The magnitude of DNA transfer between plasmids and chromosomes in Prokaryotes

Dissertation
Submitted in fulfillment of the requirements for the degree
Doktor der Naturwissenschaften (Dr. rer. nat.)
in the Faculty of Mathematics and Natural Sciences
of the Christian-Albrechts University Kiel

Submitted by
A. Samer Kadibalban

Kiel, 2021

First referee: Prof. Dr. Tal Dagan

Second referee: Dr. Julien Dutheil

Date of the oral examination: 03.12.2020

“If you torture the data long enough, it will confess to anything”

Ronald Harry Coase

(December 29, 1910 – September 2, 2013)

Declaration

I hereby declare that the thesis entitled “The magnitude of DNA transfer between plasmids and chromosomes in Prokaryotes“ has been carried out in the Institute of General Microbiology at the Christian-Albrechts University of Kiel, Kiel, Germany, under the guidance of Prof. Dr. Tal Dagan and Dr. Giddy Landan. The work is original and has not been submitted in part or full by me for any degree at any other University. I further declare that the material obtained from other sources has been duly acknowledged in the thesis. My work has been produced in compliance to the principles of good scientific practice in accordance with the guidelines of the German science foundation. I hereby assure that I have not been revoked any of my academic degrees.

Kiel, 2021

Ahmad Samer Kadib Alban

TABLE OF CONTENTS

1 ABSTRACT

Plasmids are extrachromosomal genetic elements that are abundantly found in prokaryotic organisms. Plasmids are considered a major contributor to prokaryotic genome evolution due to their ability to transfer DNA between cells and across taxonomic boundaries. Anecdotal evidence shows that genetic material can transfer between co-resident plasmids and chromosomes, either by mobile genetic elements (transposons, integrons, prophages and IS elements) or by homologous recombination. Nonetheless, a comprehensive view on the magnitude and characteristics of the transferred DNA between plasmids and chromosomes is still lacking. Here I developed a novel comparative genomics approach based on clustering of adjacent local similarity regions (BLAST and MUMmer) by segmentation into parsimonious regions of transfer events. Applying our approach to 3,264 chromosome-plasmid pairs that co-inhabit the same prokaryotic host uncovered 332,944 shared regions in 2,272 (69,6%) pairs comprising 2,272 (69,6%) plasmids and 1,157 (82,6%) chromosomes belonging to 1,157 (82,6%) isolates. The shared regions correspond to DNA transfer events that constitute 51,866 plasmid loci and 111,124 chromosomal loci. The homologous loci on each replicon had varying sequence copy number on the other replicon ranging between 1 and 457 chromosomal copies for plasmid loci and between 1 and 109 plasmid copies for chromosomal loci. The high copy number homologous loci usually correspond to transposable elements. Many plasmid loci with chromosomal homology carry complete genes (10,660 loci). Among those loci, 4,303 loci include more than one complete gene with an overall of 15,782 genes that were putatively co-transferred as pairs or longer gene clusters. Characterizing those genes, I found that the majority belong to classes of mobile genetic elements or to hypothetical proteins with unknown functions. Among genes that were laterally transferred between the plasmid and chromosome of the same cell, I also found essential genes such as RNA genes as well as beneficial genes, like antimicrobial resistance genes (AMR) with incidents of multidrug resistance introduced to the genome by plasmids. Chromosome-plasmid pairs with no sequence homology between the plasmid and chromosome, can be characterized by a lower content of coding sequence and a higher genomic

complexity. Finally, I observed and characterized several events of plasmid integration into the chromosomal genome. Our results suggest a common but not very frequent DNA transfer between plasmids and chromosomes, with some exceptions. The majority of gene transfer is facilitated by mobile elements, indicating that plasmids play a role in the dissemination of transposons much more than they do antibiotic resistance. My results implicate plasmids as mediators of transposon invasion of prokaryotic genomes.

2 ZUSAMMENFASSUNG

Plasmide sind extrachromosomale genetische Elemente, welche vielzählig in prokaryotischen Organismen vorkommen. Plasmide werden aufgrund ihrer Fähigkeit, DNA zwischen Zellen und über taxonomische Grenzen hinweg zu transportieren, als Hauptverantwortliche der prokaryotischen Genomevolution betrachtet. Vereinzelt wurde bewiesen, dass genetisches Material zwischen Plasmiden einer Zelle und Chromosomen durch mobile genetische Elemente (Transposons, Integrons, Prophagen und IS-Elemente) oder homologe Rekombination übertragen werden kann. Dennoch existiert bisher kein zusammenfassender Überblick über das Ausmaß und die Charakteristik der zwischen Plasmiden und Chromosomen transferierten DNA. In dieser Arbeit entwickle ich einen neuen Ansatz zur vergleichenden Genomik. Dieser basiert auf der Gruppierung von benachbarten, lokalen Alignment-Regionen (BLAST und MUMer), welche in ‚parsimonious'-Sequenzen von Transferereignissen eingeteilt werden. Der Ansatz wurde auf 3,264 Chromosomen-Plasmid-Paare, welche denselben prokaryotischen Wirt bewohnen, angewandt. 332,944 Regionen wurden aufgedeckt, welche sowohl im Plasmid als auch im Chromosom von 2,272 (69,6%) Paaren gefunden wurde, dies umfasst 2,272 (69,6%) Plasmide und 1,157 (82,6%) der Chromosomen. Diese Sequenzen gehören zu DNA-Übertragungsereignissen von 51,866 Plasmid-Loci und 111,124 Chromosomen-Loci. Die homologen Loci auf jedem der Replikons haben eine andere Anzahl an Sequenzkopien auf dem jeweils anderem Replikon. Die Anzahl der Sequenzkopien variiert dabei zwischen einem bis 457 chromosomalen Kopien für Plasmide und einem bis 109 Plasmidkopien für chromosomale Loci. Die hohe Anzahl an Kopien der homologen Loci wird oft mit mobilen genetischen Elementen in Verbindung gebracht. Viele der Plasmid-Loci mit chromosomaler Homologie tragen vollständige Gene (10,660 Loci). Von diesen Loci umfassen 4,303 Loci mehr als ein vollständiges Gen, dabei handelt es sich insgesamt um 15,782 Gene, welche vermutlich als Paare großer Gen-Cluster zusammen transferiert wurden. Durch die Charakterisierung dieser Gene, habe ich herausgefunden, dass die Mehrheit dieser zu der Klasse der mobilen genetischen Elemente oder hypothetischen Proteine mit unbekannter Funktion gehören. Unter den Genen, die lateral zwischen dem Plasmid und dem

Chromosom transferiert wurden, konnte ich essentielle Gene, wie RNA-Gene oder auch vorteilhaften Gene, wie antimikrobielle Resistenz-Gene (AMR), finden. Einige der Chromosomen erhielten durch die Plasmide mehrere Resistenz-Gene. Chromosomen-Plasmid-Paare, die keine homologen Sequenzen zwischen dem Plasmid und dem Chromosom aufweisen, können durch ihren geringeren Anteil codierender Gene und einer höheren Komplexität charakterisiert werden. Schließlich habe ich mehrere Ereignisse beobachten und charakterisieren können, in denen das vollständige Plasmid in das Chromosom intergiert wurde. Meine Ereignisse lassen auf einen allgemeinen aber nicht häufigen DNA-Transfer zwischen Plasmiden und Chromosomen schließen, wobei es Ausnahmen gibt. Der Hauptanteil der Gene wurde durch mobile genetische Elemente transferiert. Dies lässt darauf schließen, dass Plasmide eine größere Rolle in der Verbreitung von Transposons spielen als in der Verbreitung von Antibiotika-Resistenz-Genen. Mein Ergebnis impliziert, dass Plasmide eine Hauptrolle bei der Transposon-Invasion in das prokaryotische Genom einnehmen.

3 INTRODUCTION

3.1 Plasmid evolution

Plasmids are extra-chromosomal genetic elements that colonize and replicate in prokaryotic cells. They are considered a major driving force of prokaryote evolution as they can migrate between populations, making them potent agents of lateral gene transfer (Lederberg and Tatum, 1946; Thomas and Nielsen, 2005). Many plasmids described in the literature encode for a plethora of resistance mechanisms to growth limiting conditions (e.g., antibiotics), which are beneficial to the host depending on the environmental conditions. In contrast, plasmids that supply the host with essential functions, i.e., whose benefit to the host is independent of the environmental conditions, are somewhat more rarely reported. Examples include plasmids that encode for functions essential for the host life-style (Brinkmann *et al.*, 2018; Michael *et al.*, 2016) and plasmids that encode functions that are indispensable for the host viability (Anda *et al.*, 2015; diCenzo *et al.*, 2013; Gil *et al.*, 2006). Major transitions in plasmid evolution are typically related to their invasion, persistence in the host population, adaptation to a new host, and, in some cases, evolution into secondary chromosomes (Hülter *et al.*, 2017) (see illustration in Figure 1).

Plasmid genome content comprises backbone genes that encode the plasmid replication and transfer mechanisms, and accessory genes that are generally considered as functions related to their persistence in the host population. Such functions may include catabolic enzymes (e.g., toluate 1,2-deoxygenase that plays a major role in toluene degradation and 2,4-dichlorophenoxyacetic acid degradation enzymes; reviewed in (Schmidt *et al.*, 2011)), genes encoding for antibiotic resistance mechanisms (e.g., beta-lactams, aminoglycosides, quinolones, sulphonamides and dihydrofolate reductase inhibitor; reviewed in (Porse *et al.*, 2016)), genes encoding for resistance to heavy metals (e.g., copper, sulfate, arsenate, arsenite, cadmium, zinc, cobalt, and mercury reviewed in (Dziewit *et al.*, 2015; Gullberg *et al.*, 2014)) or virulence genes (Couchman *et al.*, 2015; Hille *et al.*, 1984).

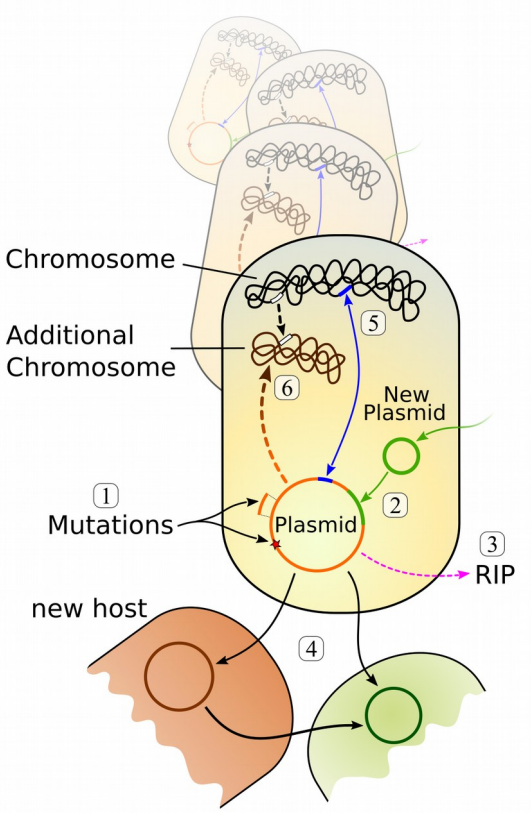


Figure 1: Major transitions in plasmid evolution. (1) Genome evolution (e.g., mutations and DNA acquisition). (2) Invasion occurs via lateral gene transfer mechanisms. (3) Plasmid extinction. (4) Evolution of the host range is largely determined by the plasmid replicon type and gene content. (5) Gene transfer from the plasmid to the chromosome may decrease the plasmid's significance for the host, while the translocation of chromosomal genes to the plasmid can lead to high persistence. (6) Transition of the plasmid into a secondary/additional chromosome. (Figure modified from Hülter N, Ilhan J, Wein T, Kadibalban AS, Hammerschmidt K, Dagan T. 2017. An evolutionary perspective on plasmid lifestyle modes. *Curr Opin Microbiol* 38:74–80; My contribution to that review was, in part, adapted in this thesis introduction; Illustration by Nils Hülter).

Plasmid backbone genes that are responsible for the plasmid replication and segregation appear to be generally highly conserved (although their annotation may vary (Thomas *et al.*, 2017)) . Additionally, plasmids may encode a "survival kit" that includes an active partitioning mechanism and a multimeric resolution system that ensures reliable inheritance of plasmids to daughter cells over generations (Baxter and Funnell, 2014; Zielenkiewicz and Cegłowski, 2001). Plasmid backbone genes may also include survival mechanisms that often rely on post-segregational killing of

plasmid-free cells (Naito *et al.*, 1995), for example, various toxin-antitoxin (TA) systems (Kopfmann *et al.*, 2016).

Self-transmissible plasmids encode the proteins required for their transfer and can be transferred via conjugation (Cabezón *et al.*, 2015) or – as shown for the Antarctic haloarchaeon *Halorubrum lacusprofundi* – also by outer membrane vesicles (OMVs) (Erdmann *et al.*, 2017). A previous large scale survey for the presence of genes encoding the conjugation machinery in plasmid genomes suggested that only ca. 25% of the known plasmids encode the full set of conjugation genes, hence they are likely self-transmissible via conjugation (Smillie *et al.*, 2010). Mobilizable plasmids encode a set of mobility genes that enable them to transfer in the presence of a conjugative plasmid (Ramsay and Firth, 2017). About half of the known plasmids are lacking mobility genes (as well as the genes required for conjugation) hence they are considered as non-mobilizable (Smillie *et al.*, 2010). Nonetheless, plasmids can invade bacterial cells by alternative transfer mechanisms including natural transformation (Morikawa *et al.*, 2012; Ramirez *et al.*, 2010), OMVs (Fulsundar *et al.*, 2014; Klieve *et al.*, 2005), or nanotubes (Dubey and Ben-Yehuda, 2011). Furthermore, plasmids may hitchhike with other mobile genetic elements such as phages during generalized transduction (Hertwig *et al.*, 1999) or with gene transfer agents (Scolnik and Haselkorn, 1984).

While plasmids are often considered as beneficial to their host, the discovery of bacterial resistance mechanisms against plasmids indicates that they are can also be considered as foreign (harmful) DNA just like phages. And just like phages, the plasmid invasion into a new host can be hindered by several resistance mechanisms. These include systems for the exclusion of invading plasmids (Cooper and Heinemann, 2000; Sakuma *et al.*, 2013), restriction modification systems (Roer *et al.*, 2015; Tock and Dryden, 2005) and the CRISPR-Cas system (Marraffini and Sontheimer, 2008). Those systems function in the degradation of foreign DNA, as well as the more recently described Wadjet system (Doron *et al.*, 2018) whose exact mechanism of action is yet unknown.

Plasmid genomes vary widely in size; they can be as small as 1,000bp, encoding only the bare backbone genes (Jørgensen *et al.*, 2014) and as large as 1.7Mb and include diverse accessory functions (Tett *et al.*, 2007). Notably, mobile

plasmids are typically larger than non-mobile plasmids, furthermore, plasmid acquisition with an antibiotics resistance gene seems to be associated with a larger plasmid genome size (Wein *et al.*, 2020). Evolution of the plasmid can occur at two levels of magnitude: single-nucleotide mutations or large-scale structural modifications. Nucleotide substitutions in the backbone genes have the potential to affect plasmid replication dynamics and adaptation to the host. Experimental evolution studies of plasmid adaptation to a naïve host were performed under selective conditions for the plasmid presence. Those studies showed that SNPs in the plasmid backbone can arise rapidly and may lead to host range modification (Fernández-Tresguerres *et al.*, 1995; Maestro *et al.*, 2003; Sota *et al.*, 2010). Moreover, it was found that adaptive mutations may also occur on the host chromosome (Harrison *et al.*, 2015; Wein *et al.*, 2019). Evolution of large-scale structural variants in plasmids is often observed in the neighbourhood of transposable elements, suggesting that such elements play a prominent role in plasmid genome evolution. Transposition-mediated acquisition – or loss – of genes has the potential to tremendously impact the plasmid fate, e.g., via the acquisition of a plasmid addiction system (Loftie-Eaton *et al.*, 2016) or loss of the conjugation machinery (Porse *et al.*, 2016). The transposition of transposable elements between plasmids can also lead to plasmid fusions (He *et al.*, 2015) and can underlie the evolution of plasmids that display modular characteristics through reshuffling of structural modules (Zaleski *et al.*, 2015). Plasmid fusions may also lead to the presence of multiple origins of replication in a plasmid, which can potentially prevent plasmid incompatibility (Chen *et al.*, 2014) and facilitate interaction with a broad range of hosts (Villa *et al.*, 2010).

Recombination among plasmids and other mobile elements appears to be frequent during plasmid evolution. This includes homologous recombination (Norberg *et al.*, 2011), as well as IS-elements and transposons (He *et al.*, 2015; Szabó *et al.*, 2016). The observation of prophages in plasmid genomes indicates that lysogenic phages can mediate large insertions into plasmids. This constitutes another mechanism for the gain of accessory genes in plasmids (Roux *et al.*, 2015). Those mechanisms constitute the basis of DNA exchange or transfer between replicons co-inhabiting the same prokaryotic host (plasmids and chromosomes). Large scale

network and comparative genomics based surveys found that although DNA flows mostly in between replicons of the same type (Halary *et al.*, 2010), there is clear evidence for gene transfer between plasmids and chromosomes during microbial evolution (Halary *et al.*, 2010; Xue *et al.*, 2015; Fondi *et al.*, 2010; Zheng *et al.*, 2015). Anecdotal evidence for DNA transfer between plasmids and chromosome within a host cell shows that DNA transfer might include essential genes such as tRNAarg and engA genes that are located on the pSymB mega plasmid of *Sinorhizobium meliloti* (diCenzo *et al.*, 2013), β-galactosidase (β-gal) gene in *Lactobacillus plantarum* strains (Fernández *et al.*, 1999), and key enzymes in the tryptophan and leucine biosynthesis pathways in *Buchnera aphidicola* where found located on plasmids rather than the main chromosome. (Latorre *et al.*, 2005). Moreover, transfer between plasmids and chromosomes was found to occur frequently in the presence of transposable elements (Fondi *et al.*, 2010; Hall *et al.*, 2017). Nonetheless, it is worth mentioning; studies have suggested that the translocation of a beneficial gene from the plasmid to the chromosome may lead to a decrease in the plasmid frequency in the population or even complete extinction, due to the trade-off between the plasmid costs and benefit to the host – a concept that was termed 'the plasmid paradox' (Harrison *et al.*, 2015; Stoesser *et al.*, 2016).

3.2 Comparative genomics methods

The comparison of two or more DNA or protein sequences for detecting sequence homology is usually achieved by sequence alignment (positional pairing of matching residues). Common practice of sequence alignment includes two main approaches; global alignment such as Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) where the aim is to find the alignment that scores highest similarity along the entire length of the compared sequences, and local alignment like Smith-Waterman algorithm (Smith and Waterman, 1981) that aims to find the best scoring sub-alignments that achieve an overall best score regardless the rest of the sequence. The sequence alignment is implemented under the evolutionary assumption of a common ancestor for the compared sequences along every column of the aligned residues.

A profound challenge in aligning whole genomes (pairwise or multiple alignment of the entire sequence of replicons) in contrast to aligning a single gene

family or genomic regions is the scale of evolutionary changes the genomes undergo. Hence, genome aligners are expected to achieve the solution of the maximal similarity along the entire genome and, at the same time, account for genomic rearrangements, duplications, insertions, deletions, segmental shuffling in addition to lateral gene transfer (Graur, 2016). Another basic sequence alignment problem arises when comparing sequences characterized by a large size range. As for our objective, I aim to compare between genomic sequences ranging in size between only one thousand base pairs up to millions of base pairs.

All common genome alignment methods depend on the seed and extend workflow where the aligners find short identical sequences as their seeds. Those seeds are used afterwards as anchors to divide the genome alignment into smaller set of alignments. The sub-alignments are then fed to global alignment algorithms for the extending step. Some of the most common genomic sequence comparison approaches include Mauve (Darling *et al.*, 2004; 2010), AVID (Bray *et al.*, 2003) and NUCmer (a package by MUMmer) (Kurtz *et al.*, 2004). Those methods are based on a similar workflow diagram with some differences in their approaches. While AVID uses the highest-scoring collinear chain of local alignments as its anchor and extends along the genome, Mauve uses maximal unique matches (MUMs) as anchors for its alignment, and views genomes as a set of collinear blocks instead of assuming that the genomes are collinear in their whole. Collinear blocks are homologous regions between two genomes without internal rearrangements within the blocks. This gives mauve the power for identifying rearrangements that are commonly observed throughout biological data. Mauve carries on by using CLUSTAL W approach (Thompson *et al.*, 1994) for their progressive global alignment of collinear blocks. NUCmer on the other hand, uses exact matches in the form of MUMmer hits (described below) as its seeds, it then clusters those matches (the clustering uses a constant distance threshold between MUMmer hits that is predefined by the user). The hits that were clustered into collinear chains are then filtered also upon a constant predefined length threshold, then Smith-Waterman algorithm is used to extend the blocks. Finding the best parameters for aligning two genomes using NUCmer is achievable. However, those parameters cannot be applied on an entirely different set of biological sequences with different patterns stemming from a different

evolutionary history. Hence, an adaptive threshold is needed for an automated recovery of sequence homology along large-scale diverse genomic datasets.

According to our working hypothesis, sequence similarity between plasmids and chromosomes should be restricted to high similarity regions that correspond to DNA transfer events. Considering the large size of bacterial chromosomes, I expect that such high similarity regions are quite patchy and surrounded similarity deserts i.e. regions that share no sequence similarity between plasmids and chromosomes. consequently, tools searching for local similarity are expected to be more appropriate for our objective. As genome aligners tested against a subset of our data found many false positives in the form of spurious regions of similarity between the compared replicons. This suggests that genome aligners mentioned above are more suited for comparing genomic elements with the assumption of a more recent divergence from a common ancestor.

Local similarity approaches can be classified into three main groups according to the type of algorithm at their core, (a) local alignment searches based on heuristic Smith-Waterman algorithm (as in BLAST and its variants (Camacho *et al.*, 2009)), (b) enumeration of common identical short sequences (k-mers) based on suffix tree algorithm as in MUMmer (Kurtz *et al.*, 2004) and there have been attempts to employee HMM based algorithms for DNA sequence homology search (Wheeler and Eddy, 2013)). Those approaches result in ‘hits’ that record local segments of similarity. An outstanding problem of using local similarity hits is that they underestimate the extent of regions of homology, and report multiple syntenic hits in what appears to be continuous regions. Such groups of hits can theoretically be joined to produce a parsimonious scenario requiring a minimal number of sequence transfers and acquisitions. Our preliminary results showed that these situations are common in plasmid-chromosome comparisons, and that existing global- or genome-alignment methods do not provide a satisfactory summary view of the event-wise syntenic context of the hits.

The visualization of sequence comparison poses an additional challenge. Global alignment approaches present the results of genome similarity as collinear blocks on a unidimensional axis depending on their genomic coordinates (Figure 2). Local alignment tools largely adopted the collinear presentation where they summarize the

local hits in a list of collinear blocks. Although there is no better alternative for viewing multiple aligned sequences or genomes, a better visualization for the comparison of a pair of replicons can be achieved by a dotplot (Gibbs and Mcintyre, 1970). A dotplot represents the compared genomes in a two-dimensional representation and the sequence homology falls into the 2D space, this allows for a clear observation of patterns such as duplications, indels, rearrangements and inversions (Figure 3). However, a major challenge in the dotplot visualizations is the balance between the amount of information and noise in the presentation of the results, especially on a nucleotide level sequence comparison (Sonnhammer and Durbin, 1995). And even after sufficient adjustments, dotplots are efficient tools for detecting patterns by naked eye but fail to report the observed information in a measurable manner for statistical analysis and automated pattern detection (Figure 3).

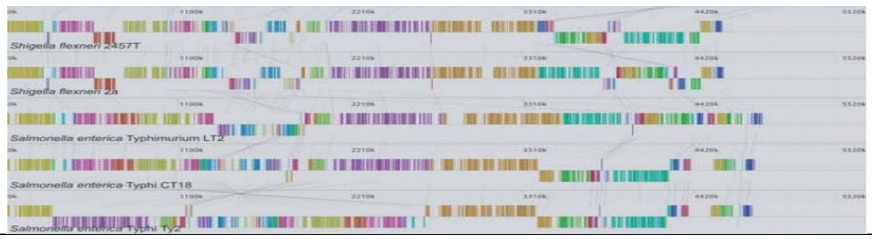


Figure 2: collinear blocks representation for sequence homology. Shown for multiple aligned genomes using MAUVE, adapted from (Darling *et al.*, 2004).

Hence, we concluded that the best way for observing homologous sequences detected from a pairwise genomic comparison is to plot the local similarity hits in a dotplot manner. Then we frame those hits in the 2D space as they were clustered by a downstream segmentation approach (we later refer to those frames as intersects) which can be quantified and characterized on a large scale of genomic data (see extended description in the materials and methods).

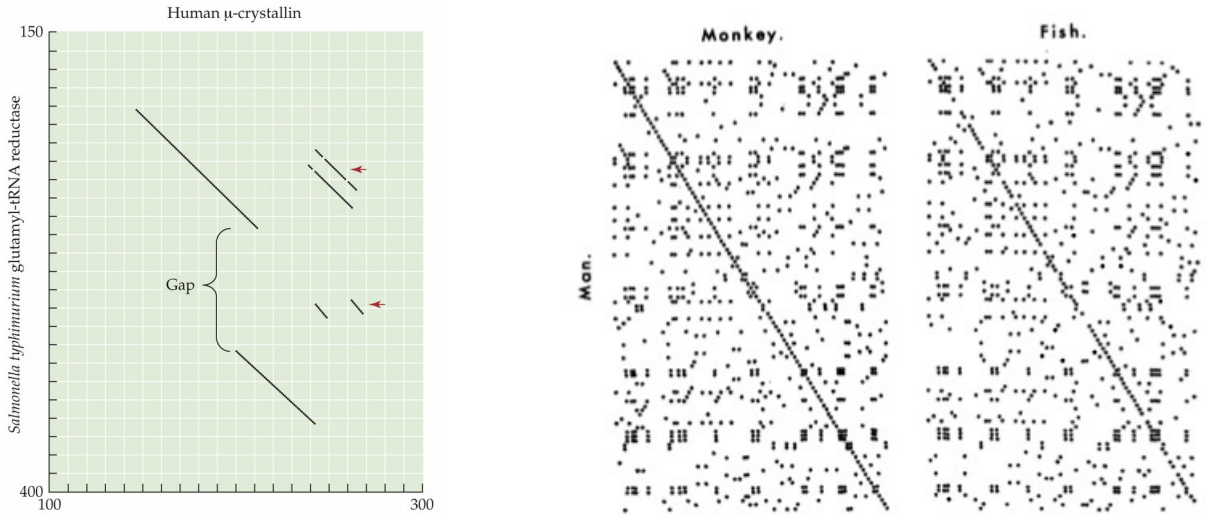


Figure 3: Dotplot matrix representation of sequence homology. The left figure shows the similarity between two genes, with incidents of duplications (pointed out by the red arrows) and a large indel detected by the gap (adapted from (Graur, 2016)). The right figure, shows dotplots of human cytochrome c compared with its homologs in monkey and fish, (adapted from (Gibbs and Mcintyre, 1970)).

3.3 Mechanisms of DNA transfer

3.3.1 Mobile genetic elements

Mobile genetic elements are DNA sequences that harbour genetic mechanisms, which enable them to relocate along genomes or duplicate or both. Mobile genetic elements include: Transmissible and mobile plasmids, bacteriophages (prophages), transposable elements (including integrative and conjugative elements ICEs), insertion sequences (IS elements) and integrons that are usually physically linked to other mobile elements. For our research objective, we study the DNA transfer between replicons within the microbial host, thus we describe plasmids in their static phase, as we study their evolution and contribution to genome evolution while they are in their microbial host. Here, we will use the term mobile elements to refer to the other four types.

Transposable elements were first described by Barbara McClintock in the 1940s as she observed differential colouring in maize kernels and after investigating, she deduced the presence of genetic elements that can jump between genomic locations. Transposition can take place within the genome or between different

genomes and it does not require any level of sequence homology between the target and donor genomic locations. Transposition locations can be characterized by inverted repeats flanking the transposable element. Those repeats are results of the DNA repair mechanism filling the gaps between the sticky end cleavage sites and the newly inserted transposable element (Graur 2016).

Three major mechanisms dictate the movement and copy number a transposable element:

(a) Conservative transposition (cut and paste model) – this model can be mediated by one of multiple transposase enzymes, namely, DDE transposase, Y transposase and S transposase, where the double stranded DNA is removed from its original location and inserted into the target (Figure 4).

(b) Replicative transposition (copy and paste model) – this mechanism is mediated by transposase of the type Y, the original DNA is replicated and the single stranded DNA is inserted to the target, then a cDNA is replicated to pair with the newly inserted single stranded DNA.

(c) Retrotransposition (RNA mediated transposition model) – this model is mediated by DNA dependant RNA polymerase (DDRP) that transcribes an RNA sequence out of the donor DNA, then a reverse transcriptase enzyme transcribes the RNA sequence back into a DNA sequence, and finally transposases of either type DDE or type Y mediate the final step of the transposition into the target location. Note that in the first model, the copy number of the transposable element remains the same, while in the other two models, the copy number increases with every event of transposition (Graur 2016).

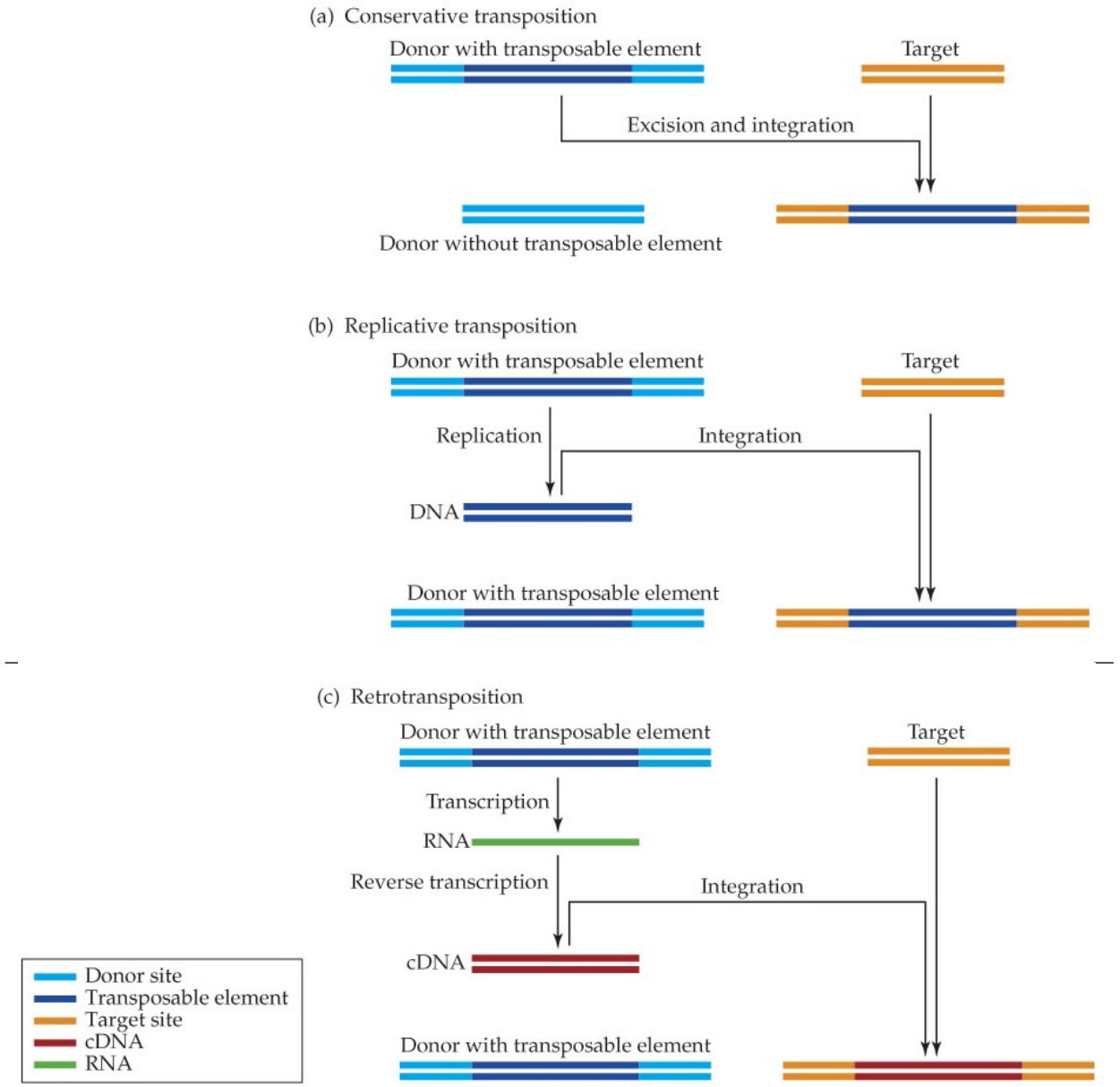


Figure 4: The three mechanisms of transposition. Steps in the excision and integration of transposable elements are illustrated. The figure was adapted from (Graur 2016).

Transposable elements can carry important genes like antibiotic resistance genes and many others, and can even be a part of transposable bacteriophages. While insertion sequences (IS elements) are the simplest form of transposable elements, as they only encode for their transposition mechanism, and are usually annotated as IS followed by a number referring to its type (Graur, 2016) (Figure 5)

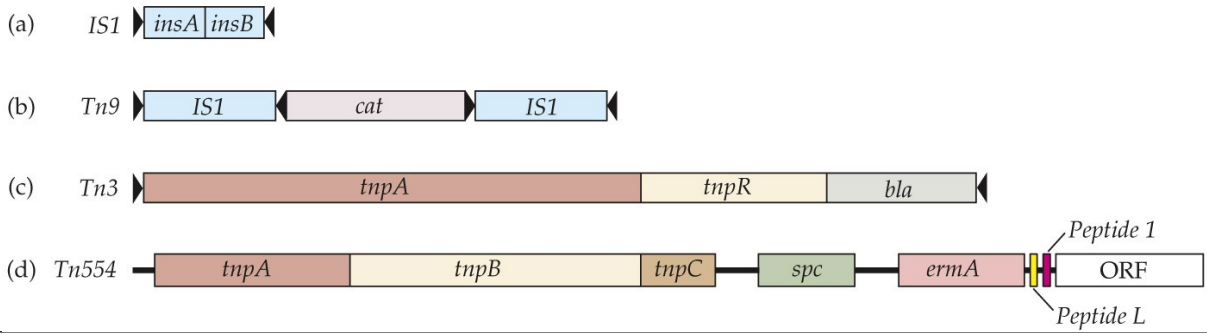


Figure 5: Structure of transposable elements. Different genes can be carried by transposable elements, while IS elements only carry transposition genes. The figure was adapted from (Graur 2016).

Integrons are genetic elements that assemble and carry a collection of coding sequences and ensures the functionality of those genes through a proper expression (Mazel, 2006). Integrons are present in many bacterial genomes (bioinformatics surveys found them in 10% to 17% of all studied genomes (Domingues *et al.*, 2012)) inhabiting different environments. They carry a tyrosin recombinase gene (also annotated as integrase) that determines the site-specific integration, a recombination site recognized by the integrase (*attC*) and many integrons carry the promoter (P_c) upstream of the recombination site.

Two types of integrons can be identified, mobile integrons, usually attached with other mobile genetic elements mediating their mobility and super-integrons that are continuously growing by adding more open reading frames to their fixed location (Figure 6) (Mazel, 2006; Domingues et al., 2012).

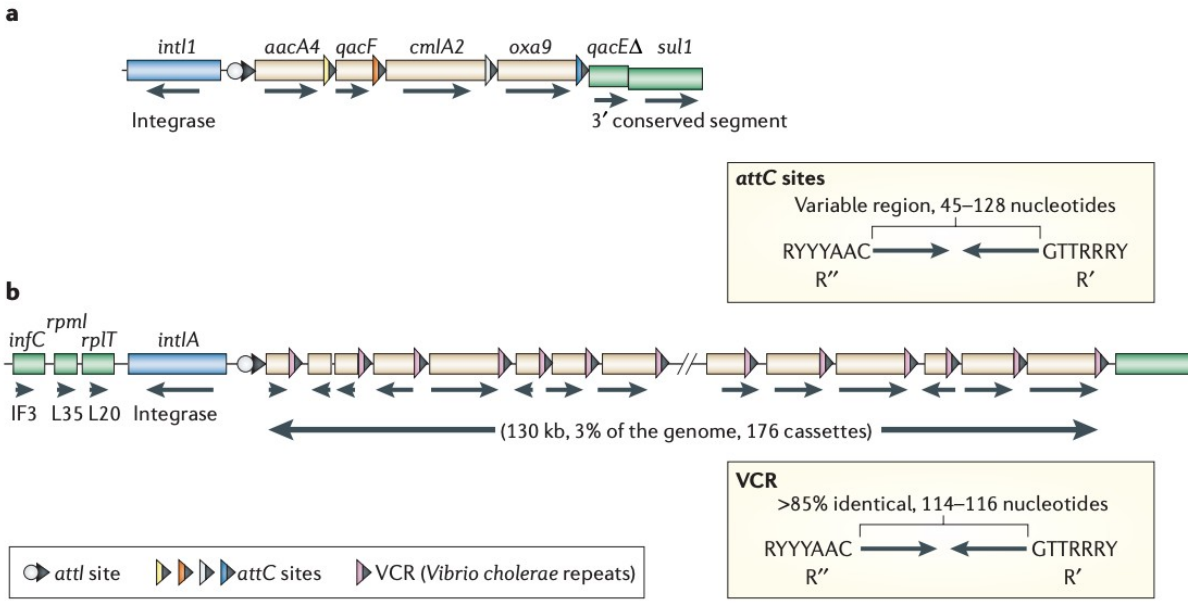


Figure 6: The two types of integrons. The mobile integron is shown on top and the superintegron on bottom, the figure is adapted from (Mazel, 2006).

Temperate phages (A.K.A lysogenic phages) integrate into prokaryotic host replicons (plasmids or chromosomes), this creates a prophage in the host genome. The phage stays within the host and replicates with it until it enters its lytic cycle. In this stage, the phage starts using the host machinery to produce new phages and release them after the host cell lysis. The new phages can then invade a new host transferring genetic material that it might have carried from its previous host (Popa *et al.*, 2015).

3.3.2 Homologous recombination

Unlike site specific recombination, homologous recombination does not need a recognition site for DNA-binding proteins and can happen between any two sites with sequence homology along the replicons (Camerini-Otero and Hsieh, 1995). Many genes (around 25 in *Escherichia coli*) are involved in the process of homologous recombination, which happens in four steps; (a) the initiation (processing), (b) the homologous pairing and DNA exchange, (c) DNA heteroduplex extension (branch migration), and (c) the resolution (those steps and their mediating enzymes are demonstrated in Figure 7).

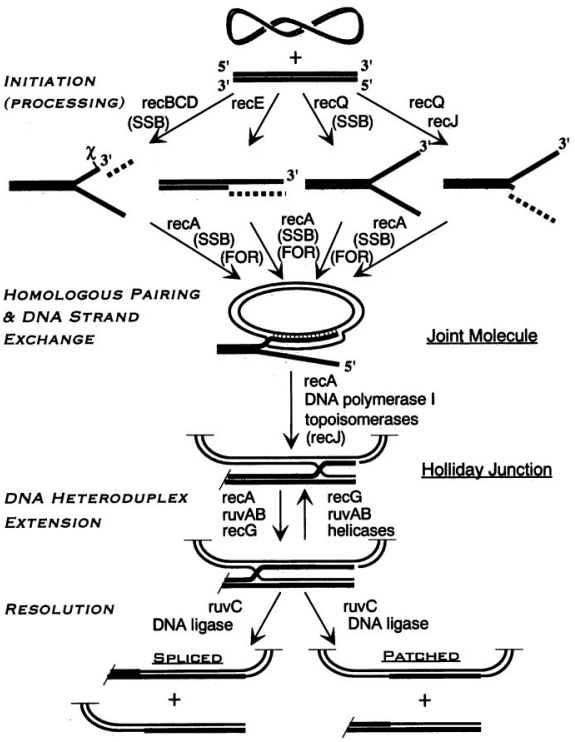


Figure 7: Process and biochemical model for homologous recombination. (The figure was adapted from (Kowalczykowski *et al.*, 1994)).

4 OBJECTIVES

In this thesis, I aimed to find the extent of DNA transfer between plasmids and chromosomes that co-inhabit the same prokaryotic host. DNA transfer between plasmids and the chromosome of their host constitute the first and last steps of plasmid mediated DNA transfer between different prokaryotes. To gain a comprehensive view on the extent of transferred elements, I aim to detect and characterize the homologous genomic regions between plasmid and chromosome replicon pairs and test for patterns and regularities in the homologous region properties.

5 DATA

Our database comprises 4,700 completely sequenced prokaryotic isolates downloaded from NCBI Genbank (Geer *et al.*, 2009) (version of January 2016) along with their annotation files. The majority of sequenced isolates in the data, 3,209 out of 4,700 (68%), do not harbor plasmids. Those strains belong to 486 genera. Which means that 60% of the total 818 genera have strains with no plasmids. A total of 294 isolates in the data have more than one chromosome, among which, 203 isolates carry plasmids and 91 do not. We excluded those isolates from our data set, in order to avoid dealing with the possible influence of a second chromosome in the host to the DNA transfer between plasmids and the chromosome, and to keep our data structure simpler. The remaining genomes include 1,400 (30%) isolates that have a single chromosome and at least one plasmid, those isolates are classified into 332 genera and comprise 3,264 pairs of chromosome-plasmid that co-inhabit the same isolate; in other words, both replicons were documented in the same prokaryotic host population.

The genome size of the plasmids in my data ranges between 744bp and 2.7Mb with a median of 49Kb, while the size of chromosomes ranges between 48Kb and 10.6Mb with a median of 4.3Mb. Thus, plasmid size range spans over 5 orders of magnitude, while chromosome sizes in the data ranges over 3 orders of magnitude. The majority of isolates in our dataset (660 isolates (47%)) harbour a single plasmid, with 342 (24%) isolates harbouring two plasmids, and 398 isolates (28%) have between 3 and up to a maximum of 21 plasmids (Figure 8).

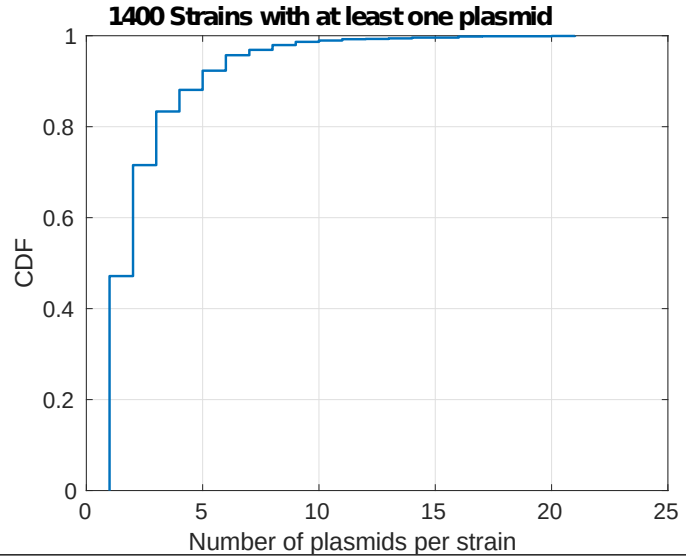


Figure 8: The cumulative distribution of the number of plasmids per isolate (strain). The distribution was calculated for 1,400 completely sequenced isolates. Isolates with no plasmids are excluded.

6 METHODS

6.1 Definition of terms

In this thesis, we use a set of terminologies, some of which might be familiar, and some we had to adapt to describe specific data structure and methodological steps that we created. The terms listed below are demonstrated in Figure 9.

Host: In this study, a completely sequenced prokaryotic isolate that has one chromosome and at least one plasmid (Figure 9.1).

Replicon: A self-replicating genetic element, a chromosome or a plasmid in this study (Figure 9.2).

Hit: A sequence of at least 20 nucleotides that is shared between two replicons with a minimum similarity of 80%, detected by local similarity methods (BLAST or MUMmer) (Figure 9.3).

Hit dot-plot: A two-dimensional representation of the shared local similarities (hits) between two replicons, where hits are plotted using their coordinates on one replicon as the x-axis (the chromosome in this study) and on the other replicon as the y-axis (the plasmid in this study) (Figure 9.4).

Segment: A continuous region on one replicon that is enriched with hits (Figure 9.5, 2.6).

Segment (an alternative definition): An inclusive continuous region of DNA on a replicon with sequence homology to another replicon allowing for a higher divergence than what is detected by local similarity hits.

Locus with homology: In the curated data, we refer to plasmid segments as plasmid loci with chromosomal homology and similarly, we refer to chromosomal segments as chromosomal loci with plasmid homology, to express the assumption of a shared origin for those loci.

Intersect: A set of hits that belong to the same segment on each of the two replicons, an intersect resembles an event of transfer or alternatively a duplication or rearrangement event of a transferred region of DNA (Figure 9.7).

Intersect (an alternative definition): A framed area from the 2D space of DNA sequence similarity between the two replicons that is an intersection between two segments on the two replicons, we consider only intersects that have hits falling into them.

Homologous pairs: In the curated data, we refer to an intersect as a homologous pair of loci.

Pattern: An arrangement of a set of hits or intersects that can be characterized and identified using their properties and represents an underlying evolutionary scenario.

Sequence Hit-Density: The proportion of a sequence (e.g. an entire replicon or a segment) that is shared with the other replicon (belongs to at least one hit).

Depth: A positional measure for the number of hits that include a given nucleotide from a replicon (the cumulative sum of hit projections on an axis (replicon)).

Intersect fullness: The proportion of hits belonging to an intersect to the size of the segment on one replicon. (hit projection size divided by segment size).

Similarity: The proportion of identical nucleotides between two sequences constituting one hit. For intersect and segment similarity, we use the median BLAST similarity of hits belonging to this intersect or segment.

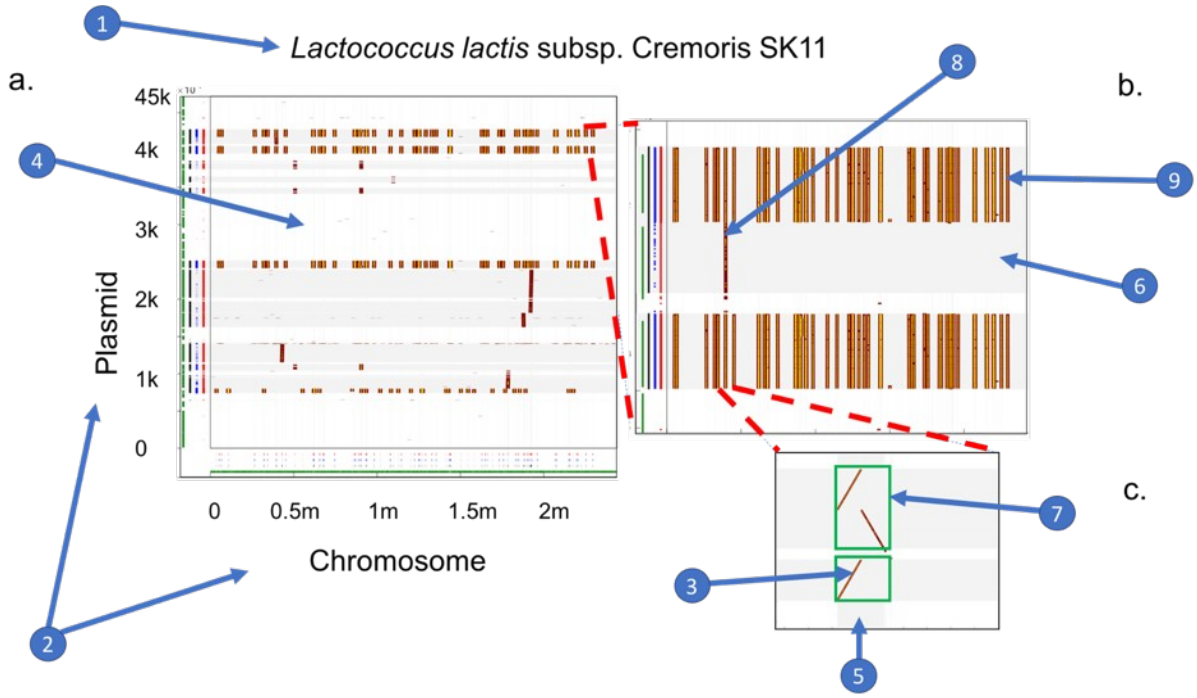


Figure 9: An example of a 2D hit-plot, depicting the different levels of sequence comparison between the two replicons, and a closer look at selected regions (zoom).
1. Prokaryotic host (genus, species, strain), 2. Chromosome-plasmid pair, 3. Local similarity hit (BLAST), 4. Hit dot plot, 5. Chromosomal segment, 6. Plasmid segment, 7. Intersect, 8. Pivot intersect, 9. Intersect echoes.
a: A hit-dotplot of the chromosome and one plasmid of *Lactococcus lactis.*
b: Zoom-in on the plasmid, focusing on two shared regions on the plasmid with many repeats on the chromosome, and a clear pivot intersect.
c: Zoom-in on the chromosome, focusing on the pivot intersect and showing a pattern for an inversion that took place after the transfer.
Notice that in all our hit-dotplots like this figure, local similarity hits are plotted in red colour for BLAST hits and yellow for MUMmer hits. The gradient of red colour for BLAST hits corresponds to their similarity score, with darker red corresponding to higher similarity. The MUMmer hits are plotted on top of the BLAST hits with a narrower width thus overlapping hits from different types can be visualized.
BLAST hits are projected on each axis in red lines, and MUMmer hits are projected on the axes in blue lines. Segments are represented on each axis in black lines, and coding sequences are represented in green lines. The extension of segments on the 2D space is highlighted in grey and the intersection of two segments from the two replicons gets a darker grey highlight colour.

Pivot intersect: The intersect with the highest fullness (pivot-f) or the highest similarity of the largest hit (pivot-s) among the group of intersects belonging to its segment on one replicon (the plasmid in our study case). The number of pivot intersects resembles the minimum number of transfer events between the replicons (Figure 9.8).

Echoes: Multiple intersects belonging to the same segment on a replicon with lower fullness or similarity than the pivot intersect, they are allegedly results of genomic duplications and rearrangements or multiple transfers of the same DNA region (Figure 9.9).

Genomic erosion: The evolutionary processes that change the original sequence of a locus on the replicon and makes it diverge from its ancestral form and from sequences sharing its common ancestor.

6.2 Detection of local sequence similarity

To find shared sequences within plasmid-chromosome pairs we conducted a comprehensive sequence similarity search of plasmid sequence against the chromosome sequence using two tools: BLAST (Zhang *et al.*, 2000) and MUMmer (Kurtz *et al.*, 2004). BLAST is powerful in detecting sequence similarity while allowing for a certain frequency of mismatches and gaps (according to a threshold), while MUMmer is useful for the detection of identical sequences only and finding small repeats such as CRISPR arrays and microsatellites. By combining hits from BLAST and MUMmer analysis, we could cover different types of shared sequences that would be ignored otherwise, as BLAST misses short separated sequences and tandem repeats and MUMmer is not sensitive for detecting divergent sequences. Aiming for a conservative estimate of sequence similarity, we further filtered the resulting hits using thresholds of ≥20bp hit length and ≥80% sequence similarity (i.e., the proportion of identical nucleotides in the hit). The analysis of the remaining hits was performed using in house scripts in MatLab©, Linux shell Perl and MySQL© (see Table 1 for a full list of scripts in the pipeline).

Table 1: a list of scripts used in this study

Script function	**Programming language**
1- Run a sequence against sequence BLAST search with the plasmid sequence as query and the chromosome sequence as a database.	Perl
2- Run plasmid against chromosome sequences MUMmer analysis	Perl
3- Data transformation	Linux shell, MySQL
4- Simulate 1,000,000 binary sequences with similar characteristics (length and density) as segmented biological sequences resulting from BLAST and MUMmer analysis	MATLAB
5- Segmentation pipeline:	Snakemake (Köster and Rahmann, 2012).
5.1 Translate BLAST and MUMmer hits on plasmid and chromosome sequences into binary sequences	MATLAB
5.2 Run the segmentation and filtration on each replicon sequence for each chromosome-plasmid pair	MATLAB
5.3 Detect the intersects between plasmid and chromosome segments for each pair	MATLAB
6- Plasmid vs chromosome sequence homology dotplots with representation for hits, segments, intersects and coding sequence	MATLAB
7- Post segmentation downstream analysis	MATLAB and MySQL
7.1 Characterize and categorize transfer data	MATLAB and MySQL
7.2 Find transferred coding sequence	MATLAB and MySQL
7.3 Find co-transferred coding sequence	MATLAB
7.4 Find transferred AMR genes	RGI (Alcock *et al.*, 2020) and MATLAB

After preliminary examination of the resulting data, we excluded 6 chromosome-plasmid pairs that showed extremely high number of local similarity hits resulting from both BLAST and MUMmer, the number of hits seemed exaggerated and it could correspond to assembly errors. However, the exclusion of those 6 pairs has not affected the number of strains in the dataset and we ended up with 3,264 chromosome-plasmid pairs.

In the 1,491 strains that carry plasmids we found a total of 1,795,221 BLAST hits and 1,810,219 MUMmer hits between the plasmids and chromosomes co-inhabiting the same strain. The further exclusion of isolates with multiple chromosomes resulted in a further decrease in the total number of shared sequences. Among the 3,264 plasmid-chromosome pairs belonging to 1400 isolates with one chromosome and at least one plasmid, we detected 1,120,750 BLAST hits and 1,195,942 MUMmer hits ranging between 0 and 40,250 BLAST hits per pair with a median of 61 hits and between 0 and 35,623 MUMmer hits per pair with a median of 34 hits (Figure 10).

Using spearman ranking correlation test we found a significant correlation between the number of BLAST hits and number of MUMmer hits within a pair and a positive correlation between the number of hits and plasmids size (Table 2).

Table 2: spearman rank correlation for number of hits per pair

	r	P value
Number of BLAST vs number of MUMmer hits	0.9	$<2.2 \times 10^{-16}$
Number of BLAST hits vs plasmids size	0.9	$<2.2 \times 10^{-16}$
Number of MUMmer hits vs plasmids size	0.8	$<2.2 \times 10^{-16}$

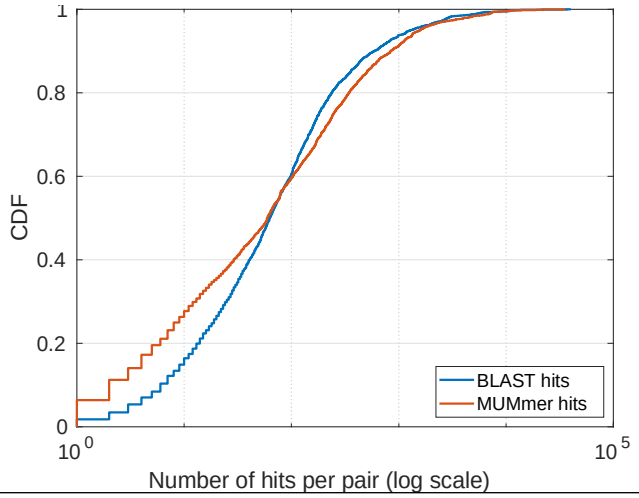


Figure 10: The cumulative distribution of the number of hits per plasmid-chromosome pair. Calculated for 3,264 plasmid-chromosome pairs.

My results show that 69% of plasmids in the dataset share less than 10% of their genome sequence with the chromosome of their host, as found by BLAST and MUMmer hits combined (as a union) (Figure 11). While 40 plasmids (1.2%), belonging to 11 genera (5 *Corynebacterium*, 2 *Lactobacillus*, 2 *Arthrobacter*, and others), share more than half of their genome sequence with the chromosome. A total of 17 of plasmids share 100% of their genome with the chromosome. Two of those plasmids that are almost entirely shared with the chromosome are carried by the same *Shewanella* strain, one of them is a small plasmid of 7,995bp and the other is larger with 116,763bp. The same strain has two other plasmids whose sequences are partially shared with the chromosome (see details in section 5.5 in the results).

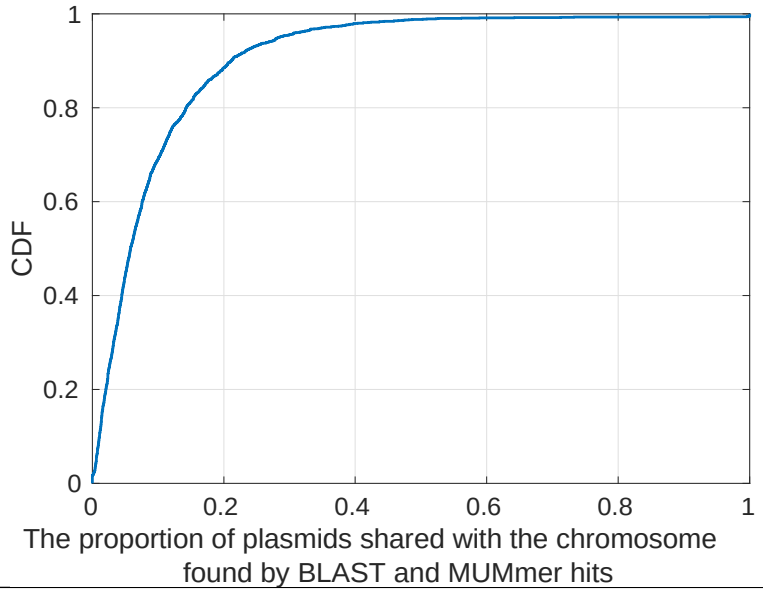


Figure 11: The proportion of plasmid shared sequence with the chromosome. This was calculated from the total shared sequence detected by wither BLAST, MUMmer or both for 3,264 plasmids in my dataset.

In several plasmids we did not find any single BLAST or MUMmer hit, including 54 plasmids (1.6% of the dataset) within 54 different strains belonging to 23 genera (18 *Arthrobacter*, 11 *Escherichia*, 5 *Klebsiella*, and others). However, other plasmids belonging to those strains (between 1 and 9 plasmids per strain) share a proportion of their sequence with the chromosome (up to 100%) (see section 5.6 in the results), hence not sharing sequence with the chromosome might be related to the plasmid type and not a property of the species per se.

A close examination of several plasmids whose genome shared most sequence similarity with the chromosomes revealed an interesting phenomenon. Very often the sequence similarity was detected as multiple hits of varying sequence similarity (see Figure 12). Indeed, these observations may be due to artefacts of genome assembly. Nonetheless, such pattern was also observed in genomes that are completely sequenced (i.e., 'closed') and genomes sequenced with PacBio (see section 5.5 in the results), which produces long reads and hence reduces the risk for mis-assembly artefacts. Consequently, we conclude that the pattern we observed is the result of genuine biological processes. Example scenarios are multiple DNA transfer events from the plasmid to the chromosome, or, alternatively, a single transfer event followed by a gradual (and variable) degradation of sequence similarity

(e.g., due to neutral selection regime). Thus, while the hits data from BLAST and MUMmer is useful in identifying the plasmid genome sequence shared with the chromosomes, for the inference of transfer events one has to join neighbouring hits into transferred segments. For that purpose, I developed a novel approach to join neighbouring sequence similarity hits into segments of shared sequence.

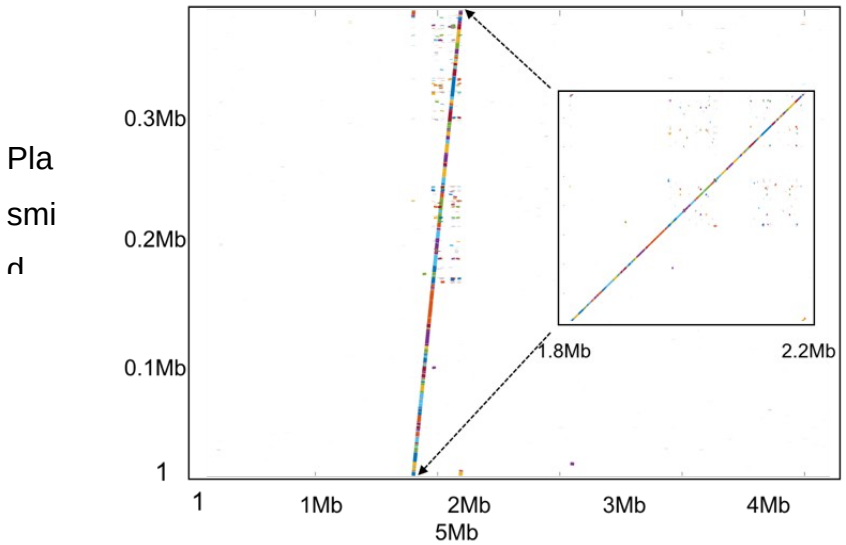


Figure 12: An example of multiple hits that form a continuous shared locus between the plasmid and chromosome. Hits are shown as explained in Figure 9, each hit is coloured differently to show that sequence similarity search tools identify smaller regions of sequence similarity. The final result for this example is demonstrated in Table 11,3.

6.3 Segmentation of sequence similarity data

In order to join neighbouring local sequence similarity hits, we have implemented a segmentation method that works on binary sequences and divides them into segments that are significantly enriched with one state (hit or no-hit). During the segmentation approach, we define a position as the space between two nucleotides. The method runs on each position of the sequence by comparing the density of one character over a predefined window size to the left with the same window size to the right (average density of the character to the left - it's average density to the right), and assigning the score of the calculated difference to each position of the sequence. We call the scoring vector differential density (*dd*). High *dd* implies that the position has a different density of the specific character on its two sides, and low density

means that this position is in the middle of homogeneous distribution of the character. The differential density (*dd*) vector has the same length as the original sequence for circular sequences (in our case of circular replicons) and a length of (sequence length - (2 $\times$ window size)) for linear sequences. In order to pick a window size for our application, we tested multiple window sizes of 200bp, 500bp and 1000bp. The character states we have in this study is (hit) for positions that belong to either a BLAST or a MUMmer hit or both, and (no hit) for positions missed by both BLAST and MUMmer. We translate the (hit / no hit) states into binary code with 1 for hits and 0 for no hits. We define true positives as positions that belong to both local similarity hits and the final produced segments, and false negatives as those belonging only to the hits. We found that the 1000bp window size produced the highest number of true positives with no large difference with the false negatives comparing to the 200bp and 500bp window sizes. Consequently, we chose the window size of 1000bp because it seemed to have the highest sensitivity with no big difference in its specificity.

The segmentation was implemented to function in a recursive manner, by splitting the sequence at the position with the local maximum *dd* values, those positions of local maximum *dd* should satisfy two criteria before being selected as segmenting points: They should be above a certain threshold and in a transition between one character to the other (0 to its left and 1 to its right or 1 to its left and 0 to its right). Segmenting at a selected position produces two sub-sequences that have the same structure as the original sequence, the two daughter sequences are fed to the segmentation again and again recursively until the termination condition is met. The segmentation terminates when the maximum *dd* is below the threshold (Figure 13).

In order to estimate a *dd* threshold, we simulated multiple binary sequences with similar characteristics (length and density) as the segment resulting from biological sequence homology data (Figure 14). We calculated the random maximum *dd* for each simulation, thereafter we could define a threshold using a certain alpha, as we chose an alpha of 0.01. Calculating the *dd* vectors for hundreds of thousands of long simulated sequences (max plasmid segment length = 393,620 bp) is very computationally extensive and time-consuming, even when we use functions for

cumulative sum of vectors, hence, we have used a convolution on vectors function. The convolution function is implemented on a very basic level of the computer hardware, that's why using it reduces the time spent in data exchange, and consequently reduces the time of the segmentation drastically.

The convolution operates on two vectors, the first vector is the binary sequence to be analysed and the second is a kernel that we designed to fit our purpose.

The Kernel is a vector of (-1) for one window size (1000 in our case) followed by a vector of (1) for one window size.

The convolution (*C*) of a position (*x*) on a sequence (*S*) with a size (s) using a kernel (*K*) with a size (k) is:

$$C(x)=\sum_{n=x}^{k+x}\left(S(n)*K(x-n+1)\right).$$

Eventually, the vector C represents the differential coverage (*dd*).

The convolution was applied to the central part of the sequence, which means:

C has the size (c=s-k+1).

Two predictor variables, the segment length (len) (in a logarithmic scale) and the sequence hit density (den) have been used to simulate the sequences and estimate the random *dd* values as the response value. The 2d plane between the two predictors was uniformly divided into 10,000 cells. Where each cell represents a combination of two values, one of each predictor (len and den) that are projections of the centre of the cell on the axes. Thereafter, 100 simulated random sequences were produced using the predictor values of each cell (total of 1000,000 sequences were simulated) and the maximum (*dd*) was recorded for each simulation (Figure 14 and Figure 15). By choosing a certain alpha (0.01 for instance) a threshold could be defined as the (99th) percentile of the simulated maximum *dd* values for one cell. So, each cell ends up with a threshold value that is higher than 99 of the 100 *dd* values produced randomly. For a given segment that has characteristics (length and coverage) within the range of a certain cell. If a position on that segment was found to score a *dd* above the threshold of this cell, this means that this position can split the sequence into two regions that are significantly different in their density and it is a valid position as a segmentation site. In other words, if hits are clustered on one part of the segment and do not follow a random distribution as expected from simulations,

there will be a significantly high *dd* for some positions that can split this segment into regions of high or low hit density. On the other hand, if the distribution of hits is uniform along the segment, every position will have a homogeneous density around it and no further segmentation is due. The resulted matrix of simulated *dd* can be used to predict any threshold using the two predictor measures, which makes the segmentation method independent from our biological data and robust to be used for different studies.

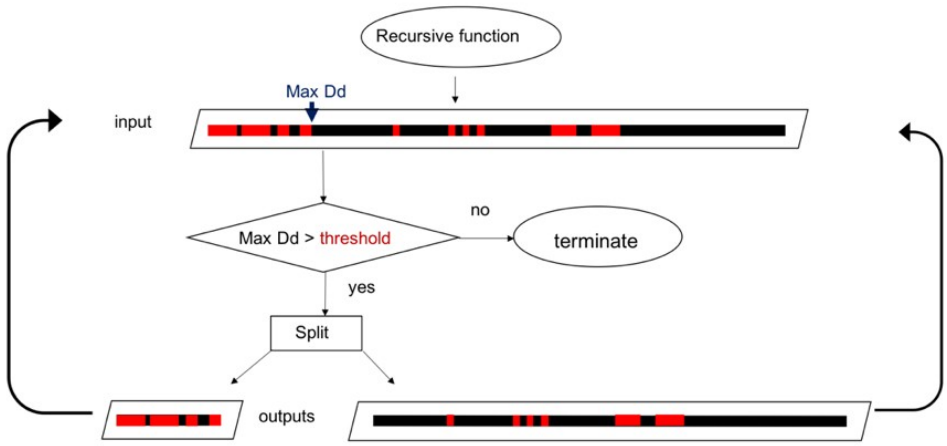


Figure 13: An illustration of the recursive segmentation approach. Horizontal bars represent a genomic locus with red and black corresponding to hit (1) and no hit (0) states respectively. The recursive segmentation aims to find the most parsimonious splitting position yielding segments that include high density of hits.

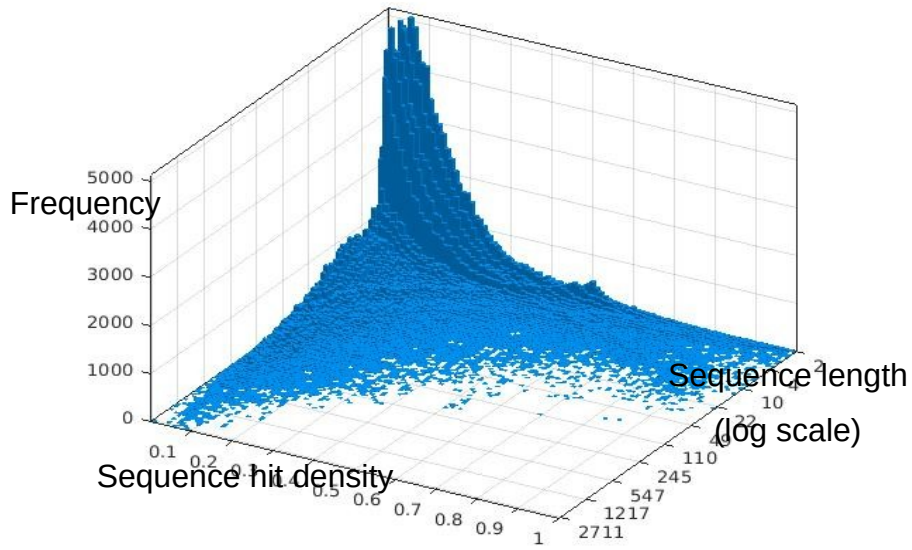


Figure 14: The empirical distribution of sequence length and hit density in our data. In order to determine the parameter space for the *dd* simulation, we recorded the sequence hit density and sequence length observed by the segmentation algorithm for genomic data. The result show that the decision points of the segmentation algorithm are frequency made for short sequences with low hit density. Using the results from this analysis we determine the parameter space for the simulations presented in Figure 15.

We applied the segmentation method on both the plasmid and chromosome of each chromosome-plasmid pair. This resulted in a total of 72,499 plasmid segments and 120,662 chromosomal segments. The segmentation procedure resulted in segments observed in 2,319 (71%) of the plasmid-chromosome pairs.

The segmentation provides a unified framework to join the two different types of local similarity hits, produced by BLAST and MUMmer. A comparison of the resulting segments to the BLAST and MUMmer data shows that out of 1,120,750 BLAST hits, 724,375 (65%) are included in the plasmid segments and 649,988 (58%) are included in the chromosome segments. Out of 1,195,942 MUMmer hits, 991,846 (83%) are included in the plasmid segments and 964,802 (81%) are included in the chromosome segments. Hence, the proportion of hits included in segments in each method is similar between plasmids and chromosomes. Furthermore, more MUMmer hits are included in segments since they are shorter than BLAST hits.

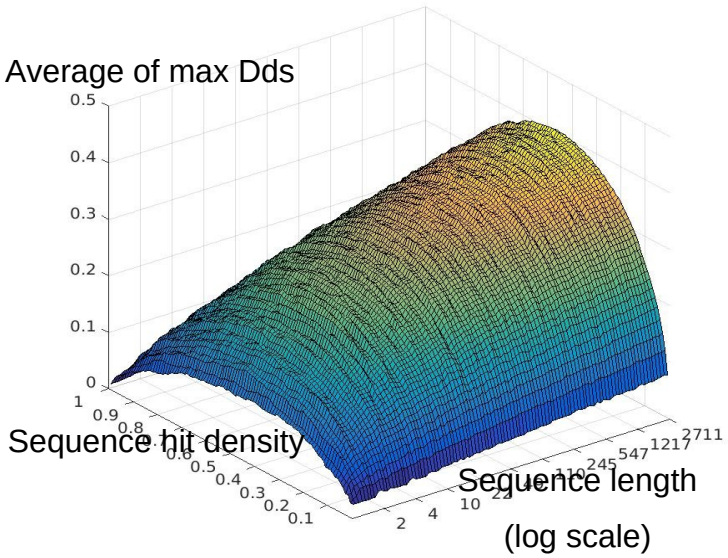


Figure 15: Surface plot of the mean *dd* values of 100 simulations for each of the 10,000 parameter combinations of sequence length and hit density. The resulting distribution of average *dd* reveals three main properties:

1. The *dd* value is dependent on both parameters, which justifies the use of an adaptive threshold for each different parameter combination.
2. The surface distribution is symmetric around 0.5 sequence density axis, this makes sense because the *dd* measure could be used in the absolute value comparing the density over the left and right flanking regions for a position.
3. Additionally, we observe an increasing dd value with sequence length.

Hits that are not included in segments are likely spurious sequence similarity patterns that cannot be grouped into a segment. The difference between BLAST and MUMmer inclusion in segments indicates that BLAST results in more spurious hits in comparison MUMmer. Indeed, we find that the E-value of BLAST hits not included in segments is significantly higher in comparison to E-values of BLAST hits included in segments ($P<2.2 \times 10^{-16}$, using KS test).

Testing the difference between segments and non-segmented BLAST hit E-values per plasmid-chromosome pair shows that in 100% of the pairs that have BLAST hits both in and outside plasmid segments (2,263 pairs), the E-value is indeed larger and the size of hits is smaller for BLAST hits not included in segments

(alpha=0.05, using KS-test and FDR). Furthermore, the comparison of MUMMER hit length between hits included or excluded from segments shows that the size of MUMmer hits is significantly larger inside segments for 100% of the pairs that have MUMmer hits both in and outside plasmid segments (1,977 pairs). Thus, shorter MUMmer hits are also likely to be spurious sequence similarity.

Notably, in 891 (27%) plasmid-chromosome pairs we observed hits but no resulting segments. Importantly, by finding hits that are clustering in a proximity and joining them into more parsimonious segments, we managed to reduce the complexity of the data such that the frequency of potential transfer events is reduced by 9.7 folds on the chromosome and 21 folds on the plasmid (from 1,120,750 BLAST hits to 120,662 chromosomal segments and 72,499 plasmid segments). Our segmentation approach is thus useful in reducing noise due to spurious sequence similarity, and in re-joining detected hits.

6.4 Characteristics and distribution of segments

While our approach can be used to segment BLAST hits on both replicons in our data (i.e., plasmids and chromosome), considering our research objectives, here we focus on plasmids. Segments in plasmid genomes are plasmid loci that share a high sequence similarity with chromosomal loci, thus they are putatively the result of LGT between plasmids and chromosomes. The number of segments per plasmid (i.e., plasmid-chromosome pair) ranges between 1 and 689 segments with a median of 13. To further characterize the performance of the segmentation approach, we examined the density of blast hits per plasmid segment. The segment hit density is calculated as the total number of nucleotides included in the hits (both BLAST and MUMmer, combined) per segments divided by the segment length.

Table 3: Descriptive statistics of plasmid segments in the total dataset (2,319 pairs). In this table, we calculate segment statistics using BLAST hits only, segments based on only MUMmer are expected to be shorter. The median segment depth of 1 means that at least 50% of the segments have a single copy on the chromosome and the median density of 1 means that at least 50% of segments are fully covered by hits.

	Min	Max	Mean	std	CV	Median
Segment size	20	393,620	572.28	2,308.96	4.03	103
Hit density per segment	0	1	0.84	0.30	0.36	1
Segment depth	1	1,082	3.28	14.32	4.36	1
No. hits per segment	1	8,456	13.35	76.01	5.69	3

Table 4: Descriptive statistics of chromosome segments in the total dataset (2,319 pairs). Similarly to the plasmid segment statistics, in this table, we calculate segment statistics using BLAST hits only, segments based on only MUMmer are expected to be shorter. the median segment depth of 1 means that at least 50% of the chromosomal segments have a single copy on the plasmid and the median density of 1 means that at least 50% of chromosomal segments are fully covered by hits.

	Min	Max	Mean	std	CV	Median
Segment size	20	1,180,829	790.82	6,053.93	7.66	433
Hit density per segment	0	1	0.83	0.30	0.37	1
Segment depth	1	88	2.05	2.98	1.45	1
No. hits per segment	1	8,456	133.59	458.15	3.43	40

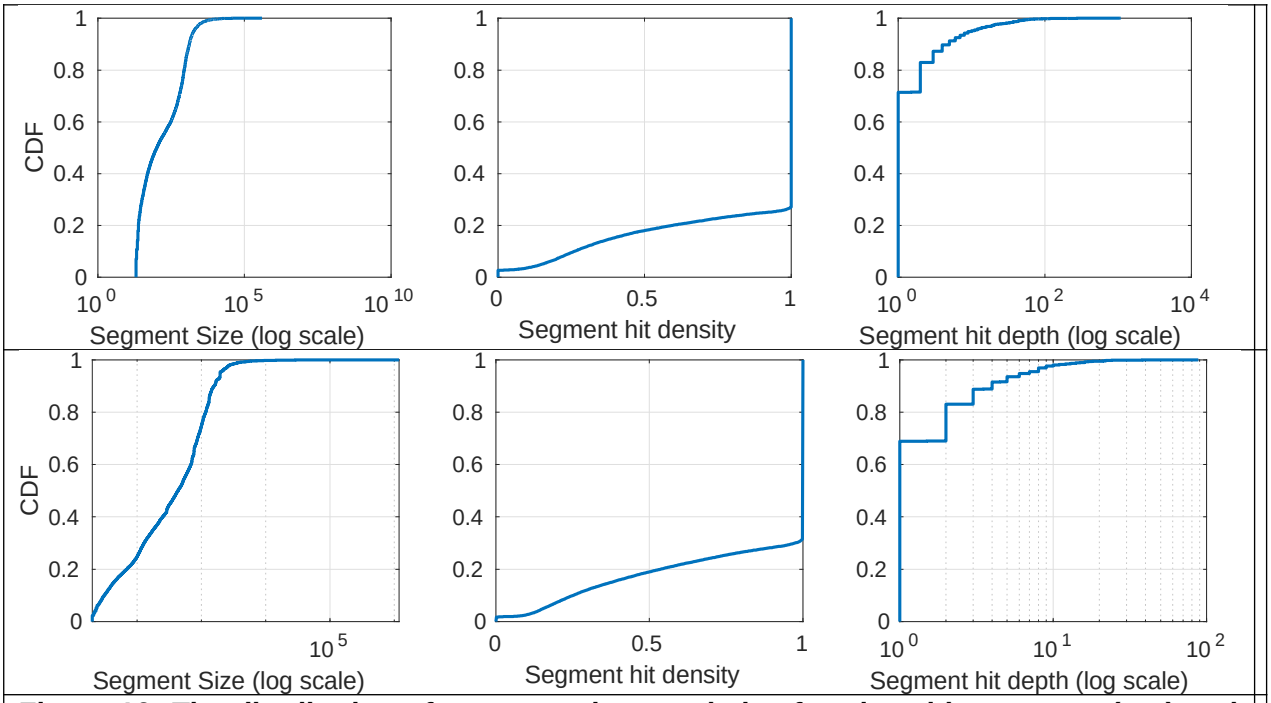


Figure 16: The distribution of segment characteristics for plasmid segments (top) and chromosome segments (bottom). Statistics are calculated using BLAST hits only; note a minority of segments with hit density of 0, those are segments that rely solely on MUMmer hits. Those properties are reported in Table 5.

The descriptive statistics of segments revealed 1,963 plasmid segments (2.7%) and 20,23 chromosome segments (1.7%) having a hit density close to 0, which are likely spurious segments created in our approach. This phenomenon is the result of low density regions; instead of optimizing our segmentation approach, we opted for the exclusion of such segments, as a part of the downstream analysis.

Table 5: Descriptive statistics of segments per pair (3,264 pairs)

		Min	Max	Mean	std	CV	Median
Plasmid Segments	Number of segments	1	689	22.21	48.66	2.19	6
	Density by segments	0	1	0.09	0.13	1.37	0.06
Chromosome segments	Number of segments	1	904	36.97	77.39	2.09	8
	Density by segments	0	0.29	0.01	0.02	2.25	0

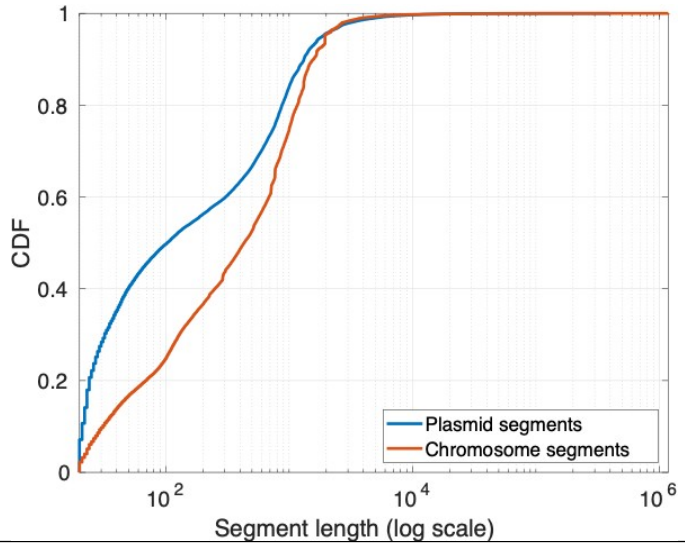


Figure 17: Cumulative distribution for the length of plasmid and chromosomal segment. A comparison of the two distributions shows that chromosomal segments are overall longer than plasmid segments.

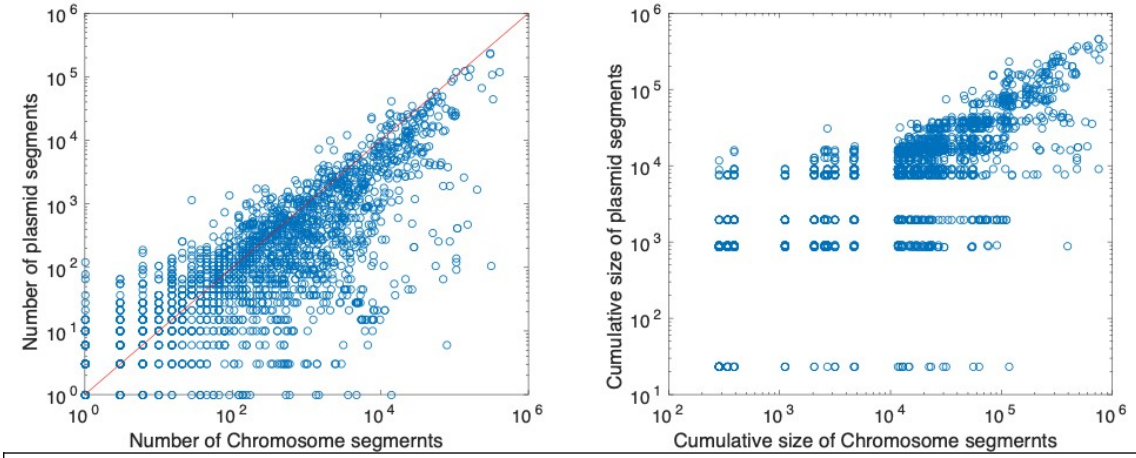


Figure 18: The frequency (left) and size (right) of segments per plasmid-chromosome. For the majority of pairs, the number of chromosomal segments is higher than the number of plasmid segments, this suggests that shared loci have multiple copies on the chromosome. Correspondingly, also the size of chromosomal segments is larger than plasmid segments.

The distribution of segments per replicon shows that in many cases, chromosomes have more segments than plasmids that cover a larger size of their DNA, this suggests that the plasmid DNA is present in multiple copies on the chromosome resulting from chromosomal DNA duplication of the transferred regions or multiple transfer events from the plasmid locus into the chromosome. Due to this

difference in the number and cumulative size of segments between plasmids and the chromosome of their hosts, the segmentation might not represent the actual extent of DNA transfer between the two replicons. Thus, we need a two-dimensional analysis of the shared DNA that pinpoints the homologous segments between plasmids and chromosomes and quantify the copy number of shared locus from one replicon on the other.

6.5 Intersections of plasmid and chromosome segments

The segmentation was applied to the binary sequence of hit positions for each replicon separately. This has resulted in consecutive alternating regions of high and low hit density on each replicon. The term segments refer to the replicon regions with high hit density. Hits that belong to the same segment on a certain replicon are expected to be clustered close to each other on that replicon, while those hits can either belong to the same segment on the other replicon or could be further away from each other and belonging to different segments (illustrated in Figure 19).

By gridding the 2d dot-plot of hits depending on the alternating regions of segments and non-segments for each replicon, where segments of the x-axis replicon determine columns and segments of the y-axis replicon determine rows, we acquire areas from the dot-plot that represent intersections for every pair of segments belonging to different replicons. This will give us *i* number of intersects between the two replicons, where theoretically:

$$i = Sc \times Sp$$

Sc is the number of chromosomal segments, and *Sp* is the number of plasmid segments.

A set of hits that belongs to the same segment on the chromosome and the same segment on the plasmid falls into the intersect of those two segments. Those hits potentially belong to one event of transfer, or alternatively, a duplication or a rearrangement of the transferred region. Each intersect is defined by four coordinates (a beginning and an end on each replicon), and has five characteristics; Two dimensions coming from the size of the chromosomal segment on the x-axes, and the plasmid segment on the y-axes, a hit density on the chromosome, a hit density on the plasmid, and the number of hits that it contains. The hit density is calculated by dividing the cumulative size of all hits' projections on one replicon by the size of

the segment. Theoretically, the number of all possible intersects between chromosomal and plasmid segments in our dataset is 12,636,076. However, when accounting for intersects, we eliminate spurious intersects by considering only intersects that contain at least one hit. Thus, the actual number of non-empty intersects is 332,944 in our dataset.

Segments on one replicon might not form any full intersect with any segment on the other replicon, this means that while hits within this segment are clustering together on its replicon, they are sparse on the other replicon and failed to create any segment on it. Those segments that have no intersects are considered spurious, as they do not represent an event of transfer or duplication. After eliminating spurious segments, we had 111,124 (92%) chromosomal segments and 51,866 (71%) plasmid segments that belong to 2,272 chromosome-plasmid pairs (98% of pairs that have segments).

Some intersects that contain hits might still be partially or mostly empty on either one or both of the replicons, this observation can be explained by genomic rearrangements on the replicon where the intersect is not full as the segments are formed by hits falling together on one of the replicons but could be anywhere far from each other on the other replicon (Figure 19).

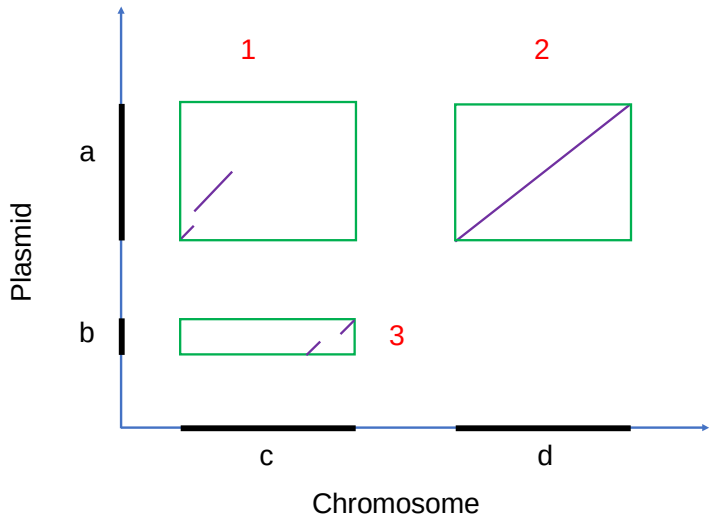


Figure 19: Intersect fullness. The chromosome and plasmid genomes are represented by the X and Y axis respectively, local similarity hits are illustrated in purple, segments on both the chromosome and the plasmid are represented as thick black lines, and intersects are illustrated as green rectangles. The intersect (1) is partially empty on segments (a) and (c) from both replicons, because the segmentation happened by joining hits on one replicon that are scattered on the other replicon. Similarly, intersect (3) is partially empty on segment (c) on the chromosome but full on segment (b) on the plasmid, which means that this might be a case of rearrangement on the plasmid after the transfer event (as we can see from intersects (1) and (3)). Finally, intersect (2) is completely full on segments (a) and (d) from both replicons and represents a transfer event that did not undergo changes, so it is probably a recent transfer event.

While the maximum chromosomal dimension of intersects is significantly larger than the maximum plasmid dimension, the median chromosomal and plasmid dimensions across all intersects are similar. Furthermore, the coefficient of variance (CV) of dimension size shows that, in the chromosomal dimension of intersects is higher in comparison to the plasmid dimension. Note that the median of BLAST hit similarity in the data that is also similar to the mean hit similarity, both are much above the minimal threshold of BLAST hit sequence similarity, this shows that with our threshold we capture shared regions that are much more conserved than 80%. Most of the data is in the range of 87% to 100% similarity in nucleotides. The clean set of intersects constitutes evidence for homologous loci between plasmids and chromosomes within the same host.

Table 6: Descriptive statistics of the intersects.

Each intersect has two dimensions, the chromosomal and plasmid, those correspond to segments on those replicons. The intersect fullness is calculated per replicon using only BLAST hits.						
	Min	Max	Mean	std	CV	Median
Chromosomal dimension of the intersect	20	1,180,829	1,300	11,063.69	8.51	677
Plasmid dimension of the intersect	20	393,620	1,564	5,881.09	3.76	779
The intersect fullness on the chromosome	0	1	0.54	0.40	0.75	0.48
The intersect fullness on the plasmid	0	1	0.49	0.40	0.81	0.38
Number of Blast hits	0	1,400	1.59	5.12	3.22	1
Number of MUMmer hits	0	2,997	2.68	8.38	3.12	1
Longest hit size	20	393,620	389	959	2.46	92
Longest hit similarity	80	100	93.12	5.99	0.06	93.75

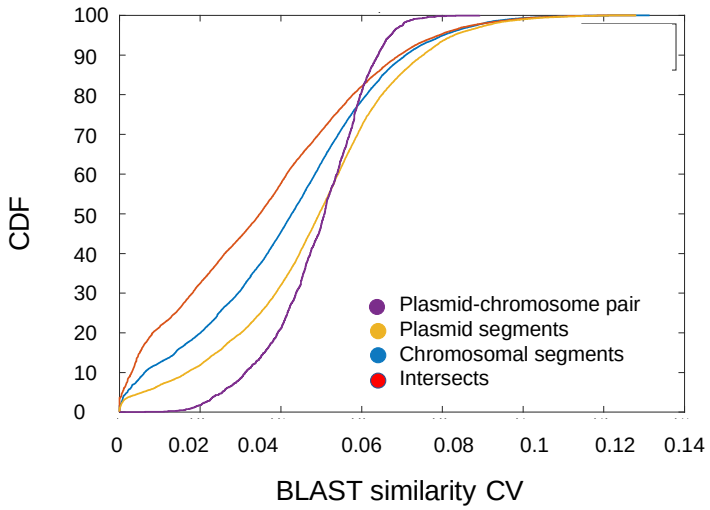


Figure 20: Cumulative distribution for the coefficient of variation (CV) for BLAST hit similarity. BLAST hits in the comparison belong to the same pair, chromosomal segment, plasmid segment or intersect.

6.5.1 General observation on the pipeline

After running the pipeline for 3,264 chromosome-plasmid pairs in our dataset, we found that only a minority of the pairs did not have sequence similarity detected by

local similarity hits. However, the number of pairs with no sequence similarity has increased by 16 folds after applying the segmentation approach. Nonetheless, the majority of pairs had shared sequence between the chromosome and plasmid detected by segments and then segment intersects (Table 7).

Notably, there is a clear correlation between the plasmid size and the level of sequence similarity it shares with the chromosome, as we found that the categories shown in Table 7 have an increasing plasmid size (Figure 21). This suggests for a positive association between the plasmid size and the frequency of shared regions between the plasmid and the chromosome.

Table 7: The frequency of plasmid-chromosome pairs after each step of the pipeline	
1. Pairs with no hits	54
2. Pairs with hits but no segments	891
3. Pairs with segments but not intersects	50
4. Pairs with intersects	2,269
Total	3,264

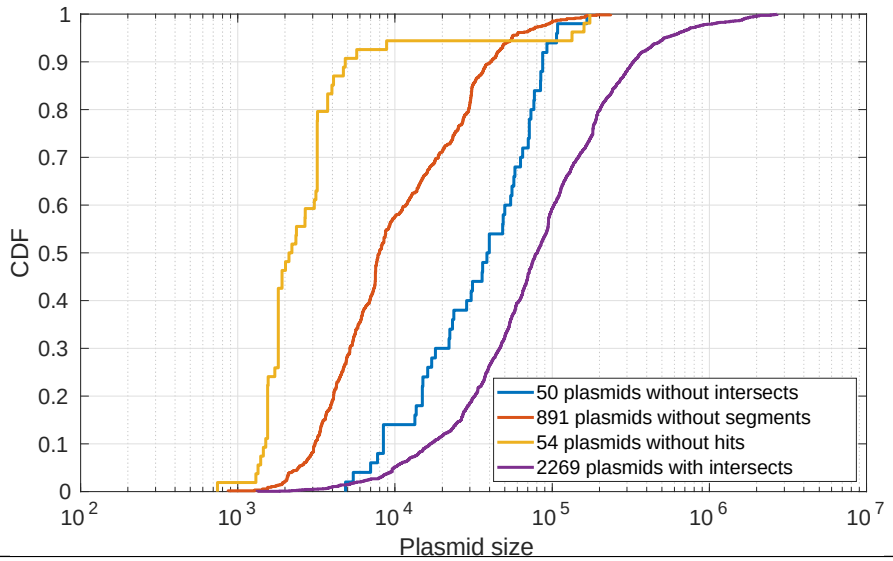


Figure 21: Cumulative distribution of plasmid size for the categories shown in Table 7. Plasmids that do not share any local similarity hits (BLAST or MUMmer) with the chromosome belong to the smallest size category, followed by plasmids that have only small and scattered hits that could not be joined by the segmentation. The next size category is plasmids that belong to pairs containing shared sequence detected by segments, but none of

the plasmid segments formed intersects with the chromosomal segments. The largest plasmid category is the one with shared sequence that belongs to the same neighbourhood on both replicons and fall into intersects that correspond to transfer events.

To further validate the segmentation results, we compared the properties of BLAST and MUMmer hits that were clustered within intersects with the properties of hits that did not make it into intersects. This shows that both the BLAST and MUMmer hit lengths in intersects are larger than hit lengths outside intersects (Table 8 and Figure 22). We also observed that BLAST hits outside intersects have much larger E-values than BLAST hits inside intersects (Figure 22). Those two observations of smaller size and larger E-value for hits outside intersects confirm that those hits represent insignificant random similarity patterns between the plasmid and chromosome (Figure 23). Using our segmentation approach, we managed to filter out this noise while keeping small hits that belong to a neighbourhood of local similarity hits and are potentially part of a DNA transfer event. This property grants the segmentation approach an advantage over hit filtration solely by size and E-value without a consideration of the distribution of those hits along the replicons.

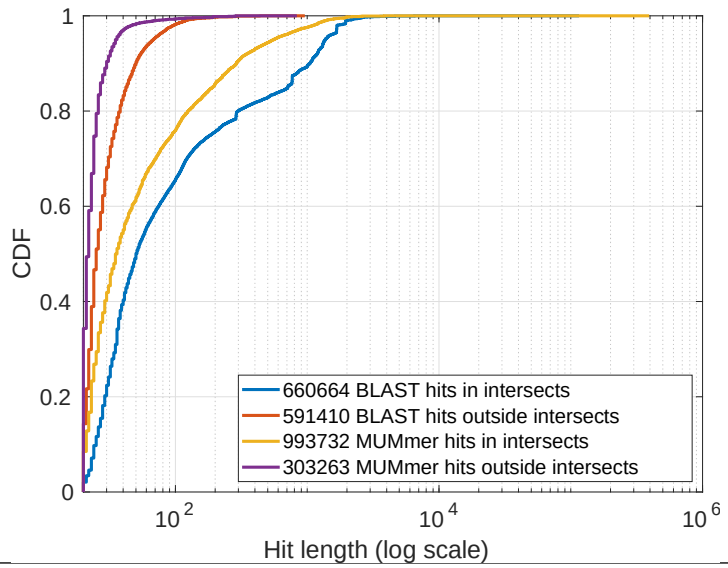


Figure 22: Cumulative distribution for BLAST and MUMmer size for hits falling inside or outside intersects. The sizes of MUMmer hits outside intersects are smaller than the sizes of MUMmer hits inside intersects. Similarly, BLAST hits outside intersects are not only smaller than BLAST hits inside intersects but also smaller than MUMmer hits inside intersects. This means that although MUMmer usually detects only unique exact matches

that are expected to be smaller than BLAST hits, a MUMmer hit that made it into an intersect has biological relevance and tends to be larger on average than random BLAST hits that were not joined within intersects.

Table 8: Statistics of BLAST and MUMmer hit size and BLAST E-value in and outside intersects. The characteristics of hits out of intersects suggest that those hits maybe considered as noise. Nonetheless, a minority of long hits (few hundred bp) ended outside intersects; this phenomenon is related to the window size in the segmentation approach and the adaptive threshold for splitting. The significant difference in the distribution of the E-value between in and out intersects indicates that the segmentation approach, overall, exclude mostly insignificant hits (Figure 23).

	In intersects	Outside intersects
Median BLAST hit length	50	25
Median MUMmer hit length	36	22
Max BLAST hit length	115,423	959
Max MUMmer hit length	393,620	834
Median BLAST hit E-value	10^{-08}	0.39

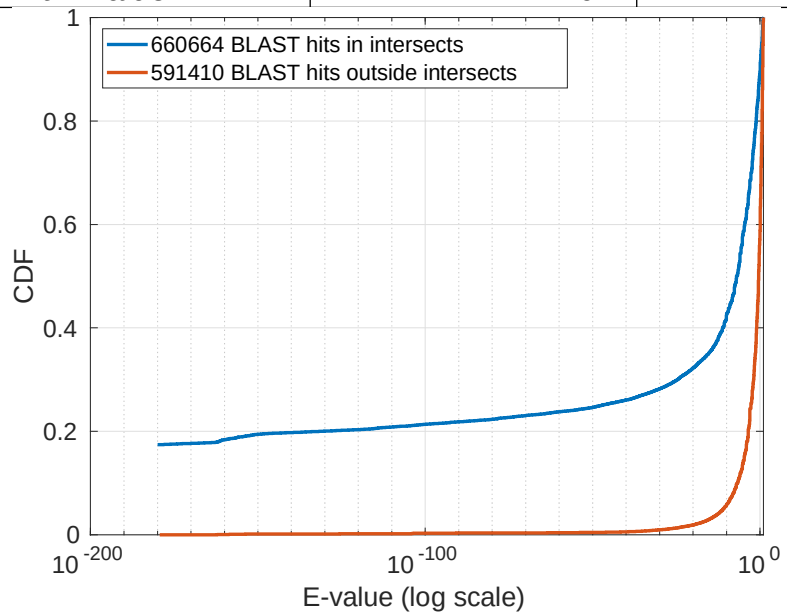


Figure 23: Cumulative distribution of BLAST hit E-value for hits falling inside or outside intersects. BLAST hits outside intersects have very high E-values which reflects a high probability that they are random, while BLAST hits inside intersects have low E-values that reflect their biological relevance as homologous sequence that were transferred from one replicon to the other.

6.5.2 Pivot Intersects

Half of the plasmid segments (50%) are shared only once with the chromosome, while the remaining segments form more than one intersect with chromosomal segments (data presented in Table 9 and Table 10). This reflects pairs where one locus on the plasmid has multiple homologous copies on the chromosome, which are either the result of duplications or rearrangements of the transferred DNA on the chromosome. Alternatively, they could reveal multiple transfers of the same locus from the plasmid to the chromosome. For now, we will discuss the first scenario and get back to the second scenario later on. If we assume one transfer event from either direction that has duplications on the chromosome, we have three possibilities;

a) Transfer from either direction followed by duplications on the chromosome
b) Duplications on the chromosome followed by transfer to the plasmid
c) Duplications on the chromosome followed by transfer to the plasmid followed by more duplications on the chromosome.

In scenario (b) we expect to observe one intersect that resembles the transfer event. We label this intersect as the pivot intersect and we expect it to have a higher similarity between its chromosomal and plasmid segments than within other intersects that we call “echoes”. In scenario (a) we expect to have a homogeneous similarity inside all intersects belonging to that certain plasmid segment. In scenario (c) we expect to find a group of intersects having higher similarities among their chromosomal and plasmid segments than the others. However, in all scenarios it is safe to assume that the intersect with the highest similarity between its chromosomal and plasmid segments resembles the actual transfer event under the assumption of a single transfer on that plasmid locus.

Table 9: The number of intersects formed by a plasmid segment with chromosomal segments. Half of the shared DNA on the plasmid have only one copy on the chromosome and the majority of the remaining plasmid segments have few copies on the chromosome. This means that most DNA transfer between plasmids and chromosomes took place as a single event.

Intersects per segment (echoes)	Number of plasmid segments	proportion
Only one intersect	26,051	0.50
Few intersects 2 .. 5	14,404	0.28
Multiple intersects 6 .. 20	7,654	0.15

Many intersects >20	3,757	0.07

Table 10: The number of intersects formed by a chromosomal segment with plasmid segments. Similarly to Table 9, about half of the shared DNA on the chromosome have only one copy on the plasmid and the majority of the remaining chromosomal segments have few copies on the plasmid.

Intersects per segment (echoes)	Number of Chromosome segments	proportion
Only one intersect	57,203	0.51
Few intersects 2 □ 5	38,792	0.35
Multiple intersects 6 □ 20	14,129	0.13
Many intersects >20	1,000	0.01

To detect the pivot intersects and discriminate them from echoes we devised two solutions. We could consider the intersect that has the highest BLAST similarity value of its largest BLAST hit, the one that has more base-pairs found by BLAST as shared between the plasmid and chromosome segments. However, our results showed that the two approaches had mismatching results regarding which intersect is the pivot. Consequently, we implemented a third way that allows combining both concepts; we produced pairwise alignments of the chromosomal and plasmid segments for each intersect and used the pairwise alignment similarity value as a measure to find the pivot intersect. In this way, we combine information regarding the sequence similarity as well as the length of the intersecting segments into a single measure.

7 RESULTS

7.1 The extent of transfer between plasmids and chromosomes

The intersects we calculated in the previous section constitute an inference of DNA transfer events between plasmids and chromosomes. In this chapter I use their frequency and properties to research the extent of LGT between plasmids and chromosomes. My results show that 2,269 (69.5%) of the plasmids in our dataset have evidence of DNA transfer with the chromosome shown by the segmentation and segment intersection results. Summarizing all intersects per plasmid shows that there is a large range for the proportion of each plasmid that is shared with the chromosome found by segments with intersects. Between 0.14% and 100% of the

plasmid sequence is shared with the chromosome with a median of 8.8% of the plasmid sequence. The copy number of plasmid segments on the chromosome – that is the echoes we detected for plasmid segments – is associated with the plasmid size, i.e., we observe less chromosomal copies of the DNA transfer in pairs including small plasmids, followed by medium size plasmids, and large plasmids have the highest copy number of their DNA on the chromosome (Figure 24).

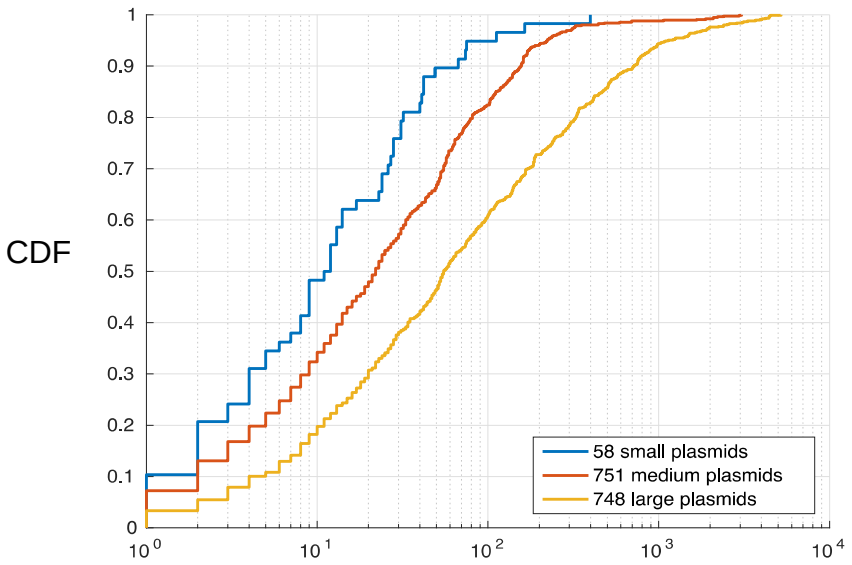


Figure 24: The cumulative distribution of the number of chromosomal copies of plasmid segments per chromosome-plasmid pair. The distribution of chromosomal copy number of plasmid loci was plotted separately for three plasmid size categories. Small plasmids have a genome size smaller than 10 kb, medium size plasmids range in size between 10 kb and 100 kb and large plasmids have a genome size larger than 100 kb. We observe a positive association between the plasmid size and the copy number of plasmid loci on the chromosome.

The finding that larger plasmids tend to have more copies of the same plasmid locus on the chromosome, may suggest that multiple transfer events end up in multiple chromosomal loci, because the alternative scenario of a single transfer followed by chromosomal duplications, should be independent of the plasmid size. We will get

back to this issue later on in the thesis while discussing the nature of genes transferred between plasmids and chromosomes.

7.2 Plasmids that are completely integrated in the chromosome

We identified 17 Plasmids as fully integrated in the chromosomal genome, those plasmids belong to a range of small and medium plasmids with a size ranging between 1,326 and 393,620 bp and a median of 16,509 bp which is lower than the median size of all plasmids (49,019 bp) (Figure 25). They mostly belong to medium and small plasmids and no clear correlation between the plasmid size and the proportion of its sequence shared with the chromosome (Figure 25). Fully integrated plasmids into the chromosome belong to different genera and some of them have multiple copies of all, or a part, of their sequence present on the chromosome. 13 of those plasmids have gene products associated with mobile genetic elements (9 plasmids have transposases, 8 have integrase genes and two plasmids have phage related genes, with no IS elements reported on any of those plasmids) as well as multiple other coding sequences (supplementary table 4).

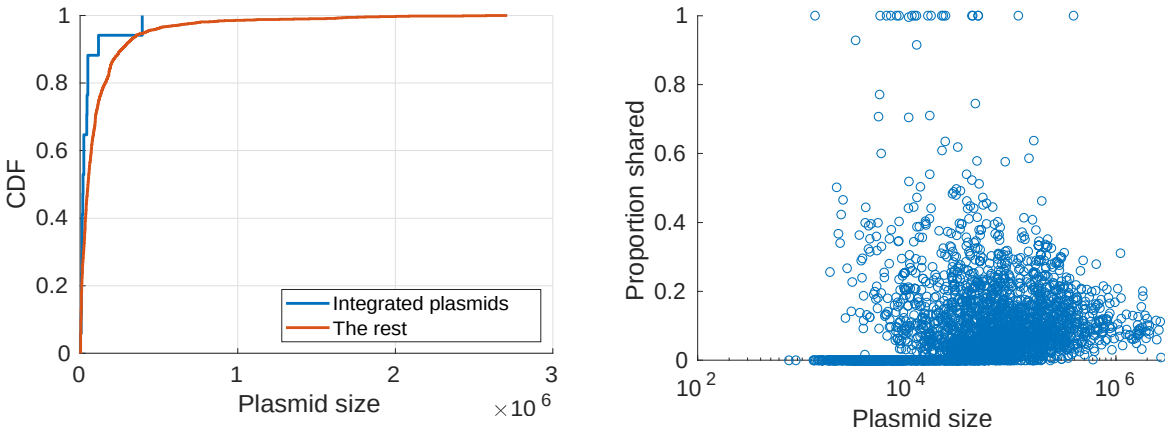
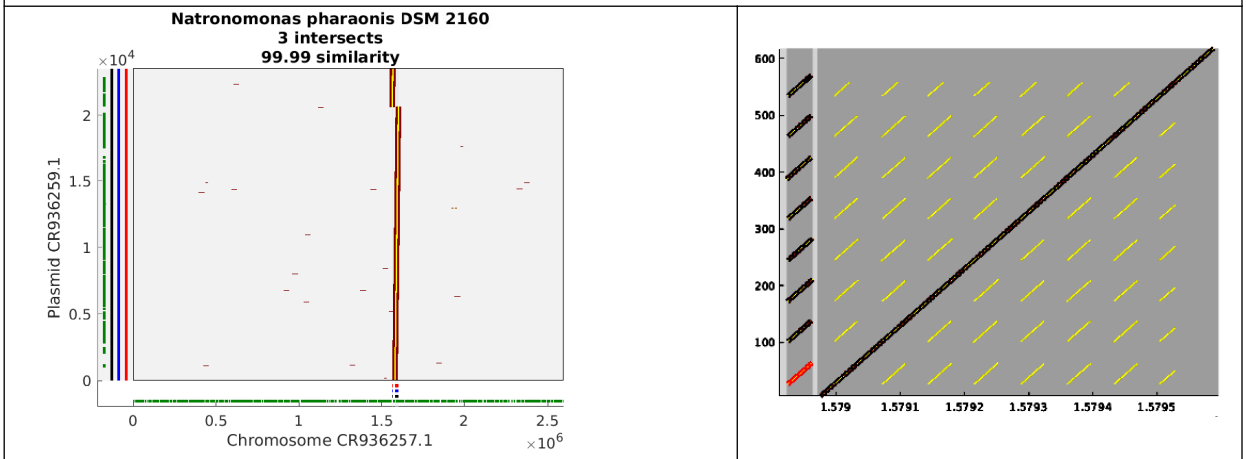


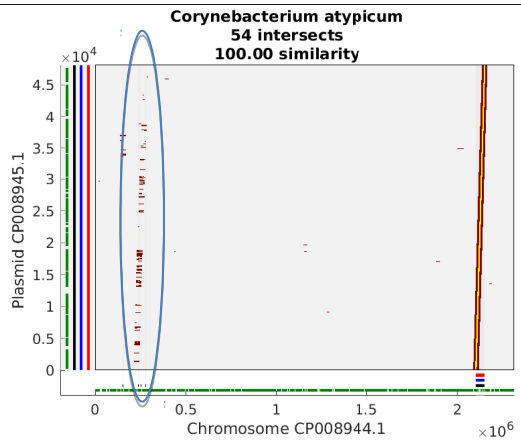
Figure 25: The size of fully integrated plasmids in the chromosome and proportion of plasmids shared with the chromosome. The cumulative distribution of plasmid size for plasmids that are fully integrated in the chromosomal sequence in comparison with other plasmids that share only a part of their sequence with the chromosome (left figure). The plasmid sizes vs. the proportion of each plasmid loci that found homologous to chromosomal sequences (right figure).

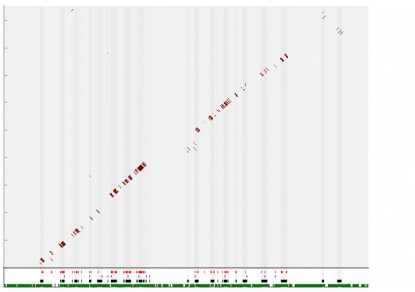
Table 11: Examples for fully integrated plasmids in the chromosomal genome of their host. We note that the plasmid integration into the chromosome might be suspected as artifacts of genome assembly. Many of the examples we report were sequenced by PacBio that produces long fragments and hence, reduces the probability for assembly artifacts. Furthermore, the sequence of integrated plasmids is never identical to the plasmid sequence itself, thus supporting the presence of a plasmid copy in the genome rather than assembly artifacts. The following plots were made as described in Figure 9.



1. A chromosome-plasmid pair belonging to *Natronomonas pharaonic* (left figure), a haloalkaliphilic archaeon Isolated from salty water with a PH of 11. The sequencing was done using a shotgun clone library. The entire plasmid is covered with one segment that corresponds to three segments on the chromosome. Two of the chromosomal segments represent one continuous sequence of DNA shared with the plasmid that got broken because of the plasmid linearization. The third segment covers an area of repeated spacer of a CRISPR array. By zooming in into this region of shared DNA in the dotplot (right figure), we can find that the entire region was found by both BLAST and MUMmer as a single conserved hit of a transferred CRISPR array. However, MUMmer is powerful in finding repeats that show in the yellow MUMmer plot lines and absent in the red lines that represent BLAST hits. In addition to the CRISPR array, the plasmid carries 36 protein coding genes and one pseudo gene, those genes include an integrase gene and a homologue to a phage protein as well as many uncharacterized products. The integrase gene suggests that the plasmid integration into the chromosome was mediated by an integron carried on

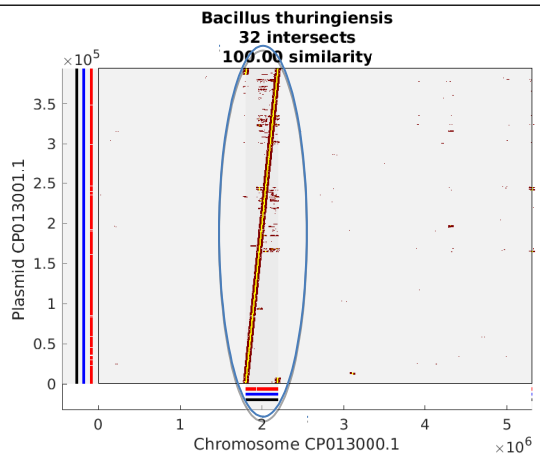
the plasmid. This event of whole plasmid integration was reported by the original publication that included the genome announcement (Falb *et al.*, 2005).

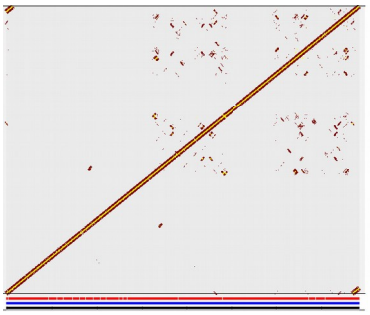




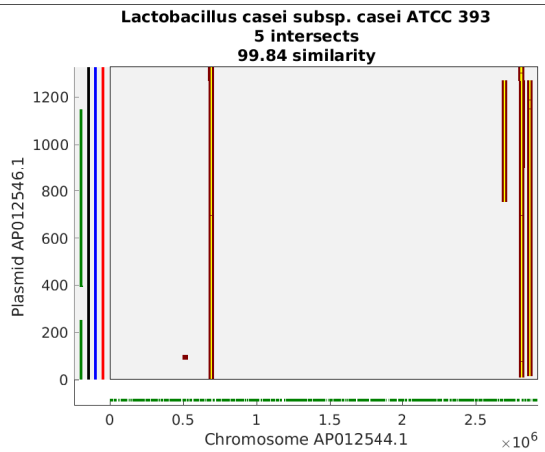
2. A chromosome plasmid pair belonging to *Corynebacterium atypicum*. The isolate comes from a clinical source from Germany and was sequenced using MiSeq desktop sequencer (Illumina). The right figure is a zoom in, focusing on the region within the blue oval. One intersect between one plasmid segment and one chromosomal segment contains the entire plasmid sequence, which means that the plasmid was integrated into the chromosome in one event of transfer. However, 53 other chromosomal segments contain syntenic BLAST hits spread along the plasmid, those hits are separated by gaps of similar sizes on both the plasmid and chromosome. They represent interrupted shared sequences that appear to the eye like a shadow or a ghost of the transfer event (the right figure is a zoom in, focusing on that region). They spread along the plasmid sequence and in the region between positions 227.4k and 280k on the chromosome. The longest BLAST hit in this region is only 68bp stretching between the positions 244,19 and 244,87 on the chromosome carrying a phage related portal protein. Another relatively large hit stretches between positions 231,21 and 231,61 carrying a coding sequence for a hydrolase enzyme (N-acetylmuramoyl-L-alanine amidase) that is important for peptidoglycan biosynthesis. Other phage related genes are also carried on the plasmid as well as a restriction modification system. A plausible explanation for the “ghost” pattern of shared sequence is that it represents an ancient event of transfer of a large part of the plasmid or the entire plasmid into the chromosome. The transferred sequence underwent genetic erosion so it appeared interrupted and with lower similarity score so it could not be detected by MUMmer for most of the hits. The original publication

of this isolate genome did not report the plasmid integration event (Tippelt *et al.*, 2014).



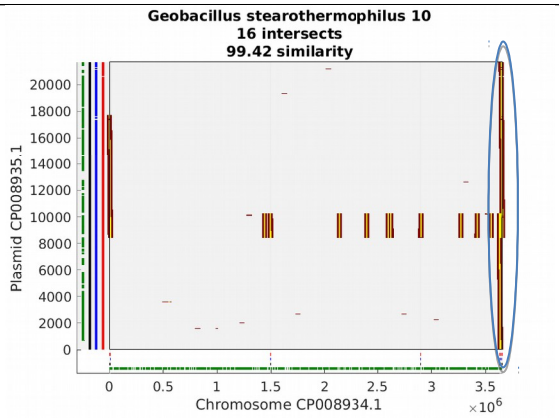


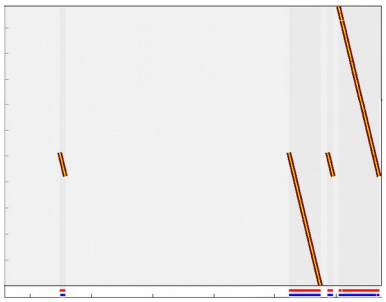
3. This pair belong to a *Bacillus thuringiensis* strain isolated from an agricultural source from China and was sequenced with Illumina PacBio. The right figure is a zoom in, focusing on the region within the blue oval. This pair illustrates another ancient event of transfer where the entire plasmid was integrated into the chromosome. 1,245 BLAST hits were joined into one plasmid segment that represent a single event of transfer that diverged with time, which caused it to appear as multiple hits. In this example, the segmentation procedure clearly succeeded in finding the more parsimonious scenario of a single transfer followed by genetic erosion. The plasmid carries 40 pseudogenes and 322 protein coding genes including multiple transposases, integrases, as well as a reverse transcriptase. Those mobile elements are probably what mediated the transfer event.
No publication is available for this isolate.



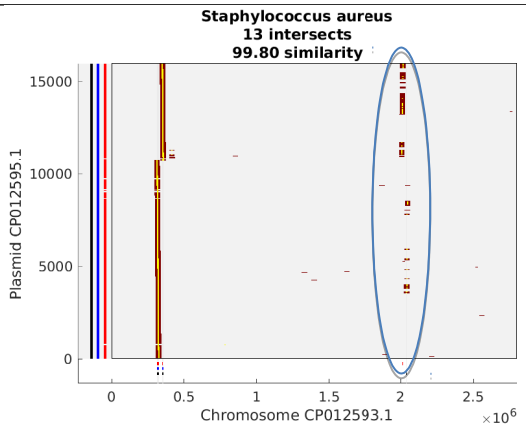
4. This pair belongs to *Lactobacillus Casei* that was isolated from a fermented milk

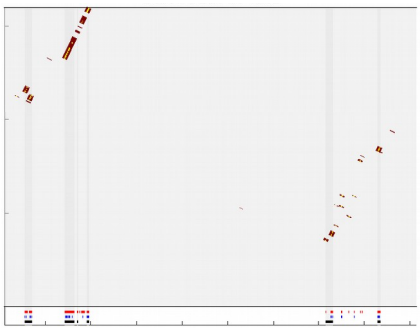
product from Japan and sequenced with Thermo Fisher 3730xl DNA Analyzer. The plasmid in this pair is small and the plasmid sequence is present on the chromosome three times and the only genes it carries are three transposases, while no replication initiation protein could be identified. This raises the suspicion that this is rather an IS element that was isolated in action. This isolate genome was announced in (Toh *et al.*, 2013).



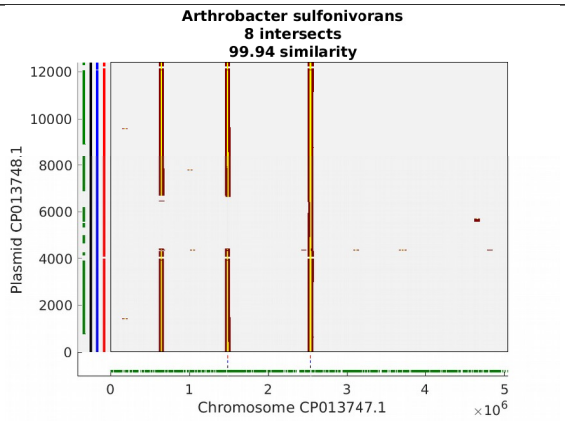


5. A pair from *Geobacillus stearothermophilus* that was isolated from hot springs in the USA and sequenced with PacBio. The right figure is a zoom in, focusing on the region within the blue oval. This pair has one plasmid segment representing the integration of the entire plasmid into the chromosome and 16 chromosomal segments that correspond to a pattern of DNA repeats or duplications on the chromosome (with very similar coordinates on the plasmid and different coordinates on the chromosome). Those repeats are indeed annotated as transposases. The plasmid carries some pseudogenes, RNase genes, a transposase and multiple metabolic pathways genes. The transfer and integration event was probably mediated by the transposon. No publication is available for this isolate.





6. A chromosome-plasmid pair belonging to a vancomycin-resistant bloodstream isolate of *Staphylococcus aureus* that was isolated from a patient in Brazil and sequenced using MiSeq PacBio (Ilumina). The right figure is a zoom in, focusing on the region within the blue oval. The plasmid is fully integrated into the chromosome possibly mediated by an inegron as it carries an integrase gene. The plasmid sequence is shared with the chromosome in two chromosomal segments with a large gap (256,698 bp), the gap carries multiple phage proteins, which means that there is a prophage that integrated into the chromosome after the plasmid integration or alternatively, the plasmid integration was mediated by a phage that was lost from the plasmid itself afterwards. In addition to the two large chromosomal segments, we can also observe the pattern of a “shadow” transfer that stretches along a large region of the plasmid as small interrupted BLAST hits interspersed with many gaps and inversions (the right figure focuses on this region). This represents another example of an ancient transfer followed by a recent one in the same plasmid-chromosome pair. The original publication of this isolate genome did not report the plasmid integration event (Panesso *et al.*, 2015).



7. A pair from *Pseudarthrobacter sulfonivorans* that is able to degrade petroleum, it was isolated in China and sequenced with Illumina Hiseq4000 and Pacbio RSII. The plasmid is fully integrated in the chromosome and carries 6 rRNA genes (s5, s16 and s23) as well as 3 other protein coding genes. The absence of mobile genetic elements suggests that the transfer was mediated by homologous recombination. No publication is available for this isolate.

7.3 Transfer of coding sequence

The shared coding sequence between plasmids and chromosomes through lateral gene transfer (LGT) can be detected and characterized using standard approaches like BLASTp and protein family clustering. In comparison, our segmentation approach provides a unique framework for the detection of co-transferred genes, including whole plasmids (as shown above), genes with no protein product, like RNA genes (rRNA and tRNA genes) (Figure 27), as well as noncoding sequence that potentially contain regulatory elements, and last but not least, genes that are partially shared between the two replicons. To that end, I compared the intersect loci with the genome annotation of plasmids and chromosomes. Among the 3,264 chromosome-plasmid pairs in our dataset, 2,989 plasmids and 1,298 chromosomes were annotated in NCBI, this yield 2,986 pairs that have annotations on both the plasmid and the chromosome, while 266 pairs had no annotation on one or both of the replicons, and are therefore excluded from the analysis.

The annotation of genes within homologous loci may vary between the plasmid and chromosome. In what follows I classify the homologous loci using the plasmid annotations only. We found 2,269 plasmids that have 41,332 (5%) protein-

coding sequences (CDSs) partially or fully included in homologous regions. A total of 22,139 plasmid CDSs are fully included within homologous loci (shared with the chromosome as complete genes) (Figure 26). The majority of homologous loci (80%) correspond to partial CDS or noncoding regions, 11.7% comprise one gene and 8.4% comprise multiple genes (two or more).

Notably, when I compare the frequency of single gene vs. multiple gene loci, I find multiple-gene-loci are highly frequent in the data in terms of number of loci as well as number of isolates where they are observed. Thus, I consider the analysis of multiple gene loci as informative for studying the frequency of transfer between plasmids and chromosomes (Table 12).

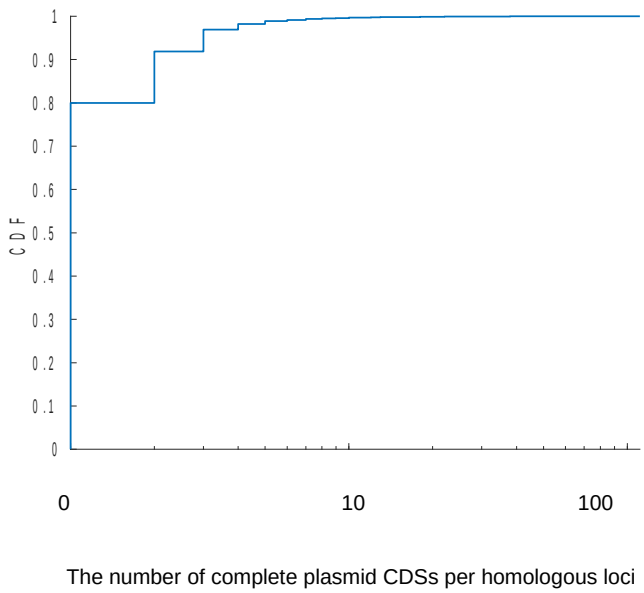


Figure 26: The cumulative distribution of number of complete plasmid CDS in homologous loci.

Table 12: Frequency of homologous loci with details on chromosomal copy number and intersect fullness, (full and partial intersects are described in Figure 19).				
		Number of segments	Number of pairs	Number of isolates
A single chromosomal copy	Full intersect	581	416	332
	Partial intersects	151	144	133
Multiple chromosomal copies	At least a full intersect	2,328	893	575
	All partial	1,243	620	452

The chromosomal copy number for plasmid loci containing genes is higher than that of loci corresponding to noncoding regions. Among plasmid homologous loci that include a coding sequence, the copy number of loci including multiple genes is higher than those that include a single gene (Figure 27).

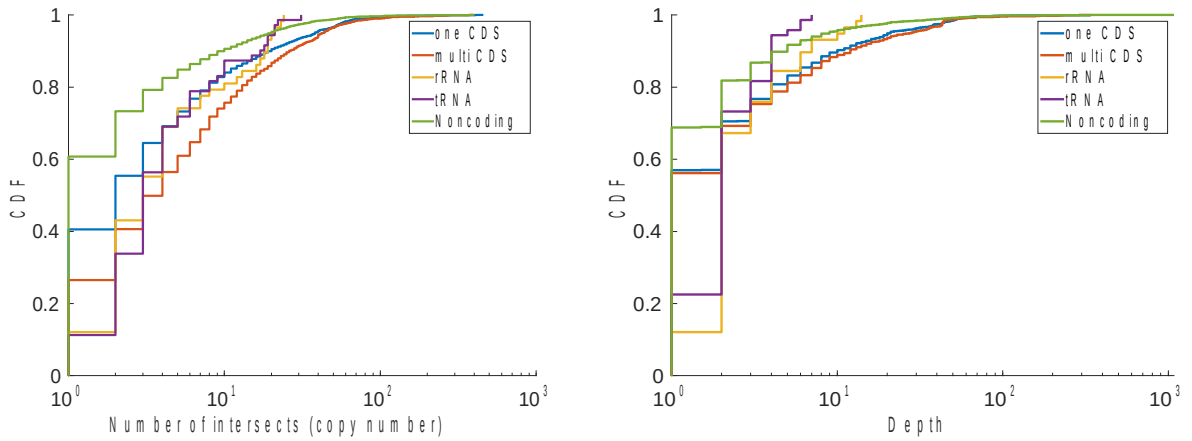


Figure 27: Cumulative distribution of chromosomal copy number of homologous loci. The copy number is calculated by two aproaches, the left figure showes the copy number estimated from echo intersects and the right figure shows an estimate of copy number using the median depth of BLAST hits that are the beasis for the homologous locus inference. These two measures yeilded simillar results.

Further comparison of the sequence similarity between plasmid and chromosome in homologous regions showed a higher sequence similarity in homologous loci that

correspond to one or more complete genes than homologous loci that include partial genes or noncoding loci. Since we have no reason to assume variable temporal dynamics of coding and noncoding DNA transfer (and duplication), this suggests that complete homologous genes between the plasmid and chromosome are conserved (i.e., likely evolving under purifying selection and may still be functional) (Figure 28). Furthermore, this can be expected and indicates rapid erosion of less essential DNA sequences to the host.

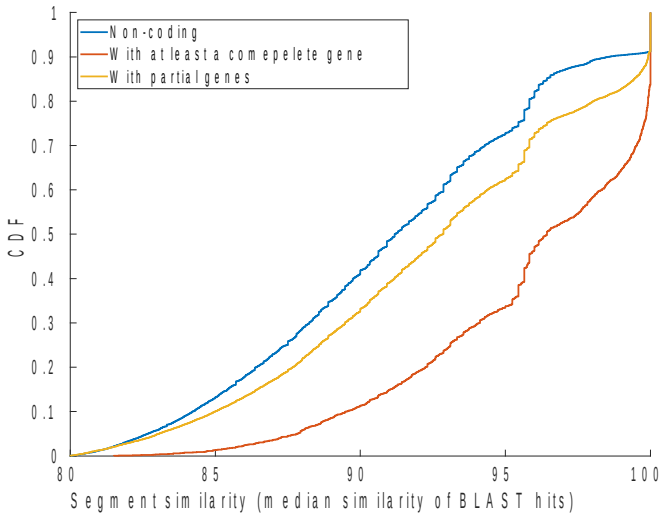


Figure 28: The cumulative distribution of homologous locus sequence similarity as calculated by the median similarity of BLAST hits

The median BLAST similarity for shared plasmid loci show a higher sequence conservation for loci carrying complete genes, followed by loci with only partial genes and then loci that are completely noncoding. The distribution of sequence similarity segments further supports our assertion that intersects containing partial genes may correspond to non-functional DNA.

7.3.1 Multigene homologous loci (MGL)

A major strength of our approach for the recovery of shared DNA between plasmids and chromosomes is the ability to recover homologous loci comprising multiple genes. Of the 10,660 plasmid homologous loci that overlap with complete genes, 4,303 loci include more than one complete gene. We found that 40% of all coding plasmid homologous loci are multigene loci (MGL), which accounts for 8.3% of all

plasmid homologous loci. In total, MGLs contain 15,782 genes, those genes were putatively co-transferred with one another in a single transfer event.

As a next step, we classified the MGLs into four classes, depending on the number of MGL copies in the chromosomes. This yielded 732 of MGLs that have a single copy on the chromosome (one intersect), termed single-copy-MGLs. The remaining 3,571 MGLs had multiple copies in the chromosome – termed here multiple-copy-MGLs.

Those loci, which we term single-copy-MGLs contain 1,626 genes with a range of 2 to 58 genes per locus and a median of 3 genes. A survey of the annotations shows that the majority of gene products in single-copy-MGLs are proteins related to mobile genetic elements (transposases, integrases, IS elements and some phage related proteins), followed by uncharacterized proteins (and conserved hypothetical proteins) in addition to various enzymes (resolvase, ATP-binding, ATPase, oxidoreductase, DNA-binding proteins, endonuclease, ribonuclease, adenylyl transferase, arsenate reductase, recombinase) and membrane proteins (Table 13).

Table 13: The most common gene products in single-copy-MGLs (top 20 products that correspond to 76.75% of the genes in single-copy-MGLs).

Gene product name	Number in single-copy-MGLs	Proportion in single-copy-MGLs	Number of segments	Number of pairs	Number of isolates
hypothetical protein	453	27.86%	250	201	165
transposase	369	22.69%	245	191	153
integrase	293	18.02%	188	162	146
IS	29	1.78%	20	16	15
DNA resolvase	14	0.86%	14	14	11
membrane protein	12	0.74%	9	9	9
ATP-binding protein	11	0.68%	11	10	10
Mobile element protein	9	0.55%	6	4	3
phage	8	0.49%	6	6	6
dihydropteroate synthase	6	0.37%	6	6	6
oxidoreductase	6	0.37%	6	6	5

Site-specific recombinase XerD	5	0.31%	3	3	3
resolvase	5	0.31%	5	5	3
ABC transporter permease	4	0.25%	3	2	2
Acyl carrier protein	4	0.25%	4	4	4
DNA-binding protein	4	0.25%	4	4	4
adenylyltransferase	4	0.25%	4	4	4
arsenate reductase	4	0.25%	4	4	4
arsenical pump membrane protein	4	0.25%	4	4	4
ethidium bromide resistance protein	4	0.25%	4	4	4

Many of the gene products found in single-copy-MGLs are hypothetical proteins, which are open reading frames in microbial genomes with an unknown function. The second most prevalent gene products are transposases that are one of the most common genes in nature (Hooper *et al.*, 2009). Other mobile elements such as IS elements and integrons made it to the top of the list as well. The remaining gene products likely constitute genes that are co-transferred with those gene transfer mechanisms.

Table 14: The most common gene products in multi-copy-MGLs (top 20 products that correspond to 86.30% of the genes in multi-copy-MGLs).					
transposase	2884	41.84%	1629	666	442
hypothetical protein	1713	24.85%	675	389	302
integrase	811	11.77%	552	342	261
IS	251	3.64%	149	56	51
helix-turn-helix family protein	56	0.81%	56	43	21
Mobile element protein	39	0.57%	22	16	10
phage	39	0.57%	15	13	11
ATPase AAA	23	0.33%	23	8	7
DNA replication protein	23	0.33%	23	16	8
isocitrate lyase	20	0.29%	20	12	11
ATP-binding protein IstB	13	0.19%	13	10	4
ATP-binding protein	11	0.16%	11	10	9
DNA-binding protein	10	0.15%	10	9	9
ABC transporter related	8	0.12%	6	5	4
cobalt/zinc/ cadmium resistance heavy metal efflux pump protein CzcA	8	0.12%	8	8	8
cobalt/zinc/ cadmium resistance heavy metal efflux pump protein CzcC	8	0.12%	8	8	8
heavy metal/cation efflux pump CzcB/HlyD	8	0.12%	8	8	8
helix-turn-helix domain protein	8	0.12%	8	7	7
putative membrane protein	8	0.12%	3	3	3
two component sensor histidine	8	0.12%	8	8	8

kinase					

Multi-copy-MGLs contain 13,517 genes with a range of 2 to 322 genes per segment and a median of 3 genes. More gene products have multiple copies on the chromosome (belonging to multiple-copy-MGLs), especially, mobile genetic elements. Multiple-copy-MGLs has several additional common enzymes; isocitrate lyase, DNA replication protein, permease, oxidoreductase, acyl-CoA dehydrogenase (Table 14).

Transposases constitute the majority of gene products in multi-copy-MGLs followed by hypothetical proteins. Also here, LGT mediating genes are on the top of the list with other gene products likely corresponding to co-transferred genes. The frequency of isolates where MGLs are observed suggests that transfer between the plasmid and chromosome as mediated by transposons (and other mediators) is not restricted to specific isolates. Hence, the transposon-mediated transfer between plasmids and chromosomes can be seen as a general phenomenon.

Table 15: The most common gene products observed in MGLs.

Gene product name	Number in single-copy-MGLs	Proportion in single-copy-MGLs	Number in multi-copy-MGLs	Proportion in multi-copy-MGLs
Transposase	369	22,69%	2884	41,84%
Hypothetical protein	453	27,86%	1713	24,85%
Integrase	293	18,02%	811	11,77%
IS	29	1,78%	251	3,64%
Mobile element	9	0,55%	39	0,57%
Phage	8	0,49%	39	0,57%
ATP-binding protein	11	0,68%	11	0,16%
DNA-binding protein	4	0,25%	10	0,15%

A comparison of the distribution of gene products between single-copy and multi-copy MGLs reveals overall similar picture with the main difference in the proportion of transposases that is higher for multi-copy-MGLs. One possible explanation for this

difference maybe the mechanism of transposition that includes a copy-paste mechanism.

Our results so far imply that non-transposase genes are being transferred between plasmids and chromosomes by transposition. To test this suggestion, we calculated the frequencies of gene product combinations that are observed in the same MGL, i.e. they are co-transferred.

Table 16: The most commonly transferred genes in multi-copy MGLs

Gene product	Transposase	Hypothetical protein	Integrase	IS	Helix-turn-helix family protein	Mobile element protein	Phage	Atpase AAA	DNA replication protein	Isocitrate lyase	ATP-binding protein istb	ATP-binding protein	DNA-binding protein	ABC transporter related	Cobalt/zinc/cadmium resistance heavy metal efflux pump czca	Cobalt/zinc/cadmium resistance heavy metal efflux pump protein czcc	Heavy metal/cation efflux pump czcb/hlyd	Helix-turn-helix domain protein	Putative membrane protein	Two component sensor histidine kinase
Transposase	1,061																			
Hypothetical protein	289	296																		
Integrase	252	175	145																	
IS	73	24	5	77																

Helix-turn-helix family protein	8	4	56	0	0															
Mobile element	3	3	1	0	0	17														
Phage	1	13	10	1	0	1	7													
Atpase AAA	5	6	13	0	0	0	0	0												
DNA replication	23	0	0	0	0	0	0	0	0											
Isocitrate lyase	20	2	1	0	0	0	0	0	0	0										
ATP-binding protein istb	13	1	0	1	0	0	0	0	0	0	0									
ATP-binding	9	0	3	0	0	0	0	0	0	0	0	0								
DNA-binding	6	5	7	0	0	0	0	0	0	0	0	0	0							
ABC transporter related	2	4	4	0	0	0	0	0	0	0	0	0	0	2						
Cobalt/zinc/ cadmium resistance heavy metal efflux pump protein czca	0	8	8	0	0	0	0	0	0	0	0	0	0	0	0					
Cobalt/zinc/ cadmium resistance heavy metal efflux pump protein czcc	0	8	8	0	0	0	0	0	0	0	0	0	0	0	8	0				
Heavy metal/cation efflux	0	8	8	0	0	0	0	0	0	0	0	0	0	0	8	8	0			

pump czcb/hlyd																				
Helix-turn-helix domain protein	2	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Putative membrane	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
histidine kinase	0	8	8	0	0	0	0	0	0	0	0	0	0	0	8	8	8	0	0	0

Table 17: The most commonly transferred genes in single-copy MGLs

Gene product	Hypothetical protein	Transposase	Integrase	IS	Dna resolvase	Membrane protein	Atp-binding	Mobile element	Phage	Dihydropteroate synthase	Oxidoreductase	Site-specific recombinase xerd	Resolvase	Abc transporter permease	Acyl carrier protein	DNA-binding	Adenylyltransferase	Arsenate reductase	Arsenical pump membrane	Ethidium bromide resistance
Hypothetical protein	105																			
Transposase	66	106																		
Integrase	62	57	78																	
IS	7	9	1	4																
DNA resolvase	0	14	0	0	0															
Membrane protein	1	0	4	0	0	3														
ATP-binding	1	8	2	0	8	0	0													
Mobile element	3	1	0	0	0	0	0	2												
Phage	3	0	1	0	0	0	0	0	2											

Dihydropteroate synthase	1	0	1	0	0	0	0	0	0	0										
Oxidoreductase	1	0	6	0	0	1	0	0	0	0	0									
Site-specific recombinase xerd	2	0	1	0	0	0	0	0	0	0	0	1								
Resolvase	2	4	1	0	0	0	0	0	0	0	0	0	0							
ABC transporter permease	0	0	1	0	0	0	0	0	0	0	0	0	0	1						
Acyl carrier	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0					
DNA-binding	1	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0				
Adenylyltransferase	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0			
Arsenate reductase	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Arsenical pump membrane	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	
Ethidium bromide resistance	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	4	0	0	0

The frequency of co-transferred gene products show that indeed most of the gene products are co-transferred with transposase. Additionally, transposases themselves are typically found in multiple copies within an MGL. This result is in agreement with the structure of many known transposons that contain two flanking transposases sandwiching other genes (Graur, 2016).

Hypothetical proteins are co-transferred with all types of gene products; this again, reflects the unsatisfying quality of bacterial genome annotations. In the single-copy-MGLs, integrases mediate the transfer of DNA and ATP binding proteins, as well as oxidoreductases and recombinases. The data of co-transferred genes in multi-copy-MGLs further reveals a high frequency of genes transferred with either IS elements or integrases. Several classes, e.g., phage proteins tend to be transferred with genes of the same gene product.

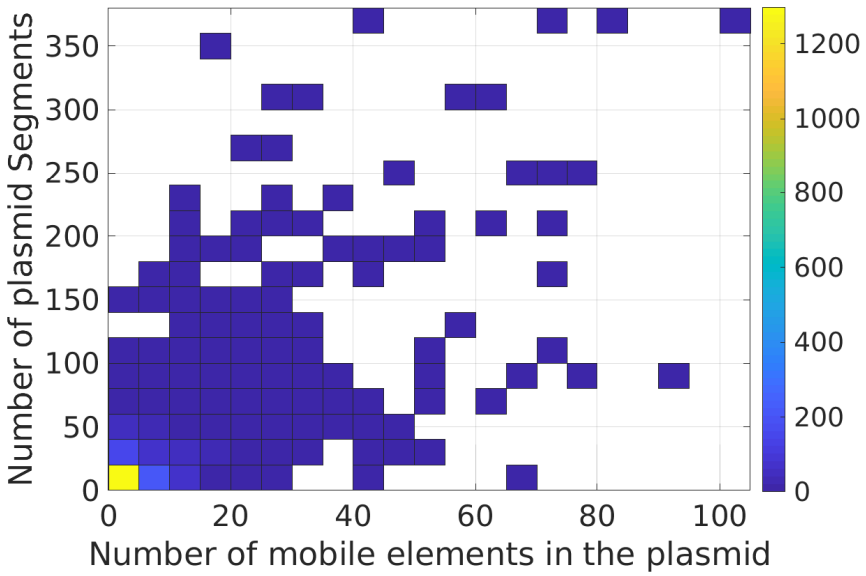


Figure 29: A heatmap depicting the correlation between the number of mobile elements in a plasmid and the number of shared loci with the chromosome. The frequency of chromome-plasmid pairs having a certain value of each axis is illustrated by the colormap.

To further confirm the significant contribution of all types of mobile elements that we observed previously in the MGLs and co-transfer patterns, the mobile element gene products were grouped together into one category (transposases, integrases, IS elements, phages and gene products annotated as mobile element). Then we tested for the correlation between the number of gene products that are related to mobile

elements and the number of plasmid loci that are shared with the chromosome (Figure 29), and found a significant positive correlation (r = 0.63, P-value < $2.2x10^{-16}$ using Spearman rank correlation test).

7.3.2 Transfer of antimicrobial resistance genes

One aspect of plasmid research is their contribution to the dissemination of antibiotics resistance genes. In what follows I examined the presence of AMR genes in the plasmid loci identified as DNA transfer (i.e., homologous loci). Using the comprehensive antibiotic resistance database (CARD) and resistance gene identifier (RGI) tool to search for AMR genes encoded in the plasmids included in our dataset (Alcock *et al.*, 2020), we found 1,655 AMR genes belonging to 86 AMR families and 54 drug classes. A total of 862 plasmids (26,4%) in the dataset encode $\geq$1 AMR gene hence they are considered as AMR plasmids. The AMR plasmids were reported in 305 (21,7%) isolates from our dataset, which are likely antibiotics resistant strains. Comparing the AMR gene loci with our inference of DNA transfer shows that 90 (5.4%) of those genes are transferred between the plasmid and the chromosome of the microbial host, they belong to 21 AMR gene families (supplementary table 1) and 14 drug classes (supplementary table 2). The AMR genes correspond to 74 homologous loci in 58 plasmid-chromosome pairs (19% of AMR carrying plasmids) in 48 isolates. According to the plasmid paradox, it has been suggested that plasmids carrying AMRs are at risk of extinction following the transfer of the AMR gene into the chromosome (Harrison and Brockhurst, 2012). Nonetheless, the results here suggest that if AMR genes are transferred from the plasmid to the chromosome, this is a relatively rare phenomenon. Furthermore, many AMR plasmids carry multiple AMR genes (Wein *et al.*, 2020), hence, it might be that a single AMR gene transfer from the plasmid to the chromosome to render the plasmid non-essential under specific conditions. The most common co-transferred gene with AMR genes is transposase (supplementary table 3). Notably, 32 AMR genes were found as co-transferred within an MGL including more than one AMR gene (Table 18). OR transferred in multiple copies to the chromosome i.e. sulfonamide resistance gene in *citrobacter freundii*, such pattern of transfer is indeed expected if the mechanism mediation AMR gene transfer is transposition.

Table 18: Loci shared between plasmids and chromosomes that correspond to AMR

genes.

Thick boxes designate single isolates. Background colours correspond to different plasmids within each isolate, starting with green for the first plasmid in the isolate, then blue and red. Text colour corresponds to different loci of shared AMR within each plasmid, with the order; green, blue, red and purple. Isolates that have only one transferred AMR gene are shown with a white background and black text.

	AMR Gene Family	Isolate	Plasmid Accession
1	RND antibiotic efflux pump	Ralstonia solanacearum, GMI1000	AL646053.1
2	TEM beta-lactamase	Escherichia fergusonii, ATCC 35469	CU928144.1
3	AAC(3)	Streptomyces hygroscopicus jinggangensis 5008	CP003276.1
4	ANT(2")	Citrobacter freundii, CFNIH1	CP007558.1
5	sulfonamide resistant sul	Citrobacter freundii, CFNIH1	CP007558.1
6	sulfonamide resistant sul	Citrobacter freundii, CFNIH1	CP007558.1
7	sulfonamide resistant sul	Citrobacter freundii, CFNIH1	CP007558.1
8	sulfonamide resistant sul	Citrobacter freundii, CFNIH1	CP007558.1
9	ATP-binding cassette (ABC) antibiotic efflux pump	Azospirillum brasilense	CP007797.1
10	tetracycline inactivation enzyme	Myroides odoratimimus	CP013691.1
11	ANT(6)	Myroides odoratimimus	CP013691.1
12	ANT(6)	Myroides odoratimimus	CP013691.1
13	AAC(3)	Streptomyces hygroscopicus, jinggangensis TL01	CP003721.1
14	sulfonamide resistant sul	Klebsiella oxytoca E718	CP003684.1
15	sulfonamide resistant sul	Klebsiella oxytoca E718	CP003684.1
16	APH(6)	Klebsiella oxytoca E718	CP003684.1
17	ANT(3")	Klebsiella oxytoca E718	CP003684.1
18	sulfonamide resistant sul	Klebsiella oxytoca E718	CP003684.1
19	AAC(3)	Klebsiella oxytoca E718	CP003684.1
20	blaZ beta-lactamase	Staphylococcus epidermidis, RP62A	CP000028.1
21	RND antibiotic efflux pump	Ralstonia solanacearum, PSI07	FP885891.2
22	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae MGH 78578	CP000649.1
23	quinolone resistance protein (qnr)	Enterobacter asburiae	CP012163.1
24	APH(3')	Klebsiella oxytoca	CP011617.1
25	ANT(3")	Klebsiella pneumoniae, pneumoniae KPNIH10	CP007729.1
26	sulfonamide resistant sul	Klebsiella pneumoniae, pneumoniae KPNIH10	CP007729.1
27	ANT(3")	Klebsiella pneumoniae, pneumoniae KPNIH1	CP008829.1
28	sulfonamide resistant sul	Klebsiella pneumoniae, pneumoniae KPNIH1	CP008829.1
29	APH(3')	Klebsiella oxytoca	CP011596.1

30	sulfonamide resistant sul	Acinetobacter baumannii, MDR-ZJ06	CP001938.1
31	ANT(3'')	Acinetobacter baumannii, MDR-ZJ06	CP001938.1
32	SHV beta-lactamase	Klebsiella pneumoniae, ATCC BAA-2146	CP006662.1
33	CTX-M beta-lactamase	Klebsiella pneumoniae, ATCC BAA-2146	CP006662.1
34	sulfonamide resistant sul	Klebsiella pneumoniae, ATCC BAA-2146	CP006661.1
35	SHV beta-lactamase	Klebsiella pneumoniae, 500 1420	CP011983.1
36	SHV beta-lactamase	Klebsiella pneumoniae, UHKPC33	CP011992.1
37	SHV beta-lactamase	Klebsiella pneumoniae, DMC1097	CP011977.1
38	SHV beta-lactamase	Klebsiella pneumoniae, UHKPC07	CP011987.1
39	APH(3')	Enterobacter cloacae	CP012170.1
40	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae Kp13	CP004000.1
41	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae KPNIH27	CP007732.1
42	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae KPNIH27	CP007733.1
43	ANT(3'')	Klebsiella pneumoniae, pneumoniae KPNIH24	CP008798.1
44	sulfonamide resistant sul	Klebsiella pneumoniae, pneumoniae KPNIH24	CP008798.1
45	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae KPNIH24	CP008799.1
46	ANT(3'')	Klebsiella pneumoniae, pneumoniae KPNIH24	CP008800.1
47	ANT(3'')	Klebsiella pneumoniae, pneumoniae KPNIH24	CP008800.1
48	AAC(6')	Klebsiella pneumoniae, pneumoniae PittNDM01	CP006799.1
49	OXA beta-lactamase	Klebsiella pneumoniae, pneumoniae PittNDM01	CP006799.1
50	AAC(3); AAC(6')	Klebsiella pneumoniae, pneumoniae PittNDM01	CP006799.1
51	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae PittNDM01	CP006801.1
52	APH(6)	Acinetobacter baumannii	CP012007.1
53	APH(3'')	Acinetobacter baumannii	CP012007.1
54	sulfonamide resistant sul	Acinetobacter baumannii	CP012007.1
55	sulfonamide resistant sul	Acinetobacter baumannii	CP012007.1
56	OXA beta-lactamase	Acinetobacter baumannii	CP012008.1
57	CTX-M beta-lactamase	Klebsiella pneumoniae	CP008933.1
58	KPC beta-lactamase	Klebsiella pneumoniae, pneumoniae	CP009773.1

59	TEM beta-lactamase	Klebsiella pneumoniae, pneumoniae	CP009776.1
60	SHV beta-lactamase	Klebsiella pneumoniae, pneumoniae	CP009776.1
61	AAC(6')	Klebsiella pneumoniae, pneumoniae	CP009776.1
62	AAC(6')	Klebsiella pneumoniae, pneumoniae	CP009778.1
63	ANT(3")	Klebsiella pneumoniae, pneumoniae	CP009778.1
64	blaZ beta-lactamase	Staphylococcus aureus, aureus 11819-97	CP003193.1
65	AAC(6')	Klebsiella pneumoniae, pneumoniae	CP009877.1
66	ANT(3")	Klebsiella pneumoniae, pneumoniae	CP009877.1
67	ANT(3")	Klebsiella pneumoniae	CP010393.1
68	sulfonamide resistant sul	Klebsiella pneumoniae	CP010393.1
69	SHV beta-lactamase	Klebsiella pneumoniae	CP010395.1
70	ANT(3")	Klebsiella pneumoniae	CP010396.1
71	ANT(3")	Klebsiella pneumoniae	CP010396.1
72	sulfonamide resistant sul	Salmonella enterica, serovar Typhimurium T000240	AP011958.1
73	ANT(3")	Salmonella enterica, serovar Typhimurium T000240	AP011958.1
74	KPC beta-lactamase	Klebsiella pneumoniae	CP011575.1
75	RND antibiotic efflux pump	Ralstonia solanacearum	CP011998.1
76	SHV beta-lactamase	Klebsiella pneumoniae	CP011622.1
77	SHV beta-lactamase	Klebsiella pneumoniae	CP011646.1
78	AAC(3); AAC(6')	Klebsiella pneumoniae	CP012754.1
79	AAC(6')	Klebsiella pneumoniae	LN824138.1
80	16S rRNA methyltransferase (G1405)	Klebsiella pneumoniae	LN824138.1
81	TEM beta-lactamase	Escherichia coli, PCN033	CP006635.1
82	SHV beta-lactamase	Klebsiella pneumoniae	CP013324.1
83	CTX-M beta-lactamase	Escherichia coli	CP009860.1
84	TEM beta-lactamase	Escherichia coli	HE610900.2
85	sulfonamide resistant sul	Salmonella enterica, serovar Typhimurium L-3553	AP014566.1
86	ANT(3")	Salmonella enterica, serovar Typhimurium L-3553	AP014566.1
87	TEM beta-lactamase	Escherichia coli	CP013024.1
88	APH(3')	Escherichia coli	CP013027.1
89	TEM beta-lactamase	Escherichia coli	CP013027.1
90	RND antibiotic efflux pump	Salmonella enterica, serovar Senftenberg	LN868945.1

Many of the organisms where we observed AMR transfer between plasmids and chromosomes are commonly reported as players in the dissemination of AMR within the hospital environment. Examples are; *Klebsiella* species as well as *Acinetobacter* and *Salmonella*. Nonetheless, one organism stands out as an exception, *Ralstonia solanacearum,* which is a potato pathogen. Indeed, previous studies reported the presence on AMR plasmids in agricultural habitats, likely as a result of antibiotics usage in animal husbandry and the application of manure for fertilization (Jechalke *et al.*, 2014).

7.3.3 Transfer of RNA genes

Among the plasmid loci that are shared with the chromosome, we observed many loci documenting genuine gene transfer events, including 58 complete rRNA genes in 36 chromosome-plasmid pairs, belonging to 27 isolates. We observed a shared region containing the full ribosomal RNA operon (Table 11 & Table 7). The 16S rRNA gene is traditionally considered as a trustworthy marker of bacterial species phylogenies; our results here show that also ribosomal RNA genes can be transferred, hence, rRNA phylogenies are not immune to reticulated evolutionary events. Additionally, we observed 71 complete tRNA genes in 40 pairs belonging to 28 isolates. The chromosomal copy number of RNA genes shared with the plasmid is demonstrated in Figure 27. The distribution of rRNA and tRNA copy number suggests frequent transfers or duplications of the plasmid-chromosome shared regions (a full list of all homologous loci between plasmids and chromosomes can be found in supplementary table 4).

7.3.4 Partial gene transfer

We define partial genes as those that have at least 40% and less than 100% of their sequence included in the homologous locus (shared with the chromosome); Methods for LGT inference that search for complete protein sequences thus miss a substantial amount of transferred DNA that comprise partial genes and non-coding sequence. Plasmid genes that are partially shared with the chromosome account for either degraded gene transfers, or shared domains between different genes that can also contribute to gene evolution. The number of coding homologous loci on the plasmid increases by 3-folds if we consider partial genes, nonetheless, there might be overlaps between the two sets as homologous loci might include both complete and partial genes. And the number of partial genes that are shared between the plasmid and chromosome is 1.5 folds larger than the number of complete genes (Table 19). The observation of high frequency of partial gene transfer reveals the two sides of evolution by gene transfer: It may lead to evolution of novelty, but at the same time may also create junk DNA (Graur *et al.*, 2015).

7.4 Transfer of noncoding sequence

Another strength of our approach is its ability to uncover DNA transfer events of non-coding loci. In 2,971 chromosome-segment pairs that have genome annotations, there are 331,649 pairs of homologous loci (intersects). Among 99,839 homologous loci on chromosomes, we found 47,010 loci (58%) that do not overlap with any chromosomal coding sequence, so those homologous loci are completely non-coding. Similarly, among 51,521 plasmid homologous loci, we found 27,255 non-coding loci (62%). Due to possible errors that might arise during the annotation of prokaryotic genome, only if no annotation on any of the two replicons was found can we trust that the transferred DNA is noncoding.

We found 72,035 pairs of homologous loci that are noncoding on both replicons, those pairs belong to 21,940 plasmid homologous loci and 32,398 chromosomal homologous loci. This type of DNA transfer may contribute to the evolution of genes (*de novo*) through transfer. Moreover, noncoding sequences can potentially comprise regulatory elements and numerous genetic elements.

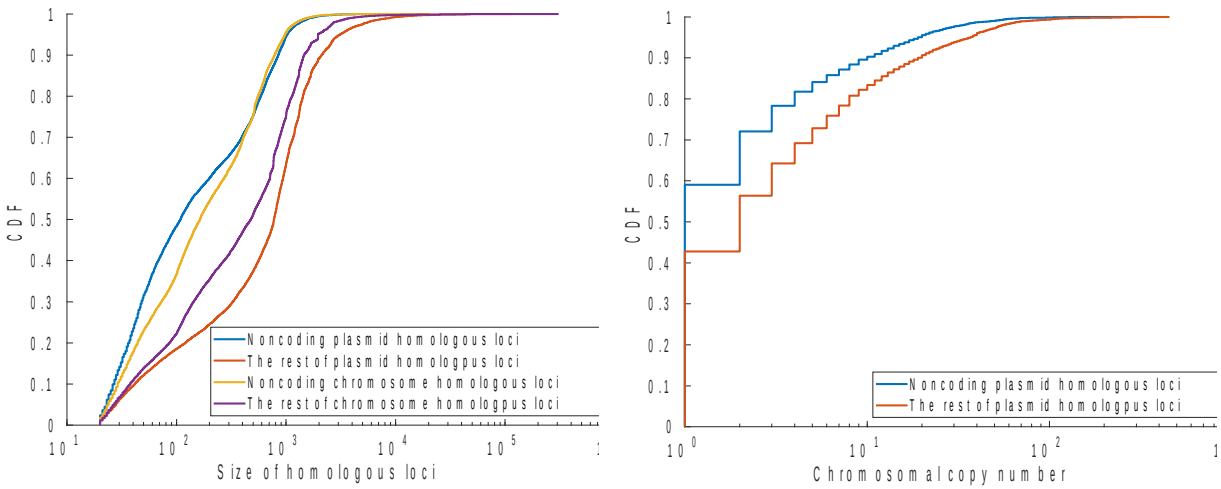


Figure 30: Comparison of noncoding and coding homologous loci. The left figure is the cumulative distribution of the size of completely noncoding plasmid and chromosomal homologous loci in comparison of the sizes of plasmid and chromosomal homologous loci that overlap fully or partially with coding sequences. The right figure is the cumulative distribution of the chromosomal copy number for noncoding plasmid loci in comparison to plasmid loci that overlap with coding sequences.

The majority of the noncoding plasmid homologous loci are small with a relatively low copy number on the chromosome

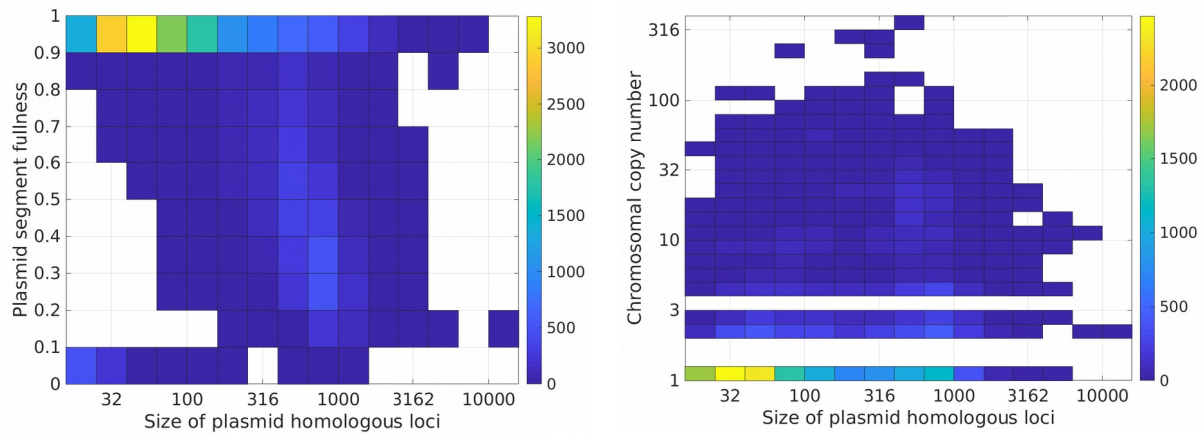


Figure 31: characteristics of noncoding plasmid homologous loci. The left figure represents a heatmap of the size of plasmid homologous loci vs. it's fullness. The right figure is a heat map of the size of noncoding plasmid homologous loci vs. their copy number on the chromosome.

Table 19: The number of shared plasmid loci categorized upon the type of sequence they contain or overlap, the number of genes they contain and the number of pairs and isolates that have each type of shared loci.

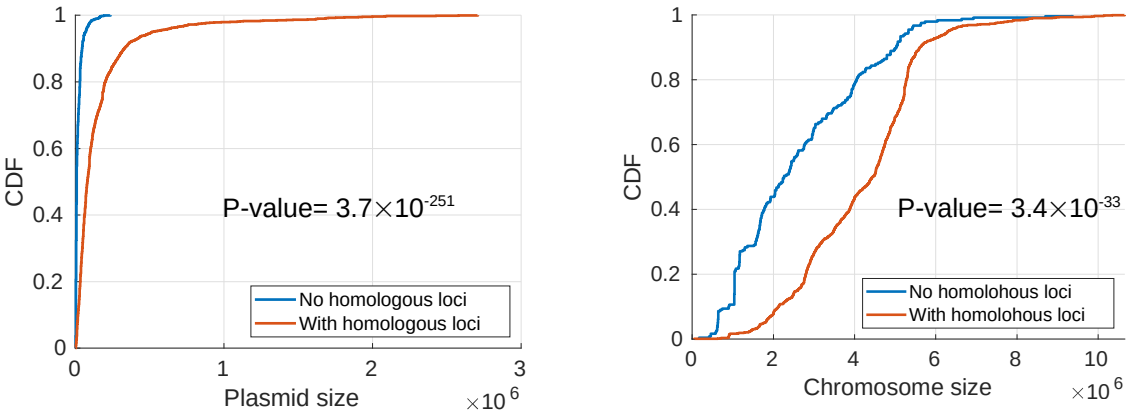
Type of homologous loci	Number of shared loci	Number of genes	Number of pairs	Number of isolates
Non-coding	11,578	0	1,375	843
With partial genes	30,039	33,349	1,832	996
With complete genes	10,392	20,825	1,724	929
MGLs	4212	14,645	1,247	746
Single-copy MGLs	3,552	12,532	1,100	671
Multi-copy MGLs	660	2,113	462	358
With AMR genes	74	90	58	48

7.5 Plasmid-Chromosome pairs with no homologous loci

We found 995 plasmids with no homologous loci with the chromosome of their host, those plasmids have a size range between 744 and 236,176 bp with a median of 7,921 bp so they mostly belong to the category of small plasmids (Figure 32) and were found significantly smaller than plasmids that have homologous loci with the

chromosome (Figure 32). By investigation other properties of those plasmids, we found that they have significantly lower GC content and lower number of coding sequences than plasmids that have homologous loci with the chromosome. The three measures, genome size, number of coding regions and the GC content are known to be causally associated. As genes usually have higher GC content than noncoding regions and they contribute to the increase of genome size. Moreover, a measure for genome complexity was adapted using the tool (macle) (Haubold *et al.*, 2009), a higher complexity score implies less repetitiveness in the genome (less duplications and repetitive elements). By applying macle to our genomic data, we found that plasmids with no homologous loci with the chromosome have a significantly higher complexity (less repetitiveness) than those that do have homologous loci with the chromosome. This result suggests that repetitive elements are major contributors to the DNA transfer and DNA homology between plasmids and chromosomes (Figure 32).

Similarly, we found that the chromosomes that do not have any homologous locus with any plasmid in its prokaryotic host are significantly smaller, have less GC content and less coding sequences than those that have homologous loci with one or more of the plasmids coinhabiting its host. Chromosomes with no homologous loci with plasmids were also found to have more genome complexity than those that have DNA homology with plasmids of its host (Figure 32). From those observations, we can also conclude that DNA transfer between plasmids and chromosomes contributes to the size increase and coding sequence enrichment of those replicons.



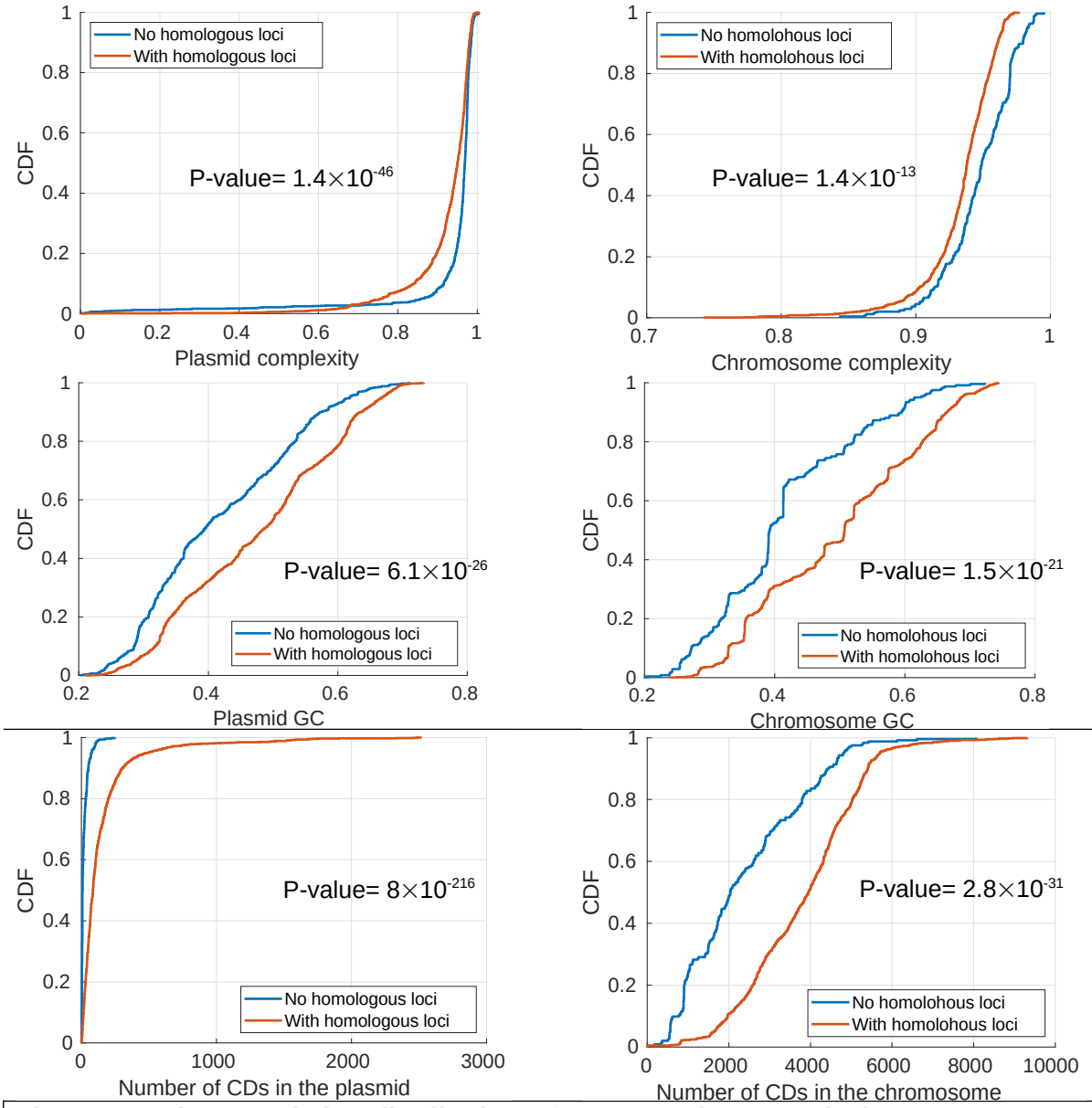


Figure 32: The cumulative distribution of genome size, complexity, GC content and number of coding sequences. In plasmids and chromosomes that have or do not have homologous loci shared with each other. P-values are calculated with kolmogorov-smirnov test to compare between replicons that have or do not have homologous loci with the null hypothesis that they were drawn from the same distribution and the alternative hypothesis that the distribution in replicons that have homologous loci is significantly different from the distribution if those that don't.

There are 881 other plasmids that co-inhabit the same host as the plasmids that have no homologous loci with the chromosome. Those plasmids share a proportion of their sequence with the chromosomes ranging between 0.2% and 100% with a

median of 10% (Figure 33). This suggests that not having homologous loci with the chromosomes is not a property of the host, but rather a property of the plasmid *per se*.

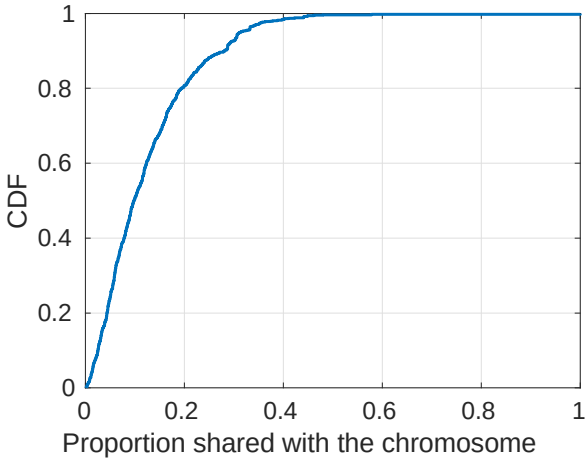


Figure 33: The cumulative distribution of the proportion of plasmids shared with the chromosome. Calculated for plasmids co-inhabiting the same host as the plasmids that have no homologous loci with their chromosomes.

Plasmids with no transfer events with the chromosome belong to 166 genera, the 30 most represented genera in our dataset are listed in Table 20 along with the number of all pairs they include as well as the number and proportion of pairs that have no sequence homology belonging to each of those genera.

We can observe differences between the different genera in the proportion of pairs with no sequence homology between the plasmid and chromosome. For instance, some genera like *Yersinia, Pseudomonas* and *Rhizobium* have the majority of their chromosome-plasmid pairs with sequence homology to each other. On another hand, genera like *Borrelia, Escherichia, Acinetobacter, Chlamydia, Acetobacter* have between 24% and 49% of their pairs with no sequence homology between the plasmid and chromosome (Table 20).

Table 20: The 30 most present genera in our dataset
The number of pairs those genera comprise as well as the number and percentage of

pairs that have no homologous loci between the plasmid and chromosome			
Genus	Number of all pairs	Number of pairs with no homologous loci	Percentage of pairs with no homologous loci
Bacillus	527	102	19%
Escherichia	339	82	24%
Salmonella	256	39	%
Klebsiella	233	40	17%
Staphylococcus	212	31	15%
Lactobacillus	211	38	18%
Streptococcus	194	4	2%
Campylobacter	170	19	11%
Yersinia	166	4	2%
Borrelia	149	74	50%
Candidatus	142	14	10%
Pseudomonas	139	5	4%
Chlamydia	135	36	27%
Mycobacterium	130	7	5%
Corynebacterium	124	4	3%
Xanthomonas	111	10	9%
Helicobacter	110	14	13%
Clostridium	101	6	6%
Acinetobacter	97	25	26%
Enterobacter	97	17	18%
Listeria	82	5	6%
Rhizobium	75	1	1%
Bifidobacterium	74	10	14%
Acetobacter	67	25	37%

Streptomyces	66	4	6%
Francisella	58	9	16%
Enterococcus	56	8	14%
Erwinia	53	5	9%
Rickettsia	51	4	8%
Paenibacillus	48	4	8%

7.6 Temporal dynamics of plasmid-chromosome transfer

We found DNA segments on the plasmid that are shared on the chromosome with various copy number and syntenic state. But we can assume that the number of transfer events can be represented with the number of plasmid segments. Moreover, we can accept the similarity range for plasmid segments for one chromosome-plasmid pair as proxy for the plasmid lifespan in the microbial host, or at least for the time since DNA transfer started between those two replicons. Taking those two proxies together (number of transfer events and time), we observe a significant positive correlation between the frequency of homologous loci and the range of homologous locus sequence similarity (r = 0.9, P-value << 0.01, using Spearman correlation). This correlation can be interpreted as a temporal signal of continuous transfer, indicating that DNA is flowing between the two replicons continuously over time.

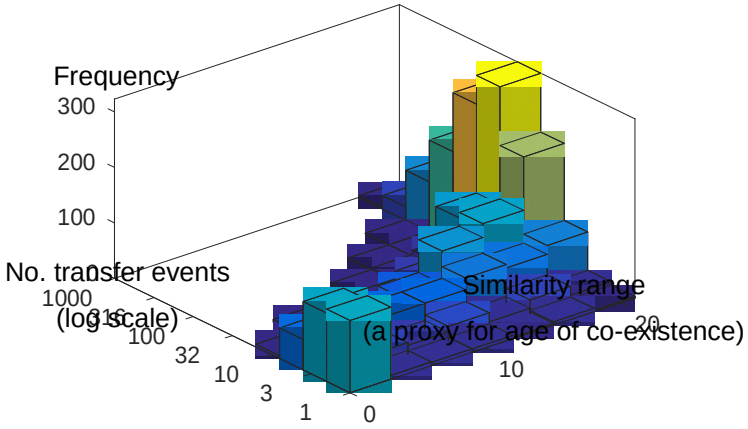


Figure 34: The correlation between the number of homologous loci and the homologous locus similarity range. The homologous locus similarity is estimated as the median BLAST similarity in each locus.

8 DISCUSSION

8.1 The segmentation approach

DNA homology between plasmids and chromosomes within the microbial host is frequent; 3,209 out of 3,264 plasmids (98% of all plasmids in our dataset) have a minimal sequence similarity with the chromosomes, as was found by sequence against sequence BLAST and MUMmer results. However, the number of BLAST or MUMmer hits between plasmids and chromosomes does not always correspond to the amount of shared DNA, which reflects the fact that BLAST and MUMmer are reporting shared sequence in multiple hits. This observation can be explained by a combination of two underlying reasons. First, the presence of multiple copies of the plasmid sequence on the chromosome due to duplications or multiple transfers of the same sequence. This makes multiple hits have the same co-ordinates on the plasmid but fall in different loci on the chromosome, which increase the number of hits found but not the actual amount of plasmid DNA shared with the chromosome. The second reason stems from genetic erosion that follows the transfer event, which causes the region of DNA sequence similarity to be split over several hits. This, besides the differences in the shared DNA found by the two different local similarity methods

(BLAST and MUMmer) made it necessary to have a deeper analysis that combines both types of hits (BLAST and MUMmer), connects hits and takes into consideration multiple copies of the shared DNA sequence.

The genomic segmentation approach we developed here results in segments of shared DNA on both of compared replicons. Segments are uni-dimensional, they are defined by a beginning and an end and contain BLAST and MUMmer hits on one replicon. Segments from both replicons taken together form intersects that are two dimensional and defined by four coordinates, two on the plasmid and two on the chromosome. Real intersects contain hits (the beginning and end of the hit falls inside the intersect on both replicons). While spurious intersects are between segments that have no hits in common.

This gives us a multiple level data structure as following:

- Each unit is one comparison pair that consists of two uni-dimensional DNA sequences that are the two replicons (a plasmid and a chromosome belonging to the same prokaryotic host).
- Within each pair, there are two sets of uni-dimensional segments, one set for each replicon.
- Each segment forms a two-dimensional intersect with every segment on the other replicon.
- The full intersects contain two-dimensional hits.

Among all intersects belonging to one plasmid segment, one intersect stands out as the pivot intersect, as it has the highest sequence similarity between the plasmid segment and all other chromosomal segments. Pivot intersects potentially represent an original transfer event, and other intersects can be debated as either duplications, genomic rearrangements or other transfer events. And this can only be investigated using phylogenomic methods.

Multiple advantages were achieved by applying the segmentation approach to our data:

- It could connect close local similarity hits that belong to the same transfer event in a parsimonious manner.
- It filters against noise more efficiently than purely using the E-value and size of BLAST hits, as it only ignores small random hits that are scattered, while keeping

those that are clustered in a close proximity with other hits. In other words, it considers the neighbourhood of small hits before deciding on their randomness.

- With the output data structure achieved by the segmentation, we could organize regions of DNA similarity in intersects and the intersects in segments.
- The data structuring and organisation makes it possible to interpret the observations systematically and infer events of transfer and genomic duplication as well as evolutionary changes on a smaller scale (within the intersects).
- The multiple levels of organization (hits, intersects, segments, pairs and isolates) made it easier to visualize the data and find patterns in a dot-plot like graphs that are condense, informative and easy to analyse and to compare between examples.
- The segmentation approach is minimalistic and independent from the input data, it uses pre-simulated data to find adaptive thresholds with a simple statistical reasoning that can apply to any binary data. Thus, it does not include many parameters that might increase the chance for errors.
- The segmentation is built on a top-down algorithm, unlike all other genome alignment methods that serve a similar purpose using seed and extend approaches. This difference gives it a potential for enhancing the detection sensitivity.
- The genomic segmentation works best for comparing the sequences of replicons with different evolutionary background (plasmids and chromosomes in our study case), and it has an advantage over genome alignment methods for analysing the sequence similarity for replicon pairs with large size difference.

8.2 The frequency of gene transfer

In this research, we focused on the magnitude of shared DNA between plasmids and chromosomes within the host. While this estimation may seem conservative, the documented proximity of both replicons in the same cell reduces the possibility for erroneous lateral gene transfer inference (e.g., due to phylogenetic artefacts (Roettger *et al.*, 2009)).

Applying comparative genomics to the entire DNA sequences of plasmids and chromosomes of the same prokaryotic host, revealed a considerable amount of shared DNA sequence with various length and similarity and copy number. A

substantial amount of the shared DNA lies in the noncoding regions of the plasmid DNA and includes many partial genes that are likely non-functional.

Our approach provided a framework to study horizontal gene transfer between different replicons despite complex evolutionary scenarios of duplications, rearrangements and genetic erosion. Furthermore, it allowed us to distinguish co-transferred genes and transfer events mediated by mobile genetic elements (transposases, IS elements and bacteriophages) and numerous antimicrobial resistance (AMR) genes have been detected as transferred between the plasmid and chromosome in multiple isolates. Those genes have evidently persisted on both replicons, probably due to advantageous dosage effect. Some of those AMR genes have co-transferred with other genes including other AMR genes causing multidrug resistance (coming from different plasmids in some cases).

The gene transfer events we have inferred here testify for a rather modest frequency of transfer between chromosomes and plasmids, nonetheless, plasmid mobility may contribute to further dispersal of acquired DNA. A correlation between the number of shared DNA regions and the range of shared DNA similarity (that serves as a proxy for the plasmid life spam) suggests a continuous dynamic of transfer of DNA. An important aspect of gene transfer between plasmids and chromosomes is the duplication of the transferred locus within the cell. Previous studies suggested that gene dose may be a barrier for lateral gen transfer specifically for complex systems where the stoichiometry of all components should be maintained constant (Sorek *et al.*, 2007; Jain *et al.*, 1999). Dose effect is likely relevant not only for the chromosome but also to the plasmid such that both replicons might experience deleterious effects as a result of gen duplication following transfer.

Our results could be furthermore used to test the validity of the plasmid paradox concept. For that, we will focus on the transfer of AMR genes between plasmid sand chromosomes: our data shows that out of 862 AMR encoding plasmids, AMR gene transfer to the chromosome occurred in only 14% of the plasmids. Furthermore, in those cases, the plasmid was not lost following the transfer event as predicted by the plasmid paradox, (at least at the sampling time point). Our data thus shows that although AMR genes are abundant on plasmids, they are not frequently transferred to the chromosome. This suggests that the importance of the

trade-off between plasmid cost and benefit may have been overestimated. Indeed, recent research in the literature indicates that plasmid acquisition may not always entail a significant cost to the host. For example, a recent study reported a mega conjugative plasmid in *Pseudomonas aeruginosa* carrying several AMR genes and reaching a size of ca. 420 kb that is neutral for the host fitness (Cazares *et al.*)

. An additional example is a study conducted using isolates from patients in a hospital in Madrid. In that study, the fitness cost of a natural plasmid encoding AMR was measured in natural isolates including *Escherichia* and *Klebsiella* strains. The results showed that the plasmid fitness cost was very often neutral to the host fitness (Valle *et al.*, 2020). We suggest an alternative explanation to the rarity of AMR gene transfer between plasmids and chromosomes: The acquisition of AMR plasmids under selection for antibiotic resistance has been shown to be accompanied by host adaptation to the plasmid (Millan *et al.*, 2017; Loftie-Eaton *et al.*, 2016). Thus, strong positive selection for the plasmid maintenance leads, most likely to rapid host-plasmid coadaptation and eventually reduces the plasmid fitness cost to the host. When the antibiotics are removed, the trade-off between plasmid cost and benefit does no longer exist, such that the plasmid can maintain a stable persistence in the population.

In summary, our results supply evidence for mechanisms that mediate DNA transfer between plasmids and chromosomes. While the transfer of many loci may be mediated by homologous recombination, the most common transfer mechanism we observe is transposition. This suggests that most of the transfer between plasmid and chromosome is mediated by active transposition or integration mechanisms. In other words, mobile elements are more likely to be transferred rather than random plasmid loci. Out of the 1,846 plasmids that contain transposons in their genome, we found 1,726 plasmids where shared transposon between the plasmid and chromosome could be observed. Plasmids play a role in the dissemination of transposons much more than they do antibiotic resistance.

9 ACKNOWLEDGEMENT

Papa, you were and will always be my role model and energy source. Mama, Rania, Noura, Sana, the distance is far and the years are long, but you are always in my mind.

Thank you Giddy, it was a true honour to be mentored by a great scientist like yourself. I have learned so much from you. Tal, I cannot thank you enough, you took me as a curious student, shaped my critical mind, and made me into what I am now, thanks for all the scientific and personal guidance along my long journey in the GMG.
Thank you, Dima, for all your support and for what you went through during my PhD.
Thank you, my Lulu, for brightening my life.
My friends from Syria (Yamen, Zak, Wael, Nijo and all the others) and Christin I'm very happy to have you all in my life.

Devani, Maxime, Fernando, Robin, Ana and Andrea, thank you for being a true family in my expatriation. Extra thanks for Christin, Devani and Fena for their help with this thesis. My great friends and colleagues at IMPRS for evolutionary biology and at the genomic microbiology group (GMG), I enjoyed all our scientific discussions and extra-scientific fun times.

I reserve especial thanks to the members of my thesis advisory committee: Dr. Julien Dutheil and Prof. Dr. Ute Hentschel for the discussions and support along my PhD. I am also thankful for the opportunity I was given to be a part of the International Max Planck Research School (IMPRS) for evolutionary biology. My gratitude is extended to Kiel university for providing an equal opportunity for foreigner students in a free high quality education and a wonderful international atmosphere, I will forever be in Kiel's debt. Finally, thanks for whoever contributed to this work through advice, criticism or fruitful discussions and to all the labs around the world that made their data publically available, without which, such research would not be possible.

10 REFERENCES

Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, *et al.* (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucl. Acids. Res.* 48: D517–D525.

Anda M, Ohtsubo Y, Okubo T, Sugawara M, Nagata Y, Tsuda M, *et al.* (2015). Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. *Proc. Natl. Acad. Sci. USA* 112: 14343–14347.

Baxter JC, Funnell BE. (2014). Plasmid partition mechanisms. *Microbiol. Spectr.*, 10.1128/microbiolspec.PLAS-0023-2014.

Bray N, Dubchak I, Pachter L. (2003). AVID: A global alignment program. *Genome Res.* 13: 97–102.

Brinkmann H, Göker M, Koblížek M, Wagner-Döbler I, Petersen J. (2018). Horizontal operon transfer, plasmids, and the evolution of photosynthesis in *Rhodobacteraceae*. *ISME J* 12: 1994–2010.

Cabezón E, Ripoll-Rozada J, Peña A, La Cruz de F, Arechaga I. (2015). Towards an integrated model of bacterial conjugation. *FEMS Microbiol. Rev.* 39: 81–95.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10: 1–9.

Camerini-Otero RD, Hsieh P. (1995). Homologous recombination proteins in prokaryotes and eukaryotes. *Annu. Rev. Genet*.

Cazares A, Moore MP, Hall J, Nature LW, 2020. A megaplasmid family driving dissemination of multidrug resistance in *Pseudomonas*. *naturecom.*,10.1038/s41467-020-15081-7.

Chen Z, Lin J, Ma C, Zhao S, She Q, Liang Y. (2014). Characterization of pMC11, a plasmid with dual origins of replication isolated from *Lactobacillus casei* MCJ and construction of shuttle vectors with each replicon. *Appl. Microbiol. Biotechnol.* 98: 5977–5989.

Cooper TF, Heinemann JA. (2000). Postsegregational killing does not increase plasmid stability but acts to mediate the exclusion of competing plasmids. *Proc. Natl. Acad Sci USA* 97: 12643–12648.

Couchman EC, Browne HP, Dunn M, Lawley TD, Songer JG, Hall V, *et al.* (2015). *Clostridium sordellii* genome analysis reveals plasmid localized toxin genes encoded within pathogenicity loci. *BMC Genomics* 16: 392–13.

Darling ACE, Mau B, Blattner FR, Perna NT. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14: 1394–1403.

Darling AE, Mau B, Perna NT. (2010). ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5: e11147.

diCenzo G, Milunovic B, Cheng J, Finan TM. (2013). The tRNAarg gene and engA are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. *J. Bacteriol.* 195: 202–212.

Domingues S, da Silva GJ, Nielsen KM. (2012). Integrons. *Mob. Genet. Elements* 2: 211–223.

Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, *et al.* (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359: eaar4120.

Dubey GP, Ben-Yehuda S. (2011). Intercellular nanotubes mediate bacterial communication. *Cell* 144: 590–600.

Dziewit L, Pyzik A, Szuplewska M, Matlakowska R, Mielnicki S, Wibberg D, *et al.* (2015). Diversity and role of plasmids in adaptation of bacteria inhabiting the Lubin copper mine in Poland, an environment rich in heavy metals. *Front. Microbiol.* 6: 152.

Erdmann S, Tschitschko B, Zhong L, Raftery MJ, Cavicchioli R. (2017). A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. *Nat. Microbiol.* 2: 1446–1455.

Falb M, Pfeiffer F, Palm P, Rodewald K, Hickmann V, Tittor J, *et al.* (2005). Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis*. *Genome Res.* 15: 1336–1343.

Fernández M, Margolles A, Suárez JE, Mayo B. (1999). Duplication of the beta-galactosidase gene in some *Lactobacillus plantarum* strains. *Int. J. Food Microbiol.* 48: 113–123.

Fernández-Tresguerres ME, Martín M, de Viedma DG, Giraldo R, Díaz-Orejas R. (1995). Host growth temperature and a conservative amino acid substitution in the replication protein of pPS10 influence plasmid host range. *J. Bacteriol.* 177: 4377–4384.

Fondi M, Bacci G, Brilli M, Papaleo MC, Mengoni A, Vaneechoutte M, *et al.* (2010). Exploring the evolutionary dynamics of plasmids: the *Acinetobacter* pan-plasmidome. *BMC Evol. Biol.* 10: 59.

Fulsundar S, Harms K, Flaten GE, Johnsen PJ, Chopade BA, Nielsen KM. (2014). Gene transfer potential of outer membrane vesicles of *Acinetobacter baylyi* and effects of stress on vesiculation. *Appl. Environ. Microbiol.* 80: 3469–3483.

Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, *et al.* (2009). The NCBI BioSystems database. *Nuc.l Acids. Res.* 38: gkp858–D496.

Gibbs AJ, Mcintyre GA. (1970). The Diagram, a method for comparing sequences. *FEBS J.* 16: 1–11.

Gil R, Sabater-Muñoz B, Perez-Brocal V, Silva FJ, Latorre A. (2006). Plasmids in the aphid endosymbiont *Buchnera aphidicola* with the smallest genomes. A puzzling evolutionary story. *Gene.* 370: 17–25.

Graur D. (2016). Molecular and genome evolution. First edition. *Sinauer Associates, Inc*.

Graur D, Zheng Y, Azevedo RBR. (2015). An evolutionary classification of genomic function. *Genome biol. evol.* 7: 642–645.

Gullberg E, Albrecht LM, Karlsson C, Sandegren L, Andersson DI. (2014). Selection of a multidrug resistance plasmid by sublethal levels of antibiotics and heavy metals. *MBio.* 5: e01918–14.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. (2010). Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci. USA* 107: 127–132.

Hall JPJ, Williams D, Paterson S, Harrison E, Brockhurst MA. (2017). Positive selection inhibits gene mobilisation and transfer in soil bacterial communities. *Nat. Ecol. Evol.* 1: 1348–1353.

Harrison E, Brockhurst MA. (2012). Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol.* 20: 262–267.

Harrison E, Guymer D, Spiers AJ, Paterson S, Brockhurst MA. (2015). Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. *Curr. Biol.* 25: 2034–2039.

Haubold B, Pfaffelhuber P, Domazet-Loso M, Wiehe T. (2009). Estimating mutation distances from unaligned genomes. *J. Comput. Biol.* 16: 1487–1500.

He S, Hickman AB, Varani AM, Siguier P, Chandler M, Dekker JP, *et al.* (2015). Insertion sequence IS26 reorganizes plasmids in clinically isolated multidrug-resistant bacteria by replicative transposition. *MBio.* 6: e00762.

Hertwig S, Popp A, Freytag B, Lurz R, Appel B. (1999). Generalized transduction of small *Yersinia enterocolitica* plasmids. *Appl. Environ. Microbiol.* 65: 3862–3866.

Hille J, Van Kan J, Schilperoort R. (1984). Trans-Acting virulence functions of the octopine Ti plasmid from *Agrobacterium tumefaciens. J. Bacteriol.* 158: 754–756.

Hooper SD, Mavromatis K, Kyrpides NC. (2009). Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol.* 10: R45.

Hülter N, Ilhan J, Wein T, Kadibalban AS, Hammerschmidt K, Dagan T. (2017). An evolutionary perspective on plasmid lifestyle modes. *Curr. Opin. Microbiol.* 38: 74–80.

Jain R, Rivera MC, Lake JA. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* 96: 3801–3806.

Jechalke S, Heuer H, Siemens J, Amelung W, Smalla K. (2014). Fate and effects of veterinary antibiotics in soil. *Trends Microbiol.* 22: 536–545.

Jørgensen TS, Xu Z, Hansen MA, Sørensen SJ, Hansen LH. (2014). Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS ONE* 9: e87924.

Klieve AV, Yokoyama MT, Forster RJ, Ouwerkerk D, Bain PA, Mawhinney EL. (2005). Naturally occurring DNA transfer system associated with membrane vesicles in cellulolytic *Ruminococcus* spp. of ruminal origin. *Appl. Environ. Microbiol.* 71: 4248–4253.

Kopfmann S, Roesch SK, Hess WR. (2016). Type II Toxin-Antitoxin Systems in the Unicellular Cyanobacterium Synechocystis sp. PCC 6803. *Toxins (Basel)* 8: 228.

Kowalczykowski SC, Dixon DA, Eggleston AK, Lauder SD, Rehrauer WM. (1994). Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol. Rev.* 58: 401–465.

Köster J, Rahmann S. (2012). Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* 28: 2520–2522.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, *et al.* (2004). Versatile and open software for comparing large genomes. *Genome. Biol.* 5: R12.

Latorre A, Gil R, Silva FJ, Moya A. (2005). Chromosomal stasis versus plasmid plasticity in aphid endosymbiont *Buchnera aphidicola*. *Heredity* 95: 339–347.

Lederberg J, Tatum EL. (1946). Gene recombination in *Escherichia coli. Nature* 158: 558–558.

Loftie-Eaton W, Yano H, Burleigh S, Simmons RS, Hughes JM, Rogers LM, *et al.* (2016). Evolutionary paths that expand plasmid host-range: Implications for spread of antibiotic resistance. *Mol. Biol. Evol.* 33: 885–897.

Maestro B, Sanz JM, Díaz-Orejas R, Fernández-Tresguerres E. (2003). Modulation of pPS10 host range by plasmid-encoded RepA initiator protein. *J. Bacteriol.* 185: 1367–1375.

Marraffini LA, Sontheimer EJ. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322: 1843–1845.

Mazel D. (2006). Integrons: agents of bacterial evolution. *Na.t Rev. Microbiol.* 4: 608–620.

Michael V, Frank O, Bartling P, Scheuner C, Göker M, Brinkmann H, *et al.* (2016). Biofilm plasmids with a rhamnose operon are widely distributed determinants of the 'swim-or-stick' lifestyle in roseobacters. *ISME J* 10: 2498–2513.

Millan AS, Escudero JA, Gifford DR, Mazel D, MacLean RC. (2017). Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Na.t Ecol. Evol.* 1: 0010.

Morikawa K, Takemura AJ, Inose Y, Tsai M, Le Thuy Nguyen Thi, Ohta T, *et al.* (2012). Expression of a cryptic secondary sigma factor gene unveils natural competence for DNA transformation in *Staphylococcus aureus* . *PLOS Pathogens* 8: e1003003.

Naito T, Kusano K, Kobayashi I. (1995). Selfish behavior of restriction-modification systems. *Science* 267: 897–899.

Needleman SB, Wunsch CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48: 443–453.

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. *Nat. Commun.* 2: 268–11.

Panesso D, Planet PJ, Diaz L, Hugonnet J-E, Tran TT, Narechania A, *et al.* (2015). Methicillin-susceptible, vancomycin-resistant *Staphylococcus aureus*, Brazil. *Emerg. Infect. Dis.* 21: 1844–1848.

Popa O, Klösges T, Landan G, Dagan T. (2015). Phylogenomic networks of microbial genome evolution. In: *Manual of environmental microbiology, 4th edition*. American Society of Microbiology, pp 4.1.1–1–4.1.1–18.

Porse A, Schønning K, Munck C, Sommer MOA. (2016). Survival and evolution of a large multidrug resistance plasmid in new clinical bacterial hosts. *Mol. Bio.l Evol*, 10.1093/molbev/msw163.

Ramirez MS, Don M, Merkier AK, Bistué AJS, Zorreguieta A, Centrón D, *et al.* (2010). Naturally competent *Acinetobacter baumannii* clinical isolate as a convenient model for genetic studies. *J. Clin. Microbiol.* 48: 1488–1490.

Ramsay JP, Firth N. (2017). Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* 38: 1–9.

Roer L, Aarestrup FM, Hasman H. (2015). The EcoKI type I restriction-modification system in *Escherichia coli* affects but is not an absolute barrier for conjugation. Gourse RL (ed). *J Bacteriol.* 197: 337–342.

Roettger M, Martin W, Dagan T. (2009). A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Mol. Biol. Evol.* 26: 1931–1939.

Roux S, Hallam SJ, Woyke T, Sullivan MB. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4: 1.

Sakuma T, Tazumi S, Furuya N, Komano T. (2013). ExcA proteins of IncI1 plasmid R64 and IncIy plasmid R621a recognize different segments of their cognate TraY proteins in entry exclusion. *Plasmid* 69: 138–145.

Schmidt R, Ahmetagic A, Philip DS, Pemberton JM. (2011). Catabolic plasmids. *eLs.*

Scolnik PA, Haselkorn R. (1984). Activation of extra copies of genes coding for nitrogenase in *Rhodopseudomonas capsulata*. *Nature.* 307: 289–292.

Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, La Cruz de F. (2010). Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74: 434–452.

Smith TF, Waterman MS. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*

Sonnhammer ELL, Durbin R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: GC1–GC10.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318: 1449–1452.

Sota M, Yano H, Hughes JM, Daughdrill GW, Abdo Z, Forney LJ, *et al.* (2010). Shifts in the host range of a promiscuous plasmid through parallel evolution of its replication initiation protein. *ISME J* 4: 1568–1580.

Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, *et al.* (2016). Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *MBio.* 7: e02162.

Szabó M, Nagy T, Wilk T, Farkas T, Hegyi A, Olasz F, *et al.* (2016). Characterization of two multidrug-resistant IncA/C Plasmids from the 1960s by using the minION sequencer device. *Antimicrob. Agents. Chemother.* 60: 6780–6786.

Tett A, Spiers AJ, Crossman LC, Ager D, Ciric L, Dow JM, *et al.* (2007). Sequence-based analysis of pQBR103; a representative of a unique, transfer-proficient mega plasmid resident in the microbial community of sugar beet. *ISME J* 1: 331–340.

Thomas CM, Nielsen KM. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3: 711–721.

Thomas CM, Thomson NR, Cerdeño-Tárraga AM, Brown CJ, Top EM, Frost LS. (2017). Annotation of plasmid genes. *Plasmid* 91: 61–67.

Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22: 4673–4680.

Tippelt A, Möllmann S, Albersmeier A, Jaenicke S, Rückert C, Tauch A. (2014). Mycolic acid biosynthesis genes in the genome sequence of *Corynebacterium atypicum* DSM 44849. *Genome Announc.* 2: 349.

Tock MR, Dryden DTF. (2005). The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* 8: 466–472.

Toh H, Oshima K, Nakano A, Takahata M, Murakami M, Takaki T, *et al.* (2013). Genomic Adaptation of the *Lactobacillus casei* Group Schacherer J (ed). *PLoS ONE* 8: e75073.

Valle AA-D, León-Sampedro R, Rodríguez-Beltrán J, DelaFuente J, Hernández-García M, Ruiz-Garbajosa P, *et al.* (2020). The distribution of plasmid fitness effects explains plasmid persistence in bacterial communities. *bioRxiv.* 4: 2020.08.01.230672.

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

Wein T, Hülter NF, Mizrahi I, Dagan T. (2019). Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. *Nat. Commun.* 10: 2595–13.

Wein T, Wang Y, Hülter NF, Hammerschmidt K, Dagan T. (2020). Antibiotics Interfere with the Evolution of Plasmid Stability. *Curr. Biol*. 10.1016/j.cub.2020.07.019.

Wheeler TJ, Eddy SR. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29: 2487–2489.

Xue H, Cordero OX, Camas FM, Trimble W, Meyer F, Guglielmini J, *et al.* (2015). Eco-evolutionary dynamics of episomes among ecologically cohesive bacterial populations. *MBio.* 6: e00552–15.

Zaleski P, Wawrzyniak P, Sobolewska A, Łukasiewicz N, Baran P, Romańczuk K, *et al.* (2015). pIGWZ12--A cryptic plasmid with a modular structure. *Plasmid* 79: 37–47.

Zhang Z, Schwartz S, Wagner L, Miller W. (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol.* 7: 203–214.

Zheng J, Guan Z, Cao S, Peng D, Ruan L, Jiang D, *et al.* (2015). Plasmids are vectors for redundant chromosomal genes in the *Bacillus cereus* group. *BMC Genomics* 16: 6.

Zielenkiewicz U, Cegłowski P. (2001). Mechanisms of plasmid stable maintenance with special focus on plasmid addiction systems. *Acta Biochim Pol* 48: 1003–1023.

11 SUPPLEMENTARY DATA

Supplementary table 1: 90 transferred AMR genes belonging to 21 families	
AMR gene family	Number of transferred genes
Sulfonamide resistant sul	17
SHV beta-lactamase	16
ANT(3)	14
TEM beta-lactamase	6
AAC(6)	5
Resistance-nodulation-cell division (RND) antibiotic efflux pump	4
APH(3)	4
AAC(3)	3
CTX-M beta-lactamase	3
blaZ beta-lactamase	2
APH(6)	2
OXA beta-lactamase	2
AAC(3); AAC(6)	2
KPC beta-lactamase	2
ANT(6)	2
ATP-binding cassette (ABC) antibiotic efflux Pump; major facilitator superfamily (MFS) antibiotic efflux pump	1
ANT(2)	1
Quinolone resistance protein (qnr)	1
APH(3)	1
16S rRNA methyltransferase (G1405)	1
Tetracycline inactivation enzyme	1

Supplementary table 2: transferred AMR belonging to 14 drug classes	
AMR drug class	Number of transferred genes
Aminoglycoside antibiotic	34
Sulfonamide antibiotic	17
Carbapenem; cephalosporin; penam	16
Monobactam; cephalosporin; penam; penem	6
Fluoroquinolone antibiotic; tetracycline antibiotic	3

Cephalosporin	3
Penam	2
Cephalosporin; penam	2
Monobactam; carbapenem; cephalosporin; penam	2
Fluoroquinolone antibiotic; cephalosporin; glycylcycline; penam; tetracycline antibiotic; acridine dye; rifamycin antibiotic; ph... <preview truncated at 128 characters>	1
Fluoroquinolone antibiotic; aminoglycoside antibiotic	1
Fluoroquinolone antibiotic	1
Aminoglycoside antibiotic; aminocoumarin antibiotic	1
Glycylcycline; tetracycline antibiotic	1

Supplementary table 3: the most common products that are co-transferred with AMR genes	
Transposase	137
Hypothetical protein	38
Oxidoreductase	9
Integrase	8
Membrane protein	4
Recombinase	3
Aldehyde reductase	3
Aldolase	2
Atpase AAA	2
Regulatory protein blar1	1
Multidrug efflux protein	1
Qacedelta	1
Atpase	1
Resolvase	1
Protein tniq	1
Endonuclease	1
Group II intron-encoded protein ltra	1
PIN domain protein	1
L-rhamnose mutarotase	1
Lactaldehyde reductase	1
L-rhamnose isomerase	1
Rhamnulokinase	1
Transcriptional activator rhar	1
Protein yiiy	1
Aminoimidazole riboside kinase	1
ADP-ribosylglycohydrolase	1
Autoinducer kinase	1
Epimerase	1
Putative uncharacterized protein yiiq	1
Ferredoxin--NADP reductase	1
Glycerol metabolic protein	1
Glycerol kinase	1
Glycerol uptake facilitator protein	1
Cell division protein ftsn	1
Arylsulfate sulfotransferase	1
O-succinylhomoserine (thiol)-lyase	1
Mechanosensitive channel mscs	1
5-nucleotidase	1
5%2C10 methylenetetrahydrofolate reductase	1
Catalase	1
NADH:quinone oxidoreductase	1
Glycerol dehydrogenase	1

Transaldolase	1

Supplementary table 4: The number of mobile genetic elements found in plasmids that are fully integrated within the chromosomal genome.

Isolate	Plasmid accession	Transposase	Integrase	IS	Phage
Shewanella baltica OS155	CP000564.1	0	3	0	0
Shewanella baltica OS155	CP000567.1	0	0	0	0
Natronomonas pharaonis DSM 2160	CR936259.1	0	1	0	0
Erwinia sp. Ejp617	CP002127.1	0	0	0	0
Corynebacterium falsenii DSM 44353	CP007157.1	2	0	0	0
Corynebacterium atypicum	CP008945.1	4	4	0	2
Corynebacterium ureicelerivorans	CP009216.1	1	2	0	0
Bacillus thuringiensis	CP013001.1	39	29	0	0
Lactobacillus casei subsp. casei ATCC 393	AP012546.1	3	0	0	0
Paenibacillus sp. IHBB 10380	CP010977.1	0	3	0	0
Corynebacterium mustelae	CP011544.1	4	0	0	3
Geobacillus stearothermophilus 10	CP008935.1	6	0	0	0
Staphylococcus aureus	CP012595.1	3	4	0	0
Lactobacillus plantarum	CP012653.1	0	1	0	0
Arthrobacter alpinus	CP013201.1	0	0	0	0

Chryseobacterium sp. IHB B 17019	CP013294.1	0	0	0	0
Arthrobacter sulfonivorans	CP013748.1	1	0	0	0

Supplementary table 5 (attached as an e-copy): Information on all the homologous loci between plasmids and chromosomes.

Supplementary table 6 (attached as an e-copy): Information on all the plasmid genes that are fully or partially shared with the chromosome.

Supplementary table 7 (attached as an e-copy): Information on all the plasmid genes that are co-transferred with other genes between plasmids and chromosomes.

Supplementary figures (attached as an e-copy): 2D plots for 2,296 pairs of chromosome-plasmid that belong to the same prokaryotic strain and have shared DNA sequences among them.