique in the reference genome (Marçais et al., 2018). To examine the influence of the seed parameters on ANI calculations and runtime, MUMmer was run with default parameters and with high sensitivity parameters; specifically, a seed length of 11 bp (--minmatch 11) and using all seeds to initiate alignments (--maxmatch). Software versions and commands used are given in Table 3. All software tools were run with 20 threads on a dual 10-core Intel Ivy Bridge computer, with 160GB of 1833MHz RAM. Pairwise correlation between ANI calculations from OrthoANIb and other tools was assessed using Pearson's correlation coefficient. An optimal ANI calculation method should show strong linear correlation with the OrthoANIb benchmark. Hence, Pearson's correlation coefficient was judged to be an informative statistic, and indeed, it has been used previously in similar benchmarking studies (Li et al., 2015; Lee et al., 2016; Yoon et al., 2017; Jain et al., 2018). In addition to ANI calculation, ATCG and MUMmer also calculate alignment adjacency breakpoints which can be expressed as a breakpoint distance. MUMmer partitions adjacency breakpoints into relocation, translocation, and inversion breakpoints; the total number of MUMmer alignment adjacency

breakpoints was calculated as the sum of these metrics, for comparability with ATCG's alignment adjacency breakpoint metric. Correlation analysis between ATCG and MUMmer breakpoint distances was performed. All statistical analyses were performed in R (https://www.r-project.org/).

Table 3. Algorithms used in ANI benchmarking, including software versions and the commands run

Algorithm	Software version*	Command
ATCG BLASTN	ATCG v0.2.0	compare.py -t 20 -p genomepairpaths.tsv -o output --task blastn --wordsize 11 --breakpoint --alnlenstats
ATCG BLASTN word size 38	ATCG v0.2.0	compare.py -t 20 -p genomepairpaths.tsv -o output --task blastn --wordsize 38 --breakpoint --alnlenstats
ATCG dc-megaBLAST	ATCG v0.2.0	compare.py -t 20 -p genomepairpaths.tsv -o output --task dc-megablast --wordsize 11 --breakpoint --alnlenstats
ATCG dc-megaBLAST pre-trim approach**	ATCG v0.2.0	as above
ATCG megaBLAST	ATCG v0.2.0	compare.py -t 20 -p genomepairpaths.tsv -o output --task megablast --wordsize 28 --breakpoint --alnlenstats
MUMmer default	MUMmer v4.0.0.beta2	nucmer -t 20 --minmatch 20 -p output genomepair1path genomepair2path && dnadiff -p output -d output/.delta && python extractani.py
MUMmer high sensitivity	MUMmer v4.0.0.beta2	nucmer -t 20 --minmatch 11 --maxmatch -p output genomepair1path genomepair2path && dnadiff -p output -d output/.delta && python extractani.py
OrthoANIb	OrthoANI v1.40	java -jar OAT_cmd.jar -blastplus_dir ncbi-blast+/2.6.0/bin -num_threads 20 -fasta1 genomepair1path -fasta2 genomepair2path | python extractani.py
OrthoANIu	OrthoANI usearch v1.2	java -jar OAU.jar -u usearch/usearch -n 20 -f1 genomepair1path -f2 genomepair2path -t tempdir | python extractani.py

MUMmer, OrthoANIb and OrthoANIu commands were run in a for loop (iterating through genome pairs), and custom scripts (designated "extractani.py") were used to extract ANI and other statistics from output, as well as remove temporary files at each iteration. MUMmer ANI is based on average percent identity of 1-to-1 alignments, as reported by dnadiff.

*BLAST version 2.6.0 was used for all BLAST-based methods.

**For ATCG dc-megaBLAST (a well-performing ATCG method, see Results and discussion), I also investigated an alternative approach to calculating ANI. Rather than calculating ANI from reciprocally-trimmed query alignment intervals as normal (step 3 in Figure 3), ANI was calculated as an average of query and subject genome coverage breadth prior to reciprocal-trimming (step 2 in Figure 3).

4.3.3 Analysis of Manchester *bla*$_{KPC}$ outbreak data

To demonstrate how ATCG can be applied, I conducted comparative genomic analysis of long-read sequenced plasmids, collected from a study focusing on a hospital outbreak of *bla*$_{KPC}$-encoding Enterobacterales in Manchester, UK (see Decraene et al., 2018). Decraene et al. had previously analysed *bla*$_{KPC}$ *Escherichia coli* samples using Illumina data, which revealed the importance of a clonal strain of *E. coli* ST216, especially in relation to outbreaks at the Manchester Royal Infirmary in 2012 (gerontology wards 45 and 46) and 2015 (Manchester Heart Centre [MHC]) (Decraene et al., 2018). However, at the time, only two samples were long-read sequenced, limiting understanding of plasmid dynamics, and analysis was restricted to *E. coli* samples only. Here, I analysed 41 samples which were sequenced using both short (Illumina) and long (Oxford Nanopore, PacBio) sequencing reads (see below). The samples were selected with the aim of incorporating a range of species and settings. Specifically, there were 19 *E. coli* ST216 samples from the gerontology wards and the MHC, which had been previously analysed by Decraene et al. using short-read data. The remaining samples were from various Enterobacterales species (including *Klebsiella pneumoniae* and *Enterobater cloacae*) and were collected from Manchester, regional, and national hospital settings. Kraken was used for *in silico* species identification, while BLAST-based methods were used for typing the KPC gene and the Tn*4401* transposon (Wood and Salzberg, 2014; Decraene et al., 2018). For further details on the sample set, see Results and discussion.

Wet lab sequencing work was performed by collaborators. All 41 samples were sequenced using Illumina HiSeq paired-end reads, as reported in Decraene et al., and long-read sequencing was also conducted for all samples. Specifically, 17 samples were sequenced with MinION sequencing (Oxford Nanopore Technologies, UK), and 24 samples were sequenced using PacBio SMRT sequencing. Assembly of PacBio reads was conducted by Dr Hang Phan. Specifically, PacBio reads were assembled using the HGAP pipeline (Chin et al., 2013) and resulting assembled contigs were manually curated to create closed structures where possible, by using BLASTN to identify (and remove) overlaps at the

ends of contigs. MinION sequencing was conducted using an R9.4 flow cell and a 1D library preparation kit (SQK-LSK108, following manufacturer's protocol). Some samples were multiplexed using a native barcoding kit (EXP-NBD103). Basecalling was carried out using a combination of Metrichor and Albacore basecallers (Wick et al., 2019). Porechop (v0.2.2, https://github.com/rrwick/Porechop) was used for adapter/barcode trimming and de-multiplexing of nanopore reads. Then, using Filtlong (v0.1.1, https://github.com/rrwick/Filtlong), reads were quality-filtered (using paired-end Illumina reads as a reference), and subsampled to 500 Mbp (see Supplementary Methods). Illumina and filtered nanopore reads were hybrid assembled using Unicycler (v0.4.1) (Wick et al., 2017). Unicycler was initially run in "standard" mode, but samples that were not completely circularised in standard mode, were re-assembled using "bold" mode.

bacterialBercow (v0.1.0, https://github.com/AlexOrlek/bacterialBercow) was used to classify contigs from the sample assemblies; the PlasmidFinder and rMLST databases used with bacterialBercow were downloaded on 20th Mar 2020. Plasmid contigs from each sample were extracted; any samples where the fraction of plasmid contigs was smaller than that of unknown contigs were excluded. bacterialBercow was also used for species identification (to corroborate Kraken assignments), and for identifying contaminant samples for exclusion. For remaining plasmid contigs, ATCG was used to conduct an all-vs-all comparison among plasmid contigs using dc-megaBLAST. Output was converted to a sample–sample edgelist as follows: where two different samples were linked by one or more pairwise plasmid contig comparisons showing at least 99% ANI and at least 50% coverage breadth across both contigs, a corresponding inter-sample edge was recorded. The resulting edgelist and associated sample metadata (including species, collection date, and location) was visualised as a network using the igraph R package (Kolaczyk and Csardi, 2014). After inspection of the network, sample pairs of interest were further investigated through comparative genomic visualisation of best hit alignments with Easyfig (Sullivan et al., 2011). Genes were annotated using Prokka (v. 1.14.6); resistance genes were annotated by BLASTN searching ResFinder (retrieved 25th September 2019)

(Zankari et al., 2012) using percent identity and coverage thresholds of 90% and 60%, respectively (BLASTN parameter: max_target_seqs: 5000).

In order to demonstrate the utility of the ATCG best hit filtering procedure, I used ATCG to conduct all-v-all comparisons among three plasmid genomes (plasmids pCAD3, pKPC-CAD1, pKPC-CAD2) which had previously been analysed by Decraene et al. Both original and best hit BLAST alignments were stored (the default 50% coverage breadth threshold was applied to exclude nested alignments). Genome comparisons were visualised using genoplotR (Guy et al., 2011; Seemann, 2014); two visualisations were plotted using original and best hit BLAST alignments, respectively.

4.4 Results and discussion

4.4.1 Features of ATCG in relation to other software for calculating whole-genome comparison metrics

ATCG offers numerous advantages in comparison with other tools, as indicated in Table 4. Firstly, ATCG is flexible in terms of input options. For example, it can handle input from the stdin stream as well as handling gzip compressed FASTA files. In addition, the different ATCG input flags allow for flexibility in terms of conducting comparisons (all-vs-all comparisons, query-vs-reference comparisons, pairwise comparisons specified by a list of genome pairs). In contrast, MUMmer and the OrthoANI tools permit only a single genome pair comparison per run. The GGDC web tool is similarly inflexible; each submission to the server can run either a single genome pair comparison or a query-reference comparison involving up to 75 reference genomes. Of course, with the standalone command-line tools, it is relatively trivial for a user with bioinformatics expertise to create a custom script to iterate through desired genome pair comparisons. However, this will introduce various computational inefficiencies, which might arise from re-loading software library dependencies; re-fragmenting the same input genomes (OrthoANI tools); and re-building suffix arrays (MUMmer4),

USEARCH database indexes (OrthoANIu), or BLAST databases (OrthoANIb). By explicitly accommodating different comparison options, ATCG bypasses such inefficiencies; taking the comparisons shown in Figure 1C as an example, only one BLAST database will need to be created (for genome 4). A further advantage of ATCG is that it calculates a wider range of statistics than any other tool. The OrthoANI algorithms fragment input genomes, precluding calculation of structural similarity and alignment contiguity metrics. Besides ATCG, the only other tool which provides metrics on structural similarity and alignment contiguity is MUMmer. ATCG and MUMmer are also the only tools which store genome-genome alignments, enabling downstream comparative genomic visualisation. However, MUMmer does not calculate distance scores. Resemblance and containment distances (ATCG distance scores d6 and d7) are especially informative since they capture both percent identity and coverage breadth; indeed, in recent years, these distances have been widely used for large-scale comparative genomic analyses of bacterial genomes (including plasmids), and bacterial metagenomes (Ondov et al., 2016; Phillippy and Ondov, 2017; Jesus et al., 2019; Acman et al., 2020).

Table 4. Comparison of features of ATCG and other alignment-based software tools for calculating whole-genome comparison metrics

	ATCG	**GGDC**	**MUMmer (NUCmer + dnadiff)**	**OrthoANIb/ OrthoANIu**
Standalone software available to download	✓	×	✓	✓
Web-tool available	×	✓	×	×
Flexible input and comparison options	✓	×	×	×
Stdin stream input accommodated	✓	×	✓	OrthoANIu only
Gzip file input accommodated	✓	×	×	×
ANI calculated	✓	✓*	✓	✓
Coverage breadth calculated	✓	✓*	✓	OrthoANIu only
Distance scores calculated	✓	✓*	×	×
Structural similarity/distance and alignment contiguity metrics calculated	✓	×	✓	×**
Alignments can be stored for downstream analysis e.g. visualisation	✓	×	✓	×**

Feature presence/absence is indicated by ticks/crosses. Besides ATCG, the tools summarised represent a non-exhaustive selection of widely used tools.
*GGDC calculates dDDH similarity scores from intergenomic distances. Different distance scores (and derived dDDH scores) reflect either percent identity (i.e. analogous to ANI) and/or coverage breadth (for details see main text, Introduction). The GGDC web tool only calculates three distance scores (d0, d4 and d6).
**OrthoANIb/OrthoANIu algorithms fragment input genomes; this represents a fundamental barrier to calculating structural similarity metrics or generating useful whole-genome comparative genomic visualisations from stored BLAST alignments.
The GGDC web-tool can be accessed here: https://ggdc.dsmz.de/ggdc.php#; MUMmer can be downloaded here: https://github.com/mummer4/mummer/releases; OrthoANIb and OrthoANIu can be downloaded here: https://www.ezbiocloud.net/tools

4.4.2 Computational runtime benchmarking

Figure 4 shows the results of benchmarking ANI calculation tools using the dataset of 4500 intra-genus bacterial genome pairs. Considering wall-clock times, MUMmer with default settings is the fastest tool (~5 hrs), OrthoANIb is the slowest (~40 hrs), and OrthoANIu is intermediate (~13 hrs); this corroborates previous benchmarking showing MUMmer (with default parameters) to be the fastest tool (Yoon et al., 2017). However, previous benchmarking studies have not examined MUMmer under different parameter settings. My benchmarking shows that with high-sensitivity settings, MUMmer is actually the slowest of all methods except OrthoANIb (Figure 4), highlighting the importance of investigating parameters as well as overall algorithms. Indeed, ATCG wall-clock times vary widely depending on BLAST parameters; as expected based on BLAST documentation (Camacho et al., 2009), using megaBLAST or BLASTN with long word size (38 bp) leads to relatively fast runtimes, whereas BLASTN with default word size (11 bp) is slower, although not as slow as OrthoANIb or high-sensitivity MUMmer.

The substantially slower wall-clock runtime of OrthoANIb compared with ATCG BLASTN (~40 hrs vs ~23 hrs) can be attributed to inefficiency of the OrthoANI algorithm, since both methods use the same principal BLAST settings (BLASTN, with default word size). It is likely that this inefficiency results from the fragmentation step, common to all BLAST-based ANI calculation tools except GGDC and ATCG. There will be some overhead associated with the fragmentation step per se, but in addition, running BLAST with fragmented input sequences will also be more inefficient. For instance, because

alignments cannot be extended beyond the 1 kb fragment length, more extension operations will be invoked, leading to a higher runtime (Altschul et al., 1997). Although ATCG methods perform relatively well in terms of wall-clock time, examining the CPU times demonstrates scope for further improvement. If CPUs were saturated and relatively little time was spent on input/output (I/O) operations (i.e. a CPU-bound process), then the ratio of CPU time to wall-clock time would approach 20 (since 20 threads were allocated). This is not the case, so ATCG must be primarily I/O bound and/or memory-bound; further performance profiling of ATCG, including memory usage profiling, could help home in on bottlenecks. Future developments to improve performance could include reduced file read/write operations through streaming, and improved parallelisation. For example, ATCG could be parallelised on a per-genome pair basis (i.e. splitting comparison tasks in an embarrassingly parallel fashion).

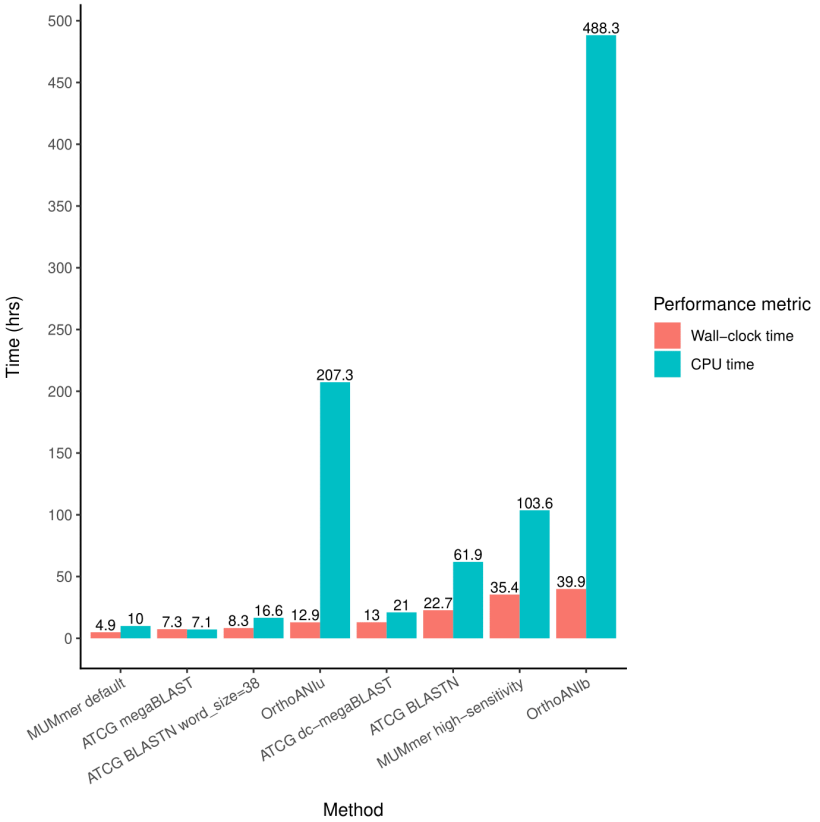


Figure 4. Wall-clock and CPU time benchmarking of different methods for ANI calculation. The methods encompass four software tools: ATCG, MUMmer (NUCmer + dnadiff), OrthoANIb, and OrthoANIu. ATCG was run

with four different parameter settings, while MUMmer was run with two different parameter settings. Benchmarking was conducted using a dataset of 4500 intra-genus pairs of bacterial genomes.

4.4.3 ANI calculation benchmarking

Relationships between ANI values calculated by OrthoANIb (reference method) and other ANI calculation methods (see Table 3 in Methods) are shown using pairwise scatterplots, arranged in a grid (Figure 5). Plots are ordered according to Pearson's correlation coefficient, from top left to bottom right (higher to lower correlation). Focusing on the different ATCG methods, ATCG run with dc-megaBLAST and BLASTN (word size 11) shows relatively strong correlation with OrthoANIb, whereas megaBLAST and BLASTN with word size 38 leads to weaker correlation. This is unsurprising given that the latter two methods have lower sensitivity parameters. Standard ATCG ANI is calculated from "reciprocally-trimmed" alignments (see Methods) but an alternative is to calculate ANI from non-overlapping optimal alignments prior to reciprocal trimming. Omitting the reciprocal trimming step will reduce runtime (to what extent has not been investigated), but as discussed in the Introduction, it will also mean that ANI values are more likely to be influenced by paralogous sequence. Figure 5 shows that for dc-megaBLAST, calculating ANI prior to reciprocal trimming ("ATCG dc-megaBLAST pre-trim") leads to slightly reduced correlation with OrthoANIb ANI values, compared with standard ATCG dc-megaBLAST. The reduction in correlation is expected given that OrthoANIb, in common with the standard ATCG approach, computes ANI from putatively orthologous alignments.

Overall, inspection of the scatterplots shows that at high OrthoANIb ANI values (approximately ≥90%) all methods show relatively strong linear correlation with OrthoANIb ANI. However, for some methods (notably, ATCG megaBLAST, ATCG BLASTN word size 38, and MUMmer default) correlation degrades at lower OrthoANIb ANI values, reflected in lower overall correlation coefficients. Specifically, the more weakly correlating methods show inflated ANI values at lower OrthoANIb ANI. These findings are in agreement with previous benchmarking studies, which have compared MUMmer under default settings to ANIb (a high-sensitivity BLAST-based method similar to OrthoANIb) (Li et al., 2015; Yoon et

al., 2017; Palmer et al., 2020). One explanation for the correlation patterns is as follows. Among closely related genome pairs, it is likely that a high proportion of homologous sequence exhibits high identity and will therefore be incorporated into the alignment sets of both high-sensitivity and lower sensitivity methods (e.g. OrthoANIb and MUMmer default, respectively); as a result, ANI calculations are similar across methods. Among distantly related genomes, high-sensitivity methods may still be able to align large tracts of the genomes, but alignments will tend to exhibit relatively low nucleotide identity. On the other hand, the low-sensitivity methods will omit low nucleotide identity alignments, and the ANI will be calculated from a restricted set of relatively high identity alignments, covering a smaller tract of the genomes (hence, the inflated ANI values of MUMmer default, at low OrthoANIb ANI). To corroborate this explanation, further analysis would be required, for example, examining how genome coverage breadth values vary across methods and across ANI values.

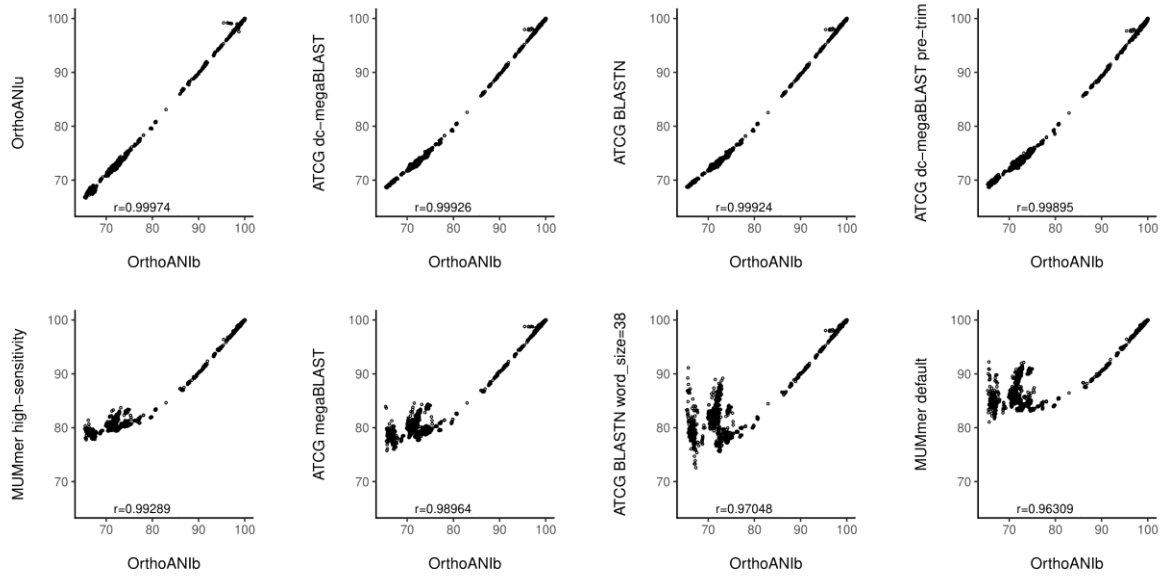


Figure 5. Pairwise scatter plots showing ANI values calculated by OrthoANIb (x-axis) in comparison with other methods (see Table 3 for details) for 4500 bacterial genome pairs. Plots are ordered by Pearson's correlation coefficient (r) from top left to bottom right (higher to lower correlation). P-values associated with correlation coefficients are not shown, since the assumption of normality is violated.

Besides ANI, another genome relatedness metric calculated by ATCG is the number of alignment adjacency breakpoints, which can be represented as a breakpoint distance in two ways – either normalising by total alignments (described here as "breakpoint distance d0") or normalising by total aligned sequence length (described here as "breakpoint distance d1"). MUMmer also reports alignment adjacency breakpoints, so equivalent breakpoint distances can be derived. Breakpoint distance d0 is commonly used in the context of gene adjacencies (Sankoff et al., 2000), and authors of GGDC have suggested that it can be applied to alignment adjacency data (Auch et al., 2006). However, in the case of alignment data, there is a potential bias associated with alignment contiguity: reduced alignment contiguity will inflate the denominator (total alignments). Consequently, breakpoint distance d0 is likely to be a less informative measure of structural conservation of alignment data. Indeed, this is supported by correlation analysis of ATCG vs MUMmer-derived breakpoint distances (Figure 6). Breakpoint distance d1 data shows higher correlation, suggesting that this metric is more robust to different alignment contiguity scenarios (arising from the different methods for generating the alignments).

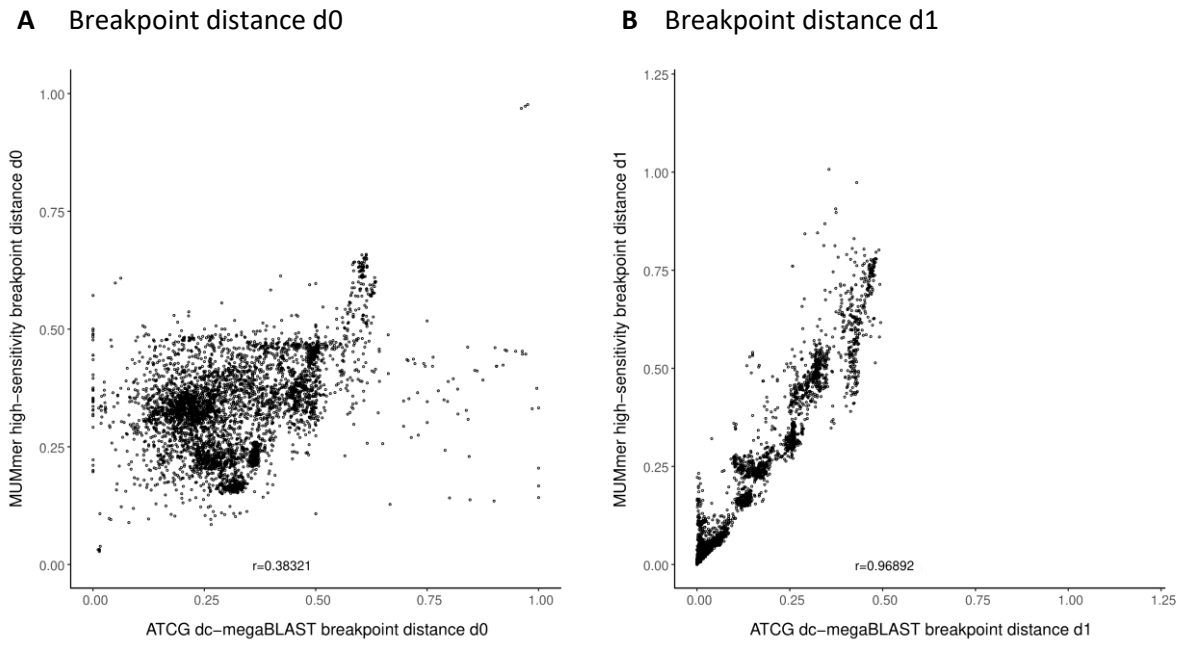


Figure 6. Scatterplots showing correlation between breakpoint distances calculated using ATCG dc-megaBLAST and MUMmer high-sensitivity methods, using (A) breakpoint distance d0, and (B) breakpoint distance d1.

Pearson's correlation coefficients (r) are indicated on each plot. P-values associated with correlation coefficients are not shown, since the assumption of normality is violated.

4.4.4 Implications of benchmarking results for users of ATCG

Based on benchmarking of computational performance and ANI calculation, dc-megaBLAST was selected as the default BLAST algorithm in the current ATCG version (v0.2.0). This is because dc-megaBLAST showed strong correlation with OrthoANIb (an established high-sensitivity method for ANI calculation), while exhibiting a lower runtime compared with default BLASTN. The benchmarking highlights USEARCH as a potentially attractive search algorithm for ANI calculation: OrthoANIu (USEARCH) substantially reduces runtime compared with OrthoANIb (BLASTN), while maintaining high correlation with OrthoANIb ANI values. Therefore, future ATCG implementations could incorporate USEARCH, as an additional option, alongside the various BLAST algorithms. Finally, regarding alignment adjacency breakpoint distance, Figure 6 demonstrates that breakpoint distance d1 should be adopted rather than breakpoint distance d0.

Although dc-megaBLAST has been adopted as an optimal BLAST algorithm in ATCG, all benchmarked ATCG methods appear to correlate well at high OrthoANIb values (e.g. >90%). Therefore, if the aim of using ATCG is simply to identify very similar genomes (e.g. in the context of transmission studies) then all benchmarked methods would be suitable, especially if used in combination with a coverage breadth threshold, which may prevent spuriously high ANI values from distantly related genomes (see above). Another common application of ANI is in bacterial genomic taxonomy. A detailed discussion of this field is beyond the scope of the chapter. Briefly, many researchers have conventionally suggested a 95–96% ANI threshold to delineate species boundaries (Richter and Rosselló-Móra, 2009). GGDC developers have calibrated their tool according to wet lab DDH, and apply their dDDH similarity scores with a 70% threshold to delineate species (Meier-Kolthoff et al., 2013). However, the notion of a single genome relatedness threshold for species delineation is flawed, for several reasons. Firstly, ANI metrics may differ according to the ANI calculation approach applied; for example, if a more

sensitive alignment method is used, lower identity alignments may be incorporated, lowering the ANI score. Secondly, intra-species diversity and inter-species genomic distances will vary across taxa. Hence, it has been suggested that ANI thresholds should be specific to a given ANI calculation method, and tuned according to taxa (Colston et al., 2014; Palmer et al., 2020). Furthermore, rather than relying solely on ANI to determine taxonomic boundaries, ANI should be used alongside phenotypic and phylogenetic approaches as part of a polyphasic taxonomy (Raina et al., 2019).

4.4.5 Comparative genomics of the Manchester *bla*$_{KPC}$ outbreak

A full description of the Manchester *bla*$_{KPC}$ outbreak sample set is provided in Appendix C: Table 3. Of the 41 assembled samples, one sample (H102700301) was excluded as a contaminant, based on bacterialBercow output which demonstrated multi-allelic rMLST loci (see Appendix C: Table 4). A further two were excluded due to poor assembly (such that there was a higher ratio of "unknown" to "plasmid" sequence). Of the 38 remaining samples, all harboured the *bla*$_{KPC-2}$ gene, according to BLAST-based typing. The predominant species represented are *E. coli* and *K. pneumoniae* and speciation assignments from Kraken were broadly concordant with assignments from bacterialBercow, although there were six minor discrepancies in total (Appendix C: Table 3). Notably, Kraken reports *E. cloacae* whereas bacterialBercow reports *E. hormaechei*; for consistency with Decraene et al. (2018), Kraken species assignments are adopted here. Arguably, species assignments from bacterialBercow are likely to be more reliable, since rMLST allele profiles are expert curated, whereas the Kraken database is derived directly from NCBI RefSeq, where (at least until very recently) species assignments are not rigorously curated (Ciufo et al., 2018; Jolley et al., 2018).

Figure 7 shows the sample-level plasmid sequence similarity network constructed based on pairwise comparisons between plasmid contigs from different samples. The 38 nodes in the network (representing bacterial samples) are linked by 184 edges; the edgelist data underlying the network is given in Appendix C: Table 5. Links were included regardless of whether contig pairs harboured a *bla*$_{KPC}$

gene, so the network reflects sharing of overall plasmid genetic content, rather than only *bla*$_{KPC}$-encoding plasmid content (although see Figure 8 below, where sharing of resistance gene-encoding plasmid content between samples of interest is investigated). Furthermore, the network needs to be interpreted cautiously given sampling bias – the 38 samples shown in the network are a small subset of total samples collected during the study period (most of which were not long-read sequenced), and sample collection methods varied during the study period, as described by Decraene et al.

Overall, the earliest samples in my dataset (collected 2009–2011) mainly comprise diverse strains of *K. pneumoniae* (~10 different sequence types [STs]) (see Appendix C: Table 3). Most of these samples (nodes 1, 2, 4, 5, 8) cluster together, consistent with shared plasmid genetic content across early *K. pneumoniae* strains. Later samples are primarily from the clonal *E. coli* ST216 outbreak focused at the Manchester Royal Infirmary (MRI). All *E. coli* ST216 samples belong to the same cluster in the network, and there is no clear division between earlier samples collected from the MRI gerontology ward (in 2012) and later samples collected from the MRI heart centre (in 2015). This suggests that the *E. coli* ST216 outbreak was associated with plasmids sharing conserved genetic content which persisted over a period of several years. The persistence of plasmid genetic content among the *E. coli* ST216 isolates was previously highlighted by Decraene et al., 2018. From limited long-read sequencing, Decraene et al. had identified an example of samples separated by a ~2.5-year time period, which nevertheless shared highly similar plasmids (samples are labelled as 15 and 34 in my network, and were collected from the MRI gerontology wards and MRI heart centre, respectively).

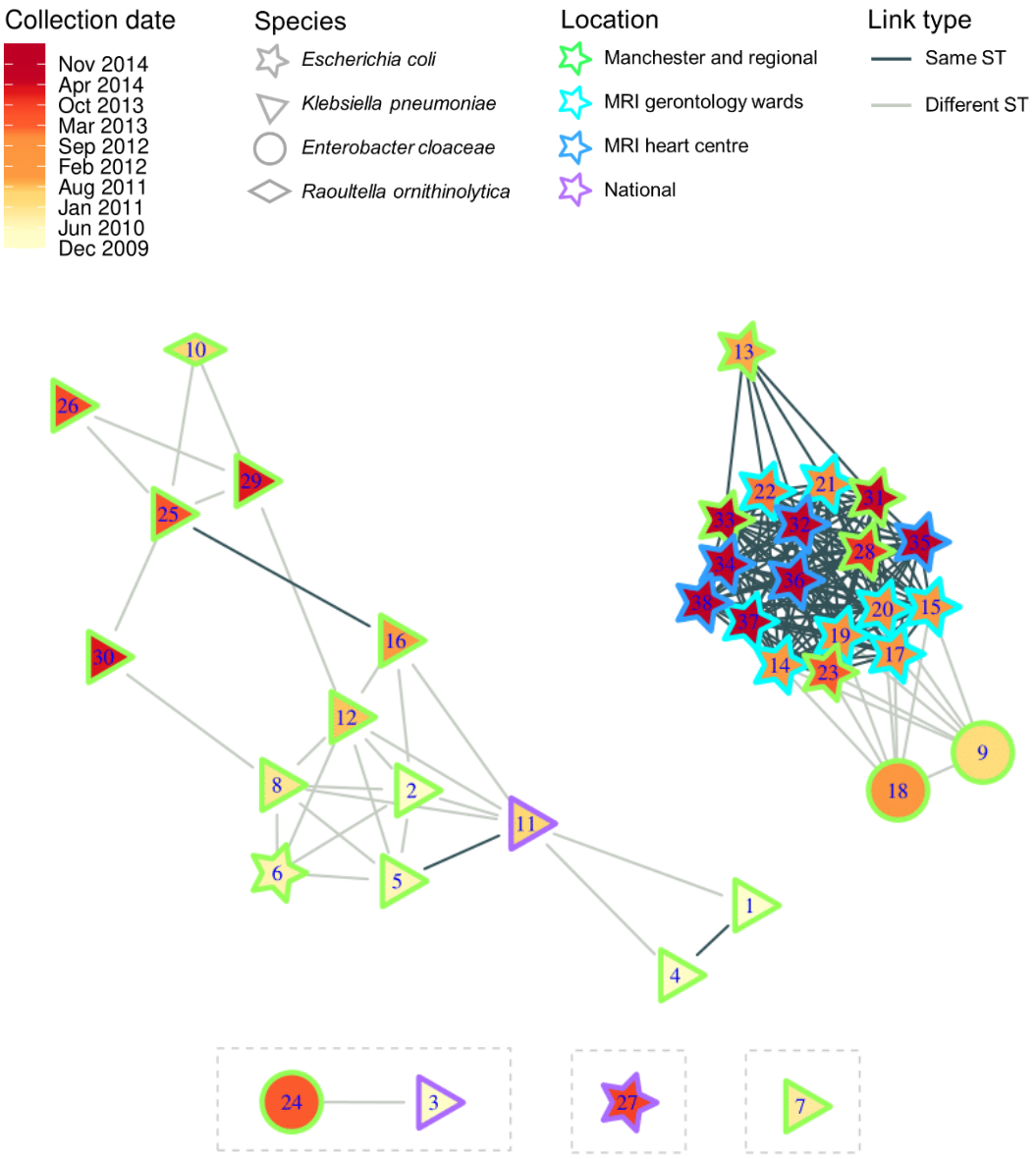


Figure 7. Sequence similarity network of sample-level plasmid genetic content; samples were collected from the *bla*$_{KPC}$ outbreak centred in Manchester, UK. Nodes (representing bacterial samples) are linked if plasmid genetic content is shared at high sequence identity (based on pairwise plasmid contig comparisons, see Methods). Node fill colour indicates sample collection date; node outline colour indicates collection location; node shape indicates species. If linked samples are from the same species and sequence type (ST), the link is coloured charcoal, otherwise, light grey. The Fruchterman-Reingold algorithm was used for node layout; four nodes that were not part of the main connected component were manually placed at the top of the network (shown within grey boxes).

Besides connectivity between *E. coli* ST216 gerontology and heart centre samples, numerous other examples of persistence of plasmid genetic content across time can be seen from the network. This includes examples of connectivity between different species. For example, the *Raoultella ornithinolytica* sample collected in 2011 is connected to a *K. pneumoniae* sample collected in 2014 (nodes 10 and 29, respectively). Other examples of plasmid genetic content shared across species are the two *E. cloacae* samples (nodes 9 and 18) which connect with the cluster of *E. coli* ST216 samples. Figure 8 shows a comparative genomic visualisation of shared plasmid genetic content between inter-species sample pairs. As can be seen from the figure, a stretch of sequence ~70kb is shared at high identity between the second largest contigs ("contig 2") from the *R. ornithinolytica*/*K. pneumoniae* samples (nodes 10 and 29). Within this ~70kb section, a ~10kb inverted section encodes four antibiotic resistance genes. In addition, between *R. ornithinolytica* sample contig 3 and *K. pneumoniae* sample contig 2, a shared inverted segment of ~10kb contains a bla_{KPC-2} resistance gene. Regarding the *E. cloacae*/*E. coli* ST216 sample pairs (nodes 9 and 17; and nodes 18 and 19), comparative visualisations appear almost identical. This is as expected since the corresponding samples (nodes 9/18 and nodes 17/19) show very high BLAST similarity (>99% identity, >98% coverage breadth). The visualisations show that long stretches of high similarity sequence are shared between sample pairs (specifically between the largest contig of the *E. coli* ST216 samples and the only contig of the *E. cloacae* samples). The shared sequence includes >10 antibiotic resistance genes, including a bla_{KPC-2} gene. Taken together, these results demonstrate sharing of long stretches of plasmid genetic content between species, which may have arisen due to horizontal dissemination of plasmid content during the course of the outbreak.

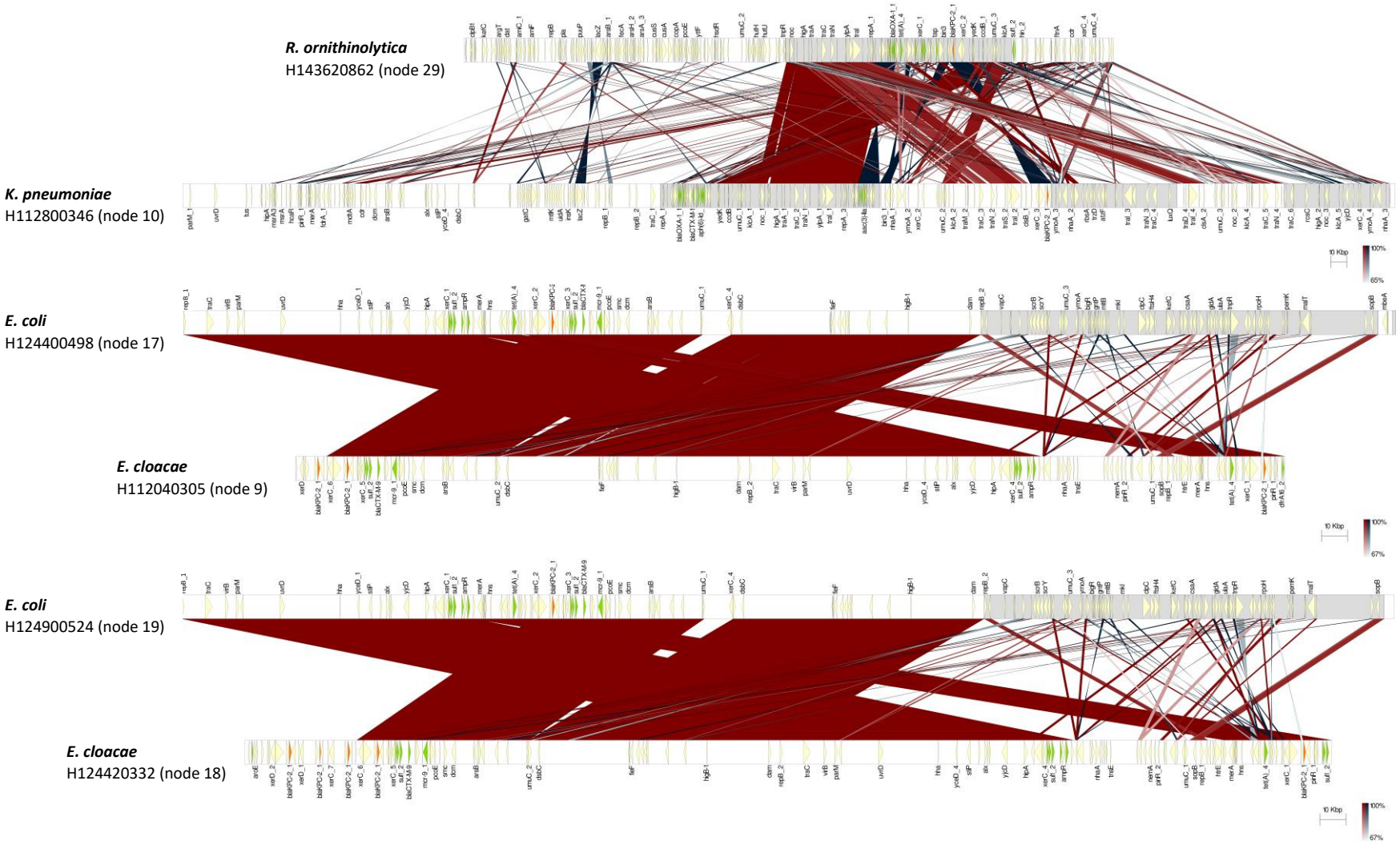


Figure 8. Comparisons of plasmid contigs between sample pairs. Species, sample names, and node labels (from Figure 7) are given on the left of each comparison diagram. Within genome tracks, contigs are delineated using white/grey shading; genes are indicated with arrows (right/left direction=forward/reverse strand; resistance gene=green; *bla*$_{KPC}$ gene resistance gene=orange; other gene=light yellow). Red/blue bands intersecting genome tracks indicate shared sequence; band shading indicates percent identity.

In order to demonstrate the utility of the ATCG best hit filtering procedure, I ran comparative genomic visualisations using original and best hit BLAST alignments generated by ATCG (Supplementary Figure S1). Comparisons were conducted among three plasmid genomes resolved from samples H124200646 and H151860951 (nodes 15 and 34), which had previously been analysed by Decraene et al. using MUMmer (see Fig. 4A in Decraene et al., 2018). The plot showing all BLAST alignments (Supplementary Figure S1B) is crowded with short alignments, which commonly nest within much longer alignments; therefore, the best hit BLAST alignments are likely to be most useful for providing a clear picture of shared genetic content at the whole-genome level.

4.5 Conclusions

Comparative genomic methods, including ANI calculation, form the backbone of many bioinformatic analyses. Indeed, papers outlining the major alignment-based ANI calculation methods are all highly cited: for example, Lee et al. (2016), describing OrthoANI, has >800 citations; Meier-Kolthoff et al. (2013), which is among various papers describing GGDC, has >2000 citations; and Richter et al. (2009), outlining ANIb and ANIm, has >2700 citations. Unfortunately, existing tools have numerous limitations, including inflexible input and comparison options. In the case of BLAST-based methods, there is unrealised potential to assess structural similarity, and store BLAST alignments for downstream comparative genomic visualisation. ATCG addresses all of these limitations. Notably, it is the only BLAST-based method to allow whole-genome (best hit) BLAST alignments to be stored. Moreover, the flexible comparison options implemented by ATCG allow it to integrate well with alignment-free comparative genomic methods, which are increasingly used to handle large datasets; specifically, a subset of genome pairs of interest can be identified using an alignment-free tool, and alignment-based comparative analysis of these pairs can then be easily and efficiently conducted using ATCG with the -p input flag. The benchmarking results demonstrate that ATCG, with the dc-megaBLAST aligner, is an attractive approach, offering favourable runtimes (similar to OrthoANIu),

while also showing strong correlation with OrthoANIb, in terms of calculated ANI values. In this chapter, I demonstrated how ATCG can be applied for comparative genomic analysis of a small set of bacterial plasmid genomes. In Chapter 5, ATCG was used in conjunction with the mash alignment-free comparison tool, to facilitate deduplication of a large NCBI plasmid dataset, by identifying similar plasmids. Besides the applications demonstrated in this thesis, ATCG could also be applied in various other situations, including for genomic taxonomy purposes. In short, ATCG is a novel comparative genomic software tool offering improved functionality compared with existing tools, and further improvements to ATCG, such as a faster runtime, will be implemented in future software versions.

4.6 Supplementary material

4.6.1 Supplementary Methods

Retrieving bacterial genome assemblies for benchmarking ANI software

The following EDirect command was used to retrieve bacterial genome assemblies:

```
cat assemblyaccessions.txt | epost -db assembly -format acc | efetch -format docsum | xtract -pattern DocumentSummary -element FtpPath_GenBank | awk -F"/" '{print $0"/"$NF"_genomic.fna.gz"}' | xargs wget -P fastas
```

Quality-filtering and subsampling of nanopore reads

The following Filtlong command was used to quality-filter and subsample nanopore reads from each sample, prior to hybrid assembly:

```
filtlong -1 illumina_reads1.fq.gz -2 illumina_reads2.fq.gz --min_length 1000 --keep_percent 90 --target_bases 500000000 --trim --split 250 nanopore_reads.fq.gz | gzip > subsampled.fq.gz
```

4.6.2 Supplementary Tables

Table S1. Calculation of similarity and distance statistics from pairwise query-subject alignment data, at different stages of the alignment parsing procedure

Statistic	Formula	Description
Stage A: Statistics calculated from optimal query/subject alignment intervals		
Query genome coverage breadth	$\frac{L_{GQ}}{\lambda(GQ)}$	Proportion of query genome covered by optimal query alignment intervals
Subject genome coverage breadth	$\frac{L_{GS}}{\lambda(GS)}$	Proportion of subject genome covered by optimal subject alignment intervals
Query genome percent identity	$\frac{N_{GQ}}{L_{GQ}}$	Number of identical bases across optimal query genome alignment intervals relative to query genome alignment length
Subject genome percent identity	$\frac{N_{GS}}{L_{GS}}$	Number of identical bases across optimal subject genome alignment intervals relative to subject genome alignment length
Stage B: Statistics calculated from a subset of optimal query/subject alignment intervals from stage A, including only alignments represented in both query and subject intervals		
Breakpoints	NA	Number of cases where an adjacent pair of alignments in one genome is not adjacent in the same relative order in the other genome
Alignments	NA	Total number of alignments per genome
Alignment pairs	NA	For each aligned sequence, alignment pairs = alignments – 1 (assuming linear sequence topology). e.g. given 4 alignments A,B,C,D; the alignment pairs are A,B; B,C; C,D
Breakpoint distance d0	$\frac{breakpoints}{alignment\ pairs}$	If the denominator is 0 (1 alignment), breakpoint distance is assigned 0
Breakpoint distance d1	$\frac{breakpoints}{L/1000}$	Breakpoints expressed per kilobase of aligned sequence
L50	NA	After ordering alignments from large to small, L50 is the number of alignments that must be cumulatively summed to reach 50% of total alignment length
N50	NA	As above, but the size of the alignment at the 50% quantile is given
Stage C: Statistics calculated from reciprocally-trimmed query alignment intervals		
Coverage breadth	$\frac{L}{\lambda(G)/2}$	Total query alignment length relative to mean genome length
Coverage breadth of smaller genome	$\frac{L}{\lambda_{min}(G)}$	Total query alignment length relative to length of the smaller genome

Distance score d0	$1 - \frac{L}{\lambda(G)/2}$	Distance metric of coverage breadth
Distance score d1	$1 - \frac{L}{\lambda_{min}(G)}$	Distance metric of coverage breadth of the smaller genome
Distance score d2	$-log\ \frac{L}{\lambda(G)/2}$	Rescaled variant of distance score d0
Distance score d3	$-log\ \frac{L}{\lambda_{min}(G)}$	Rescaled variant of distance score d1
Average nucleotide identity	$\frac{N}{L}$	Number of identical bases across query alignments relative to total query alignment length
Distance score d4	$1 - \frac{N}{L}$	Distance metric of average nucleotide identity
Distance score d5	$-log\ \frac{N}{L}$	Rescaled variant of distance score d4
Distance score d6	$1 - \frac{N}{\lambda(G)/2}$	Number of identical bases across query alignments relative to mean genome length
Distance score d7	$1 - \frac{N}{\lambda_{min}(G)}$	Number of identical bases across query alignments relative to length of the smaller genome
Distance score d8	$-log\ \frac{N}{\lambda(G)}$	Rescaled variant of distance score d6
Distance score d9	$-log\ \frac{N}{\lambda_{min}(G)}$	Rescaled variant of distance score d7

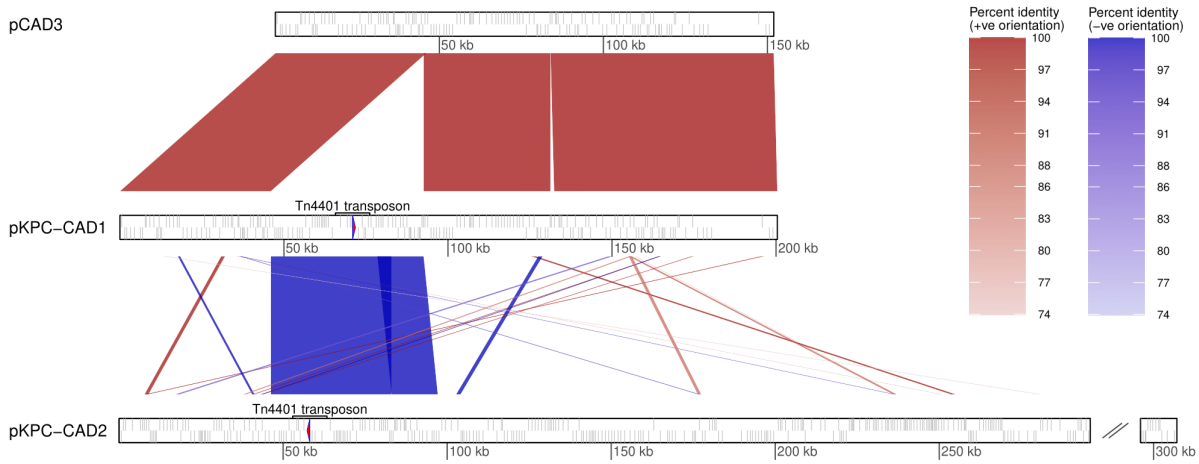
Stage A formulae nomenclature: GQ and GS are the query and subject genomes; λ(GQ) and λ(GS) are the lengths of query and subject genomes, respectively. L_{GQ} and L_{GS} are the total length of optimal alignment intervals covering query and subject genomes, respectively; N_{GQ} and N_{GS} are the total number of identical bases across optimal query and subject alignment intervals.

Stage B and C formulae nomenclature: λ(G)/2 is the mean genome length (averaged across query and subject genomes); $\lambda_{min}(G)$ is the length of the smaller genome of the query-subject pair; L is the total length of reciprocally-trimmed query alignment intervals; N is the total number of identical bases across reciprocally-trimmed query alignment intervals.

N.B If bidirectional BLAST is run, alignment length statistics (L_{GQ}, L_{GS}, L), identical base statistics (N_{GQ}, N_{GS}, N), and structural similarity/alignment contiguity statistics (breakpoints, alignments, alignment pairs, L50, N50) are calculated as a mean of the statistics obtained from both directions.

4.6.3 Supplementary Figures

A Best hit BLAST alignments



B Original (unfiltered) BLAST alignments

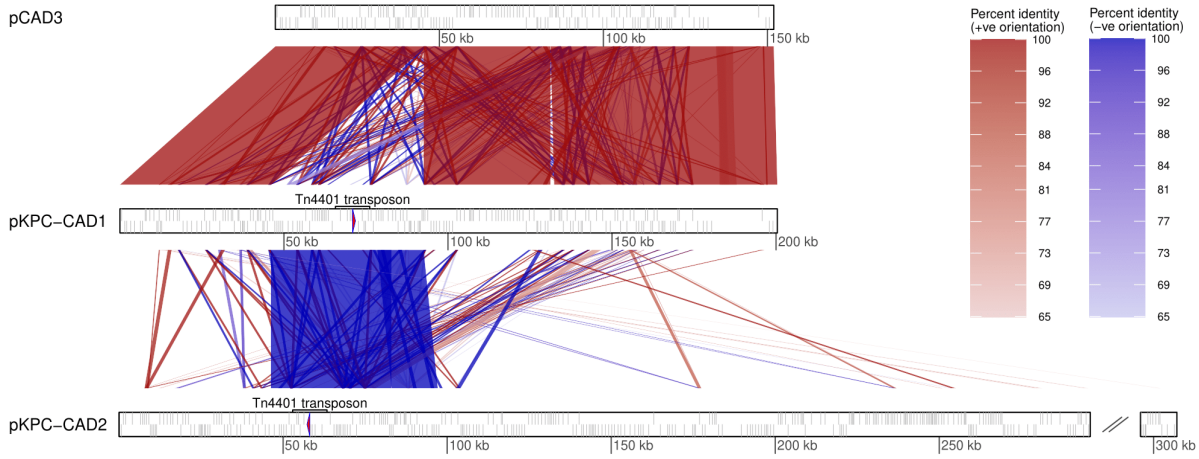


Figure S1. Visualisation of comparisons among three plasmid genomes (pCAD3, pKPC-CAD1, pKPC-CAD2), generated using (A) ATCG best hit BLAST alignments and (B) ATCG original (unfiltered) BLAST alignments. Genomes are represented by rectangular strip(s); pKPC-CAD2 comprises two contigs so is represented as two rectangular strips. The *bla*$_{KPC}$ gene (encoded by pKPC-CAD1 and pKPC-CAD2) is indicated as a triangle (pointing right or left to represent positive/negative strand orientation, respectively). Other annotated genes are indicated as grey lines (offset above and below to represent positive/negative strand orientation, respectively). Alignments are coloured red and blue to reflect forward and reverse complement orientation, respectively. Alignment colour shade and corresponding legends indicate percent identity of forward/reverse complement alignments; colours are transparent but higher scoring alignments are more prominent since they are overlaid on top of lower scoring alignments.

Bibliography

Abudahab, K., Prada, J. M., Yang, Z., Bentley, S. D., Croucher, N. J., Corander, J., et al. (2019). PANINI: Pangenome neighbour identification for bacterial populations. *Microb. Genomics*. doi:10.1099/mgen.0.000220.

Acman, M., van Dorp, L., Santini, J. M., and Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* doi:10.1038/s41467-020-16282-w.

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12, 402. doi:10.1186/1471-2164-12-402.

Altenhoff, A. M., Glover, N. M., and Dessimoz, C. (2019). "Inferring orthology and paralogy," in *Evolutionary Genomics: Statistical and Computational Methods*, ed. M. Animosova (New York, NY: Humana Press), 149–176.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi:10.1093/nar/25.17.3389.

Arahal, D. R. (2014). "Whole-Genome Analyses: Average Nucleotide Identity," in *New Approaches to Prokaryotic Systematics, Volume 41*, eds. M. Goodfellow, I. Sutcliffe, and J. Chun (Academic Press), 103–122.

Argimón, S., and Aanensen, D. M. (2016). Species Mash-up. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro.2016.175.

Armstrong, J., Fiddes, I. T., Diekhans, M., and Paten, B. (2019). Whole-Genome Alignment and Comparative Annotation. *Annu. Rev. Anim. Biosci.* doi:10.1146/annurev-animal-020518-115005.

Auch, A. F., Henz, S. R., Holland, B. R., and Göker, M. (2006). Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. *BMC Bioinformatics* 7, 1–16. doi:10.1186/1471-2105-7-350.

Auch, A. F., Klenk, H. P., and Göker, M. (2010a). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand. Genomic Sci.* 2, 142–148. doi:10.4056/sigs.541628.

Auch, A. F., von Jan, M., Klenk, H. P., and Göker, M. (2010b). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.* 2, 117–134. doi:10.4056/sigs.531120.

Baker, D. N., and Langmead, B. (2019). Dashing: Fast and accurate genomic distances with HyperLogLog. *Genome Biol.* 20. doi:10.1186/s13059-019-1875-0.

Bernt, M., Braband, A., Middendorf, M., Misof, B., Rota-Stabelli, O., and Stadler, P. F. (2013). Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. *Mol. Phylogenet. Evol.* 69, 320–327. doi:10.1016/j.ympev.2012.09.019.

bioperl.org (2019). Tiling HOWTO. Available at: https://bioperl.org/howtos/Tiling_HOWTO.html [Accessed May 13, 2020].

Blanchette, M., Kunisawa, T., and Sankoff, D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.* 49, 193–203. doi:10.1007/PL00006542.

Blom, J., Kreis, J., Spänig, S., Juhre, T., Bertelli, C., Ernst, C., et al. (2016). EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Res.* 44, W22–W28. doi:10.1093/nar/gkw255.

Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Liò, P., and Fani, R. (2008). Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics* 9, 551. doi:10.1186/1471-2105-9-551.

Brockhurst, M. A., Harrison, E., Hall, J. P. J., Richards, T., McNally, A., and MacLean, C. (2019). The Ecology and Evolution of Pangenomes. *Curr. Biol.* doi:10.1016/j.cub.2019.08.012.

Broder, A. Z. (1997). On the resemblance and containment of documents. *Proc. Int. Conf. Compression Complex. Seq.*, 21–29. doi:10.1109/sequen.1997.666900.

Bromham, L. (2016). *An Introduction to Molecular Evolution and Phylogenetics*. 2nd Editio. Oxford: OUP.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*. doi:10.1186/1471-2105-10-421.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., and Parkhill, J. (2005). ACT: The Artemis comparison tool. *Bioinformatics* 21, 3422–3423. doi:10.1093/bioinformatics/bti553.

Chin, C.-S. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi:10.1038/nmeth.2474.

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., et al. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* 68, 2386–2392. doi:10.1099/ijsem.0.002809.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163.

Colston, S. M., Fullmer, M. S., Beka, L., Lamy, B., Peter Gogarten, J., and Graf, J. (2014). Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using aeromonas as a test case. *MBio*. doi:10.1128/mBio.02136-14.

Conlan, S., Park, M., Deming, C., and Thomas, P. J. (2016). Plasmid Dynamics in KPC-Positive Klebsiella pneumoniae during Long-Term Patient Colonization. *MBio* 7, 1–9.

de Been, M., Lanza, V. F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., et al. (2014). Dissemination of Cephalosporin Resistance Genes between Escherichia coli Strains from Farm Animals and Humans by Specific Plasmid Lineages. *PLoS Genet.* 10. doi:10.1371/journal.pgen.1004776.

Decraene, V., Phan, H. T. T., George, R., Wyllie, D. H., Akinremi, O., Aiken, Z., et al. (2018). A large, refractory nosocomial outbreak of klebsiella pneumoniae carbapenemase-producing Escherichia coli demonstrates carbapenemase gene outbreaks involving sink sites require novel approaches to infection control. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01689-18.

Desper, R., and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the Minimum-Evolution principle. *J. Comput. Biol.* doi:10.1089/106652702761034136.

Dewey, C. N. (2019). "Whole-genome alignment," in *Evolutionary Genomics: Statistical and Computational Methods*, ed. M. Animosova (New York, NY: Humana Press), 121–148.

Dicks, J., and Savva, G. (2007). “Comparative genomics,” in *Handbook of Statistical Genetics, Volume 1*, eds. D. J. Balding, M. Bishop, and C. Cannings (Chichester, UK: John Wiley & Sons, Ltd), 179–184.

Edwards, D. J., and Holt, K. E. (2013). Beginner’s guide to comparative bacterial genome analysis using next-generation sequence data. *Microb. Inform. Exp.* 3, 2. doi:10.1186/2042-5783-3-2.

Eyre, D. W., Peto, T. E. A., Crook, D. W., Sarah Walker, A., and Wilcox, M. H. (2020). Hash-based core genome multilocus sequence typing for Clostridium difficile. *J. Clin. Microbiol.* doi:10.1128/JCM.01037-19.

Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M. P., and de la Cruz, F. (2017). Towards a taxonomy of conjugative plasmids. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2017.05.005.

Fricke, W. F., Welch, T. J., McDermott, P. F., Mammel, M. K., LeClerc, J. E., White, D. G., et al. (2009). Comparative genomics of the IncA/C multidrug resistance plasmid family. *J. Bacteriol.* 191, 4750–4757. doi:10.1128/JB.00189-09.

Gabrielaite, M., and Marvig, R. L. (2019). GenAPI: a tool for gene absence-presence identification in fragmented bacterial genome sequences. *bioRxiv*. doi:10.1101/658476.

Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019). PLSDB: A resource of complete bacterial plasmids. *Nucleic Acids Res.* doi:10.1093/nar/gky1050.

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2018). Microbial genome analysis: The COG approach. *Brief. Bioinform.* doi:10.1093/bib/bbx117.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi:10.1099/ijs.0.64483-0.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*. doi:10.1093/bioinformatics/btt086.

Guy, L., Kultima, J. R., Andersson, S. G. E., and Quackenbush, J. (2011). GenoPlotR: comparative gene and genome visualization in R. in *Bioinformatics* doi:10.1093/bioinformatics/btq413.

Hannenhalli, S., and Pevzner, P. A. (1995). Transforming men into mice (polynomial algorithm for genomic distance problem). in *Annual Symposium on Foundations of Computer Science - Proceedings* doi:10.1109/sfcs.1995.492588.

Hartmann, T., Middendorf, M., and Bernt, M. (2018). “Genome Rearrangement Anaysis: Cut and Join Rearrangements and Gene Cluster Preserving Approaches,” in *Comparative genomics: Methods and Protocols*, eds. J. Setubal, J. Stoye, and P. Stadler (New York, NY: Humana Press).

Hayashi Sant’Anna, F., Bach, E., Porto, R. Z., Guella, F., Hayashi Sant’Anna, E., and Passaglia, L. M. P. (2019). Genomic metrics made easy: what to do and where to go in the new era of bacterial taxonomy. *Crit. Rev. Microbiol.* doi:10.1080/1040841X.2019.1569587.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., and Schuster, S. C. (2005). Whole-genome prokaryotic phylogeny. *Bioinformatics* 21, 2329–2335. doi:10.1093/bioinformatics/bth324.

Hilker, R., Sickinger, C., Pedersen, C. N. S., and Stoye, J. (2012). UniMoG-a unifying framework for genomic distance calculation: And sorting based on DCJ. *Bioinformatics* 28, 2509–2511. doi:10.1093/bioinformatics/bts440.

Hosseini, M., Pratas, D., Morgenstern, B., and Pinho, A. J. (2019). Smash++: an alignment-free and memory-efficient tool to find genomic rearrangements. *bioRxiv*, 1–24.

doi:10.1101/2019.12.23.887349.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 1–8. doi:10.1038/s41467-018-07641-9.

Jesus, T. F., Ribeiro-Gonçalves, B., Silva, D. N., Bortolaia, V., Ramirez, M., and Carriço, J. A. (2019). Plasmid ATLAS: Plasmid visual analytics and identification in high-Throughput sequencing data. *Nucleic Acids Res.* doi:10.1093/nar/gky1073.

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Res.* doi:10.12688/wellcomeopenres.14826.1.

Kans, J. (2013). Entrez Direct: E-utilities on the UNIX Command Line. In: Entrez Programming Utilities Help. *Entrez Program. Util. Help*. Available at: http://www.ncbi.nlm.nih.gov/books/NBK179288/ [Accessed November 24, 2016].

Keogh, R. S., Seoighe, C., and Wolfe, K. H. (1998). Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi. *Yeast*. doi:10.1002/(SICI)1097-0061(19980330)14:5<443::AID-YEA243>3.0.CO;2-L.

Knudsen, P. K., Gammelsrud, K. W., Alfsnes, K., Steinbakk, M., Abrahamsen, T. G., Müller, F., et al. (2018). Transfer of a bla CTX-M-1-carrying plasmid between different Escherichia coli strains within the human gut explored by whole genome sequencing analyses. *Sci. Rep.* 8, 1–10. doi:10.1038/s41598-017-18659-2.

Kolaczyk, E. D., and Csardi, G. (2014). *Statistical Analysis of Network Data with R*. New York, NY: Springer.

Koonin, E. V. (2009). Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* doi:10.1016/j.biocel.2008.09.015.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1003118.

Lee, I., Kim, Y. O., Park, S. C., and Chun, J. (2016). OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* 66, 1100–1103. doi:10.1099/ijsem.0.000760.

Leimeister, C. A., Dencker, T., and Morgenstern, B. (2019). Accurate multiple alignment of distantly related genome sequences using filtered spaced word matches as anchor points. *Bioinformatics*. doi:10.1093/bioinformatics/bty592.

Li, X., Huang, Y., and Whitman, W. B. (2015). The relationship of the whole genome sequence identity to DNA hybridization varies between genera of prokaryotes. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* doi:10.1007/s10482-014-0322-1.

Ludden, C., Raven, K. E., Jamrozy, D., Gouliouris, T., Blane, B., Coll, F., et al. (2019). One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *MBio*. doi:10.1128/mBio.02693-18.

Lynch, T., Petkau, A., Knox, N., Graham, M., and Van Domselaar, G. (2016). A Primer on Infectious Disease Bacterial Genomics. *Clin. Microbiol. Rev.* 29, 881–913. doi:10.1128/CMR.00001-16.

Marçais, G., Deblasio, D., Pandey, P., and Kingsford, C. (2019). Locality-sensitive hashing for the edit distance. *Bioinformatics* 35, i127–i135. doi:10.1093/bioinformatics/btz354.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1005944.

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., et al. (2016). Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. *PLoS Genet.* doi:10.1371/journal.pgen.1006280.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-60.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., and Göker, M. (2014a). Highly parallelized inference of large genome-based phylogenies. in *Concurrency Computation Practice and Experience* doi:10.1002/cpe.3112.

Meier-Kolthoff, J. P., Auch, A., Wittmann, J., Klenk, H.-P., and Göker, M. (2014b). GBDP for genome-based phylogenetic analysis and classification of viruses. in *EMBO conference – Viruses of microbes: Structure and function, from molecules to communities.* Available at: http://jan.meier-kolthoff.com/wp-content/uploads/2015/07/GBDP_poster_EMBO_conference_on_viruses_Zurich.pdf.

Meier-Kolthoff, J. P., and Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat. Commun.* 10. doi:10.1038/s41467-019-10210-3.

Moret, B. B. E., Lin, Y., and Tang, J. (2013). Rearrangements in Phylogenetic Inference: Compare, Model, or Encode? *Model. Algorithms Genome Evol. SE - 7* 19, 147–171. doi:10.1007/978-1-4471-5298-9_7.

Morrison, D. A. (2014). Is the tree of life the best metaphor, Model, or heuristic for Phylogenetics? *Syst. Biol.* 63, 628–638. doi:10.1093/sysbio/syu026.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science (80-.).* 347. doi:10.1126/science.1258522.

Noll, N., Urich, E., Wüthrich, D., Hinic, V., Egli, A., and Neher, R. (2018). Resolving structural diversity of Carbapenemase-producing gram-negative bacteria using single molecule sequencing. *bioRxiv*, 456897. doi:10.1101/456897.

Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., et al. (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biol.* doi:10.1186/s13059-019-1841-x.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* doi:10.1186/s13059-016-0997-x.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. doi:10.1093/bioinformatics/btv421.

Palmer, M., Steenkamp, E. T., Blom, J., Hedlund, B. P., and Venter, S. N. (2020). All anis are not created equal: Implications for prokaryotic species boundaries and integration of anis into polyphasic taxonomy. *Int. J. Syst. Evol. Microbiol.* doi:10.1099/ijsem.0.004124.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R

language. *Bioinformatics*. doi:10.1093/bioinformatics/btg412.

Pevsner, J. (2015). "Chapter 17 Completed Genomes: Bacteria and Archaea," in *Bioinformatics and Functional Genomics* (John Wiley & Sons), 833–834.

Phillippy, A., and Ondov, B. (2017). Mash Screen: what's in my sequencing run? Available at: https://genomeinformatics.github.io/mash-screen/.

Poptsova, M. S., and Gogarten, J. P. (2010). Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology*. doi:10.1099/mic.0.033811-0.

Pritchard, L. (2014). Whole Genome Alignments [github tutorial]. Available at: https://github.com/widdowquinn/Teaching/blob/master/Comparative_Genomics_and_Visualisation/Part_1/whole_genome_alignment/whole_genome_alignments_A.md [Accessed June 10, 2020].

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., and Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. *Anal. Methods*. doi:10.1039/c5ay02550h.

Pritchard, L., White, J. A., Birch, P. R. J., and Toth, I. K. (2006). GenomeDiagram: A python package for the visualization of large-scale genomic data. *Bioinformatics*. doi:10.1093/bioinformatics/btk021.

Raina, V., Nayak, T., Ray, L., Kumari, K., and Suar, M. (2019). "Chapter 9 - A Polyphasic Taxonomic Approach for Designation and Description of Novel Microbial Species," in *Microbial Diversity in the Genomic Era*, eds. D. Surajit and H. R. Dash (London, UK: Academic Press), 137–152.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. (2016). JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32, 929–931. doi:10.1093/bioinformatics/btv681.

Rocha, E. P. C. (2006). Inference and analysis of the relative stability of bacterial chromosomes. *Mol. Biol. Evol.* 23, 513–522. doi:10.1093/molbev/msj052.

Rosselló-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. in *FEMS Microbiology Reviews* doi:10.1016/S0168-6445(00)00040-1.

Sankoff, D., and Blanchette, M. (1998). Breakpoint Phylogeny. *J. Comput. Biol.* 5, 555–570.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C., and Turmel, M. (2000). Chloroplast Gene Order and the Divergence of Plants and Algae, from the Normalized Number of Induced Breakpoints. 89–98. doi:10.1007/978-94-011-4309-7_10.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. doi:10.1093/bioinformatics/btu153.

Setubal, J. C., Almeida, N. F., and Wattam, A. R. (2018). "Comparative Genomics for Prokaryotes," in *Comparative genomics: Methods and Protocols*, eds. J. C. Setubal, J. Stoye, and P. F. Stadler (New York, NY: Humana Press).

Snel, B., Huynen, M. A., and Dutilh, B. E. (2005). GENOME TREES AND THE NATURE OF GENOME EVOLUTION. *Annu. Rev. Microbiol.* doi:10.1146/annurev.micro.59.030804.121233.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* doi:10.1101/gr.361602.

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: A genome comparison visualizer. *Bioinformatics*. doi:10.1093/bioinformatics/btr039.

Suzuki, M., Doi, Y., and Arakawa, Y. (2020). Plasmid ORF-based binarized structure network analysis of plasmids (OSNAp), a novel approach to core gene-independent plasmid phylogeny. *Plasmid* 108, 102477. doi:10.1016/j.plasmid.2019.102477.

Tripp, H. J., Sutton, G., White, O., Wortman, J., Pati, A., Mikhailova, N., et al. (2015). Toward a standard in structural genome annotation for prokaryotes. *Stand. Genomic Sci.* doi:10.1186/s40793-015-0034-9.

Uchiyama, I., Albritton, J., Fukuyo, M., Kojima, K. K., Yahara, K., and Kobayashi, I. (2016). A Novel Approach to Helicobacter pylori Pan-Genome analysis for Identification of genomic Islands. *PLoS One*. doi:10.1371/journal.pone.0159419.

UKRI (2018). General Data Protection Regulation guidance for researchers. *UK Res. Innov. News*. Available at: https://www.ukri.org/news/general-data-protection-regulation-guidance-for-researchers/ [Accessed March 3, 2020].

Wang, X., Liu, D., Luo, Y., Zhao, L., Liu, Z., Chou, M., et al. (2018). Comparative analysis of rhizobial chromosomes and plasmids to estimate their evolutionary relationships. *Plasmid*. doi:10.1016/j.plasmid.2018.03.001.

Weingarten, R. A., Johnson, R. C., Conlan, S., Ramsburg, A. M., Dekker, J. P., Lau, A. F., et al. (2018). Genomic analysis of hospital plumbing reveals diverse reservoir of bacterial plasmids conferring carbapenem resistance. *MBio* 9, 1–16. doi:10.1128/mBio.02011-17.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1005595.

Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* doi:10.1186/s13059-019-1727-y.

Wickham, H. (2011a). ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* doi:10.1002/wics.147.

Wickham, H. (2011b). The split-apply-combine strategy for data analysis. *J. Stat. Softw.* doi:10.18637/jss.v040.i01.

Wolf, Y. I., and Koonin, E. V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* doi:10.1093/gbe/evs100.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* doi:10.1186/gb-2014-15-3-r46.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21, 3340–3346. doi:10.1093/bioinformatics/bti535.

Yoon, S. H., Ha, S. min, Lim, J., Kwon, S., and Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* 110, 1281–1286. doi:10.1007/s10482-017-0844-4.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–

2644. doi:10.1093/jac/dks261.

Zhang, D., Zhao, Y., Feng, J., Hu, L., Jiang, X., Zhan, Z., et al. (2019). Replicon-based typing of IncI-complex plasmids, and comparative genomics analysis of IncIγ/K1 plasmids. *Front. Microbiol.* doi:10.3389/fmicb.2019.00048.

Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* 18, 1–17. doi:10.1186/s13059-017-1319-7.

5 Investigating biological and abiotic factors associated with plasmid antibiotic resistance gene carriage using a curated dataset of NCBI plasmid sequences and metadata

5.1 Summary

A key question in plasmid epidemiology is whether there are particular plasmid genetic traits or abiotic characteristics that are associated with antibiotic resistance gene carriage. NCBI Nucleotide and BioSample databases represent a rich source of data with which to investigate such associations. Few existing studies of NCBI plasmids incorporate statistical analysis of linked BioSample metadata; a major bottleneck lies in the need to curate metadata to facilitate downstream analyses. In this chapter, a curated dataset of plasmids and associated metadata was first compiled from NCBI. During plasmid curation, non-natural (e.g. laboratory) plasmid sequences were excluded. In addition, to address the statistical problem of 'pseudoreplication', similar plasmid sequences, sharing certain items of metadata (species/location/collection date), were deduplicated. During curation of BioSample metadata, the issue of erroneous metadata was explored by 1) examining internal consistency between latitude/longitude and geographic location name; 2) externally validating metadata for culture collection samples using the Bac*Dive* metadata database. Notably, of 236 culture collection samples with linked Bac*Dive* metadata, there were nine cases where species authority date (i.e. date of taxonomic description) was given instead of collection date, leading to spuriously early collection dates (as early as 1884). The deduplicated set of curated plasmids (14143 plasmids) and associated curated metadata was used to fit logistic Generalised Additive Models (GAMs). Genetic factors (e.g. presence of integrons, virulence genes, and biocide/metal resistance genes) and abiotic factors (e.g. isolation source, geography, collection date) were used as explanatory variables, while antibiotic resistance carriage was modelled as a binary outcome. Separate GAMs were built to model 10 major resistance class outcomes. Overall, GAM modelling results were consistent with antibiotic

selection pressure as a driver of plasmid resistance carriage. Notably, resistance patterns in different isolation sources (human/livestock), across time (collection years), and across different host bacterial taxa reflected reported antibiotic usage patterns (determined from literature reports, e.g. clinical/veterinary antibiotic sales data, for each antibiotic class). In addition, for a given antibiotic resistance class, integrons, biocide/metal resistance genes, and resistance genes of other classes were generally positively associated with antibiotic resistance carriage. The size of the association was particularly pronounced for sulphonamide resistance carriage. Sulphonamide antibiotics were widely used historically, but clinical usage is now restricted. Therefore, co-linkage of sulphonamide resistance genes with other adaptive genes (on the same plasmid/within integrons), provides a plausible explanation for the persistence of sulphonamide resistance – through co-selection – despite reduced direct selection.

5.2 Introduction

Given the critical role of plasmids as vectors of antibiotic resistance, an important knowledge gap is whether particular characteristics are more likely to be associated with antibiotic resistance plasmids, relative to plasmids not harbouring resistance. Relevant characteristics might include plasmid genetic factors, as well as higher-level characteristics such as isolation source (since antibiotic usage may differ between isolation sources) (Harbarth and Samore, 2005). An understanding of the processes determining acquisition, maintenance, and spread of plasmid-mediated resistance can be used to infer factors that might plausibly be associated with resistance carriage (Martínez et al., 2007), as described below.

Plasmid resistance gene acquisition from co-resident plasmids or the host chromosome, is commonly facilitated by nested mobile genetic elements, such as transposons and integron-associated mobile gene cassettes (as described in Chapter 1). The role of integrons and transposons in promoting gene mobility has been elucidated in the laboratory (Lartigue et al., 2006; Snyder et al., 2013). Comparative

genetic analyses of resistance genes and their genetic contexts further support the central role of transposons and integrons in plasmid resistance gene acquisition (Davies and Davies, 2010; Jacoby et al., 2011; Potron et al., 2011; Forsberg et al., 2012; D'Andrea et al., 2013; Ghaly et al., 2017). If a particular transposable element or integron is present in multiple copies within a cell, homologous recombination can promote intra-cellular dissemination of resistance genes between different replicons (Stokes and Gillings, 2011). Finally, horizontal plasmid transfer (e.g. conjugation) brings together different replicons, allowing more opportunities for resistance gene acquisition by the mechanisms described above (Baquero and Canton, 2017).

Following resistance gene acquisition, the success of a resistance plasmid depends on vertical stability and/or horizontal transmission (Cottell et al., 2014; Martínez and Baquero, 2014). Under antibiotic selection pressure, a resistance plasmid may confer a benefit on its host, and therefore, disseminate vertically through clonal expansion (Ghaly and Gillings, 2018). However, antibiotic selection pressure may be weak or transient, even in clinical settings, in which case resistance plasmids may impose fitness costs (Vogwill and Maclean, 2015; San Millan and MacLean, 2017; San Millan et al., 2018). In the absence of strong antibiotic selection pressure, various mechanisms promote resistance plasmid persistence, including conjugative transfer and compensatory mutations, as demonstrated by experimental studies (Porse et al., 2016; Lopatkin et al., 2017; Stevenson et al., 2017; San Millan, 2018). Conjugation may enable plasmids to persist as infectious genetic parasites, while plasmid and/or chromosomal compensatory mutations alleviate fitness costs. Transposable elements – i.e. insertion sequences and transposons – are commonly implicated in genome rearrangement, deletion, and insertional inactivation events, which may be compensatory (Porse et al., 2016; Vandecraen et al., 2017). Finally, co-selection between different adaptive genes (e.g. 'co-resistance' between multiple resistance genes) (Baker-Austin et al., 2006) can also promote resistance plasmid persistence. Specifically, at a given timepoint, while one resistance gene may not be actively under direct selection, it may 'hitchhike' with a genetically linked gene that is under active selection (Hughes and Andersson, 2017; Goswami et al., 2020).

The biological processes described above have generally been studied through experiments and observational genetic analyses, with individual studies based on a restricted number of isolates. It is unclear how these processes play out in nature on a global scale to shape the emergence and spread of resistance plasmids. For example, laboratory strains may not represent natural strains (San Millan, 2018). Furthermore, besides various deterministic plasmid factors, stochasticity (e.g. founder effects) and host cell genetic factors may also influence resistance plasmid success (Porse et al., 2016; San Millan, 2018). Therefore, studies of plasmids from a small number of isolates do not provide general insights into the role of plasmid factors, especially in situations where non-plasmid factors are the major drivers of plasmid success. Lastly, potentially relevant higher-level abiotic factors such as socioeconomic conditions, antibiotic stewardship, and isolation source (e.g. human/livestock) are generally not represented in small-scale studies.

As an alternative to small-scale mechanistic studies, large datasets encompassing a broad geographic scope can be used to assess patterns of association between plasmid resistance carriage and various biological and abiotic factors. This approach complements mechanistic studies since factors associated with resistance carriage can be interpreted in the light of mechanistic understanding. For example, analyses of the association between prevalence of antibiotic resistance among bacterial isolates and abiotic factors have identified higher antibiotic usage, as well as socioeconomic factors (e.g. low-income, poor governance/infrastructure) as associated factors (Bell et al., 2014; Collignon et al., 2015, 2018; ECDC/EFSA/EMA, 2017). The role of socioeconomic factors was interpreted in terms of issues such as poor sanitation and water supply contamination, which could enable resistance spread (Collignon and Beggs, 2019). However, studies focusing specifically on plasmids and/or incorporating both genetic and higher-level abiotic factors are lacking. The association between antibiotic resistance gene classes and plasmid replicon types has been investigated descriptively, by literature review of Enterobacteriaceae resistance plasmids (Rozwandowicz et al., 2018), and analysis of resistance plasmids in NCBI (Orlek et al., 2017). However, without incorporating other possible contributing factors into a multivariate statistical analysis of plasmid resistance carriage, the importance of replicon

type cannot be robustly assessed, due to confounding bias. While replicon type is a commonly studied plasmid trait, it remains unclear whether it is informative in the context of resistance carriage. NCBI databases represent a rich source of plasmid sequences and metadata for multivariate analysis of plasmid resistance carriage, but analysis of this data is challenging.

A drawback of analysing NCBI plasmid datasets is that sequencing efforts to date have been biased towards plasmids from culturable bacteria, isolated from anthropogenic (e.g. clinical and livestock) settings. Consequently, statistical inferences should be interpreted cautiously; results of analyses may not be generalisable to plasmids from taxa or isolation sources which have yet to be sampled, and there may be insufficient power to detect influential effects of less frequently sampled plasmid characteristics. A second drawback of analysing NCBI plasmid datasets is that accessions do not necessarily represent independent sampling units. The NCBI nucleotide database comprises GenBank and RefSeq source databases; since primary GenBank accessions are propagated to RefSeq (a more curated and non-redundant database), redundancy is inevitable if retrieving from the complete NCBI nucleotide database. Redundancy also occurs in the RefSeq database since curation is imperfect: identical or similar sequence data may be submitted by different researchers, and revised submissions may occur without removal of older submissions (Chen et al., 2017). Aside from straightforward redundancies, non-independent samples can arise when bacteria are densely sampled in time and space during a sequencing project, leading to biased sampling of identical or similar plasmid sequences. If non-independence is not minimised, analyses will suffer from 'pseudoreplication', which increases the risk of biased inference and false-positive results (Colegrave and Ruxton, 2018).

Current plasmid curation methods have applied arbitrary sequence similarity thresholds (e.g. 100% identity/coverage breadth) to exclude duplicates (Orlek et al., 2017; Brandt et al., 2019; Galata et al., 2019; Douarre et al., 2020). However, non-identical but related sequences may be collected during a single sequencing project (e.g. evolution of plasmid sequences during a hospital outbreak study). Therefore, a more principled approach for minimising pseudoreplication might incorporate sample

metadata, with the aim of delimiting independent sampling units (Chen et al., 2017). For example, after retrieving bacterial metagenomes from public sequence repositories, Fondi et al. deduplicated samples according to habitat and geographic location (using a 20 km distance threshold) (Fondi et al., 2016).

NCBI nucleotide accessions are often linked to sample metadata from the NCBI BioSample database (e.g. isolation source, collection date) (Barrett et al., 2012). However, studies of curated plasmid datasets have so far not capitalised on the variety of available linked sample metadata such as isolation source, collection date, and geographic location. Unfortunately, the use of NCBI BioSample metadata is impeded by issues of low quality and missing data (Gonçalves and Musen, 2019; Douarre et al., 2020). The proliferation of uncontrolled user-defined field names complicates metadata retrieval. For NCBI-defined field names, values are not always valid (e.g. recognised ontology terms are not always used) (Gonçalves and Musen, 2019). More insidiously, even if the data conform to required formats or ontologies, values may be erroneous. For example, Douarre et al. remark that the earliest plasmid in their curated dataset was collected in 1884 (accession NZ_CP020021.1; *Staphylococcus aureus* strain ATCC 6538); this is highly suspect considering that the earliest public culture collection was established in 1890 (Uruburu, 2003; Douarre et al., 2020).

In this chapter, I first compiled a nucleotide sequence dataset of curated bacterial plasmids from the NCBI Nucleotide database. In addition, linked BioSample metadata (including geographic location, isolation source, and collection date) was curated with manual oversight, and erroneous metadata was identified and corrected. I revealed previously unreported systematic errors in the NCBI BioSample database, attributed to poor reporting of bacterial culture collection samples (e.g. providing date of species taxonomic description from the culture collection record, instead of the sample collection date). The curated sample metadata was also used to deduplicate similar plasmid sequences in order to reduce pseudoreplication. Using the curated deduplicated dataset, multivariate models were constructed for major classes of antibiotic resistance, using biological and abiotic factors

as explanatory variables, and plasmid resistance carriage as a binary outcome variable. By interpreting associations using biological and clinical knowledge, I was able to make interesting inferences about potential drivers of plasmid antibiotic resistance gene carriage.

5.3 Methods

5.3.1 Retrieval and curation of NCBI bacterial plasmids, and deduplication of identical sequences

An initial curated dataset of complete bacterial plasmids was compiled from the NCBI nucleotide database on 1st May 2019, using the bacterialBercow pipeline (v0.1.0, https://github.com/AlexOrlek/bacterialBercow; options: --taxonomyquery bacteria[porgn:__txid2], --dbsource refseq_genbank). The PlasmidFinder and rMLST databases used with bacterialBercow were downloaded on 9th Apr 2019. During bacterialBercow curation, where there were identical plasmid sequences, at least one accession was retained; other sequences were stringently excluded as duplicates if they shared metadata with a retained accession, or were missing metadata. Specifically, automated deduplication using bacterialBercow was based on the following metadata fields: BioSample accession id, primary BioProject accession id (the cognate Genbank BioProject accession id in the case of RefSeq accessions), submitter contact name, submitter affiliation name (option: --deduplicationmethod both). Manual oversight was used to guide further deduplication of remaining (potentially non-duplicate) identical sequences. To gain additional insight into duplication among accessions, and to validate the automated deduplication strategy, a subset of accession clusters which had been deduplicated automatically, were selected for retrospective manual investigation. Only clusters including accessions with differing BioSample and GenBank BioProject accession ids were selected – according to NCBI documentation, BioSample and BioProject accessions should be used to delineate different isolates and sequencing projects, respectively (Barrett et al., 2012). Further

curation and deduplication of nucleotide accessions took place after retrieval of BioSample metadata (see below).

5.3.2 Curation of plasmid accession metadata

Additional accession metadata was retrieved from NCBI using Edirect (Kans, 2013); the code has been packaged as a GitHub repository (https://github.com/AlexOrlek/getNCBImetadata). Source taxon name and source taxonomic id ('taxid') were retrieved from NCBI nucleotide. The ete3 Python package (Huerta-Cepas et al., 2016) was used to derive hierarchical taxonomy data (phylum through to species) from NCBI taxids (taxids are stable unique identifiers, which should therefore be robust to nomenclature change) (Federhen, 2012). For nucleotide accessions with linked BioSample accession ids, metadata was also retrieved from NCBI BioSample. Specifically, BioSample attribute metadata was retrieved for canonical attribute names, including *collection_date*, *isolation_source*, *geo_loc_name*, *lat_lon*, *culture_collection* (for a full list see Supplementary Methods). Curated geographic location metadata was generated from geographic location names (*geo_loc_name*) in conjunction with latitude/longitude coordinates (*lat_lon*). Missing/invalid locations (e.g. "-", "n/a", "not determined") were excluded, while valid latitude/longitude coordinates were extracted using a 'regex' (a regular expression) (Supplementary Methods). To address cases where a location name was present but geospatial coordinates were missing, location names were geocoded using the Google Places API (https://developers.google.com/places/web-service/overview), accessed via the googleway R package (https://github.com/SymbolixAU/googleway). To ensure high-accuracy geocoding of all location names, a combination of automated and manual geocoding was employed (see Supplementary Methods and Appendix D: File 1).

Curation of the geocoded locations (determined for each unique location name) was conducted on a per-BioSample accession basis. For a given BioSample accession, geocoded latitude/longitude was compared with corresponding BioSample coordinates (*lat_lon* attribute) where available, in order to

identify discrepancies. Geodesic distances between coordinates were calculated using the geopy Python package (https://pypi.org/project/geopy/). In addition, "viewport" bounds of geocoded place predictions were determined (the viewport is the map region displayed for a given place) (Google Maps Platform, 2020b). A discrepancy was deemed to occur if the following logical conjunction was true: 1) coordinates were separated by >50 km *and* 2) the BioSample coordinate fell outside the viewport of the geocoded place prediction (where available – manually geocoded locations lacked Google place prediction data). Discrepancies were resolved manually. For downstream statistical analysis (see below), curated latitude/longitude coordinates were reverse geocoded at the level of country (googleway argument: result_type="country") and higher-level groupings: European Union (EU) countries (including the UK); and income groupings, defined according to the World Bank Atlas Method (2018 version) (The World Bank, 2018).

Following geocoding, collection date and host/isolation source metadata were curated. Validly formatted collection dates (*collection_date* attribute) were extracted using regexes (Supplementary Methods; Appendix D: File 1). Very early collection dates (arbitrarily, pre-1950) were manually checked. Isolation source metadata was identified using the following BioSample attribute fields: *host*, *isolation_source*, *host_description*, *env_broad_scale*, *env_local_scale*, *env_medium*, *description*. Regexes were used to identify human and livestock samples (note that 'sample' is used interchangeably with 'BioSample accession' when discussing plasmid dataset curation). The following globally important livestock species were searched for: cattle, pigs, chicken, sheep, goats, ducks, turkeys (Alvarez-Kalverkamp et al., 2014; Gilbert et al., 2018) (Supplementary Methods; Appendix D: File 1). To curate the initially identified host categories, manual oversight was used to exclude spurious matches. Livestock sample isolation sources of interest included live animal swabs, blood and faecal samples, and raw egg. Retail meat isolates were included by default, but if metadata suggested a processed animal product (e.g. dairy product, animal hide, processed meat such as ham), samples were excluded. This is because bacterial composition is likely to be substantially changed by food processing and is presumably less likely to reflect farm/livestock bacterial composition. Besides human

and livestock categories, additional isolation sources – aquaculture, agriculture (non-livestock), sewage (restricting to human wastewater where possible) – were also identified using regexes and manual curation (Supplementary Methods; Appendix D: File 1).

5.3.3 Curation of culture collection samples: identification, metadata curation, and deduplication

Culture collection samples were considered to warrant particular attention for several reasons. Firstly, a given culture collection sample may be sequenced by many independent research groups, which could produce duplicate plasmid sequences. Deduplication is not straightforward since the same isolate is commonly shared across multiple collections and may therefore be labelled with different but synonymous culture collection identifiers. Secondly, culture collection metadata may be especially prone to errors. For example, in lieu of the original collection date and source geographic location, submitters may provide the date on which they acquired the culture specimen and the location of their research institute. To address these problems, culture collection samples were identified, enabling subsequent metadata curation and deduplication. Firstly, culture collection acronyms (e.g. ATCC) were compiled and incorporated into a regex which matched culture collection identifiers in the *culture_collection* BioSample attribute field, which is designated for culture collection identifiers (NCBI, 2020a). To accommodate samples where culture collection was only indicated in a non-designated field, other potentially relevant fields such as *strain* and *sample_name* were searched (for details see Supplementary Methods).

The Bac*Dive* database contains aggregated bacterial strain metadata, derived and curated from primary culture collection databases (Reimer et al., 2017, 2019). As a result, Bac*Dive* should be a reliable source of metadata, which can be used to externally validate NCBI BioSample metadata. Metadata on isolation source, geographic location, and culture collection synonyms was downloaded from Bac*Dive* (https://bacdive.dsmz.de/isolation-sources; accessed 4th Jul 2019). In addition, available

collection date metadata (not currently published in Bac*Dive*) was kindly provided by Dr Lorenz Reimer. Bac*Dive* metadata was used to guide manual curation of BioSample metadata (host, isolation source, geographic location, latitude/longitude, collection date); additions and corrections were implemented in accordance with Bac*Dive* metadata. Furthermore, culture collection synonym information from *BacDive* was used to facilitate manual deduplication of culture collection samples sharing synonymous culture collection identifiers. Deduplication of samples with identical/synonymous culture collection identifiers favoured retention of samples with higher assembly contiguity. Exclusions were propagated to the nucleotide level: plasmids linked to excluded BioSample accessions were also excluded.

5.3.4 Exclusion of non-natural plasmid accessions

BioSample accessions considered to represent non-natural samples were removed with the intention that downstream analyses would reflect natural plasmids. Samples from model laboratory organisms, as well as samples representing commercial strains (e.g. biopesticides) were excluded. Specifically, a list of animal and plant laboratory model organisms was compiled and used to search the *host* and *lab_host* attribute fields. In addition, "lab" and "laboratory" were used as search terms (Supplementary Methods). Exclusions were guided by manual oversight to prevent false-positive exclusion of natural samples. Finally, all BioSample fields were manually searched with the following case-insensitive search terms to detect and exclude commercial isolates: "commercial", "biocontrol", "cide" (as a suffix e.g. biocide). As with exclusion of culture collection duplicates, exclusions were propagated to the level of plasmid nucleotide accessions.

Further curation was also conducted at the nucleotide accession level, with the aim of excluding vector contaminant sequences (e.g. plasmid or phage cloning vectors) (Schäffer et al., 2018). Plasmids were queried against the UniVec database using megaBLAST, with 95% percent identity and 50% query coverage thresholds (parameters: evalue=1e-8, max_target_seqs=10000) (NCBI, 2020b). Plasmids

matching a UniVec sequence were excluded as putative contaminants. In addition, plasmids linked to the same BioSample accession as an excluded plasmid were also excluded (regardless of similarity to UniVec sequences).

5.3.5 Deduplicating plasmids based on sequence similarity and metadata sharing

Besides deduplicating identical plasmid sequences (Section 5.3.1) and plasmids from duplicate culture collection samples (Section 5.3.3), a further deduplication step was applied to plasmids prior to downstream statistical analyses (see Section 5.3.7). The intention of this final deduplication step was to deduplicate plasmids with similar sequences, except where metadata (country, collection date, species, BioSample accession, primary BioProject accession) indicated that the similar plasmids represented independent sampling units. Consequently, pseudoreplication should be reduced in downstream analyses. Similar plasmids were identified using an initial all-vs-all mash (mash distance <0.1) (Ondov et al., 2016) followed by pairwise BLASTN comparisons with ATCG (>95% ANI and >50% mean coverage breadth across a plasmid pair) (v0.1.0, https://github.com/AlexOrlek/ATCG). Pairwise links between similar plasmids were filtered according to metadata, with the aim of retaining links between duplicates. For example, links between plasmids associated with different species were excluded, while links were retained if country and collection date metadata was identical (see Supplementary Methods for details). Using the igraph R package, a network was constructed from retained links (edge-weighted by pairwise ANI), and clusters were detected using the infomap algorithm (Fortunato and Hric, 2016). A single representative was selected from each cluster, favouring selection of plasmids linked to more populated BioSample metadata (collection date, isolation source, geographic location). The final deduplicated plasmid dataset therefore included the deduplicated cluster representatives, as well as plasmids that did not show sufficient sequence similarity to any other plasmids.

5.3.6 Annotation of plasmid sequences

Sequence annotation was performed on the curated plasmid dataset, prior to the final deduplication step described above. To detect antibiotic resistance genes, the complete ResFinder database (retrieved 25th September 2019) was queried using percent identity and coverage thresholds of 90% and 60%, respectively (Zankari et al., 2012) (BLASTN parameter: max_target_seqs=5000). Using the ResFinder phenotypes text file (which maps resistance genes to resistance phenotypes), detected beta-lactam resistance genes were divided into three sub-classes: carbapenem resistance, extended-spectrum beta-lactamase (ESBL) resistance, other beta-lactam resistance. Specifically, carbapenem resistance genes were defined as those predicted to confer resistance to at least one carbapenem (ertapenem, meropenem, or imipenem); ESBLs were defined as genes not conferring carbapenem resistance but conferring resistance to aztreonam and at least one third-generation cephalosporin (cefotaxime, ceftazidime and ceftriaxone) (Ghafourian et al., 2014). Antibacterial biocide and metal resistance genes were detected by BLASTX querying plasmids against experimentally validated proteins in the BacMet database (retrieved 30th September 2019) (Pal et al., 2014); BLASTX parameters (evalue=0.001, max_target_seqs=1000), and best hit filtering thresholds (90% identity, 30 bp alignment length) were selected based on previous methods (Bengtsson-Palme et al., 2016; Ma et al., 2016). Replicon typing was conducted by BLASTN querying plasmids against the PlasmidFinder Enterobacteriaceae and Gram-positive databases (retrieved 25th September 2019). Hits were filtered using a percent identity threshold of 80% and a replicon sequence coverage threshold of 60% (Carattoli and Hasman, 2020). Detection of conjugative systems was conducted using the CONJscan module of MacSyFinder (parameter: *--replicon-topology* circular) (Cury et al., 2020). Firstly, plasmid protein-coding genes were annotated with Prodigal (Hyatt et al., 2010), using "anonymous mode" for smaller plasmids and "normal mode" for larger plasmids (≥100 kb). Plasmid proteomes were searched using the profiles defined previously (Cury et al., 2020), with an additional profile for the novel MOBM relaxase, downloaded from MOBfamDB (Garcillan-Barcia et al., 2020). Complete conjugative systems were defined according to previously described gene content and inter-gene distance stipulations

(Cury et al., 2017). Virulence genes were detected by BLASTX querying plasmids against experimentally validated genes of the virulence factor database (VFDB) (retrieved 15th October 2019) (Chen et al., 2016). BLASTX parameters (evalue=1e-5, max_target_seqs=5000), and best hit filtering thresholds (75% identity, 50% coverage) were selected in accordance with previous methods (Hannigan et al., 2015; Huang et al., 2018). Integrons were detected using IntegronFinder (version 2.0), run with default parameters, except for *--local-max* (for increased sensitivity) and *--circ* (to set replicon topology as circular) (Cury et al., 2016). Insertion sequences were detected using ISEScan (v1.7.1) with default settings (Xie and Tang, 2017).

5.3.7 Statistical analysis of factors associated with plasmid antibiotic resistance carriage

Prior to the final deduplication step (Section 5.3.5), the epidemiology of resistance plasmids was visualised (in time and space) using Microreact (Argimón et al., 2016). All subsequent analyses were conducted on deduplicated plasmids using R (https://www.r-project.org/) (version 3.6.2). Logistic Generalised Additive Models (GAMs), implemented using the mgcv package (Wood, 2001), were used to investigate the association between explanatory variables and plasmid resistance carriage, modelled as a binary outcome variable. Logistic GAMs are extensions of standard logistic regression models. Unlike linear models, GAMs model numeric explanatory variables using nonlinear smooths constructed from 'basis functions' (Wood, 2017b).

Separate GAMs were built for the following 10 resistance (sub-)classes: carbapenem, ESBL, and other beta-lactam resistance (i.e. three beta-lactam sub-classes), aminoglycoside, sulphonamide, tetracycline, trimethoprim, phenicol, macrolide, and quinolone resistance. Explanatory variables, outlined in Table 1, were selected based on *a priori* biological knowledge (see Introduction). Categorical factor levels with low frequency were collapsed in order to balance the bias-variance trade-off (see Table 1). Missing collection dates were singly imputed by using the accession 'create date' (Nahin, 2008) as a proxy for collection date.

Table 1. Explanatory variables used in GAM models built using deduplicated plasmid datasets; categorical variable factor levels and pre-modelling transformations of numeric variables are described.

Explanatory variables	Notes
Numeric variables	
Plasmid size (kb)	*Baseline: 10 kb* Prior to modelling, the left and right tails were winsorised at the 5th and 95th percentiles (i.e. truncated back at the tails); the 5th and 95th percentiles corresponded to 2.9 kb and 311.7 kb, respectively. Then, values were $\log_{10}$ transformed, and centred on 10 kb (i.e. $\log_{10}$ - 4).
Insertion sequence density	*Baseline: 0* An insertion sequence density score was calculated as the number of insertion sequences per 10 kb. Prior to modelling, the right-tail was winsorised at the 95th percentile. The 95th percentile corresponded to 1.05 insertion sequences per 10 kb.
Number of other resistance gene classes	*Baseline: 0* Prior to modelling, the right-tail was winsorised at the 95th percentile. The 95th percentile corresponded to four other resistance gene classes for aminoglycoside and sulphonamide models, and five other resistance gene classes for all other models.
Collection date (years since initial collection year)	*Baseline: 1994* Missing collection years were singly imputed using accession create date. Prior to modelling, the left-tail was winsorised at the 5th percentile, then the data was re-scaled to represent years since initial collection year. The 5th percentile corresponded to a collection date of 1994.
Binary categorical variables	
Integron presence/absence	*Baseline: absence* Only 'complete integrons' (encoding an integrase and at least one *attC* site) were considered (Cury et al., 2016).
Biocide and metal resistance gene presence/absence	*Baseline: absence*
Virulence gene presence/absence	*Baseline: absence*
Categorical variables	
Conjugative system	*Levels: non-mobilisable (baseline), mobilisable, conjugative* Plasmids encoding a complete conjugative system were 'conjugative'; plasmids encoding a relaxase but lacking a complete conjugative system were 'mobilisable'; all other plasmids were 'non-mobilisable'.
Replicon type multiplicity	*Levels: untyped (baseline), single-replicon type, multi-replicon type* Factor level is determined by the number of unique replicon types detected. E.g. 'IncFIB,IncFIC' type is a multi-replicon type plasmid, whereas 'IncFIC,IncFIC' is a single-replicon type plasmid.
Host taxonomy	*Levels: Enterobacteriaceae (baseline), non-Enterobacteriaceae Proteobacteria, Firmicutes, other* 'other' includes plasmids with missing taxonomy data or rare factor levels not falling within the Proteobacteria or Firmicutes phyla.
Geographic location	*Levels: high-income (baseline), middle-income, EU, China, United States, other* 'high-income': World Bank high-income countries (not included in other categories); 'middle-income': World Bank lower-middle income and World Bank upper-middle income countries combined (not included in other categories); 'EU': the EU28 countries (includes the United Kingdom); 'other' includes plasmids with missing location data, missing World Bank income categorisation, or rare income category (World Bank low-income countries).
Isolation source	*Levels: human (baseline), livestock, other* 'other' includes plasmids with missing isolation source data, uncategorised isolation source metadata, or rare factor level (agriculture).

Prior to fitting GAMs, exploratory univariate analyses were conducted, including calculation of unadjusted odds ratios for categorical variables. After univariate analysis, long-tailed distributions were winsorised to reduce outlier influence, and collection year was re-scaled to represent years since the initial collection year (given winsorising, this was the collection year at the 5th percentile) (see Table 1). I then proceeded to multivariate analysis; the advantage of multivariate-adjusted odds ratios is that they account for confounding bias, and so estimate the independent effects of explanatory variables (provided that collinearity between explanatory variables is low). To explore inter-relationships among variables and identify collinearity, association statistics between explanatory variables were calculated: Spearman's correlation (numeric–numeric, and numeric–binary); Cramér's V statistic (categorical–categorical); and Kruskal-Wallis eta statistic (numeric–categorical). A threshold of >0.7 was used to determine whether explanatory variables should be discarded from modelling (Dormann et al., 2013). Next, logistic GAM models, implemented using the mgcv R package, were constructed using the deduplicated plasmid dataset. GAM models (10 models, one for each antibiotic resistance outcome variable) took the following structure:

```
gam(resistance ~ s(log10PlasmidSize, k = 5) + s(InsertionSequenceDensity, k = 5) +
s(NumOtherResistanceClasses, k = 5) + s(CollectionDate, k = 3) + Integron +
BiocideMetalResistance + Virulence + ConjugativeSystem + RepliconMultiplicity +
HostTaxonomy + GeographicLocation + IsolationSource, family = 'binomial',
data = DeduplicatedPlasmids.tsv, method = 'REML', gamma = 1.5)
```

The s() function fits smooth terms, and all other terms are parametric categorical variables. The number of basis functions for a smooth ('basis dimensionality', default k=10), was restricted to limit the number of degrees of freedom. The gam.check() function was used to test for non-random patterns in residuals, indicative of insufficient basis dimensionality. In addition, gam.check() was used to confirm model convergence. Smoothing parameters were selected using the maximum likelihood (ML) method initially; once a final model structure was confirmed, restricted maximum likelihood (REML) was used. ML/REML have been shown to reduce overfitting compared with other available methods (Wood, 2011). The gamma parameter was set to 1.5 to reduce overfitting (Wood, 2017a).

For each gam model, the statistical significance of smooth and parametric terms was tested using the anova.gam() and summary.gam() functions, which implement Wald tests (Wood, 2017b). Smooth and parametric effect plots were visualised using the mgcViz package (Fasiolo et al., 2020). A Bonferroni-adjusted alpha of $p<0.00025$ (17 parametric factor levels and 3 smooth terms per model, across 10 models) was used to guide interpretation of multivariate results. The same model was fitted to each antibiotic resistance outcome, to ensure that the same potential confounders were adjusted for in each model. After fitting data with the main model structure (above), alternative models were fitted with various terms removed, in order to investigate confounding. The terms selected for removal were guided by the exploratory association statistics, and by the magnitude of observed differences between the effect size of coefficients in univariate vs multivariate analysis.

5.4 Results and discussion

5.4.1 Results of initial plasmid curation and deduplication steps

After running the bacterialBercow plasmid curation pipeline, there were 16275 complete bacterial plasmid accessions. A further five accessions were excluded following manual inspection of identical sequence clusters (13440 plasmids had already been automatically deduplicated by bacterialBercow). For more details, see Figure 1, which gives an overview of the complete curation and deduplication pipeline. Findings from the initial deduplication of identical plasmids are described below. Of the 29871 accessions prior to deduplication, there were 25953 accessions represented across 12503 clusters of identical sequences (Appendix D: Table 1A). The majority of clusters (12495) were automatically deduplicated, retaining one representative per cluster (of these clusters, 12150 comprised a cognate pair of RefSeq/GenBank redundant duplicates). The remaining eight clusters comprised 16 potentially non-duplicate accessions; following manual oversight, four accession clusters were retained as non-duplicate, based on literature support or submitter correspondence

(Supplementary Table S1; Appendix D: Table 1B). Reference-guided assembly may lead to omission of single-nucleotide polymorphisms (SNPs) and structural variants. Consequently, identical plasmids may be found more commonly than might otherwise be expected (Pightling et al., 2014; Garg et al., 2018). It is therefore worth noting that some of the non-duplicate identical accessions represented a reference sequence and a corresponding reference-guided assembly sequence (Supplementary Table S1).

Besides the four retained non-duplicate clusters, other clusters were deduplicated or excluded completely (cluster 2). Cluster 2 accessions were excluded because at least one of the plasmids was from a commercial biopesticide strain of *Bacillus thuringiensis*, and the focus of downstream analyses was on natural plasmids. Of the duplicate clusters, clusters 3 and 6 comprised redundant accessions with semantically equivalent metadata, while cluster 8 accessions were mutant derivatives (Supplementary Table S1). The same categories of duplicate accessions emerge when inspecting accessions from clusters that were automatically deduplicated, but differed in terms of BioSample/GenBank BioProject accession (I examined eight of 121 such clusters). Cluster 10 comprised redundant accessions with semantically equivalent submitter metadata, whereas clusters 11, 13, and 14 comprised laboratory mutant derivative accessions (Supplementary Table S2; Appendix D: Table 1C). In total, only four of eight clusters in Supplementary Table S2 were considered plausibly non-duplicate, so the stringent approach towards automated deduplication (using submitter metadata in addition to BioSample/BioProject accession ids) appears justified.

Overall, these results highlight the limitations of simplistic approaches towards deduplication, which have been used by previous researchers. Indeed, the results informed the deduplication strategy for non-identical plasmids, where multiple pieces of metadata are used in conjunction with sequence similarity thresholds (see Methods, Section 5.3.5). Notably, GenBank BioProject accession ids do not always delimit independent sequencing projects (Supplementary Table S2) and should therefore not be used as a sole deduplication criterion. Likewise, redundant re-uploading, or sequencing of

laboratory mutant derivatives, means that accessions may be assigned different strain names yet represent duplicates. In a recent study, Douarre et al. identified 234 accession clusters comprising identical plasmids associated with the same species but different strain names, and concluded that these represented cases of intra-species plasmid dissemination (Douarre et al., 2020). Given findings presented here, this conclusion is clearly flawed; indeed, clusters 3, 8, 10, 11, and 13 which were found to be duplicate clusters following manual investigation are included among the 234 "plasmid dissemination" clusters (see Table S2 in Douarre et al.). Manual cluster examination has also demonstrated the limitations of using sequence similarity thresholds as a sole deduplication criterion. For example, a 100% identity/coverage breadth threshold will succeed in eliminating redundant duplicates (i.e. sequence re-uploads) but laboratory mutant derivatives or similar plasmids from a single site study will not be deduplicated. On the other hand, numerous clusters comprised identical plasmids isolated independently from different locations (clusters 1, 4, 7, 9, and 16); applying a 100% similarity threshold in these cases would lead to false-positive deduplication. The false-positive deduplication rate would increase further with relaxation of the similarity threshold.

5.4.2 Description of curated metadata and identification of erroneous metadata

Following bacterialBercow curation and deduplication of identical sequences, there were 16270 plasmid nucleotide accessions. Source taxon and source taxid metadata was available for all plasmids. 11848 (71%) plasmid nucleotide accessions were linked to a BioSample accession. In total, there were 5195 BioSample accessions. Of the 5195 total samples, there were 3604 (69%) with a geographic location, and 1207 (23%) with a latitude/longitude coordinate (excluding missing/invalid values). This represented 8363 and 2792 nucleotide accessions, respectively (51% and 17% of total nucleotide accessions). Of the 1207 samples with a latitude/longitude coordinate, all except six also had a geographic location. Of the 3604 locations, there were 1258 unique location names, which were geocoded. Automated geocoding was successful for 513 unique location names; the remainder were

manually geocoded due to lack of exact matching to a Google Places description (734 locations) or place name ambiguity (11 locations).

Among the 1201 locations with both a geocoded coordinate and a BioSample coordinate (*lat_lon* attribute), there were 43 discrepancies (Appendix D: Table 2). Many discrepancies were relatively minor, possibly reflecting different levels of geospatial precision (e.g. nearest city given as the location name for a rural sampling location). There were six major discrepancies (>500 km inter-coordinate distance) which could not be attributed to issues of precision. In some cases, one coordinate appears to represent the research institute, while the other represents a distant sampling location (e.g. SAMN08102921). In other cases, the discrepancy seems inexplicable (e.g. SAMN03154423, where the primary BioSample coordinate for a plant pathogen is a remote ocean location) (Appendix D: Table 2).

There were 655 samples with a detected culture collection identifier (for 437 of these samples, the identifier was found in the designated *culture_collection* field). Bac*Dive* metadata was available for 236 of the 655 samples. Following Bac*Dive*-guided curation of the 236 samples, additions and/or corrections were made to metadata fields of 116 samples (additions were implemented for 91 samples and corrections were implemented for 33 samples). The majority of discrepancies requiring correction occurred in the collection date field (Appendix D: Table 3). For 9 of 29 collection date discrepancies, the error could be attributed to species taxonomic authority date (i.e. date of taxonomic description) being given instead of the genuine collection date (Appendix D: Table 3). This was the case for the plasmid accession mentioned in the Introduction, where a dubious collection date of 1884 had been given, and subsequently reproduced in the literature (nucleotide accession NZ_CP020021.1; BioSample accession SAMN06480665; *Staphylococcus aureus* strain ATCC 6538). Manual oversight reveals that 1884 is the date *S. aureus* was originally described, whereas the ATCC 6538 strain appears to have been collected in 1951 (Peacock, 2006; Makarova et al., 2017). Confusion presumably occurs because species authority date is commonly listed on culture collection webpages (*pers. obs.*). For remaining collection date discrepancies, the reason for the discrepancy was unclear;

discrepancies may have arisen due to the date of isolate acquisition or the date of deposition in a culture collection being mistakenly provided as the collection date. Further manual examination of collection dates (focusing on pre-1950 dates) revealed an additional four discrepancies attributable to authority/collection date confusion. On the other hand, for 21 of 45 samples with a pre-1950 collection date, there was literature support for the collection date, with the earliest apparently genuine date being 1911 (samples SAMEA3368254 and SAMN02943517) (Appendix D: Table 4).

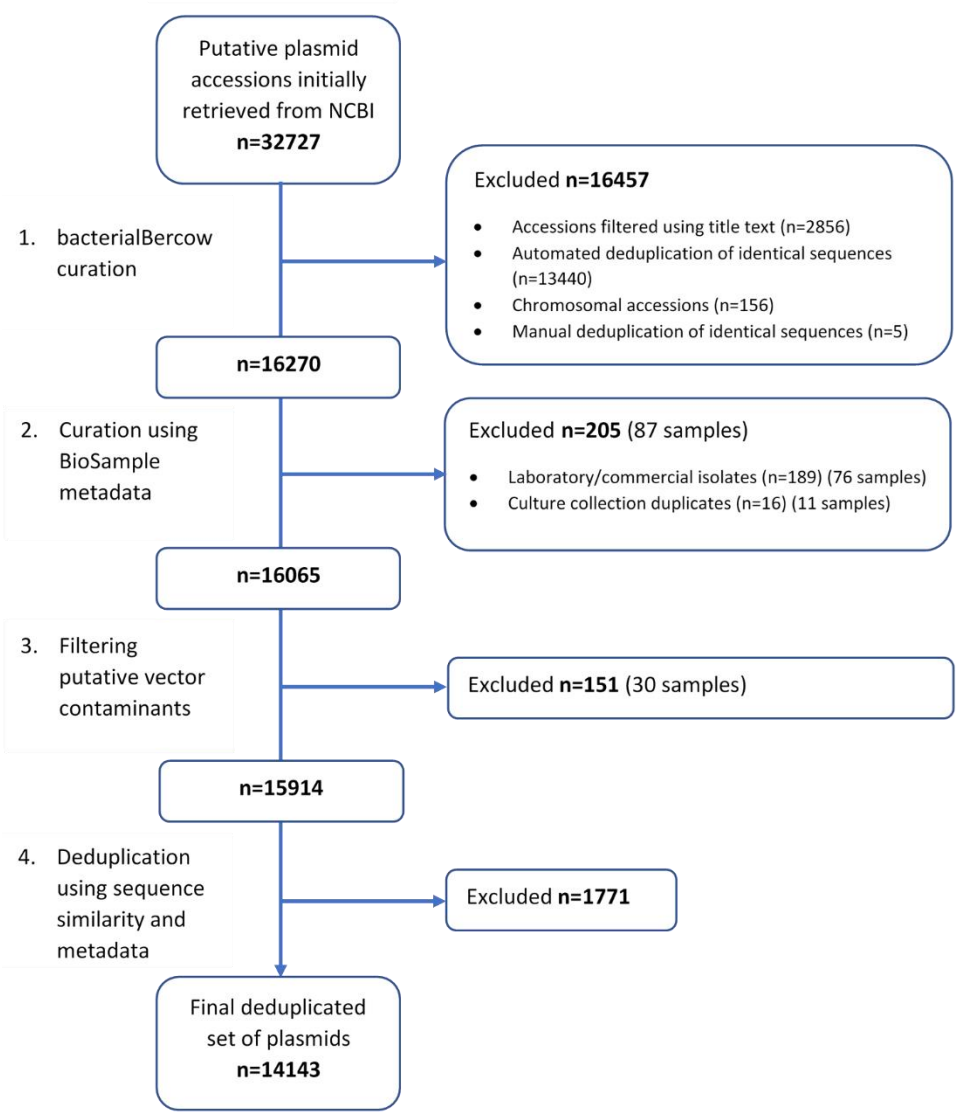


Figure 1. Plasmid curation flowchart, indicating numbers of plasmid accessions during the curation process from initial retrieval (n=32727) to the final deduplicated dataset of curated plasmids (n=14143). The number of plasmids excluded at curation steps is shown on the right. Curation steps are described in text on the left. From the dataset of 15914 plasmids, a subset of 4668 that encoded antibiotic resistance gene(s), were visualised using Microreact, to show the global epidemiology of resistance plasmids. All other downstream statistical analyses were performed using the final dataset of 14143 deduplicated plasmids.

Based on examination of BioSample metadata, 76 samples were excluded as non-natural (laboratory/commercial) isolates, while 11 samples were excluded as culture collection duplicates. The 87 excluded BioSample accessions represented 205 excluded plasmid nucleotide accessions. A further 151 plasmids were excluded as putative vector contaminants. Finally, after delineating non-independent sequences based on sequence similarity and metadata (see Methods), 1771 plasmids were excluded as duplicates, leaving 14143 in the fully-deduplicated plasmid dataset (Figure 1). Datasets showing retained and excluded plasmid accessions, as well as associated sample metadata are provided in Appendix D: Table 5A – D.

Overall, for BioSample accessions where both geographic location name and latitude/longitude was available, internal consistency was high (consistency rate=96.4%). Because only ~23% of samples had latitude/longitude data (compared to ~69% with a location name), geocoding is currently required to capitalise on available BioSample geographic metadata. Unfortunately, geocoding generally requires time-consuming manual oversight to achieve high accuracy (McDonald et al., 2017), as demonstrated in this study. Therefore, future improvements to NCBI BioSample could include implementing a controlled vocabulary of place names (e.g. https://www.geonames.org/), and/or encouraging submission of latitude/longitude coordinates. External validation of BioSample metadata using the Bac*Dive* database highlighted particular problems associated with culture collection samples and erroneous collection dates. Links to Bac*Dive* were recently included in NCBI nucleotide (Reimer et al., 2019); an obvious next step would be to pull metadata from Bac*Dive* rather than relying on submitters to provide accurate metadata for culture collection samples. Culture collection samples made up only ~13% of total samples. Nevertheless, erroneous collection dates and geospatial data, even if relatively rare, can be problematic for descriptive analyses, as demonstrated by the example of the earliest plasmid collection date being erroneously reported as 1884 (see above). The full extent of erroneous metadata is unclear. Of the total 5195 BioSample accessions, only 23% had both a latitude/longitude and a geographic location description, and of 655 culture collection samples, only 36% had linked Bac*Dive* metadata; consequently, internal/external consistency checks to assess validity of metadata

were restricted to a fraction of samples. Further, the internal consistency check (latitude/longitude vs location description) will fail to detect erroneous metadata if both items are incorrect. Finally, it is worth remembering that the errors identified reflect a single-timepoint snapshot of the BioSample database. It is unclear to what extent ongoing NCBI curation efforts may have resolved the errors identified.

5.4.3 Global epidemiology of the curated antibiotic resistance plasmids

Of the 15914 curated plasmids (see Figure 1), 4668 encoded one or more antibiotic resistance genes. The Microreact visualisation of these resistance plasmids can be accessed here: https://microreact.org/project/hgs14AXBUmjV6ZikdV2KvA. By interactively filtering the data, the epidemiology of antibiotic resistance genes/classes of interest can be explored. Taking the *bla*$_{KPC}$ carbapenemase as an example, the first *bla*$_{KPC}$ plasmid in the dataset was collected in 2007 in the USA and since then *bla*$_{KPC}$ plasmids occurred predominantly in North America, South America, China, and Europe. This geographic distribution fits with literature reports (see Figure in Munoz-Price et al., 2013); however, plasmid-mediated *bla*$_{KPC}$ is reported in the literature much earlier (e.g. Yigit et al., 2001). Clearly, a drawback of assessing only complete plasmid sequences is that some confirmed cases of plasmid-mediated resistance will be missed. On the other hand, there will be cases of plasmid-mediated resistance which have not been reported in the literature (at least not currently), but which can be detected from complete plasmid sequences submitted to NCBI nucleotide. Indeed, looking at the epidemiology of *mcr* genes in Microreact, suggests that plasmid-mediated *mcr* colistin resistance emerged as early as 2001 – an *mcr-5*-encoding plasmid, collected from livestock in Japan (accession NZ_AP018799.1). On the other hand, a recent literature review supports ~2006/2007 as the earliest recorded collection date for *mcr-5*-encoding plasmids (see Table S1 in Luo et al., 2020) (García et al., 2018; García-Meniño et al., 2019).

5.4.4 Exploratory analysis of plasmid resistance carriage patterns across major resistance classes

Exploratory analyses of resistance carriage were conducted on the deduplicated set of 14143 plasmids. Figure 2 shows a two-way contingency table of resistance gene class carriage. The major resistance classes are ordered from most frequent (aminoglycoside) to least frequent (quinolone), in terms of the number of plasmids encoding one or more resistance genes of a given class. The most closely overlapping classes are aminoglycoside and sulphonamide (the class intersection represents 66% and 92% of aminoglycoside and sulphonamide-encoding plasmids respectively). However, the fact that there were no cases of (near) complete overlap between classes justified building separate models for each class.

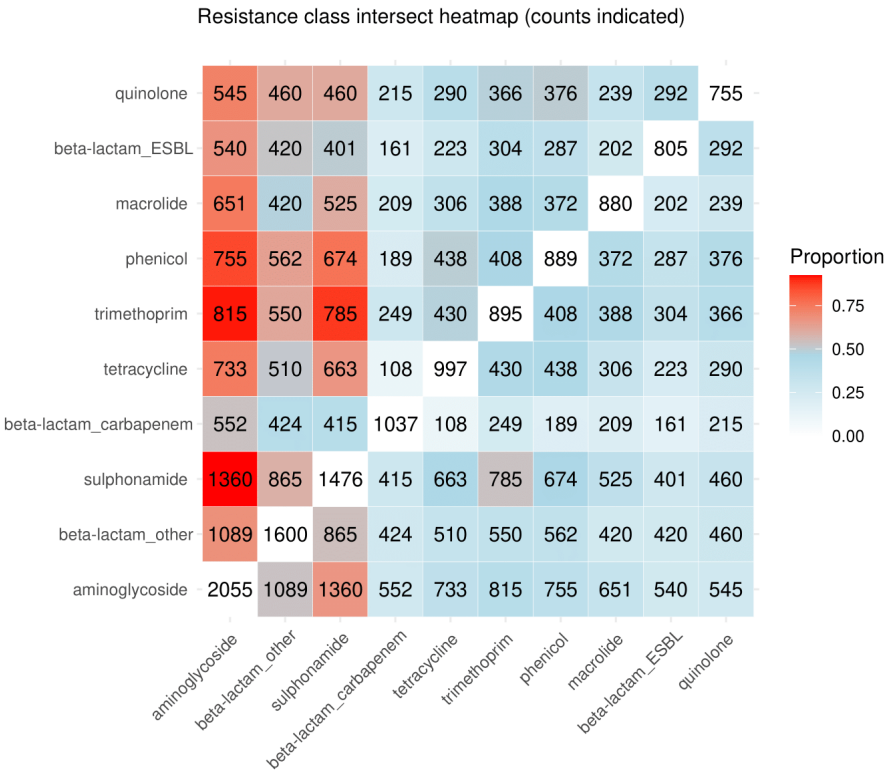


Figure 2. Heatmap contingency table showing pairwise overlaps of resistance class carriage in the deduplicated plasmid dataset. Resistance classes are ordered by frequency. Counts indicate total numbers of plasmids carrying a given resistance class (along the diagonal) and number of plasmids with a given pairwise resistance class combination. Colouring represents counts as a proportion of the less frequent (top left) and more frequent (lower right) resistance class.

To investigate the reliability of using accession 'create dates' to impute missing collection dates, the correlation between the two variables was investigated for the subset of plasmids with a non-missing collection date (Supplementary Figure S1). This indicated weak correlation (Pearson's r=0.314); given that 7768 (55%) plasmid collection dates were imputed, associations between resistance carriage and the collection date variable should be interpreted cautiously.

For categorical variables, raw counts and unadjusted odds ratios are given in Appendix D: Table 6; unadjusted odds ratios are discussed below (Section 5.4.6) in conjunction with adjusted odds ratios. Analysis of numeric variables was not conducted assuming linear relationships; instead, presumed non-linear relationships were investigated using GAMs (Section 5.4.6). The transformed/re-coded variables used as input for GAM models are provided in Appendix D: Table 5E.

5.4.5 Re-coding of factor variables, and multivariate model checking

Prior to modelling, factor variables with rare categories were re-coded. Re-coding host taxonomy categories was essential since there were rare categories at all taxonomic levels (even at phyla-level there were 23 categories). The decision to use the factor levels Enterobacteriaceae, Proteobacteria (not including Enterobacteriaceae), Firmicutes, and "other", was taken in order to partition major pathogens associated with antibiotic resistance. For example, taking the ESKAPE pathogens, *Enterococcus faecium* and *Staphylococcus aureus* belong to Gram-positive Firmicutes; *Klebsiella pneumoniae* and *Enterobacter* spp. belong to Enterobacteriaceae; *Acinetobacter baumannii* and *Pseudomonas aeruginosa* belong to non-Enterobacteriaceae Proteobacteria. Supplementary Table S3 lists the most common species represented in the dataset within each factor level.

Isolation source and geographic location factor levels were also re-coded to produce broader aggregate categories (Supplementary Tables S4 and S5). The original replicon type variable comprising 232 replicon types (Supplementary Figure S2), was re-coded to a new factor variable ("replicon type

multiplicity"), with the following factor-levels: untyped, single-replicon, multi-replicon (see Table 1). These broader categories still enable interesting biological interpretations (see Section 5.4.6).

Collinearity was investigated using association statistics. Initially, the number of insertion sequences was considered as an explanatory variable, but this showed strong correlation with plasmid size (Spearman's ρ = 0.74) (Appendix D: Table 7A) (Touchon and Rocha, 2007). After expressing the number of insertion sequences as an insertion sequence density score, all association statistics were below the commonly used <0.7 threshold (Dormann et al., 2013) (Appendix D: Table 7A). Therefore, I proceeded to construct multivariate logistic GAMs, using the deduplicated plasmid dataset (10 GAMs for the 10 resistance class outcomes). Output from gam.check showed that all models converged (Appendix D: Table 7B). Basis dimensionality checking indicated that there were non-random patterns in the residuals for the $\log_{10}$ plasmid size smooths across most models ($p<0.05$ for all models except phenicol) and for the insertion sequence density smooths across some models ($p<0.05$ for all models except macrolide, phenicol, sulphonamide, and trimethoprim) (Appendix D: Table 7B). This was not resolved by increasing the basis dimensionality. The $\log_{10}$ plasmid size and insertion sequence density terms were retained, but the smooths should be interpreted cautiously. Model R^2 values ranged from 0.266 (beta-lactam_ESBL) to 0.765 (sulphonamide) (Appendix D: Table 8A).

5.4.6 Factors associated with plasmid antibiotic resistance carriage

Unadjusted and multivariate-adjusted odds ratios and p-values are detailed in Appendix D: Tables 6 and 8, respectively. Adjusted odds ratios indicate the independent effect of a given numeric (smooth) or categorical explanatory variable on resistance carriage; they are visualised on the log-odds scale in the main text below (a value above 0 indicates a positive association with resistance carriage, and vice-versa). While log-odds ratios communicate the direction and magnitude of effects across models, they cannot be interpreted directly as probabilities. Therefore, log-odds ratios are also presented on the probability scale (Appendix D: Table 8B, Supplementary Figures 3, 8B, 14B, 15B). Overall, for some

explanatory variables, unadjusted and adjusted odds ratios were similar, notably for geographic location and virulence gene presence. For other explanatory variables (especially, integron presence, biocide/metal resistance gene presence, and host taxonomy), there were marked differences between unadjusted/adjusted odds ratios (Supplementary Table 8B). The latter suggests strong confounding. Below, I focus on describing and interpreting results of the multivariate regression analysis. I also discuss how confounding-interrelationships between explanatory variables can explain differences between univariate and multivariate effects. A detailed discussion of the limitations of the analyses is reserved for Section 5.4.7.

Antibiotic selection pressure is thought to be important in promoting the emergence and persistence of antibiotic resistance. Usage of different classes of antibiotic (and hence, antibiotic selection pressure) varies according to higher-level abiotic factors such as isolation source (human/livestock) and collection date. Selection pressure will also vary in relation to host taxonomy, notably, because different antibiotic classes have differing activity spectra across taxa. In my analysis, the adjusted odds ratios for three presumed proxies of antibiotic selection pressure (isolation source, collection date, and host taxonomy) are consistent with antibiotic selection pressure as an important driver of plasmid-mediated antibiotic resistance. Regarding isolation source, only 0.0042% (3/708) livestock plasmids encoded carbapenem resistance, compared with 12% (356/2954) human plasmids, and this is reflected in a strongly negative adjusted odds ratio for carbapenem resistance carriage in livestock relative to human isolation source (Figure 3A). Conversely, relative to plasmids isolated from humans, adjusted odds ratios indicated higher carriage of phenicol, sulphonamide, and tetracycline resistance, among livestock plasmids (Figure 3A). These results are consistent with available data on antibiotic usage. Specifically, antibiotic sales/prescriptions data suggest that tetracycline is the most heavily used antibiotic class in livestock; phenicol and sulphonamide antibiotics are also used more heavily in livestock than in humans. In contrast, carbapenem antibiotics are primarily reserved for usage in human medicine (Economou and Gousia, 2015; ECDC/EFSA/EMA, 2017; HM Government, 2019, 2020).

Across most antibiotic classes, there is no evidence for a strong relationship between resistance carriage and sample collection year. However, the log-odds (and predicted probability) of quinolone, carbapenem, and ESBL resistance carriage has increased since the baseline collection year (1994) (Figure 3B and Supplementary Figure S3). As mentioned in Section 5.4.4, collection date data is low quality, but the large effect sizes, especially in the carbapenem and quinolone resistance models, are striking. The results are concordant with reports in the literature, and appear to reflect the history of antibiotic usage. The quinolone class was first discovered in the early 1960s (nalidixic acid), however widespread use only began with the introduction of improved broad-spectrum fluoroquinolones from the early 1980s (e.g. ciprofloxacin) (Emmerson and Jones, 2003). Plasmid-mediated quinolone resistance was not discovered until 1998 (Martínez-Martínez et al., 1998). Plasmid-mediated carbapenem resistance is also a relatively recent phenomenon (reports were rare until ~late 2000s); carbapenem usage increased following the emergence of ESBL resistance in the 1980s (Paterson and Bonomo, 2005; Patel and Bonomo, 2013). Compared with quinolone and carbapenem antibiotics, the other antibiotic classes gained earlier widespread clinical adoption, and reports of plasmid-mediated resistance to these classes emerged earlier – at least as far back as the early 1980s, and substantially earlier in the case of resistance to antibiotics such as sulphonamides, tetracyclines, and penicillins (penicillin resistance is included in the "beta-lactam_other" resistance model) (Welch et al., 1979; Davies, 1983; Roberts, 1996; Silver, 2011; Van Hoek et al., 2011). The increases in quinolone and ESBL resistance carriage appear to slow in later years, and plateau in the last ~5 years of the study period. The increase in carbapenem resistance carriage also appears to have slowed in later years but remains on an increasing trajectory in the latest years of the study period.

Regarding host taxonomy, the unadjusted analysis showed negative associations with resistance carriage for Proteobacteria, Firmicutes, and other bacteria, relative to Enterobacteriaceae (the baseline factor level), across all resistance class outcomes. However, in the adjusted analysis, negative associations were attenuated or reversed (Supplementary Figure S4, Appendix D: Table 8B), suggesting that the unadjusted effects of host taxonomy could be explained at least in part by other

explanatory variables (notably, replicon type multiplicity, and the number of other resistance gene classes, as explained below). The association statistics indicated that host taxonomy was most strongly associated with replicon type multiplicity and the number of other resistance gene classes (Appendix D: Table 7A). Moreover, when these two factors were removed from the multivariate model, the host taxonomy adjusted log-odds ratios resembled the unadjusted log-odds ratios more closely (Supplementary Figure S4, Appendix D: Table 9A). Further exploratory analysis indicated that Enterobacteriaceae plasmids were more likely to encode one or more replicon types, and that for a given resistance class outcome, they were more likely to encode other resistance gene classes (Supplementary Figure S5). Subsequent analysis showed that replicon carriage and carriage of other resistance gene classes are both positively associated with resistance carriage (see below). Therefore, it makes sense that when replicon type multiplicity and the number of other resistance gene classes are adjusted for, the negative unadjusted log-odds ratios observed in non-Enterobacteriaceae taxa are attenuated/reversed. While the magnitudes of the log-odds ratios differ between unadjusted/adjusted analyses, the overall pattern across resistance classes is similar, and can be interpreted in the context of host taxa resistance mechanisms and antibiotic selection pressure. Below, I focus on interpreting the adjusted log-odds ratios presented in Figure 3C.

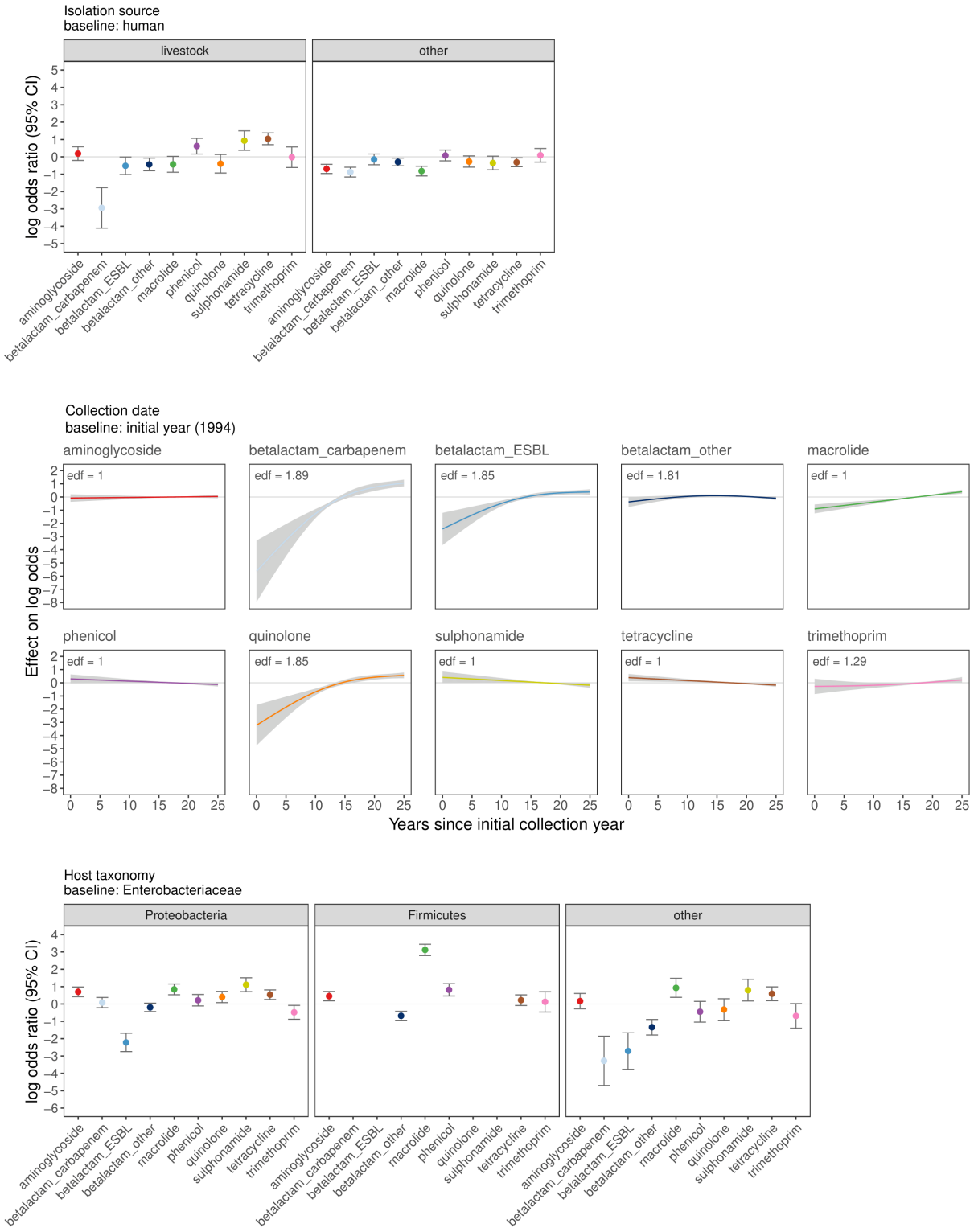


Figure 3. The multivariate-adjusted association between presumed proxies of antibiotic selection pressure (A, isolation source; B, collection date; C, host taxonomy) and the log-odds of antibiotic resistance carriage (y-axis), across 10 resistance classes. For factor variables (isolation source, host taxonomy), log-odds ratios indicate the effect of a given factor level, relative to baseline factor level, and error bars show 95% confidence intervals. For collection date, smooth curves indicate the predicted effect of sample collection date (expressed as years since baseline collection year, 1994). An effective degrees of freedom (edf) value >1 indicates a non-linear smooth. Shaded areas around smooths show 95% confidence intervals. Note that no Firmicutes plasmids encoded known

quinolone, sulphonamide, or ESBL resistance genes, and only 2 Firmicutes plasmids encoded known carbapenem resistance genes; therefore, corresponding odds ratios are not shown.

There were no plasmids with known quinolone or sulphonamide resistance genes among Firmicutes (Gram-positives); this is consistent with available literature which suggests that quinolone and sulphonamide resistance are predominantly non-plasmid-mediated in Gram-positive bacteria (Sköld, 2000; Jacoby et al., 2014; Rodríguez-Martínez et al., 2016; Foster, 2017). There were also no Firmicutes plasmids encoding ESBL resistance, and only two encoding carbapenem resistance. Again, this is consistent with known epidemiology: in Gram-positive bacteria, reported beta-lactam resistance mechanisms are: 1) modification to penicillin-binding proteins, and 2) plasmid-mediated penicillinases. In contrast, Enterobacteriaceae plasmids are known to encode beta-lactamases capable of hydrolysing cephalosporins, monobactams, and carbapenems, as well as penicillins (Bush, 2018). There was a strong positive association with macrolide resistance carriage for Proteobacteria, and especially, Firmicutes plasmids. The particularly large effect size observed in Firmicutes aligns with the fact that macrolides are more commonly used to treat Gram-positive infections such as *Staphylococcus*, whereas they are less widely used for Gram-negatives due to low membrane permeability, which confers intrinsic resistance (Gomes et al., 2017; Golkar et al., 2018). In summary, these findings indicate that patterns of plasmid-mediated resistance vary according to the biology of host taxa resistance mechanisms (chromosomal, plasmid, intrinsic). Ultimately, this links with antibiotic selection pressure: there will not be strong selective pressure for plasmid-mediated resistance to a given antibiotic class, if resistance occurs chromosomally/intrinsically.

A priori, I expected that geographic location might influence the log-odds of resistance carriage, given varying regional patterns of antibiotic usage. For example, in the EU, bans on using antibiotics as growth promoters in livestock were introduced as early as the 1970s (for tetracycline, penicillin, and streptomycin), whereas the United States and China have been slower to introduce strict controls (e.g. the United States banned tetracycline for growth promotion in 2017) (Granados-Chinchilla and Rodríguez, 2017; Kirchhelle, 2018; Ma et al., 2021). Previous studies have suggested that geographic

variation in antibiotic resistance prevalence can also be explained by socioeconomic risk factors – low-income, poor sanitation, and poor governance – which promote the spread of antibiotic resistance (Collignon and Beggs, 2019). However, in my analysis, geographic location (modelled as high-income, middle-income, EU, China, United States, other) does not exert major influence on the likelihood of plasmid-mediated resistance carriage (Supplementary Figure S6). These findings should be interpreted very cautiously though. Firstly, location metadata may be inaccurate (see Section 5.4.2). More concerning is the problem of biased sampling (discussed further in Section 5.4.7). Briefly, patterns of plasmid resistance carriage observed among NCBI plasmids from a given geographic location will be linked to available research funding for sequencing projects and the interests of research groups within that location. It is possible that these factors will outweigh any genuine underlying differences in patterns of resistance carriage between different countries or socioeconomic groupings.

Previous analyses have indicated that plasmid antibiotic resistance genes are positively co-associated with biocide/metal resistance genes (Pal et al., 2015). Furthermore, plasmid integrons are known to capture and accumulate multiple adaptive genes including antibiotic resistance genes (Stokes and Gillings, 2011; Gillings, 2014). Finally, resistance plasmids often carry multiple antibiotic resistance genes of different classes (Carattoli, 2013; Mathers et al., 2015). Based on this knowledge of the genetic architecture of plasmid-mediated antibiotic resistance, I expected *a priori*, that there might be positive associations between antibiotic resistance carriage and the following explanatory variables: integron presence, biocide/metal resistance gene presence, and the number of other resistance gene classes encoded on a plasmid (for a given resistance class outcome). In addition, as explained in the Introduction, if antibiotic resistance genes are associated (genetically linked) with other adaptive genes on the same plasmid or within an integron, there is also the potential that co-selection promotes the maintenance of these genetically linked genes, therefore re-enforcing positive co-associations (Hughes and Andersson, 2017).

In unadjusted analysis, all three explanatory variables (integron presence, biocide/metal presence, number of other resistance gene classes) showed strong positive associations with resistance carriage, across all resistance class outcomes. In the adjusted analysis, log-odds remained positive across most outcome classes, but there was strong attenuation (Figure 4, Supplementary Figure S8–S10, Appendix D: Table 8B). This suggests confounding between the effects of the three explanatory variables. This is not surprising given that these explanatory variables were found to be positively co-associated (Appendix D: Table 7A, Supplementary Figure S7). Removing two of the three co-associated explanatory variables from the model reduces attenuation of the effects of the remaining variable in each case (Supplementary Figures S8–S10, Appendix D: Table 9B–D), consistent with confounding bias. Overall, the results of my analysis align with established knowledge on the genetic architecture of plasmid-mediated resistance – integrons, biocide/metal resistance genes, and resistance genes of other classes were generally positively associated with antibiotic resistance carriage. Furthermore, differences in log-odds observed across different resistance class outcomes can in some cases be linked with existing biological knowledge. For example, the strong positive association between integron presence and trimethoprim resistance (Figure 4C) fits with literature reports (many trimethoprim resistance genes are known to be found primarily on integron cassettes) (Grape et al., 2005). Finally, the results support the role of co-selection in maintaining certain antibiotic resistance genes (namely, sulphonamide resistance genes), as explained below.

For integron presence, biocide/metal resistance gene presence, and the number of other resistance gene classes, large effect sizes were seen for the positive association with sulphonamide resistance carriage (Figure 4). This holds true across unadjusted and adjusted analyses (Supplementary Figure S8–S10). This indicates that relative to other outcome classes, there is a strong tendency for genetic linkage between sulphonamide resistance genes and biocide/metal resistance genes, antibiotic resistance genes of other classes, and integrons. One interpretation would be that co-selection has been especially important in promoting persistence of plasmid-borne sulphonamide resistance genes. This interpretation aligns with the known history of clinical sulphonamide usage. Specifically,

sulphonamide antibiotics were the first class of antibiotics to be introduced and were in widespread clinical use through to the 1980s. Since then, due to emergence of high-levels of resistance, as well as side-effects, clinical use of sulphonamides became more restricted, especially in high-income countries (Enne et al., 2001; Skold and Swedberg, 2017; WHO, 2018). Co-linkage of sulphonamide resistance genes with other adaptive genes (on the same plasmid/within integrons), provides an explanation for the persistence of sulphonamide resistance – through co-selection – despite reduced direct selection. Previous small-scale studies have reported co-linkage between integrons and sulphonamide resistance genes (notably class 1 integrons harbouring *sul1*), as well as between sulphonamide resistance genes and other classes of resistance gene (Bean et al., 2009; Domínguez et al., 2019; Goswami et al., 2020).

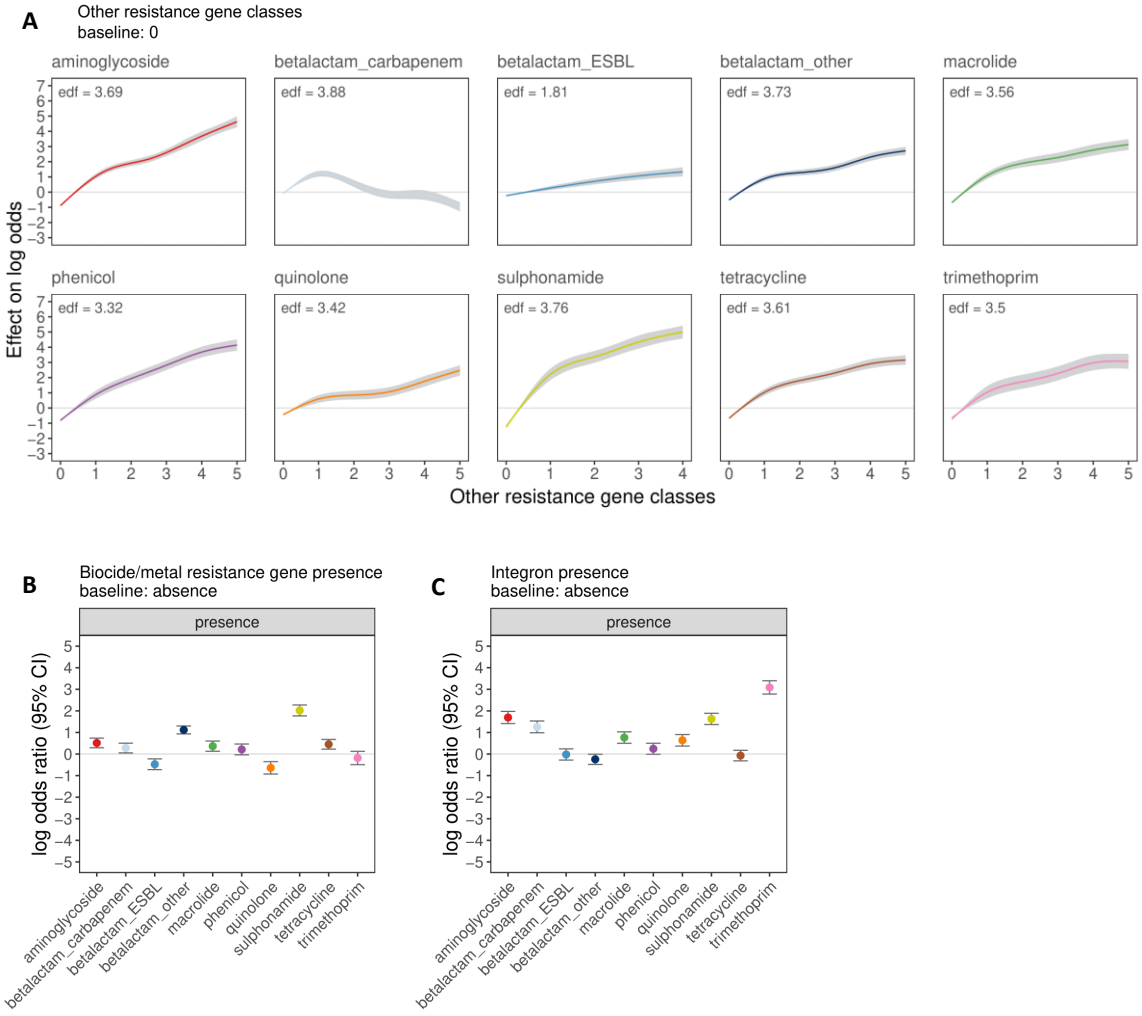


Figure 4. The multivariate-adjusted association between A) other resistance gene classes; B) biocide/metal resistance gene presence; and C) integron presence, and the log-odds of antibiotic resistance carriage (y-axis), across 10 resistance classes. For factor variables (biocide/metal resistance gene presence, integron presence), log-odds ratios indicate the effect of a given factor level, relative to baseline factor level (absence), and error bars show 95% confidence intervals. For the smooth explanatory variable (other resistance gene classes), smooth curves indicate the predicted effect of each value of the explanatory variable. An effective degrees of freedom (edf) value >1 indicates a non-linear smooth. Shaded areas around smooths show 95% confidence intervals.

There is evidence for strong negative co-association between virulence and carbapenem resistance (Supplementary Figure S11). This may link with reports that carbapenem-resistant *K. pneumoniae* strains typically show low virulence (Hennequin and Robin, 2016; Krapp et al., 2017). Overall though, the relationship between virulence and resistance appears to be complex, with no general trend across outcome classes (Supplementary Figure S11). This is likely to reflect the complex evolutionary pressures governing virulence: on the one hand, some authors have suggested that co-selection may

promote co-association of virulence and antibiotic resistance genes (Beceiro et al., 2013). On the other hand, if a virulent antibiotic resistant infection can be treated (e.g. using a broad-spectrum antibiotic) then reduced virulence might be adaptive since infected people will be less likely to seek treatment (Galvani, 2003).

Replicon carriage (single- or multi-replicon carriage) was hypothesised to be positively associated with resistance carriage, based on previous work (Chapter 2; Orlek et al., 2017). One interpretation is that plasmids which can be replicon typed tend to be well-studied plasmids from anthropogenic settings, where resistance carriage is more likely to be adaptive. There was strong evidence of positive association between both single- and multi-replicon carriage and resistance carriage, across all classes in unadjusted analyses. In the adjusted analysis, positive associations were strongly attenuated, although a trend towards positivity remained across most outcome classes (Supplementary Figure S12). The attenuation was likely due to confounding with variables including biocide/metal resistance presence and the number of other resistance classes (Appendix D: Table 7A, Supplementary Figure S12C). *A priori*, multi-replicon plasmids were considered more likely to carry resistance relative to single-replicon plasmids. This is because multi-replicon plasmids may emerge through a cointegration (plasmid fusion) event; fusion events can promote resistance gene acquisition and adaptive evolution (Carattoli, 2009; He et al., 2019). However, the results do not provide strong support for increased resistance gene carriage among multi-replicon plasmids (Supplementary Figure S12).

Insertion sequences and plasmid transmissibility have both been implicated in acquisition and persistence of plasmid-mediated resistance (see Introduction). A previous study demonstrated a link between conjugative plasmids and resistance carriage, based on descriptive univariate analysis (Pal et al., 2015). Here, the univariate analysis supports the link between plasmid transmissibility (i.e. conjugative/mobilisable plasmids) and resistance carriage across all outcome classes, especially in the case of conjugative plasmids, relative to baseline non-mobilisable plasmids (Supplementary Figure S13A). However, in the adjusted analysis, positive associations were attenuated; across most classes

there was no evidence for an association between transmissibility and resistance carriage, although there remained a positive association between conjugative plasmids and carbapenem resistance carriage (Supplementary Figure S13B). The difference between unadjusted/adjusted results suggests that the unadjusted effects of plasmid transmissibility could be explained at least in part by other explanatory variables.

The density of insertion sequences showed evidence of strong positive association with ESBL, and especially, carbapenem resistance carriage in multivariate models (Supplementary Figure S14). In the case of ESBL resistance there appears to be a saturation effect; at the point where plasmids already carry around two or three insertion sequences per 100 kb, the positive effect of higher insertion sequence density on ESBL resistance carriage diminished. Overall, the results fit with literature reports that major carbapenemase and ESBL resistance genes are associated with insertion sequences and transposons (Poirel et al., 2012; Brandt et al., 2019; Acman et al., 2021). A limitation of this analysis is that insertion sequences and composite transposons were not distinguished, despite biological differences – the latter can carry antibiotic resistance genes and other adaptive genes as cargo genes.

Given the model checking results discussed above (Section 5.4.5), the smooths depicting relationships between plasmid size and resistance carriage should be interpreted cautiously. The term was included to address the hypothesis that larger plasmids would carry more resistance genes. The multivariate-adjusted smooths (Supplementary Figure S15) indicate that there is no consistent pattern across resistance outcome classes. In the case of ESBL resistance, there appears to be a simple positive relationship between plasmid size and resistance carriage. For other classes, the relationship is complex (e.g. betalactam_other), or weak (e.g. trimethoprim resistance). In the case of quinolone resistance, there is a relatively strong positive association between small plasmids (< ~10 kb) and resistance carriage. This aligns with literature reports that small ColE-like plasmids predominantly carry *qnrS1* and *qnrB19* quinolone resistance genes (Rozwandowicz et al., 2018). A limitation of this analysis is that plasmid size may be confounded with effects of host taxonomy, which are not

completely adjusted for, given that only higher-level taxa (Enterobacteriaceae, Firmicutes etc.) are included as host taxonomy factor levels.

5.4.7 Critical appraisal of the analyses of factors associated with plasmid resistance carriage

Although efforts were made to reduce sampling bias by minimising pseudoreplication (see Methods), the plasmid dataset is inherently biased due to non-systematic sampling, and this is a major limitation of the analyses. An obvious form of sampling bias occurs due to use of selective culture media, which allows researchers to isolate particular taxa of interest, as well as to select for isolates encoding resistance to specific antibiotics (Bonnet et al., 2020). The plasmid dataset is dominated by plasmids from clinically relevant bacterial species (e.g. *Escherichia coli* and *Klebsiella pneumoniae*, see Supplementary Table S3), because many researchers choose to selectively culture clinical pathogens. Likewise, antibiotic-selective culture is likely to skew the abundance of resistance plasmids within the dataset, which may affect the reliability of the results. For example, research interest in clinical carbapenem resistance may have biased the log-odds ratio for livestock vs human carbapenem resistance carriage (Figure 3A).

The plasmid dataset is also biased in terms of isolation sources and geographic locations. Geographic sampling bias means that the plasmids are not necessarily representative of global plasmids. Unsurprisingly, high-income countries were over-represented (Supplementary Table S5), whereas there were only 70 plasmids from low-income countries (hence, it was not possible to model low-income countries as a geographic location category). Likewise, lower-middle income countries (n = 362) were combined with upper-middle income countries (n = 593) to produce the middle-income category. I found no strong evidence to suggest that socioeconomic conditions (high-income vs middle-income countries) influence the likelihood of resistance carriage, in contrast to findings of previous authors (Collignon et al., 2018). However, if there had been sufficient data to model low-

income countries, or to model lower-middle income countries as an individual category, my findings may have been different.

Around 35% of plasmids could be assigned to coherent isolation source categories (human and livestock). Remaining plasmids were not assigned to a specific category either because they were from rarer isolation sources or were missing isolation source information. The fact that human/livestock were the largest coherent categories is consistent with a bias towards collecting plasmids from anthropogenic environments (as is the predominance of clinically relevant taxa, mentioned above). Consequently, findings may not be generalisable to plasmids from more natural settings, where antibiotics, biocides, and metal products may be less frequently/intensely used. For example, in natural environments, co-selection between antibiotic resistance genes and biocide/metal resistance genes may be less likely. Instead, plasmid antibiotic resistance carriage might persist as a selfish trait, perhaps facilitated by conjugation and/or toxin-antitoxin (TA) systems. Investigating this hypothesis would require not only better representation of plasmids from natural environments, but also improved isolation source metadata. The Bac*Dive* database uses an ontology to categorise isolation sources (Reimer et al., 2019), and NCBI BioSample metadata could benefit from use of an isolation source ontology too.

Besides sampling bias, several other limitations affect the analyses. Due to insufficient data, and for simplicity, resistance carriage was modelled at the level of resistance (sub-)class rather than at the level of resistance gene type (*bla*KPC, CTX-M-15 etc.). The factors associated with resistance carriage may vary between different resistance genes within a given resistance class, but this could not be determined from this analysis. Finally, the higher-level statistical association analysis presented in this study would be complemented by more in-depth genetic analyses. For example, analysis of the genetic structures of integrons in my dataset, or similar large-scale plasmid datasets, would help to increase understanding of potential mechanisms of sulphonamide co-selection.

5.5 Conclusions

This chapter began by curating NCBI metadata, producing a cleaned dataset for downstream analyses. However, the curation step was not simply a means to an end. A recent landmark study investigated NCBI BioSample metadata quality from the perspective of adherence to expected formats (Gonçalves and Musen, 2019); but correctly formatted metadata may be erroneous. Here, erroneous metadata was identified using internal and external validation approaches, representing the first comprehensive analysis of erroneous NCBI metadata. Culture collection samples were highlighted as a key source of erroneous metadata. With a curated dataset of plasmids and linked BioSample metadata, a comprehensive analysis of the factors associated with plasmid resistance carriage – combining biological and abiotic explanatory variables – could be conducted. This analysis was novel, with *a priori* expectations often based on small-scale experimental or observational studies, as opposed to previous large-scale analyses of plasmid datasets. Overall, the analysis was consistent with antibiotic usage as a driver of plasmid-mediated resistance. Notably, resistance patterns in different isolation sources (human/livestock), across time (collection years), and across different host bacterial taxa reflected antibiotic usage patterns, and/or activity spectra in the case of host taxa, as determined from literature reports (e.g. clinical/veterinary antibiotic sales data). Previous studies have used antibiotic usage and resistance data to suggest a link between usage and resistance, but this study is unique in focusing on plasmids, and demonstrating the link using multiple presumed proxies of antibiotic usage. The results of the analysis are also consistent with suggestions from previous smaller-scale studies that genetic linkage between sulphonamide resistance genes and other resistance genes can promote the maintenance of sulphonamide resistance genes through co-selection, despite reduced direct selection pressure. Overall, the analysis demonstrates the value of statistical/bioinformatic analysis in summarising and enhancing knowledge of plasmid-mediated resistance – for example, relative to literature reports, the origin date of plasmid-mediated *mcr-5* colistin resistance was pushed back ~5 years. This analysis, like other studies based on NCBI data, is limited by the problem of sampling bias.

In future, global routine genomic surveillance programmes should help to provide the most reliable and up-to-date insights into resistance plasmid epidemiology (Petersen et al., 2015; Aarestrup and Woolhouse, 2020).

5.6 Supplementary material

5.6.1 Supplementary Methods

Canonical ('harmonized') BioSample attribute names used for metadata retrieval

collection_date, host, lab_host, isolation_source, substrate, tissue, host_body_habitat, host_body_product, host_description, host_disease, host_substrate, host_tissue_sampled, plant_body_site, plant_product, sample_name, strain, sample_type, geo_loc_name, lat_lon, env_broad_scale, env_local_scale, env_medium, env_package, project_name, culture_collection, biomaterial_provider, ref_biomaterial, specimen_voucher, reference_material, derived_from, description

Delimiting valid location names and latitude/longitude coordinates

Case-insensitive matching to the following list of words was used to exclude invalid locations and latitude/longitude coordinates:

'missing', '-', 'n/a', 'unknown', 'not collected', 'not applicable', 'na', 'none', 'not available', 'not determined', 'not recorded', 'n.a.'

In addition, the following Python regex was used to confirm that coordinates were correctly formatted, and extract these coordinates:

```
regexObj=re.compile(r'\d+\.?\d* [NS] \d+\.?\d* [WE]')
match=re.search(regexObj,latlon)
```

Geocoding methods

Unique valid location names were used to retrieve geocoding place predictions from the Google Place Autocomplete service (googleway argument: place_type='geocode'); this returns up to five place predictions, ordered by relevance. Associated geocoded latitude and longitude coordinates were subsequently retrieved through a Google Place Details request. If the top hit geocoded place name description was an exact match to the BioSample place name, and was unambiguous (see below for details), the geocoded place prediction and associated coordinates were accepted. An exact but ambiguous top hit would arise if, for instance, the BioSample text was "London", without further specification as Ontario/England. If automated geocoding failed to produce a matching and unambiguous place prediction, then manual geocoding was conducted: either the top hit was confirmed, or a different geocoding was determined manually.

Determining whether top hit geocoded place name predictions (from Google Place Autocomplete) match exactly to BioSample geographic location names: Matching between top hit geocoded place name descriptions and BioSample place names was assessed by searching for matching substrings using the stringr R package (see code snippet 1 in Appendix D: File 1). An exact match occurred when there were no unmatched substrings in either the geocoded place description or BioSample place name.

Determining whether top hit place name predictions are ambiguous: If an exact match was confirmed (see above), the place prediction was still not accepted unless the place name was unambiguous. Ambiguity was assessed based on "main_text" place names (derived from a structured formatting of

the place description) (Google Maps Platform, 2020c) and a hierarchy of place types, from country through to local-level vicinities (see Table below). Specifically, if the top hit main_text place name was identical to that of other place prediction(s), and these prediction(s) were also equal or higher in the place type hierarchy, the top hit was flagged as ambiguous. For simplicity, only the first place type of a compound place type was used to assess place type hierarchy (e.g. only "locality" for a compound place type such as "locality,political,geocode") (for background information see Google Maps Platform, 2020a).

Table showing a hierarchy of Google Place types from country through to precise local-level locations

Place type(s)	Hierarchical index	Hierarchical category
political colloquial_area geocode	NA*	vague location
country	1	hierarchical location
administrative_area administrative_area_level_1	2	hierarchical location
administrative_area_level_2	3	hierarchical location
administrative_area_level_3	4	hierarchical location
administrative_area_level_4	5	hierarchical location
administrative_area_level_5	6	hierarchical location
postal_town locality	7	hierarchical location
sublocality sublocality_level_1	8	hierarchical location
sublocality_level_2	9	hierarchical location
sublocality_level_3	10	hierarchical location
sublocality_level_4	11	hierarchical location
sublocality_level_5	12	hierarchical location
neighborhood	13	hierarchical location
premise subpremise postal_code natural_feature airport park point_of_interest street_address route intersection	14	local-level location

Place types are Google Place geocoding place types:
(https://developers.google.com/maps/documentation/geocoding/overview#Types). The hierarchical ordering of these place types was determined by this author.
*If the top hit autocomplete place prediction is a "vague location" and has the same place name (main_text) as another hit, it is flagged as ambiguous, unless all other hits are local-level locations.

Curation of collection date and host/isolation source metadata

Valid collection dates were extracted using regexes given in code snippet 2, in Appendix D: File 1. Regexes were also used to curate host/isolation source metadata. Regexes used to identify human

and livestock samples are given in code snippets 3 and 4 in Appendix D: File 1. Human isolation source subcategories were also curated (blood, cerebrospinal, faecal, nasal, oral, skin/wound, sputum, tracheobronchial, urine), however these were not used in downstream analyses. Regex terms for identifying livestock samples covered globally important livestock species (cattle, pigs, chicken, sheep, goats, ducks, turkeys).

For samples with no human/livestock isolation source identified following regex searches, metadata was searched for additional isolation sources (aquaculture, (non-livestock) agriculture, (human) sewage), and all matches were manually curated. The regex given in code snippet 5 in Appendix D: File 1 was used to identify aquaculture samples. Regex terms were compiled to reflect major aquaculture species (based on Table 7 in FAO, 2018). The regex given in code snippet 6 in Appendix D: File 1 was used to identify agriculture samples. Regex terms represent crop species (compiled from FAOSTAT; http://www.fao.org/faostat/en/#data/), bacterial pathogens of crops (Mansfield et al., 2012), and agriculture-related words ('fish farm' was excluded from matches to 'farm'). Matches to wild plants or processed agricultural produce (e.g. sugar) were manually excluded. If metadata indicated that crops were from a laboratory setting, samples and associated plasmid accessions were excluded from the curated plasmid dataset (see Section 5.3.4). Sewage samples were identified using the regex given in code snippet 7 in Appendix D: File 1. As far as possible (using available metadata) matches were restricted to human sewage samples; if metadata indicated industrial wastewater or livestock/agricultural wastewater, matches were excluded.

Identification of culture collection samples

Culture collection acronyms and names were compiled from the Word Federation for Culture Collection website (http://www.wfcc.info/ccinfo/collection/by_acronym/; accessed 6th Jun 2019); additional acronyms were compiled from a list produced by the International Journal of Systematic and Evolutionary Microbiology (IJSEM, 2015). The following regex was used to identify culture collection identifiers in the *culture_collection* attribute field:

```
#Generating a regex from a list of acronyms ("acronyms")   #Example matches:
acronym_regex=[r'\b%s.{0,3}[0-9]*\b'%a for a in acronyms]  #ATCC43845 or ATCC : 43845
combined_acronym_regex=re.compile(r'|'.join(acronyms))     #(Used combined regex to search)
```

Where a match was not found using the above approach, other BioSample fields were investigated:

Sample name identifier (xml element: Id[@db_label="Sample name"])
Title (xml element: .//Description/Title)
Comment (xml element: .//Description/Comment/Paragraph)
As well as the following attribute fields: *isolation_source*, *sample_name*, *strain*, *geo_loc_name*, *project_name*, *ref_biomaterial*, *description*, *reference_material, biomaterial_provider*

The fields were searched using the acronym regex (above), and the descriptive regex below:

```
descriptive_regex=re.compile(r'culture collection|type strain',re.IGNORECASE)
```

In addition, fuzzy string matching was used to search for culture collection names (e.g. "American Type Culture Collection") using the "partial_ratio" method from the Python fuzz module (Kazil and Jarmul, 2016). If the length of the attribute text was >70% the length of the culture collection name *and* the

string match score was >90%, a match was flagged for inspection. All matches involving the non-designated fields were manually curated.

Identification of laboratory samples (from laboratory hosts)

Plant and animal model laboratory organisms were compiled based on literature (Kostic et al., 2013; Pitzschke, 2013; Dietrich et al., 2014; Tsai et al., 2016). Then, *host* and *lab_host* attribute fields were searched using the following regex, in order to flag samples associated with laboratory model organisms:

```
lab_model_regex=re.compile(r'Mus musculus|M.? musculus|mouse|Rattus|R.? norvegicus|\brat\b|
Danio rerio|D.? rerio|zebrafish|Drosophila|D.? melanogaster|fruit fly|Caenorhabditis elegan
s|C.? elegans|nematode|Sepiola atlantica|S.? atlantica|Bobtail squid|Arabidopsis|A.? thalia
na|thale cress|Galleria mellonella|G.? mellonella|greater wax moth|honeycomb moth|Lotus jap
onicus|L.? japonicus|Medicago|M.? trunculata|M.? sativa|barrelclover|alfalfa|Brachypodium|B
.? distachyon|false brome|Oryza|\brice\b|Nicotiana|N.? benthamiana|Glycine|soybean|Triticum
|\bwheat\b|Zea mays|maize|Brassica napus|rapeseed|oilseed rape|Populus|P.? trichocarpa',re.
IGNORECASE)
```

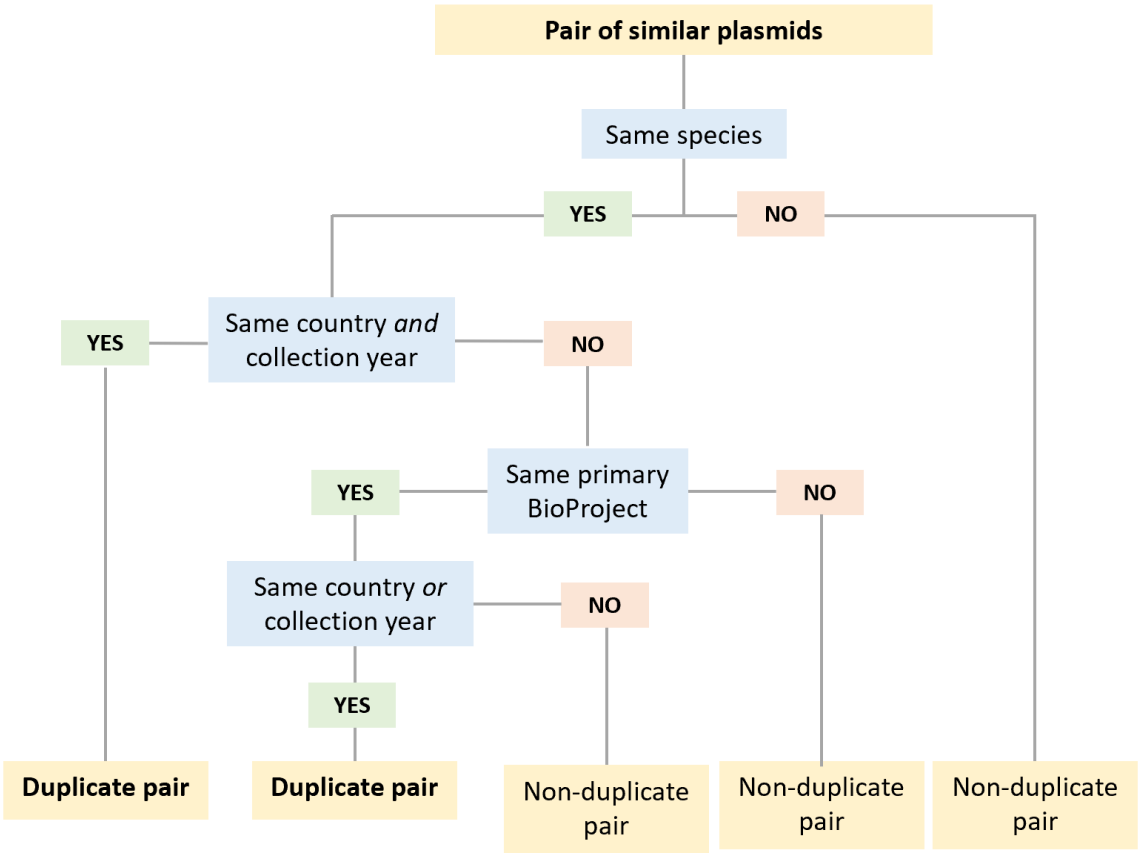
Additional laboratory samples were flagged using the regex below, which was searched against the following fields:

Title (xml element: .//Description/Title)
Comment (xml element: .//Description/Comment/Paragraph)
As well as the following attribute fields: *isolation_source*, *sample_type*, *env_broad_scale*, *env_local_scale*, *env_medium*

```
lab_regex=re.compile(r'\blab\b|laboratory',re.IGNORECASE)
```

Deduplication of plasmids, determined by pairwise similarity, metadata, and graph clustering

First, pairs of plasmids with high sequence similarity were identified using mash followed by BLASTN comparison (see main text Methods). Then, retention of links between similar plasmids was determined by metadata, according to the decision tree shown below. The network of remaining linked plasmids was clustered using the infomap algorithm, and one representative plasmid per cluster was selected for inclusion in the deduplicated plasmid dataset (along with the plasmids that did not share sequence similarity with one or more other plasmids, and were therefore not subjected to the described deduplication steps).



Decision tree for determining which pairwise links between similar plasmids represented duplicate pairs; links between duplicate pairs were retained for deduplication using a clustering approach. Note that a species/country/collection year/primary BioProject was not considered the same if data was missing for both members of the pair (i.e. discounting shared missingness).

5.6.2 Supplementary Tables

Table S1. Clusters of identical accessions which were investigated manually, following automated deduplication

Cluster accessions (deduplicated)	BioSample accession id	Primary BioProject accession id	Submitter name	Affiliation name	Retained
Cluster 1					
MG710483.1	SAMN10679998	512490	Fabricio Campos	Federal University of Tocantins	✓
NZ_CP010009.1	SAMN03216682	238238	Hajnalka Daligault	Los Alamos National Laboratory	✓
The plasmid accessions were isolated independently from different *Bacillus thuringiensis* serovars and from different isolation sources. According to BioSample metadata, accession MG710483.1 (strain Bti-UFT6.51; plasmid pBtiUFT6.51.2; *B. thuringiensis* serovar israelensis) was isolated from soil in Brazil in 2016; accession NZ_CP010009.1 (strain HD 1i; plasmid unnamed11; *B. thuringiensis* serovar kurstaki) was isolated from insect larvae in 2000. Consequently, both accessions were retained. Note that I could not find detailed information on NZ_CP010009.1. Therefore, I cannot rule out the possibility that NZ_CP010009.1 is from commercial bioinsecticide which was used on the soil from which MG710483.1 was collected (c.f. notes on Cluster 2). More details are provided in Campos et al., 2019, who sequenced accession MG710483: "In this work, we sequenced two plasmids found in a Brazilian *Bacillus thuringiensis* serovar israelensis strain which showed 100% nucleotide identities with *Bacillus thuringiensis* serovar kurstaki plasmids." The other accessions mentioned as identical in Campos et al. (MG710485 and NZ_CP004874.1) actually show 99.9% identity and are therefore not detected as identical duplicates following bacterialBercow curation. Accession MG710483.1 was Illumina sequenced and assembled using a reference-mapping approach (using accession NZ_CP010009.1 as the reference): "DNA sequence assembly using the map to reference function in Geneious version 9.1.8 was used" (Campos et al., 2019).					
Cluster 2					
NZ_CP013283.1	SAMN04288432	303961	Alexei Sorokin	MICALIS INRA	×
NZ_CP009344.1	SAMN03010437	236049	Shannon Johnson	Los Alamos National Laboratory	×
NZ_CP013283.1 is a plasmid from a commercial strain of bioinsecticide (*B. thuringiensis* serovar israelensis strain AM65-52) (Bolotin et al., 2017). My interest is in naturally occurring plasmids, so this accession was excluded. NZ_CP009344.1 was isolated from a sewage sample according to BioSample metadata. However, I decided to exclude this accession too in case of a transmission link with the plasmid from the commercial strain.					
Cluster 3					
NZ_CP030795.1	SAMN09534371	230403	Peyton Smith	CDC	✓
NZ_CP018773.2	SAMN06159501	218110	Rebecca Lindsey	Centers for Disease Control and Prevention	×

I used exact matching of metadata to determine deduplication. In this case, both accessions were submitted by the same institution, but indicated with different metadata text (CDC vs Centers for Disease Control and Prevention). Therefore, these accessions were deduplicated.					
Cluster 4					
NZ_CP016508.1	SAMN04334629	305824	Caroline Vincent	Laboratoire de sante publique du Quebec	✓
NZ_CP016523.1	SAMN05263513	298211	Roger Johnson	National Microbiology Laboratory at Guelph	✓
NZ_CP016508.1 is from *Salmonella enterica* subsp. *Enterica* serovar Heidelberg, strain SH12-003, which was isolated from a Canadian (Quebec) hospital patient in 2012, according to BioSample metadata. NZ_CP016523.1 is from *S. enterica* subsp. *Enterica* serovar Heidelberg, strain SA02DT09004001, which was isolated from chicken meat from Canada (British Columbia) in 2009. *S*. Heidelberg is known to be a clonal serovar (Labbé et al., 2016), so these accessions may represent a vertical transmission link between humans and poultry (Genevieve Labbé, pers. comm.). It was confirmed that the accessions were not redundant (i.e. did not represent re-submission of the same sequencing data) (Genevieve Labbé, pers. comm.). NZ_CP016508.1 was Illumina sequenced and assembled using the MIRA assembler (https://sourceforge.net/p/mira-assembler/wiki/Home/), with a reference-mapping approach (although the reference was NZ_CP016511.1 not NZ_CP016523.1) (Genevieve Labbé, pers. comm.).					
Cluster 5					
NZ_CP025496.1	SAMN04966146	321117	Douglas Merrell	Uniformed Services University	✓
NZ_CP025490.1	SAMN03787328	287576	Ryan Johnson	Uniformed Services University of the Health Sciences	✓
NZ_CP025496.1 (*Staphylococcus aureus* subsp. *Aureus*, strain 3020.C01) and NZ_CP025490.1 (*S. aureus* subsp. *Aureus*, strain 2014.C01) are plasmids from clinical isolates. The isolates were collected from the same military battalion on 19th and 12th July 2011, respectively (LaBreck et al., 2018). It was confirmed that these were non-redundant sequences from independent isolates (D. Scott Merrell, pers. comm.), so both accessions were retained. Both accessions were Illumina sequenced, and it appears from LaBreck et al. that contigs were assigned as plasmid/chromosomal using a reference-guided approach: "contig sequences were compared to each other, to published reference genomes, and to PCR, Sanger sequencing, and agarose gel electrophoresis results of restriction enzyme digested and non-digested DNA in order to correctly assign contigs as chromosomal or plasmid as well as to look for assembly artifacts."					
Cluster 6					
NZ_CP009457.1	SAMN03078687	260989	Woori Kwak	Seoul National Univerisity	✓
NZ_CP011119.1	SAMN03434891	279015	Gnanasekaran Gopalsamy	Seoul National University	×
I used exact matching of metadata to determine deduplication. In this case, both accessions were submitted by the same institution, but indicated with different metadata text due to a typo). Therefore, these accessions were deduplicated.					
Cluster 7					
NZ_CP016567.1	SAMN05263514	298211	Roger Johnson	National Microbiology Laboratory at Guelph	✓
NZ_CP016509.1	SAMN04334629	305824	Caroline Vincent	Laboratoire de sante publique du Quebec	✓

NZ_CP016567.1 is from *S.* Heidelberg, strain AMR588-04-00318, which was isolated from chicken faeces from Canada (Ontario) in 2013, according to BioSample metadata. NZ_CP016509.1 is from *S.* Heidelberg, strain SH12-003, which was isolated from a hospital patient in Canada (Quebec) in 2012. For discussion of *S.* Heidelberg epidemiology and transmission, see notes on cluster 4. It was confirmed that the accessions were not redundant (Genevieve Labbé, pers. comm.). NZ_CP016509.1 was Illumina sequenced and assembled using the MIRA assembler (https://sourceforge.net/p/mira-assembler/wiki/Home/), with a reference-mapping approach (although the reference was NZ_CP016583.1 not NZ_CP016567.1) (Genevieve Labbé, pers. comm.).					
Cluster 8					
NZ_CP013346.1	SAMN04288116	303954	Fusako Kawai	Kyoto Institute of Technology	✓
NZ_CP009431.1	SAMN03031197	260764	Yoshiyuki Ohtsubo	Tohoku university	×
These accessions (NZ_CP013346.1, strain 203N, culture collection NBRC 111659; NZ_CP009431.1, strain 203, culture collection NBRC 15033) are not from independent strains (see Ohtsubo et al., 2015, 2016). "The complete genome of NBRC 15033 was determined, but the genes for PEG utilization were missing, and repeated cultivation was assumed to be the reason for the loss. From a laboratory stock, we recovered a strain, designated 203N, harboring the *pegA* gene and capable of growing on PEG." (Ohtsubo et al. 2016).					

Following automated deduplication, eight clusters retained more than one accession. These accessions were retained since they did not share any metadata (BioSample accession id, primary BioProject accession id, submitter name/affiliation). With the exception of cluster 2 (see text), for each cluster, one accession was retained, while the second accession was either retained (✓) or excluded as a duplicate (×), based on manual investigation. Text in the table details reasons for retention/exclusion. For additional metadata on the accessions, see Appendix D: Table 1B.

Acknowledgements: Sincere thanks to the following researchers for their correspondences regarding identical plasmid sequences: Genevieve Labbé and Roger Johnson (Cluster 4 and 7); Douglas Scott Merrell (Cluster 5).

Table S2. A subset of the clusters of identical accessions, which were deduplicated automatically, and later investigated manually

Cluster accessions (RefSeq/GenBank duplicates excluded)	BioSample accession id	Primary BioProject accession id	Submitter name	Affiliation name	Plausibly non-duplicate
Cluster 9					
MH785255.1	SAMN09846914	486725	Michal Bukowski	Jagiellonian University	✓
MH785230.1	SAMN09846907	486718	Michal Bukowski	Jagiellonian University	
Accessions MH785255.1 and MH785230.1 are from different strains (*Staphylococcus aureus* strains tu1 and ch8, respectively) (Bukowski et al., 2019). Whether they represent independent *de novo* assemblies is unclear.					
Cluster 10					
NZ_CP018678.1	SAMN05362953	328023	-	JCVI	×
NZ_CP007713.1	SAMN02709859	242902	Hao Xu	California State University, Los Angeles	
The associated publication (Ou et al., 2015) states that "the LAC-4 genome consists of a circular chromosome of 3,954,354 base pairs and two circular plasmids, one with 8,006 base pairs while the other with 6,076 base pairs." Accessions NZ_CP018678.1 and NZ_CP007713.1 are both 8,006 bp and the recorded strain is "LAC4" and					

"LAC-4" respectively. My interpretation is that the accessions are redundant and result from re-uploading the same plasmid sequence with a slightly different strain name.					
Cluster 11					
NZ_CP011933.1	SAMN03780437	287300	Kazuhito SATOU	Okinawa Institute of Advanced Sciences	×
NZ_CP011936.1	SAMN03780438	287301	Kazuhito SATOU	Okinawa Institute of Advanced Sciences	
NZ_CP011933.1 and NZ_CP011936.1 (*Leptospira interrogans* serovar Manilae, strains UP-MMC-NIID LP and UP-MMC-NIID HP) are laboratory derivatives (Low and high passage [LP/HP]) of the same ancestral strain, as indicated in Satou et al. (2015): "*L. interrogans* serovar Manilae strain UP-MMC-NIID examined in this study had originally been isolated from the blood of a patient with severe leptospirosis. The virulent and avirulent variants were derived by serial subculture after 1 (low) and 67 (high) passages, respectively."					
Cluster 12					
NZ_CP010765.1	SAMN03294493	273605	Boyke Bunk	Leibniz Institute DSMZ	✓
NZ_CP010607.1	SAMN03294494	273606	Boyke Bunk	Leibniz Institute DSMZ	
NZ_CP010765.1 and NZ_CP010607.1 appear to be from distinct strains (*Phaeobacter inhibens* strains P80 and P83, respectively), isolated from the same location in Spain (Freese et al., 2017).					
Cluster 13					
NC_017720.1	SAMEA3138382	50407	-	EBI	×
NC_016858.1	SAMN02602988	56087	-	NCBI	
These accessions represent a parental strain and its derivative, as indicated in Kröger et al., 2012: "Bacterial strain S. enterica serovar Typhimurium SL1344 [accession NC_017720.1] and its parental strain ST4/74 [accession NC_016858.1] were used throughout the study".					
Cluster 14					
NC_017151.1	SAMD00060949	31141	-	DDBJ	×
NC_017135.1	SAMD00060948	31139	-	DDBJ	
NC_017147.1	SAMD00060947	31137	-	DDBJ	
NC_017127.1	SAMD00060946	31135	-	DDBJ	
NC_017110.1	SAMD00060945	31133	-	DDBJ	
NC_017119.1	SAMD00060944	31131	-	DDBJ	
NC_017114.1	SAMD00061107	32203	-	DDBJ	
NC_013212.1	SAMD00060943	31129	-	DDBJ	
These accessions are mutant derivatives from an experimental genome evolution study (Azuma et al., 2009).					
Cluster 15					
NZ_CP010759.1	SAMN03294493	273605	Boyke Bunk	Leibniz Institute DSMZ	✓
NZ_CP010602.1	SAMN03294494	273606	Boyke Bunk	Leibniz Institute DSMZ	

These accessions are plasmids from *Phaeobacter inhibens* strains P80 and P83, respectively (c.f. cluster 12).					
Cluster 16					
NC_022605.1	SAMN02370325	222409	Feng-Jui Chen	National Health Research Institutes	✓
NC_017332.1	SAMEA2272282	36647	-	EBI	
Accessions NC_022605.1 and NC_017332.1 are from distinct strains of *Staphylococcus aureus* (strains Z172 and TW20, respectively) isolated from Taiwan and England, respectively (Holden et al., 2010; Chen et al., 2013).					

The eight clusters were automatically deduplicated to leave only one accession per cluster, because the accessions all shared metadata and/or were missing metadata in the submitter name/submitter affiliation fields (see Methods). The table shows clusters prior to automated deduplication; only accessions with differing BioSample and GenBank BioProject accession ids (i.e. excluding cognate GenBank accessions) are shown. Since BioProjects are intended to correspond to sequencing project (as per NCBI documentation), it would be reasonable to expect that identical sequences with different BioProject accessions might be non-redundant. Following manual investigation, cluster accessions were assigned as plausibly non-duplicate (✓) or duplicate (✗). In contrast to clusters 1–8, submitters were not contacted to confirm whether or not plausibly non-duplicate cluster accessions were indeed non-duplicates. For additional metadata on the accessions, see Appendix D: Table 1C.

Table S3. The top 10 most frequent taxa (at species-level) within each factor level of the host taxonomy explanatory variable (based on the dataset of 14143 plasmids)

Enterobacteriaceae		**Proteobacteria**		**Firmicutes**		**other**	
Species	Freq	Species	Freq	Species	Freq	Species	Freq
Escherichia coli	1943	*Acinetobacter baumannii*	245	*Staphylococcus aureus*	316	*Borreliella burgdorferi*	285
Klebsiella pneumoniae	1103	*Yersinia pestis*	116	*Bacillus thuringiensis*	306	uncultured bacterium	267
Salmonella enterica	596	*Xanthomonas citri*	80	*Lactobacillus plantarum*	251	*Borreliella afzelii*	54
Enterobacter cloacae	136	*Helicobacter pylori*	68	*Enterococcus faecium*	159	*Borreliella garinii*	33
Citrobacter freundii	115	*Pseudomonas aeruginosa*	64	*Lactococcus lactis*	151	*Rhodococcus hoagii*	29
Shigella sonnei	88	*Phaeobacter inhibens*	62	*Bacillus cereus*	104	*Chlamydia trachomatis*	27
Enterobacter hormaechei	85	*Piscirickettsia salmonis*	57	*Bacillus anthracis*	70	*Bifidobacterium longum*	27
Klebsiella oxytoca	64	*Zymomonas mobilis*	51	*Staphylococcus epidermidis*	65	*Mycobacterium chimaera*	21
Klebsiella aerogenes	41	*Serratia marcescens*	49	*Enterococcus faecalis*	60	*Corynebacterium glutamicum*	21
Shigella flexneri	40	*Rhizobium leguminosarum*	49	*Clostridium botulinum*	60	*Salinibacter ruber*	20

Table S4. The top 10 most frequent isolation source sub-categories within each factor level of the isolation source explanatory variable (based on the dataset of 14143 plasmids)

human		**livestock**		**other**	
Sub-category	Freq	Sub-category	Freq	Sub-category	Freq
human	2645	aquaculture	192	-	9858
sewage*	309	cow	156	agriculture**	623
		chicken	155		
		pig	135		
		turkey	26		
		sheep	16		
		poultry	13		
		goat	6		
		goose	4		
		duck	3		

*Sewage sub-category was manually curated with the aim of including only human-derived wastewater

**Non-livestock agriculture.

"-" indicates isolation source information could not be extracted from BioSample metadata (either information was missing, or it could not be assigned to a curated category).

Table S5. The top 10 most frequent countries within each factor-level of the geographic location explanatory variable (based on the dataset of 14143 plasmids)

high-income		middle-income		EU		China	
Country	Freq	Country	Freq	Country	Freq	Country	Freq
South Korea	552	Mexico	150	Germany	328	China	1248
Japan	355	India	147	United Kingdom	157		
Canada	207	Russia	107	Spain	129		
Australia	115	Brazil	101	France	121		
Switzerland	99	Vietnam	51	Netherlands	85		
Chile	66	Thailand	49	Denmark	79		
Argentina	62	Malaysia	37	Sweden	66		
Norway	60	Colombia	30	Italy	53		
Hong Kong	42	South Africa	24	Finland	32		
New Zealand	32	Nigeria	24	Ireland	29		

United States		other	
Country	Freq	Country	Freq
United States	1491	-	7284
		Taiwan	93
		Antarctica	58
		Réunion	10
		Nepal	9
		Raas Cabaad, Somalia	8
		Tanzania	7
		Syria	6
		Rwanda	6
		The Gambia	5

"-" indicates geographic location information could not be extracted from BioSample metadata.
"high-income": World Bank high-income countries (not included in other categories).
"middle-income": World Bank lower-middle income countries (n = 362) and World Bank upper-middle income countries (n = 593) combined (not included in other categories).
"EU": the EU28 countries (includes the United Kingdom).
"other": includes plasmids with missing location data, missing World Bank income categorisation, or rare income category (specifically, World Bank low-income countries, n = 70).

5.6.3 Supplementary Figures

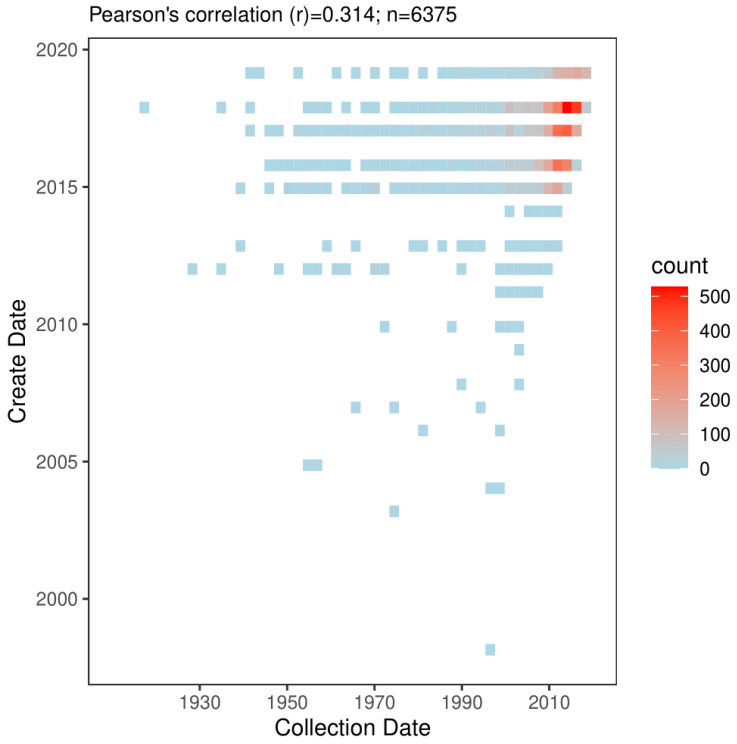


Figure S1. Density scatter plot showing the relationship between collection date and create date. To handle overplotting, the datapoints were binned; the density of points within a bin is indicated using a blue-red colour gradient.

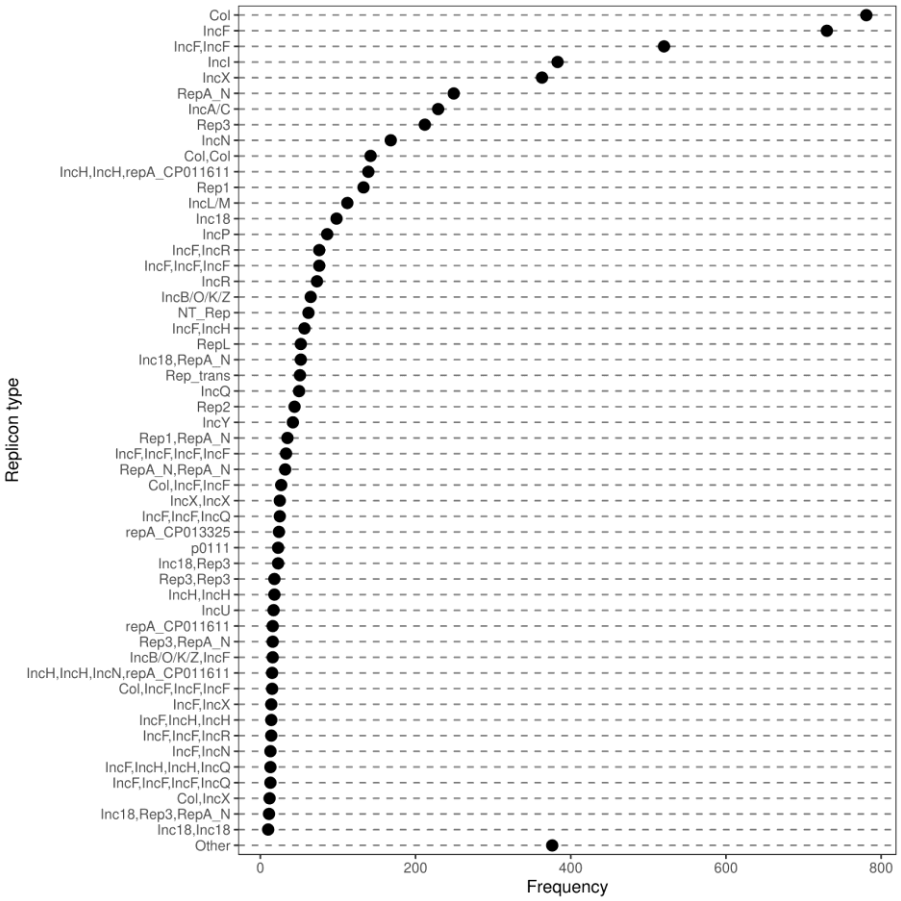


Figure S2. Cleveland dotplot showing the frequency of plasmid replicon types in the dataset of 14143 plasmids. Replicon types represented by fewer than 10 plasmids are categorised as "other". 8230 plasmids were not assigned a replicon type.

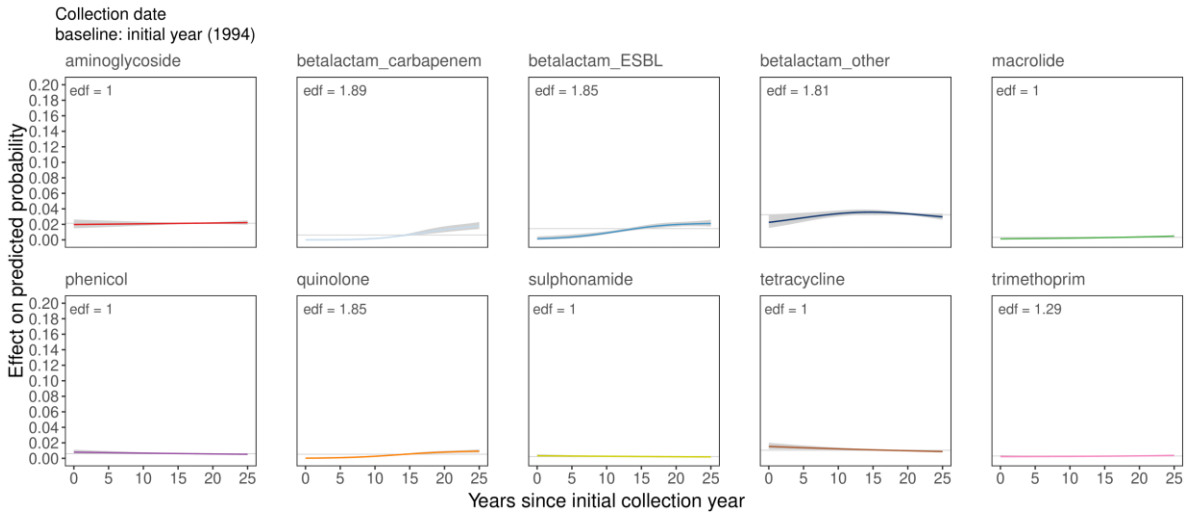
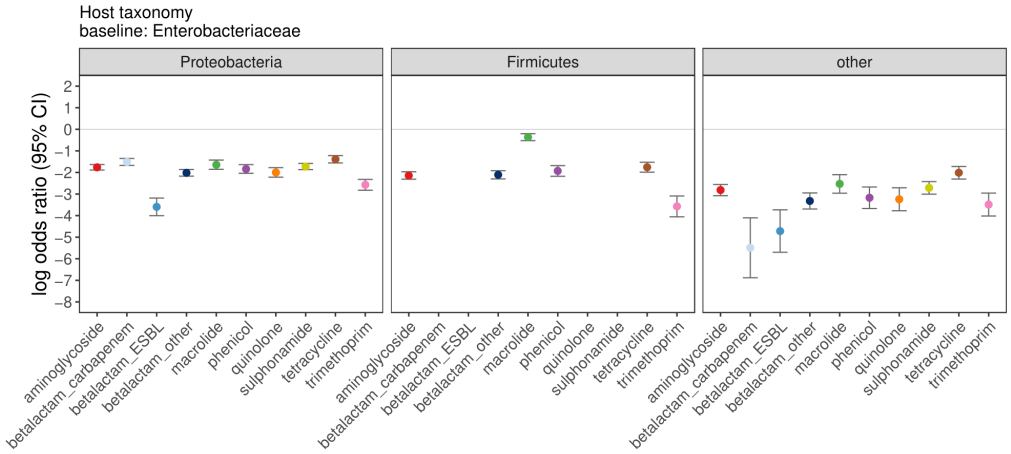
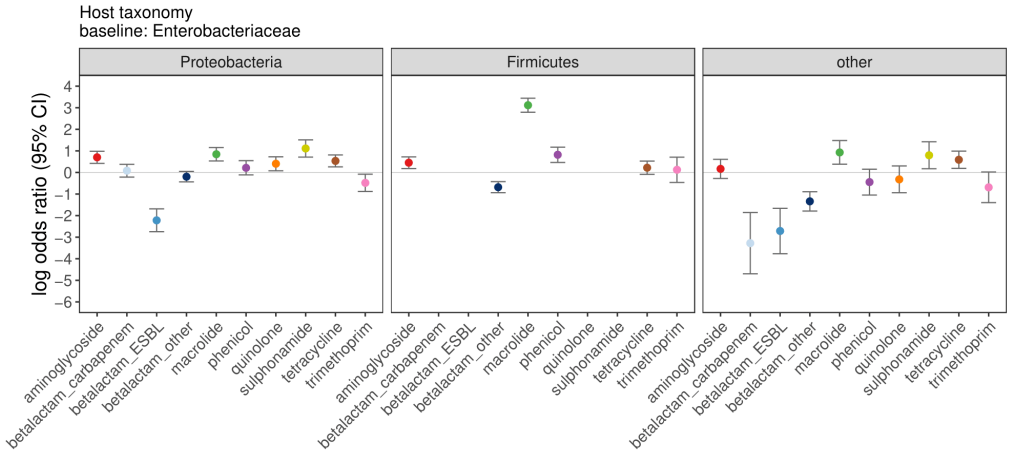


Figure S3. Multivariate-adjusted association between collection date (expressed as years since baseline collection year, 1994) and the predicted probability of antibiotic resistance carriage (y-axis), across 10 resistance classes. Shaded areas around the smooths show 95% confidence intervals. The predicted probability scale is centred on the model intercept, so the smooths can be interpreted as showing the probability if all other explanatory variables are at their reference value.

A Unadjusted



B Adjusted



C Adjusted

- replicon type multiplicity
- other resistance gene classes

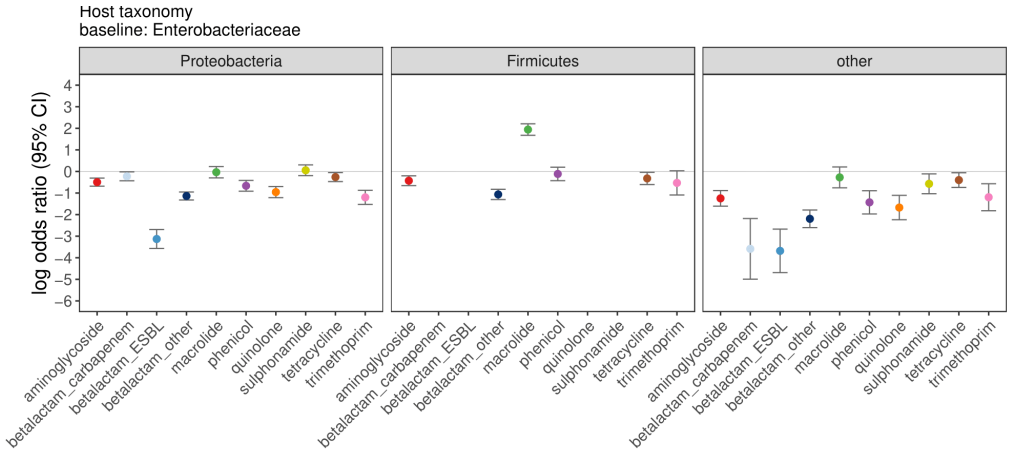


Figure S4. The association between host taxonomy (with Enterobacteriaceae as the baseline factor level) and the log-odds of antibiotic resistance carriage (y-axis), compared across 3 different analyses. A) log-odds ratios from the unadjusted analysis. B) log-odds ratios from the main adjusted model (the full model, as presented in the main text, Figure 3). C) log-odds ratios from an adjusted model which omitted the following two explanatory variables: replicon type multiplicity, and other resistance gene classes. Error bars show 95% confidence intervals.

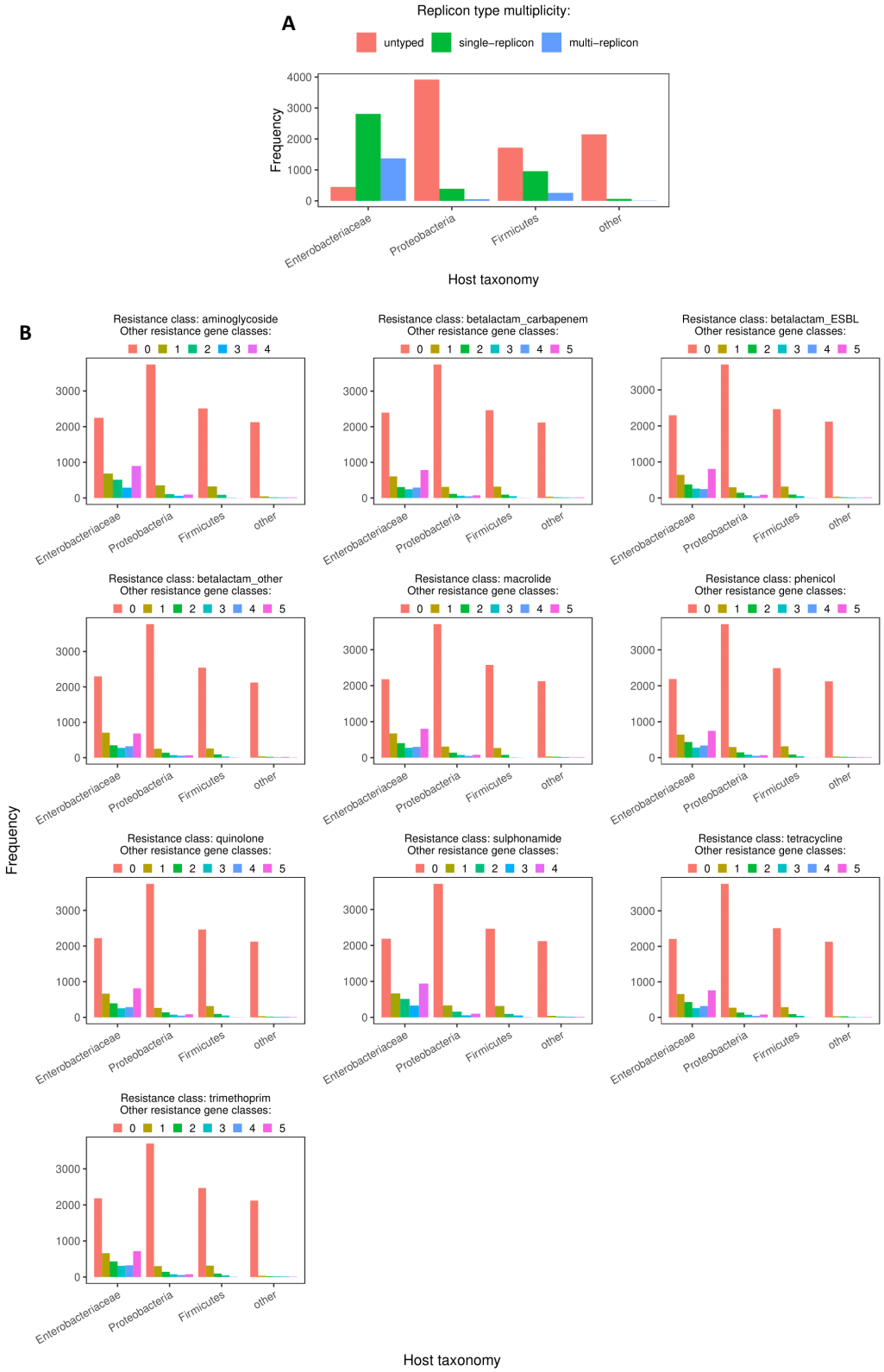


Figure S5. The association between host taxonomy and A) replicon type multiplicity (untyped, single-replicon, multi-replicon); B) for a given resistance outcome class, the number of other antibiotic resistance gene classes encoded on the same plasmid (across the 10 resistance class outcomes). Enterobacteriaceae plasmids tend to encode one or more replicon loci (single/multi replicon carriage) whereas other taxa are most frequently untyped. Enterobacteriaceae plasmids encoding a resistance gene from a given class more frequently encode one or more resistance genes from other classes, in comparison with plasmids from other taxa.

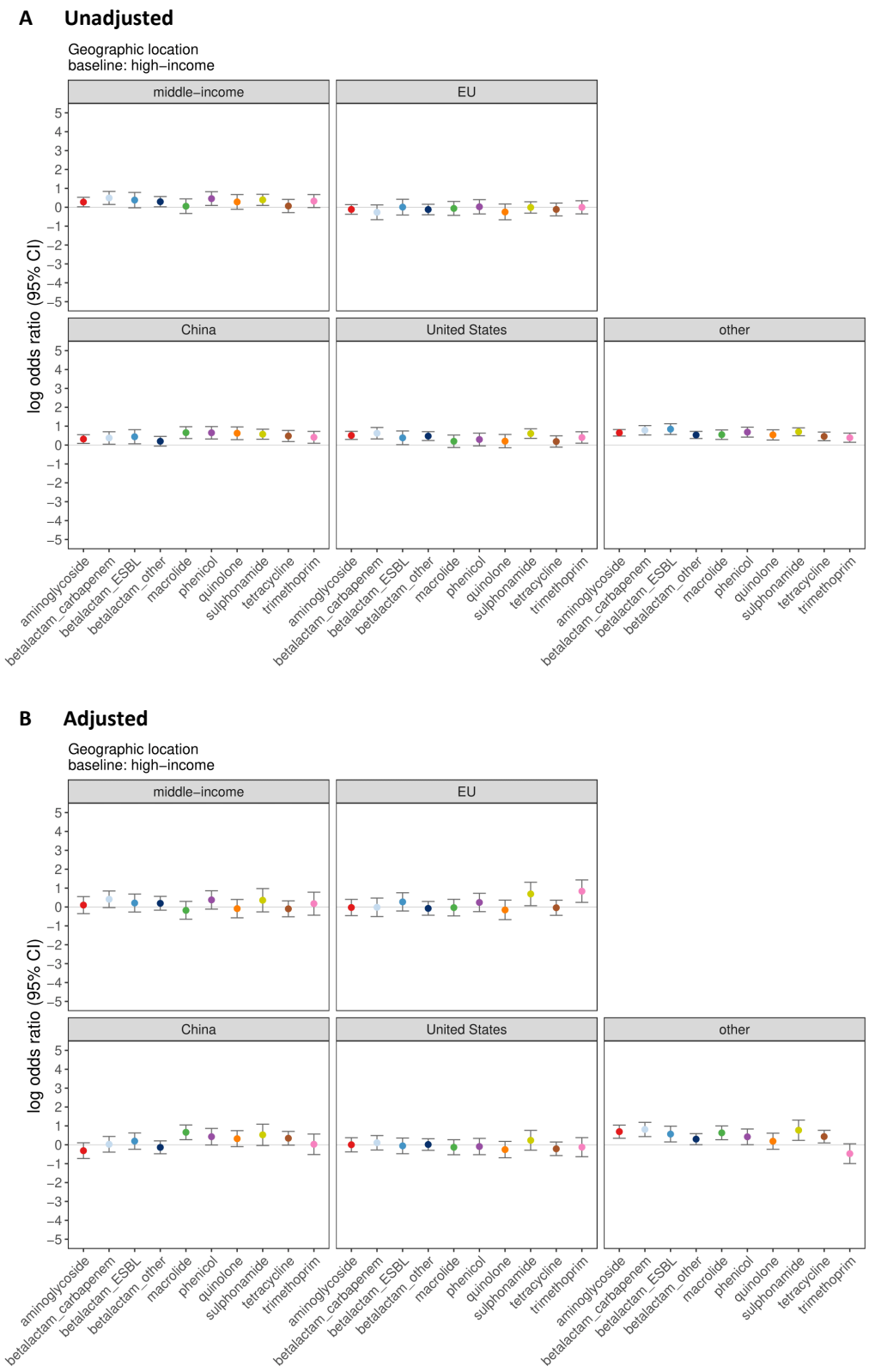


Figure S6. The association between geographic location and the log-odds of antibiotic resistance carriage (y-axis), across 10 resistance classes. A) log-odds ratios from the unadjusted analysis. B) log-odds ratios from the adjusted analysis. Log-odds ratios indicate the effect of a given factor level, relative to baseline (high-income), and error bars show 95% confidence intervals.

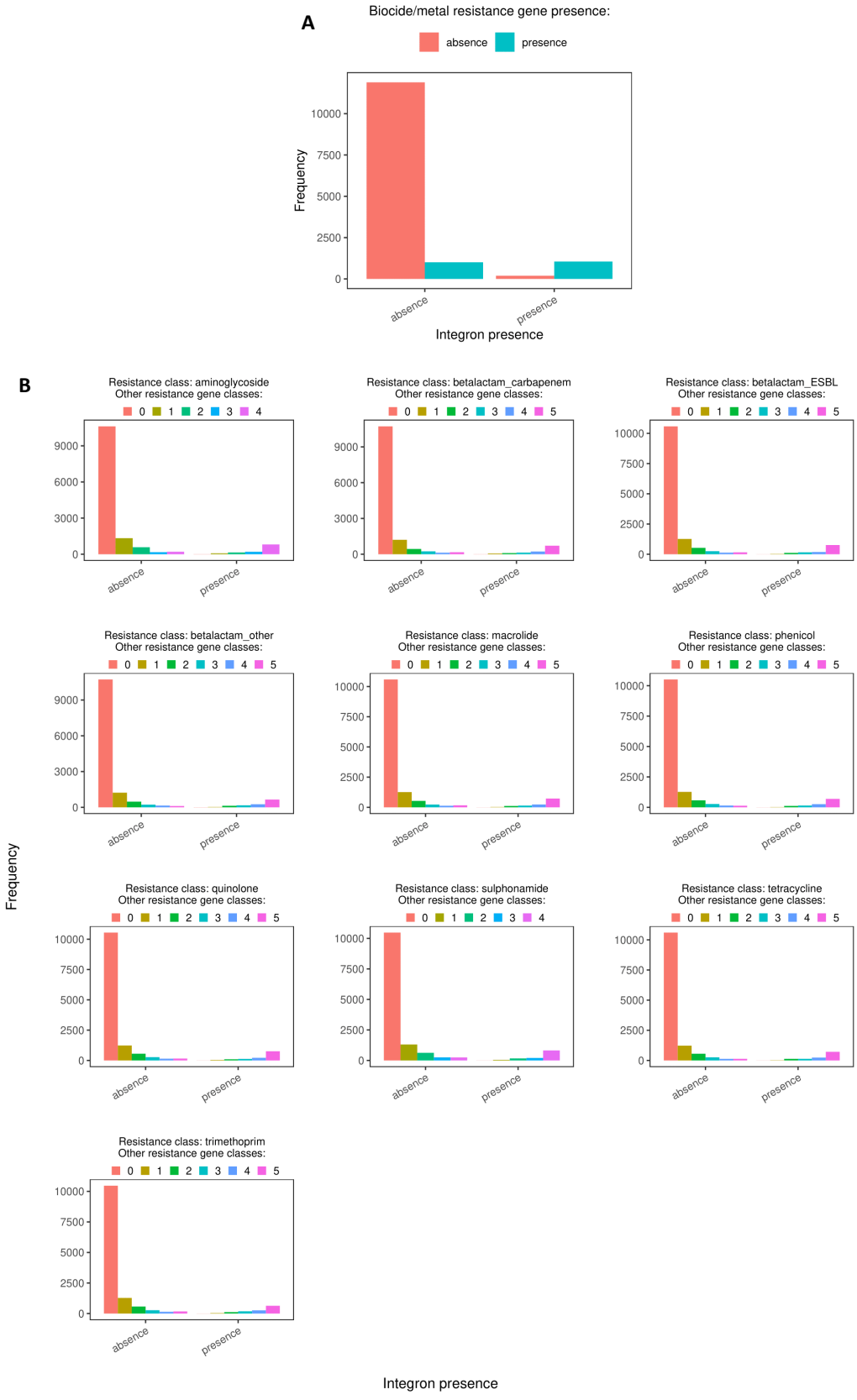


Figure S7. Continued overleaf

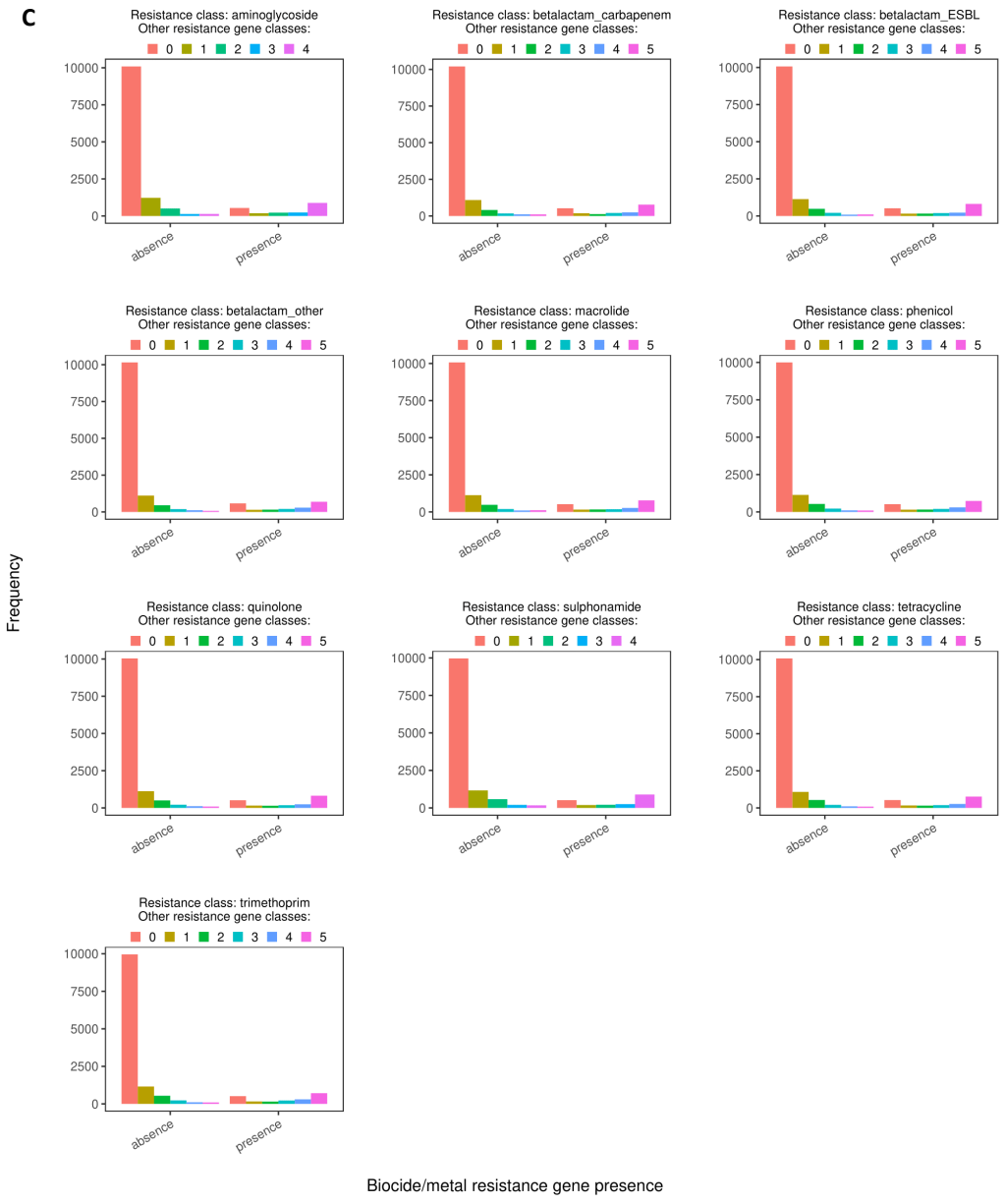


Figure S7. The association between A) presence of integrons and presence of biocide/metal resistance genes; B) for a given resistance class outcome, the presence of integrons and the number of other resistance gene classes encoded on the same plasmid (across 10 resistance class outcomes); C) for a given resistance class outcome, the presence of biocide/metal resistance genes and the number of other resistance gene classes encoded on the same plasmid (across 10 resistance class outcomes). Positive co-associations are found between all three explanatory variables (integron presence, biocide/metal resistance gene presence, the number of other resistance gene classes).

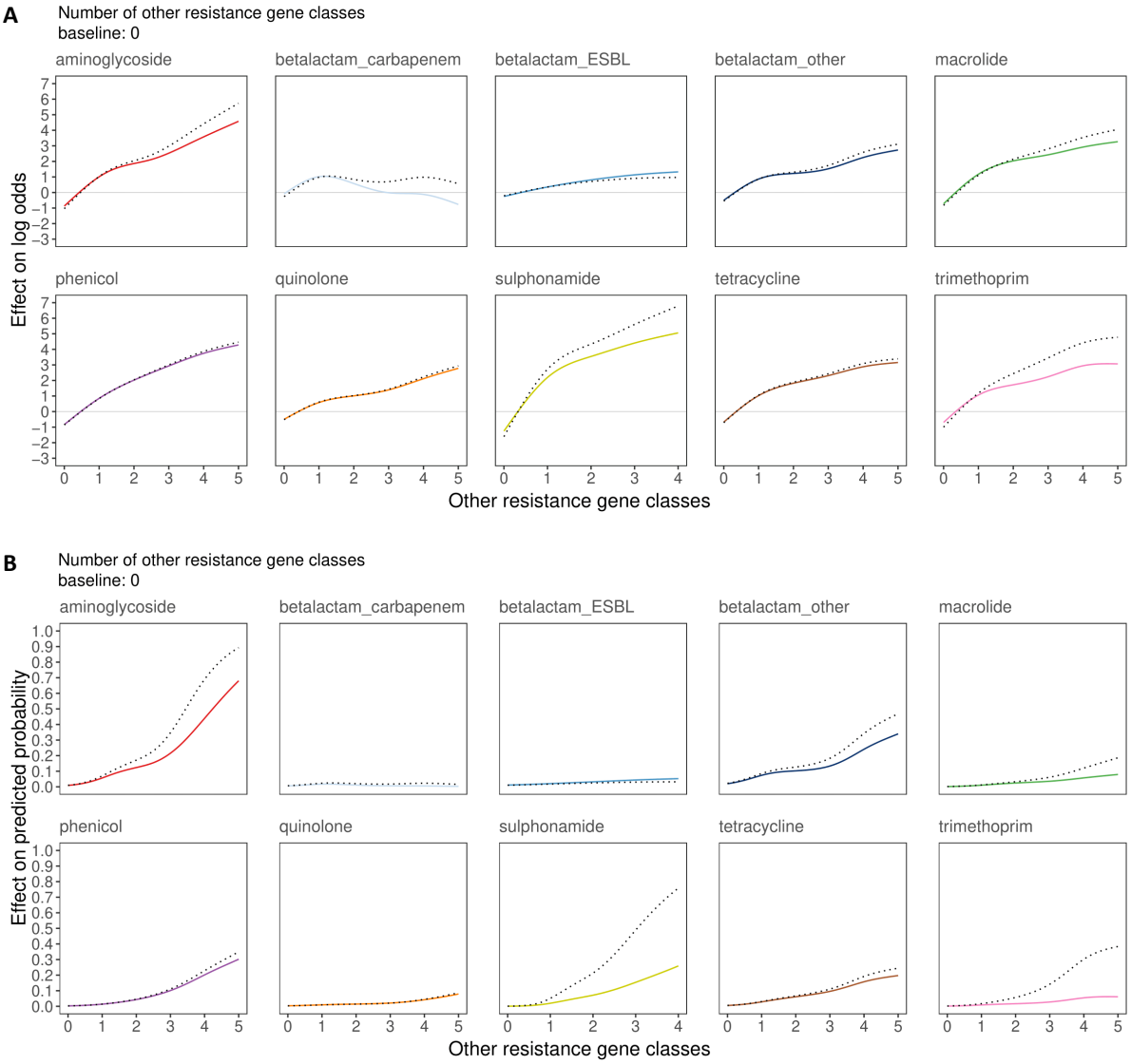


Figure S8. The association between the number of other resistance gene classes and A) the log-odds B) the predicted probability of antibiotic resistance carriage (y-axis), with all other factors at their reference value. Smooths displayed with solid, coloured lines are from the main adjusted model (the full model, as presented in the main text, Figure 4). Smooths displayed with dotted, black lines are from an adjusted model which omitted the following two explanatory variables: integron presence, and biocide/metal resistance gene presence. The predicted probability scale is centred on the model intercept, so the smooths can be interpreted as showing the probability if all other explanatory variables are at their reference value.

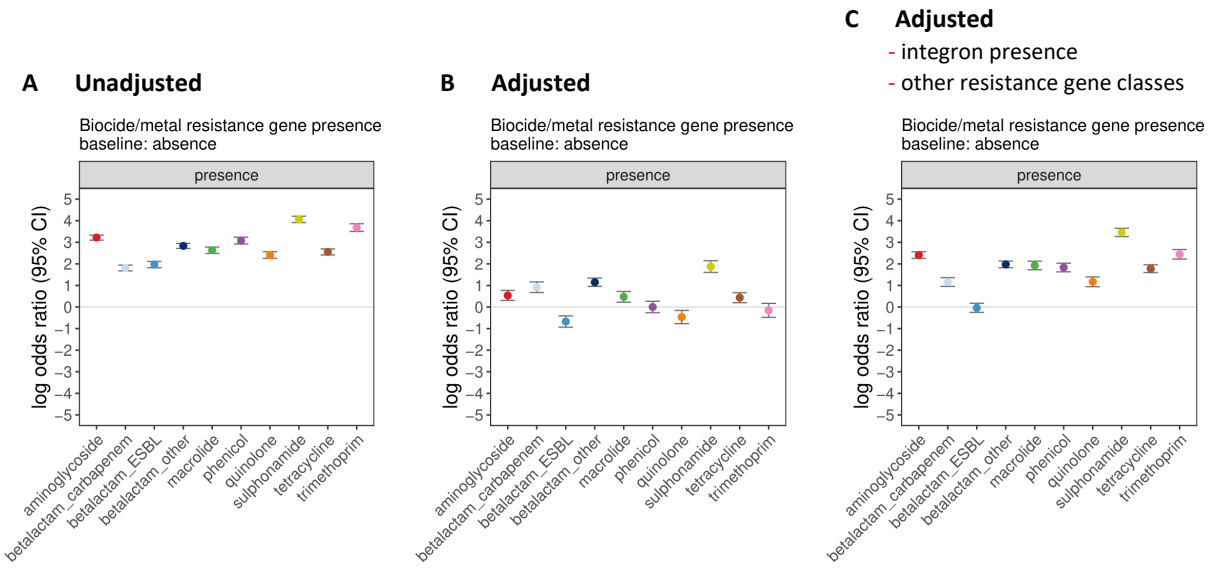


Figure S9. The association between biocide/metal resistance gene presence (vs absence) and the log-odds of antibiotic resistance carriage (y-axis), compared across 3 different analyses. A) log-odds ratios from the unadjusted analysis. B) log-odds ratios from the main adjusted model (the full model, as presented in the main text, Figure 4). C) log-odds ratios from an adjusted model which omitted the following two explanatory variables: integron presence, and other resistance gene classes. Error bars show 95% confidence intervals.

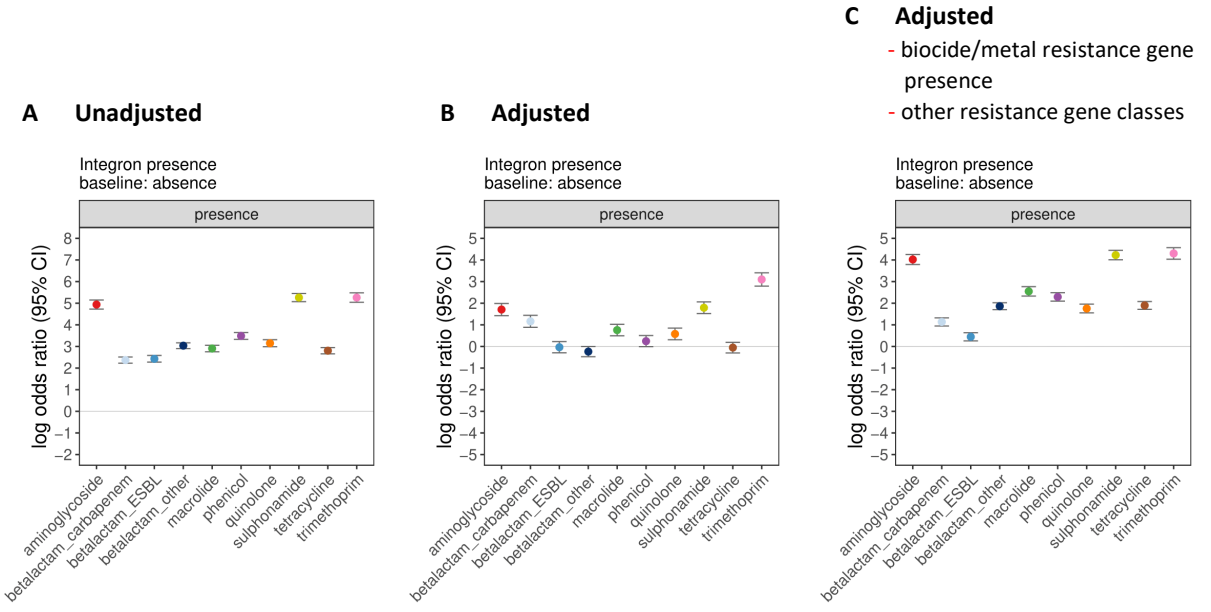


Figure S10. The association between integron presence (vs absence) and the log-odds of antibiotic resistance carriage (y-axis), compared across 3 different analyses. A) log-odds ratios from the unadjusted analysis. B) log-odds ratios from the main adjusted model (the full model, as presented in the main text, Figure 4). C) log-odds ratios from an adjusted model which omitted the following two explanatory variables: biocide/metal resistance gene presence, and other resistance gene classes. Error bars show 95% confidence intervals.

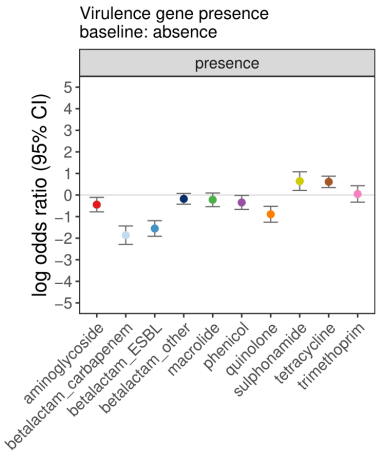
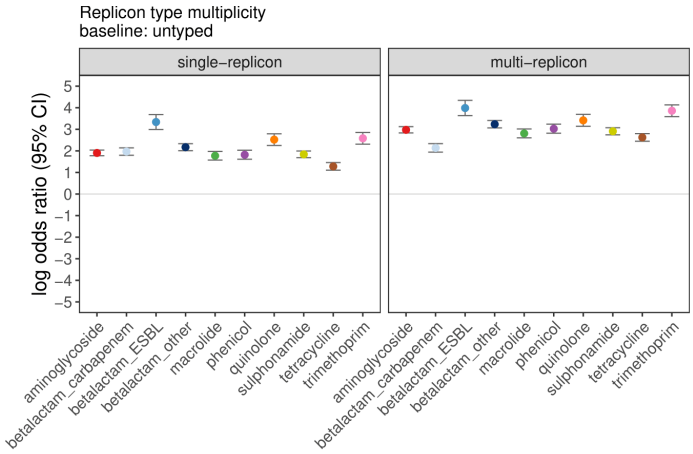
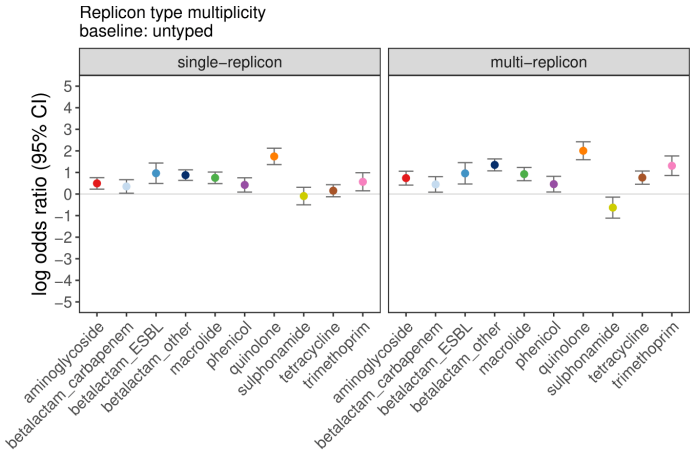


Figure S11. The multivariate-adjusted association between virulence presence (vs absence) and the log-odds of antibiotic resistance carriage (y-axis), across 10 resistance classes. Log-odds ratios indicate the effect of virulence presence relative to baseline (absence), and error bars show 95% confidence intervals.

A Unadjusted



B Adjusted



C Adjusted

- biocide/metal resistance gene presence
- other resistance gene classes

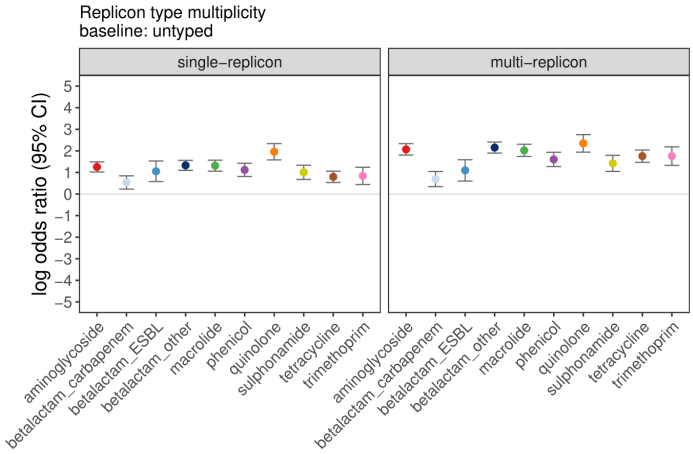


Figure S12. The association between replicon type multiplicity (untyped [reference], single-replicon, multi-replicon type plasmid) and the log-odds of antibiotic resistance carriage (y-axis), compared across three different analyses. A) log-odds ratios from the unadjusted analysis. B) log-odds ratios from the adjusted model (the full model). C) log-odds from an adjusted model which omitted the following two variables: biocide/metal resistance gene presence, and other resistance gene classes. Error bars show 95% confidence intervals.

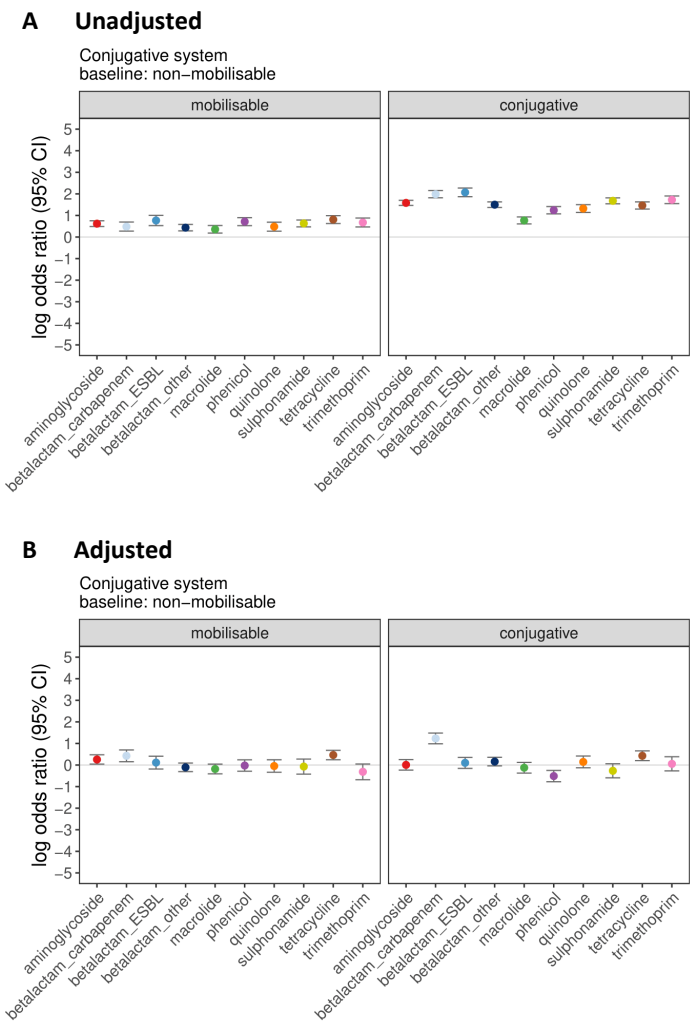


Figure S13. The association between conjugative system (non-mobilisable [reference], mobilisable, conjugative) and the log-odds of antibiotic resistance carriage (y-axis). A) log-odds ratios from the unadjusted analysis. B) log-odds ratios from the adjusted model (the full model). Error bars show 95% confidence intervals.

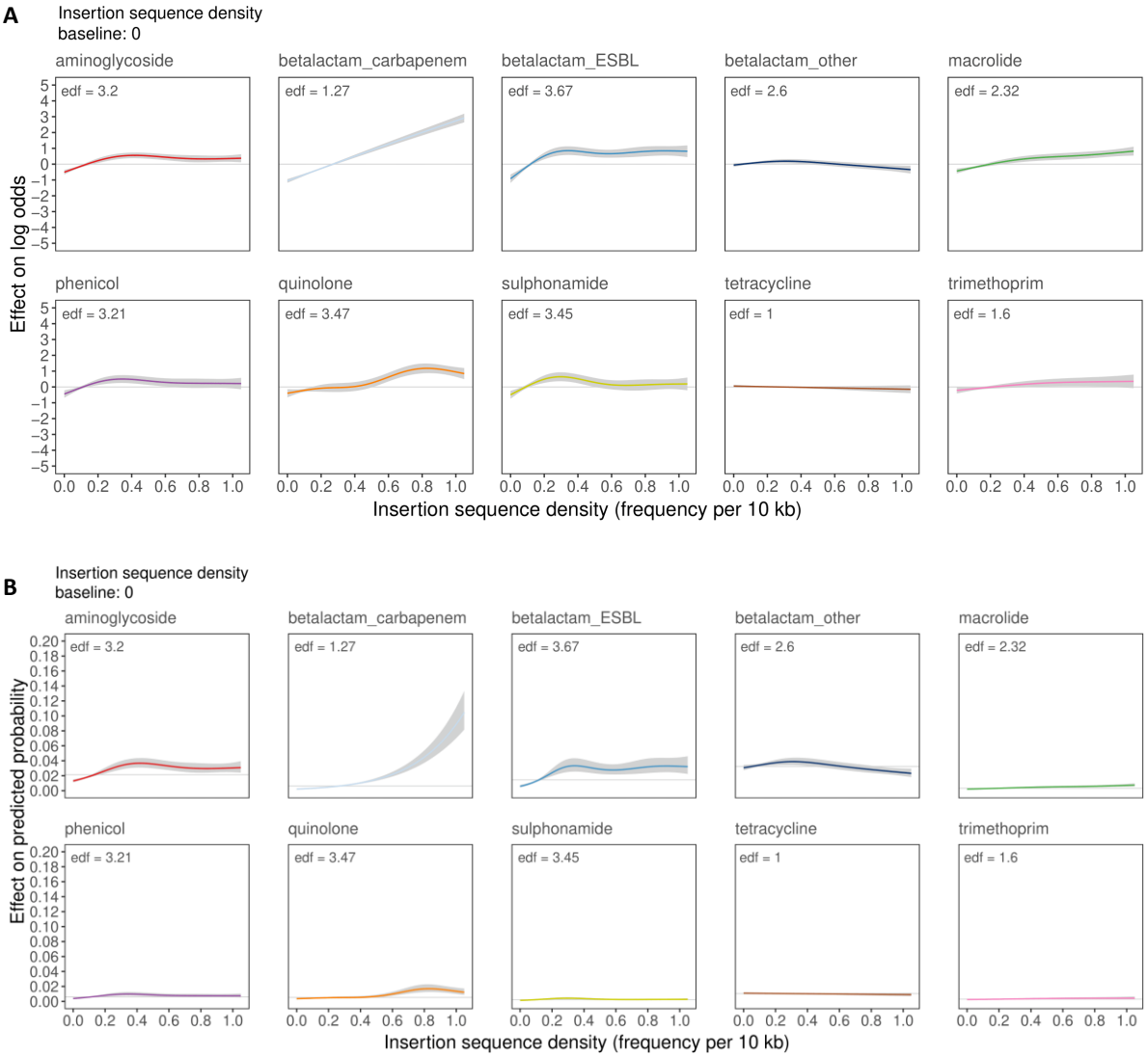


Figure S14. The multivariate-adjusted association between the number of insertion sequences and A) the log-odds B) the predicted probability of antibiotic resistance carriage (y-axis). The predicted probability scale is centred on the model intercept, so the smooths can be interpreted as showing the probability if all other explanatory variables are at their reference value.

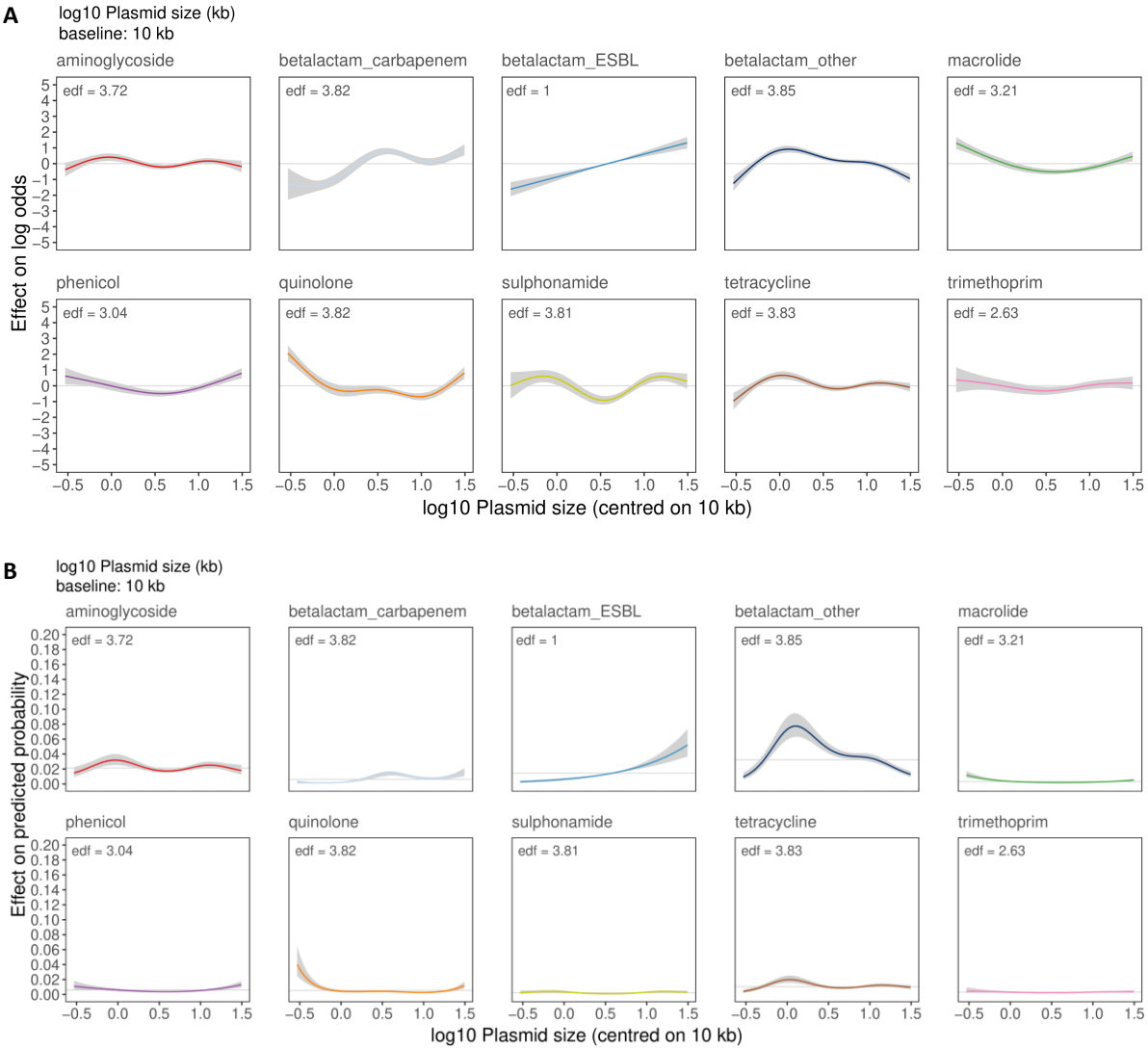


Figure S15. The multivariate-adjusted association between plasmid size (log10-transformed and centred on 10 kb) and A) the log-odds B) the predicted probability of antibiotic resistance carriage (y-axis). The predicted probability scale is centred on the model intercept, so the smooths can be interpreted as showing the probability if all other explanatory variables are at their reference value.

Bibliography

Aarestrup, F. M., and Woolhouse, M. E. J. (2020). Using sewage for surveillance of antimicrobial resistance. *Science (80-.).* doi:10.1126/science.aba3432.

Acman, M., Wang, R., Van Dorp, L., Shaw, L. P., Wang, Q., Luhmann, N., et al. (2021). Role of the mobilome in the global dissemination of the carbapenem resistance gene bla NDM 2. *bioRxiv*.

Alvarez-Kalverkamp, M., Bayer, W., Becheva, S., Benning, R., Börnecke, S., Chemnitz, C., et al. (2014). *Meat atlas, Facts and figures about the animals we eat*.

Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., et al. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. genomics*. doi:10.1099/mgen.0.000093.

Azuma, Y., Hosoyama, A., Matsutani, M., Furuya, N., Horikawa, H., Harada, T., et al. (2009). Whole-genome analyses reveal genetic instability of Acetobacter pasteurianus. *Nucleic Acids Res.* doi:10.1093/nar/gkp612.

Baker-Austin, C., Wright, M. S., Stepanauskas, R., and McArthur, J. V. (2006). Co-selection of antibiotic and metal resistance. *Trends Microbiol.* doi:10.1016/j.tim.2006.02.006.

Baquero, F., and Canton, R. (2017). "Chapter 2: Evolutionary Biology of Drug Resistance," in *Antimicrobial Drug Resistance: Mechanisms of Drug Resistance, Volume 1*, eds. D. L. Mayers, J. D. Sobel, M. Ouellette, K. S. Kaye, and D. Marchaim (Cham, Switzerland: Springer), 9–36.

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., et al. (2012). BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Res.* 40. doi:10.1093/nar/gkr1163.

Bean, D. C., Livermore, D. M., and Hall, L. M. C. (2009). Plasmids imparting sulfonamide resistance in Escherichia coli: Implications for persistence. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.00800-08.

Beceiro, A., Tomás, M., and Bou, G. (2013). Antimicrobial resistance and virulence: A successful or deleterious association in the bacterial world? *Clin. Microbiol. Rev.* 26. doi:10.1128/CMR.00059-12.

Bell, B. G., Schellevis, F., Stobberingh, E., Goossens, H., and Pringle, M. (2014). A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infect. Dis.* 14. doi:10.1186/1471-2334-14-13.

Bengtsson-Palme, J., Hammarén, R., Pal, C., Östman, M., Björlenius, B., Flach, C. F., et al. (2016). Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci. Total Environ.* doi:10.1016/j.scitotenv.2016.06.228.

Bolotin, A., Gillis, A., Sanchis, V., Nielsen-LeRoux, C., Mahillon, J., Lereclus, D., et al. (2017). Comparative genomics of extrachromosomal elements in Bacillus thuringiensis subsp. israelensis. *Res. Microbiol.* doi:10.1016/j.resmic.2016.10.008.

Bonnet, M., Lagier, J. C., Raoult, D., and Khelaifia, S. (2020). Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. *New Microbes New Infect.* 34. doi:10.1016/j.nmni.2019.100622.

Brandt, C., Viehweger, A., Singh, A., Pletz, M. W., Wibberg, D., Kalinowski, J., et al. (2019). Assessing genetic diversity and similarity of 435 KPC-carrying plasmids. *Sci. Rep.* doi:10.1038/s41598-019-

47758-5.

Bukowski, M., Piwowarczyk, R., Madry, A., Zagorski-Przybylo, R., Hydzik, M., and Wladyka, B. (2019). Prevalence of Antibiotic and Heavy Metal Resistance Determinants and Virulence-Related Genetic Elements in Plasmids of Staphylococcus aureus. *Front. Microbiol.* doi:10.3389/fmicb.2019.00805.

Bush, K. (2018). Past and present perspectives on β-lactamases. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.01076-18.

Campos, F. S., Cerqueira, F. B., Santos, G. R., Pereira, E. J. G., Corrêia, R. F. T., Cangussu, A. S. R., et al. (2019). Complete Sequences of Two Plasmids Found in a Brazilian Bacillus thuringiensis Serovar israelensis Strain . *Microbiol. Resour. Announc.* doi:10.1128/mra.00051-19.

Carattoli, A. (2009). Resistance plasmid families in Enterobacteriaceae. *Antimicrob. Agents Chemother.* 53, 2227–2238. doi:10.1128/AAC.01707-08.

Carattoli, A. (2013). Plasmids and the spread of resistance. *Int. J. Med. Microbiol.* 303, 298–304. doi:10.1016/j.ijmm.2013.02.001.

Carattoli, A., and Hasman, H. (2020). "PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS)," in *Methods in Molecular Biology* doi:10.1007/978-1-4939-9877-7_20.

Chen, F. J., Lauderdale, T. L., Wang, L. S., and Huang, I. W. (2013). Complete genome sequence of Staphylococcus aureus Z172, a vancomycin-intermediate and daptomycinnonsusceptible methicillin-resistant strain isolated in Taiwan. *Genome Announc.* doi:10.1128/genomeA.01011-13.

Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). VFDB 2016: Hierarchical and refined dataset for big data analysis - 10 years on. *Nucleic Acids Res.* doi:10.1093/nar/gkv1239.

Chen, Q., Zobel, J., and Verspoor, K. (2017). Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study. *Database*. doi:10.1093/database/baw163.

Colegrave, N., and Ruxton, G. D. (2018). Using Biological Insight and Pragmatism When Thinking about Pseudoreplication. *Trends Ecol. Evol.* 33, 28–35. doi:10.1016/j.tree.2017.10.007.

Collignon, P., Athukorala, P. C., Senanayake, S., and Khan, F. (2015). Antimicrobial resistance: The major contribution of poor governance and corruption to this growing problem. *PLoS One*. doi:10.1371/journal.pone.0116746.

Collignon, P., and Beggs, J. J. (2019). Socioeconomic enablers for contagion: Factors impelling the antimicrobial resistance epidemic. *Antibiotics*. doi:10.3390/antibiotics8030086.

Collignon, P., Beggs, J. J., Walsh, T. R., Gandra, S., and Laxminarayan, R. (2018). Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. *Lancet Planet. Heal.* doi:10.1016/S2542-5196(18)30186-4.

Cottell, J. L., Saw, H. T. H., Webber, M. A., and Piddock, L. J. V. (2014). Functional genomics to identify the factors contributing to successful persistence and global spread of an antibiotic resistance plasmid. *BMC Microbiol.* doi:10.1186/1471-2180-14-168.

Cury, J., Abby, S. S., Doppelt-Azeroual, O., Neron, B., and Rocha, E. P. C. (2020). "Chapter 19: Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan," in *Horizontal Gene Transfer: Methods and Protocols*, ed. F. de la Cruz (New York, NY: Humana Press), 265–283.

Cury, J., Jové, T., Touchon, M., Néron, B., and Rocha, E. P. (2016). Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* doi:10.1093/nar/gkw319.

Cury, J., Touchon, M., and Rocha, E. P. C. (2017). Integrative and conjugative elements and their hosts: Composition, distribution and organization. *Nucleic Acids Res.* doi:10.1093/nar/gkx607.

D'Andrea, M. M., Arena, F., Pallecchi, L., and Rossolini, G. M. (2013). CTX-M-type β-lactamases: A successful story of antibiotic resistance. *Int. J. Med. Microbiol.* doi:10.1016/j.ijmm.2013.02.008.

Davies, J., and Davies, D. (2010). Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.* 74, 417–433. doi:10.1128/MMBR.00016.

Davies, J. E. (1983). Resistance to Aminoglycosides: Mechanisms and Frequency. *Clin. Infect. Dis.* 5, S261–S267. doi:10.1093/clinids/5.supplement_2.s261.

Dietrich, M. R., Ankeny, R. A., and Chen, P. M. (2014). Publication trends in model organism research. *Genetics*. doi:10.1534/genetics.114.169714.

Domínguez, M., Miranda, C. D., Fuentes, O., De La Fuente, M., Godoy, F. A., Bello-Toledo, H., et al. (2019). Occurrence of transferable integrons and suland dfrgenes among sulfonamide-and/or trimethoprim-resistant bacteria isolated from chilean salmonid farms. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.00748.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography (Cop.).* 36. doi:10.1111/j.1600-0587.2012.07348.x.

Douarre, P. E., Mallet, L., Radomski, N., Felten, A., and Mistou, M. Y. (2020). Analysis of COMPASS, a New Comprehensive Plasmid Database Revealed Prevalence of Multireplicon and Extensive Diversity of IncF Plasmids. *Front. Microbiol.* doi:10.3389/fmicb.2020.00483.

ECDC/EFSA/EMA (2017). ECDC/EFSA/EMA second joint report on the integrated analysis of the consumption of antimicrobial agents and occurrence of antimicrobial resistance in bacteria from humans and food-producing animals: Joint Interagency Antimicrobial Consumption and Resistan. *EFSA J.* doi:10.2903/j.efsa.2017.4872.

Economou, V., and Gousia, P. (2015). Agriculture and food animals as a source of antimicrobial-resistant bacteria. *Infect. Drug Resist.* doi:10.2147/IDR.S55778.

Emmerson, A. M., and Jones, A. M. (2003). The quinolones: Decades of development and use. *J. Antimicrob. Chemother.* doi:10.1093/jac/dkg208.

Enne, V. I., Livermore, D. M., Stephens, P., and Hall, L. M. C. (2001). Persistence of sulphonamide resistance in Escherichia coli in the UK despite national prescribing restriction. *Lancet*. doi:10.1016/S0140-6736(00)04519-0.

FAO (2018). *The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals*.

Fasiolo, M., Nedellec, R., Goude, Y., and Wood, S. N. (2020). Scalable Visualization Methods for Modern Generalized Additive Models. *J. Comput. Graph. Stat.* doi:10.1080/10618600.2019.1629942.

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Res.* doi:10.1093/nar/gkr1178.

Fondi, M., Karkman, A., Tamminen, M., Bosi, E., Virta, M., Fani, R., et al. (2016). Every gene is everywhere but the environment selects : Global geo-localization of gene sharing in

environmental samples through network analysis. *Genome Biol. Evol.* 8, evw077. doi:10.1093/gbe/evw077.

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O. A., and Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. *Science (80-.).* doi:10.1126/science.1220761.

Fortunato, S., and Hric, D. (2016). Community detection in networks: A user guide. *Phys. Rep.* doi:10.1016/j.physrep.2016.09.002.

Foster, T. J. (2017). Antibiotic resistance in Staphylococcus aureus. Current status and future prospects. *FEMS Microbiol. Rev.* doi:10.1093/femsre/fux007.

Freese, H. M., Sikorski, J., Bunk, B., Scheuner, C., Meier-Kolthoff, J. P., Spröer, C., et al. (2017). Trajectories and Drivers of Genome Evolution in Surface-Associated Marine Phaeobacter. *Genome Biol. Evol.* doi:10.1093/gbe/evx249.

Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2019). PLSDB: A resource of complete bacterial plasmids. *Nucleic Acids Res.* doi:10.1093/nar/gky1050.

Galvani, A. P. (2003). Epidemiology meets evolutionary ecology. *Trends Ecol. Evol.* doi:10.1016/S0169-5347(02)00050-2.

García-Meniño, I., Díaz-Jiménez, D., García, V., de Toro, M., Flament-Simon, S. C., Blanco, J., et al. (2019). Genomic Characterization of Prevalent mcr-1, mcr-4, and mcr-5 Escherichia coli Within Swine Enteric Colibacillosis in Spain. *Front. Microbiol.* doi:10.3389/fmicb.2019.02469.

García, V., García-Meniño, I., Mora, A., Flament-Simon, S. C., Díaz-Jiménez, D., Blanco, J. E., et al. (2018). Co-occurrence of mcr-1, mcr-4 and mcr-5 genes in multidrug-resistant ST10 Enterotoxigenic and Shiga toxin-producing Escherichia coli in Spain (2006-2017). *Int. J. Antimicrob. Agents*. doi:10.1016/j.ijantimicag.2018.03.022.

Garcillan-Barcia, M. P., Redondo-Salvo, S., Vielva, L., and de la Cruz, F. (2020). "Chapter 21: MOBscan: Automated Annotation of MOB Relaxases," in *Horizontal Gene Transfer: Methods and Protocols*, ed. F. de la Cruz (New York, NY), 295–308.

Garg, S., Rautiainen, M., Novak, A. M., Garrison, E., Durbin, R., and Marschall, T. (2018). A graph-based approach to diploid genome assembly. in *Bioinformatics* doi:10.1093/bioinformatics/bty279.

Ghafourian, S., Sadeghifard, N., Soheili, S., and Sekawi, Z. (2014). Extended spectrum beta-lactamases: Definition, classification and epidemiology. *Curr. Issues Mol. Biol.* 17, 11–22. doi:10.21775/cimb.017.011.

Ghaly, T. M., Chow, L., Asher, A. J., Waldron, L. S., and Gillings, M. R. (2017). Evolution of class 1 integrons: Mobilization and dispersal via food-borne bacteria. *PLoS One*. doi:10.1371/journal.pone.0179169.

Ghaly, T. M., and Gillings, M. R. (2018). Mobile DNAs as Ecologically and Evolutionarily Independent Units of Life. *Trends Microbiol.* doi:10.1016/j.tim.2018.05.008.

Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T. P., Vanwambeke, S. O., Wint, G. R. W., et al. (2018). Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data*. doi:10.1038/sdata.2018.227.

Gillings, M. R. (2014). Integrons: past, present, and future. *Microbiol. Mol. Biol. Rev.* 78, 257–77. doi:10.1128/mmbr.00056-13.

Golkar, T., Zielinski, M., and Berghuis, A. M. (2018). Look and outlook on enzyme-mediated macrolide resistance. *Front. Microbiol.* doi:10.3389/fmicb.2018.01942.

Gomes, C., Martínez-Puchol, S., Palma, N., Horna, G., Ruiz-Roldán, L., Pons, M. J., et al. (2017). Macrolide resistance mechanisms in Enterobacteriaceae: Focus on azithromycin. *Crit. Rev. Microbiol.* doi:10.3109/1040841X.2015.1136261.

Gonçalves, R. S., and Musen, M. A. (2019). Analysis: The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data*. doi:10.1038/sdata.2019.21.

Google Maps Platform (2020a). Geocoding Service: Address Types and Address Component Types. Available at: https://developers.google.com/maps/documentation/javascript/geocoding#GeocodingAddressTypes [Accessed August 29, 2020].

Google Maps Platform (2020b). Place Details: Place Details Results. Available at: https://developers.google.com/places/web-service/details#PlaceDetailsResults [Accessed August 29, 2020].

Google Maps Platform (2020c). Places Autocomplete Service: StructuredFormatting interface. Available at: https://developers.google.com/maps/documentation/javascript/reference/places-autocomplete-service#StructuredFormatting [Accessed August 29, 2020].

Goswami, C., Fox, S., Holden, M. T. G., Connor, M., Leanord, A., and Evans, T. J. (2020). Origin, maintenance and spread of antibiotic resistance genes within plasmids and chromosomes of bloodstream isolates of escherichia coli. *Microb. Genomics*. doi:10.1099/mgen.0.000353.

Granados-Chinchilla, F., and Rodríguez, C. (2017). Tetracyclines in Food and Feedingstuffs: From Regulation to Analytical Methods, Bacterial Resistance, and Environmental and Health Implications. *J. Anal. Methods Chem.* 2017. doi:10.1155/2017/1315497.

Grape, M., Farra, A., Kronvall, G., and Sundström, L. (2005). Integrons and gene cassettes in clinical isolates of co-trimoxazole-resistant Gram-negative bacteria. *Clin. Microbiol. Infect.* 11. doi:10.1111/j.1469-0691.2004.01059.x.

Hannigan, G. D., Meisel, J. S., Tyldsley, A. S., Zheng, Q., Hodkinson, B. P., Sanmiguel, A. J., et al. (2015). The human skin double-stranded DNA virome: Topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio*. doi:10.1128/mBio.01578-15.

Harbarth, S., and Samore, M. H. (2005). Antimicrobial resistance determinants and future control. *Emerg. Infect. Dis.* doi:10.3201/eid1106.050167.

He, D., Zhu, Y., Li, R., Pan, Y., Liu, J., Yuan, L., et al. (2019). Emergence of a hybrid plasmid derived from IncN1-F33:A-:B- And mcr-1-bearing plasmids mediated by IS26. *J. Antimicrob. Chemother.* doi:10.1093/jac/dkz327.

Hennequin, C., and Robin, F. (2016). Correlation between antimicrobial resistance and virulence in Klebsiella pneumoniae. *Eur. J. Clin. Microbiol. Infect. Dis.* 35, 333–341. doi:10.1007/s10096-015-2559-7.

HM Government (2019). Joint report on antibiotic use and antibiotic resistance, 2013 – 2017. *UK One Heal. Rep.*

HM Government (2020). Veterinary Antimicrobial Resistance and Sales Surveillance Report: UK-VARSS 2019.

Holden, M. T. G., Lindsay, J. A., Corton, C., Quail, M. A., Cockfield, J. D., Pathak, S., et al. (2010). Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). *J. Bacteriol.* doi:10.1128/JB.01255-09.

Huang, Y. T., Tang, Y. Y., Cheng, J. F., Wu, Z. Y., Mao, Y. C., and Liu, P. Y. (2018). Genome analysis of multidrug-resistant Shewanella algae isolated from human soft tissue sample. *Front. Pharmacol.* doi:10.3389/fphar.2018.00419.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* doi:10.1093/molbev/msw046.

Hughes, D., and Andersson, D. I. (2017). Evolutionary Trajectories to Antibiotic Resistance. *Annu. Rev. Microbiol.* doi:10.1146/annurev-micro-090816-093813.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. doi:10.1186/1471-2105-11-119.

IJSEM (2015). IJSEM Culture Collection Abbreviations: Abbreviations of culture collections cited in the Validation Lists. Available at: https://www.microbiologyresearch.org/marketing/editorial/IJSEM_Culture_Collection_Abbreviation_14082015.pdf [Accessed September 2, 2020].

Jacoby, G. A., Griffin, C. M., and Hooper, D. C. (2011). Citrobacter spp. as a source of qnrB alleles. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.05187-11.

Jacoby, G. A., Strahilevitz, J., and Hooper, D. C. (2014). Plasmid-Mediated Quinolone Resistance. *Microbiol. Spectr.* 2. doi:10.1128/microbiolspec.PLAS-0006-2013.

Kans, J. (2013). Entrez Direct: E-utilities on the UNIX Command Line. In: Entrez Programming Utilities Help. *Entrez Program. Util. Help*. Available at: http://www.ncbi.nlm.nih.gov/books/NBK179288/ [Accessed November 24, 2016].

Kazil, J., and Jarmul, K. (2016). “Chapter 7: Data Cleanup: Investigation, Matching, and Formatting,” in *Data Wrangling with Python: Tips and Tools to Make Your Life Easier* (Sebastopol, California, USA: O’Reilly Media), 178–181.

Kirchhelle, C. (2018). Pharming animals: a global history of antibiotics in food production (1935–2017). *Palgrave Commun.* 4. doi:10.1057/s41599-018-0152-2.

Kostic, A. D., Howitt, M. R., and Garrett, W. S. (2013). Exploring host-microbiota interactions in animal models and humans. *Genes Dev.* doi:10.1101/gad.212522.112.

Krapp, F., Morris, A. R., Ozer, E. A., and Hauser, A. R. (2017). Virulence Characteristics of Carbapenem-Resistant Klebsiella pneumoniae Strains from Patients with Necrotizing Skin and Soft Tissue Infections. *Sci. Rep.* 7. doi:10.1038/s41598-017-13524-8.

Kröger, C., Dillon, S. C., Cameron, A. D. S., Papenfort, K., Sivasankaran, S. K., Hokamp, K., et al. (2012). The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1201061109.

Labbé, G., Ziebell, K., Bekal, S., Macdonald, K. A., Parmley, E. J., Agunos, A., et al. (2016). Complete genome sequences of 17 Canadian isolates of Salmonella enterica subsp. enterica serovar Heidelberg from human, animal, and food sources. *Genome Announc.* doi:10.1128/genomeA.00990-16.

LaBreck, P. T., Rice, G. K., Paskey, A. C., Elassal, E. M., Cer, R. Z., Law, N. N., et al. (2018). Conjugative

transfer of a novel staphylococcal plasmid encoding the biocide resistance gene, QacA. *Front. Microbiol.* doi:10.3389/fmicb.2018.02664.

Lartigue, M. F., Poirel, L., Aubert, D., and Nordmann, P. (2006). In vitro analysis of ISEcp1B-mediated mobilization of naturally occurring β-lactamase gene blaCTX-M of Kluyvem ascorbata. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.50.4.1282-1286.2006.

Lopatkin, A. J., Meredith, H. R., Srimani, J. K., Pfeiffer, C., Durrett, R., and You, L. (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat. Commun.* doi:10.1038/s41467-017-01532-1.

Luo, Q., Wang, Y., and Xiao, Y. (2020). Prevalence and transmission of mobilized colistin resistance (mcr) gene in bacteria common to animals and humans. *Biosaf. Heal.* doi:10.1016/j.bsheal.2020.05.001.

Ma, F., Xu, S., Tang, Z., Li, Z., and Zhang, L. (2021). Use of antimicrobials in food animals and impact of transmission of antimicrobial resistance on humans. *Biosaf. Heal.* 3. doi:10.1016/j.bsheal.2020.09.004.

Ma, Y., Metch, J. W., Yang, Y., Pruden, A., and Zhang, T. (2016). Shift in antibiotic resistance gene profiles associated with nanosilver during wastewater treatment. *FEMS Microbiol. Ecol.* doi:10.1093/femsec/fiw022.

Makarova, O., Johnston, P., Walther, B., Rolff, J., and Roesler, U. (2017). Complete genome sequence of the disinfectant susceptibility testing reference strain staphylococcus aureus subsp. aureus ATCC 6538. *Genome Announc.* doi:10.1128/genomeA.00293-17.

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., et al. (2012). Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. Plant Pathol.* doi:10.1111/j.1364-3703.2012.00804.x.

Martínez-Martínez, L., Pascual, A., and Jacoby, G. A. (1998). Quinolone resistance from a transferable plasmid. *Lancet*. doi:10.1016/S0140-6736(97)07322-4.

Martínez, J. L., and Baquero, F. (2014). Emergence and spread of antibiotic resistance: Setting a parameter space. *Ups. J. Med. Sci.* 119, 68–77. doi:10.3109/03009734.2014.901444.

Martínez, J. L., Baquero, F., and Andersson, D. I. (2007). Predicting antibiotic resistance. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro1796.

Mathers, A. J., Peirano, G., and Pitout, J. D. (2015). The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clin.Microbiol.Rev.* 28, 565–591. doi:10.1128/CMR.00116-14.

McDonald, Y. J., Schwind, M., Goldberg, D. W., Lampley, A., and Wheeler, C. M. (2017). An analysis of the process and results of manual geocode correction. *Geospat. Health*. doi:10.4081/gh.2017.526.

Munoz-Price, L. S., Poirel, L., Bonomo, R. A., Schwaber, M. J., Daikos, G. L., Cormican, M., et al. (2013). Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. *Lancet Infect. Dis.* doi:10.1016/S1473-3099(13)70190-7.

Nahin, A. (2008). Create Date — New Field Indicates When Record Added to PubMed. *NLM Tech. Bull.* Available at: https://www.nlm.nih.gov/pubs/techbull/nd08/nd08_pm_new_date_field.html [Accessed November 24, 2016].

NCBI (2020a). BioSample Attributes. Available at:

https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/ [Accessed September 2, 2020].

NCBI (2020b). The UniVec Database. Available at: https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/ [Accessed September 2, 2020].

Ohtsubo, Y., Nagata, Y., Numata, M., Tsuchikane, K., Hosoyama, A., Yamazoe, A., et al. (2015). Complete genome sequence of Sphingopyxis macrogoltabida type strain NBRC 15033, originally isolated as a polyethylene glycol degrader. *Genome Announc.* doi:10.1128/genomeA.01401-15.

Ohtsubo, Y., Nonoyama, S., Nagata, Y., Numata, M., Tsuchikane, K., Hosoyama, A., et al. (2016). Complete genome sequence of Sphingopyxis macrogoltabida strain 203N (NBRC 111659), a polyethylene glycol degrader. *Genome Announc.* doi:10.1128/genomeA.00529-16.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* doi:10.1186/s13059-016-0997-x.

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. (2017). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* 91, 42–52. doi:10.1016/j.plasmid.2017.03.002.

Ou, H. Y., Kuang, S. N., He, X., Molgora, B. M., Ewing, P. J., Deng, Z., et al. (2015). Complete genome sequence of hypervirulent and outbreak-associated Acinetobacter baumannii strain LAC-4: Epidemiology, resistance genetic determinants and potential virulence factors. *Sci. Rep.* doi:10.1038/srep08643.

Pal, C., Bengtsson-Palme, J., Kristiansson, E., and Larsson, D. G. J. J. (2015). Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics* 16, 964. doi:10.1186/s12864-015-2153-5.

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., and Larsson, D. G. J. (2014). BacMet: Antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* doi:10.1093/nar/gkt1252.

Patel, G., and Bonomo, R. A. (2013). “Stormy waters ahead”: Global emergence of carbapenemases. *Front. Microbiol.* doi:10.3389/fmicb.2013.00048.

Paterson, D. L., and Bonomo, R. A. (2005). Extended-spectrum β-lactamases: A clinical update. *Clin. Microbiol. Rev.* 18. doi:10.1128/CMR.18.4.657-686.2005.

Peacock, S. J. (2006). “Staphylococcus aureus,” in *Principles and Practice of Clinical Bacteriology*, eds. S. H. Gillespie and P. M. Hawkey (Chichester, UK: John Wiley & Sons, Ltd), 73–98.

Petersen, T. N., Rasmussen, S., Hasman, H., Carøe, C., Bælum, J., Charlotte Schultz, A., et al. (2015). Meta-genomic analysis of toilet waste from long distance flights; A step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci. Rep.* doi:10.1038/srep11444.

Pightling, A. W., Petronella, N., and Pagotto, F. (2014). Choice of reference sequence and assembler for alignment of Listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. *PLoS One*. doi:10.1371/journal.pone.0104579.

Pitzschke, A. (2013). From Bench to Barn: Plant Model Research and its Applications in Agriculture. *Adv. Genet. Eng.* doi:10.4172/2169-0111.1000110.

Poirel, L., Bonnin, R. A., and Nordmann, P. (2012). Genetic support and diversity of acquired extended-spectrum β-lactamases in Gram-negative rods. *Infect. Genet. Evol.* 12. doi:10.1016/j.meegid.2012.02.008.

Porse, A., Schønning, K., Munck, C., and Sommer, M. O. A. (2016). Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. *Mol. Biol. Evol.* 33, 2860–2873. doi:10.1093/molbev/msw163.

Potron, A., Poirel, L., and Nordmann, P. (2011). Origin of OXA-181, an emerging carbapenem-hydrolyzing oxacillinase, as a chromosomal gene in Shewanella xiamenensis. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.00681-11.

Reimer, L. C., Söhngen, C., Vetcininova, A., and Overmann, J. (2017). Mobilization and integration of bacterial phenotypic data—Enabling next generation biodiversity analysis through the BacDive metadatabase. *J. Biotechnol.* doi:10.1016/j.jbiotec.2017.05.004.

Reimer, L. C., Vetcininova, A., Carbasse, J. S., Söhngen, C., Gleim, D., Ebeling, C., et al. (2019). BacDive in 2019: Bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.* doi:10.1093/nar/gky879.

Roberts, M. C. (1996). Tetracycline resistance determinants: Mechanisms of action, regulation of expression, genetic mobility, and distribution. *FEMS Microbiol. Rev.* 19. doi:10.1016/0168-6445(96)00021-6.

Rodríguez-Martínez, J. M., Machuca, J., Cano, M. E., Calvo, J., Martínez-Martínez, L., and Pascual, A. (2016). Plasmid-mediated quinolone resistance: Two decades on. *Drug Resist. Updat.* doi:10.1016/j.drup.2016.09.001.

Rozwandowicz, M., Brouwer, M. S. M., Fischer, J., Wagenaar, J. A., Gonzalez-Zorn, B., Guerra, B., et al. (2018). Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *J. Antimicrob. Chemother.* 73, 1121–1137. doi:10.1093/jac/dkx488.

San Millan, A. (2018). Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. *Trends Microbiol.* 26, 978–985. doi:10.1016/j.tim.2018.06.007.

San Millan, A., and MacLean, R. C. (2017). Fitness Costs of Plasmids: A Limit to Plasmid Transmission. *Microb. Transm.*, 65–79. doi:10.1128/microbiolspec.mtbp-0016-2017.

San Millan, A., Toll-Riera, M., Qi, Q., Betts, A., Hopkinson, R. J., McCullagh, J., et al. (2018). Integrative analysis of fitness and metabolic effects of plasmids in Pseudomonas aeruginosa PAO1. *ISME J.* doi:10.1038/s41396-018-0224-8.

Satou, K., Shimoji, M., Tamotsu, H., Juan, A., Ashimine, N., Shinzato, M., et al. (2015). Complete genome sequences of low-passage virulent and high-passage avirulent variants of pathogenic Leptospira interrogans serovar Manilae strain UP-MMCNIID, originally isolated from a patient with severe leptospirosis, determined using PacBio single-mole. *Genome Announc.* doi:10.1128/genomeA.00882-15.

Schäffer, A. A., Nawrocki, E. P., Choi, Y., Kitts, P. A., Karsch-Mizrachi, I., and McVeigh, R. (2018). VecScreen-plus-taxonomy: Imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics*. doi:10.1093/bioinformatics/btx669.

Silver, L. L. (2011). Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* doi:10.1128/CMR.00030-10.

Sköld, O. (2000). Sulfonamide resistance: Mechanisms and trends. *Drug Resist. Updat.* 3, 155–160. doi:10.1054/drup.2000.0146.

Skold, O., and Swedberg, G. (2017). "Chapter 24: Sulfonamides and Trimethoprim," in *Antimicrobial Drug Resistance: Mechanisms of Drug Resistance, Volume 1*, eds. D. L. Mayers, J. D. Sobel, M. Ouellette, K. S. Kaye, and D. Marchaim (Cham, Switzerland: Springer), 345–358.

Snyder, L., Peters, J. E., Henkin, T. M., and Champness, W. (2013). “Chapter 9 Transposition, site-specific recombination, and families of recombinases,” in *Molecular Genetics of Bacteria* (Washington, DC: ASM Press), 361–402.

Stevenson, C., Hall, J. P. J., Harrison, E., Wood, A. J., and Brockhurst, M. A. (2017). Gene mobility promotes the spread of resistance in bacterial populations. *ISME J.* doi:10.1038/ismej.2017.42.

Stokes, H. W., and Gillings, M. R. (2011). Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* 35, 790–819. doi:10.1111/j.1574-6976.2011.00273.x.

The World Bank (2018). World Bank Country and Lending Groups. *World Bank Ctry. Lend. Groups*. Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.

Touchon, M., and Rocha, E. P. C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.* 24. doi:10.1093/molbev/msm014.

Tsai, C. J. Y., Loh, J. M. S., and Proft, T. (2016). Galleria mellonella infection models for the study of bacterial diseases and for antimicrobial drug testing. *Virulence*. doi:10.1080/21505594.2015.1135289.

Uruburu, F. (2003). History and services of culture collections. *Int. Microbiol.* doi:10.1007/s10123-003-0115-2.

Van Hoek, A. H. A. M., Mevius, D., Guerra, B., Mullany, P., Roberts, A. P., and Aarts, H. J. M. (2011). Acquired antibiotic resistance genes: An overview. *Front. Microbiol.* 2, 1–27. doi:10.3389/fmicb.2011.00203.

Vandecraen, J., Chandler, M., Aertsen, A., and Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.* doi:10.1080/1040841X.2017.1303661.

Vogwill, T., and Maclean, R. C. (2015). The genetic basis of the fitness costs of antimicrobial resistance: A meta-analysis approach. *Evol. Appl.* doi:10.1111/eva.12202.

Welch, R. A., Jones, K. R., and Macrina, F. L. (1979). Transferable lincosamide-macrolide resistance in Bacteroides. *Plasmid* 2, 261–268. doi:10.1016/0147-619X(79)90044-1.

WHO (2018). WHO Report on Surveillance of Antibiotic Consumption: 2016-2018 Early Implementation. Geneva, Switzerland.

Wood, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* doi:10.1111/j.1467-9868.2010.00749.x.

Wood, S. N. (2017a). “Chapter 4: Introducing GAMs,” in *Generalized Additive Models: An Introduction with R, Second Edition*, 185.

Wood, S. N. (2017b). *Generalized additive models: An introduction with R, second edition*. doi:10.1201/9781315370279.

Xie, Z., and Tang, H. (2017). ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*. doi:10.1093/bioinformatics/btx433.

Yigit, H., Queenan, A. M., Anderson, G. J., Domenech-Sanchez, A., Biddle, J. W., Steward, C. D., et al.

(2001). Novel carbapenem-hydrolyzing β-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.45.4.1151-1161.2001.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi:10.1093/jac/dks261.

6 Discussion and future directions

In the early stages of research (in 2016), I conducted a literature review of whole genome sequencing (WGS)-based studies of plasmid-mediated antibiotic resistance. Most studies were based on short-read sequencing data, and used plasmid typing – especially replicon typing – to analyse plasmid epidemiology. In addition, the studies were generally small-scale analyses of in-house datasets. My thesis has been conducted within the context of massive growth in the availability of bacterial WGS data, which is stored in public sequence repositories such as NCBI. Alongside continued short-read sequencing, long-read sequencing has emerged, facilitating resolution of (near) complete bacterial genome assemblies. Given this context, much of my research is unified by two major goals. Firstly, rather than focusing on replicon typing, I wanted to analyse the wider plasmid genome to gain more detailed insight into the biology and epidemiology of resistance plasmids. Secondly, as well as performing small-scale WGS-based analyses of short- and long-read WGS data, I wanted to take advantage of the large number of plasmid sequences available within NCBI. I discuss these goals in more detail below, as I summarise my research and consider future directions.

As discussed in the Introduction (Chapter 1), replicon typing lacks resolution as a typing tool for plasmid epidemiology. Nevertheless, replicon typing remains useful for plasmid identification, since plasmid replicon loci are plasmid-specific markers. Indeed, in Chapter 2, I developed a tool called bacterialBercow which uses detection of plasmid replicon and rMLST loci as a basis for delineating plasmid and chromosomal contigs in (near) complete bacterial genome assemblies. A related function of bacterialBercow is the retrieval and curation of complete plasmid sequences from NCBI. In Chapter 2, I used a curated plasmid dataset to assess the performance of *in silico* plasmid typing schemes (in 2017 and again in 2020). As part of this analysis, I investigated the proportion of plasmids typed ('typeability') by replicon typing. This demonstrated that replicon typing failed to accommodate the diversity of plasmids within NCBI; typeability was especially poor for plasmids that fell outside

relatively well-studied (clinically relevant) taxa, such as Enterobacteriaceae. A recent study of bacterial plasmids showed that around two-thirds of genes were uncharacterised (encoding hypothetical proteins) (Acman et al., 2020). Future work to functionally characterise plasmid genes is required; improved characterisation of plasmid replication machinery would simplify delineation of plasmid and chromosomal contigs.

Given the limitations of replicon typing for plasmid epidemiology, I developed novel WGS analysis methods, capitalising on the wider plasmid genome. In Chapter 3, I developed a reference-based method for comparative analysis of plasmidome content across fragmented short-read assemblies. In Chapter 4, I developed ATCG, a tool intended for comparative genomics of (near) complete (e.g. long-read resolved) bacterial assemblies. I applied ATCG to analyse available long-read sequenced samples of *bla*$_{KPC}$-encoding Enterobacterales, collected in 2012–2016, from hospitals in and around Manchester. One of the advantages of analysing long-read resolved assemblies is the potential to infer putative transmission links (according to stringent identity and coverage breath thresholds). For example, in Chapter 4, after applying ATCG, I found evidence for persistence of conserved plasmids across a period of several years, as well as examples of inferred horizontal plasmid transfer between different species and strains. Another advantage of analysing complete plasmid assemblies is that the co-association of plasmid genetic traits can be investigated (as demonstrated in Chapter 5). In future, long-read sequencing data will increasingly drive knowledge of plasmid biology and epidemiology. Nevertheless, if/when short-read technology is completely superseded, the associated software ecosystem is likely to remain useful for analysis of historical sequence collections.

ATCG and bacterialBercow were the principal software tools I developed; I demonstrated their broad utility with multiple use cases. Specifically, bacterialBercow was used for contig classification, species identification, and contamination detection (Chapter 4), as well as to curate NCBI plasmids (Chapter 5). Likewise, ATCG was used to conduct all-vs-all comparative analysis of in-house data (Chapter 4), and to facilitate deduplication of a large plasmid dataset, by identifying similar plasmids, in

conjunction with the mash alignment-free comparison tool (Chapter 5). To empower other researchers to use my software, I tried to follow good software programming practices. For example, I implemented argument parsing, error handling, and GitHub version control, and provided detailed documentation (Seemann, 2013). However, additional improvements that I could implement in future include automated unit testing (as opposed to manually testing the code) and containerisation (e.g. using Docker) (Georgeson et al., 2019). Besides general issues around software development good practice, there are also specific limitations of bacterialBercow and ATCG, which could be addressed in future, as described below.

A current limitation of bacterialBercow is that it does not scale well, since it uses BLAST to detect rMLST and replicon loci. Hence, an improvement would be to first screen for similar rMLST/replicon loci using a fast alignment-free tool such as mash screen (Ondov et al., 2019); subsequent BLAST searches could then be restricted to a subset of candidate rMLST/replicon loci. An additional limitation of bacterialBercow is that it does not detect phage sequences. Phage identification is challenging since phages are immensely diverse and lack conserved genetic markers; in addition, many phage groups are underrepresented in reference databases (Dion et al., 2020; Uelze et al., 2020). However, putative phage contigs could be identified by applying existing tools such as VirSorter. VirSorter uses a combination of reference-based and reference-independent metrics to identify putative phage sequences (Roux et al., 2015). As mentioned in Chapter 4, future developments to improve ATCG performance could alleviate bottlenecks by improving parallelisation and reducing read/write operations. Faster runtime could also be achieved by using USEARCH rather than BLAST (see Chapter 4). Comparative genomic visualisation was not a major focus of ATCG development (although an experimental script was developed). However, there is an important need for improved comparative genomic visualisation tools, which can scale to large datasets (Aurisano et al., 2015; Sullivan and van Bakel, 2020).

The Manchester outbreak analysis presented in Chapter 4 was restricted in scale and scope (taxonomic and spatio-temporal scope). Many WGS studies of plasmid-mediated antibiotic resistance are similarly restricted (Larsson et al., 2018). Such studies can provide insights into local plasmid epidemiology and transmission dynamics (Tang et al., 2017), as I demonstrated in Chapter 4. Small-scale studies may even illuminate previously unknown or underappreciated aspects of plasmid biology – for example, through a hospital outbreak study, Sheppard et al. demonstrated that frequent mobility of plasmids and their nested mobile elements can occur over relatively short timescales (Sheppard et al., 2016). However, local studies may not represent global dynamics. To gain robust general insights into plasmid biology, large diverse plasmid datasets are required. Therefore, in Chapter 5, I used a global dataset of curated NCBI plasmids and associated curated NCBI BioSample metadata to investigate factors associated with plasmid antibiotic resistance carriage. This research was the first large-scale statistical analysis of plasmids to assess both plasmid genetic features and higher-level metadata characteristics (isolation source, collection date, geographic location). I found that patterns of plasmid resistance carriage in different isolation sources (human/livestock), across time, and across different host bacterial taxa were concordant with known patterns of antibiotic usage. This finding is consistent with antibiotic usage as an important driver of resistance, and supports previous analyses of surveillance data (primarily from Europe and the USA), which link strain-level antibiotic resistance with antibiotic consumption, in clinical and livestock settings (ECDC/EFSA/EMA, 2017; Olesen et al., 2018). Hence, antibiotic stewardship in human and livestock settings is of central importance; in future, an improved evidence-base around antibiotic prescribing practice may help prevent unnecessary antibiotic use (Davey et al., 2017; Llewelyn et al., 2017).

Chapters 2 and 5 underscore the importance of curation, when analysing large public datasets of plasmids and associated metadata (Howe et al., 2008). In both chapters, curation was important to obtain a reliable dataset of complete plasmids (excluding incomplete and chromosomal sequences). Additionally, in Chapter 5, I identified and corrected erroneous BioSample metadata, following internal and external consistency checking. Previous work has highlighted BioSample metadata quality

issues in terms of *ad hoc* attribute names, missing attribute values, and attribute values not adhering to recommended ontologies (Gonçalves and Musen, 2019; Schriml et al., 2020). In Chapter 5, to account for missing/invalid attribute values, regexes were used to apply inclusion/exclusion criteria. For example, terms such as 'missing' and 'N/A' were excluded from the geographic location name field prior to geocoding, whereas values with valid date formats were extracted from the collection date field. Given high levels of manual oversight, it was not possible to produce an automated curation workflow to encompass the metadata curation methods described in Chapter 5. Hopefully, future improvements to NCBI in-house biocuration procedures will reduce the need for individual researchers to invest time in curation (Ammari et al., 2018; Tang et al., 2019). For example, as mentioned in Chapter 5, given a valid culture collection identifier from a submitter, metadata could be automatically retrieved from the Bac*Dive* metadata database, rather than relying on submitters to provide culture collection sample metadata. Other improvements could include increased use of ontologies (e.g. there is currently no ontology for isolation source), and enforcement of existing ontologies (Gonçalves and Musen, 2019). Ultimately, submitters must be encouraged to provide rich and accurate metadata, while continuing to allow for privacy concerns (e.g. currently, BioSample geographic locations need not be specified more precisely than country) (NCBI, 2020; PHG Foundation, 2020; Schriml et al., 2020).

Aside from data quality issues, a more fundamental limitation of analysing NCBI plasmids is biased sampling. To gain the most reliable and up-to-date insights into resistance plasmid epidemiology, global routine surveillance programmes are needed (Petersen et al., 2015; Aarestrup and Woolhouse, 2020). A current programme which could potentially address this need is the Global Antimicrobial Resistance and Use Surveillance System (GLASS), launched by the World Health Organisation in 2015. It aims to integrate and build upon existing regional surveillance networks. So far, GLASS has reported on the global frequency of resistance in high priority clinical pathogens, based on laboratory antimicrobial susceptibility testing. Future plans include reporting antimicrobial usage data and One Health resistance data (extended-spectrum beta-lactamase resistance in *Escherichia coli* across

human, livestock, and environmental settings) (WHO, 2020). There are no current plans for generating or publishing WGS datasets as part of GLASS, although this would be one mechanism for facilitating global routine surveillance of resistance plasmid epidemiology. Given that GLASS does not currently focus on resistance plasmids, analyses of NCBI plasmids will remain valuable for the foreseeable future.

A major question requiring further investigation is the role of plasmids in the spread of antibiotic resistance on a global scale, and across different settings (human/livestock/environment). Previous studies have investigated horizontal transfer between bacterial genomes, using public datasets. For example, using a 99% identity threshold to define transmission links, Smillie et al. found that isolation source category (human/non-human) was more important than geography, in structuring bacterial gene exchange (Smillie et al., 2011). A similar approach could be applied to NCBI plasmids and associated BioSample metadata. However, the reliability of such analyses is undermined by sampling bias – intense sampling of epidemic resistance plasmid lineages in clinical settings may inflate human-human transmission links in a way that cannot easily be controlled for (Soucy et al., 2015). To better understand resistance plasmid transmission, future WGS studies should focus on sampling diverse bacteria across different settings. Settings of particular interest include wastewater treatment plants, and the (human/livestock) gut microbiome, since these are considered potential hotspots of plasmid-mediated antibiotic resistance transmission (Von Wintersdorff et al., 2016; San Millan, 2018; Che et al., 2019).

An example of a WGS study enabling insights into resistance plasmid transmission is the REHAB project (UKRI, 2020), currently being conducted by members of my research group and collaborating groups. During the project, long-read sequencing has been applied to study Enterobacterales isolates cultured from various livestock settings (cattle, pig, and sheep farms) and wastewater treatment plants, across multiple timepoints. Because near complete genomes have been resolved through long-read sequencing, plasmid contigs can be distinguished, and consequently, putative plasmid transmission

events can be inferred by conducting all-vs-all pairwise comparisons (Shaw et al., 2020). Findings from this study alone are likely to lack generalisability (all sampling sites were within 30 km); however, as more projects like this are conducted around the world, systematic reviews and meta-analyses could help to uncover general patterns of inferred plasmid transmission in livestock/wastewater settings. The REHAB project was restricted to culturable Enterobacterales; more comprehensive insight into plasmid transmission can potentially be gained through culture-independent metagenomic sequencing (Hiraoka et al., 2016; Bickhart et al., 2019; Kent et al., 2020). Finally, more narrowly focused experimental studies of *in situ* plasmid transfer can complement large-scale multi-site sequencing studies (McInnes et al., 2020; Saak et al., 2020). For example, a plasmid can be engineered with a conditionally-expressible reporter gene, expressing fluorescent protein in recipient but not donor cells; transfer events can therefore be tracked using fluorescence-activated cell sorting (FACS) (Klümper et al., 2014). Experimental approaches are likely to be especially useful in understanding how different factors influence the rate of plasmid transfer. For example, *in situ* plasmid transfer experiments have suggested that host gut inflammation and/or dysbiosis promotes plasmid transfer (Sitaraman, 2018).

An in-depth understanding of the hotspots of plasmid transfer and the factors influencing rates of plasmid transfer in different environments will provide an evidence-base for targeting interventions against plasmid-mediated antibiotic resistance transmission. Various strategies for eliminating or managing resistance plasmids have emerged in recent years. These include plasmid curing mechanisms, conjugation inhibitors, and CRISPR/Cas-based methods (Buckner et al., 2018; Getino and de la Cruz, 2018). In future, one can imagine a scenario where plasmid targeting interventions are widely used to control high-risk antibiotic resistance plasmids in clinical settings. For example, prior to antibiotic treatment, a patient could be given a therapy targeting resistance plasmids residing in their gut microbiome. An effective therapy would prevent transfer of resistance plasmids from gut commensals to pathogenic species, thereby protecting patients, and contributing towards antibiotic stewardship.

Bibliography

Aarestrup, F. M., and Woolhouse, M. E. J. (2020). Using sewage for surveillance of antimicrobial resistance. *Science (80-.).* doi:10.1126/science.aba3432.

Acman, M., van Dorp, L., Santini, J. M., and Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* doi:10.1038/s41467-020-16282-w.

Ammari, M., Chatr Aryamontri, A., Attrill, H., Bairoch, A., Berardini, T., Blake, J., et al. (2018). Biocuration: Distilling data into knowledge. *PLoS Biol.* doi:10.1371/journal.pbio.2002846.

Aurisano, J., Reda, K., Johnson, A., Marai, E. G., and Leigh, J. (2015). BactoGeNIE: A large-scale comparative genome visualization for big displays. *BMC Bioinformatics*. doi:10.1186/1471-2105-16-S11-S6.

Bickhart, D. M., Watson, M., Koren, S., Panke-Buisse, K., Cersosimo, L. M., Press, M. O., et al. (2019). Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol.* doi:10.1186/s13059-019-1760-x.

Buckner, M. M. C., Ciusa, M. L., and Piddock, L. J. V. (2018). Strategies to combat antimicrobial resistance: Anti-plasmid and plasmid curing. *FEMS Microbiol. Rev.* doi:10.1093/femsre/fuy031.

Che, Y., Xia, Y., Liu, L., Li, A. D., Yang, Y., and Zhang, T. (2019). Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome*. doi:10.1186/s40168-019-0663-0.

Davey, P., Marwick, C. A., Scott, C. L., Charani, E., Mcneil, K., Brown, E., et al. (2017). Interventions to improve antibiotic prescribing practices for hospital inpatients. *Cochrane Database Syst. Rev.* doi:10.1002/14651858.CD003543.pub4.

Dion, M. B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* doi:10.1038/s41579-019-0311-5.

ECDC/EFSA/EMA (2017). ECDC/EFSA/EMA second joint report on the integrated analysis of the consumption of antimicrobial agents and occurrence of antimicrobial resistance in bacteria from humans and food-producing animals: Joint Interagency Antimicrobial Consumption and Resistan. *EFSA J.* doi:10.2903/j.efsa.2017.4872.

Georgeson, P., Syme, A., Sloggett, C., Chung, J., Dashnow, H., Milton, M., et al. (2019). Bionitio: Demonstrating and facilitating best practices for bioinformatics command-line software. *Gigascience*. doi:10.1093/gigascience/giz109.

Getino, M., and de la Cruz, F. (2018). "Natural and Artificial Strategies to Control the Conjugative Transmission of Plasmids," in *Microbial Transmission* doi:10.1128/microbiolspec.mtbp-0015-2016.

Gonçalves, R. S., and Musen, M. A. (2019). Analysis: The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data*. doi:10.1038/sdata.2019.21.

Hiraoka, S., Yang, C. C., and Iwasaki, W. (2016). Metagenomics and bioinformatics in microbial ecology: Current status and beyond. *Microbes Environ.* 00, 204–212. doi:10.1264/jsme2.ME16024.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., et al. (2008). Big data: The future of biocuration. *Nature*. doi:10.1038/455047a.

Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J., and Brito, I. L. (2020). Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* doi:10.1038/s41467-020-18164-7.

Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L. H., Sørensen, S. J., et al. (2014). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* 9, 934–945. doi:10.1038/ismej.2014.191.

Larsson, D. G. J., Andremont, A., Bengtsson-Palme, J., Brandt, K. K., de Roda Husman, A. M., Fagerstedt, P., et al. (2018). Critical knowledge gaps and research needs related to the environmental dimensions of antibiotic resistance. *Environ. Int.* 117, 132–138. doi:10.1016/j.envint.2018.04.041.

Llewelyn, M. J., Fitzpatrick, J. M., Darwin, E., Sarahtonkin-Crine, Gorton, C., Paul, J., et al. (2017). The antibiotic course has had its day. *BMJ*. doi:10.1136/bmj.j3418.

McInnes, R. S., McCallum, G. E., Lamberte, L. E., and van Schaik, W. (2020). Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2020.02.002.

NCBI (2020). BioSample Attributes. Available at: https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/ [Accessed September 2, 2020].

Olesen, S. W., Barnett, M. L., Macfadden, D. R., Brownstein, J. S., Hernández-Díaz, S., Lipsitch, M., et al. (2018). The distribution of antibiotic use and its association with antibiotic resistance. *Elife*. doi:10.7554/eLife.39435.

Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., et al. (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biol.* doi:10.1186/s13059-019-1841-x.

Petersen, T. N., Rasmussen, S., Hasman, H., Carøe, C., Bælum, J., Charlotte Schultz, A., et al. (2015). Meta-genomic analysis of toilet waste from long distance flights; A step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci. Rep.* doi:10.1038/srep11444.

PHG Foundation (2020). The GDPR and genomic data: The impact of the GDPR and DPA 2018 on genomic healthcare and research.

Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*. doi:10.7717/peerj.985.

Saak, C. C., Dinh, C. B., and Dutton, R. J. (2020). Experimental approaches to tracking mobile genetic elements in microbial communities. *FEMS Microbiol. Rev.* doi:10.1093/femsre/fuaa025.

San Millan, A. (2018). Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. *Trends Microbiol.* 26, 978–985. doi:10.1016/j.tim.2018.06.007.

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., et al. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data*. doi:10.1038/s41597-020-0524-5.

Seemann, T. (2013). Ten recommendations for creating usable bioinformatics command line software. *Gigascience*. doi:10.1186/2047-217X-2-15.

Shaw, L. P., Chau, K. K., Kavanagh, J., AbuOun, M., Stubberfield, E., Gweon, H. S., et al. (2020). Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *bioRxiv*. doi:10.1101/2020.07.23.215756.

Sheppard, A. E., Stoesser, N., Wilson, D. J., Sebra, R., Kasarskis, A., Anson, L. W., et al. (2016). Nested Russian Doll-like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob. Agents Chemother.* doi:10.1128/AAC.00464-16.

Sitaraman, R. (2018). Prokaryotic horizontal gene transfer within the human holobiont: Ecological-evolutionary inferences, implications and possibilities 06 Biological Sciences 0604 Genetics. *Microbiome*. doi:10.1186/s40168-018-0551-z.

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., and Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244. doi:10.1038/nature10571.

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16, 472–482. doi:10.1038/nrg3962.

Sullivan, M., and van Bakel, H. (2020). Chromatiblock: scalable whole-genome visualization of structural differences in prokaryotes. *J. Open Source Softw.* 5, 2451. doi:10.21105/joss.02451.

Tang, P., Croxen, M. A., Hasan, M. R., Hsiao, W. W. L., and Hoang, L. M. (2017). Infection control in the new age of genomic epidemiology. *Am. J. Infect. Control*. doi:10.1016/j.ajic.2016.05.015.

Tang, Y. A., Pichler, K., Füllgrabe, A., Lomax, J., Malone, J., Munoz-Torres, M. C., et al. (2019). Ten quick tips for biocuration. *PLoS Comput. Biol.* doi:10.1371/journal.pcbi.1006906.

Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., et al. (2020). Typing methods based on whole genome sequencing data. *One Heal. Outlook*. doi:10.1186/s42522-020-0010-1.

UKRI (2020). NEC05836 The environmental REsistome: confluence of Human and Animal Biota in antibiotic resistance spread (REHAB). Available at: https://gtr.ukri.org/projects?ref=NE%2FN019660%2F1#/tabOverview [Accessed October 1, 2020].

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., et al. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* 7, 1–10. doi:10.3389/fmicb.2016.00173.

WHO (2020). Global Antimicrobial Resistance Surveillance System (GLASS) Report: Early Implementation 2020.

Appendices

Appendix files are provided as Additional Materials as part of the digital thesis submission, however Appendix C: File 1 (file size: 3 Gb) is omitted from the submission. All Appendix files, including Appendix C: File 1 have also been uploaded to a Figshare repository:

https://doi.org/10.6084/m9.figshare.13076135

Appendix A

Methods 1

Details the methods used for typing plasmids of the Enterobacteriaceae plasmid dataset, retrieved in 2016.

Methods 2

Methods used to select optimal PSI-BLAST E-value thresholds for MOB typing are demonstrated.

Methods 3

Details the methods underlying the bacterialBercow pipeline which was used for retrieval, curation, and analysis of the bacterial plasmid dataset, retrieved in 2019.

Table 1

Details of the curated plasmids are provided including assigned replicon and MOB types, and detected resistance genes. The plasmids that were filtered at different stages of the curation process are shown.

Table 2

PSI-BLAST alignments prior to best hit selection are shown. Hits involving different MOB queries aligning to the same locus on a given plasmid are indicated.

Table 3

MOB typing results produced with alternative MOB queries are compared with results produced from original MOB queries.

File 1

File 1 is a folder containing a spreadsheet file ('conserved domain data') showing the conserved domains associated with replication proteins of the PlasmidFinder database. The relationship between replicon types in terms of sharing of replication protein domains is visualised using networks. The networks are provided as separate files within the folder. The interactive network (html file) opens in a web browser and is a bipartite network – one set of nodes represents replicon types, another set of nodes represents conserved domains. The other network is a static network (PDF file) derived from the bipartite network; it shows replicon type nodes only.

Appendix B

Table 1

Metadata relating to the 60 Enterobacterales samples used in the benchmarking analysis presented in Figure 1.

Table 2

Output from running bacterialBercow on Unicycler assemblies of the 125 short-read sequenced *E. coli* ST216 Manchester outbreak samples. Contigs are classified as “chromosomal”, “complete plasmid”, “putative plasmid”, or “unknown”, according to the bacterialBercow contig classification decision tree (see Chapter 2, Supplementary Methods S2).

Appendix C

Table 1

BLAST alignment data at each step in the ATCG alignment parsing process are shown. The data corresponds to Figure 3 in the main text and is derived from comparison of two closely related plasmids (query genome accession: NZ_CP026395.1; subject genome accession: NZ_CP026188.1).

Table 2

Pairs of bacterial genomes used for benchmarking ATCG and other software for ANI calculation; genomes are represented by NCBI assembly accessions. In total, there are 4500 intra-genus pairs from 9 genera, comprising 2721 genomes.

Table 3

Metadata relating to the 41 Manchester outbreak samples assembled using short and long sequencing reads. 38 of these 41 samples are represented in the network in Figure 7; network node labels and corresponding sample names (used by Decraene et al.) are given. Metadata is also provided for the three samples which were excluded prior to network construction due to contamination (one sample) and poor assembly (two samples).

Table 4

Output from running bacterialBercow on the 41 Manchester outbreak samples assembled using short and long sequencing reads. Output is given on a per-sample basis, which is more informative in terms of rMLST-based species assignments and identification of putative contaminant samples. Sample H102700301 was flagged as a contaminant sample given that there were 51 multi-allelic rMLST loci.

Table 5

The edgelist underlying the network in Figure 7, comprising 191 edges between samples. Edges represent putative plasmid transmission links (see main text for details). Edges are represented by sample name pairs as well as corresponding node label pairs.

File 1

Compressed archive file containing 2721 FASTA files, representing the bacterial genome assemblies used for benchmarking ATCG and other ANI calculation software.

Appendix D

Table 1

Table 1A shows the 12503 clusters of identical plasmid sequences. For each cluster, accessions and associated metadata are delimited by '|'. A duplicate category label is assigned to clusters according to accession metadata and the result of automated deduplication (see Methods for details). "Potentially non-duplicate" clusters retained multiple accessions after automated deduplication and were manually inspected. Other clusters were left with a single representative after automated deduplication. If such a cluster comprised a single RefSeq/GenBank cognate accession pair it was labelled "redundant duplicate: RefSeq/GenBank cognate pair"; otherwise it was labelled as a "putative duplicate" cluster. Table 1B shows "potentially non-duplicate" cluster accessions, associated metadata, and the outcome of manual deduplication. Table 1C shows a selection of duplicate cluster accessions which were manually inspected (see Methods for details); accession metadata as well as a decision on whether the clusters are plausibly non-duplicate is indicated.

Table 2

Geocoding results for BioSample accessions for which there was a discrepancy between the geocoded latitude/longitude and the primary BioSample latitude/longitude. Discrepancies were identified based on distance (>50 km) between coordinates and geocoded place prediction viewport bounds (see Methods). A discrepancy category is assigned to indicate whether discrepancies are likely to reflect primary coordinate error, or geocoding error, or the reason is unclear (respectively labelled: biosample latlon is invalid, biosample latlon is valid, biosample latlon is discrepant). Six major discrepancies are highlighted (see main text for details).

Table 3

Culture collection samples with linked Bac*Dive* metadata are shown; the Bac*Dive*-guided curation of host, isolation source, and geographic location metadata is indicated. BioSample metadata prior to Bac*Dive*-guided curation has a pink header while post-Bac*Dive* curated metadata has a green header. Bac*Dive* metadata used for curation is given to the right. In the green headed section, yellow fill indicates addition of metadata where metadata was previously missing; orange fill indicates correction to previous metadata; blue fill indicates addition/clarification to previous metadata. The rightmost columns document the curation process and where possible explain discrepancies. In the Notes column, yellow fill indicates cases where species authority date is given instead of genuine collection date.

Table 4

Samples with early (pre-1950) collection dates are shown. Two collection date fields are given (blue text). The *collection_date* field contains the original collection date metadata retrieved from NCBI BioSample. The *collection_date_curated* field contains the collection metadata after curation (e.g. removing invalid metadata such as 'unknown'; Bac*Dive*-guided curation). The rightmost columns document the manual curation of the pre-1950 dates. During manual curation, incorrect dates were

removed from the *collection_date_curated* field, as indicated in the Notes column. In the Notes column, yellow fill indicates cases where species authority date is given instead of genuine collection date.

Table 5

Tables 5A and B show BioSample accessions which were included and excluded (respectively) following metadata curation and vector contaminant filtering steps (117 samples were excluded; see Figure 1, steps 2 and 3). Associated raw and "_curated" metadata columns are provided. Table 5C lists the 356 plasmids which were excluded following the metadata curation and vector contaminant filtering steps, with reason for exclusion indicated. Table 5D shows the set of 15914 plasmids (post-curation but prior to deduplication; see Figure 1). Plasmids which were included in the deduplicated dataset used for downstream statistical analyses are indicated (column labelled "InDeduplicatedDataset"). Table 5E shows the set of 14143 plasmids used for statistical analysis; transformed explanatory variables (following winsorising and re-factoring) are provided.

Table 6

Cross tabulations of categorical predictor variables and resistance gene presence/absence, across all resistance class outcomes. Unadjusted odds ratios, 95% confidence intervals, and p-values are also provided. Odds ratios and 95% confidence intervals are also presented on the log-scale and the probability scale (after applying the inverse logit function to the log-odds).

Table 7

Table 7A shows association statistics between explanatory variables. Table 7B shows output produced by the gam.check function; output includes model convergence and basis dimension checking results. The model checking functions were applied to the 10 GAM models (for 10 resistance class outcomes) built using the deduplicated plasmid dataset (14143 plasmids); models followed the model structure outlined in Section 1.3.7 (main text).

Table 8

Table 8A shows GAM model results from anova.gam and summary.gam functions. The functions were applied to the 10 GAM models (for 10 resistance class outcomes) built using the deduplicated plasmid dataset; models followed the model structure outlined in Section 1.3.7 (main text). Output from anova.gam indicates whole term significance for smooth and parametric terms, whereas summary.gam provides more details on parametric terms; the significance of individual categorical factor levels rather than whole term significance is given. In addition, summary.gam provides model fit measures such as R^2. For summary.gam output, significance of terms is indicated using standard R significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Table 8B provides odds ratios, 95% confidence intervals, and p-values for each parametric term. Odds ratios and 95% confidence intervals are also presented on the log-scale, the probability scale (after applying the inverse logit function to the log-odds), and the intercept-centred probability scale (after applying the inverse logit function to the log-odds + intercept). Finally, the difference between unadjusted and adjusted coefficients on the log-odds scale is provided; grey shading indicates the magnitude of the difference (darker indicates a larger difference between unadjusted and adjusted coefficients); positive/negative differences (relative to unadjusted coefficients) are indicated using red/green font, respectively.

Table 9

Table 9 presents results across various adjusted models with different parameters removed, relative to the full model (see 'Contents' tab for details). Odds ratios, 95% confidence intervals, and p-values are provided for each parametric term. In addition, odds ratios and 95% confidence intervals are presented on the log-scale, the probability scale (after applying the inverse logit function to the log-odds), and the intercept-centred probability scale (after applying the inverse logit function to the log-odds + intercept). Finally, coefficients are compared to unadjusted coefficients and adjusted (full model) coefficients on the log-odds scale; grey shading indicates the magnitude of the difference (darker indicates a larger difference); positive/negative differences are indicated using red/green font, respectively.

File 1

Contains code snippets 1–7 which were used to curate BioSample metadata (see Supplementary Methods for details).