UNIVERSITY COLLEGE LONDON

Role of mobile genetic elements in the global network of bacterial horizontal gene transfer

Mislav Acman

A thesis submitted to University College London for the degree of
DOCTOR OF PHILOSOPHY

September 2021

Declaration

I, Mislav Acman, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Mislav Acman,

London, September 2021

Abstract

Many bacteria can exchange genetic material through horizontal gene transfer (HGT) mediated by plasmids and plasmid-borne transposable elements. One grave consequence of this exchange is the rapid spread of antibiotic resistance determinants among bacterial communities across the world. In this thesis, I make use of large datasets of publicly available bacterial genomes and various analytical approaches to improve our understanding of the nature and the impact of HGT at a global scale. In the first part, I study the population structure and dynamics of over 10,000 bacterial plasmids. By reconstructing and analysing a network of plasmids based on their shared k-mer content, I was able to sort them into biologically meaningful clusters. This network-based analysis allowed me to make further inferences into global network of HGT and opened up prospect for a natural and exhaustive classification framework of bacterial plasmids. The second part focuses on global spreading of bla_{NDM} – an important antibiotic resistance gene. To this end, I compiled a dataset of over 6000 bacterial genomes harbouring this element and developed a novel computational approach to track structural variants surrounding bla_{NDM} across bacterial genomes. This facilitated identification of prevalent genomic contexts of bla_{NDM} and reconstruction of key mobile genetic elements and events which led to its global dissemination. Taken together, my results highlight transposable elements as the main drivers of HGT at broad phylogenetic and geographical scales with plasmid exchange being much more spatially restricted due to the adaptation to specific bacterial hosts and evolutionary pressures.

Impact Statement

Over the past decade, a major public health crisis is arising due to the growing presence of antibiotic resistant pathogens coupled with diminishing incentives of big pharmaceutical companies to develop new antibiotics. For clinically important pathogens, determinants of antibiotic resistance are usually obtained from the environment through a complex and multi-layered network of mobile genetic elements (MGEs) which drive gene exchange between bacteria. Finding alternative ways to combat this imminent threat calls for a better understanding of the selective forces contributing to the spread of antibiotic resistance and the nature of the two main protagonists driving gene exchange: plasmids and transposable genetic elements. In this thesis, I describe two novel methods for studying the network of gene exchange. The first one clusters similar (i.e., related) plasmids and depicts their global population structure as a network thus indirectly mapping the pathways of gene exchange. Aside from providing a broad overview of the latter, the method allows more focused studies informing on the correlates and potential chokepoints and courses of action. The second approach aims to uncover specific genetic events that contributed to the dissemination of a specific resistance element; for instance, mobilization of a gene by a transposable element. This provides mechanistic perspective, gives context to resistance spread and can be further leveraged to estimate the date of underlying events. Both methods facilitate tracking of antibiotic resistance spreading and as such are of clinical relevance, but in essence provide different perspectives on gene exchange. Here, I exemplify their utility on two exhaustive global bacterial genomic datasets and discuss the results while highlighting individual roles of plasmids and transposable elements in the global network of gene exchange.

Acknowledgments

First and foremost, I would like to thank my supervisors who inspired me and helped me develop as a researcher. I experienced my first excitement for science in the laboratory of Prof. Petra Korać at the University of Zagreb. During my studies in Zagreb, Prof. Kristian Vlahoviček was solely responsible for inspiring my early interest in bioinformatics and statistics. I received further exposure to scientific research in Paris while studying at the Centre for Research and Interdisciplinarity (CRI), under the supervision of Ariel B. Lindner and Pascal Hersen. I learned many useful skills at CRI, both soft and research-specific, which undoubtedly enhanced my academic progress thereafter. Prof. Joanne M. Santini, my secondary PhD supervisor at the University College London (UCL), was always there for me when I needed help, whether it was a specific research question, a bit of career advice, or giving an introduction to a specialist in the field. However, my closest scientific mentor throughout my PhD studies has been Prof. François Balloux, who has given tremendous support, both to me personally and to the research presented in this thesis. He gave me freedom to pursue my own ideas and was there to help guide my research when difficulties arose.

I believe every research graduate has few guardian angels hiding in the shadows – beautiful souls there to provide comfort and guidance through the hard times. Zoran S. Marinković and Jeanette Nguyen were my guardian angels in Paris. They provided selfless support throughout my master's studies and helped me adjust to my first time living abroad. My guardian angel in London was Lucy van Dorp – an invaluable mentor, collaborator, and friend during my doctoral training at UCL.

I would also like to extend my gratitude to other mentors I worked with during my university education: Prof. Leonidas Bleris, my supervisor at University of Texas at Dallas (UTD); Benno Schwikowski, my supervisor at Institute Pasteur; Prof. Chris Barnes, my thesis committee chair and advisor; and Georgiadis Vassilis, my industry placement supervisor who provided some exceptional career advice; and Nadine Mogford and Lazaros Foukas for their crucial administrative and student-related support. Furthermore, I would like to thank my friends and colleagues from Office 212, UCL's GEE department, and everyone else who indirectly made my studies a very enjoyable

experience, especially: Liam P. Shaw, Christopher J. Owen, Arturo Torres Ortiz, Sam Morris, dr. dr. Prof. Dave Curtis, Ciaran Bench, and Suryateja Sarma.

I am grateful to the University College London and London Interdisciplinary Biosciences Consortium (LIDo) for supporting me and my research both administratively and financially. It has been an honour to work and study at such esteemed institution.

My doctoral studies would not have been the same without the family who helped me cope with failure and always celebrated my achievements. Even though everyone from my extended family provided an unquestionable support for which I am wholeheartedly thankful, special gratitude goes to my sister Marta, brother Marko and my parents Miroslav and Marijana. Final thanks goes to Ana Marija, a very special person and my endless source of inspiration and motivation who tirelessly accompanies me on my journeys.

Hvala vam od srca.

Table of Contents

List of Figures

List of Tables

Acronyms

AMR	antimicrobial resistance.
ANOVA	analysis of variance.
API	application programming interface.
att	attachment site.
BEAST	Bayesian evolutionary analysis by sampling trees.
BHR	broad host range.
BIGSI	bitsliced genomic signature index.
bla	beta-lactamase (gene).
BLAST	basic local alignment search tool.
bp	base pairs.
CDS	coding sequence.
CRISPR	clustered regularly interspaced short palindromic repeats.
CRKP	carbapenem-resistant *Klebsiella pneumoniae*.
DNA	deoxyribonucleic acid.
DR	direct repeat.
dsDNA	double stranded DNA.
dso	double stranded origin (of replication).
DTR	DNA transfer replication.
ESS	effective sample size.
GC	guanine-cytosine.
GI	genomic island.
GO	gene ontology.
GTR	generalised time reversible.
HGT	horizontal gene transfer.
HMM	hidden Markov model.

ICE	integrative conjugtive element.
IME	integrative mobilizable element.
Inc	incompatibility.
IR	inverted repeat.
IS	insertion sequence.
ISCR	IS common region.
JD	Jaccard distance.
JI	Jaccard index.
kb	kilobase pairs.
KPC	*Klebsiella pneumoniae* carbapenemase.
MCMC	Markov chain Monte Carlo.
mcr-1	mobilized colistin resistance 1 (gene).
MGE	mobile genetic element.
MIC	mobile insertion cassette.
MIC	minimum inhibitory concentration.
MITE	miniature inverted repeat.
ML	maximum likelihood.
MOB	mobility/relaxase (protein).
MPF	mating-pair formation.
MSA	multiple sequence alignment.
NCBI	National Center for Biotechnology Information.
NDM	New Delhi metallo-β-lactamase.
NHR	narrow host range.
NMI	normalized mutual information.
OMV	outer-membrane vesicle.
orf	open reading frame.
oriIS	origin of IS transfer
oriT	origin of (conjugative) transfer.

oriV	origin of vegetative replication.
OSLOM	order statistics local optimization method.
PBRT	PCR-based replicon typing.
PCR	polymerase chain reaction.
pMLST	plasmid multi-locus sequence typing.
PSI-BLAST	position-specific iterative BLAST.
RC	rolling circle.
rep	replication (gene).
res	resolution (gene).
SD	standard deviation.
SNP	single-nucleotide polymorphism.
SRA	Sequence Read Archive.
ssDNA	single stranded DNA.
sso	single stranded origin (of replication).
T4CP	type IV coupling protein.
T4SS	type IV secretion system.
TE	transposable element.
terIS	termination of IS transfer.
Tn	transposon.
Tnp	transposase (protein).
tra	transfer (gene).
TU	translocatable unit.
WGS	whole genome sequencing.

Chapter 1

Introduction

Bacteria are microscopic single-cell organisms which are an integral part of world's ecosystems. They are responsible for many biological processes essential for maintaining life on Earth, for example synthesis and decomposition of nutrients, immunity, nitrogen fixation and photosynthesis. As is always in life, some are harmful to other organisms and act as pathogens causing diseases. One remarkable ability of bacteria, and especially pathogens, is the exchange of genetic information, i.e., horizontal gene transfer (HGT). This feature allows bacteria to rapidly adapt to new environmental conditions by recruiting genes encoding specific functions, such as alternative metabolic pathways, virulence factors, resistance to toxic metals, stress response factors, and, most importantly, antibiotic resistance. It is estimated more than 10^9 bacterial species inhabit the biosphere. With this immense diversity, one can only imagine the complexity and the scope of the gene exchange among bacteria.

In this thesis, I aim to elucidate some basic properties of bacterial gene exchange and bring about better understanding of how gene transmission networks operate at a global scale. The introductory Chapter 1 talks about the concept and current perception of HGT with specific focus on two main actors: bacterial plasmids and transposable genetic elements (TEs). These two elements are heavily interdependent which poses certain challenges when attempting to pick apart their evolution and role in HGT. Hence, in Chapter 2 and Chapter 3, I present a new method for clustering related bacterial plasmids which opens new prospects for studying plasmid evolution as well as global network of HGT. On the other hand, Chapter 4 zooms-in on a specific case of worldwide resistance spreading focusing more on local genome reshuffling and 'plasmid hopping' caused by TEs. Taken together, results presented in this thesis expand our understanding of the interplay between plasmids and TEs and incite a paradigm shift regarding bacterial gene exchange – all of which is discusses in detail in Chapter 5. In Appendix C, I list some additional scientific research I contributed to during my doctoral training.

1.1. Horizontal gene transfer in bacteria

Horizontal gene transfer (HGT), also called lateral gene transfer, is defined as the exchange of genetic information between organisms that are not in a direct parent-offspring relationship (Snyder & Snyder, 2013). It occurs across and between all three domains of life (Husnik & McCutcheon, 2018). However, it is especially important in Bacteria and Archaea where it represents a major evolutionary driver by contributing to the fitness of different lineages and allowing organisms to adapt to various environmental stresses (Soucy et al., 2015; Vos et al., 2015). For instance, in a clinical setting, a pathogenic bacterial strain can compromise antibiotic treatment by acquiring a particular resistance gene which is why antibiotic resistance is considered one of the biggest threats to global health, food security, and development (WHO, 2021). In bacteria, HGT is driven by mobile genetic elements (MGEs) which are defined as segments of DNA encoding functions that mediate the displacement of DNA within or between bacterial cells (Frost et al., 2005; Snyder & Snyder, 2013; Soucy et al., 2015). MGEs are broadly split into three categories: bacteriophages (or phages for short), plasmids and transposable genetic elements (TEs) (Frost et al., 2005; Thomas & Nielsen, 2005).

Bacteriophages are bacterial viruses. They can exist in the nature as a virion, i.e. virus particle, but their life cycle is tightly bound to a bacterial cell from which they hijack resources and protein synthesis machinery (Mc Garth & Sinderen, 2007; Snyder & Snyder, 2013). Phages are mostly host specific and have some features resembling a living organism. Once inside a bacterial host, they replicate and pack their genomes into a synthesized protein capsids thus forming a virion. Apart from recently discovered huge phages, their genomes as well as capsids are typically much smaller compared to their bacterial hosts (Al-Shayeb et al., 2020; Yuan & Gao, 2017). They are known to contribute to the spread of antibiotic resistance or factors linked to pathogenicity (Balcazar, 2014; Penadés et al., 2015). However, the hallmark of phages is their immense influence on global ecosystems which they achieve by sweeping through bacterial populations (Breitbart et al., 2018; Emerson et al., 2018). Not surprisingly, phages are considered the most widely distributed biological entities with an estimated global viral population of 10^{31} (Hatfull, 2008) which is ten billion times more than the number of stars in the universe.

Moving down the ladder of complexity are plasmids and TEs. These naked molecules of DNA reside and replicate within the bacterial host. Their existence is thought to be primarily driven by successful transmission and survival (Park & Zhang, 2012; Soucy et al., 2015). To achieve this, plasmids and TEs employ different strategies such as multiple propagation mechanisms and addiction systems, but mostly they thrive by harbouring non-essential (accessory) genes that modulate the fitness and consequently the evolution of their bacterial host (Frost et al., 2005). Some prominent examples are accessory genes encoding toxin-antitoxin systems, virulence factors, alternative metabolic pathways, antibiotic biosynthesis, metal resistance and antimicrobial resistance (AMR). Plasmids, in their basic form, are small and generally circular DNA molecules residing separate from the host chromosome and rely on the host replication and protein synthesis machinery (Snyder & Snyder, 2013). TEs are even less autonomous as they only exist embedded in another DNA sequence usually a host chromosome or a plasmid. TEs move by 'jumping' between different DNA molecules and spread horizontally usually by hitchhiking on other mobile elements, primarily plasmids.

From a mechanistic perspective, the process of HGT is generally split into three categories: conjugation, transformation, and transduction (Dale & Park, 2010; Snyder & Snyder, 2013; Soucy et al., 2015). Conjugation primarily concerns transfer of conjugative plasmids between bacterial cells, but the concept expands to conjugative transposons too (Burrus et al., 2002). Conjugation is often regarded as the bacterial equivalent of sexual mating as it is the only mechanism which involves direct contact and exchange of genetic information between two unrelated bacterial cells. Conjugation involves a donor cell which contains a conjugative plasmid and a recipient cell which receives the plasmid via a rod-like protein complex called the sex pilus. Plasmids and TEs containing all the necessary genes to complete the conjugation are called self-transmissible (or transmissible) (Garcillán-Barcia & de la Cruz, 2013). Mobilizable MGEs encode some parts of the conjugation machinery and rely on self-transmissible MGEs to provide the missing functions. Other MGEs are referred to as non-mobilizable. Conversely, transformation relates to bacterial cell directly up taking genetic material from the environment, for example plasmids (I. Chen & Dubnau, 2004). In nature, bacteria generally employ transformation under stressful environmental conditions, such as heath shock, ion imbalance, DNA damage, or exposure to chemical agents, but some bacteria are naturally transformable (i.e., competent) (Mell & Redfield, 2014). Finally, transduction is a consequence of phage replication cycle where while packing or

replicating their viral genomes, bacteriophages can mistakenly pick-up and disseminate DNA of the host they infected (Snyder & Snyder, 2013).

Although the origin and the evolution of MGEs (and other viruses) is shared and highly intertwined (Koonin et al., 2020; Kazlauskas et al., 2019; Krupovic et al., 2009), the life cycle of phages seems more autonomous and consistent. Plasmids and TEs are, on the other hand, much more interdependent, and their genomes are subject to frequent rearrangements thanks to their simpler nature and lack of physical constraint of a capsid. Furthermore, plasmids and TEs are considered the more dominant entities driving HGT, especially among pathogenic bacteria (Dolejska & Papagiannitsis, 2018; Lerminiaux & Cameron, 2019). This in turn yields some complex but fascinating dynamics of gene exchange which will be further explored in forthcoming chapters.

1.1.1. Shoutout to pioneering discoveries

In 1928, the British bacteriologist Frederick Griffith demonstrated that a nonvirulent strain of *Streptococcus pneumoniae* (pneumococcus) can acquire factors of virulence from a virulent strain previously killed by heat (Griffith, 1928). This was the first evidence of bacterial transformation. Griffith lucidly described this phenomenon: *"In the nidus thus formed the pneumococcus gradually builds up from material furnished by its disintegrating companions an anti-genic structure with invasive properties sufficient to cope with the resistance of its host."* (Griffith, 1928). However, it took almost 20 years until Avery, MacLeod, and McCarty recognized that DNA was encoding information related to pneumococcus virulence (Avery et al., 1944), and for Tatum and Lederberg to recognise transformation as a new mode of inheritance in bacteria (Jacob & Wollman, 1961; Tatum & Lederberg, 1947).

Around the same time, Barbara McClintock was making her breakthrough discovery of TEs that can change position on chromosomes of maize (McClintock, 1950). The first prokaryotic TE was described much later in 1963 by Austin L. Taylor (Shapiro, 1983; Taylor, 1963), as the bacteriophage Mu which, once integrated in the host genome, was found capable of replicative transposition. Insertion Sequences (IS), i.e., prokaryotic TEs in their simplest form, were discovered several years after (Jordan et al., 1968; Shapiro, 1969, 1983).

Interestingly, Mu was in fact not the first bacteriophage to have been discovered. Their existence has been noted many times throughout history (Abedon et al., 2011), and even as early as 19th century (Frankland, 1895; Hankin, 1896). Still bacteriophage were not formally recognized as bacterial viruses prior to the works of Twort (Twort, 1915) and D'Hérelle (D'Hérelle, 1917) likely due to their small size and complex culturing procedure.

The last element to be discovered were plasmids. The name *plasmid* was introduced by Joshua Lederberg and was intended to encompass all extrachromosomal hereditary determinants (Lederberg, 1952), but was later refined to the term used today. Conjugation via F-plasmid was first observed in 1946 (Lederberg & Tatum, 1946; Snyder & Snyder, 2013), but in these early years scientists could not distinguish it from a bacteriophage due to similar experimental phenotypes (Summers, 1996). This notion persisted until 1961 when Marmur et al. showed that F-plasmids are a DNA molecule existing as an entity separate to the host chromosome (Clowes, 1972; Marmur et al., 1961).

There certainly were many other remarkable scientific advancements that have contributed to our understanding of HGT and MGEs. Nevertheless, the early discoveries mentioned here paved the way for research into HGT, spurred new scientific disciplines such as plasmid and phage biology, and made a profound impact on our understanding of microbial genetics and evolution and beyond.

1.2. Bacterial plasmids

For the most part, plasmids are replicons, i.e. (semi-)autonomously replicating DNA molecules (Pinto et al., 2012). They exist in a circular form apart from few linear plasmids and can vary in size from a few thousand to several hundred thousand base pair (bp) long mega-plasmids (Sitter et al., 2021; Stolz, 2014). Aside from their circular form, in rare occasions plasmids can exist integrated into their host chromosome. This mechanism is thought to prevent the loss of helpful genes in the bacterial population (Carroll & Wong, 2018; Hülter et al., 2017). Plasmids can propagate horizontally via conjugation or transduction but are also inherited vertically from parent to offspring in which case plasmid replication and partitioning machineries play an important role.

1.2.1. Replication

The genes encoded by plasmids are generally split into two groups: accessory genes, such as those involved in resistance or virulence, and core genes which encode functions necessary for plasmid conjugation, replication, partitioning and maintenance (Snyder & Snyder, 2013; Tolmasky & Alonso, 2015). The replication of a plasmid begins at the origin of replication: *oriV* for vegetative (intracellular) replication, and *oriT* in case of plasmid conjugation (Figure 1.1). In alphaproteobacterial plasmids, *oriV* together with the surrounding replication initiation (*repABC* or other *dnaA*-like genes), partitioning (*par*), and other genes form the plasmids' replicon region (Petersen, 2011; Pinto et al., 2012). This set of genes can constitute: (i) a *basic replicon* if it encodes the necessary genes and controls plasmid replication through a regulatory network; or (ii) a *minimal replicon* if the replication is not completely autonomous which is reflected in alternating plasmid copy number (Lilly & Camps, 2015). Plasmids whose replication is regulated by non-plasmid initiation factors are referred to as *trans-ori*.

The three described types of vegetative replication are theta, rolling-circle (RC), and strand displacement which is limited to the IncQ-family of plasmids (del Solar et al., 1998; Loftie-Eaton & Rawlings, 2012; Snyder & Snyder, 2013). Theta replication (Figure 1.1A) is the most common form of plasmid replication in Gram-negative bacteria, such as proteobacteria (Snyder & Snyder, 2013). At the beginning of the theta replication, the circular DNA duplex is opened at the *oriV* and the replisome – a protein complex that carries out DNA replication – is built from host encoded DNA polymerases and other factors (Lilly & Camps, 2015; Snyder & Snyder, 2013). Plasmid replication resembles the bacterial chromosomal replication in as much as an RNA primer is bound, the replisome is formed, and the replication continues either unidirectionally (i.e., only one replication fork is formed) or bidirectionally. During replication, an intermediate DNA structure resembling the Greek letter θ (theta) is formed, which served as an inspiration for the name.

RC plasmids are prevalent in gram-positive bacteria such as *Staphylococcus*, *Streptococcus*, *Bacillus*, *Clostridium*, *Lactococcus* and others, but are also found in Gram-negatives (Khan, 1997, 2005). RC replication (Figure 1.1B) is initiated when the Rep protein nicks one strand of the plasmid duplex at the double-strand origin (*dso*) (Ruiz-Masó et al., 2015). The nick generates a 3′-OH which *de facto* replaces the RNA primer and allows the host polymerases to initiate the leading strand replication. During replication, the nicked DNA strand is displaced "hanging" attached to the Rep protein

complex. Upon completion, the nicked strand is re-joined and released as a circular single-stranded (ss)DNA. The synthesis of a dsDNA from the ssDNA intermediate is initiated without the primer at a single-strand origin (*sso*).

The plasmid replicon region is essential in controlling plasmid copy number. They contain motifs with varying affinity for binding of Rep proteins and host factors, for instance iterons, AT rich areas, GC rich areas, and methylation patterns (Lilly & Camps, 2015). Iterons (Figure 1.1A) are 17-22bp direct repeat DNA segments which bind Rep proteins to initiate plasmid replication but also serve as negative regulators, meaning plasmid copy number decreases when iteron concentration is high (Chattoraj, 2000; Snyder & Snyder, 2013). This is done via two mechanisms: (i) transcriptional autoregulation where the Rep protein is inhibiting its own promoter by binding a proximal iteron sequence; (ii) a handcuffing mechanism where plasmids in high concentration are coupled (i.e., "handcuffed") together via iteron-bound Rep proteins thus inhibiting replication initiation (Chattoraj, 2000; Snyder & Snyder, 2013). Finally, plasmid replication can be negatively regulated by antisense RNA in a process sometimes called counter transcribed RNA (ctRNA) (Brantl, 2014; del Solar et al., 1998; Pinto et al., 2012). In this case, any change in plasmid concentration is reflected in the change of concentration of antisense RNAs which can increase inhibition of a function essential for replication usually by targeting gene transcription.

1.2.2. Maintenance and partitioning

Plasmids present a metabolic burden for the bacterial host as they require the cells' resources and energy to be maintained (Shintani & Suzuki, 2019; G. Wu et al., 2016). In nature, plasmids evolved several mechanisms to counter this negative selection pressure and to prevent being cured from the bacterial population (Snyder & Snyder, 2013). These include plasmid copy number optimisation, harbouring of beneficial genes and addiction systems, resolving of plasmid multimers, and controlled plasmid partitioning during cell division.

As mentioned earlier, copy number is regulated by the plasmid's replicon region. In terms of copy number control, plasmids are considered either relaxed or stringent (Dale & Park, 2010). Plasmids with relaxed copy number control usually regulate only the upper limit of the plasmid count, they are smaller in most cases, and have a more variable, but higher copy number. Stringent plasmids, such as P1, R, and F plasmids, come in lower

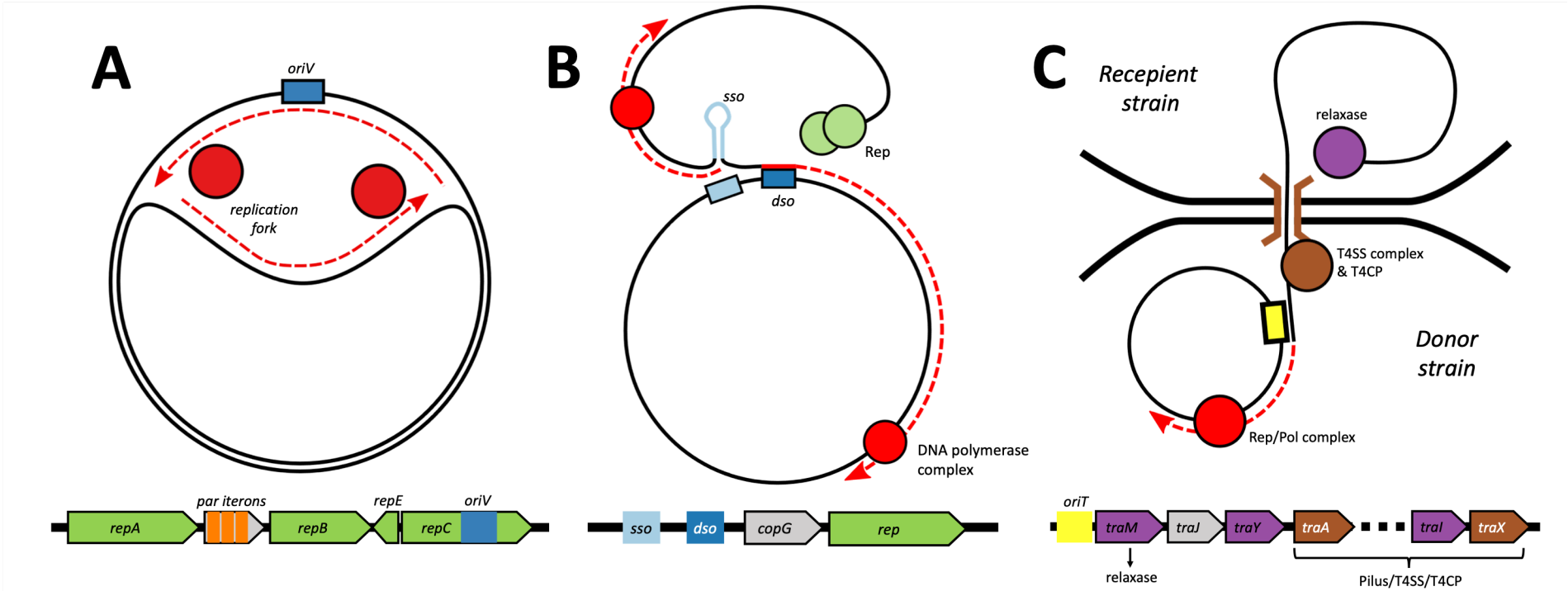


Figure 1.1. Modes of plasmid replication and representative gene encoding regions. (A) Theta replication starts at the *oriV* site where DNA duplex is melted, and RNA primers are bound. The replication proceeds from *oriV* either unidirectionally or bidirectionally. (B) Rolling circle (RC) replication is initiated by a single stranded nick at the *dso* site by the Rep protein. Free 3'-OH is attached to the *dso* serves as a primer for DNA polymerisation which continues around the circular DNA strand and unwinds the nicked DNA strand. Upon completion, another nick is made at *dso* releasing a linear ssDNA. The Rep protein rejoints the two ends forming a second circular ssDNA. The replication is completed after polymerisation of the second strand which is initiated at the *sso* hairpin. Iterons (A) and *copG* (B) are regulatory elements sometimes found in plasmid replicons which control the plasmid copy numbers. (C) During plasmid conjugation one DNA strand is nicked at *oriT* site by the relaxase and unwound in a fashion resembling RC replication. The relaxase carries the nicked DNA strand through the mating-pair formation (MPF) into the recipient strain where the strand is rejointed and the new strand polymerised thus completing the conjugation.

copy numbers and exert more control over their replication as they often present a higher burden for the cell. During bacterial cell division, high-copy-number plasmids are more likely to be inherited with the cytoplasm by chance thanks to their sheer abundance inside the cell. Low-copy-number plasmids use a partitioning system to ensure at least one copy of the plasmid ends up in each daughter cell (Ghosh et al., 2006; Snyder & Snyder, 2013). Plasmid partitioning function is analogous to the segregation of bacterial chromosomes, it is mostly self-contained and carried out by Par proteins. There are several classes of plasmid partitioning systems, but all have a similar underlying mechanism (Ghosh et al., 2006; Salje, 2010). In general, a smaller adaptor protein binds DNA repeats (iterons) in a centromere-like region thus aggregating plasmid copies. A larger motor protein then interacts with the complex and actively moves plasmid copies to opposite poles of the dividing cell either by filament polymerisation or pulling against shrinking filaments.

Furthermore, like bacterial chromosomes, plasmids are prone to forming dimers and multimers post replication due to efficient homologous recombination systems in bacteria. This can compromise plasmid stability within a bacterial population (Friehs, 2004; Summers & Sherratt, 1984). Therefore, plasmids rely on a variety of site-specific resolution (*res*) systems which broadly fall into serine or tyrosine recombinase families (Crozat et al., 2014; Partridge et al., 2018).

An additional mechanism used by plasmids to ensure their persistence are addiction systems, also called toxin-antitoxin or post-segregational killing systems (Carroll & Wong, 2018; Tsang, 2017). Generally, the system works such that the toxin is chromosomally produced or longer lived than the antitoxin. Therefore, individual bacterial cells which inhibit translation or fail to inherit a copy of antitoxin-carrying plasmid are wiped out from the population. Prominent examples of such systems include MazEF, PhD-Doc systems, *hok*/*sok*, and restriction-modification system found in *Bacillus* species (Engelberg-Kulka et al., 2005; Gerdes et al., 1986; Hazan et al., 2001; Kulakauskas et al., 1995).

1.2.3. Conjugation

Conjugation is a unique feature of many plasmids and some TEs that allows them to replicate and transfer between two unrelated bacterial cells which are in direct contact. Conjugation for the most part seems restricted to Proteobacteria and Firmicutes phyla and the conjugation machinery which enables this direct exchange of DNA the falls into a

larger type IV secretion system (T4SS) superfamily of protein complexes (Grohmann et al., 2018; Smillie et al., 2010). Conjugation is a complex function encoded by the transfer (*tra*) region of a plasmid sequence (Clewell, 1993; Dale & Park, 2010; Snyder & Snyder, 2013), and the process unravels in two steps: Mating-pair formation (MPF) and DNA transfer replication (DTR).

MPF occurs when donor and recipient bacterial cells come in contact and T4SS is formed (Figure 1.1C) (Grohmann et al., 2018). The T4SS-pilus plays an important role. The pilus is a multimeric tube-like protein structure encoded by the pilin protein. It comes in all shapes and sizes and is presumed to be involved in mate-seeking, formation of the T4SS, and DNA channelling during conjugation, and even passively in biofilm formation in the environment (Babić et al., 2008; Ghigo, 2001; Grohmann et al., 2018; B. Hu et al., 2019). Another important constituent of MPF is a coupling protein, often labelled as type IV coupling protein (T4CP), serving as a link between DTR and MPF which results in the final translocation of donor DNA through a cytoplasmic membrane (Guglielmini et al., 2014; Smillie et al., 2010).

DTR concerns the preparation of the plasmid DNA molecule for transfer (Figure 1.1C). In many ways DTR resembles and likely evolved from the RC replication systems in plasmids (Garcillán-Barcia et al., 2009). During DTR, the relaxosome is the central protein complex which is formed at the *oriT* site and together with the coupling protein coordinates all the subsequent activities (Grohmann et al., 2018; Smillie et al., 2010). A site-specific endonuclease called relaxase (MOB for short) creates a single-stranded nick within the *oriT* site. Then, a plasmid-encoded helicase unwinds the nicked strand. Resulting ssDNA, together with the covalently bound relaxase, is transferred into a recipient cell via the aforementioned T4SS. Once there, the relaxase seals the nick and the DNA-polymerase synthesizes the missing DNA chain thus completing conjugation.

1.2.4. Classification schemes

Despite being prone to frequent genome rearrangements, some core plasmid genes stay relatively conserved across different plasmid backbones, and hence can be used to infer evolutionary histories (Garcillán-Barcia et al., 2009). The presence of core plasmid genes has led to the development of several plasmid classification or typing schemes, with the two most widely used relying on the plasmid replicon region and MOB genes (Orlek, Stoesser, et al., 2017).

Plasmids that share the same replication and partitioning mechanisms cannot stably coexist within a bacterial population as they compete for the common elements (Couturier et al., 1988). Historically, such plasmids were deemed incompatible. Incompatibility of plasmid replicons can be determined using lab-based methods such as PCR-based replicon typing (PBRT) (Carattoli et al., 2005) or replicon probe hybridisation (Couturier et al., 1988). However, genome sequencing and the advancement of bioinformatics analyses has led to several *in silico* typing resources, such as PlasmidFinder (Carattoli et al., 2014) or the plasmid Multi Locus Sequence Typing (pMLST) database. These tools inherently rely on experimentally verified replicon sequences and consequently have a taxonomic range limited to culturable or pathogenic bacteria from the *Enterobacteriaceae* family and several well-studied genera of gram-positive bacteria (Jensen et al., 2010; Lozano et al., 2012; Orlek, Stoesser, et al., 2017; Shintani & Suzuki, 2019). In addition, plasmids can be associated to multiple replication types due to the plasticity of their genomes and frequent genome rearrangements (Orlek, Stoesser, et al., 2017; Shintani et al., 2015). Presently, there are 132 and 141 replicon sequences available on PlasmidFinder for classification of plasmids from *Enterobacteriaceae* and Gram-positive bacteria respectively (Carattoli & Hasman, 2020). As an example, these can classify plasmids to approximately 27 Inc groups in the *Enterobacteriaceae* family, and 14 and 18 in the *Pseudomonas* and *Staphylococcus* genera, respectively (Shintani & Suzuki, 2019).

Based on their ability to conjugate, plasmids are also considered self-transmissible, mobilizable, or non-mobilizable (Smillie et al., 2010). Self-transmissible plasmids encode all the proteins required for conjugation, while mobilizable plasmids encode only a subset and consequently need to borrow parts of the conjugation machinery from other MGEs. Mobilizable and self-transmissible plasmids can be classified using MOB typing scheme into six MOB types (MOB_C, MOB_F, MOB_H, MOB_P, MOB_Q, and MOB_V) (Garcillán-Barcia et al., 2009; Smillie et al., 2010). The MOB typing scheme relies on the conserved N-terminal sequence of the aforementioned relaxase protein. Similar to replicon typing, there are some database-reliant *in silico* MOB typing resources available, such as MOB-suite (Robertson & Nash, 2018), MOBscan (Garcillán-Barcia et al., 2020), or MOBtyping software (Orlek, Phan, et al., 2017b). In addition, self-transmissible and mobilizable plasmids can be classified according to their MPF machinery into MPF_F, MPF_G, MPF_I, and MPF_T classes (Smillie et al., 2010). However, this form of plasmid typing is not frequently used.

1.2.5. Plasmid host range

Genes governing plasmid replication and conjugation are considered to be the major determinants of plasmid host range, i.e. the range of bacterial species in which a plasmid can persist and replicate (Shintani & Suzuki, 2019). The scientific community traditionally bins plasmids into two polar groups: so-called narrow-host-range (NHR) plasmids which only persist in closely related organisms, and broad-host-range (BHR) plasmids which occur in more distantly related hosts. In fact, some BHR plasmids can cross the domains of life. For example, Ti plasmids can transfer part of themselves from *Agrobacterium tumefaciens* into a plant cell via mechanism resembling conjugation (Brencic & Winans, 2005). Nevertheless, the definitions of plasmid range broadness and narrowness are still controversial for a series of reasons (Shintani & Suzuki, 2019). Plasmid host range is likely more a transient feature than a fixed one, and there are multiple factors affecting host range, such as dependence on a host replication machinery and transcriptional signalling, fitness costs, and presence of multiple plasmid replicons (A. Jain & Srivastava, 2013; Shintani & Suzuki, 2019). Furthermore, plasmid host range remains determined primarily based on limited empirical observations.

Despite the lack of clear-cut definitions, it is sensible to consider BHR plasmids as likely the most important for dissemination of AMR and other bacterial phenotypes (A. Jain & Srivastava, 2013). Some prominent examples of such BHR plasmids include those belonging to the IncP, IncW, IncN, and IncQ replicon types. These plasmids are found across the *Enterobacteriaceae* family and in some *Pseudomonas* strains (Götz et al., 1996; Popowska & Krawczyk-Balska, 2013). A list of other well-characterized BHR plasmids is provided in a review by Jain and Srivastava (2013).

Aside from genes involved in replication and conjugation, other features of the plasmid sequence such as plasmid size, nucleotide composition, and replication strand asymmetry could help determine likely bacterial hosts (Nishida, 2012; Shintani & Suzuki, 2019). For instance, it has been observed that plasmid GC content tends to be slightly lower than, but strongly correlated with, the GC content of the host chromosome (Nishida, 2012). Currently, there are two observations / hypotheses touching on this phenomenon: (i) a lower energy cost of A/T codons for the bacterial host (Rocha & Danchin, 2002), and (ii) xenogeneic silencing of the plasmid (San Millan & MacLean, 2017; Suzuki-Minakuchi & Navarre, 2019). Xenogeneic silencing in bacteria pertains to H-NS (heat-stable nucleoid structuring) proteins which silence transcription of foreign DNA atypically rich in AT compared to the host chromosome (Suzuki-Minakuchi & Navarre,

2019). By employing xenogeneic silencing, a bacterial host can lower the cost of maintaining a particular MGE and thus preserve a particularly useful virulence or AMR gene. More importantly, the nucleotide composition of plasmids could have evolved as a trade-off between GC energy costs and xenogeneic silencing (San Millan & MacLean, 2017).

Another potentially interesting predictor of plasmid host range is the oligonucleotide composition where plasmids tend to have a similar oligonucleotide composition to their known host chromosomes (Suzuki et al., 2008). Nevertheless, despite interesting correlations found between features of the plasmid backbone and the corresponding bacterial hosts, the factors influencing the host range are still vast, intertwined, and difficult to capture using a finite set of rules (R. J. Sheppard et al., 2020; Shintani & Suzuki, 2019).

1.3. Transposable genetic elements

Transposable genetic elements (TEs) are a type of MGEs which can repeatedly move or copy within a genome, not necessarily within the same DNA molecule. Unlike most bacteriophages or plasmids, TEs are not free form but integrated into a host DNA molecule. The mechanism by which a TE moves is called transposition, and the proteins enabling this movement are called transposases (Tnp) (Dale & Park, 2010; Kusumoto & Hayashi, 2019; Snyder & Snyder, 2013). Broadly speaking there are two mechanisms of transposition: *copy-and-paste* where a TE makes a copy of itself in a new location in the DNA and *cut-and-paste* where the whole TE is excised and moved. Thus far, *Copy-and-paste*, or replicative transposition, appears to be the most common mechanism of TE movement in prokaryotes (Kusumoto & Hayashi, 2019).

Depending on their structural organisation, most common TEs are categorized into insertion sequences (ISs), composite transposons, and non-composite (or unit) transposons (Figure 1.2) (Dale & Park, 2010; Snyder & Snyder, 2013). In addition to these common forms, there are other TEs which can, to some extent, catalyse their self-integration (Roberts et al., 2008). These include Integrative Conjugative Elements (ICEs) and Integrative Mobilizable Elements (IMEs), genomic islands, integrons, integrated prophages, and IStrons. Databases and other online resources are readily available to provide classification for various forms of TEs. Most widely used examples include the Transposon Registry (Tansirichaiya et al., 2019) and ISFinder (Siguier et al., 2006).

1.3.1. Insertion sequences (IS)

Insertion sequences are the simplest form of TEs (Figure 1.2A). They are small and genetically compact sequences raging 0.7-2.5 kb (Siguier et al., 2014, 2015; Snyder & Snyder, 2013). They are comprised of a single *orf* which encodes for Tnp. If a second gene is present, it likely encodes a regulatory protein. Flanking this coding region are two inverted repeats (IRs), approximately 10-40 bp in size and named IRL (left) and IRR (right) depending on the direction of *tnp* gene transcription. These mark the borders of the IS as they are recognized and cleaved by the Tnp. An additional feature of ISs are 3-14 bp long direct repeats (DR) on either side of the IS which are occasional artefacts of the transposition process.

At the time of writing, prokaryotic ISs are classified into 26 families most of which are defined based on the Tnp they use and some based on conserved catalytic sites or conserved IRs (Siguier et al., 2006). There are three main types of Tnps in bacteria named after their catalytic sites: DDE, DEDD and HUH (Siguier et al., 2015). DDE Tnps have a catalytic site with conserved triad of amino acids (Asp, Asp, Glu) that coordinate two Mg^{2+} ions essential in the reactions of DNA cleavage and integration (Hickman & Dyda, 2015). They share crystal structure like retroviral integrases and eukaryotic TEs and are thought to be most widely spread in bacteria. DEDD Tnps share similar structural topology in their catalytic sites as DDE Tnps and likely have a similar transposition chemistry, but the overall mechanism is presumed to be different (Siguier et al., 2015). DEDD Tnps are limited to the IS*110* family whose member have smaller IRs, do not produce DR upon insertion, and involve Holiday-junction intermediates (Siguier et al., 2017). Lastly, HUH Tnps have a His-hydrophobe-His amino acid triad in their catalytic site (Chandler et al., 2013; Siguier et al., 2015). In prokaryotes, HUH Tnps are limited to IS91 and IS200/IS605 families. Aside from these, the HUH protein superfamily of single stranded nucleases also includes Rep proteins involved in bacteriophage and plasmid RC replication and MOB proteins involved in conjugation. More details about various molecular mechanisms of transposition are given in reviews by Hickman & Dyda (2015, 2016) and Siguier et al (2014, 2015, 2017).

The IS91 family employing HUH-Tnps is particularly interesting as it includes IS Common Region elements (ISCR; Figure 1.2C) which are found proximal to many AMR genes (Toleman et al., 2006). ISCR elements, like other IS91 elements, do not contain IRs and are presumed to employ a RC form of transposition (Ilyina, 2012; Toleman & Walsh, 2010). The transposition of ISCR element is initiated at the *oriIS* located

approximately 300 bp downstream of the *tnp* gene and stops at *terIS* (found ~120 bp upstream in case of ISCR2). It is thought that the slippage at *terIS* during RC transposition enables mobilization of genes located upstream of the element (Ilyina, 2012; Toleman et al., 2006, 2012).

Miniature Inverted Repeat Transposable Elements (MITEs; Figure 1.2B) are another special case of IS (Delihas, 2011; Siguier et al., 2015). These are small non-autonomous fragments of IS containing just IRs. Thus, they only transpose in *trans* by a related Tnp. MITEs were found to carry function/structure protein motifs and promoter regions. Nevertheless, other special cases of MITEs exist, such as integron mobilization units (IMUs) (Poirel et al., 2009) and mobile insertion cassettes (MICs) (Y. Chen et al., 1999) which are larger and carry integron structures and gene cassettes respectively.

1.3.2. Composite and non-composite transposons

When two matching ISs sequences enclose a set of genes they can form a composite (or compound) transposon (Figure 1.2D) (Snyder & Snyder, 2013). In that case, Tnp can mobilise the whole cassette by acting upon the farthest IRs (the outer ends). While the flanking ISs may maintain some independence and excise themselves from the cassette, persistent mutations can render one of them non-autonomous and leave only one *tnp* gene active or functional. Similarly, mutations in one or both inner IRs can enhance transposition of the composite transposon by reducing the chances of excision of individual ISs. Composite transposons can have the bracketing ISs facing the same way or in the opposite direction. One basic example is Tn*10* (Figure 1.2D). Tn*10* was one of the first composite transposons to be described that employs a *cut-and-paste* mechanism of transposition (Bender & Kleckner, 1986). In its basic form it encodes the *tetR* AMR gene which is flanked by two inverted IS*10*.

Non-composite (also called unit or non-compound) transposons could be considered a more complex form of a basic IS (Figure 1.2E). In this respect, accessory (usually resistance) genes are an integral part of a functional IS, meaning they are flanked by just the two IRs with no possibility of an alternative transposition unlike in the case of composite transposons (Snyder & Snyder, 2013). One example is Tn*3* (Figure 1.2E), a well-studied and prevalent non-composite replicative transposon family among prokaryotes (Chandler, 2016; Siguier et al., 2014). Like many other non-composite

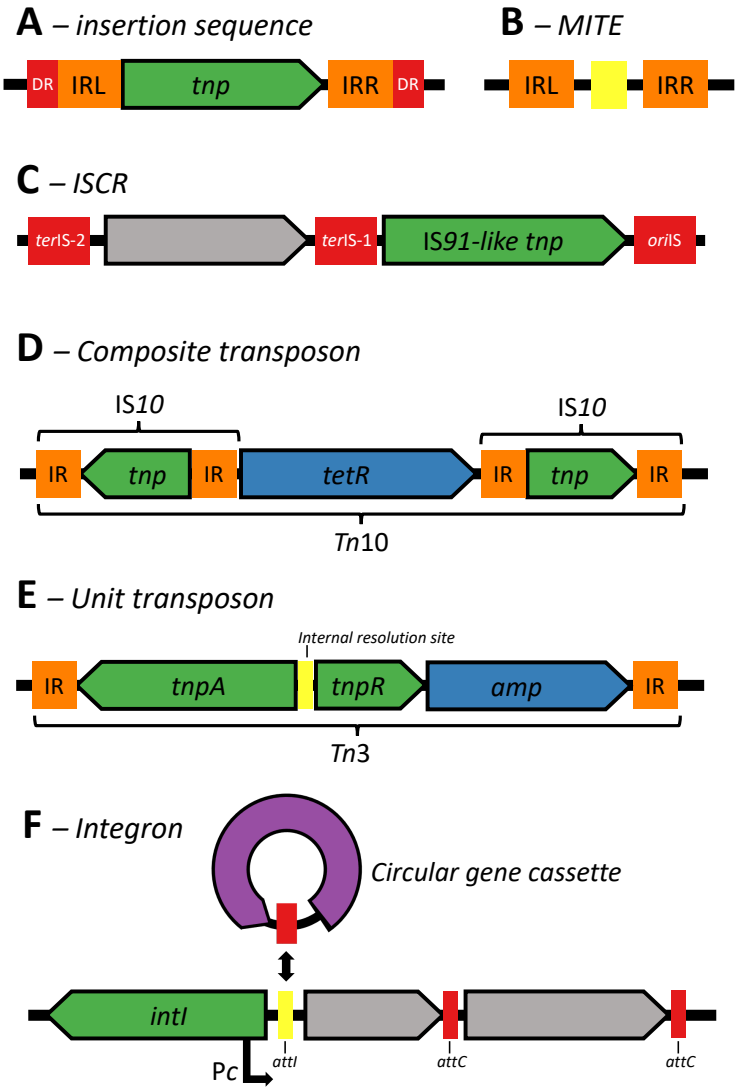


Figure 1.2. Schematic representation of the possible genetic assortments of different transposable genetic elements (TEs). Insertion sequence (IS) is the basic form of TE (**A**). It is comprised of transposase gene (*tnp*) and left (IRL) and right (IRR) inverted repeats. In some cases, direct repeats (DR) are present. These are short artefacts of the transposition process. Miniature Inverted Repeat Transposable Elements (MITEs; **B**) employ transposases of other TEs as they are only comprised of IRs and occasional accessory genes. Insertion sequence common region (ISCR) elements employ a rolling circle transposition (**C**). ISCR transposition starts at the origin of transposition (*ori*IS) and ends at the terminus site (*ter*IS). Occasional slipping at the *ter*IS allows mobilization of genes upstream of ISCR. Composite transposons (D) are made of two flanking ISs. In this example, *Tn*10 transposon is made of tetracycline repressor (*tetR*) gene flanked by two IS*10*. Unit transposons (E) can be regarded as a complex form of IS. *Tn*3, for example, is comprised of IRs, transposase gene (*tnpA*), resolvase (*tnpR*) and a beta-lactamase (*amp*) gene. Integrons (F) carry mobile gene cassettes which are inserted/excised via recombination between *attI* and *attC* sites by site-specific recombinase (*intI*). Promoter P***c*** is used for transcription of the cassettes.

transposons, they can often carry integron recombination sites (*attI*) or complete integron platforms allowing them to incorporate various gene cassettes (Escudero et al., 2015).

1.3.3. Other TEs

There are other types of TEs in bacteria which do not conform to the above-described structural and mechanistic norms and blur the lines between TEs and other MGEs. Prominent examples introduced below include previously mentioned integrons, ICEs and IMEs, integrated prophages, and genomic islands.

Integrons are important drivers of bacterial evolution and adaptation (Souque et al., 2021). In their simplest form, integrons are platforms for capturing genes (Figure 1.2F) (Escudero et al., 2015; Stokes & Hall, 1989). They are comprised of an integrase (*intI*) gene which encodes the function for integrating and excising gene cassettes, an attachment (*attI*) site for integration of a gene cassette usually found upstream of the *intI*, and a promoter which acts upon integrated cassettes. Prior to integration, gene cassettes are comprised of a promotor-less gene with an *attC* recombination site which enables integration at *attI*. While the origin of integrons and gene cassettes is still unknown, it is beyond doubt that such genetic structures allow for rapid acquisition of valuable accessory functions with an associated low maintenance cost for the bacterial cell.

Integrative and conjugative elements (ICEs), also called conjugative transposons, are TEs which encode a functional conjugation system (Delavat et al., 2017; Johnson & Grossman, 2015). As the name suggest, these elements can horizontally spread via conjugation, and they can integrate in the host chromosome after which they can be vertically inherited from parent to offspring. ICEs which encode only a portion of the conjugation machinery and can be mobilized in *trans* are referred to as Integrative and Mobilizable Elements (IMEs). ICEs share several features of integrons and conjugative plasmids. Excised ICEs come in a circular form and contain *attI* site which allows reintegration into *attB* site on the host chromosome via site-specific recombination catalysed by an integrase protein. Many aspects of ICE biology are poorly understood, such as their influence on AMR spread and transmission and evolutionary dynamics in microbial communities (Botelho & Schulenburg, 2021).

Genomes of bacteriophages integrated into bacterial host chromosome or a plasmid are termed prophages (Snyder & Snyder, 2013). Integrated prophages can propagate via bacterial reproduction and plasmid replication, but under certain (usually stressful) environmental conditions they can escape this quiescent state (Díaz-Muñoz & Koskella, 2014). Like ICEs and integrons, phage genomes excise and integrate using a site-specific integrase which targets phage (*attP*) and bacterial (*attB*) attachment sites (Groth & Calos, 2004). Once integrated, a prophage can undergo excessive recombination picking up multiple accessory and other genes thus participating in microbial HGT network.

Prophages, ICEs, and IMEs can all be considered as members of an even more elusive group of TEs termed genomic islands (GIs). These are 10-200 kb-long chromosomal segments which carry beneficial accessory genes and are acquired by HGT (Bellanger et al., 2014; Hacker & Kaper, 2000; Lu & Leong, 2016). GIs can harbour hundreds of accessory genes, however, the exact transfer mechanisms for many GIs is rarely identified (Bellanger et al., 2014).

1.4. A wider perspective on bacterial gene mobility

Aside from clonal bacteria such as *Mycobacterium Tuberculosis*, pathogenic bacterial strains almost exclusively acquire pre-existing AMR and virulence determinants from the environment (Partridge et al., 2018). Acquisition and prevalence of these elements is a direct response to selective pressures such as excessive antibiotic usage in healthcare, agriculture, livestock, fish farming, and others (Partridge et al., 2018; Rodriguez-Mozaz et al., 2015). Dissemination of AMR and other important determinants of bacterial phenotypes is an essential feature of bacterial adaptation and evolution driven by a complex and multi-layered network of MGEs. Plasmids are important carriers of accessory genes and play a key role in the flow of genes within the network of bacterial hosts as efficient vehicles for both horizontal and vertical transmission. To maximize their fitness and ensure survival within a bacterial population, plasmid genomes undergo constant reshuffling by capturing, losing, silencing, and duplicating parts of their genomes, individual genes, and other elements. Important contributors to the displacement of beneficiary genes across plasmid backbones and bacterial chromosomes are TEs which have varying levels of activity and can act as layers of nested insertions.

1.4.1. Spread of antibiotic resistance genes

There are many MGEs contributing to the global spread of AMR genes (see review by Partridge et al., 2018a). Some of the most prominent TEs associated with multiple AMR genes include: IS*26* (Harmer et al., 2014) and related IS*257* and IS*1216* which are associated with both Gram-negative and Gram-positive bacteria; IS*1380* family elements, previously mentioned IS*Apl1* and ISCR elements, and Tn*3* and Tn*7*-like families of unit transposons. In most cases, the mobility of the bacterial accessory genome can be summarised using a so-called *Russian doll* model where dissemination of a particular gene is a consequence of mobility at multiple nested genetic levels (A. E. Sheppard et al., 2016). For instance, AMR gene-carrying transposons jump between various plasmid backbones; these plasmids can then spread between different pathogenic bacterial strains which in turn can cause havoc in hospital wards.

One example of such dynamics is the spread of the *bla*$_{\text{KPC}}$ beta-lactam AMR gene. *bla*$_{\text{KPC}}$ encodes a carbapenemase enzyme, capable of hydrolysing all members of the beta-lactam class of antibiotics including penicillins (penems), cepalosphorins, carbapenems and others (Malmir et al., 2018; Munoz-Price et al., 2013; Zaman et al., 2017). Beta-lactams are one of most widely used antibiotics in healthcare and carbapenem resistance is of particular concern with increasing global presence of resistance genes such as *bla*$_{\text{CTX-M}}$, *bla*$_{\text{KPC}}$, *bla*$_{\text{NDM}}$, and *bla*$_{\text{oxa-48}}$ in recent years (Bush, 2018; Tooke et al., 2019). These genes are known to rapidly spread in both Gram-negative and Gram-positive bacterial strains carried by multitude of MGEs (Partridge et al., 2018). As evidence of this, *bla*$_{\text{KPC}}$ has been found in 13 different bacterial species (62 distinct strains) and carried by a multitude of plasmid backbones all within a single hospital between 2007-2012 (A. E. Sheppard et al., 2016). This diversity of bacterial and plasmid hosts has been attributed to frequent Tn*4401* transposon jumps. In fact, such mobility of *bla*$_{\text{KPC}}$ is not an isolated case. In Peking University People's Hospital in China, *bla*$_{\text{KPC}}$-carrying carbapenem resistant *Klebsiella pneumoniae* (CRKP) caused an outbreak in 2016 (Van Dorp et al., 2019). Within 14 months since the first reported case, *bla*$_{\text{KPC}}$ spread throughout the hospital wards (some of which were 14 km apart) carried by at least four different plasmid backbones. A more complex interplay between TEs and plasmids has also been reported for the global dissemination of *bla*$_{\text{NDM}}$ carbapenem resistance gene (W. Wu et al., 2019). Mobility of this gene is further explored in Chapter 4.

Another example of *Russian doll*-like gene mobility is provided by the spread of *mcr-1*-mediated resistance against colistin, a last-resort antibiotic against multi-drug-

resistant bacterial infections (Liu et al., 2016). Identified in 2011, *mcr-1* spread across the globe within 8-12 years (R. Wang, Van Dorp, et al., 2018). This was caused by the IS*Apl1*-flanked composite transposon which first mobilized *mcr-1* into 13 different plasmid backbones thus enabling the exchange of resistance gene between multiple bacterial strains.

1.4.2. Bioinformatics approaches to studying HGT in bacteria

Experimental studies of MGEs are undoubtedly responsible for our understanding of the basic principles of HGT in bacteria and represent the foundation of all subsequent computational analyses. However, these studies primarily lack scalability, meaning they are excessively laborious or unable to provide insight from a broader perspective. Bioinformatics analyses based on whole genome sequencing (WGS) are increasingly affordable and convenient alternatives, and in some cases, essential complements to experimental studies. Nowadays, short-read mainly Illumina-based WGS of clinical isolates and environmental samples is considered a primary method for studying unculturable microbes, microbial communities, pathogenic bacteria and AMR (Gu et al., 2019; Orlek, Stoesser, et al., 2017; Quainoo et al., 2017; Sohn & Nam, 2018). In addition, improvement in the efficacy of third generation WGS platforms, such as Nanopore and PacBio, and novel and improved long-read and hybrid based *de novo* assembly methods are providing better quality genomes (M. Jain et al., 2016; Rhoads & Au, 2015; van Dijk et al., 2018). However, certain challenges remain when employing these methods to study HGT in bacteria. In particular, due to their accessory nature, MGEs are rarely present in all strains analysed which undermines many current epidemiological or phylogenetic analyses based on sequence alignments. Furthermore, *de novo* assembly of genomes prone to frequent movement of TEs, recombination and gene duplication can yield shorter contigs or result in misassemby of plasmids and other genomic regions.

A usual next step following the assembly process is the annotation of *orf*s and other genetic elements which helps provide context for the sequence in question. Two well established methods include NCBI's prokaryotic genome annotation pipeline (Tatusova et al., 2016) and the Prokka-Roary pipeline (Page et al., 2015; Seemann, 2014). Both rely on built-in databases of known genetic elements frequently found across bacterial genomes. Nevertheless, many other databases exist for custom annotation of MGEs (Table 1.1).

Table 1.1. Commonly used *in silico* resources for annotation of MGEs.

Type	Resource	Description	Reference
AMR profiling	**ABRicate**	A tool bundled with several databases for mass screening of assembled contigs for AMR and virulence genes.	https://github.com/tseemann/abricate
	CARD	Web-based platforms and databases useful for identifying genes and chromosomal mutations conveying AMR.	(Alcock et al., 2020; McArthur et al., 2013)
	ResFinder		(Bortolaia et al., 2020; Zankari et al., 2017)
	SStar	BLAST-based and user-friendly sequence search tool for screening AMR determinants.	(de Man & Limbago, 2016)
	TB-Profiler	Webservice and stand-alone software for infering *Mycobacterium tuberculosis* lineage and AMR resistance profile.	(Phelan et al., 2019)
AMR & Virulence profiling	**Kleborate**	A tool designed for screening of genome assemblies of *Klebsiella pneumonae* species for virulence, AMR and other determinants.	(Lam et al., 2021)
Annottating TEs	**ISFinder**	A curated and searchable database of prokaryotic ISs.	(Siguier et al., 2006)
	Transposon Registry	A curated and searchable database of all known prokaryotic transposons.	(Tansirichaiya et al., 2019)
Plasmid classification	**MOBscan**	A more sophisticated tool for MOB-based plasmid typing based on HMM sequence profiling (HMMER).	(Garcillán-Barcia et al., 2020)
	PlasmidFinder	Web-based platform and database for identifying plasmid types of Gram-positive bacteria and *Enterobacteriaceae* in raw sequencing reads and assembled contigs.	(Carattoli et al., 2014; Carattoli & Hasman, 2020)
	pMLST	Method for subtyping of specific plasmids based on multi-locus sequence typing. Currently, the method is limited to IncA/C, IncI1, IncHI1, IncHI2, IncF and IncN plasmid types.	(Carattoli et al., 2014; Jolley et al., 2018)
Plasmid classification & reconstruction	**MOB-suite**	Suite of tools for replicon- and MOB-based typing and reconstruction of plasmid sequences from WGS assemblies.	(Robertson & Nash, 2018)
Plasmid reconstruction	**Bandage**	Interactive tool for visualization of *de novo* assemblies useful for manual untangling of chromosomal and plasmid sequences.	(Wick et al., 2015)
	PLACNETw	a web-based tool for plasmid reconstruction from pair-end WGS reads.	(Vielva et al., 2017)
	PlasFlow	A pipeline for predicting plasmid sequences in metagenomic data.	(Krawczyk et al., 2018)
	plasmidSPAdes	Software for assembling plasmids from raw WGS data.	(Antipov et al., 2016)
	Unicycler	SPAdes-based pipeline for *de novo* assembly of bacterial genomes from WGS data. It aims to produce circularized contigs hence facilitating reconstruction and identification of plasmids.	(Bankevich et al., 2012; Wick et al., 2017)
Plasmid reconstruction & annotation	**Platon**	A tool for prediction and annotation of plasmid sequences from short-read WGS assemblies based on distributions of replicon protein-coding genes.	(Schwengers et al., 2020)
Toxin-antitoxin systems	**TASmania**	HMM-based pipeline and a database for annotating bacterial toxin-antitoxin systems.	(Akarsu et al., 2019)
Virulence profiling	**VFDB**	A BLAST-searchable comprehensive database of virulence factors covering 74 genera of pathogenic bacteria.	(L. Chen et al., 2005)
	VirulenceFinder	Web-based platform and database for prediction of virulence factors from WGS data of *Listeria*, *Staphylococcus aureus*, *Escherichia coli* and *Enteroccocus* isolates.	(Joensen et al., 2014)

Aside from genome annotation, homologous genomic regions can be aligned using multiple-sequence alignment (MSA) tools which comprise the first step in establishing phylogenetic relationships across samples. Alternatively, bacterial WGS samples with more clonal inheritance (i.e., fewer recombination events) can be mapped to a known reference sequence from which single nucleotide polymorphisms (SNPs) can be determined. Conversely, highly dissimilar or recombining sequences cannot be aligned or mapped to a reference which prompts the use of alignment-free sequence comparison tools like Mash or BinDash (Ondov et al., 2016; X. Zhao, 2019). Reconstruction of the phylogenetic tree from MSA or SNPs opens prospects for other types of analyses such as: computation of mutation rates; molecular dating of evolutionary events (Bouckaert et al., 2019; Didelot et al., 2018; Rieux & Balloux, 2016; R. Wang, Van Dorp, et al., 2018); or detecting recombination events (Didelot & Wilson, 2015). Results of the above-mentioned analyses can then be correlated with known metadata such as the presence of specific MGEs, AMR genes or regions of pathogenicity, certain phenotypes or disease outcomes, sampling locations or sampling sources, all of which can help further uncover the dynamics of HGT events or the progress of an outbreak.

Highlighted above are some basic principles, tools and resources used when investigating gene mobility or an outbreak of AMR genes, virulence factors or pathogenic bacterial strains: starting from the collection of samples for the dataset; determining genetic similarity or establishing phylogenetic relationships between samples; and finally correlating those with specific observations and events. Variations of this approach can be found in two studies presented in this thesis. In Chapter 2 and Chapter 3, I describe the analysis of a large dataset of complete bacterial plasmids. Following alignment-free similarity assessment between all pairs of bacterial plasmids, a network-based approach is used to uncover their underlying population structure. A more standardized analysis of global dissemination of *bla*$_{\text{NDM}}$ gene is implemented in Chapter 4 which helped uncover roles of specific MGEs in *bla*$_{\text{NDM}}$ mobility.

Chapter 2

Uncovering population structure of bacterial plasmids

Declaration of contributions

Francois Balloux and I conceived the project and conceptualized the methods. I performed all the analyses under the guidance of Lucy van Dorp and Francois Balloux. Joanne M. Santini advised on plasmid biology.

Publication

This work has been published in Nature Communications as Acman et al. (2020):

Acman, M., van Dorp, L., Santini, J. M., & Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. Nature Communications, 11(1), 1–11. https://doi.org/10.1038/s41467-020-16282-w

2.1. Introduction

Plasmids are extra-chromosomal DNA molecules found across all three Domains of Life. In bacteria, they are considered one of the main mediators of horizontal gene transfer (HGT) through the processes of conjugation and transformation (Halary et al., 2010; Shintani & Suzuki, 2019; Von Wintersdorff et al., 2016). Plasmids generally harbour non-essential genes that can modulate the fitness of their bacterial host. These accessory genes can be located on transposable elements involved in lateral gene transfer across genomes and can thus lead to a highly mosaic structure of plasmid genomes (Stokes & Gillings, 2011). The mix of vertical and horizontal inheritance of plasmids, together with exchanges of plasmid-borne genes, generates complex dynamics that are difficult to capture with classical population genetics tools and make it challenging to classify plasmids within a coherent universal framework.

As discussed in the introduction (Chapter 1), there are two well-established plasmid classification schemes which attempt to bin plasmids according to their propagation mechanisms, while indirectly capturing some features of the plasmid backbone. The first scheme is based on replicon types (Carattoli et al., 2005) and the second on mobility (MOB) groups (Garcillán-Barcia et al., 2009). Replicon-based typing relies on relatively conserved genes of the replicon region which encode the plasmid replication and partitioning machinery (Carattoli et al., 2005). Plasmids with matching replication or partitioning systems cannot stably coexist within the same cell. Conversely, MOB typing is used to classify self-transmissible and mobilizable plasmids into six MOB types (Garcillán-Barcia et al., 2009). The MOB typing scheme relies on the conserved N-terminal sequence of the relaxase, a site-specific DNA endonuclease which binds to the origin of transfer (*oriT*) cleaving at the *nic* site and is essential for plasmid conjugation.

Despite being widely used and informative, these typing schemes only work within a limited taxonomic range (Orlek, Phan, et al., 2017a; Orlek, Stoesser, et al., 2017; Shintani et al., 2015). Replicon typing is dependent on the availability of prior experimental evidence and remains restricted to culturable bacteria from the family *Enterobacteriaceae* and well-studied genera of gram-positive bacteria (Carattoli et al., 2014; Jensen et al., 2010; Lozano et al., 2012; Shintani & Suzuki, 2019). Furthermore, this approach can lead to ambiguous classification, even for experimentally validated replicons, as recently demonstrated by the discovery of compatible plasmids assigned to

the same replicon type, which led to the further subdivision of the IncK type into IncK1 and IncK2 (Rozwandowicz et al., 2017), and IncA/C type into IncA and IncC (Ambrose et al., 2018). In addition, plasmids can carry genes from more than one replication machinery leading to assignment to multiple replicon types, further reducing interpretability (Orlek, Stoesser, et al., 2017; Shintani et al., 2015). MOB typing schemes generate fewer multiple assignments and can cover a potentially wider taxonomic range, however they are not applicable to the classification of non-mobilizable plasmids. These two typing schemes have inspired several *in silico* classification tools, such as PlasmidFinder (Carattoli et al., 2014), the plasmid Multi Locus Sequence Typing (pMLST) database, and MOB-suite (Robertson & Nash, 2018). However, all those tools intrinsically rely on the completeness of their reference sequence databases, which typically lack representatives from understudied and/or unculturable bacterial hosts.

As bacterial plasmids undergo extensive recombination and HGT, their evolutionary history is not well captured by phylogenetic trees, which are designed for the analysis of point mutations in sequence alignments (Bapteste et al., 2009; Brilli et al., 2008). Network models offer an attractive alternative given they can incorporate both horizontal and vertical inheritance (Bernard et al., 2018; Corel et al., 2016), and can deal with point mutations as well as structural variants. Networks have gained much attention in the past decade as an alternative method for studying prokaryotic evolution, including plasmids (Bernard et al., 2018; Corel et al., 2016; Dagan et al., 2008; Halary et al., 2010; Orlek, Stoesser, et al., 2017). Plasmid gene-sharing networks have proven a useful means to track AMR and virulence dissemination yielding deeper insights into HGT events (Brilli et al., 2008; Tamminen et al., 2012; Yamashita et al., 2014). However, the main drawback of previous work relying on plasmid sequence alignments is the exclusion of important non-coding elements such as non-coding RNAs, promoter regions, CRISPRs, stretches of homologous sequences, or putative, disrupted and currently unannotated genes. A more comprehensive approach could consider a plasmid network based on estimates of alignment-free sequence similarity (Zielezinski et al., 2017). Alignment-free genetic distance methods are becoming established tools for the analysis of large genomic datasets, and their usefulness has been validated in both prokaryotes and eukaryotes (Bernard et al., 2018, 2019; Ren et al., 2018; Zielezinski et al., 2017, 2019). A recently published Plasmid ATLAS tool by Jesus *et al.* (Jesus et al., 2019) provides an illustration of such an approach, with a network of plasmids constructed based on pairwise genetic distances estimated using alignment-free *k*-mer matching methods implemented in Mash (Ondov et al., 2016).

In this chapter, I present a new approach to classifying bacterial plasmids using a community detection algorithm. To this end, I have curated a dataset containing more than 10,000 bacterial plasmids obtained from the publicly available NCBI's RefSeq database (O'Leary et al., 2016). I quantified the genetic similarity between pairs of plasmids and constructed a network which reflects their relatedness based on shared *k*-mer content. Applying a community detection algorithm to the network enabled clustering of plasmids with high genetic similarity into cliques (complete subgraphs) revealing a strong underlying population structure.

2.2. Methods

2.2.1. Assembling a dataset of complete bacterial plasmids

A dataset of complete plasmids was downloaded from NCBI's RefSeq release repository (O'Leary et al., 2016) on 26th of September 2018. The metadata accompanying each plasmid sequence was parsed from the associated GenBank files. The resulting dataset was then systematically curated to include only those plasmids sequenced from a bacterial host and with a sequence description which implies a complete plasmid sequence (regular expression term used: `"plasmid.*complete sequence"`). This is a simpler, but similar approach to a previously reported curation effort by Orlek and colleagues (Orlek, Phan, et al., 2017a). Nevertheless, a large portion of unsuitable entries, such as gene sequences, partial plasmid genomes, whole genomes, non-bacterial sequences and other poorly annotated sequences, were removed. The final dataset included 10,696 complete bacterial plasmids.

Information about the taxonomic hierarchy of plasmid bacterial hosts was obtained with the *ncbi_taxonomy* module from the ETE 3 Python toolkit (Huerta-Cepas et al., 2016). To determine the replicon and MOB types of plasmids included in the dataset I used the PlasmidFinder replicon database (version: 2018-09-04) (Carattoli et al., 2014) and MOBtyping software (Orlek, Phan, et al., 2017b). The PlasmidFinder database was screened using nucleotide BLAST (Altschul et al., 1990) with a minimum coverage and percentage identity of 95%. In cases where two or more replicon hits were found at overlapping positions on a plasmid, the one with the higher percentage identity was retained. For determining the plasmid MOB type, MOBtyping software was used with the recommended settings of 14 PSI-BLAST iterations.

Plasmid CDSs were annotated using Prokka (version 1.13.3) (Seemann, 2014) and Roary (version 3.12.0) (Page et al., 2015) pipelines run with default parameters. The identified CDSs were further associated with Gene Ontology (GO) terms (Ashburner et al., 2000; Carbon et al., 2019) to facilitate downstream gene content analysis. Since Prokka uses a variety of databases to annotate identified CDSs, different resources have been used to append the corresponding GO terms. For example, GO terms for CDSs with a known protein product have been obtained using Uniprot's 'Retrieve/ID Mapping' tool (H. Huang et al., 2011), while the GO terms for CDSs with just the HAMAP family were obtained with the hamap2go mapping table (Lima et al., 2009) (version date: 2019/05/04). CDSs annotated with the ISfinder database were given GO terms GO:0070893 and GO:0004803 in order to associate them with transposition. Similarly, CDS annotated with Aragorn, MinCED, and BARRGD were given GO:0006412, GO:0099048, and GO0046677 terms respectively.

2.2.2. Assessing similarity between pairs of plasmids

The exact Jaccard index (JI) was used as a measure of similarity between all possible plasmid pairs. Each plasmid sequence was converted to a set of 21 bp *k*-mers. The JI was then calculated as the fraction of shared *k*-mers between two sets. JI thus takes a value between 0 and 1, where 1 indicates 100% *k*-mer similarity, and 0 indicates no *k*-mers shared. This allows balanced comparison of diverse plasmid genomes and universality. Also, JI does not weight k-mers based on their abundance, like the popular D_2^* and D_2^S statistics (Reinert et al., 2009), which would exacerbate the inherent sampling biases towards well-studied species in the dataset. Bindash (X. Zhao, 2019) was used to calculate the exact JI which resulted in the creation of a plasmid adjacency matrix which was used to build the network. All networks presented in this chapter have been explored and visualized using the Cytoscape software (Shannon et al., 2003).

2.2.3. Implementing OSLOM community detection algorithm

OSLOM (Ordered Statistics Local Optimization Model version 2.5) was applied to identify cliques (complete subgraphs) with high internal JI similarity in the plasmid network (Lancichinetti et al., 2011). OSLOM aims to identify highly cohesive clusters of vertices (communities) which may or may not be cliques (complete subgraphs). The

statistical significance of a cluster is measured as the probability of finding the cluster in a configuration model which is designed to build random networks while preserving the degrees (number of neighbours) of each vertex. The method locally optimizes the statistical significance with respect to vertices directly neighbouring a particular cluster. In brief, OSLOM starts by randomly choosing vertices from a network which are regarded as clusters of size one. These small clusters alongside their neighbouring vertices are assessed. Vertices are scored based on their connection strength with a particular cluster and are either added or removed from the cluster. The process continues until the entire network is covered. Due to the stochastic nature of the algorithm, this network assessment goes through many iterations after which the frequently emerging significant clusters (i.e., communities) are kept. The algorithm then proceeds to assess the clusters of the next hierarchical level; vertices belonging to the significant clusters are condensed into super-vertices with weighted edges connecting them. The process of cluster assessment is repeated at higher hierarchical levels until no more significant clusters are recovered.

OSLOM was executed for an undirected and weighted network with the following parameters:

```
oslom_undir -w -t 0.05 -r 50 -cp 0 -singlet -hr 0 -seed 1
```

Clusters were considered significant if their *p*-value was lower than 0.05 (`-t 0.05`). The number of iterations required before the recovery of significant clusters was set to 50 during the search for the optimally sparse network (`-r 50`), and 250 for the final network analysis after the introduction of the 0.3 JI threshold (`-r 250`). After the iteration process, OSLOM considers merging similar significant clusters if the significance of their union is high enough. This feature can potentially yield a lower number of cliques and was suppressed with the coverage parameter set to zero (`-cp 0`) thus forcing OSLOM to opt for the biggest and most significant cluster from a set of similar clusters. In addition, OSLOM tries to place all vertices of a network in clusters which is also unfavourable for clique recovery and was suppressed with option `-singlet`. Lastly, cliques can only be recovered at the first hierarchical level. Therefore, the OSLOM analysis of the higher hierarchical levels was disregarded (`-hr 0`).

OSLOM is a non-deterministic algorithm, and the initial single-vertex clusters are chosen at random. While looking for the optimally sparse network, five OSLOM runs were executed to assess every JI threshold and were given seeds for a random number generator (`-seed`) of 1, 5, 42, 93, and 212. The final network analysis was performed

with a seed equal to 42, after which only cliques were considered**,** with non-complete communities disregarded.

2.3. Results

2.3.1. Makeup of the dataset of complete bacterial plasmids

A dataset of complete bacterial plasmids was compiled comprising 10,696 sequences found in bacteria from 22 phyla and over 400 genera (Figure 2.1A and Figure A.1). The composition of plasmid hosts reflects current research interests, with the Proteobacteria and Firmicutes phyla together representing over 84% of plasmid sequences. Though, the dataset includes plasmids from a diversity of bacterial hosts, with 66 plasmids from unknown bacterial families, 14 from uncultured bacteria and 37 samples from *candidatus* species.

In total, 510,463 different Coding Sequences (CDSs) were identified in the plasmid dataset. 66.01% of the CDSs were predicted to encode a hypothetical protein, 27.9% had a known product with Gene Ontology (GO) biological process annotation, with the remaining 6.09% encoded a known protein product with unknown biological function (Figure 2.1B). There are 3,328,916 bacterial genes available in the RefSeq database (NCBI Gene Statistics accessed on June 19th, 2019), meaning that roughly one in twenty of the currently known bacterial genes are plasmid-borne. The GO biological processes associated with plasmid CDSs are diverse. After accounting for multiple occurrences of annotated CDSs in the dataset, the dominant associated GO terms relate to catabolic and biosynthetic processes (20.64% relative to total number of annotated CDSs), transposon mobility (17.09%) and positive and negative regulation of transcription (7.70%).

Replicon-based typing classified 27.66% of the plasmids into 163 different replicon types (Figure 2.1C and Figure A.2). However, 31.67% of these classified plasmids were assigned to multiple replicon types. MOB typing was more comprehensive, successfully classifying 32.63% of the plasmids into six MOB types of which 9.48% were assigned to multiple types (Figure 2.1C). Unsurprisingly, classification by these two methods performed best for well-studied plasmids of the phyla Proteobacteria and Firmicutes.

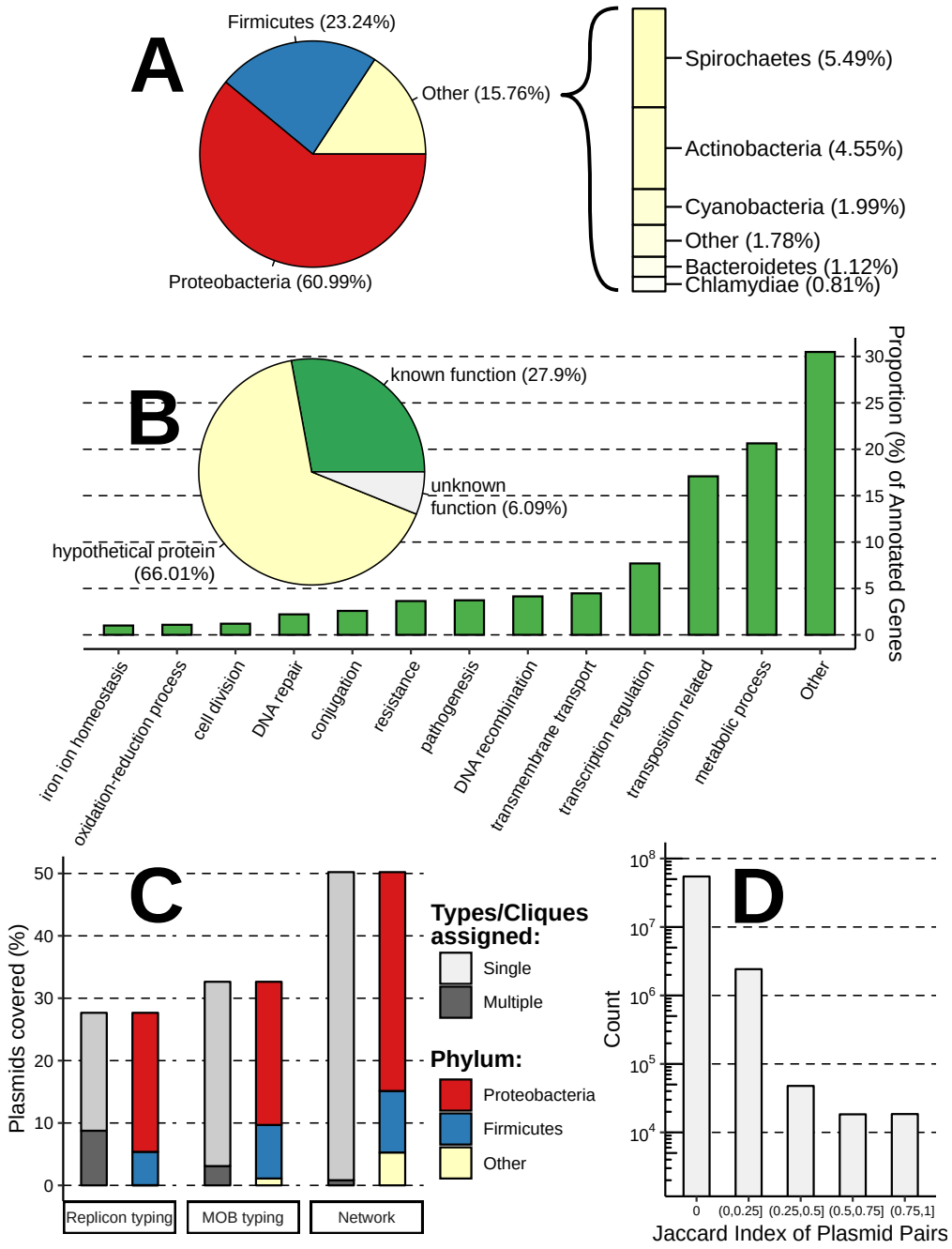


Figure 2.1. Summary of the dataset of complete bacterial plasmids. (A) The distribution of host phylum represented in the plasmid dataset. (B) Functional annotation of plasmid-borne genes. The pie chart shows the proportion of unique CDSs with hypothetical function as predicted by Prokka, and CDSs (genes) with known/unknown biological function based on GO annotation. The bar chart provides the most common biological functions associated with plasmid-borne genes also considering the respective frequency of these genes on plasmid genomes. (C) The percentage of plasmids covered by the three classification methods: replicon and MOB typing schemes, and clique assignment. (D) The distribution of pairwise plasmid similarities (Jaccard Index).

2.3.2. Analysing the network of plasmids

A network was constructed based on the plasmid pairwise sequence similarities. This represents a weighted, undirected network with plasmids (vertices) connected by edges indicating similarity (Figure 2.2). Similarity was scored using the exact Jaccard index (JI), defined as the size of the intersection divided by the size of the union of two sets of *k*-mers. Plasmid pairs which shared less than 100 *k*-mers were considered to have a JI equal to zero. This cut-off value was implemented since the majority of CDSs found on plasmids have lengths greater than 100 bp, thus only a fraction of the functional genome is common between plasmids with low shared *k*-mer count (Figure 2.3 and Figure 2.4). The majority of plasmid pairs shared little to no similarity (Figure 2.1D). 6.14% (657) of the plasmids were singletons, whilst 3.31% (354) were connected to only one other plasmid, illustrating the high levels of diversity across bacterial plasmid genomes. It follows that plasmids with more *k*-mers in common are more likely to share the same functional genetic elements and hence participate in similar biological processes falling within the same host niche (Figure 2.4). Such plasmids are presumed to form cliques within the network with higher internal JI score. The objective is then to identify cliques which contain plasmids with markedly higher similarity between themselves, relative to their immediate network neighbourhood.

Listing all cliques of the large plasmid network and assessing their internal similarity is computationally intractable with current tools (Karp, 1972). A solution for a single clique can be quickly verified, but the time required to process all possible cliques scales rapidly as the size of the network increases. As an alternative solution, a stochastic community detection algorithm OSLOM (Ordered Statistics Local Optimization Method) was implemented (Lancichinetti et al., 2011). OSLOM detects communities (i.e., densely interconnected subgraphs) with statistical significance, meaning that they have a low probability of being encountered by chance in a random network with similar features to the plasmid network. OSLOM is well suited for this task since it can be used to analyse undirected networks with overlapping communities or hierarchical structures. In addition, OSLOM shows similar performance to other widely used methods such as Infomap or Louvain (Hric et al., 2014; Lancichinetti et al., 2011) which, unlike OSLOM, were unable to analyse this dataset due to computational limitations. To validate the results from the

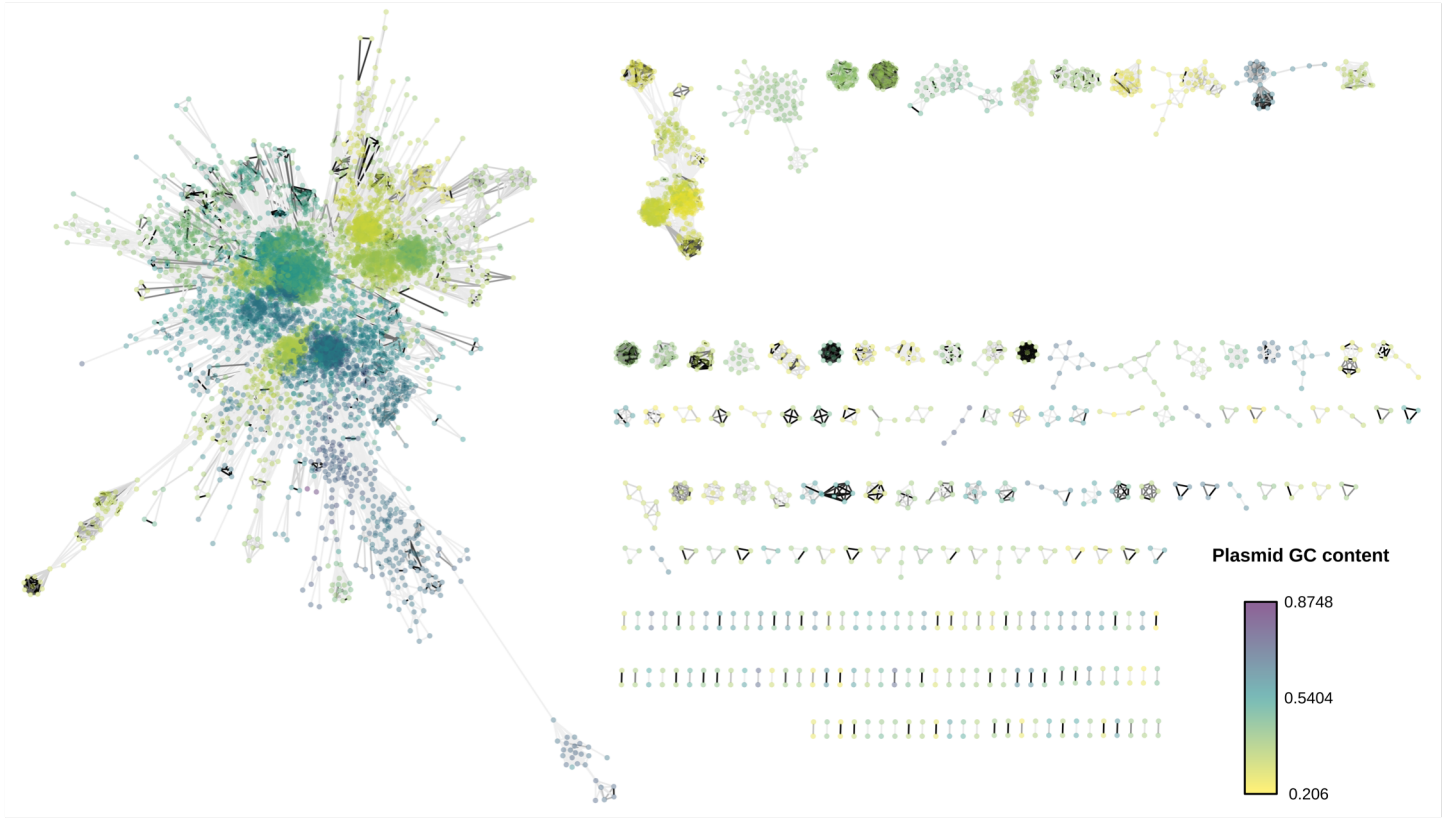


Figure 2.2. A network of plasmids (network density = 0.0438). 10,696 complete plasmids (represented as nodes) are connected by weighted edges where grey-scale colour gradient specifies the Jaccard Index (JI) similarity. JI is calculated as the proportion of shared k-mers between pairs of plasmids with a darker shade indicating higher similarity. Plasmid pairs which share less than 100 k-mers are considered to have a JI equal to zero. The colour gradient of the nodes indicates the GC content of plasmids. This representation depicts clustering of plasmids according to their GC content and hints at an underlying population structure.

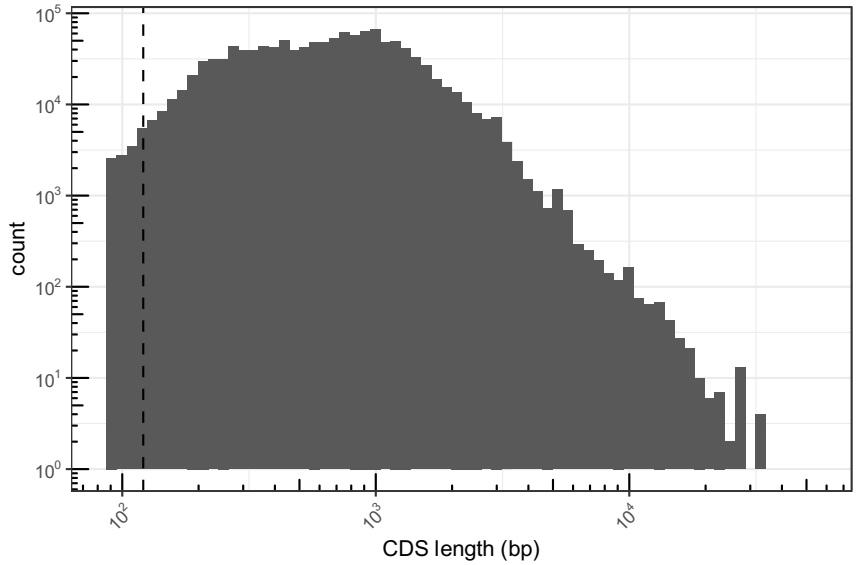


Figure 2.3. The distribution of the lengths of the plasmid-borne coding sequences (CDSs). Both axes are in logarithmic scale. The vertical dashed line represents the cut-off value (<100 k-mers) applied while calculating JI similarity between plasmid pairs. Since the length of k-mers used was 21 bp, the effective cut-off value applied is 121 bp. Thus, few CDSs shared between plasmids with length less than 121 bp may have been omitted as a result of the implementation of this cut-off.

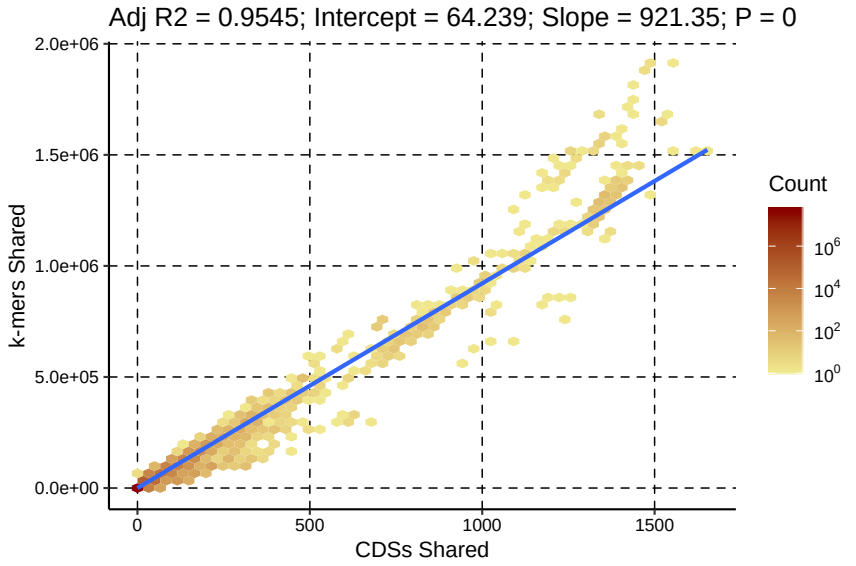


Figure 2.4. Linear correlation between number of shared CDSs and number of shared k-mers in plasmid pairs. To facilitate visual interpretation, the data points were grouped into hexagonal areas of equal size. The colour intensity of each hexagon reflects the density (count scale) of the data points in that particular area. An intercept (64.239) of the regression line (blue) suggests that a plasmid pair sharing around 64 k-mers on average do not have any CDSs in common.

stochastic clique assignment, all communities of size three or more detected by OSLOM were assessed for their completeness (i.e., whether they form cliques) against the original plasmid network (Figure 2.2).

Despite the notable dissimilarity among plasmids, the original network was too dense (network density = 0.0438) to yield a consistent performance for every OSLOM run (Figure 2.5 and Figure A.3). Furthermore, a large proportion of communities detected did not form cliques and would have to be disregarded (Figure 2.5A). A JI threshold was therefore introduced to increase the sparsity of the network and to upweight more similar plasmids, thus optimizing the performance of OSLOM. A range of thresholds were assessed based on the following criteria: (i) the clique to community ratio (Figure 2.5A), (ii) the proportion of plasmids assigned to cliques (Figure 2.5B), (iii) the congruence with replicon-based typing (Figure 2.5C), and (iv) the consistency of OSLOM performance (Figure 2.5 and Figure A.3). The optimum threshold was consistently obtained at a JI of 0.3. The resulting sparse network is shown in Figure 2.6 (network density = 0.00128).

2.3.3. Validating plasmid classification

Analysis of the sparse network with OSLOM successfully assigned 50.21% (5371) of the plasmids into 561 cliques of size three or more (Figure 2.1C, Figure 2.6, and Figure A.4). Only 1.64% (88) of these plasmids were assigned to multiple cliques, and these were found in the densest regions of the network and at the interfaces between cliques indicating the presence of ‘chimeric plasmids’ (i.e. hybrid plasmids generated through merging of two different plasmids), large-scale transposition or recombination events, or extensive repeated transposition/recombination (Figure 2.1C and Figure 2.6). In addition, this approach covered 564 plasmids from phyla other than the Proteobacteria and Firmicutes, namely from Spirochaetes, Chlamydiae, Actinobacteria, Tenericutes, Bacteroidetes, Cyanobacteria, and Fusobacteria. Interestingly, after applying the 0.3 JI threshold, 38.01% (4,066/10,696) of plasmids that could not be assigned to cliques of size three or more, were separated from the network as singletons, while 10.10% (1080) shared an edge with a single plasmid. Therefore, only 1.67% (179) of plasmids in the dataset were effectively left unassigned. Nonetheless, due to the apparent lack of shared genetic signal, plasmid singletons and pairs were not considered in any subsequent analyses.

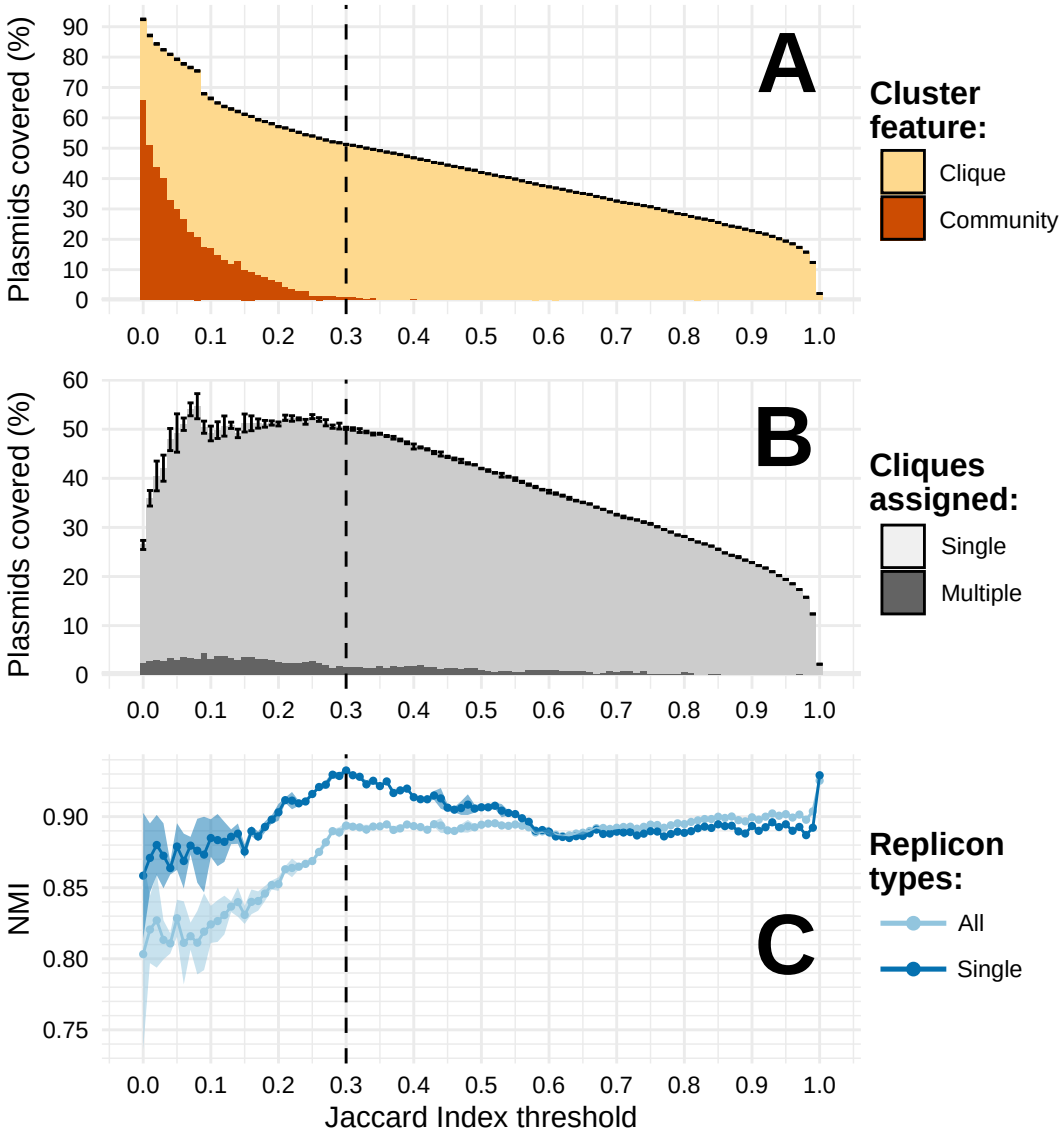


Figure 2.5. Optimization of OSLOM performance. A range of Jaccard Index (JI) thresholds were applied to the original plasmid network (Figure 2.2) with edges below a particular threshold being removed prior to OSLOM analysis. During the process, several criteria were considered: (A) clique to community ratio; (B) percentage of plasmids covered by the cliques; (C) the congruence with replicon typing measured by NMI score. NMI was calculated for all cliques containing plasmids assigned to a single or multiple replicon types (legend: All) and just to a single replicon type (legend: Single). Error bars (A and B) and light-coloured shading (C) provide +/- 2 standard deviations (SD) of uncertainty. Standard deviation around every value on the y-axis across all JI thresholds assessed (points and bars) was calculated based on results of n=5 iterations of OSLOM software (see Methods). The dashed vertical line indicates the selected optimal JI threshold of 0.3.

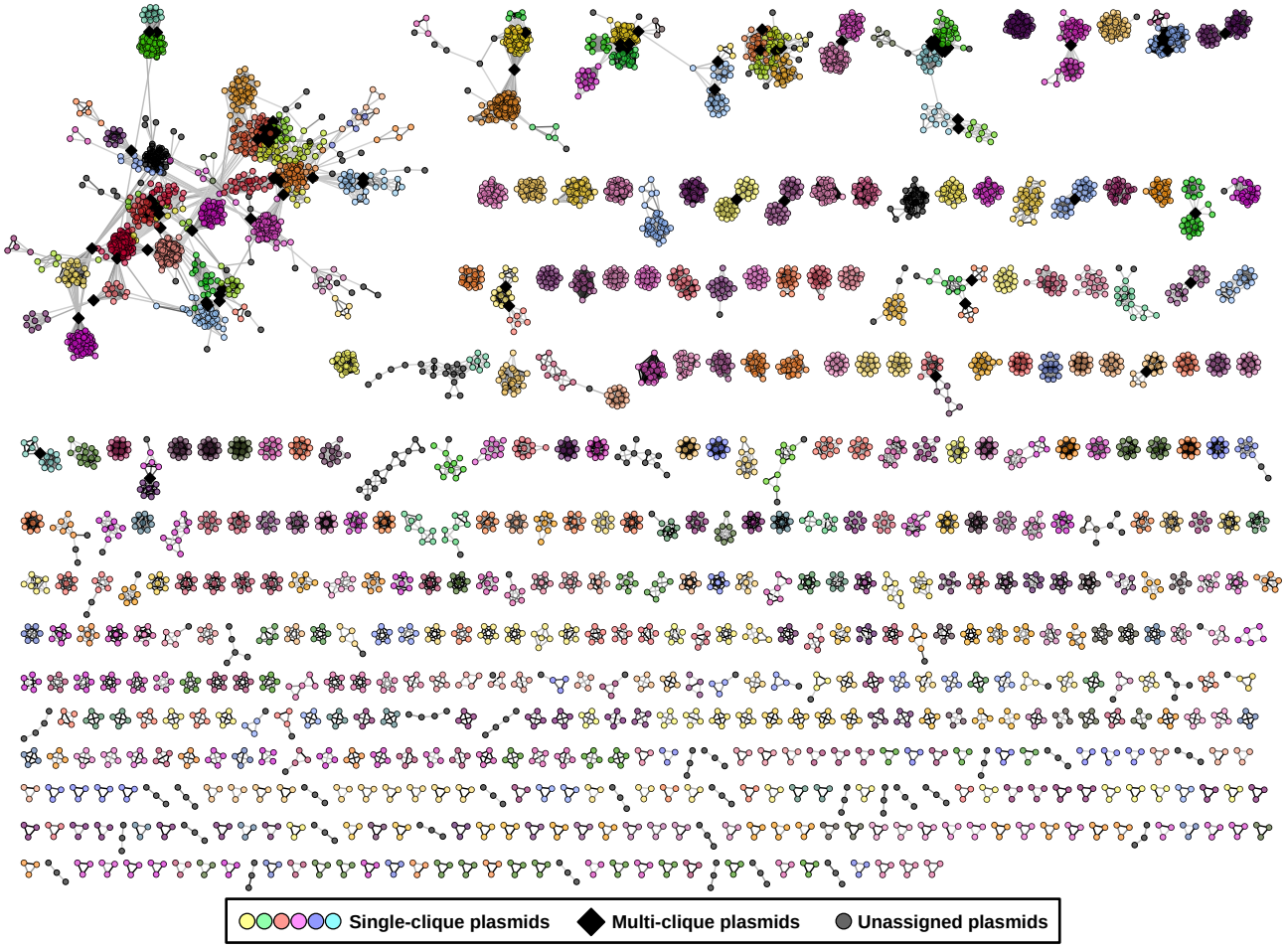


Figure 2.6. Sparse network of plasmids assigned to cliques by OSLOM algorithm (network density = 0.00128). The network includes 5,371 plasmids (nodes) assigned into 561 cliques (complete sub-graphs). The completeness of identified cliques was evaluated based on the original network (Figure 2.2). 5,008 unassigned plasmids, which formed disjoined singletons and pairs, were removed from the network. The plasmids in the network are connected by weighted edges where grey-scale colour gradient specifies the Jaccard Index (JI) similarity. Coloured nodes indicate plasmids assigned to a single clique.

The OSLOM-guided clique detection algorithm offers flexibility and identifies cliques of plasmids with a wide range of internal similarity scores (Figure A.5). I assessed the importance of considering pairwise JI distances as a continuous variable by reanalysing the dataset with the Bron-Kerbosch Max-clique algorithm (Bron & Kerbosch, 1973), implemented in the graph-tool Python library (Peixoto, 2014). The Bron-Kerbosch algorithm is computationally highly effective, but the pairwise distances between plasmids are treated as binary values defined by the given threshold. Applied across a range of JI thresholds, the Max-clique approach systematically identifies a very large

number of cliques (Figure A.6A), assigns a large proportion of plasmids to multiple cliques (Figure A.6B) and leads to a low correlation between resulting cliques and plasmid replicon types (Figure A.6C).

Lastly, robustness of the classification was assessed by evaluating the extent to which 'mobile elements' shared between plasmids affect their classification into cliques. In particular, the clique assignment analysis was repeated after removing all accessory CDSs (29,913) associated with transposition, pathogenesis, or resistance (Figure A.7). Pruning these genes did not markedly affect the assignment of plasmids into cliques, which gives support to the genetic signal being driven by the genetic similarity of plasmid backbones rather than shared mobile genetic elements.

2.4. Discussion

Plasmids are one of the main contributors to HGT in bacteria as they have the capacity to harbour and disseminate resistance and virulence genes. In fact, 5% of all annotated bacterial genes are currently in circulation on plasmids. Underlying this phenomenon is the plasticity of plasmid genomes which undergo frequent genome rearrangements and integrating DNA segments of various origins. The ability to distinguish a meaningful phylogenetic signal within populations of bacterial plasmids opens new prospects to understand plasmid evolution and the nature of HGT and enables building of a more coherent classification system as well as tracking the dissemination of specific genetic elements. A network-based representation of sequence similarities condenses both vertical and horizontal evolutionary histories thus offering an attractive solution to studying bacterial plasmids.

By using an alignment-free sequence similarity comparison and subsequent network analysis I uncovered strong population structure in bacterial plasmids. This approach was applied to a comprehensive set of complete bacterial plasmids that covered a wide genetic and host diversity. The analysis yielded a network in which over half of the plasmids were classified into cliques which represents a significant improvement in coverage over existing plasmid typing methods.

Jaccard index (i.e., the fraction of shared *k*-mers) was chosen as a measure of sequence similarity between pairs of plasmids due to it being a straightforward metric which considers genome sequences as a whole, embodying both point mutations and large-scale genome rearrangements. As a result, it is not biased by the ability to annotate

genes, open reading frames, or other genetic elements. In addition, it is not prone to errors and biases intrinsically associated with alignment-based methods, such as: *a priori* assumptions about the sequence evolution, higher inaccuracy when comparing more dissimilar sequences, or suboptimal alignments (Zielezinski et al., 2017). JI can in principle provide fine-scale resolution when comparing small genomes, a characteristic common to the majority of plasmids. Conversely, JI is sensitive to varying genome sizes (Ondov et al., 2016) and plasmids are known to differ more than 1000-fold in sequence length (Shintani et al., 2015; Smillie et al., 2010). While differences in plasmid genome size can lead to a drop in JI score even when high proportions of *k*-mers are shared, sequence length variation did not seem to impact clique structure with cliques found to comprise plasmids of different lengths (Figure 3.6 C and D).

Assessing the statistical significance of all cliques is computationally intractable given the size and the density of the plasmid network. Hence, the OSLOM community detection algorithm was employed to uncover cliques of plasmids with high genetic similarity. In an effort to optimize the performance of the OSLOM algorithm and maximize the number of biologically meaningful cliques, all edges with a JI value below 0.3 were removed from the network prior to the analysis. This threshold was chosen to maximise compliance with replicon-based typing as well as several other criteria. Whilst the classification of plasmids into cliques is fairly robust to this exact JI threshold, I appreciate that a 0.3 JI threshold remains somewhat arbitrary. This being said, any taxonomy based on sequence similarity will be partly subjective. As such, the 0.3 JI threshold is comparable in its subjectivity to the 95% average nucleotide identity (ANI) which was set over a decade ago and is routinely used to define species boundaries in prokaryotes (Goris et al., 2007). However, depending on the question pursued, enforcing a strict JI threshold may not be necessary, and it could be left to plasmid sequences in the network to solely inform the cut-offs. Some boundaries are likely to be blurrier than others, largely reflecting the extensive variation of genetic inheritance in different bacterial hosts.

Finally, the presented network-based analysis was found to be robust to removal of a large fraction of accessory genes suggesting the classification is based on meaningful (phylo-) genetic signal. In addition, the uncovered cliques include plasmids over varying range of similarities thus reflecting highly mosaic structure of plasmid genomes. All code used in this chapter is available at https://github.com/macman123/plasmid_network_analysis.

Chapter 3

Biological significance of plasmid cliques

Declaration of contributions

Francois Balloux and I conceived the project and conceptualized the methods. I performed all the analyses under the guidance of Lucy van Dorp and Francois Balloux. Joanne M. Santini advised on plasmid biology.

Publication

This work has been published in Nature Communications as Acman et al. (2020):

Acman, M., van Dorp, L., Santini, J. M., & Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. Nature Communications, 11(1), 1–11. https://doi.org/10.1038/s41467-020-16282-w

3.1. Introduction

In the previous chapter (Chapter 2), I introduced a novel network-based approach for classification of bacterial plasmids as well as a comprehensive dataset of 10,696 complete bacterial plasmids assembled for this purpose. The genetic similarity between all pairs of plasmids from the dataset was quantified using a Jaccard index (JI) and represented as a network. The network is undirected with plasmids (vertices) connected by edges indicating similarity. A 0.3 JI threshold was introduced to increase the sparsity of the network which was followed by the analysis using OSLOM (Lancichinetti et al., 2011) community detection algorithm. This resulted in classification of highly similar plasmids into cliques and revealed a strong underlying population structure (Figure 2.6).

In this chapter, I establish and discuss the biological meaning of cliques of bacterial plasmids which show high correlation with replicon (Carattoli et al., 2005) and mobility (MOB) (Garcillán-Barcia et al., 2009) based classification schemes. I also demonstrate how to leverage this correlation to discover candidates of yet-undescribed replicon genes. Furthermore, I demonstrate that plasmids within cliques exhibit a high degree of phylogenetic relatedness as suggested by matching bacterial hosts, GC content, sequence length, and gene content. Clustering of plasmids into these more phylogenetically related groups allowed further insight into the dynamics of HGT and helped identify transposable genetic elements as the main drivers of HGT at broad phylogenetic scales. Taken together, results presented in this chapter illustrate the potential of network-based analyses of plasmid sequences and the prospect of a natural, exhaustive classification framework for bacterial plasmids.

3.2. Methods

3.2.1. Scoring normalized mutual information (NMI) and purity

The correlation between plasmid cliques identified in Chapter 2 and replicon and MOB typing schemes was assessed by measuring the Normalized Mutual Information (NMI) and purity between them. NMI is a commonly used method to assess the performance of clustering algorithms (Fortunato & Hric, 2016). For the two clustering/classification schemes ($\mathbf{C}_1$ and $\mathbf{C}_2$) NMI is defined as (Fred & Jain, 2003):

$$\mathrm{NMI}(\mathbf{C_1}, \mathbf{C_2}) = \frac{\mathrm{I}(\mathbf{C_1}, \mathbf{C_2})}{\frac{[\mathrm{H}(\mathbf{C_1}) + \mathrm{H}(\mathbf{C_2})]}{2}} \,. \tag{3.1}$$

In equation (3.1), the mutual information, also known as the information gain and denoted as I($\mathbf{C_1}$,$\mathbf{C_2}$), is an information theory concept which measures the reduction of uncertainty around $\mathbf{C_1}$ given knowledge about the $\mathbf{C_2}$, and vice versa. It is normalized by the averaged Shannon entropy (H) between $\mathbf{C_1}$ and $\mathbf{C_2}$. Shannon entropy tends to be larger as the number of classes in $\mathbf{C_1}$ or $\mathbf{C_2}$ approach the size of the dataset in question. Consequently, the NMI is sensitive to differences in the number of classes between $\mathbf{C_1}$ and $\mathbf{C_2}$, and to extensively fragmented classifications. The NMI equals one if the two classifications yield identical partitioning of the dataset, whereas a value of zero indicates complete incoherence. The NMI was measured using the *R* package *NMI* (version 2.0). During the assessment, plasmids which were not classified by replication or MOB typing schemes were disregarded.

Purity was used to estimate the homogeneity of cliques for replicon or MOB types, and plasmid host taxa. For a set of cliques **C**, and a plasmid typing scheme **T**, purity is defined as:

$$\mathrm{purity}(\mathbf{C}, \mathbf{T}) = \frac{1}{N} \sum_{c_i \in \mathbf{C}} \max_{t_j \in \mathbf{T}} \left| c_i \cap t_j \right| \tag{3.2}$$

where N is the total number of plasmids covered by a set of cliques, $\mathbf{C} = \{ c_1, c_2, \ldots, c_i \}$ is a set of cliques in which plasmids were placed, and $\mathbf{T} = \{ t_1, t_2, \ldots, t_j \}$ are the types associated with plasmids. Similar to NMI, the purity scores take a value between 0 and 1 with high purity indicating high homogeneity of classes in the dataset for a given set of plasmid types. The purity was only assessed for cliques which contain at least one typed plasmid. Untyped plasmids found within the assessed cliques were disregarded.

Please refer to Chapter 2 for more information on the preceding network analysis and description of the dataset of complete bacterial plasmids.

3.3. Results

3.3.1. Plasmid cliques agree with current typing schemes

The clique purity and Normalized Mutual Information (NMI) metrics presented above, were used to assess the congruence of clique-based classification with replicon and MOB typing schemes. These metrics were calculated for cliques comprising plasmids with identified replicon type, plasmids carrying a single identified replicon type, or plasmids with assigned MOB type (Figure 3.1). Untyped plasmids were disregarded. The observed purity scores were high (>85%) indicating the homogeneity of cliques for a particular plasmid type (Figure 3.1). This was particularly the case for MOB types (purity = 0.9887) and plasmids assigned to a single replicon type (purity = 0.9522). NMI provides an entropy-based measure of the similarity between two classification systems where a score equal to one indicates identical partitioning into classes while zero means independent classification. NMI penalizes differences in the number of assignment classes which justifies the low score observed when assessing clique-based versus MOB-based typing which recognizes only six MOB types (NMI = 0.5223). Nevertheless, high NMI scores were obtained when considering a replicon-based classification scheme (NMI = 0.9044 all types, and NMI = 0.9336 for single replicon types). It follows that plasmids with the same replicon type often fall together within the same clique. This is also supported by the high correlation between the clique membership size and the number of plasmids assigned to the corresponding replicon class (Figure 3.2, R^2=0.862 for plasmids assigned to a single replicon types).

There are exceptions where plasmids from larger replicon classes are further resolved into a few smaller evolutionary related cliques (Figure 3.2 – area below $y=x$ line). One such example is provided by the 22 ‘broad-host-range’ IncP plasmids which have been split into three cliques (14, 118 and 332; Figure 3.3A). While plasmids within these cliques share notably high JI similarity, the similarities between cliques remain low. This is especially true for clique 332 and 14, for which between-clique similarity is zero. Interestingly, plasmids from clique 332 have been exclusively associated with Gammaproteobacteria, while the ones from cliques 118 and 14 are mostly found in hosts from the Betaproteobacteria class. This arrangement of IncP into multiple cliques with a more constrained host-range is in line with previous findings of weaker incompatibilities in IncP (Chikami et al., 1985) and the existence of multiple genetically distinct IncP sub-

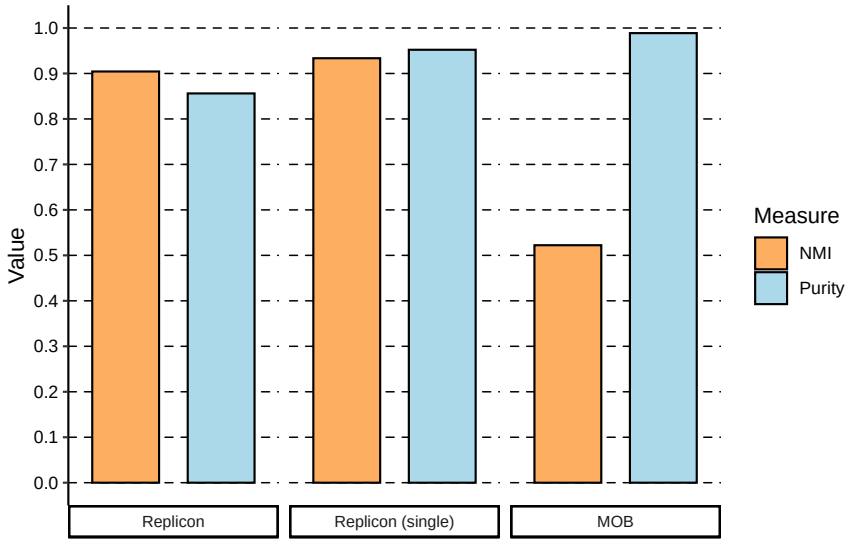


Figure 3.1. Concordance of plasmid clique assignment with replicon and MOB typing schemes. Normalized Mutual Information (NMI) and purity (y-axis: see legend) were calculated for cliques containing plasmids with identified replicon type, plasmids carrying a single identified replicon type, and plasmids assigned to MOB types.

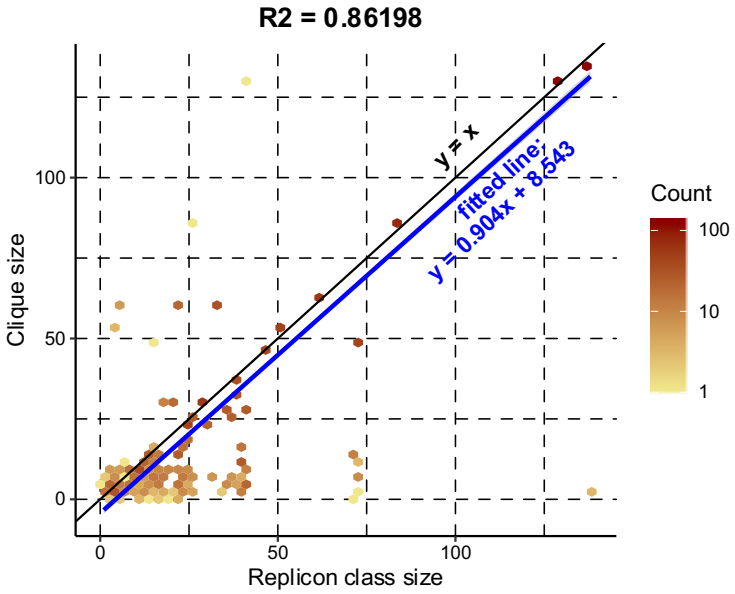


Figure 3.2. Plasmid clique size as a function of replicon class size. For each single-type plasmid in the dataset, the size of its corresponding clique and replicon class size were determined, y and x axis respectively. The colour intensity of each hexagon reflects the density of the data points (count scale). If all plasmids from a particular replicon type are encompassed by a single clique, the points corresponding to those plasmids would fall on the line y=x. The coefficient of determination (R^2) for the function y=x was estimated to be 0.86198. In addition, the slope of the function (y=0.904x + 8.543) derived from the data points by linear regression is less than that of a y=x which reflects the trend of plasmids from the same replicon class being sorted into multiple smaller cliques.

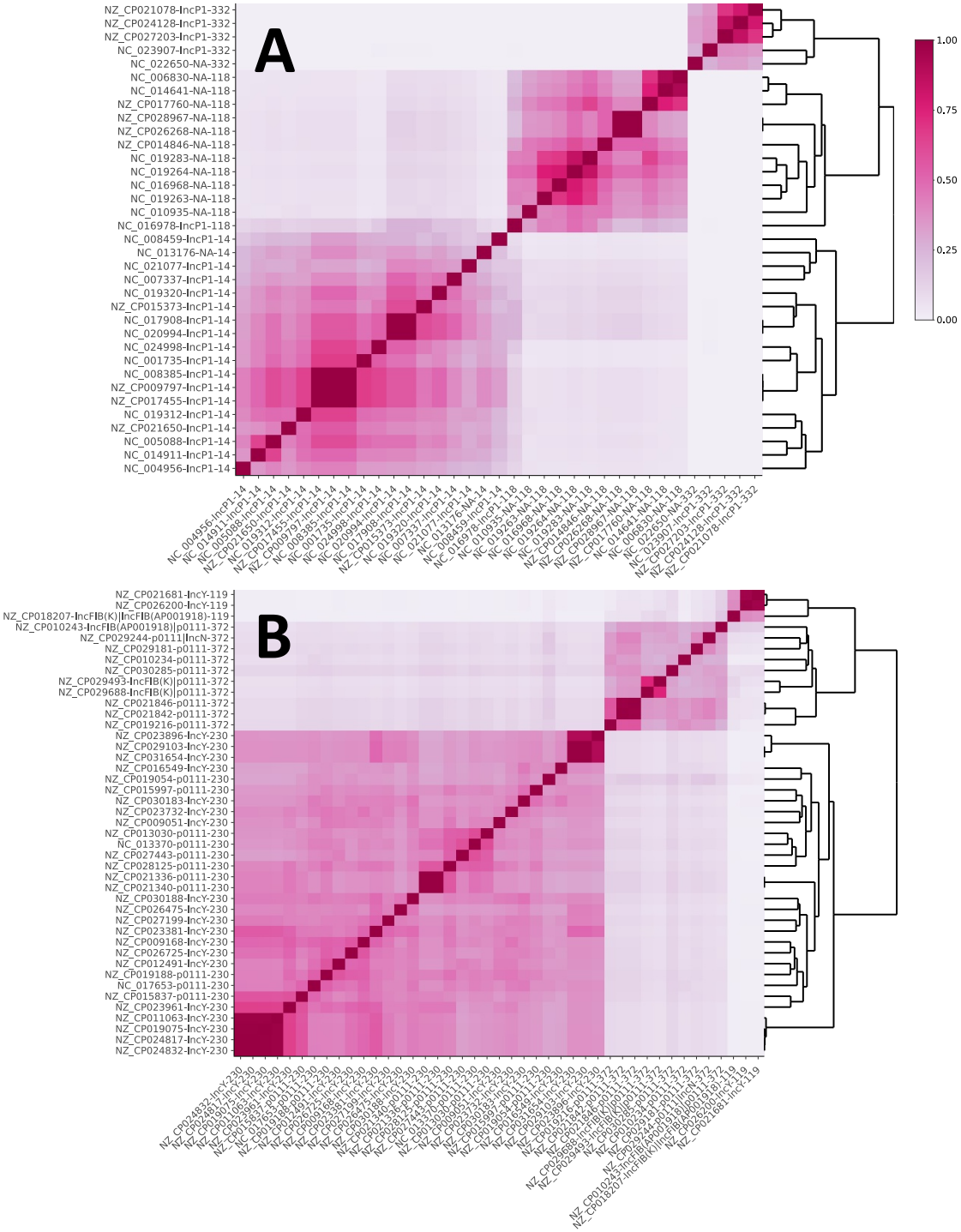


Figure 3.3. Heatmap of pairwise JI between plasmids from cliques containing IncP (A) and IncY and p0111 (B) replicon types. Plasmids were ordered using hierarchical clustering as provided by the dendrograms. The legend on the right matches the colour gradient of the heatmap with the corresponding JI value. The accession number of each plasmid sequence can be found on two symmetrical axes and it is followed by the replicon type and the clique number. *NA* denotes plasmids with unknown replicon type. Additional information about analysed plasmid sequences can be found in Supplementary Data 1 of Acman et al., 2020.

lineages whose backbone is coadapted to their host (Norberg et al., 2011). Another example of a genetically heterogeneous replicon type is provided by IncY and p0111 plasmids collected from *E. coli* strains which fall into three cliques (119, 230 and 372; Figure 3.3B). Clique 119 and 372 cluster IncY and p0111 plasmids respectively, with a single, possibly misassigned IncFIB plasmid. Conversely, clique 230 comprises both IncY and p0111 plasmids, with a remarkably related genetic backbone. The latter result raises questions on the distinctiveness of IncY and p0111 plasmid types.

3.3.2. Candidate replicon genes recovered from untyped plasmids

The majority of plasmids with unknown replicon types formed small cliques (Figure A.4). In fact, 81.02% of the smallest cliques (carrying three to five plasmids) contain exclusively untyped plasmids. Together with more than 5,000 singletons and lone plasmid pairs present in the network (Figure 2.6), this trend highlights the many understudied and underrepresented plasmids in sequence databases. Accordingly, the next objective was to investigate the genetic content of untyped cliques to determine candidate replicon genes and further traits of biological relevance.

In total, there are 388 cliques with no assigned replicon types. As the cliques tend to be homogeneous for a replicon type, only the core genes (i.e., genes occurring on all plasmids of a particular clique) found on untyped cliques were considered. Core genes were translated into protein sequences and screened against the translated PlasmidFinder database using TBLASTN (Altschul et al., 1990). A range of e-values were assessed to determine the threshold maximizing the discovery of replicon candidates while minimizing false positives (Figure 3.4). The majority of plasmids were assigned to one replicon type with some plasmids having hits to a maximum of three to four different types. Accordingly, the optimal e-value threshold was selected when the total number of core gene hits started to diverge from the number of untyped cliques covered. A conservative e-value threshold of 0.001 was chosen which resulted in the identification of 105 candidate genes from 106 plasmid cliques. The accession numbers and positions of candidate genes are listed in Table A.1.

To verify the plausibility of the identified gene candidates, HMMER (version 3.2.1) was used to scan amino acid sequences for known protein domain families found in the Pfam database (version 32.0) (El-Gebali et al., 2019). 166 families, with e-values lower than 0.001, were identified on 97 protein sequences and were most commonly associated

with replication initiation (Figure 3.5). Moreover, the majority of functions associated with the discovered protein families relate to plasmid replicon proteins. For example, domains with helix-turn-helix motifs are important for DNA binding of replicon proteins and allow some proteins to regulate their own transcription (del Solar et al., 1998). Other examples of transcriptional regulators also exist in plasmid replicon regions, while DNA primase activity has been found on the RepB replicon protein (del Solar et al., 1998). Interestingly, replicon proteins involved in rolling-circle replication (a mechanism of plasmid replication) share some of their motifs with proteins involved in plasmid transfer and mobilization (del Solar et al., 1998). This could explain why some of the discovered domain families are linked to plasmid mobilization. On the whole, the candidate replicon genes are highly specific to a particular clique of plasmids and should assist description of new incompatibility types.

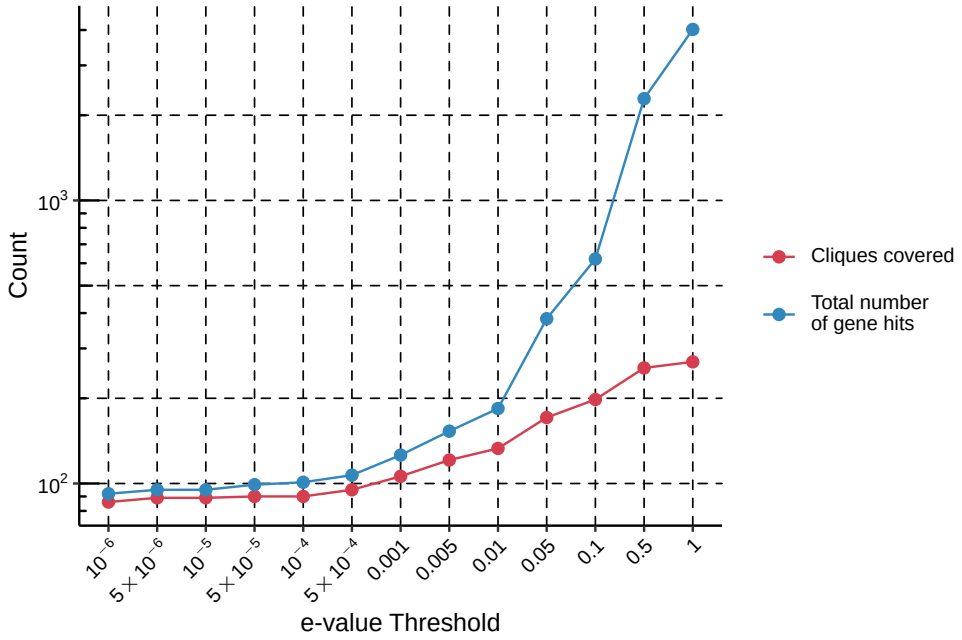


Figure 3.4. Finding the optimal e-value threshold for discovery of candidate replicon genes within untyped plasmid cliques. The core genes presented in untyped cliques were screened against the PlasmidFinder replicon database using TBLASTN. The number of cliques covered by the blast search and the total number of core gene hits were recorded. The majority of plasmids within the dataset are assigned to a single replicon type though with some plasmids assigned to a maximum of three to four different types. Upon reaching the e-value of 0.01, the number of gene hits obtained starts to rapidly diverge from the number of cliques covered disclosing false positives. Therefore, a conservative e-value of 0.001 was chosen as the final threshold.

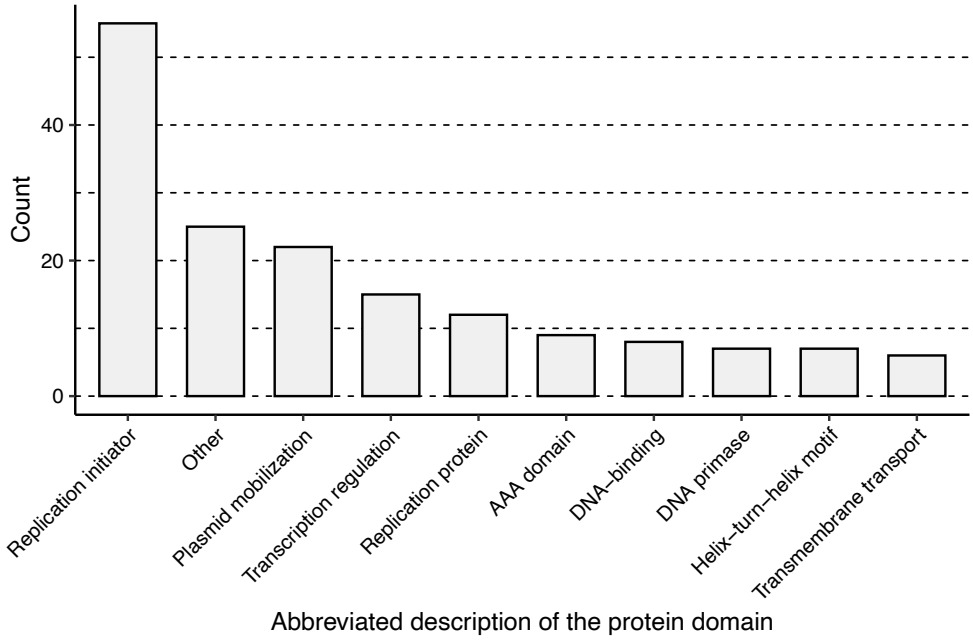


Figure 3.5. Protein domain families found associated with the candidate replicon genes. Protein sequences of candidate replicon genes were screened using HMMER tool against the Pfam database. The domain families discovered were binned according to their abbreviated description.

3.3.3. Cliques exhibit common GC content and bacterial hosts

In the previous chapter, I demonstrated that the unprocessed plasmid network (Figure 2.2) recovered pronounced structure in terms of the plasmid nucleotide composition, measured by the GC content. This trend is also reflected in the clique composition (Figure 3.6A). Within a clique, the standard deviation of GC content rarely exceeds 0.02 and is weakly correlated with the clique size (R^2 = 0.0155) (Figure 3.6B). Moreover, a significant difference in GC content is often found between cliques. Analysis of variance (ANOVA), followed by a Tukey test, found that 85.3% of the time the GC content between two cliques differs significantly (adjusted p-value < 0.001). In contrast, the sequence lengths of plasmids within a clique are more variable but are also not strongly correlated with clique size (R^2 = 0.029) (Figure 3.6C and D). Similarly, a Tukey test showed that a significant difference in plasmid length between cliques is observed less than 34% of the time (adjusted p-value < 0.001).

Plasmid GC content has been shown to be strongly correlated to the base composition of the bacterial host's chromosome (Nishida, 2012). Indeed, the cliques showed a very high homogeneity (purity) relative to their hosts (Figure 3.7), a trend which has been identified in other plasmid network reconstruction efforts (Tamminen et al.,

2012). At higher taxonomic levels, cliques have near perfect purity scores (>0.99). The purity score slightly decreases at the level of the plasmid host family, reaching a value of 0.807 at the species level. Therefore, plasmids with high genetic similarity rarely transcend the level of the bacterial genus, which suggests a limited host range for the vast majority of plasmids. However, these results need to be carefully considered due to inherent biases in the dataset, especially in terms of the predominance of well-studied taxa. Overall, the plasmid cliques show a strong intrinsic propensity towards confined GC content and are found in a limited range of bacterial hosts.

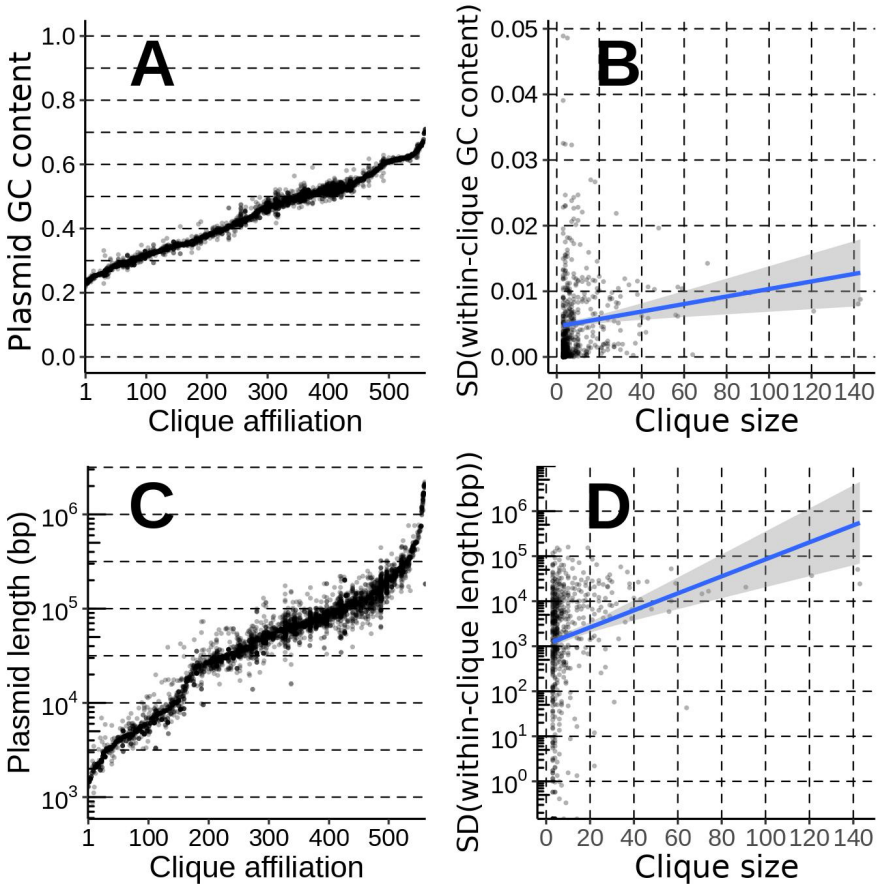


Figure 3.6. Variability of plasmids within cliques in GC content and length. The GC content (A) and the length (C) of a particular plasmid (y-axis) relative to its clique affiliation (x-axis columns). The clique affiliation of plasmids has been ordered based on the average clique GC content and the average clique length in panels A and C respectively. Hence, the numbers on both x-axes in panels A and C are arbitrary and do not correspond to the same clique. To further explore within-clique variability, the standard deviation (SD) of the within-clique GC content (B) and length (D) were considered with respect to the clique size. A weak linear correlation with clique size was found for both SD measurements (R^2 = 0.0155 and R^2 = 0.029 for B and D respectively).

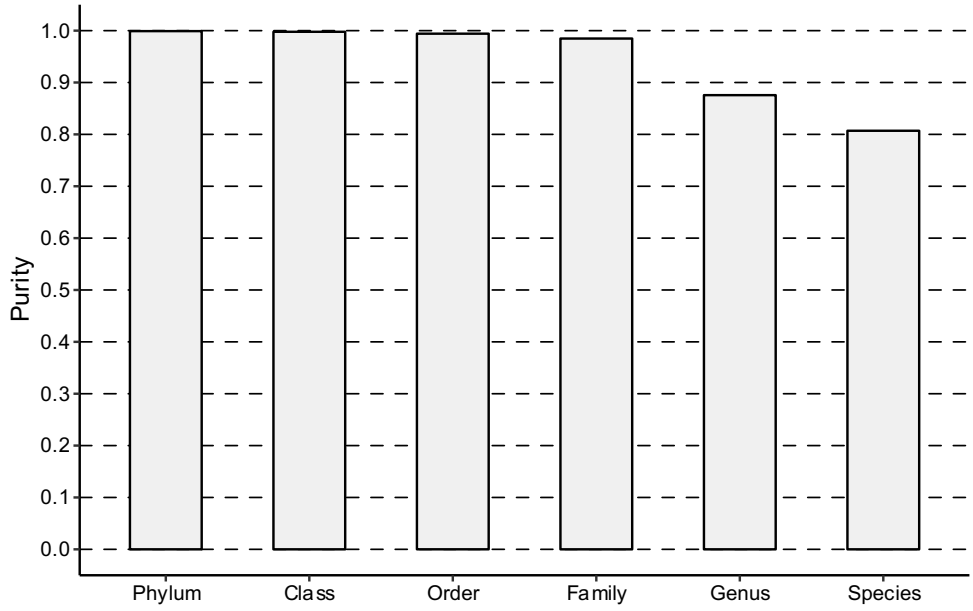


Figure 3.7. Purity of cliques relative to the taxonomic level of the plasmid bacterial host. For each clique, the frequency of the most abundant taxonomic level in question was calculated. Purity score represents the average of those frequencies (see section 3.2.1). Therefore, a purity of one would indicate that all plasmids from any clique belong to the same genus, species, or other taxonomic level.

3.3.4. Plasmids within cliques have uniform gene content

The gene content of cliques was assessed for all genes occurring five or more times in the dataset. This threshold was chosen to facilitate computation, and to adequately characterize prevalent genes. In total, 15,851 out of 35,883 (44.17%) of the assessed genes were 'core' genes, meaning they had a within-clique frequency equal to one, suggesting an overall uniformity of gene content in cliques (Figure 3.8). Furthermore, 6,577 (18.33%) of the genes were considered 'private'. Private genes are those found in only one clique, with a frequency of one, and their relatively high abundance in the dataset suggests the uniqueness of some cliques with respect to their gene content. However, there is an inherent bias. Plasmids within larger cliques tend to be more dissimilar and share proportionally fewer genes (Figure A.8). This pattern can in part be explained by the broader gene content of large cliques and the high sequence similarity required for same-gene clustering (95%) within the default implementation of the Prokka-Roary annotation pipeline. 31.94% of cliques containing five or more plasmids were found to have between one and 10 core genes. However, cliques exhibited a wide range in the number of core genes with 7.74% of cliques carrying over 100 shared genes. Interestingly, 13.55% (42) of cliques had no core genes which could also be an artefact of the gene

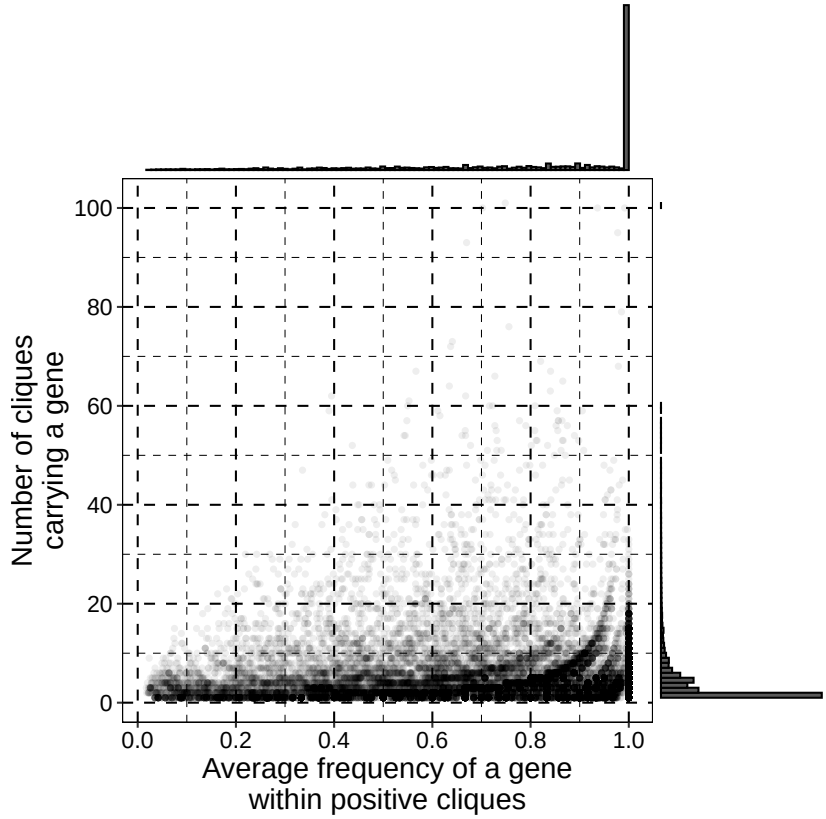


Figure 3.8. Assessing the frequency of genes within cliques. The average within-clique frequency of all genes with five or more occurrences in the dataset was calculated and plotted against the number of cliques in which a particular gene occurs. Within-plasmid gene duplications were disregarded and counted as a single occurrence. The histograms on the top and the right-hand side provide the distribution of values for the two axes.

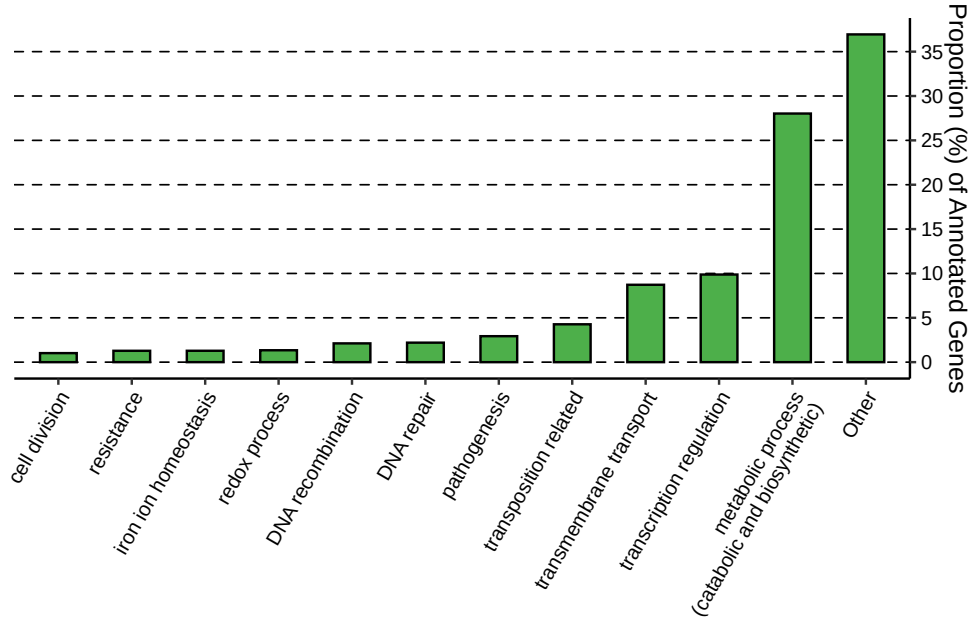


Figure 3.9. Distribution of the biological functions associated with the core genes in plasmid cliques. The respective frequency (%) of each gene was considered.

annotation pipeline or poor-quality assemblies. For instance, plasmids from 19 cliques carried no recognized genes from the pool of 35,883 assessed genes.

Functionally, core genes were found to be more often associated with various metabolic processes, transcription regulation and transmembrane transport (Figure 3.9) when compared to the overall distribution of GO terms, shown in Figure 2.1B of the previous chapter. Similarly, fewer core genes were involved in transposon movement, pathogenesis, and resistance.

3.3.5. Inferring horizontal gene transfer through clique interactions

Gene content was also considered in the context of clique structure and interconnectedness. To do so, the original network of plasmids (Figure 2.2) was rearranged such that: (i) plasmids assigned to the same clique were clustered under a single vertex; (ii) plasmids assigned to multiple cliques were left as solitary vertices anchoring the cliques; (iii) unassigned plasmids were removed. The resulting network is shown in Figure 3.10. As highlighted earlier and visible in the figure, large cliques generally show lower internal similarity compared to the smaller ones. It is important to note that an arbitrary JI threshold of 0.01 was introduced in Figure 3.10 to assist visual interpretation, but the unfiltered version of the network is provided in Figure A.9.

The clustering of cliques in Figure 3.10 shows high concordance with the phylogenetic hierarchy of the bacterial hosts. On a global scale, there are four large, interconnected clusters (three corresponding to cliques from the phylum Firmicutes and one from the Proteobacteria), eight disjointed clusters, and a dozen singled-out triplets and pairs. The clique clusters mostly contain plasmids from a specific genus with some minor deviations – hence the cluster naming. The only two exceptions are the large and diverse Proteobacteria cluster which harbours plasmids mainly from the genera *Escherichia*, *Klebsiella*, and *Salmonella*, and the lactic acid bacteria cluster here referred to as ‘the Dairy bacteria’.

The majority of genes identified in the four large clusters were those functionally involved in transposition. Specifically, 26.4% of the genes in the Proteobacteria cluster were transposition related. In addition, 9.66% of the genes in the Proteobacteria were involved in some form of AMR or metal resistance, and 7.38% in pathogenesis, which may reflect the high number of pathogens found in this phylum (Rizzatti et al., 2017).

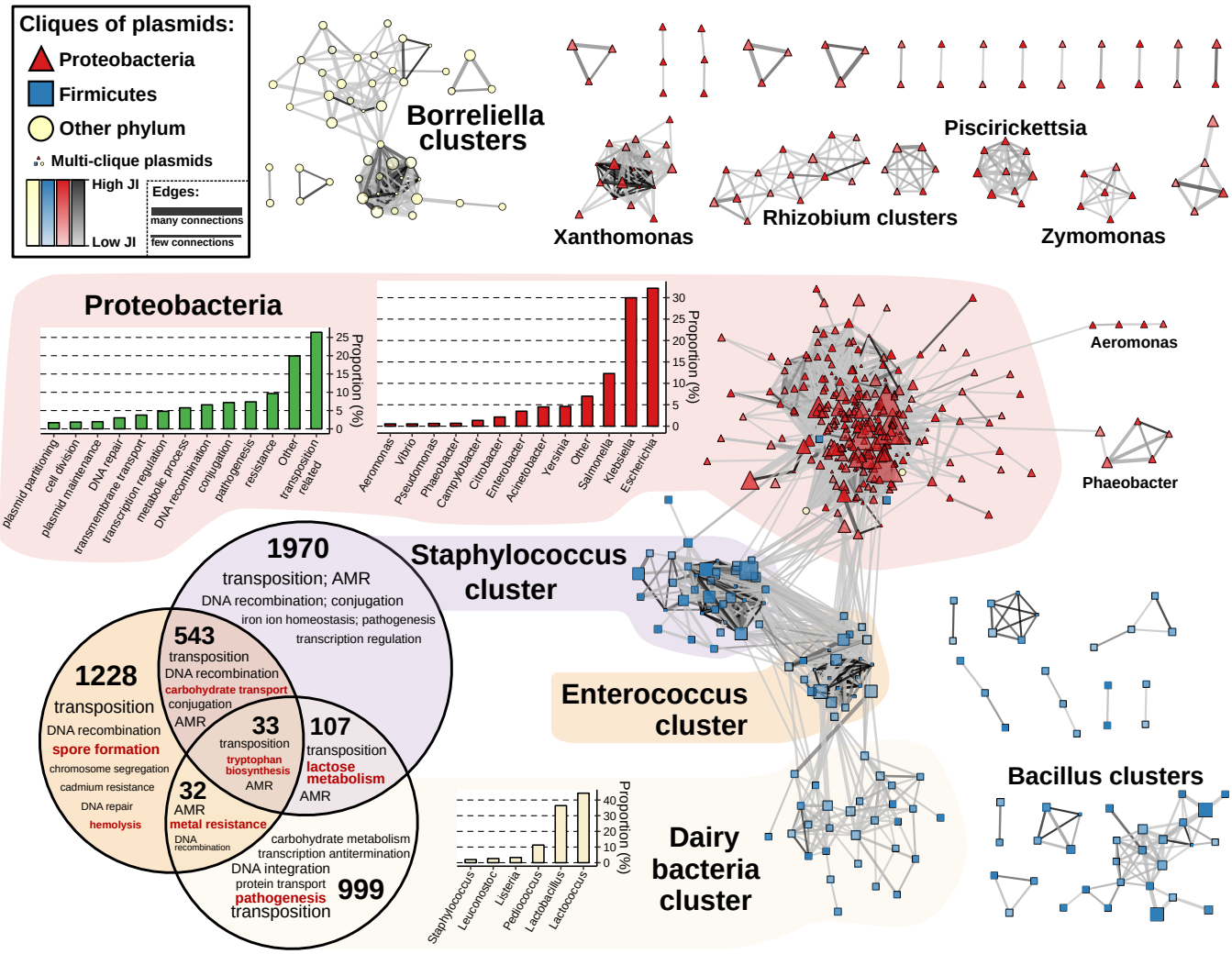


Figure 3.10. The network of cliques. Cliques, represented as vertices, are connected with an edge if the average Jaccard Index (JI) between plasmids of two cliques is higher than 0.01. The colour of the edges indicates the average JI while the width is proportional to the number of connections between a pair of cliques. The shape and colour of the cliques indicates the phylum of the predominant bacterial host. The size and the transparency are proportional to the clique size and the internal JI respectively. The cliques form multiple clusters which have been named based on the genus of the bacterial host characteristic for a particular cluster. There are two exceptions – the Proteobacteria and the Dairy (Lactic) cluster whose respective genera distributions have been provided. The most common GO biological functions of the genes found on plasmids of Proteobacteria, *Staphylococcus*, *Enterococcus* and Dairy clusters were further assessed. During the assessment, the respective frequencies of the genes were considered. In case of Proteobacteria, the bar chart distribution of the biological functions is provided. The shared and core gene content of *Staphylococcus*, *Enterococcus* and Dairy clusters is presented in the Venn diagram with the numbers in the diagram indicating the number of core and shared genes.

The core and shared gene content of the three Firmicutes clusters (*Staphylococcus*, *Enterococcus* and Dairy) was also assessed (Figure 3.10, Venn diagram). Gene sharing was most common between the clusters associated with the genera *Staphylococcus* and *Enterococcus* indicating a high frequency of HGT between them, and the least between the *Staphylococcus* and Dairy bacteria cluster. Analysing the content of these shared genes provides insight into both plasmid function and dynamics, such as the identification of HGT events. For example, the same lactose metabolism genes were found in association with both *Staphylococcus* and Dairy bacteria plasmids. Also, the *trpF* gene, involved in tryptophan biosynthesis, and the bla_{NDM} AMR gene were previously associated with the Tn*3000* and Tn*125* transposable elements (Campos et al., 2015; H. Hu et al., 2012) and are found on plasmids in all three clusters. In contrast to these, the more disjoint clusters of plasmid cliques observed for other genera may be driven by the species' ecology and life history, which may lead to limited opportunities for contact between lineages. Such an explanation seems plausible for strict pathogens with restricted host range, such as *Xanthomonas* or *Borrellelia*. Conversely, for lineages with a wider environmental niche like *Bacillus*, the lower connectivity between cliques may be due to intrinsic genetic factors leading to lower between-plasmid recombination and/or transposition rates.

3.4. Discussion

The network analysis conducted in Chapter 2 classified over 5,000 complete bacterial plasmids into 561 cliques of size three or more. In this chapter, I present evidence these cliques capture biologically meaningful information. In particular, plasmids assigned to the same clique show strong correlation with replicon and MOB based typing schemes as evaluated by NMI and purity scores. Cliques show high homogeneity in terms of their respective bacterial hosts, as well as similar GC and gene content. Moreover, the genes shared among plasmids from the same clique (i.e., core genes) were found to be less associated with functions frequently involved in plasmid genome rearrangement, such as transposon movement, pathogenesis, and resistance.

Taken together, these findings imply cliques delineate clusters of plasmids with shared evolutionary histories which in turn allows for further inferences on the nature of HGT and plasmid function. The strong, host-constrained population structure documented for the majority of bacterial plasmids points to transposable genetic elements

as the main drivers of HGT on a broader phylogenetic scale. This is further corroborated by the excess of transposable elements shared between cliques of different taxa, and likely extends to so-called 'broad-host-range' plasmids such as IncP, whose representatives in the dataset fell into three genetically distinct cliques associated to different host species.

Furthermore, evolutionary structured plasmid cliques facilitated identification of new replicon gene candidates, as well as detailed investigation of the distribution of other plasmid-borne genetic determinants of incompatibility, mobility, AMR, virulence, and transposon carriage. Such meta-information could be incorporated within the network framework thanks to a plethora of well-maintained bioinformatics tools, ever growing genetic databases, and gene ontology efforts to systematize gene annotation. One of the more interesting examples are *trpF* gene and bla_{NDM} AMR gene which are found coupled on plasmids across *Enterococcus*, *Staphylococcus*, and other dairy bacteria genera. The two are often incorporated within the *Tn*3000 and *Tn*125 transposons, and the story about their global dissemination is unravelled in the following chapter.

Presented results suggest it should be possible to devise a 'natural', global sequence-based classification scheme for bacterial plasmids. That being said, my findings do not diminish the relevance of replicon and MOB typing schemes, rather they build upon these prior classification schemes and may extend them to plasmids from understudied and uncultured bacteria. Beyond just plasmid classification, the presented network-based approach also has potential to infer key features of plasmid groupings. Indeed, plasmid clique assignment can be completely automated and inspection of any particular area of the network facilitates biological inference about plasmid dynamics and their biological features within various groups of bacterial hosts.

Chapter 4

Tracing the global dissemination of the *bla*$_{NDM}$ resistance gene

Declaration of contributions

Francois Balloux, Lucy van Dorp, Hui Wang and I conceived the project and designed the experiments and analyses. Lucy van Dorp, Nina Luhmann and I collected data from online repositories. Ruobing Wang, Yuyao Yin, Qi Wang, Shijun Sun, and Hongbin Chen sequenced samples from Chinese hospitals. Lucy van Dorp, Ruobing Wang and I *de novo* assembled all the genomes. I performed all the analyses and conceptualized the computational methods under the guidance of Lucy van Dorp and Francois Balloux.

Publication

This work has been published in Nature Communications as Acman et al. (2022):

Acman, M., Wang, R., van Dorp, L., Shaw, L.P., Wang, Q., Luhmann, N., … & Balloux, F. (2022). Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene blaNDM. Nature Communications, 13(1), 1-3. https://doi.org/10.1038/s41467-022-28819-2

4.1. Introduction

As discussed in Chapter 1, AMR can be conferred by vertically inherited point mutations or via the acquisition of horizontally transmitted 'accessory' genes, often located in transposons and plasmids. The *bla*$_{NDM}$ gene represents a typical example of a mobile AMR gene (W. Wu et al., 2019). *bla*$_{NDM}$ encodes a metallo-β-lactamase capable of hydrolysing most β-lactam antibiotics. These antibiotics are used as a first-line treatment for severe infections and to treat multidrug-resistant Gram-negative bacterial infections. As such, the global prevalence of bacteria carrying *bla*$_{NDM}$ represents a major public health concern.

bla$_{NDM}$ was first described in 2008 from a *Klebsiella pneumoniae* isolated from a urinary tract infection in a Swedish patient returning from New Delhi, India (Yong et al., 2009). While *bla*$_{NDM}$ now has a worldwide distribution, most of the earliest cases have been linked to the Indian subcontinent, leading to this region being suggested as the likely location for the initial mobilisation event (Castanheira et al., 2011; Kumarasamy et al., 2010; Poirel, Dortet, et al., 2011; Struelens et al., 2010; W. Wu et al., 2019). NDM-positive *Acinetobacter baumannii* isolates have been retrospectively identified from an Indian hospital in 2005 (Jones et al., 2014), which remain the earliest observations to date. Nevertheless, an NDM-positive *Acinetobacter pittii* isolate was also isolated in 2006 from a Turkish patient with no history of travel outside Turkey (Roca et al., 2014).

Although no complete genome sequences are publicly available from these earliest observations, the first NDM-positive isolates from 2005 were shown to carry *bla*$_{NDM}$ on multiple non-conjugative, but potentially mobilizable plasmid backbones (Jones et al., 2014). In addition, *bla*$_{NDM}$ in these early isolates was positioned within a complete Tn*125* transposon with IS*26* insertion sequences (ISs) as well as ISCR*27* (IS-containing common region 27), suggesting the possibility of complex patterns of mobility since the gene's initial integration. Subsequent NDM-positive isolates across multiple species consistently harbour either a complete or fragmented IS*Aba125* (an IS constituting Tn*125*), immediately upstream of *bla*$_{NDM}$, which provides a promoter region (Poirel, Bonnin, et al., 2011; Poirel, Dortet, et al., 2011; Toleman et al., 2012; W. Wu et al., 2019). The presence of IS*Aba125* in some form in all NDM-positive isolates to date and the early observations in *A. baumannii* have led to Tn*125* being proposed as the ancestral

transposon responsible for the mobilization of *bla*$_{NDM}$, and *A. baumannii* as the ancestral host (Partridge & Iredell, 2012; Toleman et al., 2012).

The NDM enzyme itself is of possible chimeric origin (Partridge & Iredell, 2012; Toleman et al., 2012), with the first six amino acids in NDM matching to those in *aphA6*, a gene providing aminoglycoside resistance and also flanked by IS*Aba125*. It is hypothesised that ISCR*27*, which uses a rolling-circle (RC) transposition mechanism (Ilyina, 2012; Toleman et al., 2006), initially mobilized a progenitor of *bla*$_{NDM}$ in *Xanthomonas sp.* and placed it downstream of IS*Aba125* (Partridge & Iredell, 2012; Poirel et al., 2012; Sekizuka et al., 2011; Toleman et al., 2012). At least 29 distinct sequence variants of the NDM enzyme have been described to date (McArthur et al., 2013; W. Wu et al., 2019). The most prevalent of these variants is the first to have been characterised, denoted NDM-1 (Basu, 2020). Different NDM variants are mostly distinguished by a single amino-acid substitution, apart from NDM-18 which carries a tandem repeat of five amino acids. None of the observed substitutions occur in the active site and their functional impact remains under debate (W. Wu et al., 2019).

bla$_{NDM}$ is found in at least 11 bacterial families and NDM-positive isolates have heterogeneous clonal backgrounds, supporting multiple independent acquisitions of *bla*$_{NDM}$ (W. Wu et al., 2019). Although *bla*$_{NDM}$ has been observed on bacterial chromosomes (Baraniak et al., 2016; Rahman et al., 2018) it is most commonly found on plasmids, comprising multiple different backbones or types. Thus far, *bla*$_{NDM}$ has been associated to at least 20 different plasmid types, predominantly IncFIB, IncFII, IncA/C (IncC), IncX3, IncH, and IncL/M, and also in untyped plasmids (H. Hu et al., 2012; Kumarasamy et al., 2010; Rasheed et al., 2013; Wailan et al., 2015; W. Wu et al., 2019; Yang et al., 2015). Furthermore, even within the same plasmid type, *bla*$_{NDM}$ can be found in a variety of genomic contexts, often interspersed by multiple ISs and composite transposons (Partridge & Iredell, 2012; W. Wu et al., 2019). The immediate environment of *bla*$_{NDM}$ has been reported to vary even in isolates from the same patient (Wailan et al., 2015). Many mobile elements are thought to play important roles in dissemination, including IS*Aba125*, IS*3000*, IS*26*, IS*5*, ISCR1, Tn*3*, Tn*125*, and Tn*3000* (Campos et al., 2015; Feng et al., 2018; Poirel et al., 2012; Wailan et al., 2015; W. Wu et al., 2019; Q.-Y. Zhao et al., 2021). It is therefore clear that the spread of *bla*$_{NDM}$ was, and is, a multi-layer process involving multiple mobile genetic elements – 'the mobilome'. *bla*$_{NDM}$ mobility involves diverse processes, including genetic recombination (T. W. Huang et al., 2013; Lynch et al., 2016), transposition, conjugation and transformation of plasmids

(Datta et al., 2017), transduction (Krahn et al., 2016), and transfer through outer-membrane vesicles (OMVs) (Chatterjee et al., 2017; González et al., 2016) (see Chapter 1 for a description of some of these mechanisms).

Previous surveys of *bla*$_{NDM}$-positive genomes have led to a better understanding of its evolution (W. Wu et al., 2019). However, a major difficulty, as for other AMR genes, is relating the diverse genomic contexts to temporal evolution. In this chapter, I outline an alignment-based method to identify flanking structural variants and use it to build a history of the insertion and mobilization events. To this end, I compiled a global dataset of more than 6,000 NDM-positive isolates. In line with previous studies, Tn*125*, IS*26* and Tn*3000* were identified as the main contributors to *bla*$_{NDM}$ mobility. Nevertheless, I go further and estimate the timing of the initial emergence of *bla*$_{NDM}$ to pre-1990, around two decades prior to its first detection and rapid dissemination worldwide. These findings suggest that this global spread was driven primarily by transposons, with plasmids playing more of a role in local transmission.

4.2. Methods

4.2.1. Compiling the curated dataset of NDM sequences

An extensive dataset of 6,155 bacterial genomes carrying the *bla*$_{NDM}$ gene was compiled from several publicly available databases. 2,632, 1,158, and 1,379 fully assembled genomes were downloaded from NCBI Reference Sequence Database (O'Leary et al., 2016; Pruitt et al., 2007; RefSeq; accessed on 15th of April 2021), NCBI's GenBank (Benson et al., 2013; accessed on 15th of April 2021), and EnteroBase (Zhou et al., 2020; accessed on 27th of April 2021) respectively. The EnteroBase repository was screened for *bla*$_{NDM}$ using BlastFrost (v1.0.0; Luhmann et al., 2020) allowing for one mismatch. In addition, the Bitsliced Genomic Signature Index (BIGSI) tool (v0.3; Bradley et al., 2019) was used to identify all Sequence Read Archive (SRA) unassembled reads which carry the *bla*$_{NDM}$ gene. At the time of writing, a publicly available BIGSI demo did not include sequencing datasets from after December 2016. Therefore, I manually indexed and screened an additional 355,375 SRA bacterial sequencing datasets starting from January 2017 to January 2019. The presence of 95% of *bla*$_{NDM-1}$ *k*-mers was required to identify NDM-positive samples from the raw SRA reads. This led to the inclusion of a further 882 isolates. The dataset also included 104 NDM-positive genomes from 79 hospitalized patients across China and 25 livestock farms selected from two

previous studies (Q. Wang et al., 2018; R. Wang, Liu, et al., 2018). These were sequenced using paired-end Illumina (Illumina HiSeq 2500) and PacBio (PacBio RS II). The sequencing reads are available on the Short Read Archive (SRA) under accession number PRJNA761884. All reads were *de novo* assembled using Unicycler (v0.4.8; Wick et al., 2017) with default parameters while also specifying hybrid mode for those isolates for which both Illumina short-read and PacBio long read sequencing data was available. Spades (v3.11.1; Bankevich et al., 2012) was applied, without additional polishing, for cases where Unicycler assemblies failed to resolve.

Assembled genomes were retained when they derived from a single BioSample identifier. Contigs carrying the *bla*$_{\text{NDM}}$ gene were identified using Mega BLAST (v2.10.1+; Camacho et al., 2009). Forty-eight contigs were found to carry more than one copy of *bla*$_{\text{NDM}}$ and were not included in the analysis and eighty-eight contigs were excluded due to having partial (<90%) *bla*$_{\text{NDM}}$ hits. Fourteen assemblies had a single *bla*$_{\text{NDM}}$ gene split into two contigs; these 28 contigs were also excluded. Several contigs were also removed due to poor assembly quality. The filtering resulted in a dataset of 7,148 contigs (6,155 samples) which were used in all subsequent analyses. Of these, 958 assembled genomes were found to contain *bla*$_{\text{NDM}}$ on multiple (mostly two) contigs, each harbouring a single and complete copy of *bla*$_{\text{NDM}}$. Even though the information about sequencing platform or assembly methods of most samples from RefSeq, GenBank and Enterobase databases could not be determined, the distribution of *bla*$_{\text{NDM}}$-positive contig lengths (Figure 4.1) indicates that they are likely to be based on short read assemblies

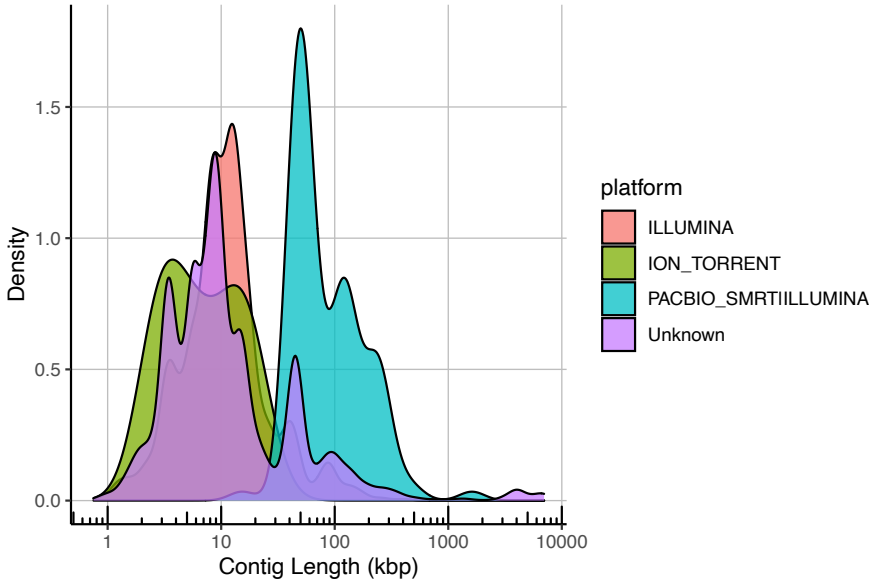


Figure 4.1. Marginal density distribution of the lengths of all assembled blaNDM-positive contigs depending on the sequencing platform.

with the minority of contigs, mostly from RefSeq, reaching the quality of a hybrid *de novo* assembly. The metadata of newly sequenced samples from mainland China is available in Table A.2.

4.2.2. Annotating the dataset

Full metadata for each genome was collected from its respective database and the R package 'taxize'(Chamberlain & Szöcs, 2013) used to retrieve taxonomic information for each sample. In the case of samples for which exact sampling coordinates were not provided, the Geocoding API from Google cloud computing services was used to retrieve coordinates based on location names.

The coding sequences (CDS) of all NDM-positive contigs were annotated using the annotation tool Prokka (v1.14.6; Seemann, 2014) and Roary (v3.13.0; Page et al., 2015) run with minimum blastp percentage identity of 90% (-i 0.9) and disabled paralog splitting (-s). To identify plasmid replicon types (Orlek, Stoesser, et al., 2017), contigs were screened against the PlasmidFinder replicon database (version 2020-02-25; Carattoli et al., 2014) using Mega BLAST (v2.10.1+; Camacho et al., 2009) where only BLAST hits with a minimum coverage of 80% and percentage identity of ≥95% were retained. In cases where two or more replicon hits were found at overlapping positions on a contig, the one with the higher percentage identity was retained. Identified plasmid types were used to cluster contigs into broader plasmid groups: IncX3, IncF, IncC, IncN2, IncHI1B, IncHI2, and other.

NDM-positive contigs were also screened against a dataset of complete bacterial plasmids. Bacterial plasmid references were obtained from RefSeq (O'Leary et al., 2016) and curated as described in the Methods section of Chapter 2. Mash, a MinHash based genome distance estimator (Ondov et al., 2016), was applied with default settings to evaluate pairwise genetic distances between contig sequences and plasmid references. Contig-reference hits with a pairwise Mash distance of less than 0.05 and a pairwise difference in length less than 20% were retained. Additional pruning was implemented such that, for each contig analysed, only the 10% of best scoring plasmid reference hits were retained. A table of pairwise genetic distances between contigs and references was represented as a network which was then analysed with the Infomap (Rosvall & Bergstrom, 2008) community detection algorithm. Contigs were annotated according to their community membership and the network was visualized using Cytoscape (Shannon et al., 2003).

4.2.3. Algorithm for resolving structural variations

A novel alignment-based approach was developed to identify stretches of homology (i.e., maximal alignable regions) as well as structural variations across all contigs upstream and downstream of *bla*$_{NDM}$ gene. A conceptual illustration of the method is presented in Figure 4.2. First, contigs carrying *bla*$_{NDM}$ were reoriented such that the *bla*$_{NDM}$ gene was located on the positive-sense DNA strand (i.e., facing 5' to 3' direction). A discontiguous Mega BLAST (v2.10.1+; Ma et al., 2002) search with default settings was then applied against all pairs of retained contigs (Figure 4.2A). This method was selected over the regular Mega BLAST implementation as it is comparably fast, but more permissive towards dissimilar sequences with frequent gaps and mismatches. BLAST hits including a complete *bla*$_{NDM}$ gene represent maximal stretches of homology around the gene for every pair of contigs. The analysis continues by considering only portions of BLAST hits at (i) the start of *bla*$_{NDM}$ gene and the downstream sequence or (ii) the end of the *bla*$_{NDM}$ gene and the upstream sequence depending on the analysis at hand, referred to as the downstream or the upstream analysis, respectively. This trimming of BLAST hits establishes *bla*$_{NDM}$ as an anchor and enables comparisons to be made across all samples.

A table of BLAST hits can be considered as a network (graph), where each pair of contigs (i.e., nodes) are connected by the edge weighted by the length of the BLAST hit (Figure 4.2 B). The algorithm proceeds with a stepwise network analysis of BLAST hits. For this purpose, a 'splitting threshold' was introduced. Starting from zero which represents the start/end position of *bla*$_{NDM}$ gene, the threshold is gradually increased by 10 bp. At each step, BLAST hits with a length lower than the value given by the 'splitting threshold' are excluded. Thus, as the 'splitting threshold' increases, a network of BLAST hits is also pruned and broken down into components – groups of interconnected nodes (contigs). It is expected that contigs within each component share a homologous region downstream (or upstream) of *bla*$_{NDM}$ at least of the length given by the threshold. It is therefore not possible for a single contig to be assigned to multiple components. Components of size <10 are labelled as 'Other Structural Variants'. Also, contigs that are shorter than the defined 'splitting threshold' and share no edge with any other contig are considered as 'cutting short'.

By tracking the splitting of the network as the 'splitting threshold' is increased, one can determine clusters of homologous contigs at any given position downstream or upstream from the anchor gene (here *bla*$_{NDM}$), as well as the homology breakpoint (Figure

4.2C). The precision of the algorithm is directly influenced by the step size, in this case 10 bp, and the alignment algorithm, in this case discontiguous Mega BLAST. I assessed the precision of the algorithm on the tree of structural variations downstream of *bla*$_{NDM}$ (show in Figure 4.13). To this end, I compared extended 50 bp sequence fragments of each branching point in the tree checking for missed homologies and comparing Mash distances between pairs of branched-out contigs. I found no similarities among 50 bp fragments of any split branches. The described algorithm is available at https://github.com/macman123/track_structural_variants.

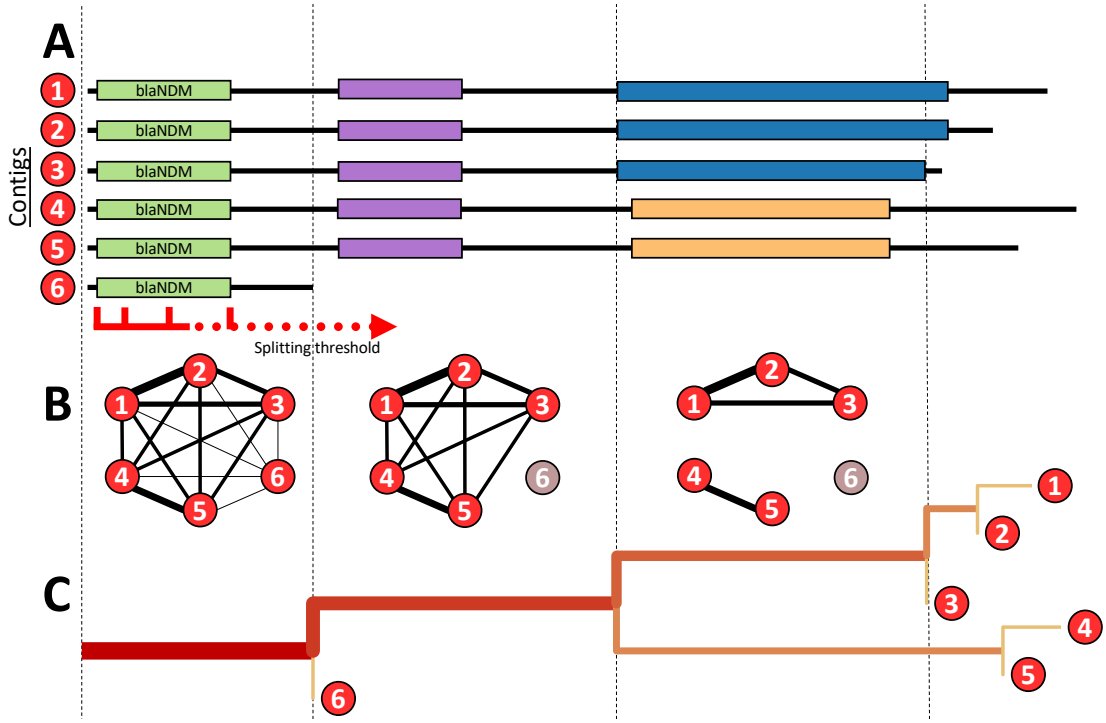


Figure 4.2. Schematic representation of the tracking algorithm splitting structural variants upstream or downstream of *bla*$_{NDM}$ gene. (A) A pairwise BLAST search is performed on all NDM-positive contigs. Starting from *bla*$_{NDM}$ and continuing downstream or upstream, the inspected region is gradually increased using the 'splitting threshold'. (B) At each step, a graph is constructed connecting contigs (nodes) that share a BLAST hit with a minimum length as given by the 'splitting threshold'. Contigs which have the same structural variant at the certain position of the threshold belong to the same graph component, while the short contigs are singled out. (C) The splitting is visualized as a tree where branch lengths are scaled to match the position within the sequence, and the thickness and the colour intensity of the branches correspond to the number of sequences carrying the homology.

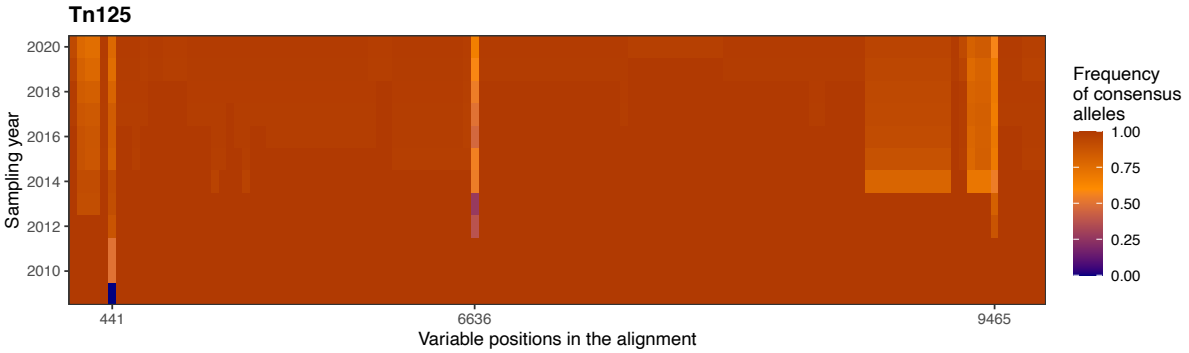
4.2.4. Analysis of overhanging reads mapping to short contigs

To investigate the reasons behind a number of distinctively short *bla*$_{NDM}$-carrying contigs, I mapped 781 raw Illumina paired-end sequencing reads (originally downloaded from SRA) back to their matching contigs. The mapping was done using BBMap (v38.59; Bushnell, 2014; *maxindel = 0* and *minratio = 0.2* settings). Within the output SAM file, only the overhanging reads with the CIGAR string matching the "*[0-9]*M[0-9]*S"* regular expression were selected. All overhangs of reads ≥50bp were screened against ISFinder database (Siguier et al., 2006).

4.2.5. Molecular tip-dating analysis.

The 112 complete Tn*125* and 73 complete Tn*3000* contigs with a known collection date and harbouring *bla*$_{NDM}$ were sequentially aligned (--pileup flag) using Clustal Omega (v1.2.3; Sievers et al., 2011) specifying the *bla*$_{NDM-1}$ sequence (FN396876.1) as a profile. Each alignment was manually inspected using UGENE (v38.0; Okonechnikov et al., 2012). The ancestral (i.e., root) sequence was determined by evaluating SNP frequencies over time (Figure 4.3 and Figure 4.4). Due to a short sampling time span and the relatively low number of mutations present, it is unlikely that any one non-ancestral SNP has become dominant in the population. Therefore, the ancestral sequence is expected to be reflected by SNPs found at high frequency in earlier years.

Under this assumption I find that, in all but two cases, the consensus sequence of an alignment displaying this behaviour. The first exception is the consensus sequence allele of Tn*125* at the variable position 441 (Figure 4.3). This allele was observed at low frequency in 2009. However, inspection of the allele frequency table for samples collected through time, one can see the low frequency is based on a single observation. Leaving out this early sample restores the desired frequency pattern; hence the consensus allele is considered ancestral in this case. The second exception is the variable position 449 in the Tn*3000* alignment (Figure 4.4). The consensus allele 'a' is not found in the early sample from 2009. Both allele 't', present in the early sample, and allele 'a' were found equally frequent in more recent samples. Thus, due to the lack of other evidence, allele 't' was considered ancestral. The determined ancestral sequences were used to evaluate temporal signal in different alignments, and in the subsequent rooting of phylogenetic trees.



Position: 441 (consensus: a)

year	a	t	g	c	–	n_samples
2009	0	0	0	1	0	1
2010	0.5	0	0	0.5	0	2
2011	0.5	0	0	0.5	0	2
2012	0.88	0	0	0.12	0	8
2013	0.91	0	0	0.09	0	11
2014	0.92	0	0	0.08	0	24
2015	0.82	0	0	0.18	0	45
2016	0.84	0	0	0.16	0	61
2017	0.84	0	0	0.16	0	68
2018	0.8	0	0	0.2	0	74
2019	0.76	0	0	0.24	0	86
2020	0.78	0	0	0.22	0	108

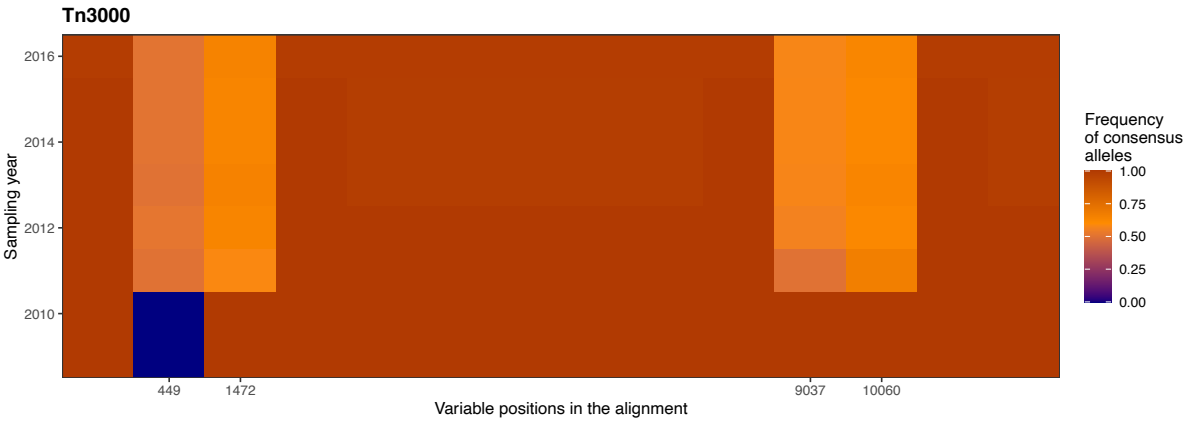
Position: 6636 (consensus: g)

year	a	t	g	c	–	n_samples
2009	0	0	1	0	0	1
2010	0	0	1	0	0	2
2011	0	0	1	0	0	2
2012	0.62	0	0.38	0	0	8
2013	0.73	0	0.27	0	0	11
2014	0.46	0	0.54	0	0	24
2015	0.44	0	0.56	0	0	45
2016	0.54	0	0.46	0	0	61
2017	0.5	0	0.5	0	0	68
2018	0.46	0	0.54	0	0	74
2019	0.42	0	0.58	0	0	86
2020	0.33	0	0.67	0	0	108

Position: 9465 (consensus: c)

year	a	t	g	c	–	n_samples
2009	0	0	0	1	0	1
2010	0	0	0	1	0	2
2011	0	0	0	1	0	2
2012	0.12	0	0	0.88	0	8
2013	0.18	0	0	0.82	0	11
2014	0.46	0	0	0.54	0	24
2015	0.31	0	0	0.69	0	45
2016	0.3	0	0	0.7	0	61
2017	0.32	0	0	0.68	0	68
2018	0.31	0	0	0.69	0	74
2019	0.31	0	0	0.69	0	86
2020	0.43	0	0	0.57	0	108

Figure 4.3. Temporal patterns across variable positions in the alignment of the *Tn*125 transposon. Heatmap providing the frequency of consensus sequence alleles in the *Tn*125 alignment over time. Allele frequency tables of more variable positions are given below the heatmap.



Position: 449 (consensus: a)

year	a	t	g	c	–	n_samples
2009	0	1	0	0	0	1
2010	0	1	0	0	0	1
2011	0.5	0.5	0	0	0	12
2012	0.52	0.48	0	0	0	66
2013	0.5	0.5	0	0	0	68
2014	0.51	0.49	0	0	0	71
2015	0.51	0.49	0	0	0	71
2016	0.51	0.49	0	0	0	73

Position: 1472 (consensus: a)

year	a	t	g	c	–	n_samples
2009	1	0	0	0	0	1
2010	1	0	0	0	0	1
2011	0.58	0.42	0	0	0	12
2012	0.64	0.36	0	0	0	66
2013	0.65	0.35	0	0	0	68
2014	0.63	0.37	0	0	0	71
2015	0.63	0.37	0	0	0	71
2016	0.64	0.36	0	0	0	73

Position: 9037 (consensus: t)

year	a	t	g	c	–	n_samples
2009	0	1	0	0	0	1
2010	0	1	0	0	0	1
2011	0.5	0.5	0	0	0	12
2012	0.44	0.56	0	0	0	66
2013	0.43	0.57	0	0	0	68
2014	0.42	0.58	0	0	0	71
2015	0.42	0.58	0	0	0	71
2016	0.42	0.58	0	0	0	73

Position: 10060 (consensus: a)

year	a	t	g	c	–	n_samples
2009	1	0	0	0	0	1
2010	1	0	0	0	0	1
2011	0.67	0.33	0	0	0	12
2012	0.62	0.38	0	0	0	66
2013	0.63	0.37	0	0	0	68
2014	0.62	0.38	0	0	0	71
2015	0.62	0.38	0	0	0	71
2016	0.63	0.37	0	0	0	73

Figure 4.4. Temporal patterns across variable positions in the alignment of the *Tn*3000 transposon. Heatmap providing the frequency of consensus sequence alleles in the *Tn*3000 alignment over time. Allele frequency tables of more variable positions are given below the heatmap.

Date randomization (10,000 iterations) and linear regression analyses were employed to estimate the presence of temporal signal in the alignment (Duchene et al., 2020; Rambaut et al., 2016; Rieux & Balloux, 2016) (Figure 4.5 and Figure 4.6). Tn*125* and Tn*3000* exhibited significant temporal signal using a simple regression approach (p=0.0356 and p=0.0456 respectively) and following date randomization (true evolutionary rate quantiles >0.95).

Bayesian based molecular dating was subsequently implemented in BEAST2 (v2.6.0; Bouckaert et al., 2019) and BactDating (Didelot et al., 2018) to infer the date of the emergence of the two transposons. Both BEAST2 and BactDating were run specifying a strict prior on the molecular clock. For BEAST2, the generalised time reversible (GTR) substitution model prior was used together with three population size models: Coalescent Constant population, Coalescent Exponential population, and Coalescent Bayesian Skyline. In addition, all BEAST2 and BactDating runs were supplied with a maximum likelihood (ML) phylogenetic tree (starting tree prior) constructed from both transposon alignments using RAxML (v8.2.12; Stamatakis, 2014) with specified GTRCAT substitution model and rooted using the inferred ancestral sequences. The chosen MCMC chain lengths for BactDating and BEAST2 runs were 10^7 and $1.5x10^9$ respectively to ensure convergence. I evaluated effective sample sizes (ESS) of the posterior distributions using *effectiveSize* function implemented in *coda* (Plummer et al., 2006) R package after discarding the first 20% of burn-in (Figure A.10 and Figure A.11). All BEAST2 and BactDating runs successfully converged with ESS of the posteriors close to or above 200. An example of the BEAST2 configuration used is available in xml format in Appendix B.

4.2.6. Estimating Shannon entropy among NDM-positive contigs

I estimated Shannon entropy ('diversity') for several categorizations of *bla*$_{\text{NDM}}$-containing contigs: country of sampling, bacterial host genera, plasmid backbones (determined by mapping to plasmid reference sequences), and ISs flanking the *bla*$_{\text{NDM}}$ gene. To estimate entropy of the population and to provide confidence intervals around the estimates, I use bootstrapping with replacement (1,000 iterations). At each iteration, entropy was calculated for a sampled set of contigs (X) classified into n unique categories according to the following formula:

$$H(\boldsymbol{X}) = -\sum_{i=1}^{n} P(x_i) log\, P(x_i), \quad (4.1)$$

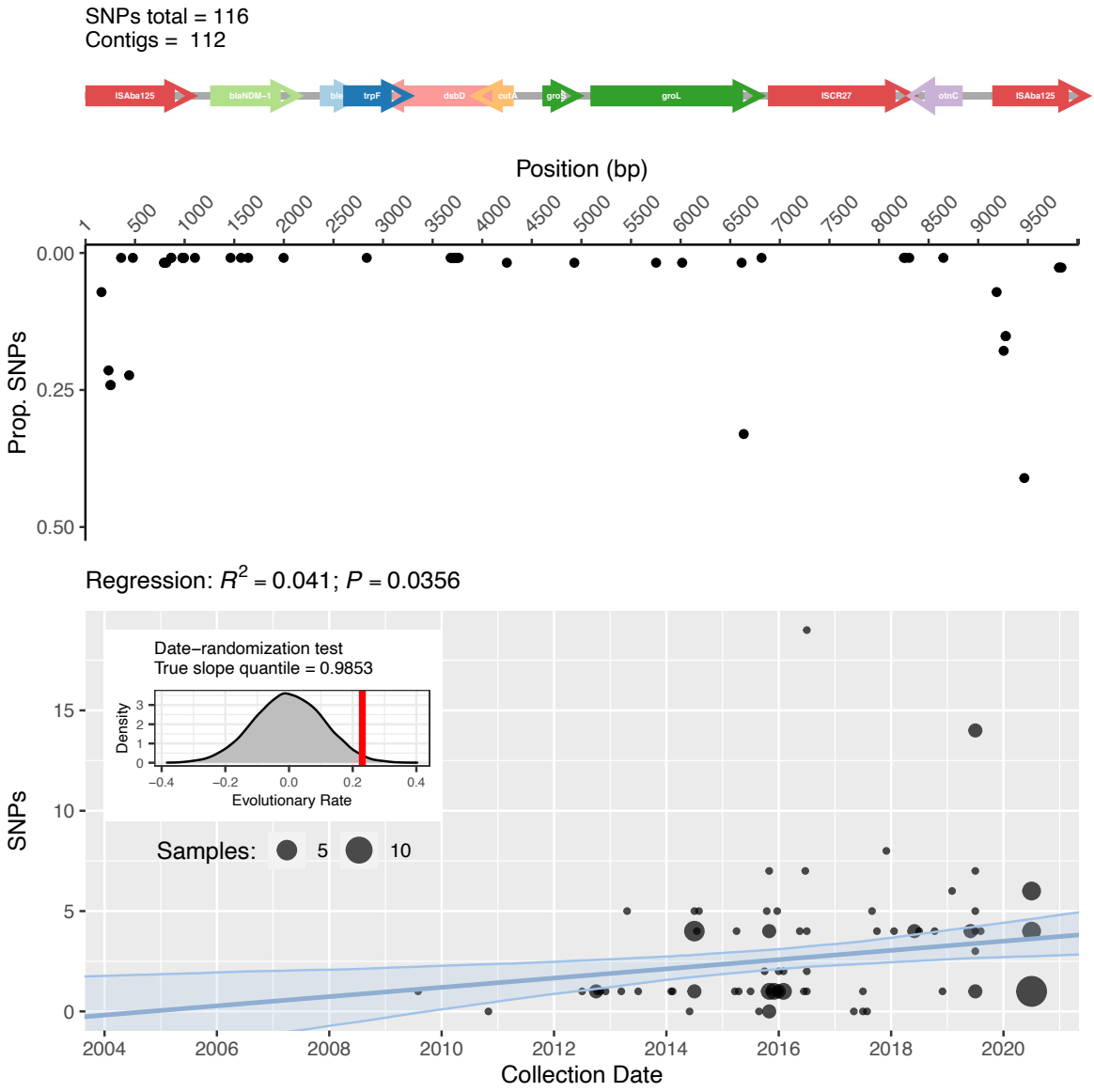


Figure 4.5. Assessment of temporal signal in the alignment of *Tn*125. Starting from the top, the figure contains the following plots: schematic representation of the aligned transposon sequences; the SNP frequency across the alignment evaluated against the inferred ancestral sequence; the linear regression analysis of number of SNPs accumulated against the year of sample collection. The ribbon surrounding the regression line provides a 95% confidence interval given by the bootstrapping the regression analysis (1,000 iterations). The inset of the regression plot shows the results of the date-randomization test. The marginal distribution of the inset indicates the regression line slope values (i.e., evolutionary rates) after 10,000 date randomizations and the red vertical line indicates the true slope value.

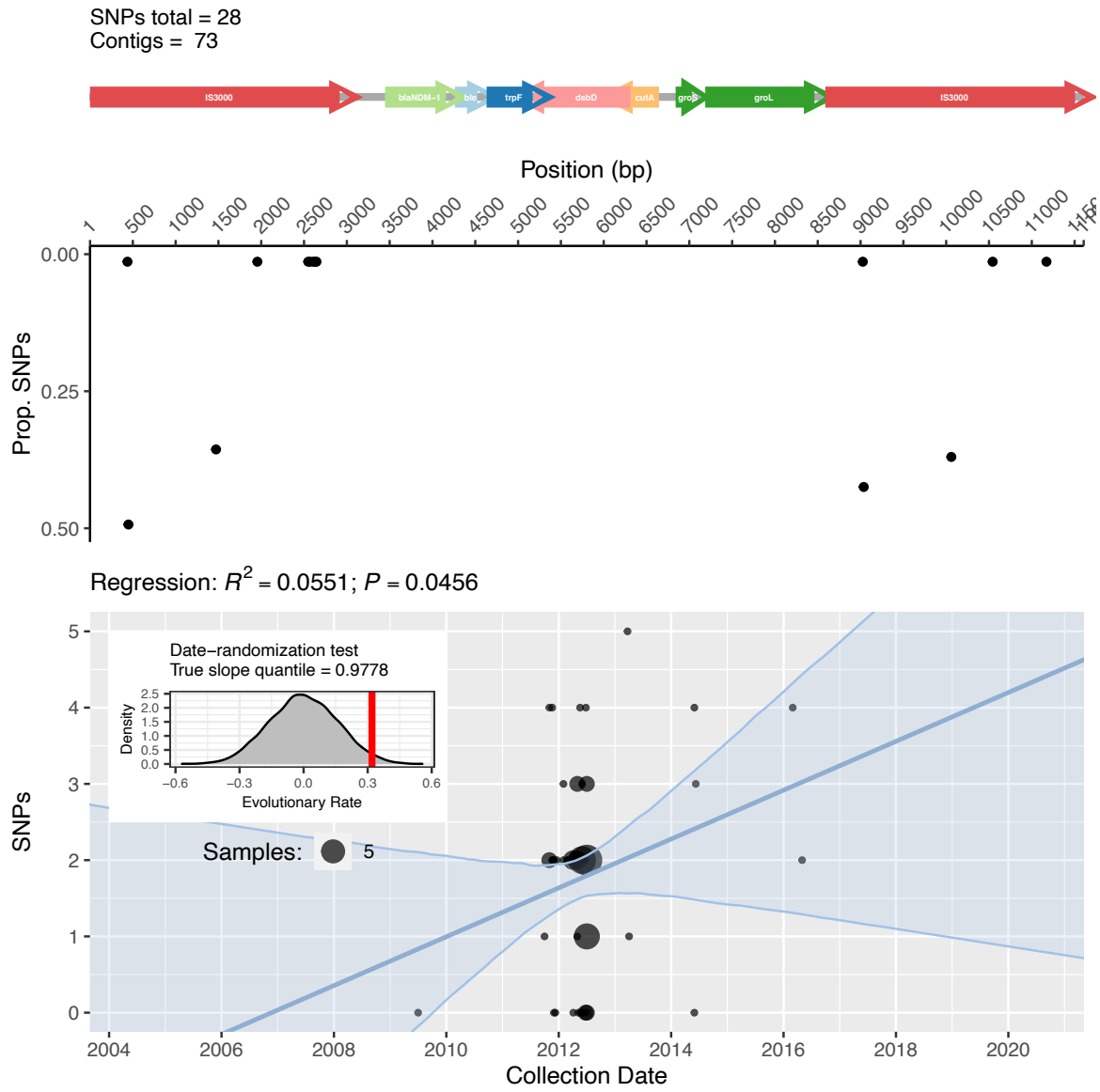


Figure 4.6. Assessment of temporal signal in the alignment of *Tn*3000. Starting from the top, the figure contains the following plots: schematic representation of the aligned transposon sequences; the SNP frequency across the alignment evaluated against the inferred ancestral sequence; the linear regression analysis of number of SNPs accumulated against the year of sample collection. The ribbon surrounding the regression line provides a 95% confidence interval given by the bootstrapping the regression analysis (1,000 iterations). The inset of the regression plot shows the results of the date-randomization test. The marginal distribution of the inset indicates the regression line slope values (i.e., evolutionary rates) after 10,000 date randomizations and the red vertical line indicates the true slope value.

where the probability $P(x_i)$ of any sample belonging to any particular category x_i (e.g., country or plasmid backbone) is approximated using the category's frequency. Accordingly, higher entropy values indicate an abundance of equally likely categories, while lower entropy indicates a limited number of categories.

4.2.7. Estimating correlation between genetic and geographic distance

Geographic distance between pairs of samples was determined using their sampling coordinates and the *geodist* (Padgham & Sumner, 2020) R package. Exact Jaccard distance (JD) was used as a measure of the genetic distance calculated using the tool Bindash (v0.2.1; X. Zhao, 2019) with *k*-mer size equal to 21 bp. The JD is defined as the fraction of total *k*-mers not shared between two contigs. JD between all pairs of contigs was first calculated on a 1,000 bp stretch of DNA downstream of bla_{NDM} start codon continuing with a 500 bp increments. At each increment, the two distance matrices (genetic and geographic) were assessed using the *mantel* function (Spearman correlation and 99 permutations) from the *vegan* (Oksanen et al., 2019) package in R. The correlation between genetic and geographic distance, was plotted as a function of distance from bla_{NDM} gene (Figure 4.19).

4.3. Results

4.3.1. A global dataset of bla_{NDM} carriers

In this chapter, I compiled a dataset of 6,155 bacterial genomes (7,148 contigs) carrying at least one copy of $bla_{\text{NDM.}}$ This included published assemblies from NCBI RefSeq (O'Leary et al., 2016) (*n*=2,632), NCBI GenBank (Benson et al., 2013) (*n*=1,158) and Enterobase (Zhou et al., 2020) (*n*=1,379); bacterial genomes assembled using short read *de novo* assembly from NCBI's Sequence Read Archive (SRA) (*n*=882); and newly generated bacterial genomes isolated from 79 hospitalized patients across China and 25 livestock farms assembled using hybrid PacBio-Illumina *de novo* assembly (*n*=104) (Table 4.1, Table A.2, Figure 4.1 and Figure 4.7). While public genomes have inherent sampling biases, using them is the most comprehensive approach available (W. Wu et al., 2019). Taken together the dataset included genetic data from 251 different Bioprojects, though with more than half the samples linked to two large-scale database refinement efforts (Souvorov et al., 2018; Tatusova et al., 2014).

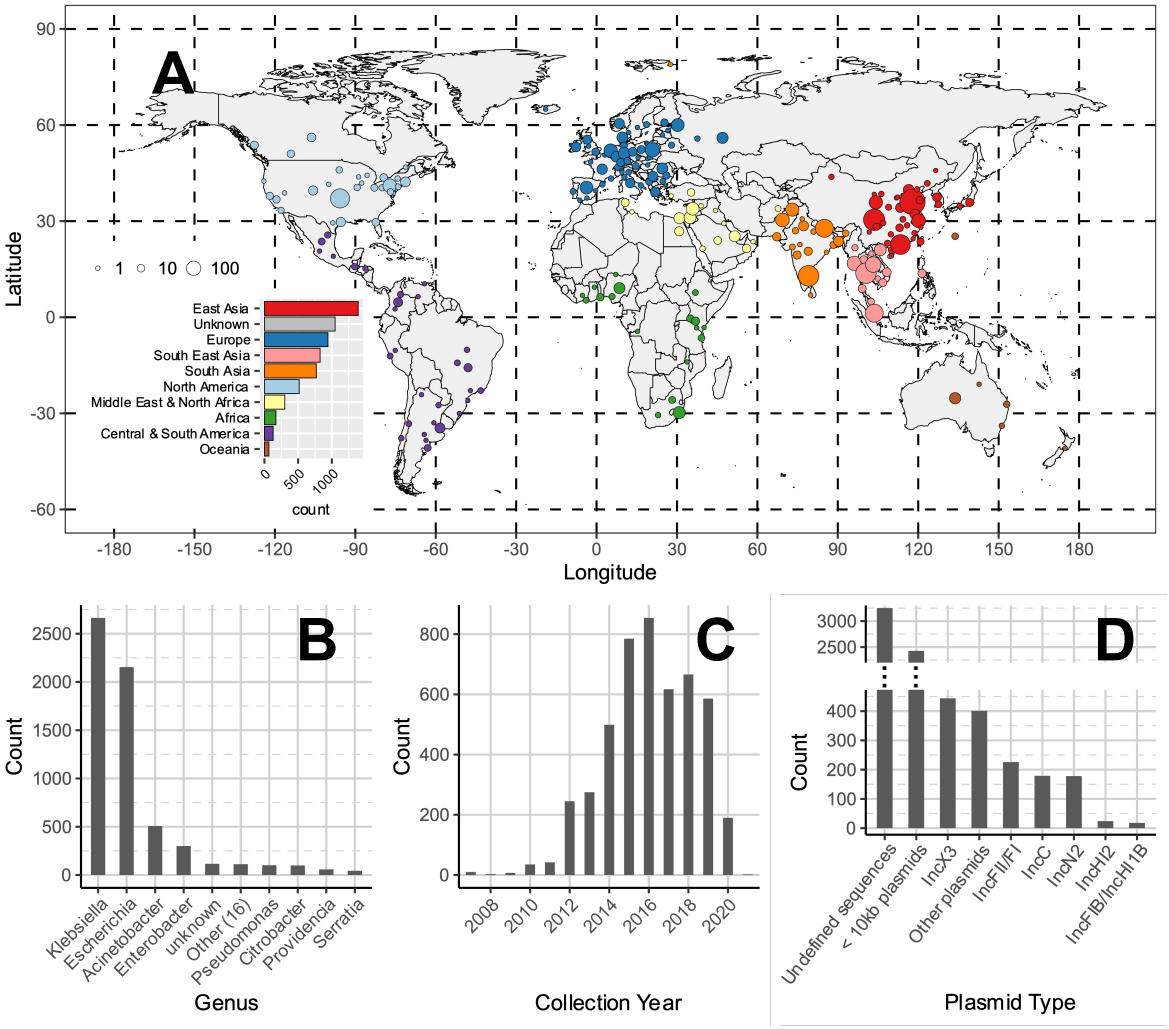


Figure 4.7. Composition of the global dataset of 6,155 NDM-positive samples.(**A**) Geographic distribution of *bla*$_{NDM}$-positive assemblies. Points are coloured by geographic region and their size reflects the number of samples they encompass. **(B)** Distribution of host bacterial genera of NDM-positive samples. **(C)** Distribution of sample collection years. **(D)** Distribution of contigs according to the plasmid backbone.

Table 4.1. NDM-positive samples (and NDM-positive contigs) stratified by where the data was sourced and the associated sequencing platform.

		Sequencing Platform			
		ILLUMINA	ION_TORRENT	PACBIO_SMRT & ILLUMINA (Hybrid assembly)	Unknown
Source/Database	China Hospitals	0	0	104 (105)	0
	Enterobase	185 (185)	0	0	1194 (1194)
	GenBank	0	0	0	1158 (2117)
	RefSeq	0	0	0	2632 (2665)
	SRA	872 (872)	10 (10)	0	0

The dataset included *bla*$_{NDM}$-positive isolates from 88 countries (Figure 4.7A) mostly collected in Asia, particularly mainland China (*n*=1,270), European countries (941), USA (461), Thailand (419) and India (361). At least 27 bacterial genera were represented, with a large fraction of *Klebsiella* and *Escherichia* isolates (2,664 and 2,154 genomes respectively; Figure 4.7B). Collection dates were recorded for 4,816 samples (78.25%). Of these, the majority were collected between 2014-2019 (71.05%, Figure 4.7C). The dataset also includes 55 genomes collected in 2010 or earlier. These include the *K. pneumoniae* isolate from 2008 Sweden in which *bla*$_{NDM}$ was first characterized (Yong et al., 2009); one *Enterobacter hormaechei* isolate from 2008 India (Chavda et al., 2016); one *S. enterica* isolate from 2008 London, UK (Ashton et al., 2016); one *A. baumannii* isolate from an individual of Balkan origin collected in Germany in 2007 (Bonnin et al., 2012; Sahl et al., 2015); and nine assembled *E. coli* genomes from urine samples collected in Greece in 2007.

The dataset contained 17 known variants of NDM. NDM-1 was the most abundant (*n*=4,127; Figure 4.8A) with NDM-5 (*n*=2,394) increasing in prevalence after 2012 (Figure 4.8B and C). Variants showed different associations with plasmid types (Figure 4.8D) and genera (Figure 4.8E) but were fairly evenly distributed across the world except for *bla*$_{NDM-4}$-carrying isolates largely collected in Southeast Asia and *bla*$_{NDM-9}$ predominantly found in East Asia (Figure 4.8F).

4.3.2. Plasmid backbones carrying *bla*$_{NDM}$

I identified 33 different replicon types on 1,222 contigs using PlasmidFinder (Carattoli et al., 2014) (Figure 4.7). The most prevalent replicon type was IncX3 (444 contigs), and the most abundant types within the dataset exhibited clear geographic structure (Figure 4.9). To further identify uncharacterised plasmid types, we mapped 3,599 contigs to a set of complete plasmid reference sequences after discarding short contigs. This revealed 181 clusters of similar putative plasmid sequences (Figure 4.10). Most clusters (*n*=105) grouped contigs of the same replicon type and contained a small number of contigs (only 27 clusters included >10 contigs), in line with a diverse and dynamic population of plasmid backbones for *bla*$_{NDM}$.

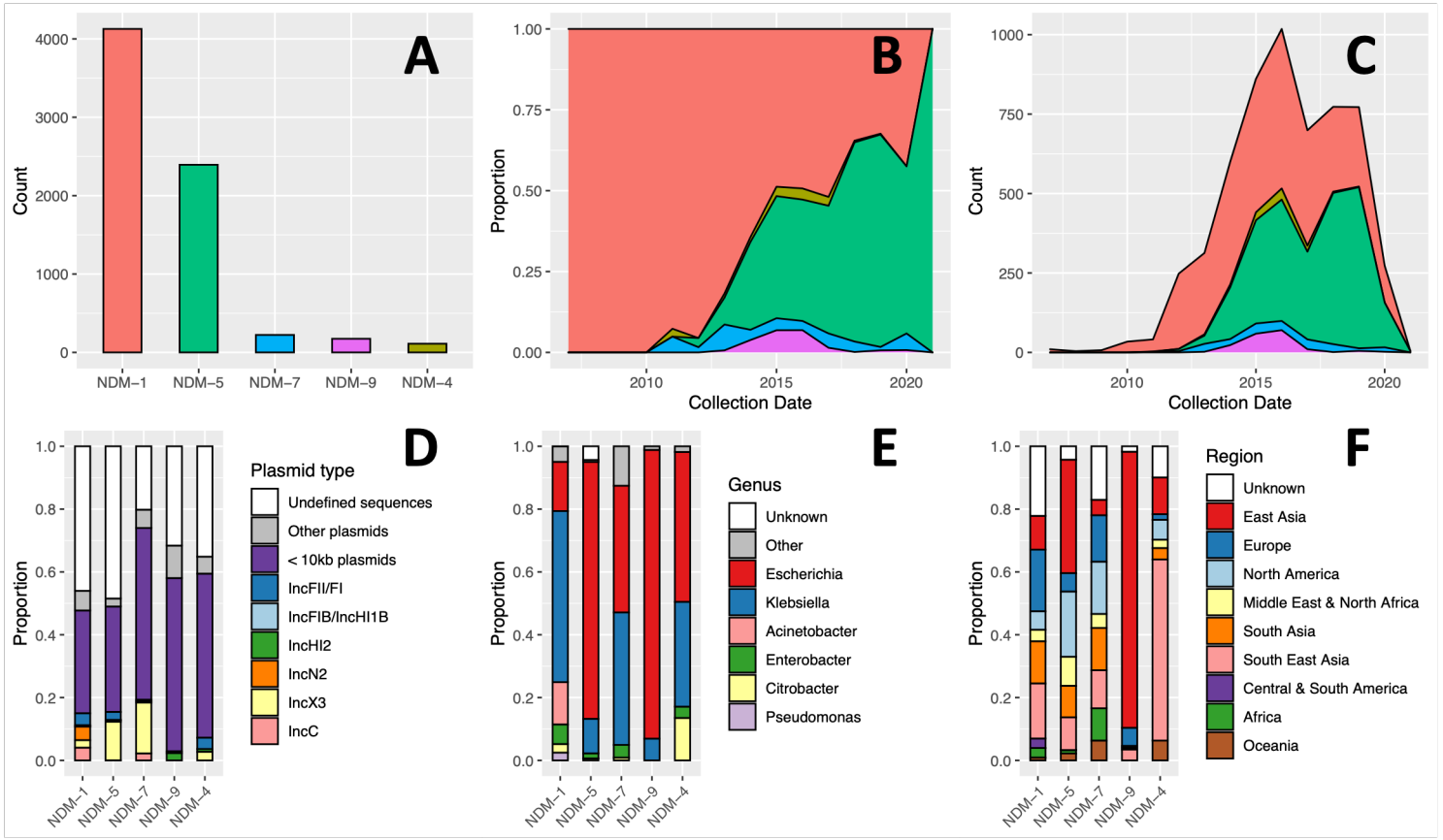


Figure 4.8. Global prevalence and genetic context of NDM variants. Panel (**A**) shows overall counts of the five most frequent NDM variants (i.e., >25 representatives) while panels (**B**) and (**C**) show their prevalence over time. Panels (**D**), (**E**) and (**F**) have different colour coding as indicated by the legends and show bar plots of proportions of plasmid backbones, bacterial genera, and sampling location respectively for each NDM variant.

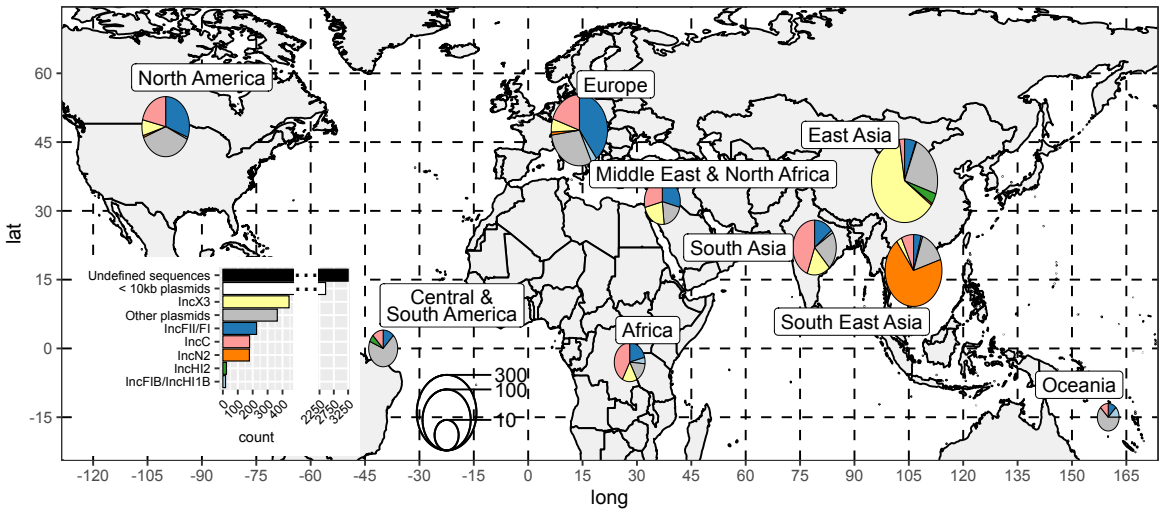


Figure 4.9. Global distribution of plasmid backbones of NDM-positive contigs. All NDM-positive contigs with an identified replicon type were grouped into several broader groups of plasmids. These have been pooled according to the geographical region and represented on the world map using pie charts. The sizes of the pie charts are log-scaled to aid interpretability. The figure inset serves as a legend and indicates counts of plasmid backbones in the dataset. The five largest geographical regions count at minimum 30 different Bioprojects each with majority of samples originating from PRJNA224116, a Refseq Prokaryotic Genome Annotation Project.

The majority (n=2,427; 68.4%) of *bla*$_{NDM}$-carrying contigs were associated with small putative plasmids (<10 kb; Figure 4.10). While this could suggest small plasmids play a key role as *bla*$_{NDM}$ carriers, this pattern could also result from consistently fragmented *de novo* assemblies due to duplicated ISs and transposons. Consistent with this latter hypothesis, 610 contigs mapped to pKP-YQ12450 which is likely a 7.8 kb fragment of a larger plasmid (Yang et al., 2015). Conversely, Roach et al. provide evidence that other small *bla*$_{NDM}$-carrying plasmids (Peruvian pKP-NDM-1_isoforms 1-5) are inherited by descent and are result of transposon-mediated plasmid fusion (Roach et al., 2020).

4.3.3. Resolving structural variants in the *bla*$_{NDM}$ flanking regions

To go beyond a static reference-based view of variation around *bla*$_{NDM}$ and gain a detailed overview of the possible events in its evolution, I developed an alignment-based approach to progressively resolve genomic variation moving upstream or downstream

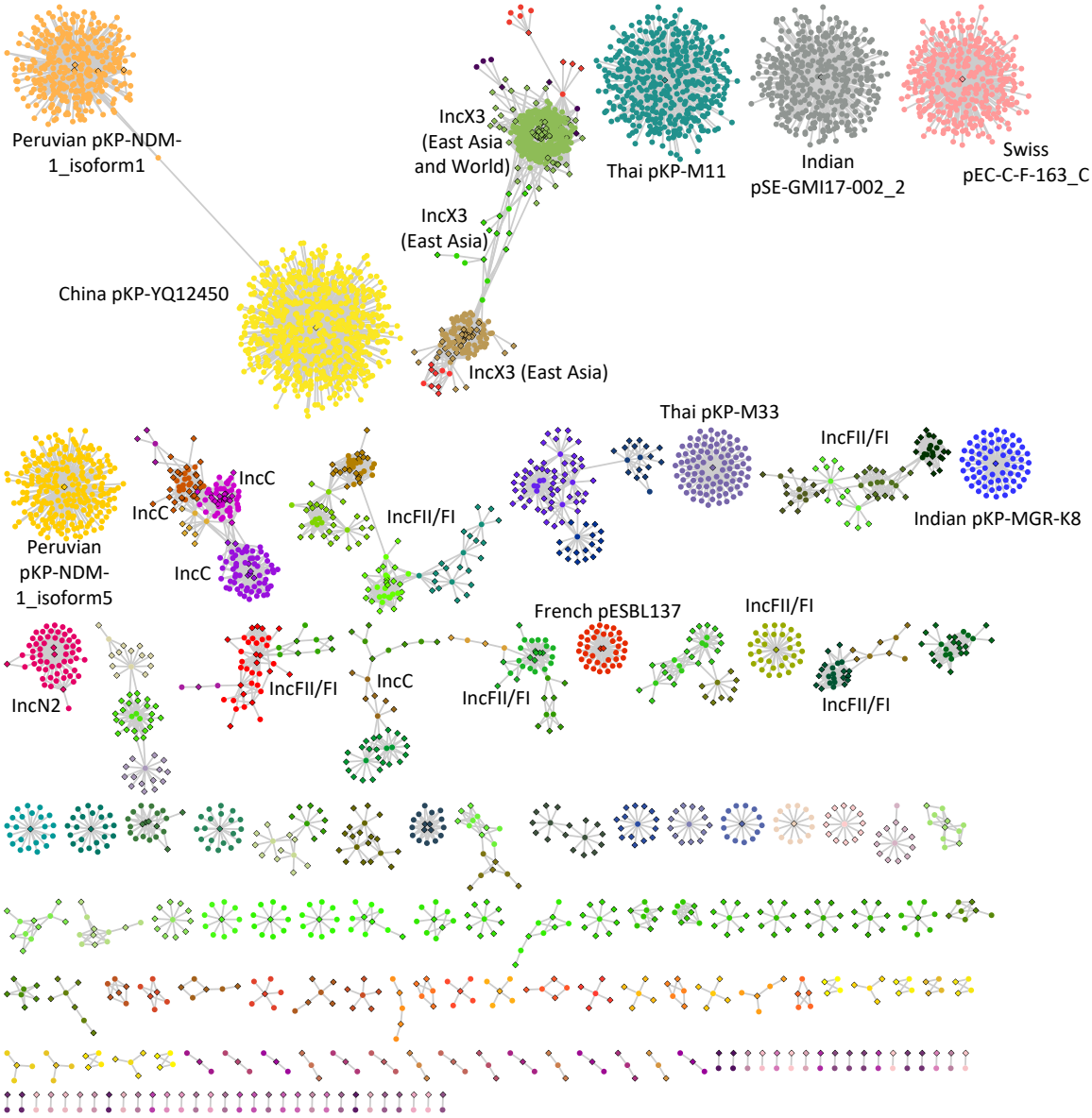


Figure 4.10. A network of *bla*$_{NDM}$-carrying contigs (circles) mapping to the bacterial plasmid reference sequences (diamonds). The network is visualized using Cytoscape and coloured according to communities identified by the Infomap algorithm. The largest communities are annotated according to the predominant plasmid type or the reference plasmid. Plasmids of <10 kb in length include China pKP-YQ12450 (*n*=610 contigs), Thai pKP-M11 (*n*=399), Indian pSE-GMI17-002_2 (*n*=354), Swiss pEC-C-F-163_C (*n*=324), Peruvian pKP-NDM-1_isoform1-4 (*n*=318), Peruvian pKP-NDM-1_isoform5 (*n*=226), Thai pKP-M33 (*n*=91), Indian pKP-MGR-K8 (*n*=66), and 39 other <10 kb putative plasmids.

from the gene. The method is described in detail in Section 4.2.3 of this chapter (Figure 4.2). In brief, a pairwise discontiguous Mega BLAST search (Camacho et al., 2009; Ma et al., 2002) is applied to all *bla*$_{NDM}$-carrying contigs to identify all possible homologous regions between each contig pair. Only BLAST hits covering the complete *bla*$_{NDM}$ gene are retained (Figure 4.2A). Next, starting from *bla*$_{NDM}$, a gradually increasing 'splitting threshold' is used to monitor structural variants as they appeared upstream or downstream of the gene. At each step, a network of contigs (nodes) that share a BLAST hit with a minimum length as given by the 'splitting threshold' is assessed (Figure 4.2B). As we move upstream or downstream and further away from the gene, the network starts to split into smaller clusters, each carrying contigs that share an uninterrupted stretch of homologous DNA, which can be represented as a tree (Figure 4.2C). This approach treats the upstream and downstream flanking regions separately rather than simultaneously and is agnostic to whether splitting into 'sequence clusters' is caused by structural variants of the same genomic background or different genomic backgrounds.

Upstream of *bla*$_{NDM}$, >98% of sufficiently long contigs included a ~75 bp fraction of IS*Aba125* (Figure 4.11 andFigure 4.12), supporting Tn*125* as an ancestral transposon of the *bla*$_{NDM}$ gene in agreement with previous work (Poirel, Bonnin, et al., 2011; Poirel, Dortet, et al., 2011; Toleman et al., 2012; W. Wu et al., 2019). However, the homology of the region upstream of *bla*$_{NDM}$ falls quickly: within a few hundred base pairs of the *bla*$_{NDM}$ start codon the region splits into multiple structural variants, none of which dominate the considered pool of contigs (Figure 4.12). I identified 141 different structural variants within 1,200 bp upstream of *bla*$_{NDM}$. The upstream region contained a high number of ISs (e.g., IS*Aba125* [*n*=243], IS*5* [*n*=426], IS*3000* [*n*=60], IS*Kpn14* [*n*=55], and IS*Ec33* [*n*=147]). This transposition hotspot probably contributes to fragmented assemblies: 2,269 contigs were excluded from further analysis for being too short (Figure 4.12).

The downstream flanking region exhibits more gradual structural diversification than the upstream region, with one dominant putative ancestral background (Figure 4.13). As illustrated by the stem of the tree of structural variants, many of the 7,014 contigs analysed contained complete sequences of the same set of genes: *ble* (6,863 contigs), *trpF* (6,038), *dsbD* (5,551), *cutA* (2,731), *groS* (2,175), *groL* (1,631). When restricted to *bla*$_{NDM}$-positive contigs of sufficient length to possibly harbour the full repertoire of these genes (*n*=3,786), almost half carry all of them (*n*=1,631; 43.1%). In addition, dominant

structural variants were found associated with various source databases and sequence lengths hence diminishing the impact of the sampling bias (Figure A.15).

4.3.4. Early events in the spread of *bla*$_{NDM}$

While I did not observe any strong overall signal in the distribution of associated plasmid backbones, bacterial genera, or sampling locations, closer examination of mobilome features common to sufficiently large numbers of isolates indicated early events in the spread of *bla*$_{NDM}$. The putative ancestral Tn*125* background, with an uninterrupted downstream IS*Aba125* element, was seen in contigs mainly from *Acinetobacter* and *Klebsiella* (Figure 4.13 top). Conversely, the diversity of bacterial genera carrying IS*Aba125* upstream is more substantial (Figure 4.12 top). Only 203 contigs carried a complete IS*Aba125* downstream of *bla*$_{NDM}$, of which 147 carried an IS*Aba125* sequence in proximity (<9 kb) to the *bla*$_{NDM}$ start codon. These account for a minority (7%; 147/2097) of isolates when sufficiently long contigs are considered. This supports the initial dissemination of *bla*$_{NDM}$ by Tn*125* to other plasmid backbones predominately being mediated by *Acinetobacter* and *Klebsiella*, after which the transposon was disrupted by other rearrangements.

IS*3000*, both upstream and downstream, was almost exclusively associated with samples from *Klebsiella* (Figure 4.12 and Figure 4.13). Thus, as suggested by Campos et al. (2015), Tn*3000* – a composite transposon made of two copies of IS*3000* – likely re-mobilized *bla*$_{NDM}$ following the 'fossilization' of Tn*125*; my findings suggest this secondary mobilization primarily happened in *Klebsiella* species. Tn*5403* was found extensively associated with IncN2 plasmids (Figure 4.13) which could have placed *bla*$_{NDM}$ in this background via cointegrate intermediate as previously suggested by Poirel et. al. (2011). Some elements of the mobilome were geographically linked e.g., IS*5* which was predominantly found upstream of *bla*$_{NDM}$ on IncX3 plasmids in *Escherichia* from East Asia (Figure 4.12). IS*5* is known to enhance transcription of nearby promoters in *E. coli* (Schnetz & Rak, 1992) and its abundance and positioning just upstream of *bla*$_{NDM}$ suggests a similar role in this case.

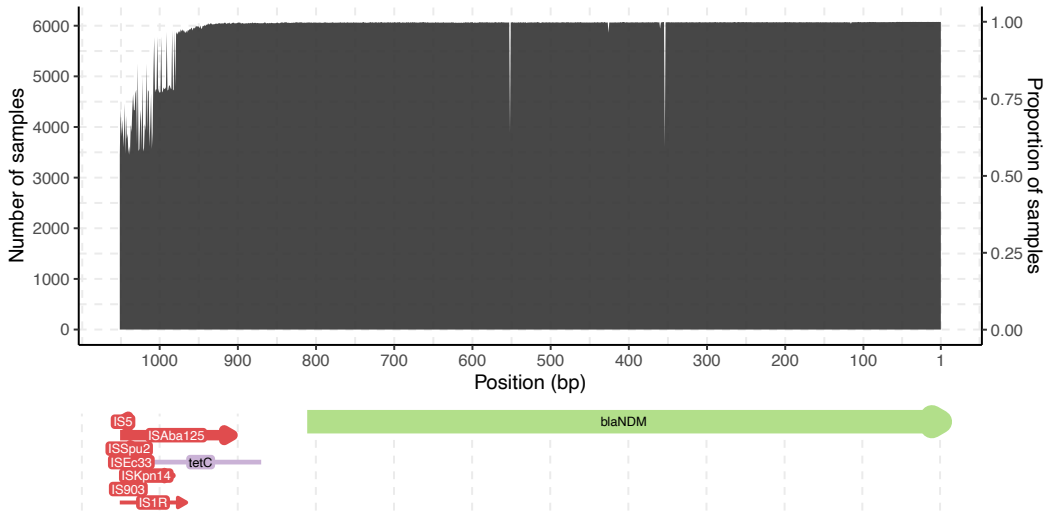
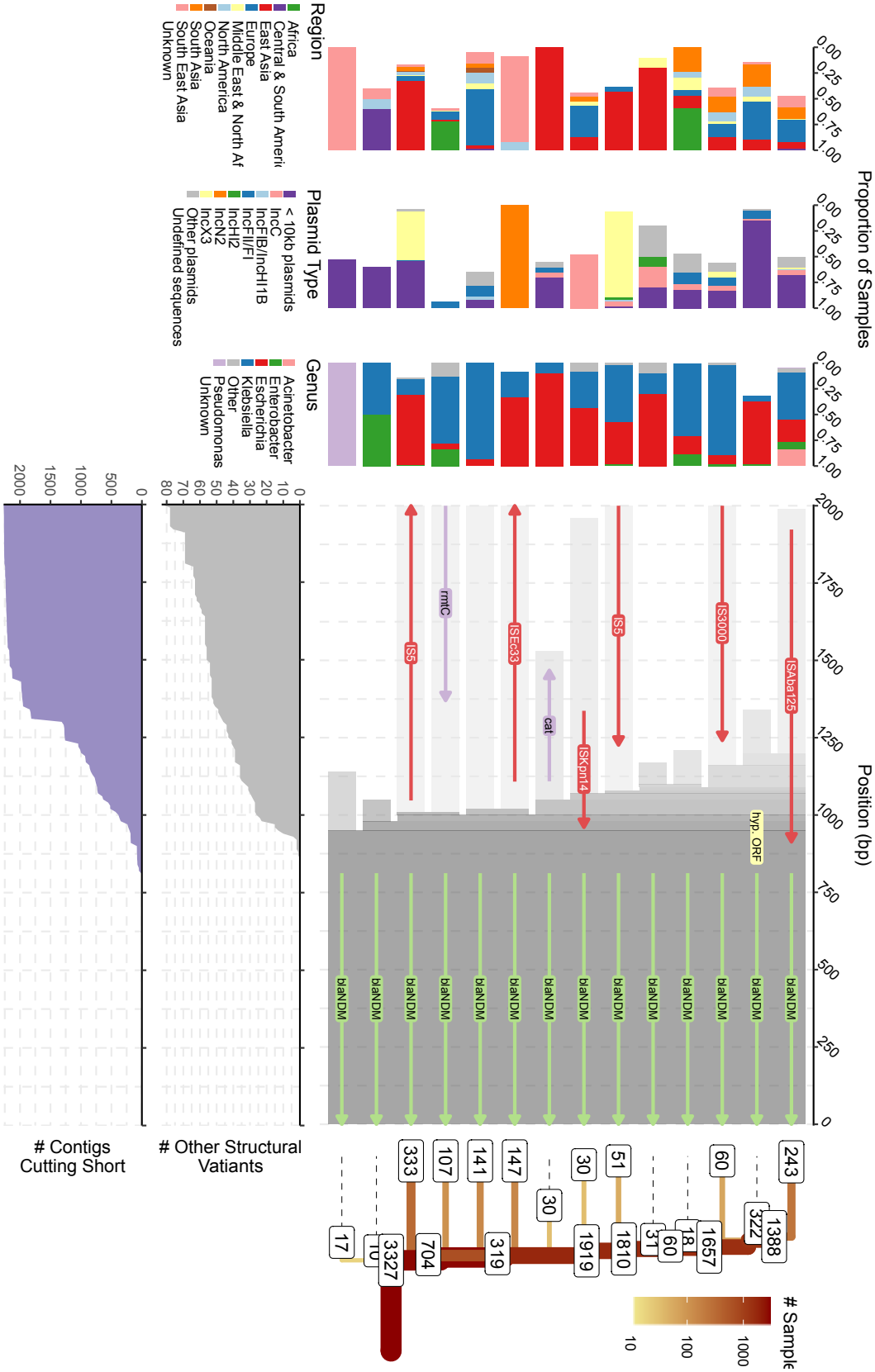


Figure 4.11. Alignment of 6,455 sufficiently long contigs 1,050 bp upstream of the *bla*$_{NDM}$ stop codon. The sequence annotations (produced by the Prokka-Roary annotation pipeline) of the aligned contigs are shown in the bottom panel. The thickness of the arrows reflects the number of contigs having a particular gene annotation. Red colouring indicates various transposable elements.

Figure 4.12. Splitting of structural variants upstream of *bla*$_{NDM}$. The 'splitting tree for the most common (i.e., ≥ 10 contigs) structural variants is shown on the right-hand side. The labels on the nodes indicate the number of contigs remaining on each branch. The other contigs either belong to other structural variants or were removed due to being too short in length. The number of contigs cutting short is indicated by the area chart at the bottom. Similarly, the number of less common structural variants is indicated by the upper area chart. The genome annotations provided by the Prokka and Roary pipelines of the most common structural variants are shown in the middle of the figure. The homologous regions across structural variants are indicated by the grey shading. Some of the structural variants and branches were intentionally cut short even though their contigs were of sufficient size. This was done to prevent excessive bifurcation and to make the tree easier to interpret. Branches with more than 75% of contigs lost due to variation and short length were truncated. The distribution of genera, plasmid types and geographical regions of samples that belong to a each of the common structural variant is shown on the left-hand side.


Region
Africa
Central & South Ameri
East Asia
Europe
Middle East & North Af
North America
Oceania
South Asia
South East Asia
Unknown
Plasmid Type
< 10kb plasmids
IncC
IncFIB/IncHI1B
IncFII/FI
IncHI2
IncN2
IncX3
Other_plasmids
Undefined sequences
Genus
Acinetobacter
Enterobacter
Escherichia
Klebsiella
Other
Pseudomonas
Unknown
Proportion of Samples
0.00
0.25
0.50
0.75
1.00
Position (bp)
0
250
500
750
1000
1250
1500
1750
2000
blaNDM
ISAba125
hyp. ORF
IS3000
IS5
ISEc33
ISKpn14
cat
rmtC
IS3
Sample
10
100
1000
243
1388
322
60
1657
18
60
31
51
1810
30
1919
30
147
319
141
107
704
333
17
10
3327
Other Structural Vatiants
Contigs Cutting Short

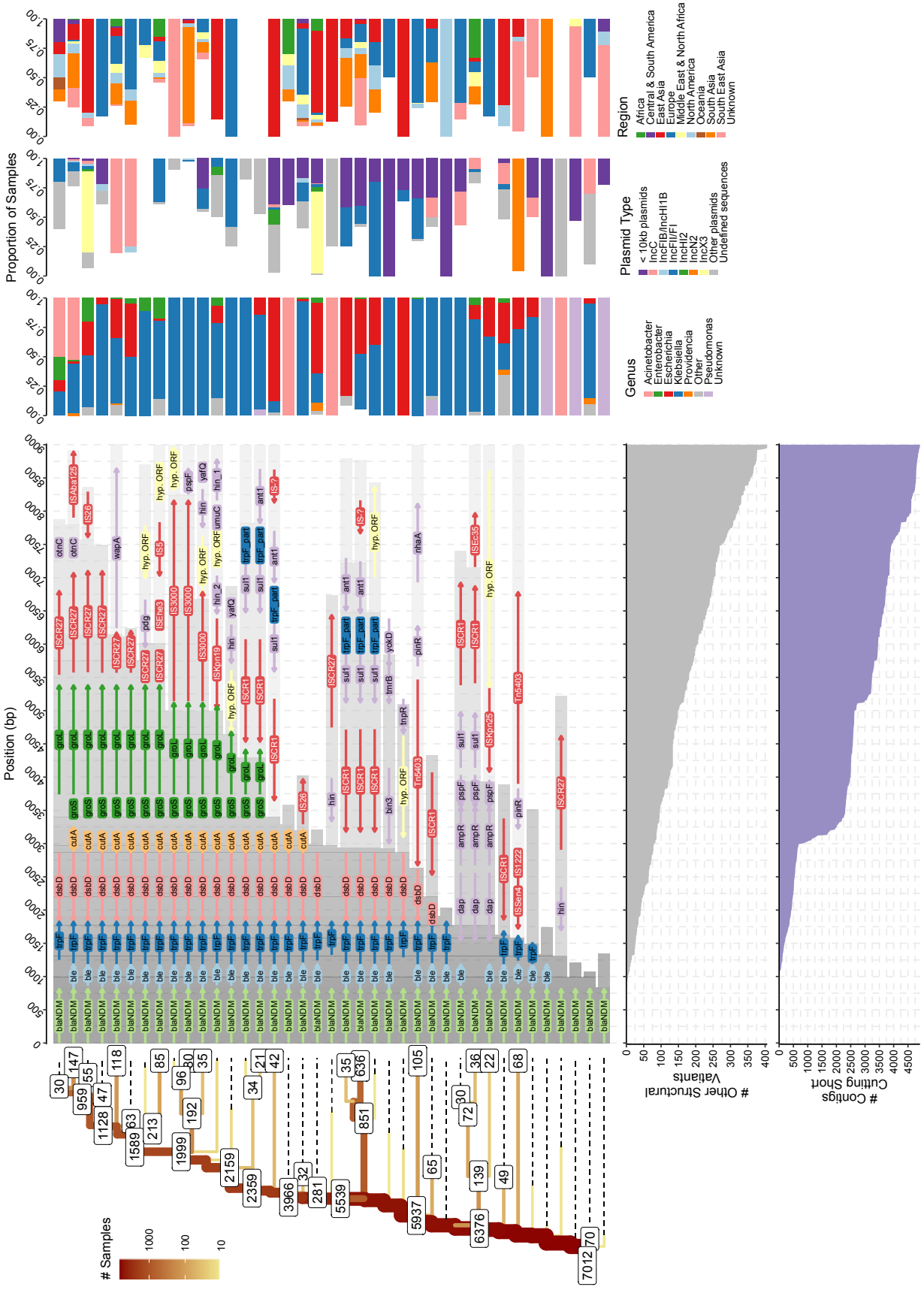

Proportion of Samples
Region
Africa
Central & South America
East Asia
Europe
Middle East & North Africa
North America
Oceania
South Asia
South East Asia
Unknown
Plasmid Type
< 10kb plasmids
IncC
IncFIB/IncHI1B
IncFII/FI
IncHI2
IncX3
Other plasmids
Undefined sequences
Genus
Acinetobacter
Enterobacter
Escherichia
Klebsiella
Providencia
Other
Pseudomonas
Unknown
Position (bp)
Samples
Other Structural Variants
Contigs Cutting Short

Figure 4.13. Splitting of structural variants downstream of *bla*$_{NDM}$. Like in figure 4.12., the 'splitting' tree for the most common (i.e., ≥10 contigs) structural variants is shown on the left-hand side with labels indicating the number of contigs remaining on each branch. Less common structural variants with <20 contigs are not shown in the tree. The number of less common structural variants is indicated by the upper area chart. Similarly, the number of contigs removed from the analysis due to short length is indicated by the area chart at the bottom. Genome annotations provided by the Prokka and Roary pipelines of the most common structural variants are shown in the middle of the figure. The homologous regions among structural variants are indicated by the grey shading. The stem of the tree of structural variants indicates a dominant genetic background of the *bla*$_{NDM}$ consisting of genes encoding: bleomycin resistance (*ble*), N-(5'-phosphoribosyl)anthranilate isomerase (*trpF*), thiol:disulfide interchange protein (*dsbD*), divalent-cation tolerance protein (*cutA*), co-chaperonin GroES (*groS*), chaperonin GroEL (*groL*), and ISCR27. Other less common genes associated with *bla*$_{NDM}$ encode for: Na(+)/H(+) antiporter (*nhaA*), serine recombinase (*pinR*), aminoglycoside N(3')-acetyltransferase-like protein (*yokD*), dihydropteroate synthase (*sul1*), tRNA(Ser)-specific nuclease (*wapA*), DNA-invertase (*hin*), spectinomycin resistance (*ant1*), putative transposon *Tn*552 DNA-invertase (*bin3*), tunicamycin resistance (*tmrB*), AmpC beta-lactamase regulator (*ampR*), succinyltransferase (*dap*), psp operon transcriptional activator (*pspF*) and ultraviolet N-glycosylase/AP lyase (*pdg*). The distribution of genera, plasmid backbones and geographical regions of samples that belong to each of the common structural variant is shown on the right-hand side.

One of the most commonly identified transposable elements in the downstream flanking region (~30% prevalence) was ISCR1 (IS91 family transposase) (Figure 4.13) always accompanied by *sul1* and occasionally in configuration with *ant1* or *pspF*, *ampR*, and *dap* genes. In some cases, a small and possibly fragmented putative IS (IS-?) is found further downstream. IS-? bears little similarity to known ISs and it is unclear what role it plays in the mobility of *bla*$_{NDM}$. ISCR1 is found at various positions downstream of *bla*$_{NDM}$ and often in *Escherichia* and *Klebsiella* species. In most cases, the orientation of ISCR1 should prevent this element from mobilizing *bla*$_{NDM}$ (Figure 4.13)(Ilyina, 2012). Nevertheless, the prevalence of this element could be due to the several AMR genes it can mobilize, such as *sul1* or *ampR*. ISCR1s are mainly found in complex class 1 integrons (Ilyina, 2012), however, not many annotated integrase genes are located within the vicinity of *bla*$_{NDM}$. In fact, only 15 contigs were found to have an integrase <50 kb away from *bla*$_{NDM}$ and none showed any consistency in integrase placement with respect to *bla*$_{NDM}$. This suggests integrases play a minor role in the dissemination of *bla*$_{NDM}$.

Another notable ISCR element is ISCR27 which is consistently found immediately downstream of the *groL* gene at high prevalence (33.1% of sufficiently long contigs; Figure 4.13). Contrary to its ISCR1 relative, ISCR27 is correctly oriented to mobilize bla_{NDM} as is presumed to have happened during the initial mobilization of the progenitor of bla_{NDM} (Toleman et al., 2012). However, I find no evidence of subsequent ISCR27 mobility. The origin of rolling-circle replication of ISCR27 (*oriIS*; GCGGTTGAAC TTCCTATACC) is located 236 bp downstream of the ISCR27 transposase stop codon. The region downstream of this stop codon in all structural variants bearing a complete ISCR27 is highly conserved for at least 750 bp (Figure 4.13).

4.3.5. Subsequent rearrangements dominated by IS*26*

Three sharp drops in the number of considered contigs at particular distances downstream of bla_{NDM} (Figure 4.13, e.g., region 3,000-3,300 bp) prompted investigation of these distinct cut-offs. I mapped 781 raw Illumina paired-end sequencing reads from the dataset back to their matching bla_{NDM} contigs. The read overhangs (≥50 bp) that mapped to the downstream end of the contigs were screened against the ISFinder database (Siguier et al., 2006). The ≥50 bp overhangs associated with 3,000-3,300 long flanks downstream of bla_{NDM} corresponding to the largest observed drop almost exclusively match the left inverted repeat (IRL) of the IS*26* sequence (Figure 4.14). Another hotspot, associated to IS*26* was found around 7,500 bp, while at around 7,800 bp a number of overhanging reads mapped to IS*Aba125*. These positions roughly match the third drop in the number of contigs observed 7500-8000 bp downstream of bla_{NDM}. No ISs were found to match the second drop in the number of contigs (5000-5250 bp).

IS*26*, although often found in two adjacent copies forming a seemingly composite transposon, is a so-called pseudo-composite (or pseudo-compound) transposon (Harmer et al., 2020). In contrast to composite transposons, a fraction of DNA flanked by the two IS*26* is mobilized either via cointegrate formation or in the form of a circular translocatable unit (TU), which consists of a single IS*26* element and a mobilized fraction of DNA, and inserts preferentially next to another IS*26* (Harmer et al., 2014, 2020). Taken together, the presented results, including Figure 4.14, suggest three possible explanations for the presence of short bla_{NDM} carrying contigs in the dataset: (i) the presence of IS*26* TUs in the host cell; (ii) other circular DNA formations mediated by plasmid

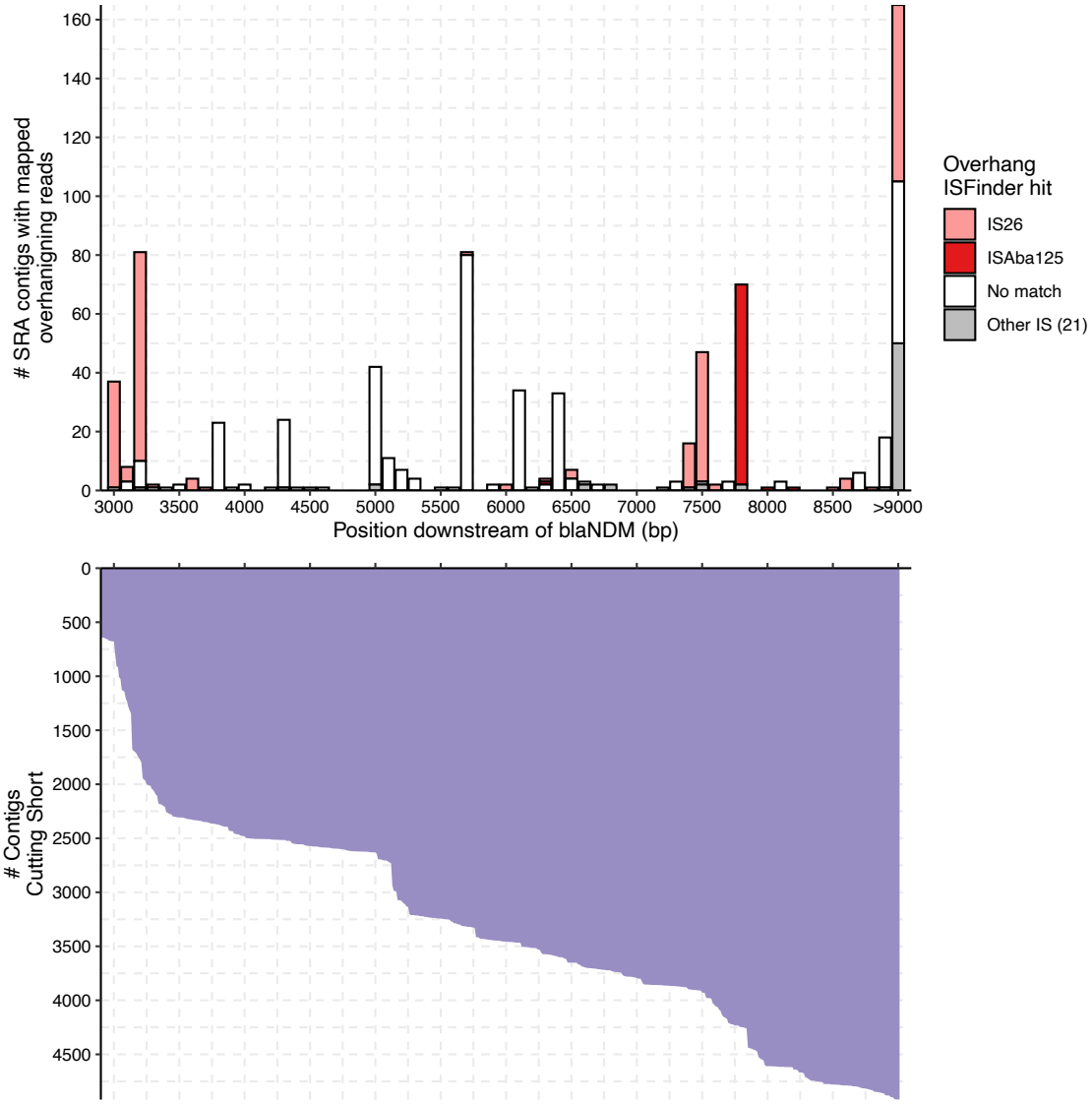


Figure 4.14. Mapping of overhangs of *bla*$_{NDM}$-carrying contigs to the ISFinder database. Reads of 781 Illumina paired-end sequencing datasets from SRA was mapped back to the corresponding contigs with length downstream of *bla*$_{NDM}$ ≥3000 bp. Downstream overhangs ≥50 bp were then screened against ISFinder database. The top panel provides the distribution of ISFinder hits over the lengths of contigs downstream of *bla*$_{NDM}$ start codon. The bottom panel is an excerpt from Figure 4.13 and shows a cumulative distribution of all contig lengths downstream of *bla*$_{NDM}$ start codon. From the bottom panel, three sharp increases in number of short contig can be distinguished at positions: 3000-3300 bp, 5000-5250 bp, and 7500-8000 bp.

recombination, transposons (Poirel, Bonnin, et al., 2011; Roach et al., 2020) or ISCR elements (Li et al., 2014; Toleman et al., 2006); (iii) missasembly of contigs due to presence of multiple copies of the same ISs (Sohn & Nam, 2018).

To further investigate the mobility of *bla*$_{NDM}$, I characterised the most common (pseudo-)composite transposons theoretically capable of mobilizing *bla*$_{NDM}$ (Figure 4.15). These were defined as stretches of DNA flanked by two matching complete or partial ISs <30 kb apart and enclosing *bla*$_{NDM}$. In total, I identified 640 composite transposons in 468 contigs which comprised 31 different types with the most frequent being: IS*26* (231 instances), IS*3000* (forming Tn*3000*; 168), IS*Aba125* (forming Tn*125*; 138 instances), and IS*15* (28) (Figure 4.15B). Interestingly, there are 80 cases where >2 of the same IS flank *bla*$_{NDM}$. These are mostly IS*26* (59) which could indicate the presence of cointegrate formation (Harmer et al., 2020) and showcases increased activity of this particular insertion element. Only 431 of the 640 putative composite transposons identified contained both complete flanking ISs, while others had at least one IS partially truncated. In addition, 1,681 ISCR27, and 150 ISCR1 were found in similar proximity and appropriate orientation to mobilize *bla*$_{NDM}$ (Figure 4.15B). However, as mentioned earlier, their role in transposition of *bla*$_{NDM}$ appears minor.

In the majority of cases, composite transposons Tn*125* and Tn*3000* were found to have a consistent length ranging from 7-10 kb (Figure 4.15A). Similarly, ISCR1 and ISCR27 are found at fixed positions downstream of *bla*$_{NDM}$. However, the lengths of transposons formed by IS*15*, a known variant of IS*26* (Harmer & Hall, 2019), and especially IS*26* were found to be more variable. Pairs of IS*26* are found to be 2.5-30 kb apart again consistent with increased activity and multiple independent insertions. IS*15* and IS*26* occur at increased presence in samples collected in East and Southeast Asia (Figure 4.15C). These occur roughly equally in *Escherichia* and *Klebsiella* genera (Figure 4.15D) and are associated to multiple plasmid backbones, but predominantly on IncF plasmids (Figure 4.15E). Tn*125* and Tn3000 have a notable predominance in the Indian subcontinent (Figure 4.15C) largely in *Acinetobacter* and *Klebsiella* genera respectively (Figure 4.15D).

4.3.6. Molecular dating of key events

As described in Section 4.2.5, I estimated the relative timing of the formation of the Tn*125* and Tn*3000* transposons. After selecting only contigs with conserved transposon configurations I aligned each transposon region and identified the likely root (ancestral)

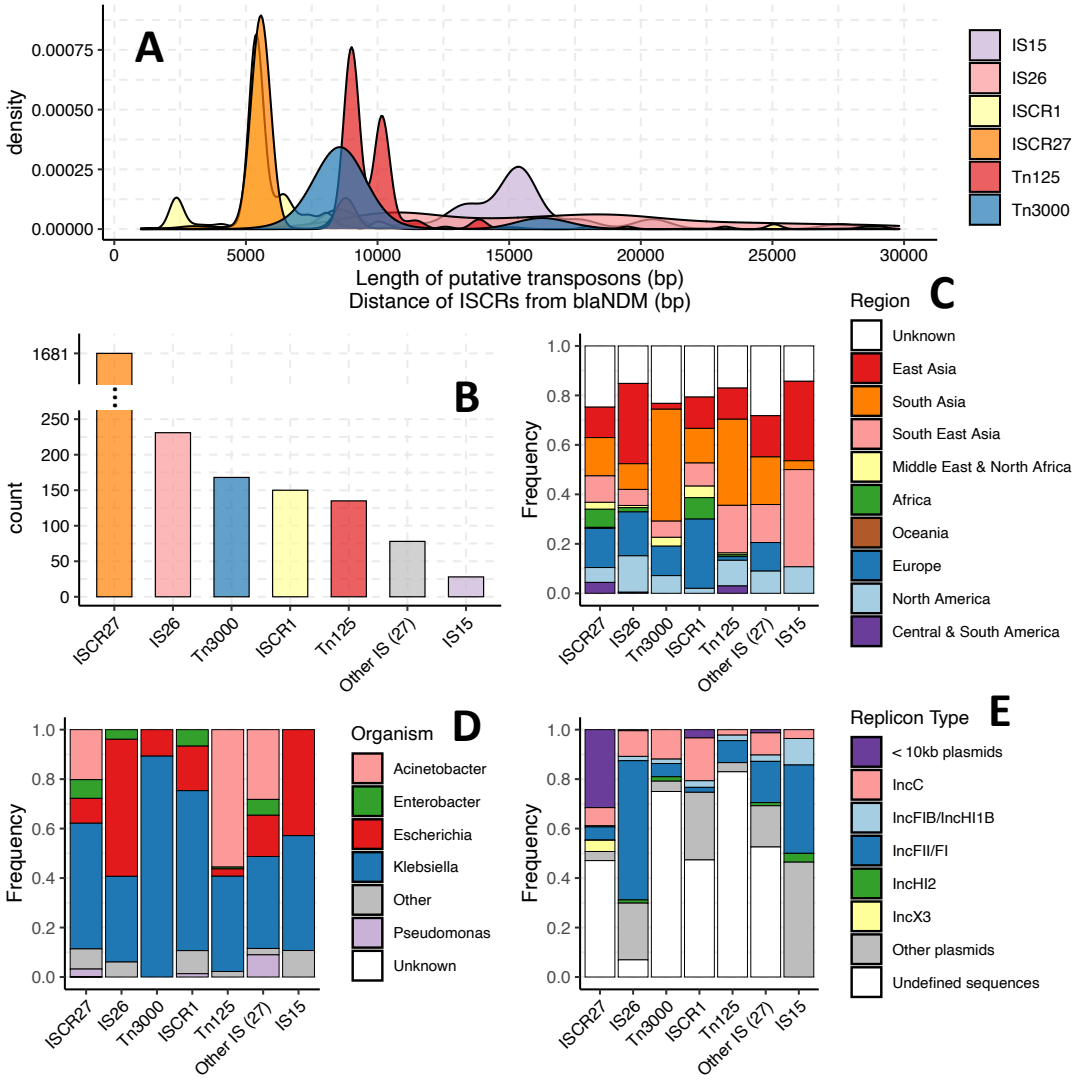


Figure 4.15. Global prevalence and genetic context of the most frequent putative (pseudo-)composite transposons and ISCRs capable of mobilizing the *bla*$_{NDM}$ gene. Transposons were defined as stretches of DNA flanked by two matching complete or partial ISs <30 kb apart and enclosing *bla*$_{NDM}$. Putative pseudo-composite transposons were labelled according to their constituent ISs (IS15 and IS26). Panel (**A**) shows the marginal distributions of transposon lengths or distances of ISCRs from the *bla*$_{NDM}$ start codon. Panel (**B**) shows the overall counts of the frequent TEs (i.e., >25 representatives). Panels (**C**), (**D**) and (**E**) are bar plots respectively indicating proportions of plasmid backbones, bacterial genera and sampling location associated with most frequent TEs.

sequence by assessing temporal patterns (Figure 4.3 and Figure 4.4). Overall, fewer SNPs, mostly located within the transposase gene, were observed in the alignment of Tn*3000* compared to Tn*125*; nevertheless, a significant temporal signal was recovered for both (Figure 4.5 and Figure 4.6). I also assessed temporal signal for three other prevalent insertion events (Figure 4.13), namely: *bla*$_{NDM}$ with downstream ISCR27, *bla*$_{NDM}$ with

correctly oriented downstream *folP*-ISCR1 (+ strand), and *bla*$_{NDM}$ – *dsbD* with downstream ISCR1 (- strand) ending with an unknown putative IS (labelled IS-?). However, no significant temporal signal was recovered for these.

This Bayesian analysis indicated that the most recent common ancestor (MRCA) of the Tn*125* transposon carrying the *bla*$_{NDM}$ gene dated to before 1990 (Figure 4.16). While the time intervals are uncertain, the results are consistent with a MRCA in the mid-20th century – half a century prior to the first reported Tn*125*-*bla*$_{NDM}$-positive isolates (Jones et al., 2014). Conversely, the mobilization of *bla*$_{NDM}$ by Tn*3000* is estimated to have happened later at the turn of the millennium (Figure 4.16). These findings are consistent with a wider narrative whereby the spread of *bla*$_{NDM}$ was initially driven by Tn*125* mobilization before subsequent transposition by Tn*3000*, IS*26* and others.

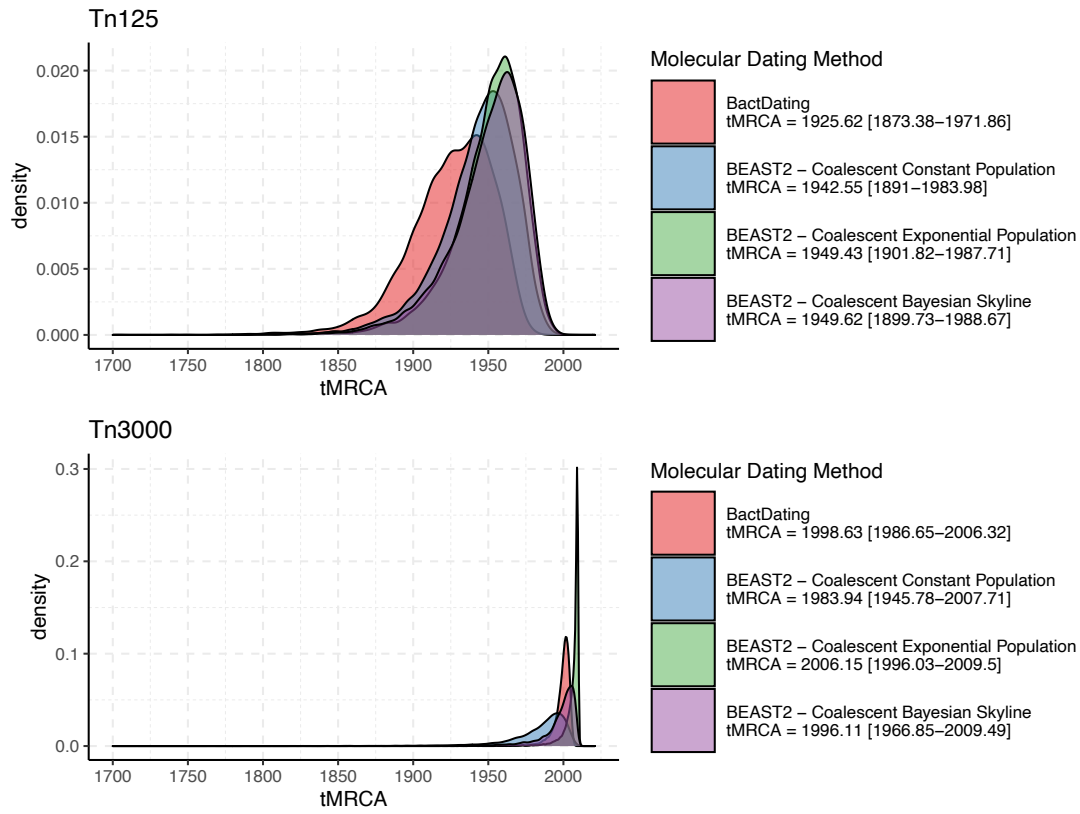


Figure 4.16. Posterior density distributions of ancestral sequence age (i.e., root height) for the *Tn125* and *Tn3000* transposons. The ancestral sequence emergence was estimated using two Bayesian tip-dating approaches implemented in BactDating and BEAST2. Three different population growth priors were used in case of BEAST2: Coalescent Constant Population, Coalescent Exponential Population, and Coalescent Bayesian Skyline as given by the colour scheme and legend at right. Median estimates with 95% highest density interval (HDI) are provided in the panel legends.

4.3.7. Temporal diversity in *bla*$_{NDM}$ isolates suggests role of plasmids

The earliest samples included in the dataset are from 2007 to 2010 and comprise 21 *bla*$_{NDM}$-positive isolates. These already encompass seven bacterial species, collected in eight countries spanning four geographic regions (17 clinical samples and four of unknown origin from South Asia, Middle East, Oceania, and Europe). Such a wide host and geographic distribution, even in the earliest available genomes, illustrates the extraordinarily high mobility of *bla*$_{NDM}$ at this stage and is consistent with the molecular dating estimates.

In order to trace the progress of *bla*$_{NDM}$'s rapid spread after 2005 (to coincide with the first published observations), I measured diversity over time for several metadata categories, including country, genera, plasmid backbone and IS presence (**Error! Reference source not found.**). The change in diversity of the countries associated to *bla*$_{NDM}$-positive isolates was used to approximate the broad patterns of global dissemination of *bla*$_{NDM}$. Our results are consistent with the spread stabilising between 2013-2015, with a gradual decline in diversity afterwards (**Error! Reference source not found.**A). This observation supports a scenario whereby the global dissemination of NDM took place over 8-10 years. Temporal diversity of bacterial genera was largely unchanged, consistent with *bla*$_{NDM}$ having been highly mobile across genera since at least 2005 (**Error! Reference source not found.**B).

The estimated change in the diversity of countries associated to *bla*$_{NDM}$-positive isolates was positively correlated with other metadata categories (Figure 4.18) suggesting this approach holds valuable information which can be leveraged to reconstruct dissemination trends. The strongest correlation was found between the diversity of countries and identified plasmid backbones ($\rho = 0.864$ [0.691-0.964]) supporting a strong dependence between the two (Figure 4.18B). To further investigate this relationship, I assessed the correlation between genetic and geographic distance between pairs of confirmed plasmid contigs (tested for IncF, IncX3, IncC, IncN2 and confirmed plasmid contigs >10 kb) as a function of the distance downstream of *bla*$_{NDM}$ gene (Figure 4.19). The approach is described in detail in Section 4.2.7.

No relationship was detected for IncX3 and IncN2 plasmids (Figure 4.19A and B) likely due to the lack of long plasmid sequences and insufficient variation in geographic distance between pairs of plasmids as both replicon types are mostly localized to China and India respectively (Figure 4.9). However, in all other cases aside from IncN2

plasmids, a peak in the correlation recovered between genetic and geographical distance was observed immediately downstream of *bla*$_{NDM}$ possibly signifying more recent and local genome reshuffling events (Figure 4.19). More importantly, in the case of IncF and IncC, and other confirmed plasmid contigs, a notable and gradual increase in the strength of correlation was noted as more downstream plasmid sequence is included in the analysis (Figure 4.19B, C, and D). These trends suggest that plasmids carrying *bla*$_{NDM}$ are geographically structured and that dissemination of *bla*$_{NDM}$ fundamentally follows an isolation by distance process, where most expansion events happen over short physical distances. This would be consistent with the existence of plasmid niches: settings to which particular plasmids are more adapted.

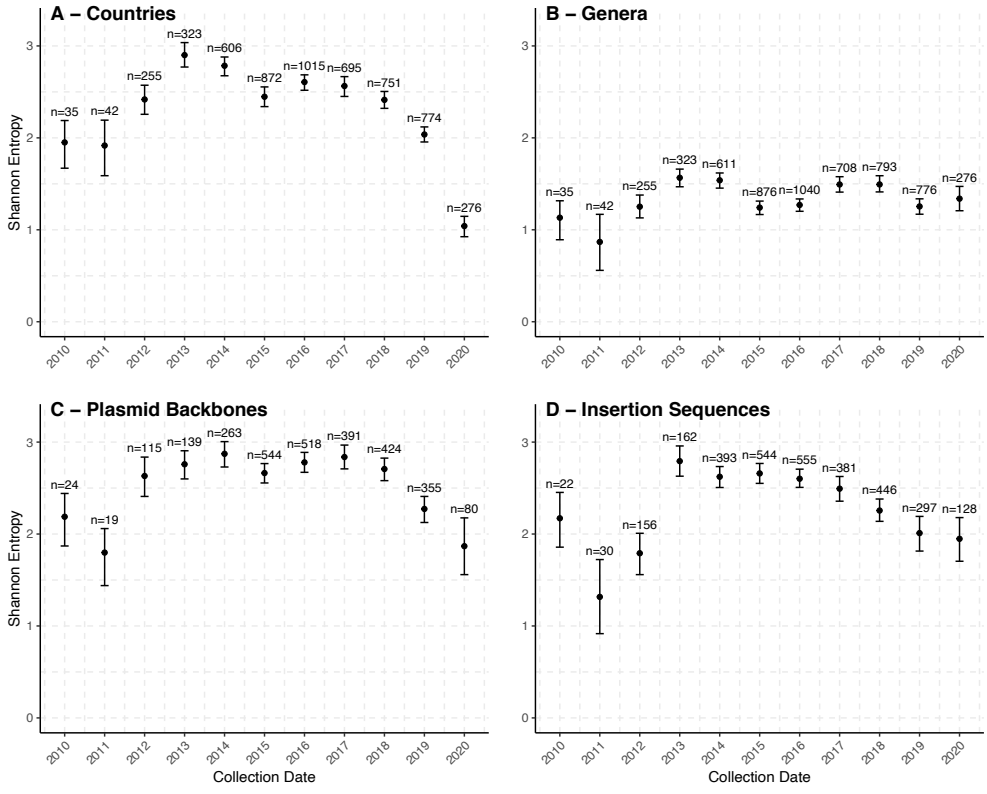


Figure 4.17. Change in Shannon entropy (diversity) over time for four categories of NDM-positive samples. More specifically, the entropy was estimated for samples' country labelling (A), bacterial genera (B), plasmid backbones bearing *bla*$_{NDM}$ gene (as determined by mapping to plasmid reference sequences, C), Insertion sequences found in the vicinity (≤ 10kb) of *bla*$_{NDM}$ gene (D). The median entropy (points) together with 95% confidence interval (error bars) was estimated using bootstrapping with replacement (1,000 iterations). Years with too few associated samples (≤10) were excluded from the analysis.

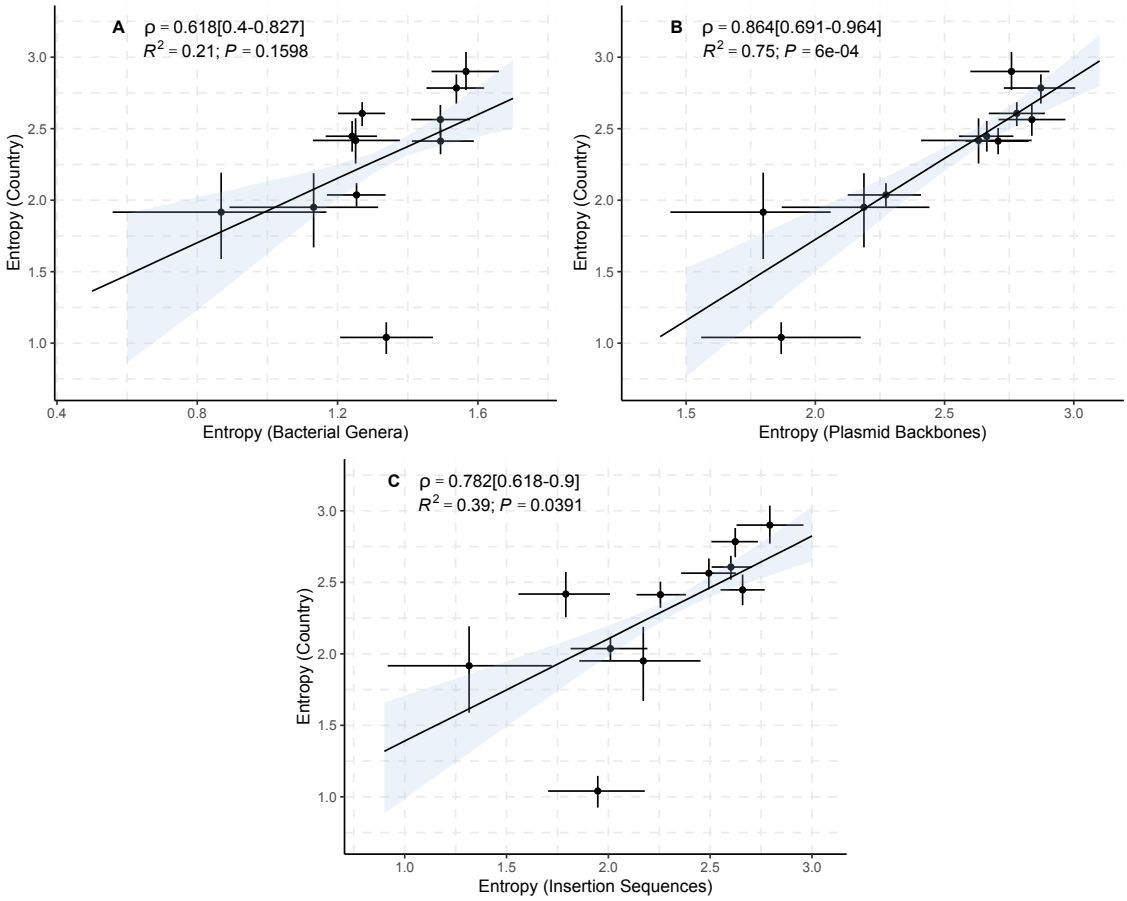


Figure 4.18. Spearman correlation and linear regression between Shannon entropy (diversity) estimates. The Shannon entropy bootstrapped values (**Error! Reference source not found.**) were used to provide a median and 95% confidence interval (CI) of Spearman correlation coefficients, as well as a median regression line with 95% CI (ribbon) between samples' country labelling and: bacterial genera (**A**), plasmid backbones (**B**), and Insertion Sequences (**C**).

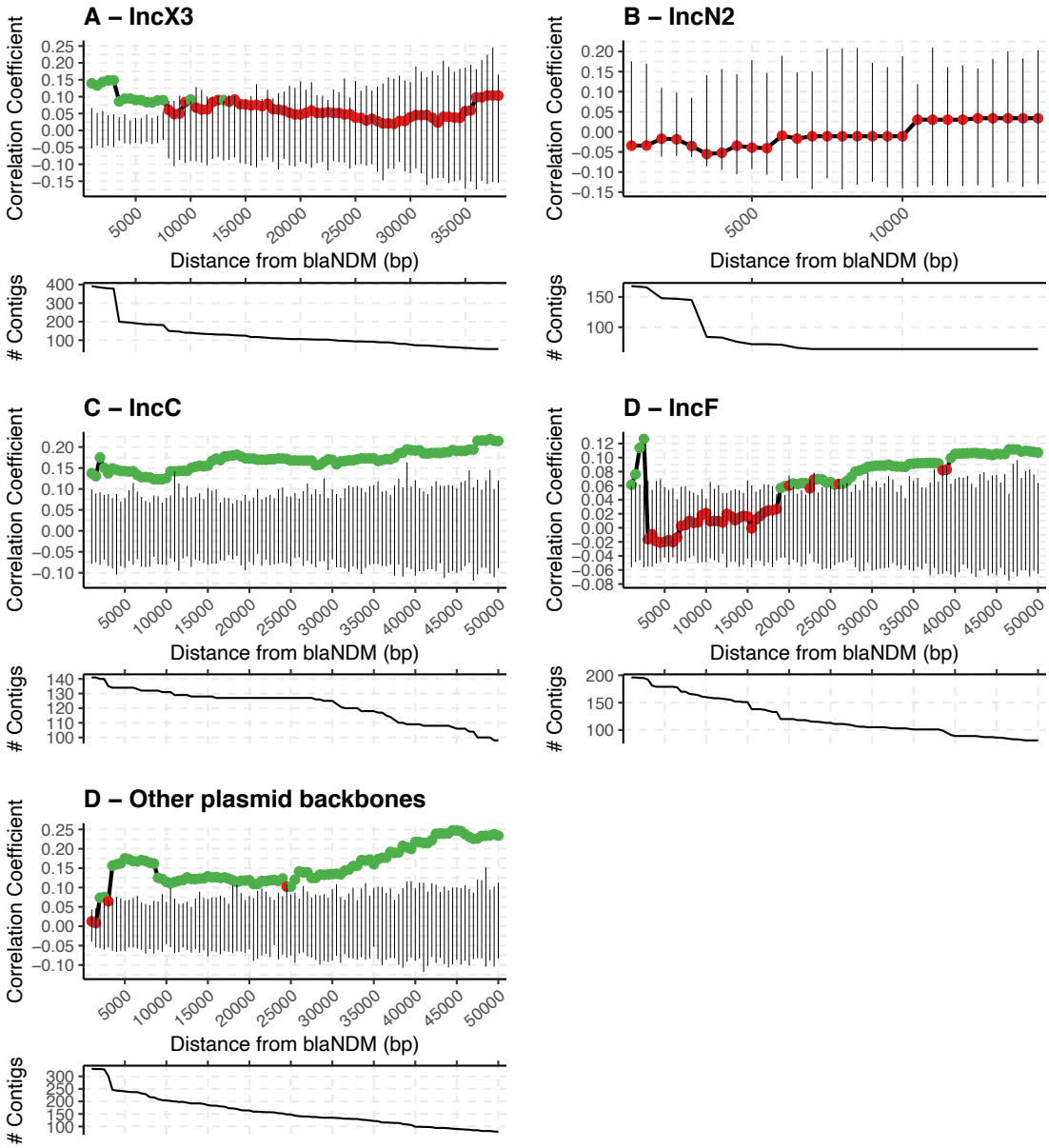


Figure 4.19. The spearman correlation estimates between genetic and geographic distance of NDM-positive contigs as the DNA sequence upon which the genetic distance is measured is increased downstream of *bla*$_{NDM}$ gene. The analysis was performed on contigs with confirmed IncX3 (**A**), IncN2 (**B**), IncC (**C**), and IncF (**D**) replicon types, and other plasmid backbones >10 kb (**D**). The genetic (JD) and geographic distance was measured between all pairs of contigs which yielded two distance matrices. The Spearman correlation was then estimated between two matrices and its significance evaluated using Mantel (randomization) test. Significant Spearman correlations (p-value <0.05) are indicated with green points and non-significant correlations with the red point, while the black vertical lines provide the 95% confidence interval of 100 Mantel test permutations. The genetic distance matrix and subsequent Spearman correlation were estimated multiple times by increasing the assessed DNA sequence starting from *bla*$_{NDM}$ gene and continuing downstream. The plot below each correlation graph indicates the number of contigs used in the correlation analysis as the assessed DNA sequence is increased.

4.4. Discussion

In this chapter, I present the extant structural variation around *bla*$_{NDM}$ in a large dataset to reconstruct its evolutionary history and the main genetic elements involved in its global spread. The results, summarized in Figure 4.13, highlight an ancestral background of *bla*$_{NDM}$ as well as several insertion events and a myriad of other genetic reshuffling processes, together pointing to an early emergence of *bla*$_{NDM}$ followed by a more recent and rapid dissemination globally. Genetic reshuffling and mobilization of *bla*$_{NDM}$ by multiple transposons aided its rapid dissemination via a multitude of plasmid backbones.

Here, I go beyond previous smaller studies by dating the MRCA of the hypothesised ancestral form – the transposon Tn*125*, together with *bla*$_{NDM}$ in its chimeric form (Toleman et al., 2012) – to pre-1990, and possibly well back into the mid-20th century. A likely scenario is an origin event in the genus *Acinetobacter* in the Indian subcontinent. Tn*125* is found mostly in *Acinetobacter* and *Klebsiella* species and it is likely this transposon played an important role in early plasmid jumps of *bla*$_{NDM}$, given it is the dominant transposon in the dataset which encompasses the ancestral genetic background of *bla*$_{NDM}$ – *groS*/*groL* genes and ISCR27 sequence. I also estimated the formation of a secondary transposon, involving Tn*3000*, which remobilized the region likely in *Klebsiella* species sometime between the 1980s and early 2000s. However, the results suggest Tn*3000* likely played a lesser role in the early spread of *bla*$_{NDM}$ as it does not include the ISCR27 found at least partially preserved in many more recent samples.

In total, 31 different putative transposons were identified within the dataset. Their role, together with integrons and other transposable elements, is likely mostly minor or disruptive, as suggested for ISCR1. However, I do identify IS*26* as of interest, given it frequently forms putative transposons in our dataset, especially in IncF plasmids. IS*26* is known for its increased activity and rearrangement of plasmids in clinical isolates (S. He et al., 2015) and has been observed to drive within-plasmid heterogeneity even in a single *E. coli* isolate (D. D. He et al., 2019). Thus, IS26 flanked pseudo-composite transposons likely represent the most important contributor to genetic reshuffling of *bla*$_{NDM}$ in recent times.

The assessment of temporal diversity in the associated of countries of origin of *bla*$_{NDM}$ positive isolates supports a globalisation peak in 2013-2015. Such a rapid 8-10

year world-wide spreading has been suggested for other important mobile resistance genes such as the *mcr-1* gene, mediating colistin resistance (R. Wang, Van Dorp, et al., 2018). Furthermore, I identified 33 different plasmid types carrying bla_{NDM} and a positive correlation between genetic distance calculated for differing lengths of plasmid backbones and geographic distances of sampling locations. Such an observation is consistent with the existence of a constraint on plasmid spread, i.e., plasmid niches, which highlights the importance of between-plasmid transposon jumps and genetic recombination for AMR spreading at a global scale.

Chapter 5

Conclusion

Some bacterial pathogens, like members of *Escherichia*, *Klebsiella*, *Enterococcus* or *Streptococcus*, have a particularly extensive accessory genomes (Brockhurst et al., 2019; Eldholm & Balloux, 2016). For instance, *E. coli* has been estimated to have ~2,200 conserved genes of a reservoir of more than 13,000 genes (Rasko et al., 2008). Furthermore, many bacterial genes only seem to be beneficial under specific environmental circumstances making them a burden in others and prone to being silenced or lost (Brockhurst et al., 2019; Price et al., 2018). This selective pressure places an emphasis on HGT as a dominant force of prokaryotic evolution with gains and losses of genes introducing variation on which evolutionary forces can act (Hall et al., 2017; Vos et al., 2015). In this process bacterial plasmids undoubtedly play an important role as carriers of genetic variation, i.e., evolutionary novelty. Nevertheless, large-scale comparative pangenome studies over wide spatial and temporal scales are difficult to conduct due to frequent genome reshuffling. As a result many questions remain unanswered as the nature of gene exchange and the impact of specific MGEs remain only partially understood (Brockhurst et al., 2019; Vos et al., 2015). Tallying the multi-layered and recombinogenic nature of MGEs, in this thesis I presented two approaches to studying HGT aimed at two layers of accessory genome complexity. The first one, presented in Chapter 2, concerns population structure of bacterial plasmids uncovered using a network-based approach. The second was presented in Chapter 4 and focuses on the detection of localized genome reshuffling events, such as transposition, helping to decipher patterns of mobility of the AMR gene bla_{NDM}.

In brief, in Chapter 2, the similarity between bacterial plasmids was represented as a network. After removing the edges connecting plasmids sharing low similarity, a community-detection algorithm was applied to identify highly interconnected regions of the network thus effectively classifying plasmids into clusters (cliques) of higher similarity. Uncovering the plasmid population structure in this manner was purely sequence-based and relied on a single presumption inherent to our understanding of gene mobility: most accessory genes are not conserved across evolutionary related plasmid backbones. The exact similarity (i.e., JI or k-mer sharing) threshold determining the

relatedness across plasmid backbones and consequently the level of gene conservation likely varies across different plasmids and depends on bacterial host and particular plasmid function. A flexible similarity threshold could be considered in more formal implementations of the method, although this would likely come with computational challenges such as developing a fast and reliable approaches for clique detection and assessment. Regardless, a 0.3 JI threshold allowed me to cluster plasmids in a manner which replicated the results and improved upon classifications by well-established replicon-based typing.

Additional evidence for the biological implications of inferring plasmid population structure was provided in Chapter 3. Classified plasmids were found to be consistent in gene content, GC content, and MOB types. Perhaps the most interesting finding was the high association of classified plasmids with specific bacterial hosts. This implies a strong constraint on between-host movement of many plasmids and questions the commonplace existence of BHR plasmids (A. Jain & Srivastava, 2013). Finally, representing relatedness of plasmids as a network opens prospects for further studies into gene mobility as highlighted in Figure 3.10 which shows the importance of transposons as a link between host-constrained clusters of plasmids, especially in pathogenic genera such as *Escherichia* and *Klebsiella*.

An in-depth case-study of global spreading of *bla*$_{NDM}$, a significant AMR gene, was presented in Chapter 4. By identifying stretches of homology across the *bla*$_{NDM}$-carrying contigs in the immediate genetic neighbourhood of the *bla*$_{NDM}$ gene, I was able to pinpoint the main actors driving its mobility. An increased activity of IS*26*-flanked pseudo-composite transposons was detected which potentially aided the rapid global dissemination of *bla*$_{NDM}$ (Jones et al., 2014; Weber et al., 2019; Zhang et al., 2018)**.** Furthermore, identified structural variants and stretches of homology informed the construction of alignments which could be subsequently used for phylogenetic molecular dating of the timing of *bla*$_{NDM}$ mobilization events. The results presented in this chapter suggest that mobilization of the *bla*$_{NDM}$ progenitor, by an ancestral Tn*125* and Tn*3000*, considerably predated the earliest reported cases of NDM-mediated resistance. This is not altogether surprising considering beta-lactamases have been demonstrated to have existed for millions of years (Dcosta et al., 2011). However, the early placement of *bla*$_{NDM}$ in the genetic background which now dominates the pool of NDM-resistant bacterial genomes hints at the existence of an event, or series of events, which triggered its more recent and rapid dissemination. One such event could be the mobilization of *bla*$_{NDM}$ by the

aforementioned IS*26* element. Another likely key event was the relocation of *bla*$_{NDM}$-carrying Tn*125* from *Acinetobacter* to *Klebsiella* species, whose expansive mobilome may have facilitated further spread of the gene.

bla$_{NDM}$ reached global prevalence in 8-10 years and its dissemination was found to be strongly correlated with an increase in diversity of plasmid backbones. Closer examination of the geographic distribution of plasmid backbones in Chapter 4 suggests the possibility of a constraint on plasmid spreading. Scientists have known for some time about the existence of a variety of mechanistic and ecological limitations to HGT where genes are mainly spread between compatible bacterial species sharing a common habitat (Brockhurst et al., 2019; Smillie et al., 2011). The results summarized in this thesis suggest these constraints are also reflected by the mobility of plasmids. Hence, I introduced a term 'plasmid niche' to define the likely complex local ecological and evolutionary pressures acting on particular plasmid backbones. The number of hypothetical constraints that may contribute to restricting plasmid range are diverse and could, in no specific order include factors such as country boundaries limiting population movement, region-specific practice in antibiotic usage, influence of co-resistance, plasmid fitness costs, conjugation rates and copy numbers, the narrow host range of most bacterial plasmids, or plasmids being associated with particular locations or environments. Thus, an introduction of another plasmid into a foreign plasmid niche may lead to plasmid loss or fast adaptation by, for instance, acquisition of resistance and other accessory elements. This hypothetical scenario also provides an opportunity for resistance to spread by transposition or recombination, by which a new resistance gene could establish itself into another plasmid niche. In the case of *bla*$_{NDM}$, this would also imply that after the initial introduction of *bla*$_{NDM}$ to a geographic region, dissemination and persistence of the gene could proceed idiosyncratically – selection for carbapenem resistance being just one of many selective pressures acting on plasmid diversity.

The importance of transposon movement in HGT has been demonstrated throughout this thesis as well as in other publications investigating plasmid networks (Redondo-Salvo et al., 2020) and promoting a Russian-doll model of resistance mobility (A. E. Sheppard et al., 2016; R. Wang, Van Dorp, et al., 2018). Considering the results presented in this thesis, I suggest an improved conceptual framework of AMR gene dissemination across genera where plasmid mobility is for the most part spatially restricted. Although plasmids can facilitate rapid spread within specific bacterial habitats, plasmid transfer may not be the main driver of widespread dissemination. Instead, most

plasmid horizontal transfers are likely only transient, with plasmids generally failing to establish themselves in the new bacterial host. Though, such aborted plasmid exchanges still provide a crucial opportunity for transposon jumps and genetic recombination to spread AMR genes across bacterial species.

Bibliography

Abedon, S. T., Thomas-Abedon, C., Thomas, A., & Mazure, H. (2011). Bacteriophage prehistory. *Bacteriophage*, *1*(3), 174–178. https://doi.org/10.4161/bact.1.3.16591

Acman, M., van Dorp, L., Santini, J. M., & Balloux, F. (2020). Large-scale network analysis captures biological features of bacterial plasmids. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-16282-w

Akarsu, H., Bordes, P., Mansour, M., Bigot, D. J., Genevaux, P., & Falquet, L. (2019). TASmania: A bacterial toxin-antitoxin systems database. *PLoS Computational Biology*, *15*(4), e1006946. https://doi.org/10.1371/journal.pcbi.1006946

Al-Shayeb, B., Sachdeva, R., Chen, L. X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., … Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems. *Nature*, *578*(7795), 425–431. https://doi.org/10.1038/s41586-020-2007-4

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., … McArthur, A. G. (2020). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *48*(D1), D517–D525. https://doi.org/10.1093/nar/gkz935

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ambrose, S. J., Harmer, C. J., & Hall, R. M. (2018). Compatibility and entry exclusion of IncA and IncC plasmids revisited: IncA and IncC plasmids are compatible. *Plasmid*, *96–97*, 7–12. https://doi.org/10.1016/j.plasmid.2018.02.002

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., & Pevzner, P. A. (2016). plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data. *Draft*, *32*(22), 3380–3387. https://doi.org/10.1101/048942

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. In *Nature Genetics* (Vol. 25, Issue 1, pp. 25–29). NIH Public Access. https://doi.org/10.1038/75556

Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., Tewolde, R., Schaefer, U., Jenkins, C., Dallman, T. J., De Pinna, E. M., & Grant, K. A. (2016). Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ*, *2016*(4), e1752. https://doi.org/10.7717/peerj.1752

Avery, O. T., Macleod, C. M., & 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. *Journal of Experimental Medicine*, *79*(2), 137–158. https://doi.org/10.1084/jem.79.2.137

Babić, A., Lindner, A. B., Vulić, M., Stewart, E. J., & Radman, M. (2008). Direct visualization of horizontal gene transfer. *Science*, *319*(5869), 1533–1536. https://doi.org/10.1126/science.1153498

Balcazar, J. L. (2014). Bacteriophages as Vehicles for Antibiotic Resistance Genes in the Environment. *PLoS Pathogens*, *10*(7), e1004219. https://doi.org/10.1371/journal.ppat.1004219

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M.,

Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F. J., Dupré, J., Dagan, T., Boucher, Y., & Martin, W. (2009). Prokaryotic evolution and the tree of life are two different things. In *Biology Direct* (Vol. 4, Issue 1, p. 34). BioMed Central. https://doi.org/10.1186/1745-6150-4-34

Baraniak, A., Izdebski, R., Fiett, J., Gawryszewska, I., Bojarska, K., Herda, M., Literacka, E., Zabicka, D., Tomczak, H., Pewińska, N., Szarata, M., Ozorowski, T., Milner, A., Hryniewicz, W., & Gniadkowski, M. (2016). NDM-producing Enterobacteriaceae in Poland, 2012-14: Inter-regional outbreak of Klebsiella pneumoniae ST11 and sporadic cases. *Journal of Antimicrobial Chemotherapy*, *71*(1), 85–91. https://doi.org/10.1093/jac/dkv282

Basu, S. (2020). Variants of the New Delhi metallo-β-lactamase: New kids on the block. In *Future Microbiology* (Vol. 15, Issue 7, pp. 465–467). Future Medicine Ltd. https://doi.org/10.2217/fmb-2020-0035

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. In *FEMS Microbiology Reviews* (Vol. 38, Issue 4, pp. 720–760). Oxford Academic. https://doi.org/10.1111/1574-6976.12058

Bender, J., & Kleckner, N. (1986). Genetic evidence that Tn10 transposes by a nonreplicative mechanism. *Cell*, *45*(6), 801–815. https://doi.org/10.1016/0092-8674(86)90555-6

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1). https://doi.org/10.1093/nar/gks1195

Bernard, G., Chan, C. X., Chan, Y. B., Chua, X. Y., Cong, Y., Hogan, J. M., Maetschke, S. R., & Ragan, M. A. (2019). Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics*, *20*(2), 426–435. https://doi.org/10.1093/bib/bbx067

Bernard, G., Greenfield, P., Ragan, M. A., & Chan, C. X. (2018). k-mer Similarity Networks of Microbial Genomes and Taxonomic Rank. *MSystems*, *3*(6). https://doi.org/10.1128/msystems.00257-18

Bonnin, R. A., Poirel, L., Naas, T., Pirs, M., Seme, K., Schrenzel, J., & Nordmann, P. (2012). Dissemination of New Delhi metallo-β-lactamase-1-producing Acinetobacter baumannii in Europe. *Clinical Microbiology and Infection*, *18*(9), E362–E365. https://doi.org/10.1111/j.1469-0691.2012.03928.x

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., Philippon, A., Allesoe, R. L., Rebelo, A. R., Florensa, A. F., Fagelhauer, L., Chakraborty, T., Neumann, B., Werner, G., Bender, J. K., Stingl, K., Nguyen, M., Coppens, J., Xavier, B. B., … Aarestrup, F. M. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, *75*(12), 3491–3500. https://doi.org/10.1093/jac/dkaa345

Botelho, J., & Schulenburg, H. (2021). The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. In *Trends in Microbiology* (Vol. 29, Issue 1, pp. 8–18). Elsevier Current Trends. https://doi.org/10.1016/j.tim.2020.05.011

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Bradley, P., den Bakker, H. C., Rocha, E. P. C., McVean, G., & Iqbal, Z. (2019). Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology*, *37*(2), 152–159. https://doi.org/10.1038/s41587-018-0010-1

Brantl, S. (2014). Plasmid Replication Control by Antisense RNAs. *Microbiology Spectrum*, *2*(4). https://doi.org/10.1128/microbiolspec.plas-0001-2013

Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. In *Nature Microbiology* (Vol. 3, Issue 7, pp. 754–766). Nature Publishing Group. https://doi.org/10.1038/s41564-018-0166-y

Brencic, A., & Winans, S. C. (2005). Detection of and Response to Signals Involved in Host-Microbe Interactions by Plant-Associated Bacteria. *Microbiology and Molecular Biology Reviews*, *69*(1), 155–194. https://doi.org/10.1128/mmbr.69.1.155-194.2005

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

Brockhurst, M. A., Harrison, E., Hall, J. P. J., Richards, T., McNally, A., & MacLean, C. (2019). The Ecology and Evolution of Pangenomes. *Current Biology*, *29*(20), R1094–R1103. https://doi.org/10.1016/J.CUB.2019.08.012

Bron, C., & Kerbosch, J. (1973). Algorithm 457: Finding All Cliques of an Undirected Graph. *Communications of the ACM*, *16*(9), 575–577. https://doi.org/10.1145/362342.362367

Burrus, V., Pavlovic, G., Decaris, B., & Guédon, G. (2002). Conjugative transposons: The tip of the iceberg. In *Molecular Microbiology* (Vol. 46, Issue 3, pp. 601–610). John Wiley & Sons, Ltd. https://doi.org/10.1046/j.1365-2958.2002.03191.x

Bush, K. (2018). Past and present perspectives on β-lactamases. In *Antimicrobial Agents and Chemotherapy* (Vol. 62, Issue 10). American Society for Microbiology. https://doi.org/10.1128/AAC.01076-18

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/biblio/1241166

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Campos, J. C., Da Silva, M. J. F., Dos Santos, P. R. N., Barros, E. M., Pereira, M. D. O., Seco, B. M. S., Magagnin, C. M., Leiroz, L. K., De Oliveira, T. G. M., De Faria-Júnior, C., Cerdeira, L. T., Barth, A. L., Sampaio, S. C. F., Zavascki, A. P., Poirel, L., & Sampaio, J. L. M. (2015). Characterization of Tn3000, a transposon responsible for blaNDM-1 dissemination among Enterobacteriaceae in Brazil, Nepal, Morocco, and India. *Antimicrobial Agents and Chemotherapy*, *59*(12), 7387–7395. https://doi.org/10.1128/AAC.01458-15

Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K. L., & Threlfall, E. J. (2005). Identification of plasmids by PCR-based replicon typing. *Journal of Microbiological Methods*, *63*(3), 219–228. https://doi.org/10.1016/j.mimet.2005.03.018

Carattoli, A., & Hasman, H. (2020). PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). In *Methods in Molecular Biology* (Vol. 2075, pp. 285–294). Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9877-7_20

Carattoli, A., Zankari, E., Garcia-Fernandez, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M. M., & Hasman, H. (2014). In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy*, *58*(7), 3895–3903. https://doi.org/10.1128/AAC.02412-14

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., … Westerfield, M. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Carroll, A. C., & Wong, A. (2018). Plasmid persistence: costs, benefits, and the plasmid paradox. In *Canadian Journal of Microbiology* (Vol. 64, Issue 5, pp. 293–304). https://doi.org/10.1139/cjm-2017-0609

Castanheira, M., Deshpande, L. M., Mathai, D., Bell, J. M., Jones, R. N., & Mendes, R. E. (2011). Early dissemination of NDM-1- and OXA-181-producing Enterobacteriaceae in Indian hospitals: Report from the SENTRY Antimicrobial Surveillance Program, 2006-2007. *Antimicrobial Agents and Chemotherapy*, *55*(3), 1274–1278. https://doi.org/10.1128/AAC.01497-10

Chamberlain, S. A., & Szöcs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Research*, *191*, 1–28. https://doi.org/10.12688/f1000research.2-191.v2

Chandler, M. (2016). Transposons: Prokaryotic. In *eLS* (pp. 1–9). American Cancer Society.

https://doi.org/10.1002/9780470015902.a0000591.pub2

Chandler, M., De La Cruz, F., Dyda, F., Hickman, A. B., Moncalian, G., & Ton-Hoang, B. (2013). Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. In *Nature Reviews Microbiology* (Vol. 11, Issue 8, pp. 525–538). Nature Publishing Group. https://doi.org/10.1038/nrmicro3067

Chatterjee, S., Mondal, A., Mitra, S., & Basu, S. (2017). Acinetobacter baumannii transfers the blaNDM-1 gene via outer membrane vesicles. *Journal of Antimicrobial Chemotherapy*, *72*(8), 2201–2207. https://doi.org/10.1093/jac/dkx131

Chattoraj, D. K. (2000). Control of plasmid DNA replication by iterons: no longer paradoxical. *Molecular Microbiology*, *37*(3), 467–476. https://doi.org/10.1046/j.1365-2958.2000.01986.x

Chavda, K. D., Chen, L., Fouts, D. E., Sutton, G., Brinkac, L., Jenkins, S. G., Bonomo, R. A., Adams, M. D., & Kreiswirth, B. N. (2016). Comprehensive genome analysis of carbapenemase-producing Enterobacter spp.: New insights into phylogeny, population structure, and resistance mechanisms. *MBio*, *7*(6). https://doi.org/10.1128/mBio.02093-16

Chen, I., & Dubnau, D. (2004). DNA uptake during bacterial transformation. In *Nature Reviews Microbiology* (Vol. 2, Issue 3, pp. 241–249). Nature Publishing Group. https://doi.org/10.1038/nrmicro844

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Research*, *33*(DATABASE ISS.). https://doi.org/10.1093/nar/gki008

Chen, Y., Braathen, P., Léonard, C., & Mahillon, J. (1999). MIC231, a naturally occurring mobile insertion cassette from Bacillus cereus. *Molecular Microbiology*, *32*(3), 657–668. https://doi.org/10.1046/j.1365-2958.1999.01388.x

Chikami, G. K., Guiney, D. G., Schmidhauser, T. J., & Helinski, D. R. (1985). Comparison of 10 IncP plasmids: Homology in the regions involved in plasmid replication. *Journal of Bacteriology*, *162*(2), 656–660. https://doi.org/10.1128/jb.162.2.656-660.1985

Clewell, D. B. (Ed.). (1993). *Bacterial Conjugation*. Springer Science+Business Media.

Clowes, R. C. (1972). Molecular structure of bacterial plasmids. *Bacteriological Reviews*, *36*(3), 361–405. https://doi.org/10.1128/mmbr.36.3.361-405.1972

Corel, E., Lopez, P., Méheust, R., & Bapteste, E. (2016). Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. In *Trends in Microbiology* (Vol. 24, Issue 3, pp. 224–237). Elsevier. https://doi.org/10.1016/j.tim.2015.12.003

Couturier, M., Bex, F., Bergquist, P. L., & Maas, W. K. (1988). Identification and classification of bacterial plasmids. In *Microbiological Reviews* (Vol. 52, Issue 3, pp. 375–395). American Society for Microbiology (ASM). https://doi.org/10.1128/mmbr.52.3.375-395.1988

Crozat, E., Fournes, F., Cornet, F., Hallet, B., & Rousseau, P. (2014). Resolution of Multimeric Forms of Circular Plasmids and Chromosomes. *Microbiology Spectrum*, *2*(5). https://doi.org/10.1128/microbiolspec.plas-0025-2014

D'Hérelle, F. (1917). Sur un microbe invisible antagoniste des bacilles dysentériques. *C R Acad Sci Ser D*, *165*, 373–375.

Dagan, T., Artzy-Randrup, Y., & Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(29), 10039–10044. https://doi.org/10.1073/pnas.0800679105

Dale, J. W., & Park, S. F. (2010). Molecular genetics of bacteria. In *John Wiley & Sons, Ltd* (5th ed.). https://doi.org/10.1080/09084280802073344

Datta, S., Mitra, S., Chattopadhyay, P., Som, T., Mukherjee, S., & Basu, S. (2017). Spread and exchange of bla NDM-1 in hospitalized neonates: role of mobilizable genetic elements. *European Journal of Clinical Microbiology and Infectious Diseases*, *36*(2), 255–265. https://doi.org/10.1007/s10096-016-2794-6

Dcosta, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., & Wright, G. D. (2011). Antibiotic

resistance is ancient. In *Nature* (Vol. 477, Issue 7365, pp. 457–461). Nature Publishing Group. https://doi.org/10.1038/nature10388

de Man, T. J. B., & Limbago, B. M. (2016). SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. *MSphere*, *1*(1). https://doi.org/10.1128/msphere.00050-15

del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M., & Díaz-Orejas, R. (1998). Replication and Control of Circular Bacterial Plasmids. *Microbiology and Molecular Biology Reviews*, *62*(2), 434–464. https://doi.org/10.1128/mmbr.62.2.434-464.1998

Delavat, F., Miyazaki, R., Carraro, N., Pradervand, N., & van der Meer, J. R. (2017). The hidden life of integrative and conjugative elements. In *FEMS microbiology reviews* (Vol. 41, Issue 4, pp. 512–537). Oxford Academic. https://doi.org/10.1093/femsre/fux008

Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. *Genome Biology and Evolution*, *3*(1), 959–973. https://doi.org/10.1093/gbe/evr077

Díaz-Muñoz, S. L., & Koskella, B. (2014). Bacteria-Phage interactions in natural environments. In *Advances in Applied Microbiology* (Vol. 89, pp. 135–183). Academic Press. https://doi.org/10.1016/B978-0-12-800259-9.00004-4

Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R., & Wilson, D. J. (2018). Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research*, *46*(22), 1–11. https://doi.org/10.1093/nar/gky783

Didelot, X., & Wilson, D. J. (2015). ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Computational Biology*, *11*(2), e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Dolejska, M., & Papagiannitsis, C. C. (2018). Plasmid-mediated resistance is going wild. In *Plasmid* (Vol. 99, pp. 99–111). Academic Press Inc. https://doi.org/10.1016/j.plasmid.2018.09.010

Duchene, S., Lemey, P., Stadler, T., Ho, S. Y. W., Duchene, D. A., Dhanasekaran, V., & Baele, G. (2020). Bayesian evaluation of temporal signal in measurably evolving populations. *Molecular Biology and Evolution*, *37*(11), 3363–3379. https://doi.org/10.1093/molbev/msaa163

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Eldholm, V., & Balloux, F. (2016). Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. In *Trends in Microbiology* (Vol. 24, Issue 8, pp. 637–648). Elsevier Current Trends. https://doi.org/10.1016/j.tim.2016.03.007

Emerson, J. B., Roux, S., Brum, J. R., Bolduc, B., Woodcroft, B. J., Jang, H. Bin, Singleton, C. M., Solden, L. M., Naas, A. E., Boyd, J. A., Hodgkins, S. B., Wilson, R. M., Trubl, G., Li, C., Frolking, S., Pope, P. B., Wrighton, K. C., Crill, P. M., Chanton, J. P., … Sullivan, M. B. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. In *Nature Microbiology* (Vol. 3, Issue 8, pp. 870–880). Nature Publishing Group. https://doi.org/10.1038/s41564-018-0190-y

Engelberg-Kulka, H., Hazan, R., & Amitai, S. (2005). mazEF: A chromosomal toxin-antitoxin module that triggers programmed cell death in bacteria. In *Journal of Cell Science* (Vol. 118, Issue 19, pp. 4327–4332). The Company of Biologists. https://doi.org/10.1242/jcs.02619

Escudero, J. A., Loot, C., Nivina, A., & Mazel, D. (2015). The Integron: Adaptation On Demand. *Microbiology Spectrum*, *3*(2). https://doi.org/10.1128/microbiolspec.mdna3-0019-2014

Feng, Y., Liu, L., McNally, A., & Zong, Z. (2018). Coexistence of two blaNDM-5 genes on an IncF plasmid as revealed by nanopore sequencing. *Antimicrobial Agents and Chemotherapy*, *62*(5). https://doi.org/10.1128/AAC.00110-18

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. In *Physics Reports* (Vol. 659, pp. 1–44). North-Holland. https://doi.org/10.1016/j.physrep.2016.09.002

Frankland, P. (1895). Ueber das Verhalten des Typhusbacillus und des Bacillus coli communis im Trinkwasser. *Zeitschrift Für Hygiene Und Infektionskrankheiten*, *19*(1), 393–407. https://doi.org/10.1007/BF02216789

Fred, A. L. N., & Jain, A. K. (2003). Robust data clustering. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, *2*, II-128-II–133. https://doi.org/10.1109/cvpr.2003.1211462

Friehs, K. (2004). Plasmid copy number and plasmid stability. In *Advances in biochemical engineering/biotechnology* (Vol. 86, pp. 47–82). Springer, Berlin, Heidelberg. https://doi.org/10.1007/b12440

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722–732. https://doi.org/10.1038/nrmicro1235

Garcillán-Barcia, M. P., & de la Cruz, F. (2013). Ordering the bestiary of genetic elements transmissible by conjugation. *Mobile Genetic Elements*, *3*(1), e24263. https://doi.org/10.4161/mge.24263

Garcillán-Barcia, M. P., Francia, M. V., & De La Cruz, F. (2009). The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiology Reviews*, *33*(3), 657–687. https://doi.org/10.1111/j.1574-6976.2009.00168.x

Garcillán-Barcia, M. P., Redondo-Salvo, S., Vielva, L., & de la Cruz, F. (2020). MOBscan: Automated Annotation of MOB Relaxases. In *Methods in Molecular Biology* (Vol. 2075, pp. 295–308). Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9877-7_21

Gerdes, K., Bech, F. W., Jørgensen, S. T., Løbner-Olesen, A., Rasmussen, P. B., Atlung, T., Boe, L., Karlstrom, O., Molin, S., & Meyenburg, K. von. (1986). Mechanism of postsegregational killing by the hok gene product of the parB system of plasmid R1 and its homology with the relF gene product of the E. coli relB operon. *The EMBO Journal*, *5*(8), 2023–2029. https://doi.org/10.1002/J.1460-2075.1986.TB04459.X

Ghigo, J. M. (2001). Natural conjugative plasmids induce bacterial biofilm development. *Nature*, *412*(6845), 442–445. https://doi.org/10.1038/35086581

Ghosh, S. K., Hajra, S., Paek, A., & Jayaram, M. (2006). Mechanisms for chromosome and plasmid segregation. In *Annual Review of Biochemistry* (Vol. 75, pp. 211–241). Annual Reviews. https://doi.org/10.1146/annurev.biochem.75.101304.124037

González, L. J., Bahr, G., Nakashige, T. G., Nolan, E. M., Bonomo, R. A., & Vila, A. J. (2016). Membrane anchoring stabilizes and favors secretion of New Delhi metallo-β-lactamase. *Nature Chemical Biology*, *12*(7), 516–522. https://doi.org/10.1038/nchembio.2083

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(1), 81–91. https://doi.org/10.1099/ijs.0.64483-0

Götz, A., Pukall, R., Smit, E., Tietze, E., Prager, R., Tschäpe, H., Van Elsas, J. D., & Smalla, K. (1996). Detection and characterization of broad-host-range plasmids in environmental bacteria by PCR. *Applied and Environmental Microbiology*, *62*(7), 2621–2628. https://doi.org/10.1128/aem.62.7.2621-2628.1996

Griffith, F. (1928). The Significance of Pneumococcal Types. *Journal of Hygiene*, *27*(2), 113–159. https://doi.org/10.1017/S0022172400031879

Grohmann, E., Christie, P. J., Waksman, G., & Backert, S. (2018). Type IV secretion in Gram-negative and Gram-positive bacteria. In *Molecular Microbiology* (Vol. 107, Issue 4, pp. 455–471). John Wiley & Sons, Ltd. https://doi.org/10.1111/mmi.13896

Groth, A. C., & Calos, M. P. (2004). Phage integrases: Biology and applications. In *Journal of Molecular Biology* (Vol. 335, Issue 3, pp. 667–678). Academic Press. https://doi.org/10.1016/j.jmb.2003.09.082

Gu, W., Miller, S., & Chiu, C. Y. (2019). Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. *Annual Review of Pathology: Mechanisms of Disease*, *14*, 319–338. https://doi.org/10.1146/annurev-pathmechdis-012418-012751

Guglielmini, J., Néron, B., Abby, S. S., Garcillán-Barcia, M. P., La Cruz, D. F., & 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 Research*, *42*(9), 5715–5727. https://doi.org/10.1093/nar/gku194

Hacker, J., & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. In *Annual Review*

of Microbiology (Vol. 54, pp. 641–679). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA. https://doi.org/10.1146/annurev.micro.54.1.641

Halary, S., Leigh, J. W., Cheaib, B., Lopez, P., & Bapteste, E. (2010). Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences*, *107*(1), 127–132. https://doi.org/10.1073/pnas.0908978107

Hall, J. P. J., Brockhurst, M. A., & Harrison, E. (2017). Sampling the mobile gene pool: Innovation via horizontal gene transfer in bacteria. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 372, Issue 1735). The Royal Society. https://doi.org/10.1098/rstb.2016.0424

Hankin, E. H. (1896). L'action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera. *Ann Inst Pasteur*, *10*(11).

Harmer, C. J., & Hall, R. M. (2019). An analysis of the IS6/IS26 family of insertion sequences: Is it a single family? *Microbial Genomics*, *5*(9). https://doi.org/10.1099/mgen.0.000291

Harmer, C. J., Moran, R. A., & Hall, R. M. (2014). Movement of IS26-Associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS26 and preferentially inserts adjacent to another IS26. *MBio*, *5*(5). https://doi.org/10.1128/mBio.01801-14

Harmer, C. J., Pong, C. H., & Hall, R. M. (2020). Structures bounded by directly-oriented members of the IS26 family are pseudo-compound transposons. In *Plasmid* (Vol. 111, p. 102530). Academic Press Inc. https://doi.org/10.1016/j.plasmid.2020.102530

Hatfull, G. F. (2008). Bacteriophage genomics. In *Current Opinion in Microbiology* (Vol. 11, Issue 5, pp. 447–453). Elsevier Current Trends. https://doi.org/10.1016/j.mib.2008.09.004

Hazan, R., Sat, B., Reches, M., & Engelberg-Kulka, H. (2001). Postsegregational killing mediated by the P1 phage "addiction module" phd-doc requires the Escherichia coli programmed cell death system mazEF. *Journal of Bacteriology*, *183*(6), 2046–2050. https://doi.org/10.1128/JB.183.6.2046-2050.2001

He, D. D., Zhao, S. Y., Wu, H., Hu, G. Z., Zhao, J. F., Zong, Z. Y., & Pan, Y. S. (2019). Antimicrobial resistance-encoding plasmid clusters with heterogeneous MDR regions driven by IS26 in a single Escherichia coli isolate. *Journal of Antimicrobial Chemotherapy*, *74*(6), 1511–1516. https://doi.org/10.1093/jac/dkz044

He, S., Hickman, A. B., Varani, A. M., Siguier, P., Chandler, M., Dekker, J. P., & Dyda, F. (2015). Insertion sequence IS26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. *MBio*, *6*(3), 1–14. https://doi.org/10.1128/mBio.00762-15

Hickman, A. B., & Dyda, F. (2015). Mechanisms of DNA Transposition. In *Mobile DNA III* (pp. 531–553). https://doi.org/10.1128/microbiolspec.mdna3-0034-2014

Hickman, A. B., & Dyda, F. (2016). DNA Transposition at Work. In *Chemical Reviews* (Vol. 116, Issue 20, pp. 12758–12784). American Chemical Society. https://doi.org/10.1021/acs.chemrev.6b00003

Hric, D., Darst, R. K., & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *90*(6), 062805. https://doi.org/10.1103/PhysRevE.90.062805

Hu, B., Khara, P., & Christie, P. J. (2019). Structural bases for F plasmid conjugation and F pilus biogenesis in Escherichia coli. *Proceedings of the National Academy of Sciences*, *116*(28), 14222–14227. https://doi.org/10.1073/pnas.1904428116

Hu, H., Hu, Y., Pan, Y., Liang, H., Wang, H., Wang, X., Hao, Q., Yang, X., Yang, X., Xiao, X., Luan, C., Yang, Y., Cui, Y., Yang, R., Gao, G. F., Song, Y., & Zhu, B. (2012). Novel plasmid and its variant harboring both a bla(NDM-1) gene and type IV secretion system in clinical isolates of Acinetobacter lwoffii. *Antimicrobial Agents and Chemotherapy*, *56*(4), 1698–1702. https://doi.org/10.1128/AAC.06199-11

Huang, H., McGarvey, P. B., Suzek, B. E., Mazumder, R., Zhang, J., Chen, Y., & Wu, C. H. (2011). A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics*, *27*(8), 1190–1191. https://doi.org/10.1093/bioinformatics/btr101

Huang, T. W., Chen, T. L., Chen, Y. T., Lauderdale, T. L., Liao, T. L., Lee, Y. T., Chen, C. P., Liu, Y. M., Lin, A. C., Chang, Y. H., Wu, K. M., Kirby, R., Lai, J. F., Tan, M. C., Siu, L. K., Chang, C. M., Fung, C. P., & Tsai, S. F. (2013). Copy Number Change of the NDM-1 Sequence in a Multidrug-Resistant

Klebsiella pneumoniae Clinical Isolate. *PLoS ONE*, *8*(4), 1–12. https://doi.org/10.1371/journal.pone.0062774

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Hülter, N., Ilhan, J., Wein, T., Kadibalban, A. S., Hammerschmidt, K., & Dagan, T. (2017). An evolutionary perspective on plasmid lifestyle modes. In *Current Opinion in Microbiology* (Vol. 38, pp. 74–80). Elsevier Ltd. https://doi.org/10.1016/j.mib.2017.05.001

Husnik, F., & McCutcheon, J. P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. In *Nature Reviews Microbiology* (Vol. 16, Issue 2, pp. 67–79). Nature Publishing Group. https://doi.org/10.1038/nrmicro.2017.137

Ilyina, T. S. (2012). Mobile ISCR elements: Structure, functions, and role in emergence, increase, and spread of blocks of bacterial multiple antibiotic resistance genes. In *Molecular Genetics, Microbiology and Virology* (Vol. 27, Issue 4, pp. 135–146). Springer. https://doi.org/10.3103/S0891416812040040

Jacob, F., & Wollman, E. L. (1961). Sexuality and the genetics of bacteria. *Sexuality and the Genetics of Bacteria.*

Jain, A., & Srivastava, P. (2013). Broad host range plasmids. *FEMS Microbiology Letters*, *348*(2), 87–96. https://doi.org/10.1111/1574-6968.12241

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 1–11. https://doi.org/10.1186/s13059-016-1103-0

Jensen, L. B., Garcia-Migura, L., Valenzuela, A. J. S., Løhr, M., Hasman, H., & Aarestrup, F. M. (2010). A classification system for plasmids from enterococci and other Gram-positive bacteria. *Journal of Microbiological Methods*, *80*(1), 25–43. https://doi.org/10.1016/j.mimet.2009.10.012

Jesus, T. F., Ribeiro-Gonçalves, B., Silva, D. N., Bortolaia, V., Ramirez, M., & Carriço, J. A. (2019). Plasmid ATLAS: Plasmid visual analytics and identification in high-Throughput sequencing data. *Nucleic Acids Research*, *47*(D1), D188–D194. https://doi.org/10.1093/nar/gky1073

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of Clinical Microbiology*, *52*(5), 1501–1510. https://doi.org/10.1128/JCM.03617-13

Johnson, C. M., & Grossman, A. D. (2015). Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annual Review of Genetics*, *49*, 577–601. https://doi.org/10.1146/annurev-genet-112414-055018

Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Research*, *3*. https://doi.org/10.12688/wellcomeopenres.14826.1

Jones, L. S., Toleman, M. A., Weeks, J. L., Howe, R. A., Walsh, T. R., & Kumarasamy, K. K. (2014). Plasmid carriage of blaNDM-1in clinical Acinetobacter baumannii isolates from India. *Antimicrobial Agents and Chemotherapy*, *58*(7), 4211–4213. https://doi.org/10.1128/AAC.02500-14

Jordan, E., Saedler, H., & Starlinger, P. (1968). Oo and strong-polar mutations in the gal operon and insertions. *MGG Molecular & General Genetics*, *102*(4), 353–363. https://doi.org/10.1007/BF00433726

Karp, R. M. (1972). Reducibility among Combinatorial Problems. In *Complexity of Computer Computations* (pp. 85–103). Springer US. https://doi.org/10.1007/978-1-4684-2001-2_9

Khan, S. A. (1997). Rolling-circle replication of bacterial plasmids. *Microbiology and Molecular Biology Reviews*, *61*(4), 442–455. https://doi.org/10.1128/mmbr.61.4.442-455.1997

Khan, S. A. (2005). Plasmid rolling-circle replication: Highlights of two decades of research. *Plasmid*, *53*(2), 126–136. https://doi.org/10.1016/j.plasmid.2004.12.008

Krahn, T., Wibberg, D., Maus, I., Winkler, A., Bontron, S., Sczyrba, A., Nordmann, P., Pühler, A., Poirel,

L., & Schlüter, A. (2016). Intraspecies transfer of the chromosomal Acinetobacter baumannii blaNDM-1 carbapenemase gene. *Antimicrobial Agents and Chemotherapy*, *60*(5), 3032–3040. https://doi.org/10.1128/AAC.00124-16

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, *46*(6), e35. https://doi.org/10.1093/nar/gkx1321

Kulakauskas, S., Lubys, A., & Ehrlich, S. D. (1995). DNA restriction-modification systems mediate plasmid maintenance. *Journal of Bacteriology*, *177*(12), 3451–3454. https://doi.org/10.1128/jb.177.12.3451-3454.1995

Kumarasamy, K. K., Toleman, M. A., Walsh, T. R., Bagaria, J., Butt, F., Balakrishnan, R., Chaudhary, U., Doumith, M., Giske, C. G., Irfan, S., Krishnan, P., Kumar, A. V., Maharjan, S., Mushtaq, S., Noorie, T., Paterson, D. L., Pearson, A., Perry, C., Pike, R., … Woodford, N. (2010). Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. *The Lancet Infectious Diseases*, *10*(9), 597–602. https://doi.org/10.1016/S1473-3099(10)70143-2

Kusumoto, M., & Hayashi, T. (2019). Bacterial Transposable Elements and IS-Excision Enhancer (IEE). In *DNA Traffic in the Environment* (pp. 197–213). Springer Singapore. https://doi.org/10.1007/978-981-13-3411-5_8

Lam, M. M. C., Wick, R. R., Watts, S. C., Cerdeira, L. T., Wyres, K. L., & Holt, K. E. (2021). A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nature Communications*, *12*(1), 1–16. https://doi.org/10.1038/s41467-021-24448-3

Lancichinetti, A., Radicchi, F., Ramasco, J. J., & Fortunato, S. (2011). Finding statistically significant communities in networks. *PLoS ONE*, *6*(4), e18961. https://doi.org/10.1371/journal.pone.0018961

Lederberg, J. (1952). Cell genetics and hereditary symbiosis. *Physiological Reviews*, *32*(4), 403–430. https://doi.org/10.1152/physrev.1952.32.4.403

Lederberg, J., & Tatum, E. L. (1946). Gene recombination in Escherichia coli. *Nature*, *158*(4016).

Lerminiaux, N. A., & Cameron, A. D. S. (2019). Horizontal transfer of antibiotic resistance genes in clinical environments. *Canadian Journal of Microbiology*, *65*(1), 34–44. https://doi.org/10.1139/cjm-2018-0275

Li, J., Lan, R., Xiong, Y., Ye, C., Yuan, M., Liu, X., Chen, X., Yu, D., Liu, B., Lin, W., Bai, X., Wang, Y., Sun, Q., Wang, Y., Zhao, H., Meng, Q., Chen, Q., Zhao, A., & Xu, J. (2014). Sequential isolation in a patient of Raoultella planticola and Escherichia coli bearing a novel IS CR1 element carrying blaNDM-1. *PLoS ONE*, *9*(3), e89893. https://doi.org/10.1371/journal.pone.0089893

Lilly, J., & Camps, M. (2015). Mechanisms of Theta Plasmid Replication. *Microbiology Spectrum*, *3*(1). https://doi.org/10.1128/microbiolspec.plas-0029-2014

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L., & Bairoch, A. (2009). HAMAP: A database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Research*, *37*(SUPPL. 1), D471–D478. https://doi.org/10.1093/nar/gkn661

Liu, Y. Y., Wang, Y., Walsh, T. R., Yi, L. X., Zhang, R., Spencer, J., Doi, Y., Tian, G., Dong, B., Huang, X., Yu, L. F., Gu, D., Ren, H., Chen, X., Lv, L., He, D., Zhou, H., Liang, Z., Liu, J. H., & Shen, J. (2016). Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: A microbiological and molecular biological study. *The Lancet Infectious Diseases*, *16*(2), 161–168. https://doi.org/10.1016/S1473-3099(15)00424-7

Loftie-Eaton, W., & Rawlings, D. E. (2012). Diversity, biology and evolution of IncQ-family plasmids. In *Plasmid* (Vol. 67, Issue 1, pp. 15–34). Academic Press. https://doi.org/10.1016/j.plasmid.2011.10.001

Lozano, C., García-Migura, L., Aspiroz, C., Zarazaga, M., Torres, C., & 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. *Applied and Environmental Microbiology*, *78*(16), 5948–5955. https://doi.org/10.1128/AEM.00870-12

Lu, B., & Leong, H. W. (2016). Computational methods for predicting genomic islands in microbial genomes. In *Computational and Structural Biotechnology Journal* (Vol. 14, pp. 200–206). Elsevier. https://doi.org/10.1016/j.csbj.2016.05.001

Luhmann, N., Holley, G., & Achtman, M. (2021). BlastFrost: fast querying of 100,000s of bacterial genomes in Bifrost graphs. *Genome Biology*, *22*(1), 1–15. https://doi.org/10.1186/s13059-020-02237-3

Lynch, T., Chen, L., Peirano, G., Gregson, D. B., Church, D. L., Conly, J., Kreiswirth, B. N., & Pitout, J. D. (2016). Molecular evolution of a Klebsiella Pneumoniae ST278 isolate harboring blaNDM-7 and involved in nosocomial transmission. *Journal of Infectious Diseases*, *214*(5), 798–806. https://doi.org/10.1093/infdis/jiw240

Ma, B., Tromp, J., & Li, M. (2002). PatternHunter: Faster and more sensitive homology search. *Bioinformatics*, *18*(3), 440–445. https://doi.org/10.1093/bioinformatics/18.3.440

Malmir, S., Bahreinian, M., Zahiri Yeganeh, S., Mirnejad, R., Moosazadeh Moghaddam, M., & Saberi, F. (2018). Molecular Mechanisms of Resistance to Conventional Antibiotics in Bacteria. *International Journal of Medical Reviews*, *5*(3), 118–129. https://doi.org/10.29252/ijmr-050305

Marmur, J., Rownd, R., Falkow, S., Baron, L. S., Schildkraut, C., & Doty, P. (1961). The nature of intergeneric episomal infection. *Proceedings of the National Academy of Sciences*, *47*(7), 972–979. https://doi.org/10.1073/pnas.47.7.972

Mc Garth, S., & Sinderen, D. Van (Eds.). (2007). *Bacteriophage Genetics and Molecular Biology*. Caister Academic Press.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., … Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, *57*(7), 3348–3357. https://doi.org/10.1128/AAC.00419-13

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), 344–355. https://doi.org/10.1073/pnas.36.6.344

Mell, J. C., & Redfield, R. J. (2014). Natural competence and the evolution of DNA uptake specificity. In *Journal of Bacteriology* (Vol. 196, Issue 8, pp. 1471–1483). American Society for Microbiology. https://doi.org/10.1128/JB.01293-13

Munoz-Price, L. S., Poirel, L., Bonomo, R. A., Schwaber, M. J., Daikos, G. L., Cormican, M., Cornaglia, G., Garau, J., Gniadkowski, M., Hayden, M. K., Kumarasamy, K., Livermore, D. M., Maya, J. J., Nordmann, P., Patel, J. B., Paterson, D. L., Pitout, J., Villegas, M. V., Wang, H., … Quinn, J. P. (2013). Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. In *The Lancet Infectious Diseases* (Vol. 13, Issue 9, pp. 785–796). Elsevier. https://doi.org/10.1016/S1473-3099(13)70190-7

Nishida, H. (2012). Comparative Analyses of Base Compositions, DNA Sizes, and Dinucleotide Frequency Profiles in Archaeal and Bacterial Chromosomes and Plasmids. *International Journal of Evolutionary Biology*, *2012*, 1–5. https://doi.org/10.1155/2012/342482

Norberg, P., Bergström, M., Jethava, V., Dubhashi, D., & Hermansson, M. (2011). The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. *Nature Communications*, *2*(1). https://doi.org/10.1038/ncomms1267

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., German Grehov, O. G., Kandrov, D., Rasputin, K., Syabro, M., & Tleukenov, T. (2012). Unipro UGENE: A unified bioinformatics toolkit. In *Bioinformatics* (Vol. 28, Issue 8, pp. 1166–1167). Oxford Academic. https://doi.org/10.1093/bioinformatics/bts091

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2019). *vegan:*

Community Ecology Package (R package version 2.5-6.). Article R package version 2.5-6. https://cran.r-project.org/package=vegan

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0997-x

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., Crook, D., Walker, A. S., Woodford, N., Anjum, M. F., & Stoesser, N. (2017a). Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid*, *91*, 42–52. https://doi.org/10.1016/j.plasmid.2017.03.002

Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., Crook, D., Walker, A. S., Woodford, N., Anjum, M. F., & Stoesser, N. (2017b). A curated dataset of complete Enterobacteriaceae plasmids compiled from the NCBI nucleotide database. *Data in Brief*, *12*, 423–426. https://doi.org/10.1016/j.dib.2017.04.024

Orlek, A., Stoesser, N., Anjum, M. F., Doumith, M., Ellington, M. J., Peto, T., Crook, D., Woodford, N., Sarah Walker, A., Phan, H., & Sheppard, A. E. (2017). Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. In *Frontiers in Microbiology* (Vol. 8, Issue FEB, pp. 1–10). https://doi.org/10.3389/fmicb.2017.00182

Padgham, M., & Sumner, M. D. (2020). *geodist: Fast, Dependency-Free Geodesic Distance Calculations.* (R package version 0.0.4). https://cran.r-project.org/package=geodist

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Park, C., & Zhang, J. (2012). High expression hampers horizontal gene transfer. *Genome Biology and Evolution*, *4*(4), 523–532. https://doi.org/10.1093/gbe/evs030

Partridge, S. R., & Iredell, J. R. (2012). Genetic Contexts of bla NDM-1. In *Antimicrobial Agents and Chemotherapy* (Vol. 56, Issue 11, pp. 6065–6067). https://doi.org/10.1128/AAC.00117-12

Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clinical Microbiology Reviews*, *31*(4), 1–61. https://doi.org/10.1128/CMR.00088-17

Peixoto, T. P. (2014). The graph-tool python library. *Figshare*. https://doi.org/10.6084/m9.figshare.1164194

Penadés, J. R., Chen, J., Quiles-Puchalt, N., Carpena, N., & Novick, R. P. (2015). Bacteriophage-mediated spread of bacterial virulence genes. In *Current Opinion in Microbiology* (Vol. 23, pp. 171–178). Elsevier Ltd. https://doi.org/10.1016/j.mib.2014.11.019

Petersen, J. (2011). Phylogeny and compatibility: Plasmid classification in the genomics era. In *Archives of Microbiology* (Vol. 193, Issue 5, pp. 313–321). Springer. https://doi.org/10.1007/s00203-011-0686-9

Phelan, J. E., O'Sullivan, D. M., Machado, D., Ramos, J., Oppong, Y. E. A., Campino, S., O'Grady, J., McNerney, R., Hibberd, M. L., Viveiros, M., Huggett, J. F., & Clark, T. G. (2019). Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Medicine*, *11*(1), 1–7. https://doi.org/10.1186/s13073-019-0650-x

Pinto, U. M., Pappas, K. M., & Winans, S. C. (2012). The ABCs of plasmid replication and segregation. In *Nature Reviews Microbiology* (Vol. 10, Issue 11, pp. 755–765). Nature Publishing Group. https://doi.org/10.1038/nrmicro2882

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC - Open Research Online. *R News*, *6*(1), 7–11. http://oro.open.ac.uk/22547/

Poirel, L., Bonnin, R. A., Boulanger, A., Schrenzel, J., Kaase, M., & Nordmann, P. (2012). Tn125-related acquisition of blaNDM-like genes in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*, *56*(2), 1087–1089. https://doi.org/10.1128/AAC.05620-11

Poirel, L., Bonnin, R. A., & Nordmann, P. (2011). Analysis of the resistome of a multidrug-resistant NDM-1-producing Escherichia coli strain by high-throughput genome sequencing. *Antimicrobial Agents and Chemotherapy*, *55*(9), 4224–4229. https://doi.org/10.1128/AAC.00165-11

Poirel, L., Carrër, A., Pitout, J. D., & Nordmann, P. (2009). Integron mobilization unit as a source of mobility of antibiotic resistance genes. *Antimicrobial Agents and Chemotherapy*, *53*(6), 2492–2498. https://doi.org/10.1128/AAC.00033-09

Poirel, L., Dortet, L., Bernabeu, S., & Nordmann, P. (2011). Genetic features of blaNDM-1-positive Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy*, *55*(11), 5403–5407. https://doi.org/10.1128/AAC.00585-11

Popowska, M., & Krawczyk-Balska, A. (2013). Broad-host-range IncP-1 plasmids and their resistance potential. In *Frontiers in Microbiology* (Vol. 4, Issue MAR, p. 44). Frontiers. https://doi.org/10.3389/fmicb.2013.00044

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. V., Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., … Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503–509. https://doi.org/10.1038/s41586-018-0124-0

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(SUPPL. 1), D61–D65. https://doi.org/10.1093/nar/gkl842

Quainoo, S., Coolen, J. P. M., van Hijum, S. A. F. T., Huynen, M. A., Melchers, W. J. G., van Schaik, W., & Wertheim, H. F. L. (2017). Whole-genome sequencing of bacterial pathogens: The future of nosocomial outbreak analysis. In *Clinical Microbiology Reviews* (Vol. 30, Issue 4, pp. 1015–1063). American Society for Microbiology. https://doi.org/10.1128/CMR.00016-17

Rahman, M., Prasad, K. N., Gupta, S., Singh, S., Singh, A., Pathak, A., Gupta, K. K., Ahmad, S., & Gonzalez-Zorn, B. (2018). Prevalence and Molecular Characterization of New Delhi Metallo-Beta-Lactamases in Multidrug-Resistant Pseudomonas aeruginosa and Acinetobacter baumannii from India. *Microbial Drug Resistance*, *24*(6), 792–798. https://doi.org/10.1089/mdr.2017.0078

Rambaut, A., Lam, T. T., Carvalho, L. M., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1), 1–7. https://doi.org/10.1093/ve/vew007

Rasheed, J. K., Kitchel, B., Zhu, W., Anderson, K. F., Clark, N. C., Ferraro, M. J., Savard, P., Humphries, R. M., Kallen, A. J., & Limbago, B. M. (2013). New Delhi metallo-β-lactamase-producing enterobacteriaceae, United States. *Emerging Infectious Diseases*, *19*(6), 870–878. https://doi.org/10.3201/eid1906.121515

Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V., & Ravel, J. (2008). The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology*, *190*(20), 6881–6893. https://doi.org/10.1128/JB.00619-08

Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P., & de la Cruz, F. (2020). Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nature Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-17278-2

Reinert, G., Chew, D., Sun, F., & Waterman, M. S. (2009). Alignment-free sequence comparison (I): Statistics and power. *Journal of Computational Biology*, *16*(12), 1615–1634. https://doi.org/10.1089/cmb.2009.0198

Ren, J., Bai, X., Lu, Y. Y., Tang, K., Wang, Y., Reinert, G., & Sun, F. (2018). Alignment-Free Sequence Analysis and Applications. *Annual Review of Biomedical Data Science*, *1*(1), 93–114. https://doi.org/10.1146/annurev-biodatasci-080917-013431

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. In *Genomics, Proteomics and Bioinformatics* (Vol. 13, Issue 5, pp. 278–289). Elsevier. https://doi.org/10.1016/j.gpb.2015.08.002

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: A review and a practical guide. *Molecular Ecology*, *25*(9), 1911–1924. https://doi.org/10.1111/mec.13586

Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C., & Gasbarrini, A. (2017). Proteobacteria: A common factor in human diseases. In *BioMed Research International* (Vol. 2017, p. 9351507). Hindawi

Limited. https://doi.org/10.1155/2017/9351507

Roach, D., Waalkes, A., Abanto, J., Zunt, J., Cucho, C., Soria, J., & Salipante, S. J. (2020). Whole genome sequencing of peruvian Klebsiella pneumoniae identifies novel plasmid vectors bearing carbapenem resistance gene NDM-1. *Open Forum Infectious Diseases*, *7*(8). https://doi.org/10.1093/ofid/ofaa266

Roberts, A. P., Chandler, M., Courvalin, P., Guédon, G., Mullany, P., Pembroke, T., Rood, J. I., Jeffery Smith, C., Summers, A. O., Tsuda, M., & Berg, D. E. (2008). Revised nomenclature for transposable genetic elements. *Plasmid*, *60*(3), 167–173. https://doi.org/10.1016/j.plasmid.2008.08.001

Robertson, J., & Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics*, *4*(8). https://doi.org/10.1099/mgen.0.000206

Roca, I., Mosqueda, N., Altun, B., Espinal, P., Akova, M., & Vila, J. (2014). Molecular characterization of NDM-1-producing Acinetobacter pittii isolated from Turkey in 2006. In *Journal of Antimicrobial Chemotherapy* (Vol. 69, Issue 12, pp. 3437–3438). Oxford University Press. https://doi.org/10.1093/jac/dku306

Rocha, E. P. C., & Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. In *Trends in Genetics* (Vol. 18, Issue 6, pp. 291–294). Elsevier. https://doi.org/10.1016/S0168-9525(02)02690-2

Rodriguez-Mozaz, S., Chamorro, S., Marti, E., Huerta, B., Gros, M., Sànchez-Melsió, A., Borrego, C. M., Barceló, D., & Balcázar, J. L. (2015). Occurrence of antibiotics and antibiotic resistance genes in hospital and urban wastewaters and their impact on the receiving river. *Water Research*, *69*, 234–242. https://doi.org/10.1016/j.watres.2014.11.021

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Rozwandowicz, M., Brouwer, M. S. M., Zomer, A. L., Bossers, A., Harders, F., Mevius, D. J., Wagenaar, J. A., & Hordijk, J. (2017). Plasmids of distinct IncK lineages show compatible phenotypes. *Antimicrobial Agents and Chemotherapy*, *61*(3), e01954-16. https://doi.org/10.1128/AAC.01954-16

Ruiz-Masó, J. A., MachóN, C., Bordanaba-Ruiseco, L., Espinosa, M., Coll, M., & Del Solar, G. (2015). Plasmid Rolling-Circle Replication. *Microbiology Spectrum*, *3*(1). https://doi.org/10.1128/microbiolspec.plas-0035-2014

Sahl, J. W., Del Franco, M., Pournaras, S., Colman, R. E., Karah, N., Dijkshoorn, L., & Zarrilli, R. (2015). Phylogenetic and genomic diversity in isolates from the globally distributed Acinetobacter baumannii ST25 lineage. *Scientific Reports*, *5*. https://doi.org/10.1038/srep15188

Salje, J. (2010). Plasmid segregation: How to survive as an extra piece of DNA. In *Critical Reviews in Biochemistry and Molecular Biology* (Vol. 45, Issue 4, pp. 296–317). Taylor & Francis. https://doi.org/10.3109/10409238.2010.494657

San Millan, A., & MacLean, R. C. (2017). Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiology Spectrum*, *5*(5). https://doi.org/10.1128/microbiolspec.mtbp-0016-2017

Schnetz, K., & Rak, B. (1992). IS5: A mobile enhancer of transcription in Escherichia coli. *Proceedings of the National Academy of Sciences*, *89*(4), 1244–1248. https://doi.org/10.1073/pnas.89.4.1244

Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., & Goesmann, A. (2020). Platon: Identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial Genomics*, *6*(10), 1–12. https://doi.org/10.1099/mgen.0.000398

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sekizuka, T., Matsui, M., Yamane, K., Takeuchi, F., Ohnishi, M., Hishinuma, A., Arakawa, Y., & Kuroda, M. (2011). Complete sequencing of the blaNDM-1-positive IncA/C plasmid from Escherichia coli ST38 isolate suggests a possible origin from plant pathogens. *PLoS ONE*, *6*(9), e25334. https://doi.org/10.1371/journal.pone.0025334

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular

interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shapiro, J. A. (1969). Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. *Journal of Molecular Biology*, *40*(1), 93–105. https://doi.org/10.1016/0022-2836(69)90298-8

Shapiro, J. A. (1983). *Mobile genetic elements*. Academic Press Inc.

Sheppard, A. E., Stoesser, N., Wilson, D. J., Sebra, R., Kasarskis, A., Anson, L. W., Giess, A., Pankhurst, L. J., Vaughan, A., Grim, C. J., Cox, H. L., Yeh, A. J., Sifri, C. D., Walker, A. S., Peto, T. E., Crook, D. W., & Mathers, A. J. (2016). Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC. *Antimicrobial Agents and Chemotherapy*, *60*(6), 3767–3778. https://doi.org/10.1128/AAC.00464-16

Sheppard, R. J., Beddis, A. E., & Barraclough, T. G. (2020). The role of hosts, plasmids and environment in determining plasmid transfer rates: A meta-analysis. *Plasmid*, *108*, 102489. https://doi.org/10.1016/j.plasmid.2020.102489

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

Shintani, M., & Suzuki, H. (2019). Plasmids and Their Hosts. In *DNA Traffic in the Environment* (pp. 109–133). Springer Singapore. https://doi.org/10.1007/978-981-13-3411-5_6

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Siguier, P., Gourbeyre, E., & Chandler, M. (2014). Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiology Reviews*, *38*(5), 865–891. https://doi.org/10.1111/1574-6976.12067

Siguier, P., Gourbeyre, E., & Chandler, M. (2017). Known knowns, known unknowns and unknown unknowns in prokaryotic transposition. In *Current Opinion in Microbiology* (Vol. 38, pp. 171–180). Elsevier Current Trends. https://doi.org/10.1016/j.mib.2017.06.005

Siguier, P., Gourbeyre, E., Varani, A., Ton-hoang, B., & Chandler, M. (2015). Everyman's Guide to Bacterial Insertion Sequences. In *Mobile DNA III* (pp. 555–590). https://doi.org/10.1128/microbiolspec.mdna3-0030-2014

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, *34*(Database issue), D32–D36. https://doi.org/10.1093/nar/gkj014

Sitter, T. L., Vaughan, A. L., Schoof, M., Jackson, S. A., Glare, T. R., Cox, M. P., Fineran, P. C., Gardner, P. P., & Hurst, M. R. H. (2021). Evolution of virulence in a novel family of transmissible mega-plasmids. *Environmental Microbiology*, 1462-2920.15595. https://doi.org/10.1111/1462-2920.15595

Smillie, C. S., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C., & de la Cruz, F. (2010). Mobility of Plasmids. *Microbiology and Molecular Biology Reviews*, *74*(3), 434–452. https://doi.org/10.1128/mmbr.00020-10

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., & Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, *480*(7376), 241–244. https://doi.org/10.1038/nature10571

Snyder, L., & Snyder, L. (2013). *Molecular genetics of bacteria*. ASM Press.

Sohn, J. Il, & Nam, J. W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23–40. https://doi.org/10.1093/bib/bbw096

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: Building the web of life. In *Nature Reviews Genetics* (Vol. 16, Issue 8, pp. 472–482). Nature Publishing Group. https://doi.org/10.1038/nrg3962

Souque, C., Escudero, J. A., & Maclean, R. C. (2021). Integron activity accelerates the evolution of antibiotic resistance. *ELife*, *10*, 1–47. https://doi.org/10.7554/eLife.62474

Souvorov, A., Agarwala, R., & Lipman, D. J. (2018). SKESA: Strategic k-mer extension for scrupulous assemblies. *Genome Biology*, *19*(1), 153. https://doi.org/10.1186/s13059-018-1540-z

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stokes, H. W., & Gillings, M. R. (2011). Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. In *FEMS Microbiology Reviews* (Vol. 35, Issue 5, pp. 790–819). Narnia. https://doi.org/10.1111/j.1574-6976.2011.00273.x

Stokes, H. W., & Hall, R. M. (1989). A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular Microbiology*, *3*(12), 1669–1683. https://doi.org/10.1111/j.1365-2958.1989.tb00153.x

Stolz, A. (2014). Degradative plasmids from sphingomonads. *FEMS Microbiology Letters*, *350*(1), 9–19. https://doi.org/10.1111/1574-6968.12283

Struelens, M. J., Monnet, D. L., Magiorakos, A. P., Santos O'Connor, F., Giesecke, J., Grisold, A., Zarfel, G., Jans, B., Velinov, T., Kantardjiev, T., Alexandrou, M., Zemlickova, H., Hrabak, J., Frimodt-Møller, N., Hammerum, A. M., Maimets, M., Ivanova, M., Jalava, J., Rummukainen, M., … Woodford, N. (2010). New Delhi metallo-beta-lactamase 1-producing Enterobacteriaceae: Emergence and response in Europe. *Eurosurveillance*, *15*(46), 19716. https://doi.org/10.2807/ese.15.46.19716-en

Summers, D. K. (1996). *The biology of plasmids*. Blackwell Science Ltd.

Summers, D. K., & Sherratt, D. J. (1984). Multimerization of high copy number plasmids causes instability: Col1 encodes a determinant essential for plasmid monomerization and stability. *Cell*, *36*(4), 1097–1103. https://doi.org/10.1016/0092-8674(84)90060-6

Suzuki-Minakuchi, C., & Navarre, W. W. (2019). Xenogeneic Silencing and Horizontal Gene Transfer. In *DNA Traffic in the Environment* (pp. 1–27). Springer Singapore. https://doi.org/10.1007/978-981-13-3411-5_1

Suzuki, H., Sota, M., Brown, C. J., & Top, E. M. (2008). Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Research*, *36*(22), e147–e147. https://doi.org/10.1093/nar/gkn753

Tamminen, M., Virta, M., Fani, R., & Fondi, M. (2012). Large-scale analysis of plasmid relationships through gene-sharing networks. *Molecular Biology and Evolution*, *29*(4), 1225–1240. https://doi.org/10.1093/molbev/msr292

Tansirichaiya, S., Rahman, M. A., & Roberts, A. P. (2019). The Transposon Registry. *Mobile DNA*, *10*(1), 1–6. https://doi.org/10.1186/s13100-019-0182-3

Tatum, E. L., & Lederberg, J. (1947). Gene Recombination in the Bacterium Escherichia coli. *Journal of Bacteriology*, *53*(6), 673–684. https://doi.org/10.1128/jb.53.6.673-684.1947

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., & Tolstoy, I. (2014). RefSeq microbial genomes database: New representation and annotation strategy. *Nucleic Acids Research*, *42*(D1), D553–D559. https://doi.org/10.1093/nar/gkt1274

Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569

Taylor, A. L. (1963). Bacteriophage-induced mutation in Escherichia coli. *Proceedings of the National Academy of Sciences*, *50*(2), 1043–1051. https://doi.org/10.1073/pnas.50.6.1043

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. In *Nature Reviews Microbiology* (Vol. 3, Issue 9, pp. 711–721). Nature Publishing Group. https://doi.org/10.1038/nrmicro1234

Toleman, M. A., Bennett, P. M., & Walsh, T. R. (2006). ISCR Elements: Novel Gene-Capturing Systems of the 21st Century? *Microbiology and Molecular Biology Reviews*, *70*(2), 296–316. https://doi.org/10.1128/mmbr.00048-05

Toleman, M. A., Spencer, J., Jones, L., & Walsha, T. R. (2012). blaNDM-1 is a chimera likely constructed in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*, *56*(5), 2773–2776.

https://doi.org/10.1128/AAC.06297-11

Toleman, M. A., & Walsh, T. R. (2010). ISCR Elements Are Key Players in IncA/C Plasmid Evolution. *Antimicrobial Agents and Chemotherapy*, *54*(8), 3534. https://doi.org/10.1128/AAC.00383-10

Tolmasky, M. E., & Alonso, J. C. (Eds.). (2015). *Plasmids biology and impact in biotechnology and discovery*. AMS Press. https://doi.org/10.1017/CBO9781107415324.004

Tooke, C. L., Hinchliffe, P., Bragginton, E. C., Colenso, C. K., Hirvonen, V. H. A., Takebayashi, Y., & Spencer, J. (2019). β-Lactamases and β-Lactamase Inhibitors in the 21st Century. In *Journal of Molecular Biology* (Vol. 431, Issue 18, pp. 3472–3500). Academic Press. https://doi.org/10.1016/j.jmb.2019.04.002

Tsang, J. (2017). Bacterial plasmid addiction systems and their implications for antibiotic drug development. *Postdoc Journal*, *5*(5), 3. https://doi.org/10.14304/surya.jpr.v5n5.2

Twort, F. W. (1915). An investigation on the nature of ultra-microscopic viruses. *The Lancet*, *186*(4814), 1241–1243. https://doi.org/10.1016/S0140-6736(01)20383-3

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. In *Trends in Genetics* (Vol. 34, Issue 9, pp. 666–681). Elsevier Current Trends. https://doi.org/10.1016/j.tig.2018.05.008

Van Dorp, L., Wang, Q., Shaw, L. P., Acman, M., Brynildsrud, O. B., Eldholm, V., Wang, R., Gao, H., Yin, Y., Chen, H., Ding, C., Farrer, R. A., Didelot, X., Balloux, F., & Wang, H. (2019). Rapid phenotypic evolution in multidrug-resistant Klebsiella pneumoniae hospital outbreak strains. *Microbial Genomics*, *5*(4), 1–11. https://doi.org/10.1099/mgen.0.000263

Vielva, L., De Toro, M., Lanza, V. F., & De La Cruz, F. (2017). PLACNETw: A web-based tool for plasmid reconstruction from bacterial genomes. *Bioinformatics*, *33*(23), 3796–3798. https://doi.org/10.1093/bioinformatics/btx462

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., Savelkoul, P. H. M., & Wolffs, P. F. G. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in Microbiology*, *7*(FEB). https://doi.org/10.3389/fmicb.2016.00173

Vos, M., Hesselman, M. C., te Beek, T. A., van Passel, M. W. J., & Eyre-Walker, A. (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? In *Trends in Microbiology* (Vol. 23, Issue 10, pp. 598–605). Elsevier Current Trends. https://doi.org/10.1016/j.tim.2015.07.006

Wailan, A. M., Sartor, A. L., Zowawi, H. M., Perry, J. D., Paterson, D. L., & Sidjabat, H. E. (2015). Genetic contexts of blaNDM-1 in patients carrying multiple NDM-producing strains. *Antimicrobial Agents and Chemotherapy*, *59*(12), 7405–7410. https://doi.org/10.1128/AAC.01319-15

Wang, Q., Wang, X., Wang, J., Ouyang, P., Jin, C., Wang, R., Zhang, Y., Jin, L., Chen, H., Wang, Z., Zhang, F., Cao, B., Xie, L., Liao, K., Gu, B., Yang, C., Liu, Z., Ma, X., Jin, L., … Wang, H. (2018). Phenotypic and Genotypic Characterization of Carbapenem-resistant Enterobacteriaceae: Data from a Longitudinal Large-scale CRE Study in China (2012-2016). *Clinical Infectious Diseases*, *67*(Suppl 2), S196–S205. https://doi.org/10.1093/cid/ciy660

Wang, R., Liu, Y., Zhang, Q., Jin, L., Wang, Q., Zhang, Y., Wang, X., Hu, M., Li, L., Qi, J., Luo, Y., & Wang, H. (2018). The prevalence of colistin resistance in Escherichia coli and Klebsiella pneumoniae isolated from food animals in China: coexistence of mcr-1 and blaNDM with low fitness cost. *International Journal of Antimicrobial Agents*, *51*(5), 739–744. https://doi.org/10.1016/j.ijantimicag.2018.01.023

Wang, R., Van Dorp, L., Shaw, L. P., Bradley, P., Wang, Q., Wang, X., Jin, L., Zhang, Q., Liu, Y., Rieux, A., Dorai-Schneiders, T., Weinert, L. A., Iqbal, Z., Didelot, X., Wang, H., & Balloux, F. (2018). The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nature Communications*, *9*(1), 1–9. https://doi.org/10.1038/s41467-018-03205-z

Weber, R. E., Pietsch, M., Frühauf, A., Pfeifer, Y., Martin, M., Luft, D., Gatermann, S., Pfennigwerth, N., Kaase, M., Werner, G., & Fuchs, S. (2019). IS26-Mediated Transfer of blaNDM–1 as the Main Route of Resistance Transmission During a Polyclonal, Multispecies Outbreak in a German Hospital. *Frontiers in Microbiology*, *10*, 2817. https://doi.org/10.3389/fmicb.2019.02817

WHO. (2021). *Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2021*.

Geneva: World Health Organization;

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6), 1–22. https://doi.org/10.1371/journal.pcbi.1005595

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wu, G., Yan, Q., Jones, J. A., Tang, Y. J., Fong, S. S., & Koffas, M. A. G. (2016). Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. In *Trends in Biotechnology* (Vol. 34, Issue 8, pp. 652–664). Elsevier Current Trends. https://doi.org/10.1016/j.tibtech.2016.02.010

Wu, W., Feng, Y., Tang, G., Qiao, F., McNally, A., & Zong, Z. (2019). NDM metallo-β-lactamases and their bacterial producers in health care settings. In *Clinical Microbiology Reviews* (Vol. 32, Issue 2). https://doi.org/10.1128/CMR.00115-18

Yamashita, A., Sekizuka, T., & Kuroda, M. (2014). Characterization of antimicrobial resistance dissemination across plasmid communities classified by network analysis. *Pathogens*, *3*(2), 356–376. https://doi.org/10.3390/pathogens3020356

Yang, Q., Fang, L., Fu, Y., Du, X., Shen, Y., & Yu, Y. (2015). Dissemination of NDM-1-producing Enterobacteriaceae mediated by the IncX3-type plasmid. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0129454

Yong, D., Toleman, M. A., Giske, C. G., Cho, H. S., Sundman, K., Lee, K., & Walsh, T. R. (2009). Characterization of a new metallo-β-lactamase gene, bla NDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India. *Antimicrobial Agents and Chemotherapy*, *53*(12), 5046–5054. https://doi.org/10.1128/AAC.00774-09

Yuan, Y., & Gao, M. (2017). Jumbo bacteriophages: An overview. *Frontiers in Microbiology*, *8*(MAR), 403. https://doi.org/10.3389/fmicb.2017.00403

Zaman, S. Bin, Hussain, M. A., Nye, R., Mehta, V., Mamun, K. T., & Hossain, N. (2017). A Review on Antibiotic Resistance: Alarm Bells are Ringing. *Cureus*, *9*(6). https://doi.org/10.7759/cureus.1403

Zankari, E., Allesøe, R., Joensen, K. G., Cavaco, L. M., Lund, O., & Aarestrup, F. M. (2017). PointFinder: A novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *Journal of Antimicrobial Chemotherapy*, *72*(10), 2764–2768. https://doi.org/10.1093/jac/dkx217

Zhang, X., Feng, Y., Zhou, W., McNally, A., & Zong, Z. (2018). Cryptic transmission of ST405 Escherichia coli carrying blaNDM-4 in hospital. *Scientific Reports*, *8*(1), 1–4. https://doi.org/10.1038/s41598-017-18910-w

Zhao, Q.-Y., Zhu, J.-H., Cai, R.-M., Zheng, X.-R., Zhang, L.-J., Chang, M.-X., Lu, Y.-W., Fang, L.-X., Sun, J., & Jiang, H.-X. (2021). IS26 Is Responsible for the Evolution and Transmission of bla NDM-Harboring Plasmids in Escherichia coli of Poultry Origin in China. *MSystems*, *6*(4). https://doi.org/10.1128/msystems.00646-21

Zhao, X. (2019). BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics*, *35*(4), 671–673. https://doi.org/10.1093/bioinformatics/bty651

Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y., & Achtman, M. (2020). The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Research*, *30*(1), 138–152. https://doi.org/10.1101/gr.251678.119

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C. A., Tang, K., Dencker, T., Lau, A. K., Röhling, S., Choi, J. J., Waterman, M. S., Comin, M., Kim, S. H., Vinga, S., Almeida, J. S., Chan, C. X., James, B. T., Sun, F., Morgenstern, B., & Karlowski, W. M. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1755-7

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, *18*(1), 1–17.

Appendix A

Additional figures and tables

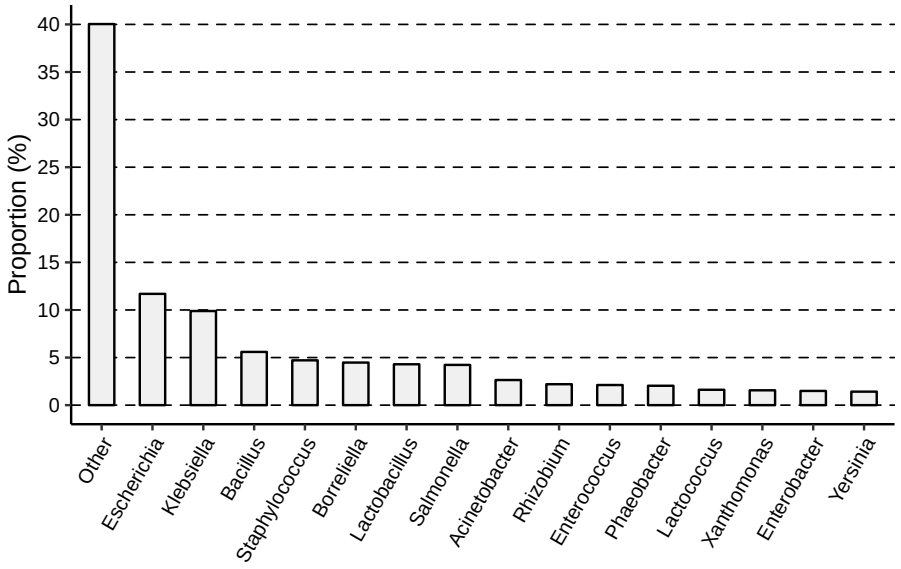


Figure A.1. Phylogenetic diversity of plasmid hosts at the genus level.

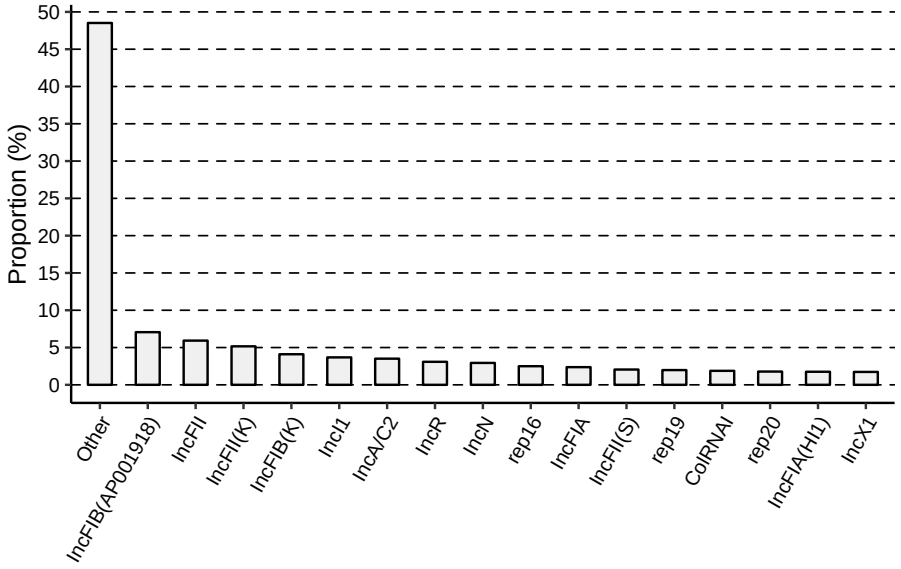


Figure A.2. Distribution of the proportion of known plasmid replicon types.

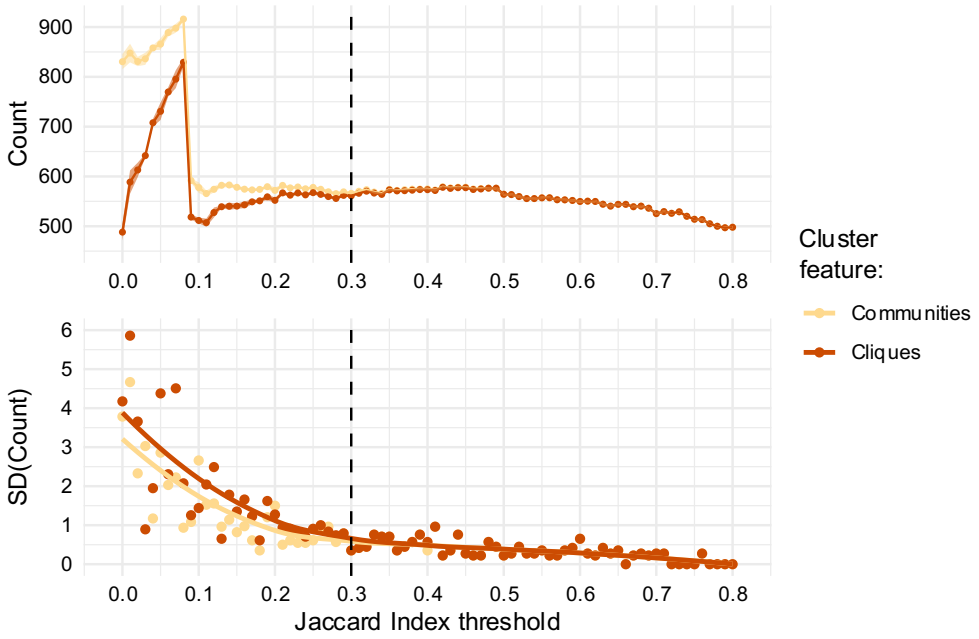


Figure A.3. Optimization of OSLOM performance. Additional criterion in determining the optimal JI threshold by assessing the consistency of OSLOM community detection performance. The upper plot shows the number of communities and cliques detected by OSLOM for each tested JI threshold. The lower plot depicts the drop in standard deviation (SD) of the number of cliques and communities detected as the network becomes sparser. The dashed vertical line represents the 0.3 JI threshold.

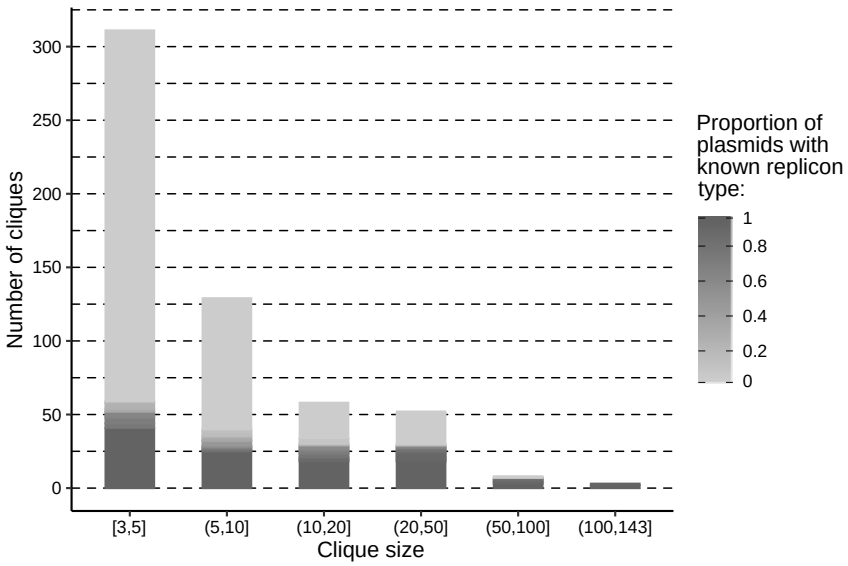


Figure A.4. The distribution of the clique sizes. The cliques were sorted into six bins based on the number of plasmids they carry (x-axis). The shading on each bin corresponds to the number of cliques within a bin that have a certain proportion of plasmids with known replicon type (as indicated by the figure legend at right).

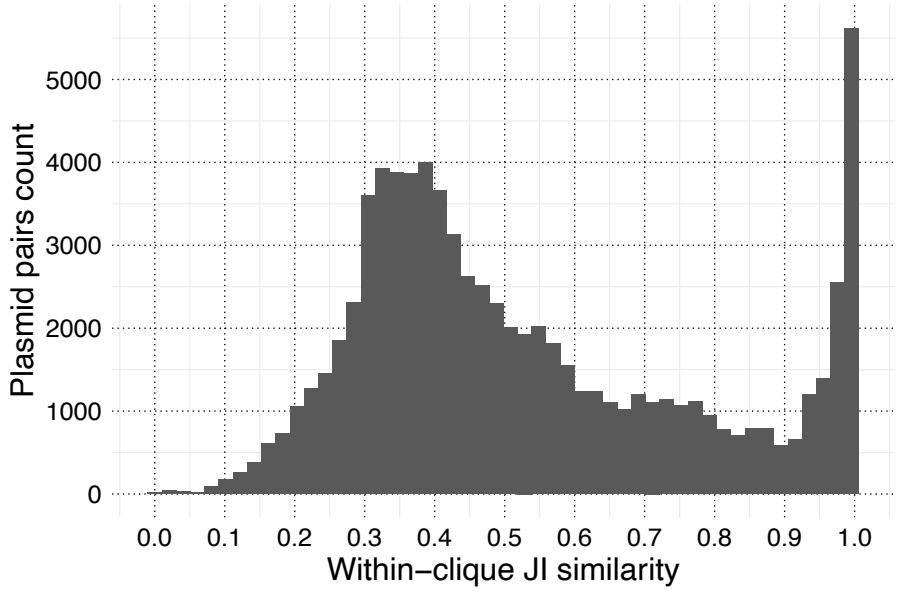


Figure A.5. The distribution of Jaccard Index (JI) similarities between plasmid pairs within cliques.

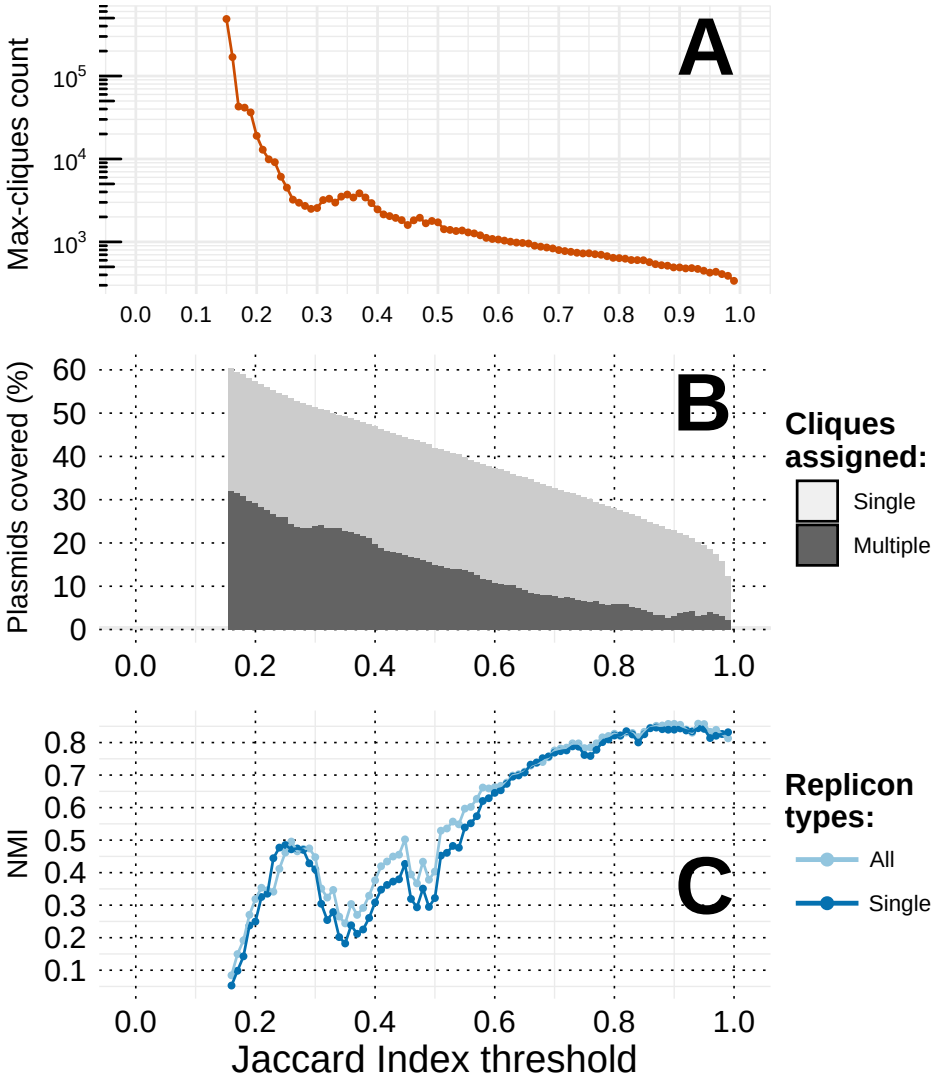


Figure A.6. Assessment of the Max-clique algorithm performance over a range of Jaccard Index (JI) thresholds. JI thresholds were used to transform the weighted plasmid network (Figure 2.2) to a binary one after which a Max-clique algorithm was used to identify all maximal cliques of size three or more. For each JI value, the results presented here show: total number of maximal cliques detected (A), percentage of plasmids covered by the cliques (B), and the congruence with replicon typing measured by NMI score (C). The analysis below JI of 0.15 was aborted due to high memory and computational requirements due to the large number of detected cliques.

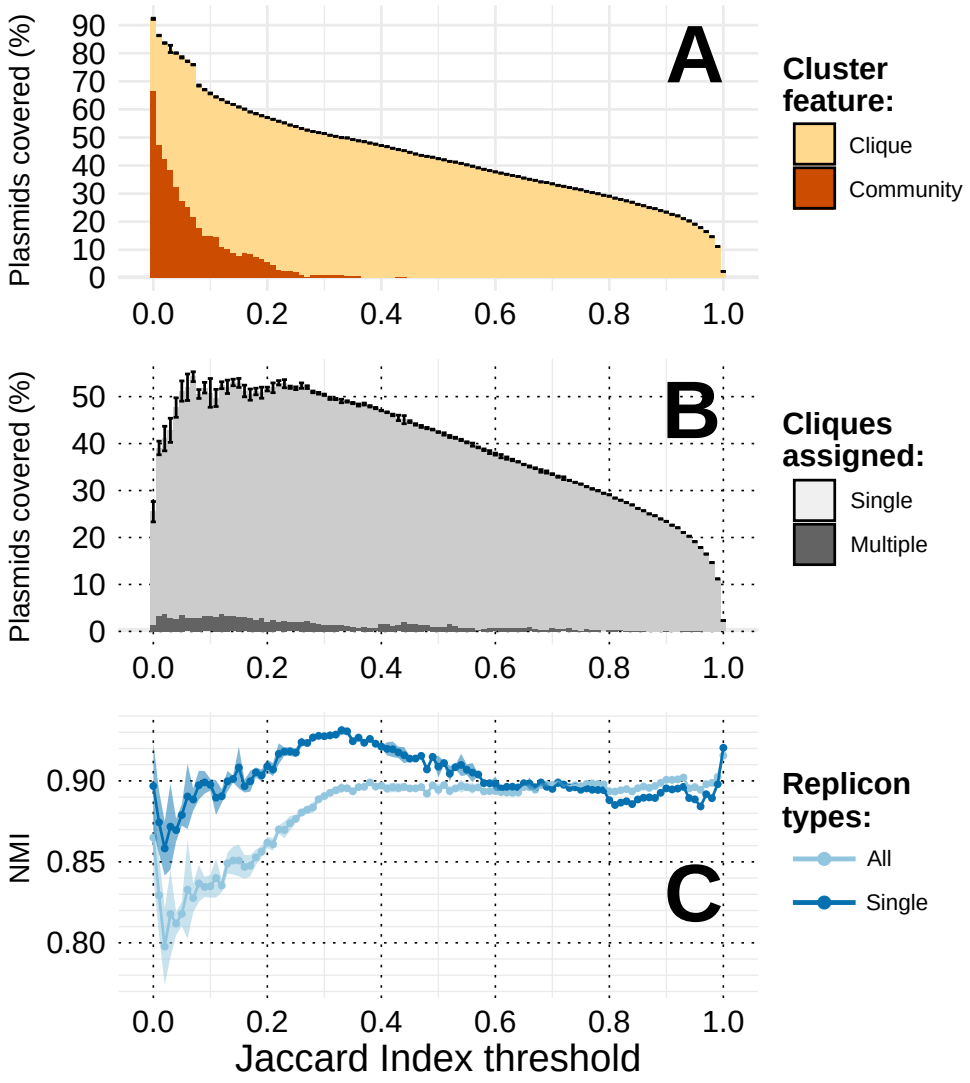


Figure A.7. OSLOM performance over a range of Jaccard Index (JI) thresholds after the removal of 29,913 accessory CDSs from the plasmid sequences. Similarly to Figure 2.5, edges with values below a particular JI threshold were removed from the original plasmid network (Figure 2.2) prior to the OSLOM analysis. The OSLOM performance was assessed based on the following criteria: (A) clique to community ratio; (B) percentage of plasmids covered by the cliques; (C) the congruence with replicon typing measured by NMI score. Error bars (A and B) and light-coloured shading (C) provide +/- 2 standard deviations (SD) of uncertainty. Standard deviation around every value on the y-axis across all JI thresholds assessed (points and bars) was calculated based on results of n=5 iterations of OSLOM software. Maximal NMI score detected was 0.9157 for plasmids assigned to a single or multiple replicon types at JI=1.0, and 0.9312 for plasmids belonging to only one replicon type at JI=0.33.

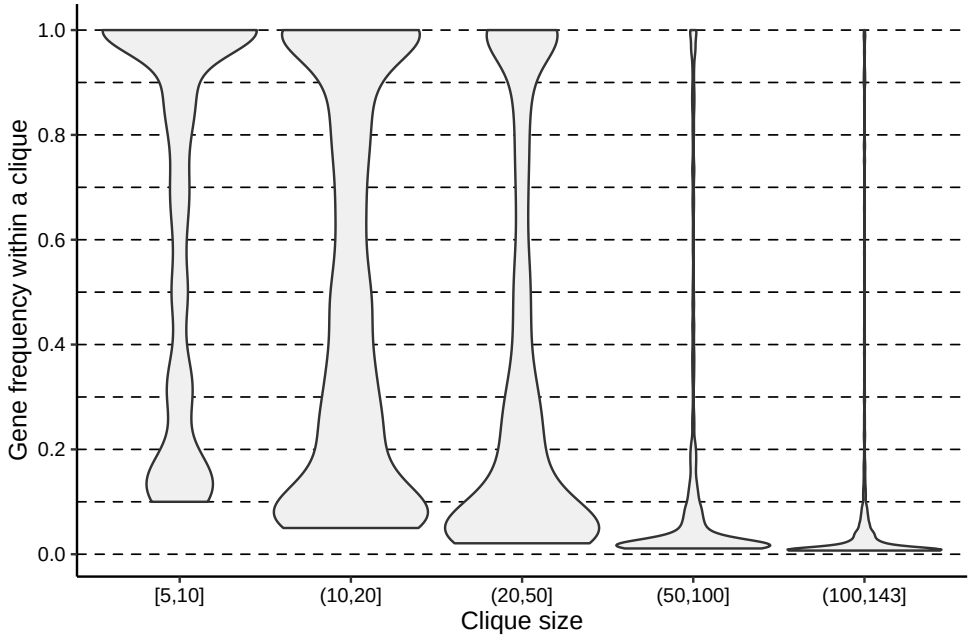


Figure A.8. The distribution of within-clique gene frequencies relative to the clique size. The assessed genes had five or more occurrences in the dataset, thus only the cliques carrying five or more plasmids were considered.

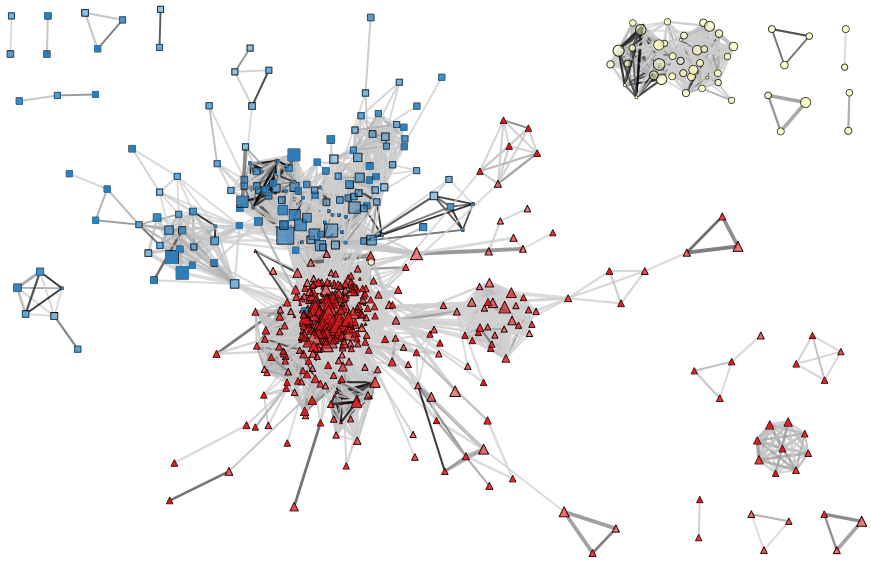


Figure A.9. The unfiltered network of plasmid cliques. Colour and shape of the cliques (vertices) indicate the phylum of the predominant bacterial host: Proteobacteria – red triangle; Firmicutes – blue squares; Other phyla – yellow circles. As noted in the legend of the Figure 3.10, the transparency of the vertex indicates the average internal JI of the clique, the colour of the edges indicates the average JI between plasmids of two cliques, and the width of the edge is proportional to the number of connections.

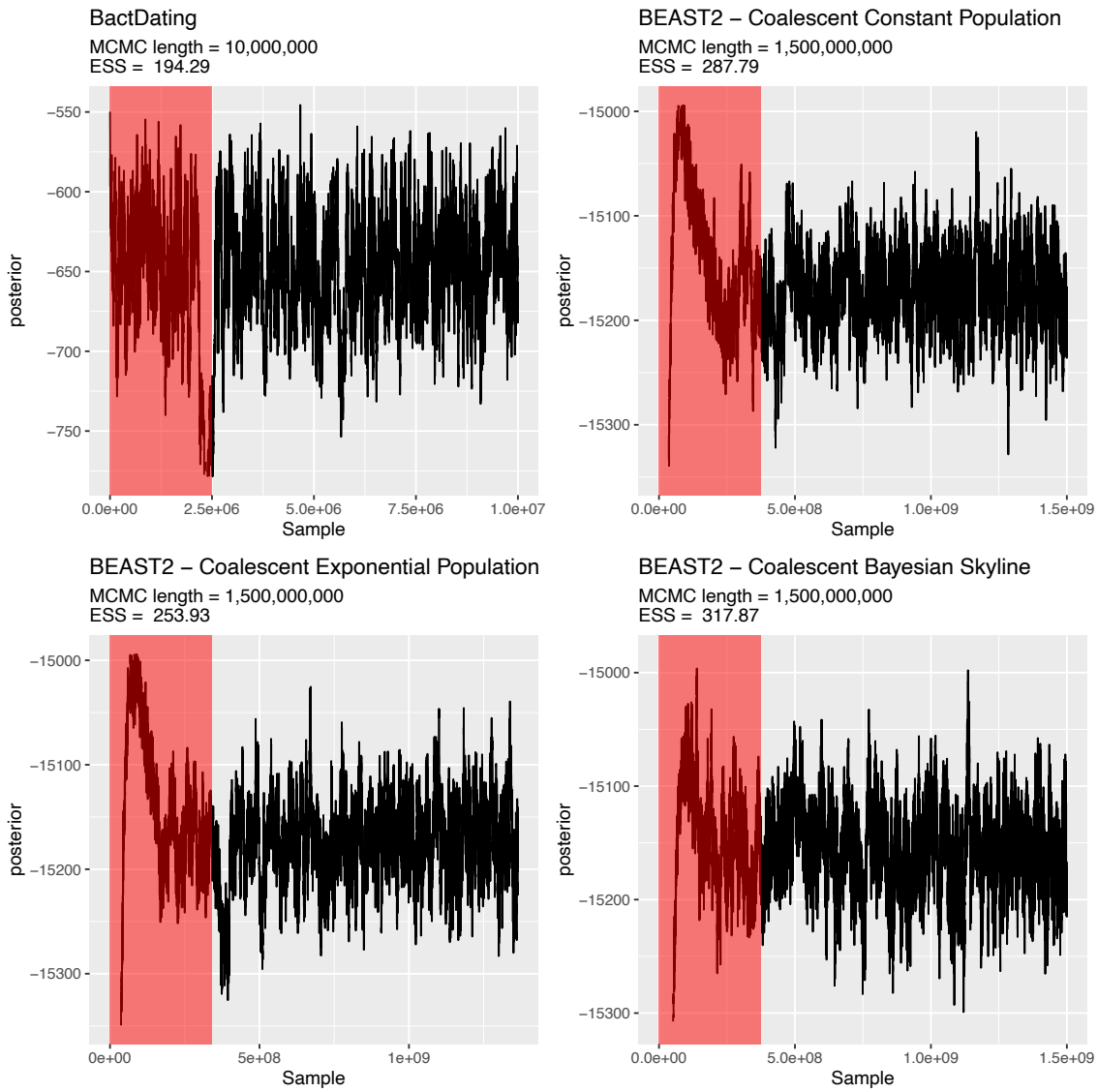


Figure A.10. MCMC trace plots of the posterior for four Bayesian molecular tip-dating analyses of *Tn*125. The MCMC lengths and effective sample sizes (ESS; evaluated using *coda* R package) are provided above each panel. The red shading indicates 20% of burn-in.

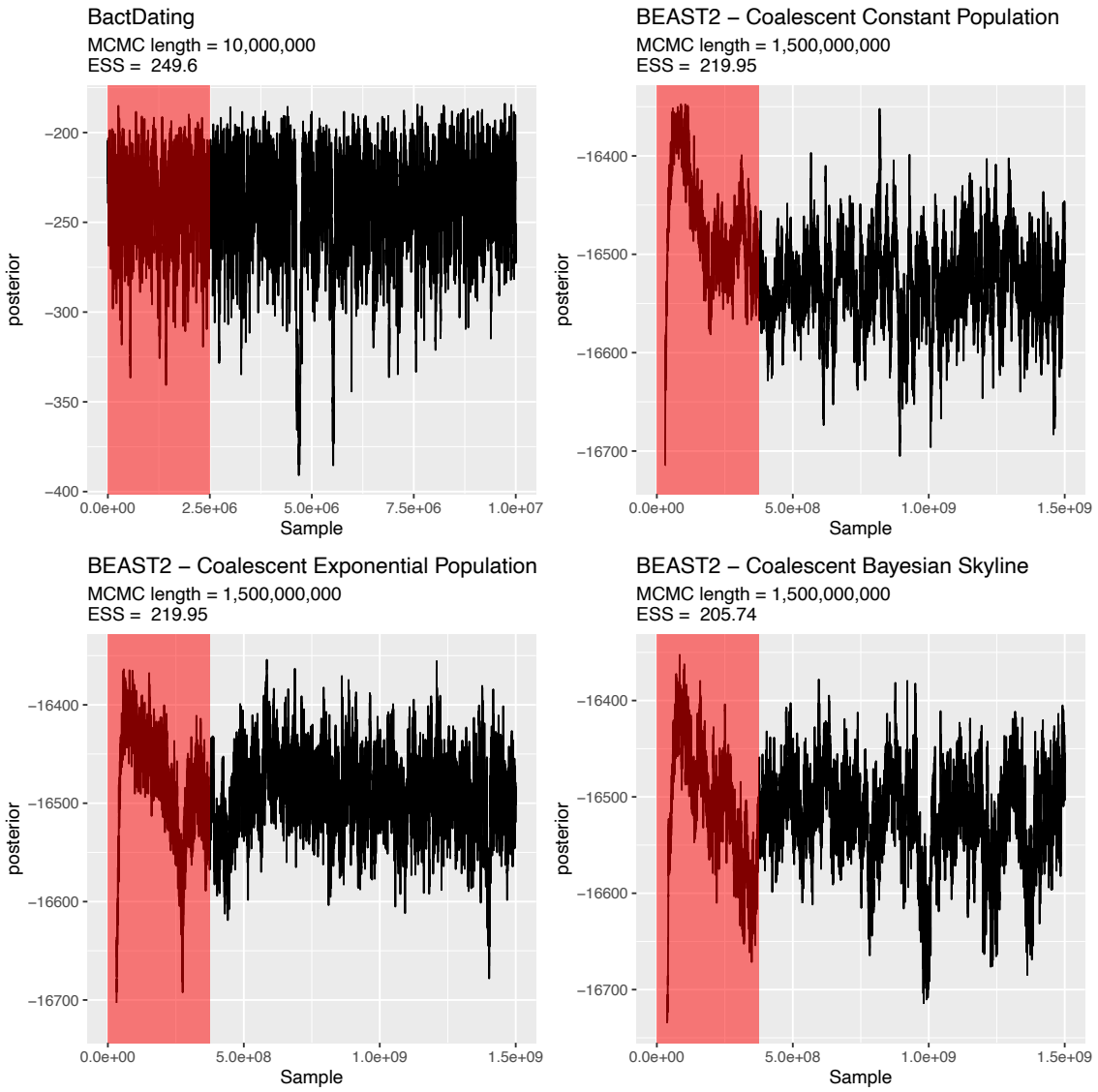


Figure A.11. MCMC trace plots of the posterior for four Bayesian molecular tip-dating analyses of *Tn*3000. The MCMC lengths and effective sample sizes (ESS; evaluated using *coda* R package) are provided above each panel. The red shading indicates 20% of burn-in.

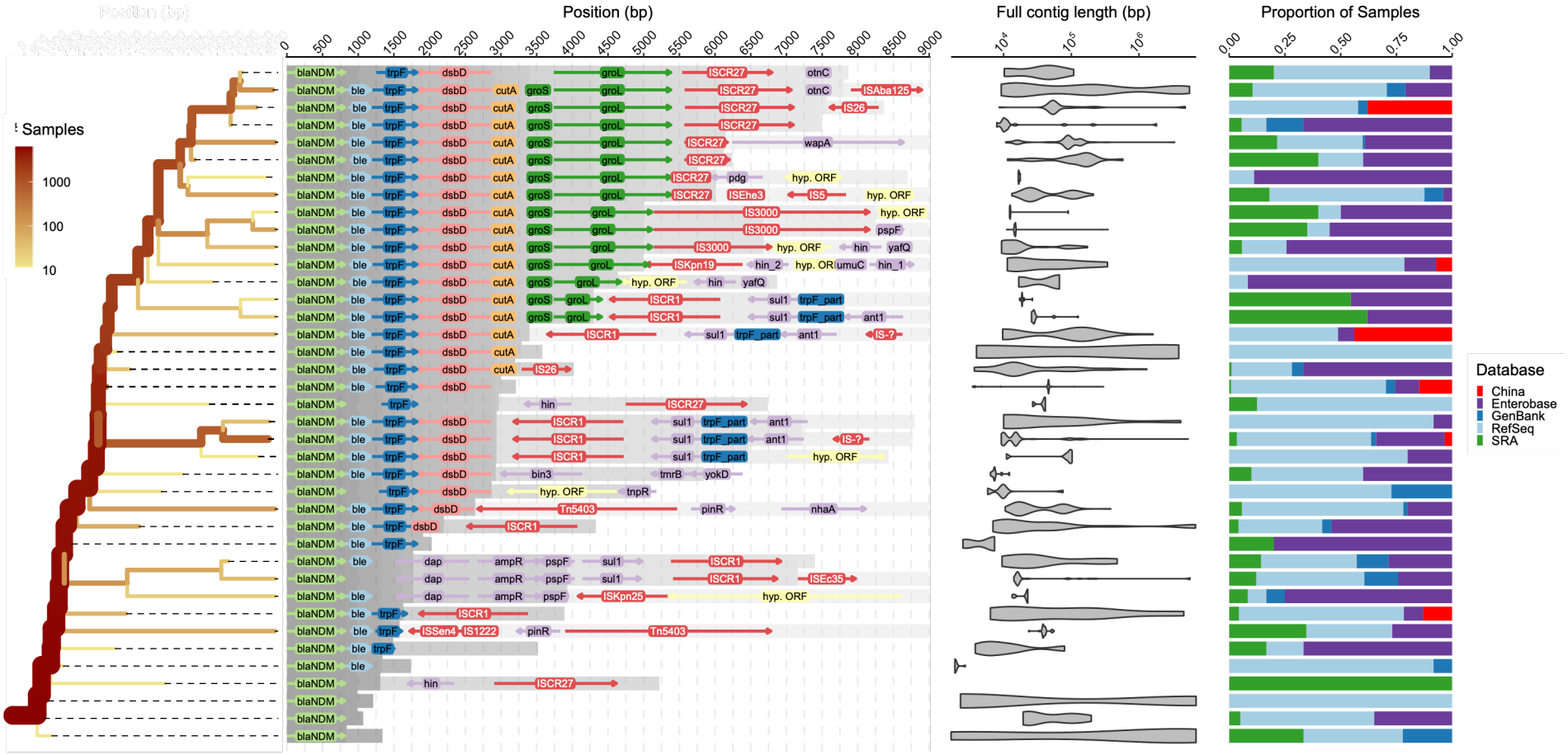


Figure A.12. Splitting of structural variants downstream of *bla*$_{NDM}$ (appendix). This is an extension of Figure 4.13. The thickness and the hue of the branches reflect the number of contigs on each branch. The genome annotations of the most common structural variants are shown in the middle of the figure and the grey shading indicates homology across structural variants. To aid interpretability, some of the structural variants and branches were intentionally cut short. The distributions of contig lengths and source databases are shown on the left-hand side.

Table A.1. A list of positions of candidate replicon genes. Please note some sequences listed here are equivalent.

Plasmid Accession	Candidate gene position
NC_011497	262-1215
NC_001371	4443-4880
NZ_CP018949	8571-8909
NC_001377	44-295
NC_001446	834-1472
NC_001774	3890-4195
NC_017197	25-870
NC_002192	813-1511
NC_002377	30831-34133
NC_003037	857628-858311
NC_019102	1395-1733
NC_003101	7842-8066
NZ_CP021449	19-1026
NZ_CP015851	115313-115732
NZ_CP020888	41561-42781
NC_004834	13-990
NC_006997	778-1545
NC_005003	4425-5249
NC_017284	4045-4452
NZ_CP013734	19606-20445
NC_018978	13-990
NC_006828	4777-5082
NC_008505	25476-25844
NC_008505	65569-66561
NZ_CP016737	1-1167
NC_011960	74675-75781
NC_009435	579-1931
NZ_CP012255	104291-105304
NC_010993	3000-3344
NC_010875	500-1156
NC_010423	1330-1995
NZ_CP014330	20308-21219
NZ_CP013556	464850-465971
NZ_CP019843	944-1288
NC_012551	50068-51588
NC_012661	18-452
NC_013056	7881-8768
NC_013056	6252-7343
NC_013368	239-586
NC_013383	2565-3008
NC_013389	10919-11716
NC_013551	10157-10693

Plasmid Accession	Candidate gene position
NC_013551	10686-11846
NC_013551	2380-2982
NC_014003	1402-1740
NC_014475	36058-38022
NZ_CP010015	77542-78405
NC_016643	1-939
NC_016817	1-930
NC_017172	3546-4313
NC_017496	49863-51215
NC_017575	294824-295243
NC_017575	322-1392
NZ_CP010731	1-1278
NZ_CP010613	17841-18638
NC_018288	101-1126
NC_019180	284-1276
NC_019311	4540-4947
NZ_CP022136	51320-52534
NC_019436	1-1146
NC_020249	42333-42920
NC_020524	565-1686
NC_020525	145-1068
NC_020544	37217-38329
NZ_CP018476	18379-18897
NZ_CP010381	3586-3930
NC_022344	179399-180586
NZ_CP028175	5220-5567
NC_023140	17772-18569
NZ_CP021044	1-1278
NC_023142	4-1005
NC_025149	1523-1867
NC_025152	35031-35771
NC_025172	1-1173
NC_025173	747-1919
NZ_CP003994	1980-2324
NZ_CP003994	2942-3112
NZ_CP004860	311588-312514
NZ_CP004871	17488-18591
NZ_CP004874	71999-73540
NZ_CP006989	274820-274978
NZ_CP021127	216923-218044
NZ_CP006991	512959-514029
NZ_CP009029	718-1845

Plasmid Accession	Candidate gene position
NZ_CP009030	31356-32582
NZ_CP028194	10603-11451
NZ_CP009851	14693-15379
NZ_CP029428	46538-47284
NZ_CP010203	415-771
NZ_CP010382	21422-22303
NZ_CP010383	99378-100244
NZ_CP010616	47239-48732
NZ_CP010646	1-1278
NZ_CP010648	1-1014
NZ_CP010722	1-1035
NZ_CP011121	56170-57093
NZ_CP011596	20993-22003
NZ_CP011596	38886-39785
NZ_CP011606	1420-1764
NZ_CP013797	183141-184112
NZ_CP011969	22635-23822
NZ_CP019733	36298-37599
NZ_CP014889	5615-6718
NZ_CP012274	20233-20742
NZ_CP012417	6632-7132
NZ_CP013822	12881-14005
NZ_CP012984	22999-23751
NZ_CP013494	694338-698999
NZ_CP013494	659924-661045
NZ_CP013632	323796-324866
NZ_CP013642	274234-275355
NZ_CP013750	41909-43441
NZ_CP013752	32528-33619
NZ_CP013842	3270-3824
NZ_CP014627	5330-6430
NZ_CP016407	1-1011
NZ_CP017850	34600-35475
NZ_CP018472	166215-167342
NZ_CP020969	10947-12440
NZ_CP021485	47806-49863
NZ_CP021485	53789-54478
NZ_CP022666	1026944-1027606
NZ_CP022666	662428-663549
NZ_CP025507	912892-917502
NZ_CP024097	66900-68456
NZ_CP026086	101610-102548

Table A.2. Metadata of 104 newly sequenced bacterial isolates collected across mainland China. The isolates were collected as a part of pathogen surveillance efforts of Peking University People's Hospital and include 69 *Escherichia coli*, 34 *Klebsiella pneumoniae* and one *Klebsiella oxytoca* samples from environmental (farm) and hospital settings.

Sample	Biosample	Organism	Collection date	Sampling location	Host	Isolation Source
C1021	SAMN21365994	K pneumoniae	11/08/2016	Beijing, China	Homo sapiens	Abdominal fluid
C1026	SAMN21365995	K pneumoniae	19/08/2016	Beijing, China	Homo sapiens	Sputum
C1044	SAMN21365996	E coli	05/09/2016	Zunhua, China	Homo sapiens	Secretion
C1074	SAMN21365997	E coli	28/05/2016	Xuzhou, China	Homo sapiens	Blood
C1107	SAMN21365998	E coli	27/06/2016	Xuzhou, China	Homo sapiens	Sputum
C1151	SAMN21365999	K pneumoniae	29/08/2016	Weifang, China	Homo sapiens	Sputum
C1157	SAMN21366000	E coli	03/10/2016	Weifang, China	Homo sapiens	Sputum
C1184	SAMN21366001	K pneumoniae	02/10/2015	Changsha, China	Homo sapiens	Drainage
C1193	SAMN21366002	K pneumoniae	14/09/2015	Changsha, China	Homo sapiens	Sputum
C1197	SAMN21366003	E coli	23/01/2015	Changsha, China	Homo sapiens	Wound
C1203	SAMN21366004	K pneumoniae	25/01/2015	Changsha, China	Homo sapiens	Blood
C1279A	SAMN21366005	E coli	06/10/2016	Liaocheng, China	Homo sapiens	Urine
C1288	SAMN21366006	E coli	01/12/2016	Zibo, China	Homo sapiens	pus
C1293	SAMN21366007	E coli	11/08/2016	Zibo, China	Homo sapiens	Urine
C133	SAMN21366008	E coli	06/03/2013	Guangzhou, China	Homo sapiens	Blood
C135	SAMN21366009	E coli	17/03/2013	Xi'an, China	Homo sapiens	Sputum
C1358	SAMN21366010	E coli	25/02/2015	Beijing, China	Homo sapiens	Sputum
C1375	SAMN21366011	E coli	29/11/2016	Beijing, China	Homo sapiens	Urine
C1376	SAMN21366012	E coli	17/02/2016	Beijing, China	Homo sapiens	Abscess
C141	SAMN21366013	E coli	12/04/2013	Wuhan, China	Homo sapiens	Blood
C1461B	SAMN21366014	E coli	09/12/2016	Zunhua, China	Homo sapiens	Urine
C1522	SAMN21366015	E coli	23/12/2016	Xiangtan, China	Homo sapiens	Abdominal wound discharge
C1555	SAMN21366016	E coli	02/07/2016	Guangzhou, China	Homo sapiens	Sputum
C1596	SAMN21366017	K pneumoniae	03/08/2016	Guangzhou, China	Homo sapiens	Broncho-alveolar lavage
C1609	SAMN21366018	E coli	25/10/2016	Guangzhou, China	Homo sapiens	Abdominal fluid
C1616	SAMN21366019	E coli	16/06/2016	Jinan, China	Homo sapiens	Urine
C174	SAMN21366020	E coli	09/06/2014	Beijing, China	Homo sapiens	Blood
C179	SAMN21366021	E coli	08/09/2014	Beijing, China	Homo sapiens	Blood
C182	SAMN21366022	K pneumoniae	18/09/2014	Beijing, China	Homo sapiens	Sputum
C184	SAMN21366023	K pneumoniae	06/10/2014	Beijing, China	Homo sapiens	Urine
C1858	SAMN21366024	E coli	03/08/2016	Xi'an, China	Homo sapiens	fine-needle
C189	SAMN21366025	E coli	13/10/2014	Beijing, China	Homo sapiens	Urine
C1972	SAMN21366026	K oxytoca	11/10/2016	Lanzhou, China	Homo sapiens	Sputum
C295	SAMN21366027	K pneumoniae	05/08/2015	Liaocheng, China	Homo sapiens	Sputum
C296	SAMN21366028	E coli	10/07/2015	Liaocheng, China	Homo sapiens	Urine
C313	SAMN21366029	E coli	13/08/2015	Beijing, China	Homo sapiens	Urine
C320	SAMN21366030	E coli	15/10/2015	Beijing, China	Homo sapiens	Blood
C406	SAMN21366031	K pneumoniae	25/05/2015	Dongguan, China	Homo sapiens	Urine
C407	SAMN21366032	K pneumoniae	28/11/2015	Dongguan, China	Homo sapiens	Sputum
C414	SAMN21366033	K pneumoniae	30/12/2015	Yinchuan, China	Homo sapiens	catheter site

C430	SAMN21366034	E coli	22/08/2015	Xiangtan, China	Homo sapiens	Urine
C435	SAMN21366035	E coli	10/12/2015	Xiangtan, China	Homo sapiens	Urine
C440	SAMN21366036	K pneumoniae	30/01/2016	Xiangtan, China	Homo sapiens	Ulcer; decubitis
C459	SAMN21366037	K pneumoniae	09/12/2015	Jinan, China	Homo sapiens	Sputum
C460	SAMN21366038	K pneumoniae	25/11/2015	Jinan, China	Homo sapiens	Pleural fluid
C461	SAMN21366039	E coli	18/12/2015	Jinan, China	Homo sapiens	Broncho-alveolar lavage
C462	SAMN21366040	E coli	18/12/2015	Jinan, China	Homo sapiens	Urine
C463	SAMN21366041	E coli	13/01/2016	Jinan, China	Homo sapiens	Sputum
C473	SAMN21366042	E coli	03/01/2015	Jinan, China	Homo sapiens	pus
C477	SAMN21366043	K pneumoniae	27/09/2014	Jinan, China	Homo sapiens	Sputum
C478	SAMN21366044	K pneumoniae	22/01/2015	Jinan, China	Homo sapiens	Blood
C481	SAMN21366045	E coli	29/11/2014	Jinan, China	Homo sapiens	Pus
C487	SAMN21366046	E coli	25/06/2015	Jinan, China	Homo sapiens	Catheter
C495	SAMN21366047	K pneumoniae	22/12/2014	Jinan, China	Homo sapiens	Sputum
C501	SAMN21366048	E coli	06/01/2015	Jinan, China	Homo sapiens	Sputum
C508	SAMN21366049	E coli	24/03/2016	Yinchuan, China	Homo sapiens	Sputum
C539	SAMN21366050	K pneumoniae	30/03/2016	Liaocheng, China	Homo sapiens	Sputum
C544	SAMN21366051	E coli	17/02/2016	Beijing, China	Homo sapiens	Sputum
C57	SAMN21366052	K pneumoniae	11/05/2012	Xi'an, China	Homo sapiens	Sputum
C597	SAMN21366053	K pneumoniae	20/04/2015	Xi'an, China	Homo sapiens	Sputum
C605	SAMN21366054	E coli	06/01/2015	Guangzhou, China	Homo sapiens	Urine
C624	SAMN21366055	E coli	07/12/2015	Guangzhou, China	Homo sapiens	Blood
C625	SAMN21366056	K pneumoniae	14/12/2015	Guangzhou, China	Homo sapiens	Drainage
C631	SAMN21366057	K pneumoniae	06/01/2016	Guangzhou, China	Homo sapiens	Urine
C684	SAMN21366058	E coli	25/04/2013	Guangzhou, China	Homo sapiens	Urine
C693	SAMN21366059	E coli	18/12/2014	Guangzhou, China	Homo sapiens	Bile
C812	SAMN21366060	E coli	04/04/2016	Tengzhou, China	Homo sapiens	Sputum
C82	SAMN21366061	K pneumoniae	11/04/2013	Beijing, China	Homo sapiens	Urine
C831	SAMN21366062	E coli	26/02/2015	Xuzhou, China	Homo sapiens	Sputum
C84	SAMN21366063	K pneumoniae	20/07/2013	Beijing, China	Homo sapiens	Retroperitoneum
C848	SAMN21366064	E coli	13/01/2016	Xuzhou, China	Homo sapiens	Urine
C889	SAMN21366065	E coli	13/10/2015	Xiamen, China	Homo sapiens	Blood
C898	SAMN21366066	E coli	22/05/2016	Liaocheng, China	Homo sapiens	Urine
C900	SAMN21366067	E coli	23/12/2015	Jinan, China	Homo sapiens	Sputum
C924	SAMN21366068	E coli	07/07/2016	Qinhuangdao, China	Homo sapiens	Urine
C953	SAMN21366069	K pneumoniae	11/04/2016	Hangzhou, China	Homo sapiens	Blood
C965	SAMN21366070	K pneumoniae	20/01/2016	Urumqi, China	Homo sapiens	Pleural fluid
C966	SAMN21366071	K pneumoniae	27/12/2015	Urumqi, China	Homo sapiens	Urine
C971	SAMN21366072	E coli	01/03/2016	Harbin, China	Homo sapiens	Blood
WFA04	SAMN21366073	E coli	05/09/2016	Weifang, China	pig	sick pig lung
WFA13	SAMN21366074	E coli	05/09/2016	Weifang, China	chicken	sick chicken lung
WFA17	SAMN21366075	E coli	05/09/2016	Weifang, China	chicken	sick chicken lung
WFA19	SAMN21366076	E coli	05/09/2016	Weifang, China	chicken	sick chicken liver
WFA24	SAMN21366077	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA27	SAMN21366078	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA29	SAMN21366079	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA31	SAMN21366080	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA35	SAMN21366081	E coli	05/09/2016	Weifang, China	chicken	chicken manure

WFA36	SAMN21366082	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA41A	SAMN21366083	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA56	SAMN21366084	E coli	05/09/2016	Weifang, China	environmental	chicken slaughterhouse splitting skin
WFA62A	SAMN21366085	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA64A	SAMN21366086	E coli	05/09/2016	Weifang, China	pig	sick pig liver
WFA65A	SAMN21366087	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA66A	SAMN21366088	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA76	SAMN21366089	K pneumoniae	05/09/2016	Weifang, China	chicken	chicken manure
WFA84	SAMN21366090	K pneumoniae	05/09/2016	Weifang, China	chicken	chicken manure
WFA85	SAMN21366091	K pneumoniae	05/09/2016	Weifang, China	chicken	chicken manure
WFA89A	SAMN21366092	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA90	SAMN21366093	E coli	05/09/2016	Weifang, China	chicken	chicken manure
WFA98	SAMN21366094	K pneumoniae	05/09/2016	Weifang, China	chicken	chicken manure
WW04	SAMN21366095	K pneumoniae	16/12/2016	Weifang, China	environmental	chicken dung channel
WW09A	SAMN21366096	E coli	18/12/2016	Weifang, China	environmental	chicken dung channel
WW12A	SAMN21366097	E coli	19/12/2016	Weifang, China	environmental	chicken dung channel

Appendix B

An example BEAST2 Bayesian Skyline configuration file used in molecular dating

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?><beast beautitemplate='Standard'
beautistatus=''namespace="beast.core:beast.evolution.alignment:beast.evolution.tree.coalescent
:beast.core.util:beast.evolution.nuc:beast.evolution.operators:beast.evolution.sitemodel:beast
.evolution.substitutionmodel:beast.evolution.likelihood" required="" version="2.6">

<data id="alignment" spec="Alignment" name="alignment">
      <!-- INSERT ALIGNED SEQUENCES (eg. bellow) -->
<sequence id="seq_DRR111572" spec="Sequence" taxon="DRR111572" totalcount="4"
value="GGCAGAGTAAAACTTGAAGTGCGACATAAACCACCTAATTAATTTAAA"/>
</data>

<map name="Uniform" >beast.math.distributions.Uniform</map>
<map name="Exponential" >beast.math.distributions.Exponential</map>
<map name="LogNormal" > beast.math.distributions.LogNormalDistributionModel</map>
<map name="Normal" >beast.math.distributions.Normal</map>
<map name="Beta" >beast.math.distributions.Beta</map>
<map name="Gamma" >beast.math.distributions.Gamma</map>
<map name="LaplaceDistribution" > beast.math.distributions.LaplaceDistribution</map>
<map name="prior" >beast.math.distributions.Prior</map>
<map name="InverseGamma" >beast.math.distributions.InverseGamma</map>
<map name="OneOnX" >beast.math.distributions.OneOnX</map>


<run id="mcmc" spec="MCMC" chainLength="1500000000">
        <state id="state" spec="State" storeEvery="5000">

        <!-- INSERT COLLECTION DATES (eg. bellow) -->
        <tree id="Tree.t:tree" spec="beast.evolution.tree.Tree" name="stateNode">
<trait id="dateTrait.t:tree" spec="beast.evolution.tree.TraitSet"
traitname="date"  value="DRR111572=2015.649315, DRR121824=2016.021858,
DRR121825=2016.008197, GCA_003428845=2015.939726">
<taxa id="TaxonSet.alignment" spec="TaxonSet">
<alignment idref="alignment"/>
</taxa>
</trait>
<taxonset idref="TaxonSet.alignment"/>
</tree>

<parameter id="clockRate.c:clock" spec="parameter.RealParameter"
name="stateNode">1.0</parameter>
<parameter id="freqParameter.s:site" spec="parameter.RealParameter" dimension="4"
lower="0.0" name="stateNode" upper="1.0">0.25</parameter>
<parameter id="rateAC.s:site" spec="parameter.RealParameter" lower="0.0"
name="stateNode">1.0</parameter>
<parameter id="rateAG.s:site" spec="parameter.RealParameter" lower="0.0"
name="stateNode">1.0</parameter>
<parameter id="rateAT.s:site" spec="parameter.RealParameter" lower="0.0"
name="stateNode">1.0</parameter>
<parameter id="rateCG.s:site" spec="parameter.RealParameter" lower="0.0"
name="stateNode">1.0</parameter>
<parameter id="rateCT.s:site" spec="parameter.RealParameter" lower="0.0"
name="stateNode">1.0</parameter>
<parameter id="rateGT.s:site" spec="parameter.RealParameter" lower="0.0"
name="stateNode">1.0</parameter>
<parameter id="bPopSizes.t:tree" spec="parameter.RealParameter" dimension="5"
lower="0.0" name="stateNode">380.0</parameter>

<stateNode id="bGroupSizes.t:tree" spec="parameter.IntegerParameter"
dimension="5">1</stateNode>
</state>


<!-- INSERT STARTING TREE IN NEWICK FORMAT -->
```

```
  <init id="NewickTree.t:tree" spec="beast.util.TreeParser" IsLabelledNewick="true"
  estimate="false" initial="@Tree.t:tree" newick="(((())));" taxa="@alignment"/>

  <distribution id="posterior" spec="util.CompoundDistribution">

<!-- Setting up prior distributions for MCMC run -->
  <distribution id="prior" spec="util.CompoundDistribution">
          <distribution id="BayesianSkyline.t:tree" spec="BayesianSkyline"
          groupSizes="@bGroupSizes.t:tree" popSizes="@bPopSizes.t:tree">
                  <treeIntervals id="BSPTreeIntervals.t:tree" spec="TreeIntervals"
                  tree="@Tree.t:tree"/>
          </distribution>


  <distribution id="MarkovChainedPopSizes.t:tree"
  spec="beast.math.distributions.MarkovChainDistribution" jeffreys="true"
  parameter="@bPopSizes.t:tree"/>

          <prior id="ClockPrior.c:clock" name="distribution" x="@clockRate.c:clock">
                  <Uniform id="Uniform.0" name="distr" upper="Infinity"/>
          </prior>

          <prior id="FrequenciesPrior.s:site" name="distribution"
          x="@freqParameter.s:site">
                  <Uniform id="Uniform.3" name="distr"/>
          </prior>

          <prior id="RateACPrior.s:site" name="distribution" x="@rateAC.s:site">
                  <Gamma id="Gamma.0" name="distr">
                          <parameter id="RealParameter.1" spec="parameter.RealParameter"
                          estimate="false" name="alpha">0.05</parameter>
                          <parameter id="RealParameter.2" spec="parameter.RealParameter"
                          estimate="false" name="beta">10.0</parameter>
                  </Gamma>
          </prior>

          <prior id="RateAGPrior.s:site" name="distribution" x="@rateAG.s:site">
                  <Gamma id="Gamma.1" name="distr">
                          <parameter id="RealParameter.3" spec="parameter.RealParameter"
                          estimate="false" name="alpha">0.05</parameter>
                          <parameter id="RealParameter.4" spec="parameter.RealParameter"
                          estimate="false" name="beta">20.0</parameter>
                  </Gamma>
          </prior>

          <prior id="RateATPrior.s:site" name="distribution" x="@rateAT.s:site">
                  <Gamma id="Gamma.2" name="distr">
                          <parameter id="RealParameter.5" spec="parameter.RealParameter"
                          estimate="false" name="alpha">0.05</parameter>
                          <parameter id="RealParameter.6" spec="parameter.RealParameter"
                          estimate="false" name="beta">10.0</parameter>
                  </Gamma>
          </prior>

          <prior id="RateCGPrior.s:site" name="distribution" x="@rateCG.s:site">
                  <Gamma id="Gamma.3" name="distr">
                          <parameter id="RealParameter.7" spec="parameter.RealParameter"
                          estimate="false" name="alpha">0.05</parameter>
                          <parameter id="RealParameter.8" spec="parameter.RealParameter"
                          estimate="false" name="beta">10.0</parameter>
                  </Gamma>
          </prior>

          <prior id="RateCTPrior.s:site" name="distribution" x="@rateCT.s:site">
                  <Gamma id="Gamma.4" name="distr">
                          <parameter id="RealParameter.9" spec="parameter.RealParameter"
                          estimate="false" name="alpha">0.05</parameter>
                          <parameter id="RealParameter.10"
                          spec="parameter.RealParameter" estimate="false"
                          name="beta">20.0</parameter>
                  </Gamma>
        </prior>

        <prior id="RateGTPrior.s:site" name="distribution" x="@rateGT.s:site">
                  <Gamma id="Gamma.5" name="distr">
                          <parameter id="RealParameter.11"
                          spec="parameter.RealParameter" estimate="false"
                          name="alpha">0.05</parameter>
                          <parameter id="RealParameter.12"
                          spec="parameter.RealParameter" estimate="false"
                          name="beta">10.0</parameter>
                  </Gamma>
          </prior>
  </distribution>


  <distribution id="likelihood" spec="util.CompoundDistribution" useThreads="true">
```

```
                <distribution id="treeLikelihood.alignment" spec="ThreadedTreeLikelihood"
                data="@alignment" tree="@Tree.t:tree">
                        <siteModel id="SiteModel.s:site" spec="SiteModel">
                                <parameter id="mutationRate.s:site"
                                spec="parameter.RealParameter" estimate="false"
                                name="mutationRate">1.0</parameter>
                                <parameter id="gammaShape.s:site"
                                spec="parameter.RealParameter" estimate="false"
                                name="shape">1.0</parameter>
                        <parameter id="proportionInvariant.s:site"
                        spec="parameter.RealParameter" estimate="false" lower="0.0"
                        name="proportionInvariant" upper="1.0">0.0</parameter>

                        <substModel id="gtr.s:site" spec="GTR" rateAC="@rateAC.s:site"
                        rateAG="@rateAG.s:site" rateAT="@rateAT.s:site"
                        rateCG="@rateCG.s:site" rateCT="@rateCT.s:site"
                        rateGT="@rateGT.s:site">
                                <frequencies id="estimatedFreqs.s:site" spec="Frequencies"
                                frequencies="@freqParameter.s:site"/>
                        </substModel>

                </siteModel>
                <branchRateModel id="StrictClock.c:clock"
                spec="beast.evolution.branchratemodel.StrictClockModel"
                clock.rate="@clockRate.c:clock"/>
            </distribution>

      </distribution>
      </distribution>


<!-- Defining operators on distributions above -->
<!-- Clock -->
      <operator id="StrictClockRateScaler.c:clock" spec="ScaleOperator"
      parameter="@clockRate.c:clock" weight="3.0"/>
      <operator id="strictClockUpDownOperator.c:clock" spec="UpDownOperator"
      scaleFactor="0.75" weight="3.0">
            <up idref="clockRate.c:clock"/>
            <down idref="Tree.t:tree"/>
      </operator>

<!-- Site -->
      <operator id="FrequenciesExchanger.s:site" spec="DeltaExchangeOperator" delta="0.01"
      weight="0.1">
            <parameter idref="freqParameter.s:site"/>
      </operator>
      <operator id="RateACScaler.s:site" spec="ScaleOperator" parameter="@rateAC.s:site"
      scaleFactor="0.5" weight="0.1"/>
      <operator id="RateAGScaler.s:site" spec="ScaleOperator" parameter="@rateAG.s:site"
      scaleFactor="0.5" weight="0.1"/>
      <operator id="RateATScaler.s:site" spec="ScaleOperator" parameter="@rateAT.s:site"
      scaleFactor="0.5" weight="0.1"/>
      <operator id="RateCGScaler.s:site" spec="ScaleOperator" parameter="@rateCG.s:site"
      scaleFactor="0.5" weight="0.1"/>
      <operator id="RateCTScaler.s:site" spec="ScaleOperator" parameter="@rateCT.s:site"
      scaleFactor="0.5" weight="0.1"/>
      <operator id="RateGTScaler.s:site" spec="ScaleOperator" parameter="@rateGT.s:site"
      scaleFactor="0.5" weight="0.1"/>

<!-- Tree -->
      <operator id="BayesianSkylineTreeScaler.t:tree" spec="ScaleOperator" scaleFactor="0.5"
      tree="@Tree.t:tree" weight="3.0"/>
      <operator id="BayesianSkylineTreeRootScaler.t:tree" spec="ScaleOperator" rootOnly="true"
      scaleFactor="0.5" tree="@Tree.t:tree" weight="3.0"/>
      <operator id="BayesianSkylineUniformOperator.t:tree" spec="Uniform" tree="@Tree.t:tree"
      weight="30.0"/>
      <operator id="BayesianSkylineSubtreeSlide.t:tree" spec="SubtreeSlide"
      tree="@Tree.t:tree" weight="15.0"/>
      <operator id="BayesianSkylineNarrow.t:tree" spec="Exchange" tree="@Tree.t:tree"
      weight="15.0"/>
      <operator id="BayesianSkylineWide.t:tree" spec="Exchange" isNarrow="false"
      tree="@Tree.t:tree" weight="3.0"/>
      <operator id="BayesianSkylineWilsonBalding.t:tree" spec="WilsonBalding"
      tree="@Tree.t:tree" weight="3.0"/>
      <operator id="popSizesScaler.t:tree" spec="ScaleOperator" parameter="@bPopSizes.t:tree"
      weight="15.0"/>
      <operator id="groupSizesDelta.t:tree" spec="DeltaExchangeOperator" integer="true"
      weight="6.0">
      <intparameter idref="bGroupSizes.t:tree"/>
      </operator>

<!-- Loggers -->
      <logger id="tracelog" spec="Logger" fileName="alignment.log" logEvery="5000"
      model="@posterior" sanitiseHeaders="true" sort="smart">
            <log idref="posterior"/>
            <log idref="likelihood"/>
            <log idref="prior"/>
```

```
            <log idref="treeLikelihood.alignment"/>
            <log id="TreeHeight.t:tree" spec="beast.evolution.tree.TreeHeightLogger"
            tree="@Tree.t:tree"/>
            <log idref="clockRate.c:clock"/>
            <log idref="freqParameter.s:site"/>
            <log idref="rateAC.s:site"/>
            <log idref="rateAG.s:site"/>
            <log idref="rateAT.s:site"/>
            <log idref="rateCG.s:site"/>
            <log idref="rateCT.s:site"/>
            <log idref="rateGT.s:site"/>
            <log idref="BayesianSkyline.t:tree"/>
            <log idref="bPopSizes.t:tree"/>
            <log idref="bGroupSizes.t:tree"/>
      </logger>

      <logger id="screenlog" spec="Logger" logEvery="5000">
            <log idref="posterior"/>
            <log idref="likelihood"/>
            <log idref="prior"/>
      </logger>

      <logger id="treelog.t:tree" spec="Logger" fileName="$(tree).trees" logEvery="5000"
      mode="tree">
        <log id="TreeWithMetaDataLogger.t:tree"
        spec="beast.evolution.tree.TreeWithMetaDataLogger" tree="@Tree.t:tree"/>
      </logger>

      <operatorschedule id="OperatorSchedule" spec="OperatorSchedule"/>

</run>
</beast>
```

Appendix C

Additional contributions to scientific research during my doctoral training

Tan, C. C. S., Acman, M., van Dorp, L., & Balloux, F. (2021). Metagenomic evidence for a polymicrobial signature of sepsis. Microbial Genomics, 7(9), 000642. https://doi.org/10.1099/mgen.0.000642

In this publication we utilized machine learning approaches on metagenomic sequencing data from blood of sepsis patients. The analysis resulted in characterization of polymicrobial bacterial community in sepsis. Apart from being involved in the conceptualization and writing of the manuscript, my role was also helping with data pre-processing, network analysis and interpretation of the results.

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., ... & Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution, 83, 104351. https://doi.org/10.1016/j.meegid.2020.104351

Inspired by the current worldwide COVID-19 pandemic, we sort to provide the scientific community with an overview of the diversity of SARS-CoV-2 virus and highlight important recurring mutations in its genome. To this end, I developed a user-friendly web-application for querying the alignment of SARS-CoV-2 genomes. The application highlights dominant mutations and homoplasies across SARS-CoV-2 genome and allows users to subset the data by sampling locations or genes and to inspect ML phylogenetic tree built from the genomes' alignment. The application is available on this website: https://macman123.shinyapps.io/ugi-scov2-alignment-screen/.

van Dorp, L., Richard, D., Tan, C. C., Shaw, L. P., Acman, M., & Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nature communications, 11(1), 1-8. https://doi.org/10.1038/s41467-020-19818-2

To study selective pressures acting on mutations in SARS-CoV-2 genome we developed a phylogeny-based metric and tested if any of observed mutations is significantly linked with increased viral transmission. Here, I was partially involved in data analysis and contributed by updating and improving upon the abovementioned SARS-Cov-2 alignment screening application. In addition, I participated in manuscript writing and regular discussions regarding methodological approaches and results.

van Dorp, L., Nimmo, C., Ortiz, A. T., Pang, J., Acman, M., Tan, C. C., ... & Balloux, F. (2020). Detection of a bedaquiline/clofazimine resistance reservoir in Mycobacterium tuberculosis predating the antibiotic era. bioRxiv. https://doi.org/10.1101/2020.10.06.328799

My contributions to this work on bedaquiline resistance in *Mycobacterium tuberculosis* (Mtb) consisted of: (i) assembling portion of the dataset of bedaquiline resistant Mtb from SRA database using BIGSI screening tool (see Section 4.2.1 for more information); (ii) setting up TBprofiler tool used in classification of Mtb resistance variants; (iii) molecular dating of the emergence of bedaquiline resistant Mtb variant using BEAST2; (iv) proofing the manuscript.

van Dorp, L., Wang, Q., Shaw, L. P., Acman, M., Brynildsrud, O. B., Eldholm, V., ... & Wang, H. (2019). Rapid phenotypic evolution in multidrug-resistant Klebsiella pneumoniae hospital outbreak strains. Microbial genomics, 5(4). https://doi.org/10.1099/mgen.0.000263

My contribution to this research was minor and mostly consisted of managing the data upload, rendering figures, and manuscript proofing.