Universidade do Minho
Escola de Engenharia
Departamento de Informática

Catarina Freitas de Sousa Santos

Development of a Database and web tool for the in silico characterization of plasmid data

November 2017
Development of a Database and web tool for the \textit{in silico} characterization of plasmid data

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Professor Miguel Francisco Almeida Pereira Rocha
Professora Cláudia Sofia Soares de Oliveira

November 2017
ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisors, Professor Miguel Rocha and Professor Cláudia Oliveira, without whose guidance this process could not be completed. I would especially like to thank Professor Cláudia Oliveira, for being such a positive influence and a constant presence throughout this journey. You have always had the right thing to say, the power to calm me down and been an inspiration and an enormous role model to me.

I have to thank my parents the most, for being the best parents I could ask for and for the inconditional love and support and enduring my (not so) occasional stress-induced insanity.

To my godmother, Carmen, and my cousin, Susana: thank you for being a part of my life and always supporting me.

To Débora, Inês, Joana, Filipa and Rita: thank you for being great friends and always being a call away.

And last, but definitely not least, to Pedro, for being my happy place. Thank you, from the bottom of my heart, for always being there for me.
ABSTRACT

Bacterial plasmids are mobile genetic structures capable of conferring selective advantages to their hosts, such as resistance to antibiotics, virulence genes and tolerance to pollutants.

By associating with other genetic elements, like integrons and transposons, plasmids provide a platform for genetic recombination and for gene transfer between different bacterial species, allowing them to colonize multiple environments and guaranteeing their persistence.

Although there are over 4000 complete plasmid sequences available in GenBank, most have absent or non-standardized (disorganized) information regarding their isolation source, environment and year and country of isolation. Furthermore, a prediction about their mobility and incompatibility group is also lacking.

The goals of this thesis are, besides completing the missing information about plasmid data, the development of a repository of fully sequenced plasmids and the development of easy-to-use web tools for the characterization of plasmid data regardless of their source environment and bacterial host. For the development of these tools, Shiny was used, which is a package from the R scientific computing environment.

The present work is organized as follows: the core concepts related to plasmids are described, their background is characterized and a critical analysis of the available web tools for plasmid classification is carried out. Then, the adopted approach and the development (implementation, outcomes) of the database and web tool are explained. Lastly, the main conclusions are highlighted.
RESUMO

Os plasmídeos bacterianos são estruturas genéticas móveis capazes de conferir vantagens seletivas ao seu hospedeiro, tais como resistência a antibióticos, genes de virulência e tolerância a poluentes.

Ao associar-se a outros elementos genéticos, como integrações e transposões, os plasmídeos constituem uma plataforma para a recombinação genética e transferência de genes entre diferentes espécies bacterianas, permitindo que colonizem múltiplos ambientes e garantindo a sua persistência.

Embora existam acima de 4000 sequências completas de plasmídeos disponíveis no GenBank, a maioria apresenta informação ausente ou não sistematizada (desorganizada) em relação à sua fonte de isolamento, ambiente e país e ano de isolamento.

Os objetivos desta tese são, para além de completar a informação em falta sobre plasmídeos, o desenvolvimento de um repositório de plasmídeos completamente sequenciados e a disponibilização de ferramentas online facilmente utilizáveis para a caracterização de plasmídeos independentemente da sua fonte de isolamento e hospedeiro bacteriano. Para o desenvolvimento destas ferramentas, foi utilizado o Shiny, que é um package do sistema de computação científica R.

Este trabalho está organizado da seguinte forma: os principais conceitos relacionados com os plasmídeos são apresentados, a sua história é caracterizada e é efetuada uma análise das ferramentas online existentes para a classificação de plasmídeos. Depois, a abordagem utilizada e a ferramenta desenvolvida são explicadas. Finalmente, as principais conclusões são destacadas.
## CONTENTS

1 INTRODUCTION 1
   1.1 Motivation 1
   1.2 Objectives 1
   1.3 Organization of the Thesis 2
2 STATE OF THE ART 3
   2.1 Plasmid Definition and Main Characteristics 3
      2.1.1 Plasmid Replication 5
      2.1.2 Replication Control 8
      2.1.3 Plasmid Stability and Partition 9
      2.1.4 Antibiotic Resistance and Virulence 10
   2.2 Mobile Genetic Elements 11
      2.2.1 Horizontal Gene Transfer 12
      2.2.2 Plasmid Mobility Groups and Mating Pair Formation 14
   2.3 Plasmid Incompatibility 15
   2.4 Plasmids at the NCBI Database 19
   2.5 Available Databases and Programs for Mobile Genetic Elements 20
3 DEVELOPMENT 22
   3.1 Construction and Structure of the Plasmid Database 22
   3.2 Populating the Database 24
      3.2.1 Obtaining the Plasmid Files 24
      3.2.2 Getting the Taxonomy Information 27
      3.2.3 Classifying According to Incompatibility and Mobility 27
   3.3 Shiny 29
   3.4 The PlasmidClassifier Application 32
      3.4.1 Lists/Search Tool and Interactive Plots Tab 38
      3.4.2 Plasmid Classification Panel 40
4 RESULTS AND DISCUSSION 48
   4.1 Plasmid Classification Results 48
   4.2 The PlasmidClassifier Application 50
      4.2.1 Generated Plots and Tables 50
      4.2.2 Classification Results 51
5 CONCLUSIONS AND FUTURE WORK 58
   5.1 Conclusions 58
5.2 Prospect for future work  

References  

A ATTACHMENTS  
A.1 User Report Without Matches  
A.2 User Report With a Single Match  
A.3 User Report With Multiple Matches
LIST OF ACRONYMS

**BLAST** Basic Local Alignment Search Tool.

**CRAN** Comprehensive R Archive Network.

**ctRNA** countertranscribed RNA.

**EER** Enhanced Entity-Relationship.

**HGT** Horizontal Gene Transfer.

**IDE** Integrated Development Environment.

**Inc** Incompatibility.

**IS** Insertion sequences.

**MGE** Mobile Genetic Elements.

**Mpf** Mating Pair Formation.

**mrs** Multimer Resolution System.

**ORFs** Open Reading Frames.

**ori** Origin of replication.

**oriV** Origin of vegetative replication.

**pRNA** primer RNA.

**PSI-BLAST** Position-Specific Iterative Basic Local Alignment Search Tool.

**PSK** Post-Segregational Killing.

**RCR** Rolling Circle Replication.

**Rep** Replication Initiation.

**RM** Restriction Modification.

**SQL** Structured Query Language.

**Tn** Transposons.

**T4CP** Type 4 Coupling Protein.

**T4SS** Type IV Secretion System.

**TA** Toxin-Antitoxin.

**TE** Transposable elements.

**ui** User-Interface.

**WGS** Whole Genome Shotgun.
LIST OF FIGURES

Figure 1  Interaction between some Mobile Genetic Elements  11
Figure 2  Lineage distribution of the complete plasmid sequences in the NCBI Plasmid Genome Database.  20
Figure 3  Enhanced Entity-Relationship (EER) diagram depicting the structure of the database.  23
Figure 4  Tooltip used in the Start Classification Process button  31
Figure 5  Disabled textInput  31
Figure 6  General aesthetics of the PlamidClassifier Application  35
Figure 7  Homepage of the PlamidClassifier application  36
Figure 8  Definitions Tab of the PlamidClassifier application  37
Figure 9  About/Contacts Tab of the PlamidClassifier application  38
Figure 10  Side panel of the Lists/Search Tool and Interactive Plots Tab highlighting the available filters and transferability options  39
Figure 11  Plot options available in the application with an empty plot in the background  40
Figure 12  Appearance of the Plamid Classification panel upon initialization  41
Figure 13  Output in the Sequence tab for existing sequence (A) without or (B) by uploading it  42
Figure 14  Application response when an unexisting sequence file is inputted  43
Figure 15  Sequence details (A) without and (B) after Classification process  44
Figure 16  Adding a new sequence to the Database  45
Figure 17  Sequences in the seqRepository table  47
Figure 18  Distribution Plots depicted in the PlasmidClassifier app when no filters are selected  53
Figure 19  Proportion Plots drawn in the PlasmidClassifier app when no filters are selected, using the ggplot2 package  54
Figure 20  First 10 entries of the Plasmid Table available in the Lists/Search and Interactive Plots panel when no filters are selected  55
Figure 21  First 10 entries of the Taxonomy Table available in the Lists/Search and Interactive Plots panel when no filters are selected  56
Figure 22  First 10 entries of the MOB Table available in the Interactive Plots panel when no filters are selected  57
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Common plasmid initiator proteins</td>
<td>5</td>
</tr>
<tr>
<td>Table 2</td>
<td>Mobilization, occurrence, host range and general characteristics of known plasmid Incompatibility groups</td>
<td>17</td>
</tr>
<tr>
<td>Table 3</td>
<td>Archetype plasmid, GenBank number, replicon typing (amplicon size and target gene region) and bibliography of known plasmid Incompatibility groups</td>
<td>18</td>
</tr>
<tr>
<td>Table 4</td>
<td>Prefixes found in the Plasmid Genome Database</td>
<td>19</td>
</tr>
<tr>
<td>Table 5</td>
<td>Websites and references for the web tools described in section 2.5</td>
<td>21</td>
</tr>
<tr>
<td>Table 6</td>
<td>File types found at the NCBI FTP repository concerning plasmid data</td>
<td>24</td>
</tr>
<tr>
<td>Table 7</td>
<td>Functions used in the Home, Definitions and About/Contacts tabs</td>
<td>33</td>
</tr>
<tr>
<td>Table 8</td>
<td>Functions used in the Lists/Search Tool &amp; Interactive Plots tabsetPanel</td>
<td>33</td>
</tr>
<tr>
<td>Table 9</td>
<td>Functions used in the Plasmid Classification tab</td>
<td>34</td>
</tr>
<tr>
<td>Table 10</td>
<td>Comparison between the results obtained in this study and Shintani et al. (2015) and Orlek et al. (2017)</td>
<td>49</td>
</tr>
<tr>
<td>Table 11</td>
<td>Transferability outcomes for the results in common between this study and Shintani et al. (2015)</td>
<td>49</td>
</tr>
</tbody>
</table>
INTRODUCTION

1.1 MOTIVATION

Bacterial plasmids are genetic structures that have essential roles in bacterial adaptation (Heuer and Smalla, 2012). By associating with other genetic structures, like integrons and transposons, plasmids provide a platform for genetic recombination and for gene transfer between different bacterial species (Ochman et al., 2000; Heuer and Smalla, 2012). So far, over 10000 plasmid sequences – of which over 4000 are complete – are available in GenBank. They have diverse genetic structures, and can be classified according to their encoded accessory traits (like, for example, resistance and catabolic plasmids), mobility (conjugative, mobilizable and non-mobilizable), or incompatibility groups (from IncA to Z).

In recent years, there have been important efforts to analyze plasmid sequence data, essentially from clinical origin, due to their important involvement with epidemiology of multidrug resistant bacteria (Carattoli et al., 2014). However, plasmid importance should not be limited to bacterial hosts associated with hospital environments. Similarly with what has been proved for resistance genes (Gibson et al., 2014), plasmid encoded functions and genetic diversity are constrained by ecology (Binh et al., 2008). The number of plasmid sequences in the database has been growing (Shintani et al., 2015; Smillie et al., 2010), but only approximately 47% (our study, unpublished results) can be associated with clinical related environments\(^1\). The remaining plasmids, coming from other environmental compartments such as soil, water, food, extreme environments, and wastewater have been poorly compared and explored. A comprehensible and unified approach to catalogue non-clinical related plasmids is crucial to properly explore plasmid diversity and functions.

1.2 OBJECTIVES

The main goal of this thesis is the development of a repository of fully sequenced plasmids, gathering information about their ecology and the development of easy-to-use web tools for

---

\(^1\) Clinical, clinical potential and veterinary
the characterization of plasmid data regardless of their source environment and bacterial host.

In details, the scientific/technological objectives are to: (i) identify and review existing tools within the field of plasmid sequence analysis, (ii) review gene targets for plasmid classification/characterization by similarity searches, (iii) create a plasmid database with ecological information and replicon classifications and (iv) create a framework for the plasmid web tool resource. The latter will include a list tab with additional data regarding plasmid ecology and a search tool to explore it and allow the in silico detection for plasmid characterization (mobility, incompatibility and putative transferability).

1.3 ORGANIZATION OF THE THESIS

This study is divided into five main chapters, all containing one or more sections:

1) **Introduction**, including the motivation behind the chosen theme and the objectives to achieve;

2) **State of the Art**, comprising the main concepts involving plasmids, that is:
   - The background and definition of plasmid and its key features – replication, stability and partition and selective advantages for its host;
   - The concept of mobile genetic elements, including horizontal gene transfer, mobility and mating pair formation;
   - The types of plasmid data available at the NCBI database;
   - The existing tools for the in silico characterization of plasmid data (and other mobile genetic elements).

3) **Development**, which includes the stages of designing the created tools:
   - Creating the database with ecological information retrieved from NCBI and supplementing it with information from related publications and with the results from similarity searches in terms of incompatibility, mobility and putative transferability;
   - Description of the software used to carry out the similarity searches and develop the web tool and the illustration of features of the application.

4) **Results**, regarding the obtained classification by similarity searches and developed application;

5) **Conclusions and Future Work**, indicating the main conclusions drawn from this work and the next steps for the improvement of the developed tools.
STATE OF THE ART

2.1 PLASMID DEFINITION AND MAIN CHARACTERISTICS

The term plasmid was first used to describe extrachromosomal genetic elements by Lederberg in 1952 (Kado, 2015). Theretofore, designations such as R, F and T factors and episome were used.

Extrachromosomal genetic elements capable of conferring antibiotic resistance to their host were called R factors, those which were conjugative (see section 2.2) were identified as T factors and the ones carrying the genetic element needed for conjugation were known as F factors (Kado, 2015; Novick et al., 1976; Summers, 1996). Episome described a genetic element that can replicate into or independent of the host chromosome (that is, autonomously in the cytoplasm). This term was used to label all the known bacterial genetic elements (including the three factors just described), which was very confusing, since it was too restrictive and, instead of defining a taxonomic entity, it merged the definition of plasmids and phages, among other elements (Kado, 2015; Novick et al., 1976).

To avoid this dubiety, plasmid became the conventionally accepted designation, episome was discarded as its synonym and the R and F factors are now known as Resistance (R) and F plasmids (Kado, 2015; Novick et al., 1976). In addition to these definitions, two other denominations are also noteworthy: cryptic and Col plasmids, which denote a plasmid without any specific traits and a colicin-producing plasmid, respectively (Novick et al., 1976).

Presently, a plasmid is defined as a double-stranded deoxyribonucleic acid (DNA) extrachromosomal genetic entity, transferable between species and capable of self-replicating (Carattoli, 2009; Carattoli et al., 2014; Novick et al., 1976).

Plasmids, mostly found in the Bacteria domain, are typically circular but occasionally linear, their size can vary from one to thousands of base pairs and their GC content also suffers major fluctuations (Novick, 1987; del Solar et al., 1998).

One of their most important specific features, which differs according to each host, is the copy number, defined as the number of molecules/concentration of a plasmid per cell, which must be maintained at a constant level so that the plasmid can persevere. For this
to happen, depending on the number of copies the plasmid requires, different strategies are adopted (see subsection 2.1.1) (Novick et al., 1976; Novick, 1987; del Solar et al., 1998; Summers, 1996). Note that a plasmid can be identified as low-copy, medium-copy or high-copy, with numbers ranging from 1 to 10, 10 to 20 and 20 to over 700, respectively (Fricke et al., 2009).

Plasmids rely on two main functions to survive: persistence and proliferation. The former requires that the replication rate of the plasmid is the same as its host’s division rate, depending on the control of replication and, for low copy numbers, on the partition system, to guarantee that at least one copy is received by the daughter cell (see subsection 2.1.1). The latter, explained in section 2.2, is the capacity of a plasmid to transfer from one cell to another (Summers, 1996).

Although usually not being essential, these entities are systematically inherited due to being able to confer a selective advantage to the host, such as antibiotic resistance and virulence genes (see subsection 2.1.4) and tolerance to pollution, production of iron (to overcome its scarcity) and other environmental constraints, allowing them to spread to different surroundings (Carattoli et al., 2014; Kado, 2015).

Plasmids have previously been classified as parasitic, implying that they harm their host, or as symbionts, sharing a reciprocal exchange of advantages with the host (Norman et al., 2009; Summers, 1996). However, they are currently classified as selfish, due to the ability to autonomously replicate and maintain themselves in a cell, spreading within the genome or between genomes (see section 2.2) and being often acquired by sexual transmission. The degree of selfishness can vary, as they can benefit themselves while increasing or not the host’s fitness, since their persistence solely depends on their replication rate. These are also capable of influencing other plasmids, as they will try to prevent the entry of other plasmids in the cell (Kado, 1998; Liu et al., 2015; Norman et al., 2009).

Because of their unique capabilities, plasmids are acknowledged as of great interest from a biotechnological standpoint, as they have caused major breakthroughs in molecular biology. The goals of researchers include studying their structure and genes, identifying their replication and partition systems and understanding the phenotypes they confer (e.g. tumor-inducing, virulence and antibiotic, heavy metal and radiation resistance), in order to use them in multiple procedures, such as reporter systems, genetic engineering and the development of gene vectors (Kado, 2015).
2.1.1 Plasmid Replication

Plasmids are characterized by the presence of one or more specific regions encoding functions capable of activating and controlling replication, called replicons (Carattoli et al., 2014). Two types of replicons should be highlighted: (i) basic replicon, the smallest DNA sequence (usually 1-3kbp) that allows replication while preserving the regulatory scheme; and (ii) minimal replicon, the smallest DNA portion enabling replication, even if regulation is flawed (Lilly and Camps, 2015; Summers, 1996). Various plasmids have only one basic replicon, while others – multireplicon plasmids – display several replicons (for example, IncF plasmids), although often only one is active at a time. The reason for this difference is unknown, but it is inferred that having multiple replicons may enable plasmids to colonize numerous hosts by allowing them to avoid incompatibility with the resident plasmid (see section 2.3) (Summers, 1996).

Plasmid replication can be divided into three stages: initiation, elongation, and termination. The first stage, which is different for each replicon, is reliant on plasmid-encoded properties, namely the origin of replication (ori) and, generally, a replication initiation (Rep) protein, some of which are represented in Table 1 (del Solar et al., 1998; Kues and Stahl, 1989). In the cases where the Rep proteins exist, the origin of replication includes directly repeated DNA sequences with specific spacing, named iterons, which are the binding sites for the Rep proteins and are essential for replication control (Lilly and Camps, 2015; del Solar et al., 1998; Summers, 1996). Iteron spacing (iterons can be separated by other sequences or be contiguous) is of special importance, since it allows the recognition of specific DNA sequences by being at a distance that matches the helical periodicity of the DNA double helix (Lilly and Camps, 2015; del Solar et al., 1998). Because initiation is characteristic of each replicon, replication control takes place at that stage (del Solar et al., 1998).

<table>
<thead>
<tr>
<th>Initiator</th>
<th>Replication mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>RepA</td>
<td>Rolling circle; theta type</td>
</tr>
<tr>
<td>RepA1</td>
<td>Theta type</td>
</tr>
<tr>
<td>RepB</td>
<td>Rolling circle</td>
</tr>
<tr>
<td>RepC</td>
<td>Rolling circle; theta type</td>
</tr>
<tr>
<td>RepD</td>
<td>Rolling circle</td>
</tr>
<tr>
<td>RepE</td>
<td>Theta type</td>
</tr>
<tr>
<td>(\pi) (\textit{pir})</td>
<td>Theta type</td>
</tr>
<tr>
<td>TrfA</td>
<td>Theta type</td>
</tr>
</tbody>
</table>
2.1. Plasmid Definition and Main Characteristics

As mentioned above, replication initiates at a location known as the origin of vegetative replication (oriV), which is commonly classified in one of four different ways: (i) the minimal cis-acting region that can support autonomous replication of the plasmid, (ii) the region where DNA strands are melted to initiate the replication process, (iii) the position/base at which the first deoxyribose base is added to the leading-strand RNA primer and/or, more strictly, (iv) the portion of sequence that is targeted by replication initiation factors in trans (Lilly and Camps, 2015; Shintani et al., 2015; del Solar et al., 1998; Summers, 1996).

Albeit competent in self-replication, plasmids still have to synchronize their replication with the growth and division of their host cell, having to replicate a certain number of times per generation (usually one) to persist, since an uncontrolled plasmid concentration can have severe consequences for both the plasmid and its host (del Solar et al., 1998; Summers, 1996).

The number of plasmids per cell is controlled during the initiation stage of replication with the help of trans-activators/inhibitors, generally Rep replicases\(^1\), encoded by plasmids (Norman et al., 2009). For an excessive copy number, to avoid an increased metabolic burden to the host, jeopardizing its viability, the average replication for generation is kept at less than one by employing trans-activators. Inversely, for insufficient copy numbers, which can lead to consecutive loss until the cell is plasmid-free (see section 2.1.3), thus requiring closer regulation, the average replication for generation is maintained at more than one with the aid of trans-inhibitors (Kues and Stahl, 1989; Norman et al., 2009; Summers, 1996).

There are three main types of plasmid replication: rolling circle, strand displacement and theta type, all of which observed in both Gram Positive and Gram Negative bacteria and explained in the next subsections.

**Rolling Circle Replication**

Rolling circle replication (RCR) has only been found on plasmids with less than 10kbp and it is prototyped by the pT181, pC194, pMV158 and pUB110 plasmids, isolated from organisms of the *Staphylococcus* genus (Shintani et al., 2015; del Solar et al., 1998). This method has a unique initiation mechanism: it is initiated from a 3’-OH primer produced by nicking one strand of the plasmid, which allows the host DNA polymerases to initiate the leading strand replication (Norman et al., 2009; Ruiz-Maso et al., 2015). Here, in order to cleave (split) and join plasmid DNA, Rep proteins have transferase\(^2\) enzymatic activity (del Solar et al., 1998).

---

1 Class of enzymes that catalyze the synthesis of an RNA molecule.
2 Class of enzymes capable of transferring functional groups for a donor to another molecule (acceptor).
2.1. Plasmid Definition and Main Characteristics

Strand Displacement Replication

The name strand displacement is due to, in this process, the synthesis of each one of the strands taking place continuously and causing the displacement of the complementary strand. This type of replication is characteristic of the broad host IncQ family, namely of their archetype plasmid, RSF1010 (Shintani et al., 2015; del Solar et al., 1998). In this case, replication has two origins, *ssiA* and *ssiB*, one per DNA strand, and initiation is promoted by three Rep encoded proteins: (i) *RepA*, which has 5′ → 3′ helicase activity, (ii) *RepB*, which catalyzes the starting point for DNA synthesis, and (iii) *RepC*, which recognizes iterons, together allowing bidirectional replication to occur (del Solar et al., 1998).

Theta Type Replication

During this type of replication, which can happen uni- or bidirectionally, the images obtained by electronic microscopy (EM) resemble the greek letter θ (Lilly and Camps, 2015; Norman et al., 2009; del Solar et al., 1998). Theta type replication has been mostly studied in gram negative bacteria, although cases of theta plasmids being isolated from gram positive bacteria have also been reported (Norman et al., 2009; del Solar et al., 1998). In this method, similarly to the process in chromosomal replication, the leading strand, synthesized continuously, and lagging strand, synthesized discontinuously, are replicated coordinately (Lilly and Camps, 2015; del Solar et al., 1998). This method can start from one to multiple origins and it includes synthesis of a primer RNA (pRNA), initiation of DNA synthesis by covalent extension to the pRNA and, generally, a plasmid-encoded Rep protein (del Solar et al., 1998).

There are four classes of theta type replication, from A to D. Plasmids from class A include iterons (differing for each plasmid), that is, they rely on Rep proteins for replication initiation (see subsection 2.1.1), namely RepA for R1, pSC101, pPS10 and P1 (even though the Rep proteins of these three plasmids are homonymous, their characteristics are different), Trf1 for RK1 and π for R6K (plasmids of reference) (Lilly and Camps, 2015). Class B includes the archetype plasmid CoIE1 and CoIE1-like plasmids, which do not require Rep proteins, but rely on their host factors, such as DNA Polymerase I, for replication initiation and primer synthesis and only rely on antisense RNA for replication control (Lilly and Camps, 2015; Shintani et al., 2015; del Solar et al., 1998). Finally, classes C, containing CoIE2 and CoIE3 plasmids, and D – with large, low-copy streptococcal plasmids that replicate in a broad range of Gram-positive bacteria, such as pAMβ1, pIP501 and pSM19035 – have similarities with both A and B classes (del Solar et al., 1998).

---

3 Class of enzymes that unwind and separate the two strands of DNA.
2.1.2 Replication Control

The goal of replication control is to maintain the plasmid copy number at a steady state, which must be in accordance with their host (Kues and Stahl, 1989; del Solar et al., 1998; Summers, 1996). In order to define and keep a specific copy number, plasmids employ negative regulatory circuits, which should have a virtually imperceptible concentration upon host colonization; after that point, avoiding fluctuations in plasmid concentration, as mentioned in subsection 2.1.1, involves adjusting the replication rate (del Solar et al., 1998; Summers, 1996). This regulation is the reason for plasmid incompatibility (see section 2.3), causing plasmids with similar regulation mechanisms to segregate within the host population (del Solar et al., 1998).

Classification of control systems is usually made according to the nature of their main replication inhibitor, which can be a protein (λ − dv, not described in natural plasmids), a small antisense RNA (ColE1, pT181 and R1), both a protein and an antisense RNA (pMV158 and pIP501), or a set of short DNA repeats, that is, DNA sites for binding initiator proteins (P1, F, RK2 and R6K) (del Solar et al., 1998; Summers, 1996).

The control by antisense RNA, found in both low and high copy plasmids, consists of a short antisense inhibitor binding close to the 5’ end of a region (longer transcript) that plays a major role in the replication process and precludes its activity, either by binding to the mRNA of a Rep protein in R1 or pT181 plasmids (prohibiting translation) or by blocking the formation of an RNA II-DNA hybrid (pre-primer RNA for replication) at the origin of replication in the ColE1 plasmid; in both cases, these inhibitors are called countertranscribed RNAs (ctRNAs) (del Solar et al., 1998; Summers, 1996).

Control by both a protein (transcriptional repressor) and antisense RNA is typical for the R1, pIP501 and pMV158 plasmids. For the latter, a repressor protein, named CopG is binded and inhibits transcription for a single promoter for the copG and repB genes and a small ctRNA is also involved and, for both these elements, mutations and deletions can increase the plasmid copy number (del Solar et al., 1998).

Lastly, besides posing as binding sites for the Rep proteins, iterons from the theta type replication also play a role in replication control when they are located at a site that is not the origin of replication. In this type of control, it is postulated that, because plasmids are able to detect the concentration of Rep proteins (titration), the frequency of replication initiation is limited. An alternative hypothesis, which considers the iteron concentration in control of replication rate, is that, when Rep proteins bind to the iteron located at the origin of replication, initiation only takes place if the copy number is low. As it increases, Rep proteins start interacting with each other and, as a consequence of cell growth, the intended plasmid concentration per cell is achieved (del Solar et al., 1998).
2.1.3 Plasmid Stability and Partition

Generally, the term plasmid stability is applied from a segregational perspective (Friehs, 2004). As highlighted throughout subsection 2.1.1, plasmids must be kept at a specific number per host cell. For an effective control mechanism, that is, a mechanism that inhibits replication for plasmids with an excessive copy number and stimulates it for plasmids with less than the average copy number, the only factor in maintaining stability is guaranteeing that there is, at least, one copy, i.e., an average of one, of the plasmid per cell (Novick et al., 1976; Summers and Sherratt, 1985; Summers, 1996). However, instability is not only caused by faulty replication, but also, for example, by damaging insertions of foreign DNA, enzymatic degradation and homologous recombination that leads to the formation of plasmid multimers (Norman et al., 2009), forcing the plasmid to resort to different strategies to persevere.

Plasmids usually impose a metabolic burden on their host. Hence, if the plasmid does not encode a specific characteristic that the host is lacking, a cured, that is, a plasmid-free lineage will be privileged by the host, which means that, if no selective pressure is applied, all the plasmid copies in that cell will be gradually eliminated (Friehs, 2004; Norman et al., 2009).

These entities can attain stability by various strategies: multimer resolution, active partitioning, plasmid addiction and, less frequently, restriction-modification (RM) systems (Baxter and Funnell, 201; Norman et al., 2009; Summers and Sherratt, 1985; Summers, 1996).

Plasmids can be distributed by random or active partitioning. Random partition can only be used by high copy plasmids, because it relies on random diffusion and will eventually cause the generation of a plasmid-free cell with a speed inversely proportional to the number of copies per cell, thus being highly dangerous for low-copy plasmids. Plasmids with a low average concentration per cell must be distributed actively, normally by a nucleoprotein complex, named segrosome, moving plasmids to a specific position before cell division, resembling the eukaryotic mitotic division (Baxter and Funnell, 201; Norman et al., 2009; Summers and Sherratt, 1985; Summers, 1996).

Multimer resolution systems (mrs), well studied in the ColE1 plasmid (Xer-cer system) consist of site-specific recombinase systems formed to avoid the damage caused by plasmid multimer formation, capable of negatively affect segregation (Norman et al., 2009).

Stability can also be accomplished by plasmid addiction, which, for competition purposes, eliminates plasmid-free lineages by post-segregational killing (PSK) or addiction systems. Customarily, this mechanism, denominated toxin-antitoxin (TA), encompasses two genes with products with different effects: the first can be responsible for growth limitation or death of the host cell, whereas the second softens these effects. Hence, stability can only be attained if the two genes are simultaneously present. One example of this method can be
observed in the RK2 plasmid, where host replication is inhibited by a toxin that suppresses gyrase unless it is coupled with its complementary antitoxin protein (Norman et al., 2009).

Finally, restriction modification (RM) systems, although typically composed of selfish elements, can confer stability to the DNA molecule in which they are inserted (Norman et al., 2009).

2.1.4 Antibiotic Resistance and Virulence

As the number of multiresistant organisms arises at a concerning pace, the comprehension of plasmid structure and functions becomes of even greater importance, not only in the medical and veterinary fields, but also from a technological point of view.

Reports of drug resistance trace back to 1907 when Trypanosoma brucei was first registered as resistant to the effect of para-rosaniline and 1912, when Streptococcus pneumoniae started resisting to ethyldihydrocupreine hydrochloride (Summers, 1996). However, multiple antibiotic resistance was first linked to plasmids in Japan in the early 1950s after an outbreak of dysentery, caused by Shigella dysenteriae (Ramirez et al., 2015; Summers, 1996).

Although multiple species of bacteria may endanger public health, the currently most threatening are those represented in the ESKAPE acronym, the culprits of most hospital infections by their persistence and resistance: Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa and Enterobacter (Ramirez et al., 2015).

Antibiotic resistance is, normally, more related with the plasmid than with the host chromosome (Summers, 1996). Plasmids may confer resistance to a wide range of antibiotics, including, but not limited to: tetracycline, phenicol, oxazolidinone, macrolide-lincosamide-streptogramin, aminoglycoside-aminocyclitol-streptothricin, pleuromutilin, mupirocin, fusidic acid, trimethoprim, vancomycin and peniciline (Ramirez et al., 2015).

Virulence can be defined as the ability to cause disease. This, however, is evidence of poor adaptation to the host, since a plasmid that establishes a long, stable relationship of parasitism limits its violent reactions, thus reducing the metabolic burden of the host. It is believed that plasmid-encoded virulence is involved in processes such as adhesion, toxinogenesis and serum resistance (Summers, 1996).

Virulence can be observed in both spore-forming and non spore-forming bacteria, being implicated in several illnesses, such as enteric and gastroenteritis syndromes, tetanus, botulism and gas gangrene (Adams et al., 2015).
2.2 MOBILE GENETIC ELEMENTS

Mobile Genetic Elements (MGE) are identified as DNA sequences, with size ranging from one to thousands of base pairs, capable of recombining and transferring within the host cell or between genomes (Leplae et al., 2004). There are several types of MGE, some of which illustrated in Figure 1, such as plasmids, bacteriophages, viruses, genomic/pathogenicity islands and integron-associated gene cassettes and transposable elements (TE), including insertion sequences (IS) and transposons (Tn) (Leplae et al., 2004; Piotrowska and Popowska, 2015; Norman et al., 2009; Smillie et al., 2010). When collectively considered in the genome, these elements are called the mobilome (Piotrowska and Popowska, 2015).

An insertion sequence (IS) is the most basic transposable element, consisting only of short sequences of inverted repeats with a transposase gene in between them. Transposons are more complex, containing genes for the expression of different genotypes in addition to the transposase (Piotrowska and Popowska, 2015).

Genomic islands are chimeric genes or entire groups of genes (Leplae et al., 2004). Integrons are genomic elements incapable of replication, often associated with plasmids and transposons (becoming mobile integrons, since these entities stimulate their propagation), capable of capturing and expressing gene cassettes. These elements, thought to be the primary way of obtaining antibiotic resistance, are composed of a gene coding for an integrase (intI), a specific recombination site (attI) and one or two promoters (Pc) controlling the expression of the captured gene (Moura et al., 2009; Piotrowska and Popowska, 2015).

The physical deslocation of DNA requires the intervention of four main enzymes, namely recombinases, which allow homologous recombination, transposases, enabling the movement and insertion of transposons, integrases, facilitating the insertion of elements into integrons and resolvases, resolving Holiday junctions resulting from recombination (Norman et al., 2009).

Figure 1.: Depiction of the interaction between some Mobile Genetic Elements, showing the elements necessary for a gene cassette to be acquired by a plasmid (Retrieved from Norman et al. (2009))
In order to identify their role and interactions in specific environmental niches, various prokaryotic-specific designations should be taken into consideration. On one hand, regarding their frequency, core genome describes the set of genes omnipresent in all strains of a given species, dispensable/flexible genome respects genes present in some strains of a species, but not in all of them, and pan genome is the sum of the previous two. However, since all of these terms disregarded interaction of genetic elements between species, a new denomination, supergenome, was created by Norman et al. (2009), defined as the total pool of genes readily available to a prokaryotic organism within a specific environment. On the other hand, concerning their availability, private pool identifies genes that are encoded by the host chromosome (available only for the resident plasmid) and communal pool characterizes genes produced by mobile genetic elements (available to most prokaryotes) (Norman et al., 2009).

2.2.1 Horizontal Gene Transfer

Bacterial genomes can obtain new genes by horizontal gene transfer (HGT), also known as bacterial sex, allowing genetic variation. Horizontal gene transfer occurs either by introduction via MGE or by direct uptake and incorporation of DNA by recombination (Norman et al., 2009). This mechanism can lead to evolution and it is important from microbiology, ecology and pathogenically transfer perspectives, since it ensures plasmid persistence and it is responsible for the acquisition of genes encoding antibiotic resistance and virulence (Norman et al., 2009; Smillie et al., 2010).

The effectiveness and selective advantages of horizontal gene transfer depend on a multiplicity of factors, such as the nature of the gene, their host restrictions and their co-inhabitant genetic elements (Norman et al., 2009; Smillie et al., 2010). Informational genes, such as those involved in transcription, translation and related processes usually belong to large molecular complexes, have a smaller probability of having been successfully transferred horizontally than operational genes, that is, genes involved in amino acid biosynthesis and other housekeeping functions, which are generally compact operons (Norman et al., 2009).

There are two sets of genes involved in plasmid propagation: (i) the set of mobility (MOB) genes, allowing conjugative DNA processing, which is indispensable; and (ii) the mating pair formation (Mpf) complex (membrane-associated), which is a form of a type 4 secretion system (T4SS), providing the mating channel (Smillie et al., 2010). Plasmids can be classified, according to the incidence of these sets, as conjugative or self-transmissible, if the MOB set is present and they encode their own Mpf genes, as mobilizable, if they have the MOB set but rely on other genetic elements to obtain the Mpf genes, and as nonmobilizable if they fit in neither of those categories (Garcillan-Barcia et al., 2009; Norman et al., 2009;
Conjugative plasmids, generally with a size over 30kb, are maintained at a low copy number per cell (typically less than ten copies per cell), while mobilizable plasmids, which are of a considerably smaller size (<15kb), can reach hundreds of copies per cell (Garcillan-Barcia et al., 2009; Norman et al., 2009).

Horizontal transfer can proceed by one of three manners: transformation, transduction and conjugation.

Conjugation can be defined as the combination of two functions, mating pair formation (Mpf) and RCR (rolling-circle replication), and it is composed of several steps. The first step is the formation (Mpf) of a channel connecting the donor to the recipient, generally through the synthesis of a type IV secretion system (T4SS), where the coupling is made by filaments named conjugation pili (the morphology of these filaments determines the conjugation host range and the medium of transfer). The second step is the formation of the relaxosome, which contains the single strand plasmid DNA, the relaxase (see subsection 2.2.2) and some other proteins. The final step is rolling-circle replication to synthesize a second strand in the donor and in the recipient (Norman et al., 2009). Conjugative plasmids are also capable of replicating by the theta-type mechanism during vegetative growth (Carattoli, 2009).

Transformation and transduction are likely to favor small plasmids and have a lower impact on the plasmid transfer rate, since they promote mainly intraspecies transfers, whereas conjugation can link remotely distant organisms (Norman et al., 2009; Smillie et al., 2010).

Transformation is the uptake of DNA into cells from the surrounding environment, depending on the existence of plasmid and/or chromosomal DNA fragments in the environment, requiring only a recipient and a mechanism for the insertion of DNA (Norman et al., 2009; Smillie et al., 2010).

Transduction is the transportation of DNA through bacteriophages, dependent on a phage replicating within the donor and, during DNA packaging, DNA fragments can be incorporated into the phage capsid (Norman et al., 2009; Smillie et al., 2010). This method is limited by the size of the phage (plasmids must have a genome with a size under or equal to it) (Smillie et al., 2010).
2.2. Mobile Genetic Elements

2.2.2 Plasmid Mobility Groups and Mating Pair Formation

The encoded transfer genes vary among different types of plasmids: while conjugative plasmids display the complete set of transfer genes, that is, MOB and a T4SS, mobilizable plasmids only include a simple MOB region, allowing them to be transported by other conjugative plasmids and merely consist of a vegetative origin of replication oriV, a relaxase protein and one or more nicking-accessory proteins. The MOB machinery only requires the presence of a relaxase, the protein responsible for the initiation and termination of conjugative DNA processing. These proteins are of small size and contain two or more protein domains, always including a domain at the N-terminus of the protein and at the C-terminus, with the latter generally including a DNA helicase and ligase or other domains. With some exceptions, replicases also contain a 3H (histidine triad) motif, which they use to bind divalent cations (Garcillan-Barcia et al., 2009). Mobilizable plasmids carry only the relaxosomal components (see subsection 2.2.1). Inversely, conjugative plasmids additionally include the type IV coupling protein (T4CP) – used to bind the relaxosome and the transport channel – and the components of the mating channel needed for T4SSs (Smillie et al., 2010).

Plasmids can be assigned into one of six MOB groups, namely MOB_C, MOB_F, MOB_H, MOB_P, MOB_Q, and MOB_V, depending on their amino acid sequence (Garcillan-Barcia et al., 2009; Shintani et al., 2015). This classification was able to organize each relaxase into a specific protein family, that is, a set of proteins sharing a biological function and similar in their sequence, for the exception of the MOB_P group, still considered unfinished (Garcillan-Barcia et al., 2009; Smillie et al., 2010). Members of MOB_F and MOB_H are mostly found in large conjugative plasmids, while the other families are predominantly encountered in small mobilizable plasmids (Garcillan-Barcia et al., 2009).

Recently, Smillie et al. (2010) classified T4SS bacteria into four mating pair formation (Mpf) groups depending on their protein homology, and each one was named according to the archetype T4SS of the group: vir system for MPF_T, F for MPF_F, R64 for MPF_I and ICEHIN1056 for MPF_G.

As a concluding remark, it should be noted that these classification systems can be used to identify the mobilization potential of a plasmid, since a specific behavior accompanied by the existence or absence of certain elements (like the presence of a Mpf group in conjugative plasmids), the similarity with known relaxases and the size of the plasmid (large for nonmobilizable, intermediate for conjugative and small for mobilizable) should lead to the recognition of its transfer system (Garcillan-Barcia et al., 2009; Smillie et al., 2010).
2.3 Plasmid Incompatibility

The term *incompatibility* was first used in the early 1960s for the F plasmid and, one decade later (1976), a formal classification scheme based on incompatibility was proposed by Novick (Carattoli, 2009; Couturier et al., 1988; Novick et al., 1976). Despite being a well-accepted plasmid classification scheme, there are still some limitations, namely if the plasmid is not transmissible or it does not contain a suitable marker gene. These difficulties are surpassed by using a set of reference miniplasmids belonging different Inc groups and having a gene for galactose utilization (Couturier et al., 1988).

Plasmid incompatibility is defined as the impossibility of two plasmids with the same replication origin (*ori*) and partition system to coexist stably in the same cell line or host without external selection (Carattoli, 2009; Novick et al., 1976; Novick, 1987; Shintani et al., 2015), that is, two plasmids are considered incompatible if the introduction of a second plasmid leads to the elimination of the first (resident). Hence, incompatibility can be perceived as an evidence of relatedness between the plasmids that share the same replication controls (Couturier et al., 1988).

There are two types of incompatibility between two plasmids: *symmetric* and *vectorial*. The former implies that the possibility of losing any of the plasmids is equiprobable, whereas the latter postulates that one of the plasmids has a higher probability of being excluded than the other. These are, usually, used in different contexts: symmetric incompatibility is commonly applicable to co-existing single replicons with similar replication and maintenance functions, while vectorial incompatibility is often associated with problems in replication (Novick, 1987).

In order to test and classify incompatibilities, the conventional method is to introduce, either by conjugation, transduction or transformation (see section 2.2), a plasmid into a strain which already carries another plasmid. If the resident plasmid is excluded, then the two plasmids belong to the same Inc group (Couturier et al., 1988). However, due to the fact that classification is based on the amino acid sequence of the replicon initiation (rep) protein (replicon typing), testing is not always necessary, since the same incompatibility group will share the same rep protein and should be assigned the same name (Shintani et al., 2015). Note that the assignment of incompatibility should take into account that the plasmids within the same host are able to mutually interact *in vivo* (Couturier et al., 1988).

Several factors influence plasmid incompatibility. The first is the origin of replication (*ori*), since two plasmids that rely on the same replication strategy cannot reside on the same cell. Another constraint is competition for replication factors, with the plasmids with selective advantages such as antibiotic resistance, less toxicity and faster replication (which should always be inversely proportional to the copy number) having a higher probability
of succeeding. The last factor is the copy number. On one hand, when two plasmids are compatible, each one produces a replication inhibitor that does not affect the other’s replication, maintaining their normal copy numbers and both are able to persist. On the other hand, if incompatible plasmids produce the inhibitor, the host cell cannot differentiate between the origins of replication and the replication of both will be affected: the total number of plasmid copies in the cell will be less than the sum of the total number of plasmid copy numbers (this happens because each plasmid adjusts its replication rate and copy number to the total inhibitor concentration). As such, replication will only be resumed when the pre-replication copy number is restored, that is, when one of the plasmids is no longer in the cell (Summers, 1996; Velappan et al., 2007).

The assignment of a plasmid to a certain incompatibility group determines its hosts and interactions and possible opportunities (e.g. acquisition of new traits) (Izmalkova et al., 1987). Most of the presently known incompatibility groups are represented in Tables 2 and 3. This information was adapted, not only from the references listed in Table 3, but also from Carattoli (2009); Clark et al. (2016); Garcillan-Barcia et al. (2009); Norman et al. (2009) and Shintani et al. (2015).
2.3. Plasmid Incompatibility

Table 2: Mobilization, occurrence, host range and general characteristics of known plasmid Incompatibility groups. Note that the members of the Resistance Plasmid Families in Enterobacteriaceae are marked as ***. The references on which this Table is based are listed in Table 3.

<table>
<thead>
<tr>
<th>Inc Group</th>
<th>Mobilization</th>
<th>Occurrence</th>
<th>Host Range</th>
<th>General Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>IncA/C</td>
<td>conjugative</td>
<td>Ex. Escherichia, Salmonella, Vibrio, Yersinia; ***</td>
<td>broad host range</td>
<td>large plasmids (144kb); transferable antimicrobial resistance to sulfonamides and tetracyclines; small size;</td>
</tr>
<tr>
<td>IncB/O</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>important role in the spread of multiple antibiotic resistance determinants</td>
</tr>
<tr>
<td>IncC, Inc6</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncD</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncE</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncF</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncG</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncH</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncI</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>plasmids conferring antibiotic resistance (producing type F pilil)</td>
</tr>
<tr>
<td>IncJ</td>
<td>conjugative</td>
<td>Present in clinical and environmental pathogens from disparate geographical locations</td>
<td>broad host range</td>
<td>large size; plasmids encoding antibiotic and heavy metal resistance genes</td>
</tr>
<tr>
<td>IncK</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>small size; confers antibiotic resistance medium size</td>
</tr>
<tr>
<td>IncL</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncL/M</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncM</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncP-1α</td>
<td>conjugative</td>
<td>Exogenous/Isolated from Pseudomonas aeruginosa</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncP-1β</td>
<td>conjugative</td>
<td>Exogenous/ Found in numerous species</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncP-1γ</td>
<td>conjugative</td>
<td>Exogenous (triparental)/ found in Burkholderia cepacia and Achromobacter xylosoxidans subsp. Denitrificans EST4002</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncP-1ε</td>
<td>conjugative</td>
<td>Exogenous</td>
<td>broad host range</td>
<td>may confer antibiotic and heavy metal resistance and virulence</td>
</tr>
<tr>
<td>IncP-7</td>
<td>conjugative</td>
<td>main natural host: Pseudomonas fluorescens</td>
<td>broad host range</td>
<td>large size; plasmids which control the biodegradation of naphthalene and salicylate small plasmids; plasmids conferring resistance to streptomycin and quinolone</td>
</tr>
<tr>
<td>IncQ</td>
<td>mobilizable</td>
<td>found in many bacterial species</td>
<td>broad host range</td>
<td>large (217kb) size</td>
</tr>
<tr>
<td>IncS</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>medium to large size; may confer quinolone resistance</td>
</tr>
<tr>
<td>IncT</td>
<td>conjugative</td>
<td>isolated from multiple strains of Escherichia coli/ Aeromonas spp.</td>
<td>broad host range</td>
<td>large size</td>
</tr>
<tr>
<td>IncU</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>code for multiple-drug-resistant efflux pumps as well as the formation of fimbriae plasmids conferring antibiotic resistance</td>
</tr>
<tr>
<td>IncV</td>
<td>conjugative</td>
<td>***</td>
<td>narrow host range</td>
<td>small plasmids; plasmids conferring resistance to streptomycin and quinolone</td>
</tr>
<tr>
<td>IncW</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>large size</td>
</tr>
<tr>
<td>IncX</td>
<td>mobilizable</td>
<td>***</td>
<td>narrow host range</td>
<td>small plasmids; plasmids conferring resistance to streptomycin and quinolone</td>
</tr>
<tr>
<td>IncY</td>
<td>conjugative</td>
<td>***</td>
<td>broad host range</td>
<td>large size</td>
</tr>
</tbody>
</table>

References on which this Table is based are listed in Table 3.
### 2.3. Plasmid Incompatibility

<table>
<thead>
<tr>
<th>Inc Group</th>
<th>Archetype plasmid</th>
<th>GenBank Number</th>
<th>Amplicon size</th>
<th>Replicon typing</th>
<th>Target gene region</th>
<th>Bibliography</th>
</tr>
</thead>
<tbody>
<tr>
<td>IncA/C</td>
<td>pRA1</td>
<td>FJ709807</td>
<td>465bp</td>
<td>repA region</td>
<td>repA</td>
<td>(Harmer and Hall, 2016; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncB/O</td>
<td>pMU720</td>
<td>M28718</td>
<td>159bp</td>
<td>RNAI (miscRNA)</td>
<td></td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncC, Inc6</td>
<td>R40a, R55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Harmer and Hall, 2016; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncFIA</td>
<td>F, R386</td>
<td>M11322</td>
<td>462bp</td>
<td>traA</td>
<td>traA</td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncFIB</td>
<td>unnamed</td>
<td>M26308</td>
<td>702bp</td>
<td>repA</td>
<td>repA</td>
<td>(Carattoli, 2005)</td>
</tr>
<tr>
<td>IncFIC</td>
<td></td>
<td></td>
<td>262bp</td>
<td>repA2</td>
<td>repA2</td>
<td>(Carattoli, 2005)</td>
</tr>
<tr>
<td>IncFII</td>
<td>R100, R1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncFIII</td>
<td>ColB-K98</td>
<td></td>
<td></td>
<td>traT region</td>
<td>traT region</td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncFIV</td>
<td>R124</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncFV</td>
<td>Folac</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncHI1</td>
<td>R27 (&lt;TP117)</td>
<td>AF250878</td>
<td>471bp</td>
<td>parA-parB</td>
<td>parA-parB</td>
<td>(Carattoli, 2005; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncHI2</td>
<td>R478</td>
<td>BX684015</td>
<td>644bp</td>
<td>iterons</td>
<td></td>
<td>(Carattoli, 2005; Gilmour et al., 2004)</td>
</tr>
<tr>
<td>IncI1, IncIa</td>
<td>ColB-P9, A, R144</td>
<td>MA0137</td>
<td>139bp</td>
<td>RNAI (miscRNA)</td>
<td></td>
<td>(Carattoli, 2005; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncI2a</td>
<td>R721</td>
<td>NC002525</td>
<td>repA (plasmid replication protein RepA)</td>
<td></td>
<td>(Novick et al., 1976)</td>
<td></td>
</tr>
<tr>
<td>IncI7</td>
<td>R621</td>
<td>NC015965</td>
<td>repA (plasmid replication protein RepA)</td>
<td></td>
<td>(Novick et al., 1976)</td>
<td></td>
</tr>
<tr>
<td>IncI</td>
<td>R931</td>
<td>U13633</td>
<td>160bp</td>
<td>RNAI (miscRNA)</td>
<td></td>
<td>(Carattoli, 2005; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncK</td>
<td>M93063</td>
<td></td>
<td>160bp</td>
<td>RNAI (miscRNA)</td>
<td></td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncL</td>
<td>R471a</td>
<td>AO027768</td>
<td>785bp</td>
<td>repA,B,C</td>
<td>repA,B,C</td>
<td>(Carattoli, 2005)</td>
</tr>
<tr>
<td>IncN, Inc2</td>
<td>N3, R15</td>
<td>NC003292</td>
<td>559bp</td>
<td>repA (plasmid replication protein RepA)</td>
<td></td>
<td>(Carattoli, 2005; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncP-1a</td>
<td>RP4</td>
<td>AAA26427</td>
<td>281bp</td>
<td>trfA region</td>
<td></td>
<td>(Bahl et al., 2009; Daletry et al., 2014; Popowska and Krawczyk-Balska, 2013; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncP-1b</td>
<td>R751</td>
<td>U67194</td>
<td>282bp</td>
<td>trfA region</td>
<td></td>
<td>(Bahl et al., 2009; Daletry et al., 2014; Popowska and Krawczyk-Balska, 2013)</td>
</tr>
<tr>
<td>IncP-1c</td>
<td>pQKH54</td>
<td>AM137767</td>
<td>283bp</td>
<td>trfA region</td>
<td></td>
<td>(Bahl et al., 2009; Daletry et al., 2014; Popowska and Krawczyk-Balska, 2013)</td>
</tr>
<tr>
<td>IncP-1d</td>
<td>pEST4011</td>
<td>AY540995</td>
<td>284bp</td>
<td>trfA region</td>
<td></td>
<td>(Bahl et al., 2009; Daletry et al., 2014; Popowska and Krawczyk-Balska, 2013)</td>
</tr>
<tr>
<td>IncP-1e</td>
<td>p3-408</td>
<td></td>
<td>285bp</td>
<td>trfA region</td>
<td></td>
<td>(Bahl et al., 2009; Daletry et al., 2014; Popowska and Krawczyk-Balska, 2013)</td>
</tr>
<tr>
<td>IncP-1f</td>
<td>pCAR1</td>
<td>AB088420</td>
<td>524bp</td>
<td>rep region</td>
<td></td>
<td>(Deletry et al., 2014; Izmalkova et al., 1987)</td>
</tr>
<tr>
<td>IncQ</td>
<td>RSF1010</td>
<td>NC001740</td>
<td>703bp</td>
<td>strA and strB (streptomycin resistance - streptomycin phosphotransferase A and B)</td>
<td></td>
<td>(Gott et al., 1996; Pezzella et al., 2004; Novick et al., 1976)</td>
</tr>
<tr>
<td>IncS</td>
<td>R478</td>
<td>BX664015</td>
<td>750bp</td>
<td>repA</td>
<td>repA</td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncT</td>
<td>Rts1</td>
<td>K00533</td>
<td>750bp</td>
<td>repA (plasmid replication protein RepA)</td>
<td></td>
<td>(Novick et al., 1976)</td>
</tr>
<tr>
<td>IncW</td>
<td>R388</td>
<td>U12441</td>
<td>242bp</td>
<td>repA (plasmid replication protein RepA)</td>
<td></td>
<td>(Carattoli, 2005)</td>
</tr>
<tr>
<td>IncX</td>
<td>R6K</td>
<td>Y00768</td>
<td>376bp</td>
<td>ori y (protein coding)</td>
<td></td>
<td>(Garcillan-Barcia et al., 2009; Carattoli, 2005)</td>
</tr>
<tr>
<td>IncY</td>
<td>K02380</td>
<td></td>
<td>765bp</td>
<td>repA (plasmid replication protein RepA)</td>
<td></td>
<td>(Carattoli, 2005)</td>
</tr>
</tbody>
</table>

**Table 3:** Archetype plasmid, GenBank number, replicon typing (amplicon size and target gene region) and bibliography of known plasmid Incompatibility groups.
2.4 Plasmids at the NCBI Database

In January of 2017, there were 8935 plasmid reference sequences (RefSeq) available in the NCBI Plasmid Genome Database (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid/), from which 4498 are of high importance, since they are complete (closed) sequences (see Table 4) (O’Leary et al., 2015).

The number of complete sequences, however, has suffered considerable fluctuations throughout the years, with many entries added and some removed. Between January of 2009 and August of 2014, this number has exponentially grown from 1600-1700 (Smillie et al., 2010; Norman et al., 2009) to 4602 (Shintani et al., 2015). Finally, since March of 2016 (4514 sequences, our study), 16 sequences have been uploaded to the database and 32 have been excluded due to missing RNA genes, deriving from environmental sources, length of the genome being too large or too small, having many frameshifted proteins, being from a mixed culture or reasons not stated (Clark et al., 2016).

Table 4: DNA prefixes found in the Plasmid Genome Database (Adapted from (O’Leary et al., 2015))

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Application</th>
<th>Number in the Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_</td>
<td>Chromosomes and linkage groups</td>
<td>4498</td>
</tr>
<tr>
<td>NG_</td>
<td>Genomic regions</td>
<td>688</td>
</tr>
<tr>
<td>NT_</td>
<td>Scaffolds</td>
<td>37</td>
</tr>
<tr>
<td>NZ_</td>
<td>Chromosomes and scaffolds</td>
<td>3712</td>
</tr>
</tbody>
</table>

The fully sequenced plasmids present in the database are quite discrepant in terms of type, lineage, size and GC content. From these, 139 are from Archaea, 4312 from Bacteria, 38 from Eukaryota, 8 from viruses and, finally, one last sequence from "other sequences plasmids". From the identified phyla, the most abundant are Proteobacteria (2110 hits, 46.92% of the sequences), Firmicutes (1100 hits, 24.46%) and Spirochaetes (413, 9.18%), with numbers much higher than the remaining cases (see Figure 2). The majority of the plasmids is circular (4028), their GC content ranges from, approximately, 0.1926 (pMR2 plasmid in Moniliophthora roreri) to 0.7562 (pFQ12 in Frankia sp. Cpl1), with an average of 0.4409 and their size has a mean of 29253.94 base pairs, varying from 21 bp (pSE11-1 in Escherichia coli SE11) to 99806 bp (pTpa001 in Tsukamurella paurometabola DSM 20162).

Still, not all the information is reachable. Only under 17% of the entries have data about the isolation source (e.g. soil, food, veterinary, human) of the plasmid and 20% about their geography, over 88% of their hosts are unidentified, 92.9% of the collection dates are unknown and merely 1.6% of their sources (environmental or clinical) are identified, which forces the user to search for this information in the respective publications (which are still
2.5. Available Databases and Programs for Mobile Genetic Elements

not a guarantee of finding the intended data). Hence, one of the focuses of this thesis is to complete the missing information (regarding bacterial plasmids), so that it becomes easily accessible to any individual.

Figure 2.: Lineage distribution of the complete plasmid sequences in the NCBI Plasmid Genome Database. Note: the group “Others” includes “bacterial environmental samples” and several phyla, such as Acidobacteria, Amoebozoa, Aquificae, Chlorobi, Chloroflexi, Deferribacteres, Elusimicrobia, Nitrospira, Planctomycetes, Rhodophyta, Synergistetes, Thermotogae and Viridiplantae.

2.5 AVAILABLE DATABASES AND PROGRAMS FOR MOBILE GENETIC ELEMENTS

To keep up with the fast-paced growing number of mobile genetic elements, several databases and web tools were created to filter and analyze important data (Carattoli et al., 2014; Curry et al., 2016; Leplae et al., 2004; Moura et al., 2009; Siguier et al., 2014; Smillie et al., 2010).

Most of these tools are, though, available for a short period of time and they are updated at a much slower rate than GenBank (Smillie et al., 2010). One example is ACLAME, a database which, although showing a great potential, has not been updated since 2013. This database was built with the purpose of collecting and classifying all the existing plasmids, transposons and phage genomes.

Regarding integrons, two main tools are known: (i) the INTEGRALL database, still updated on a regular basis (see Table 5), which has the purpose of gathering information about
DNA sequences of integrons on a single repository and (ii) the *Integron Finder* program, which searches for integrons in DNA sequences in FASTA formats, returning the input sequence with all integrons and features found, a list of all the elements detected and the representation of the discovered complete integrons.

As for miscellaneous content, two databases are of relevance: *Isfinder* is designed for the identification of Insertion Sequences and *ICEberg* has the goal of assembling information about Integrative and Conjugative Elements present in bacteria.

The *PlasmidFinder* database allows the *in silico* identification of plasmids in complete sequences and translates them to an Inc group-based classification. However, this is a restricted tool, since it is limited to the identification of plasmids in the *Enterobacteriaceae* family. The *pMLST* web tool is capable of performing a Plasmid Multilocus Sequence Typing analysis for five incompatibility groups (both PlasmidFinder and pMLST rely on the BLASTn algorithm for the homology searches).

Other currently unavailable/out-of-date databases worthy of mention are: the *Plasmid Genome Database* (unavailable), which was intended to be a repository for complete plasmids, harboring information about their core features, genetic composition and structural maps (Molbak, 2003); and the *Database of Completely Sequenced, Naturally Occurring Plasmids* (out-of-date since 2000, available at [http://www.biochem.ucl.ac.uk/bsm/PLASMID/mainpage.htm](http://www.biochem.ucl.ac.uk/bsm/PLASMID/mainpage.htm)) comprises information about a very limited number of plasmids (17 archaeal, 157 bacterial and 27 others), concerning replication regions, transfer or mobility regions, bacteriocin and antibiotic resistance coding regions and other functions.

All the websites and references for the still existing web tools just mentioned are listed in Table 5.

<table>
<thead>
<tr>
<th>Web Tool</th>
<th>Website</th>
<th>Last Modified</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACLAME</td>
<td><a href="http://aclame.ulb.ac.be/">http://aclame.ulb.ac.be/</a></td>
<td>September 17, 2013</td>
<td>(Leplae et al., 2004)</td>
</tr>
<tr>
<td>ICEberg</td>
<td><a href="http://db-mml.sjtu.edu.cn/ICEberg/">http://db-mml.sjtu.edu.cn/ICEberg/</a></td>
<td>unknown</td>
<td>(Bi and Xu, 2012)</td>
</tr>
<tr>
<td>Integron Finder</td>
<td><a href="https://github.com/gem-pasteur/Integron_Finder">https://github.com/gem-pasteur/Integron_Finder</a></td>
<td>December 20, 2016</td>
<td>(Curry et al., 2016)</td>
</tr>
<tr>
<td>Isfinder</td>
<td><a href="https://www-is.biotoul.fr/">https://www-is.biotoul.fr/</a></td>
<td>January 23, 2017</td>
<td>(Siguier et al., 2014)</td>
</tr>
<tr>
<td>PlasmidFinder</td>
<td><a href="https://cge.cbs.dtu.dk/services/PlasmidFinder/">https://cge.cbs.dtu.dk/services/PlasmidFinder/</a></td>
<td>May 31, 2016</td>
<td>Carattoli et al. (2014)</td>
</tr>
<tr>
<td>pMLST</td>
<td><a href="https://cge.cbs.dtu.dk/services/pMLST/">https://cge.cbs.dtu.dk/services/pMLST/</a></td>
<td>October 26, 2015</td>
<td>Carattoli et al. (2014)</td>
</tr>
</tbody>
</table>
DEVELOPMENT

In this chapter, the stages of the development of this thesis are described. Namely, the information used to create and populate the database and the adopted approach for the classification into Incompatibility and MOB groups and in terms of Putative Transferability are presented. In the last section, the features and functionality of the created Shiny application – named PlasmidClassifier: The Bacterial Plasmids Database – and Classification tool are characterized, the used functions from the shiny package are explained and various prints are shown.

3.1 CONSTRUCTION AND STRUCTURE OF THE PLASMID DATABASE

In order to construct the plasmid database, MySQL, an open-source relational database management system (RDBMS), was used. The database was firstly designed using the Enhanced Entity-Relationship (EER) diagram shown in Figure 3.

The plasmid database, as illustrated in the diagram, is organized in two tables:

- **plasmidinfo**: this is the main table, containing information about the plasmids’ accession number, location, size, name, type (circular or linear), GC content of its genomic sequence, taxonomy id, strain, geographical data (the country and/or region in which the plasmid was found), collection date (year of isolation), isolation source and any additional information (if the plasmid is found in extreme conditions or if it is a symbiont, for example). Their Inc and MOB groups (see subsection 3.2.3) are also included. For the MOB groups, the type of Mating Pair Formation (Mpf), Type 4 Coupling Protein (T4CP), VirB4 and their putative transferability, that is, if the plasmid is conjugative, mobilizable or non-mobilizable (see section 2.2 for details) are also comprised.

- **taxonomy**: linked to the plasmidinfo table by the plasmid’s taxonomy id, that is, an identification number which, when used in NCBI’s Taxonomy database (available at https://www.ncbi.nlm.nih.gov/taxonomy), returns the classification and nomenclature...
3.1. Construction and Structure of the Plasmid Database

The construction and structure of the plasmid. This table includes the organism and its lineage (columns L1 to L6 refer to the detailed information).

![Diagram of the plasmid database structure]

**Figure 3.** Enhanced Entity-Relationship (EER) diagram depicting the structure of the database.

The two tables have a multiplicity relationship of N:1, since a plasmid can only have one Taxonomy ID, but the same Taxonomy ID can be present in multiple plasmids: the *pIL105* (NC_000906) and *pMRC01* (NC_001949) plasmids, for example, share the Taxonomy ID 1358, which means that they were both found in the *Streptococcus lactis* species.
3.2 Populating the Database

3.2.1 Obtaining the Plasmid Files

One of the purposes of this thesis, as mentioned throughout the previous sections, was to provide a database gathering as much information about bacterial plasmids as possible. To achieve this goal, genome files containing information about plasmid reference sequences (RefSeq) were downloaded from the NCBI FTP repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid/), which contains several file types (see Table 6).

<table>
<thead>
<tr>
<th>File Type</th>
<th>Description</th>
<th>Relevant retrievable data</th>
</tr>
</thead>
<tbody>
<tr>
<td>plasmid.*.genomic.gbff</td>
<td>GenBank flat file format of the plasmids' genomic sequence(s). Includes the genomic sequence and the contig description</td>
<td>Accession number, plasmid name, size, type, location, host, lineage, isolation source, list of genes and information about protein products</td>
</tr>
<tr>
<td>plasmid.*.genomic.fna</td>
<td>DNA sequence for each plasmid (FASTA format).</td>
<td>DNA sequence, which allows the calculation of the GC content</td>
</tr>
<tr>
<td>plasmid.*.protein.gpff</td>
<td>GenPept format of the accessioned protein products annotated on the genome assembly</td>
<td>Protein's accession number, source plasmid, size, lineage, name and function of the protein product(s)</td>
</tr>
<tr>
<td>plasmid.*.protein.faa</td>
<td>Protein sequence of each protein(^c) (FASTA format)</td>
<td>Protein sequence</td>
</tr>
<tr>
<td>plasmid.*.rna.gbff</td>
<td>GenBank flat file format of RNA products annotated on the genome assembly</td>
<td>RNA's accession number, size, type and features of the RNA (source plasmid, host, molecule type)</td>
</tr>
<tr>
<td>plasmid.*.rna.fna</td>
<td>RNA sequence of the RNA products (FASTA format)</td>
<td>RNA sequence</td>
</tr>
</tbody>
</table>


\(^b\) The symbol * corresponds to the number of the file.

\(^c\) No reference to which plasmid it belongs.

From the file types listed in Table 6, only the .genomic.gbff and .genomic.fna files were used. These compiled files (containing information about multiple plasmids) were then separated into 8935 files, each one containing the GenBank information or DNA sequence, respectively, of a single plasmid, corresponding to a specific accession number (used to name each record). From these, only the files corresponding to complete genomic molecules
(accession number with a NC_, prefix, see Table 4 for details) were kept, filtering out 4437 entries (accession numbers with NG_, NT_, and NZ_ prefixes).

A python script was created to assemble the relevant information from the files, for each plasmid: (i) the accession number, location, size, plasmid name, type, strain, geography, and collection date and isolation source from the .gbff files; (ii) the GC content, calculated from the DNA sequence of the plasmid found in the .fna files; (iii) the taxonomy id, lineage and Inc and MOB groups and Putative Transferability from other sources (explained in subsections 3.2.2 and 3.2.3). The collected information was saved in a table named Plasmid-Data_Jan17.csv and used to populate the plasmidinfo table in the Plasmid Database.

Some fields, however, were not easily reachable:

- Geography and Collection Date: although, as mentioned in section 2.4, some were available, most of this information was missing. Recurring to the publications corresponding to each accession number, the remaining missing data was then manually filled (Conde, 2017). Due to this work, approximately 72% of the plasmids’ isolation countries are now available and all the collection dates were assigned, although most were assigned a time interval and only a few an exact year of isolation.

- Environment: Although most of the plasmids have been attributed one of 4 possible environments (clinical, clinical potential, environmental – air, animal, estuarine, food, freshwater, human, marine, plant and algae, soil and wastewaters – or veterinary) (Conde, 2017), 4.5% of the plasmids still have an unknown isolation source (the definitions for each source are described in this subsection).

These difficulties demonstrate the urge to standardize the available data in GenBank regarding plasmids, encouraging the authors of several publications to provide the detailed information about the plasmids with which they worked, in order to obtain useful and systematized information, as complete as possible.

Characterization of the Plasmid Sources

The characterization of the plasmid’s isolation source was based on the environmental habitat of its bacterial hosts, that is, the source from which the bacteria was isolated:

- Clinical – bacteria isolated from infected patients;
- Clinical potential – infectious bacteria isolated from a source other than an infected patient;
- Veterinary – bacteria isolated from infected animals;
- Environmental-air – obtained from airborne non-clinical bacteria;
• **Environmental-freshwater** – non-clinical bacterial sources isolated from naturally occurring water, such as lakes, ponds, rivers and streams;

• **Environmental-estuarine** – non-clinical bacterial sources from estuarine related settings, such as water samples and sediments.

• **Environmental-food** – non-clinical bacterial sources from food for both human and animal feeding purposes, encompassing the product from all the food processing steps;

• **Environmental-human/animal**: non-clinical bacterial sources isolated from healthy individuals/animals;

• **Environmental-marine** – non-clinical bacterial sources from marine related settings, including water samples and sediments.

• **Environmental-plant and algae** – non-clinical bacteria isolated from plant and algae tissues, including plant and algae endosymbionts, ectosymbionts and phytopathogenic bacteria;

• **Environmental-soil** – non-clinical bacteria isolated from soil;

• **Environmental-wastewaters** – obtained from waters directly affected by anthropogenic influence (a result of domestic, industrial, commercial or agricultural activities), including sewage and any related settings between wastewater treatment and disposal;

• **Environmental** (without subgroup) – isolated from non-clinical bacteria that does not fall into any of the previous environmental subgroups.

• **Not determined** – When the plasmid bacterial sources cannot be determined, either by absence of associated publications in the plasmid sequence report, or by absence of author’s address and email contacts in the submission, inexistence of bacterial strain identification or by other reason;

• **Extreme** – This definition relates to environmental conditions that hardly allow for bacteria to survive. These environments comprise low temperature, low content of oxygen or carbon dioxide in the atmosphere, high levels of radiation, acidity, alkalinity, and absence of water, high sugar or salt concentrations, presence of sulphur, petroleum and other toxic substances. In the plasmidInfo table, this information is in the column Comments and not in Source like the previous entries.
3.2.2 Getting the Taxonomy Information

The taxonomic data was retrieved using a python script, recurring to the `urllib.request` module, which allowed accessing the NCBI’s nuccore database and retrieving the taxonomic information (taxonomy id, organism and lineage) of each plasmid. The compiled information was then saved in a CSV file named `Taxonomy_Info_Jan17.csv` and the taxonomy id was also added to the `plasmidinfo` table. From the entries in this table, only 1517 – corresponding to unique IDs – were kept and used to populate the `taxonomy` SQL table.

3.2.3 Classifying According to Incompatibility and Mobility

To run the commands mentioned throughout this section, NCBI’s Standalone Basic Local Alignment Search Tool (BLAST), also known as BLAST+, was used. This software, available at [ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/](ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), allows the user to run BLAST applications locally from the command line and it is significantly faster than its homonymous online tools when dealing with long queries and/or database sequences (Camacho et al., 2009). Although there are several programs in BLAST+, listed in Adams et al. (2016), only `makeblastdb`, `tblastn` and `PSI-BLAST` were required in the context of this thesis.

The `makeblastdb` tool was used to create two plasmid BLAST databases. The first is a nucleotide database, which was generated by: (i) running a python script that merges the genomic sequences (`fna` files) of the plasmids present in the `plasmidInfo` table and stores them in a file called `Plasmid Database.fsa` and (ii) running the `makeblastdb` command in the command line, specifying the database type as `nucleotide`. The second is a protein database. The data needed for this database was stored in `translatedDatabase.fsa`, which is a compilation of the sequences present in the `Plasmid Database` file, translated in 6 Open Reading Frames (ORFs), obtained using python.

The `tblastn` tool was utilized to identify sequences in the Plasmid Database similar to those on a given query. This program has an alignment type of nucleotide vs. protein, dynamically translating the sequences in the database. It was used to classify the plasmids in terms of Incompatibility and Putative Transferability.

Lastly, `PSI-BLAST` (Position-Specific Iterative Basic Local Alignment Search Tool) was used to reveal distant relationships between the sequences present in the database and a specific query (Altschul, 1997; Adams et al., 2016). This tool relies on a protein vs. protein

---

1 Available at [https://www.ncbi.nlm.nih.gov/books/NBK52637/](https://www.ncbi.nlm.nih.gov/books/NBK52637/)
2 Each sequence includes a header indicating the accession number, the reading direction (`5'->3'` or `3'-> 5'`) and reading frame (1 to 3)
alignment type, which means that the sequences had to be translated prior to its execution and it was employed to classify the MOB Groups (see details in subsection 3.2.3).

Retrieving the Inc Groups

The Incompatibility groups were obtained by running a tblastn program, using the protein sequences suggested in Shintani et al. (2015) as queries\(^3\) and our nucleotide database (Plamid Database, see section 3.2.3). Then, a python script was created to filter the results which did not meet Shintani et al. (2015)’s criteria (E-value below 1e-5 and a query coverage and identity higher than 50\%) (see section 3.2.3) and to assign each plasmid, if possible, its respective Inc Group.

Attaining the MOB Groups

To assign each plasmid to its respective MOB group through PSI-BLAST, Orlek et al. (2017)’s methodology was followed, which means that no identity or query coverage filters were applied and that the appropriate E-value depends on the MOB group\(^4\) (see Orlek et al. (2017) for details). MOBC, MOBF, MOBH, MOBP, MOBQ and MOBV protein sequences were used as queries as suggested by Orlek et al. (2017) and Shintani et al. (2015) and the translatedDatabase file was used as the database. For this classification, two python scripts were created – one keeping only the best possible hit for each plasmid, that is, the hit with the lowest E-value and other using Orlek et al. (2017)’s restrictions and assigning each plasmid its corresponding MOB group, if available.

Initially, for consistency purposes, the goal was to use the tblastn program to carry out the three types of classifications (Inc, MOB and Putative Transferability). However, the results were noticeably worse, with considerably less plasmids being assigned a MOB Group (see section 4.1).

Determining the Type of Putative Transferability

To obtain the plasmids’ type of Mating Pair Formation (Mpf), Type 4 Coupling Protein (T4CP) and VirB4 and, subsequently, classifying them according to Putative Transferability, a tblastn program was executed, using the Plasmid Database file as the database and the sequences suggested in Shintani et al. (2015)’s supplementary material (table S2)\(^5\) were used as queries\(^6\). Then, a python script was executed to filter the best hits according to the proposed approach in Shintani et al. (2015) and to classify plasmids’ Putative Transferability

---

3 \text{These sequences were downloaded and merged automatically using the function urllnherit from python's urllib.request, retrieving each sequence from the NCBI's protein database.}

4 \text{MOBC: < 0.001; MOBF: < 0.01; MOBH: < 0.01; MOBP: < 1; MOBQ: < 0.0001; MOBV: < 0.01}

5 \text{Available at http://www.frontiersin.org/journal/10.3389/fmicb.2015.00242/abstract}

6 \text{These sequences were downloaded and merged using the function urlopen function from python's urllib.request module}
depending on the obtained results. In terms of Transferability, a plasmid can belong to one of 5 categories, as suggested by Shintani et al. (2015) and Smillie et al. (2010):

- For a plasmid without an assigned MOB Group (Without relaxase):
  - Non-mob: non transmissible plasmids.
  - Non-mob, protein export?: plasmids that contain T4CP, VirB4, Mpf (or any two of the three), being potentially involved in protein exportation.

- For a plasmid with an assigned MOB Group (With relaxase):
  - Mob: mobilizable plasmids that code for a relaxase, but are lacking VirB4 and Mpf.
  - Determined conj: Conjugative plasmids that code for a known type of Type IV Secretion System (T4SS), plus relaxase and T4CP.
  - Undetermined conj: Conjugative plasmids that contain a relaxase, T4CP and VirB4 but an unknown Mpf.

However, since there were some inconsistencies in Shintani et al. (2015)'s results, a new definition – Unknown – was added, applicable to plasmids that do not fall into any of the other categories, such as cases where (i) only the MOB group and Mpf type are known (e.g. NC_004464) or (ii) the MOB group and Mpf and VirB4 are available, but not the T4CP (e.g. NC_009469).

For more information about the terms and definitions used above, please check section 2.2.

3.3 shiny

To build the online plasmid repository, the R scientific computing system – resorting to RStudio, an Integrated Development Environment (IDE) – and, more specifically, the Shiny package was used.

R is an open source system (programming language and environment) designed to deal with a multiplicity of problems, especially regarding computational statistics, including basic statistical tests, linear and nonlinear modeling, classification and clustering, among others (Pruim, 2011; R Core Team, 2016). Furthermore, it is also indicated for graphical approaches, enabling the production of high-quality detailed plots (R Core Team, 2016).

Used commonly by scientists, statistics and data miners, R is freely available for Windows, MacOS and Linux. R is often updated, gaining new features on a regular basis. In addition to its core capabilities and base packages, it is extensible, that is, it allows the user to create
supplementary functions and install new packages according to any specific needs (Pruim, 2011; R Core Team, 2016). As of January of 2017, there were 10033 packages available at the Comprehensive R Archive Network (CRAN) repository, which, to be used, need to be installed upon the first utilization and always loaded by the user in each session (Pruim, 2011).

The Shiny package (Chang et al., 2017), currently available in version 1.0.5, is a framework that allows developing interactive web applications/pages with little or no previous knowledge about JavaScript, HTML or CSS (web development skills) (Chang et al., 2017). This package can be installed by running the instruction `install.packages("shiny")` and loaded by `library("shiny")` or `require(shiny)` in the command line. To obtain help about the package, the command `RunExample(example)` can be executed, showing some selected examples. Additionally, an online tutorial is available at [https://shiny.rstudio.com/tutorial/](https://shiny.rstudio.com/tutorial/).

A Shiny application is composed of two or three R scripts: (i) user-interface (ui), which controls the appearance of the application; (ii) server, with the information necessary (instructions) to build the application; and (iii) global, storing the global variables used in the application (these variables can also be defined in `server.R`, but the script becomes too extensive and difficult to read for complex applications). Since, while a Shiny application is running, R is monitoring its performance, the R session will be busy, that is, no other commands can be executed during that time (Chang et al., 2017).

Still within the Shiny application, some additional functions from different packages were used:

- `bsTooltip` from `shinybs` (Bailey, 2015), which adds a tooltip on any Shiny input or output and can be triggered by hovering, clicking or focusing. It was used in the PlasmidClassifier app to show a tooltip when hovering on the `Start Classification Process` button (see Figure 4).

- `show/hide` and `enable/disable` from `shinyjs` (Attali, 2017), which were mostly used to hide and disable buttons when it was not appropriate for them to be operational and to reset input values when needed (see Figure 5 for an example of a disabled button used in the PlasmidClassifier application).

- `render, html_document, pdf_document` and `word_document` from `rmarkdown` (Allaire et al., 2017) and `knitr` (Xie, 2017), which allowed compiling the reactive parameterized PDF, HTML and Word documents described in section 4.2 from a Rmd file.

- The `DT` package (Xie, 2016), used to format the outputted tables in the PlasmidClassifier application and to allow searching and filtering.
Finally, *Shiny* can also be used – recurring, for example, to the *RMySQL* package (Ooms et al., 2017) – in association with other languages such as SQL (Structured Query Language), which allows the integration of a relational database. This combination was useful for the development of the database of complete plasmid sequences, since it includes diverse and abundant information (see chapter 1 and section 2.4).

This tool, however, has some limitations, with highlight to:

- Each tab – or, at least, *tabsetPanel* – not having its own sub-URL by default, forcing the application developer to rely on alternative methods to create them;

- The *Clear file* input button not being completely resettable without having to create some code in *Javascript* or *HTML* or using external packages (which was the adopted approach in the *PlasmidClassifier* application). This is due to the fact that the file is not completely removed: this object starts with a *NULL* value; however, when the file is removed, it does not completely reset, that is, instead of reacquiring the *NULL* value, it gets an unknown one in its place. This is a known common issue and the ideal solution seems to be using a *UpdateFileInput* object, which has not yet been created.

---

7 set of tabs
8 which can be read and followed at [https://github.com/daattali/shinyjs/issues/104](https://github.com/daattali/shinyjs/issues/104).
3.4 The PlasmidClassifier Application

The development of the PlasmidClassifier application was accomplished by creating three R scripts (files), which contain the information to build the shiny app:

- **Global**: stores the global variables, that is, the functions and variables that are accessible throughout the session, such as the plasmidInfo and taxonomy tables (accessed through SQL queries). The required packages – such as shiny, RMySQL and rmarkdown – are also loaded in this file;

- **User-Interface (ui)**: comprises the general structure of the PlasmidClassifier app, i.e., builds the web document. It includes information about the aesthetics and organization of the app: determines the exact location in which all the objects appear and their appearance. The visual aspect of this application was edited by using some CSS and HTML coding in specific items of the tabsetPanels, such as in the headers, the Clear File button (and other buttons) and the text boxes shown in the Figures of this section, for example;

- **Server**: calls the global file and contains the instructions necessary to build the application and react to when the user changes input objects in the app. It is responsible for constructing the responsive components of the tabsetPanels and subsequent tabs, such as the depicted plots, tables and buttons.

The PlasmidClassifier Application (general aesthetics illustrated in Figure 6) is composed of five main tabs:

- **Home**: default page (depicted in Figure 7), containing information pertaining the app (describes its main features and tools);

- **Definitions** (Figure 8): contains a short summary about plasmids, Incompatibility (Inc) and MOB groups and putative transferability and explains the different sources included in the Bacterial Plasmids Database;

- **Lists/Search Tool & Interactive Plots**: comprises the data encompassed in the MySQL Bacterial Plasmids Database divided into three tables (Plasmid, Lineage and MOB) and allows the user to select multiple filters and generate plots accordingly and download the filtered data;

- **Plasmid Classification**: allows the user to verify the existence of a sequence in the database and/or classify it according to its Inc and MOB groups and putative transferability. Also offers the possibility of adding a new sequence to the Bacterial Plasmids Database;
• **About/Contacts** (Figure 9): provides details about the circumstances in which the *PlasmidClassifier* application was created and offers the possibility of contact for any user with questions, comments or suggestions.

The functions used in the *PlasmidClassifier* application are listed in a simplified version in tables 7 to 9. Please note that not all functions used in the application are shown, as the full list would be incomprehensible.

### Table 7.: Functions used in the Home, Definitions and About/Contacts tabs

<table>
<thead>
<tr>
<th><strong>Ui</strong></th>
<th><strong>Server</strong></th>
<th><strong>Purpose</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>includeHTML</em></td>
<td>- - -</td>
<td>Including the text shown in Figures 7, 8 and 9 formatted with HTML</td>
</tr>
<tr>
<td><em>htmlOutput</em></td>
<td><em>renderText</em></td>
<td>Wrapping the summary about the database in the Home tab with HTML</td>
</tr>
<tr>
<td><em>a</em>(img(src = “image.jpg”), href = &quot;<a href="https://website.com">https://website.com</a>&quot;)</td>
<td>- - -</td>
<td>Linking the icons in the About/Contacts tab to its respective website</td>
</tr>
</tbody>
</table>

### Table 8.: Functions used in the Lists/Search Tool & Interactive Plots tabsetPanel

<table>
<thead>
<tr>
<th><strong>Ui</strong></th>
<th><strong>Server</strong></th>
<th><strong>Purpose</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>selectInput</em></td>
<td>- - -</td>
<td>Including the options for the filters shown in Figure 10a</td>
</tr>
<tr>
<td><em>conditionalPanel</em></td>
<td>- - -</td>
<td>Updating the content shown in the sidebarPanel depending on the filter chosen (an example of this content is illustrated in Figure 10b)</td>
</tr>
<tr>
<td><em>checkboxGroupInput</em> updateCheckboxGroupInput</td>
<td></td>
<td>Checking/Unchecking options in the sidebarPanel</td>
</tr>
<tr>
<td><em>actionButton</em> observe</td>
<td></td>
<td>Using and updating the selected filters</td>
</tr>
<tr>
<td><em>mainPanel</em> <em>sidebarPanel</em> (sidebarLayout)</td>
<td></td>
<td>Plots tab</td>
</tr>
<tr>
<td><em>selectInput</em></td>
<td>- - -</td>
<td>Showing the available plots to draw</td>
</tr>
<tr>
<td><em>conditionalPanel</em></td>
<td>- - -</td>
<td>Updating the content shown in the tab depending on the chosen plot</td>
</tr>
<tr>
<td><em>plotOutput</em> renderPlot</td>
<td></td>
<td>Drawing the resulting plot (see Figure 11)</td>
</tr>
<tr>
<td><em>downloadButton</em> downloadHandler</td>
<td></td>
<td>Downloading the resulting plot</td>
</tr>
</tbody>
</table>

**Plasmid, Lineage and MOB tables**

<table>
<thead>
<tr>
<th><strong>Ui</strong></th>
<th><strong>Server</strong></th>
<th><strong>Purpose</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>DT::dataTableOutput</em></td>
<td><em>DT::renderDataTable</em></td>
<td>Rendering the three tables</td>
</tr>
<tr>
<td><em>downloadButton</em></td>
<td><em>downloadHandler</em></td>
<td>Downloading the filtered or unfiltered tables</td>
</tr>
</tbody>
</table>
### 3.4. The PlasmidClassifier Application

#### Table 9: Functions used in the Plasmid Classification tab

<table>
<thead>
<tr>
<th>Component</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>sidebarPanel</code></td>
<td>Displaying descriptive text to increase the comprehension of the application</td>
</tr>
<tr>
<td><code>helpText</code></td>
<td>Enabling the Refresh button</td>
</tr>
<tr>
<td><code>conditionalPanel</code></td>
<td>Conditionally building the ui depending on the user inputs (used in all tabs)</td>
</tr>
<tr>
<td><code>verbatimTextOutput</code></td>
<td>Showing text boxes depicted in Figures 13 and 14</td>
</tr>
<tr>
<td><code>bsTooltip</code></td>
<td>Showing the tooltip explained in Figure 4</td>
</tr>
<tr>
<td><code>ActionButton</code></td>
<td>Showing and enabling the Submit Sequence button</td>
</tr>
<tr>
<td><code>uiOutput</code></td>
<td>Conditionally showing output</td>
</tr>
<tr>
<td><code>radioButtons</code></td>
<td>Choosing the download type of the user report (inside the download handler, <code>rmardown::render</code> is used to generate the user report)</td>
</tr>
<tr>
<td><code>textInput</code></td>
<td>Allowing the user to fill the form to add a sequence to the Bacterial Plasmids Database</td>
</tr>
<tr>
<td><code>hidden(htmlOutput)</code></td>
<td>Conditionally showing the &quot;Your sequence was submitted successfully&quot; text</td>
</tr>
<tr>
<td><code>DT:dataTableOutput</code></td>
<td>Showing the sequence repository table</td>
</tr>
<tr>
<td><code>downloadButton</code></td>
<td>Downloading the sequence repository table</td>
</tr>
</tbody>
</table>
The PlasmidClassifier Application

General Information
The Bacterial Plasmids Database aims to comprise as much data about fully sequenced bacterial plasmids as possible, providing a Classification tool.

The plasmids included in the database were downloaded from the NCBI/FTP repository, retrieving their GenBank descriptive file and DNA sequence. Along with these publications, multiple relevant fields were obtained. These fields are organized in the App in three tables, which can be consulted, filtered, and downloaded in the Lists/Search Tool tab:

- **Plasmid Info**: this is the main table, containing information about the plasmid's accession number (identification of GenBank's descriptive file), size, name, type (circular or linear), GC content of its genomic sequence, taxonomy id, strain, geography (the country and/or region in which the plasmid was found), collection date (year of isolation), isolation source and any additional information (if the plasmid is found in extreme conditions or if it is a symbiont, for example). Their Inc and MOb groups were also included (see tab Definitions for details).

- **Taxonomy Info**: linked to the plasmidinfo table by the plasmid's taxonomy id, that is, an identification number which, when used in NCBI's **Taxonomy database**, returns the classification and nomenclature of the plasmid. This table includes the organism and its lineage (columns L1 to L5 refer to the detailed information).

- **MOB**: this table comprises the plasmid's accession number and, if available, their MOB group, type of Mating Pair Formation (Mpf), Type 4 Coupling Protein (T4CP), VirB4 and their putative transferability, that is, if the plasmid is conjugative, mobilizable or non-mobilizable.

Currently, in this database, there are:
- 4300 bacterial plasmids
- 1517 Taxonomy IDs
- 112 Geographic areas
- 15 Types of isolation environment
- 147 Types of Inc Group
- 7 Types of MOB Group
- 6 Types of Putative Transferability
3.4. The PlasmidClassifier Application

The PlasmidClassifier Application aims to compile as much data about fully sequenced bacterial plasmids as possible, providing a classification tool. The plasmids included in the database were downloaded from the NCBI FTP repository, retaining their GenBank descriptive file and DNA sequence. Along with the related publications, multiple relevant fields were obtained. These fields are organized in the app in three tables, which can be consulted, filtered, and downloaded in the Lists/Search Tool tab:

- **Plasmid Info:** This is the main table, containing information about the plasmids: accession number (identification of GenBank descriptor file), size, name, type (circular or linear), GC content of its genomic sequence, taxonomy ID, hosts, geographical location, and year in which the plasmid was found. The collection date (year of isolation), isolation source, and any additional information if the plasmid is found in extreme conditions or if it is a symbiont, for example. These Inn and MOB groups were also included (see Table Definition for details).

- **Taxonomy:** This is the secondary taxonomy table, that is, an identification number which, when used in NCBI's Taxonomy database, returns the classification and nomenclature of the plasmid. This table includes the organism and its lineage (columns 11 to 16 refer to the detailed information).

- **MOB:** This table comprises the plasmid's accession number; if available, their MOB group, type of Mating-Pair Formation (Vpf), Type 4 Coupling Protein (TCP), VPI and their putative transmissibility, that is, if the plasmid is conjugative, mobilizable or non-modifiable.

Currently, in this database, there are:
- 4398 bacterial plasmids
- 1097 Taxonomy IDs
- 12 Geographic areas
- 15 Taxonomic groups
- 147  Groups
- 7 Types of MOB Group
- 6 Types of Putative Transmissibility

### Lists/Search Tool & Interactive Plots

This tab allows data filtering by GC Content, environment, geography, lineage, MOB group size and putative transmissibility. The following plots are available: distribution by geography, lineage, MOB group and transmissibility, environmental source, year of isolation and proportion graphs for plasmid size and GC content depending on the isolation source and lineage. According to the applied filters, the resulting plots and Plasmid Info Taxonomy Info and MOB tables are created and available for downloading.

### Plasmid Classification

This tab allows you to submit a plasmid sequence and classify it according to its Inc and MOB groups and putative transmissibility. You can search for your sequence in the database:
- If it already exists, a summary with its relevant information is shown and you can download its corresponding PDF/HTML/Word report.
- If it does not exist, you can begin the classification process. Once it is completed, you can download your sequence's report.

The Plasmid Classification report includes three sections: Inc Group, MOB Group and Putative Transmissibility.
- The Inc Group chapter includes the compatibility groups to which your plasmid sequence was matched. For these groups, the three most common countries and years of isolation, average GC content and plasmid size, environmental source and typical MOB group are included. In this section, the plasmids matched Inn Groups are also illustrated.
- The MOB Group section contains information about the MOB Groups matched to your sequence, including the number of plasmids classified as that MOB Group in the Bacterial Plasmids Database, the three most common sources for that year of isolation and geography, the main environmental source, average GC content and size, the two most frequent plasms, the typical type of putative transmissibility, and the most reported Inc Group.
- Finally, the Putative Transmissibility section contains an explanation of the possible putative transmissibility results and, for your matched result, a table with your plasmid’s MOB Group, type of Mating-Pair Formation (MPF), Type 4 Coupling Protein (TCP), VPI and Putative Transmissibility is available, as well as the information mentioned above.

Figure 7.: Homepage of the PlasmidClassifier application
3.4. The PlasmidClassifier Application

Figure 8.: Definitions Tab of the PlasmidClassifier application
3.4.1 Lists/Search Tool and Interactive Plots Tab

This panel allows the user to filter data by GC Content, environment, geography, lineage, MOB Group, Putative Transferability and Size (see Figure 10) and generates data accordingly. These filters correspond to the unique entries in the plasmidInfo and taxonomy tables.

The Incompatibility groups were not included in these options, since a plasmid can have more than one Inc group, which would make the resulting plots and tables difficult to interpret (a plasmid can have numerous arrangements of Inc groups (see, for example, SAP014A (NC_013320), which was the plasmid assigned the highest number of Inc groups9).

The Lists/Search Tool & Interactive Plots panel contains four tabs:

- **Plots**: in this tab, the following plots, which support all available filters, can be depicted: distribution by geographic region, Incompatibility group, lineage (phylum), MOB group and transferability, environmental source and year of isolation and proportion graphs for plasmid size and GC content depending on the isolation source and lineage (see Figure 11). The proportion plots were generated using the R package ggplot2 (Wickham, 2009).

- **Plasmid Table**: displays the plasmidInfo table (see section 3.1) filtered by the user’s selection;

---

3.4. The PlasmidClassifier Application

Figure 10.: Side panel of the Lists/Search Tool and Interactive Plots Tab highlighting the (A) available filters and (B) transferability options.

- **Taxonomy Table:** contains the taxonomic information of the selected data (filtered from the taxonomy table – see section 3.1).

- **MOB Table:** shows the rows that match the applied filters for the MOB groups, type of MPF, T4CP and VirB4 and putative transferability (last five columns from the plasmidInfo table). This information was assigned a new table solely for visualization purposes.

The available filters are applicable to every plot and table, which are all downloadable. Unless the user clicks on the *Remove All Filters* button (see Figure 10), these options will be saved during the whole session (the user can visit the other tabs and features without losing the results).

Finally, the three tables mentioned above allow conducting and downloading the results of (i) searches by column by clicking on the boxes below the column name and/or (ii) general word searches by inserting the keyword of interest in the search box in the upper right corner of the table.
3.4.2 Plasmid Classification Panel

The Plasmid Classification panel was created with the objective of assigning, if possible, each plasmid to its respective Incompatibility (Inc) and MOB groups and type of Putative Transferability and it exhibits a different behavior depending on the user’s input.

This is the most important tool of the PlasmidClassifier Application (its initial appearance is shown in Figure 12), which was accomplished by using the BLAST+ tools and adapting the python scripts described in subsection 3.2.3 and executing them in the command line through R (using the system command).
The PlasmidClassifier Application

In this panel, the user can verify if a certain sequence already exists in the plasmid database, whether it is by submitting its file to the application (in FASTA format\(^\text{10}\)) or by inserting its accession number in a text box, which is disabled after a file is inputted (see Figure 13). If the sequence is uploaded, one of four messages can appear in the Sequence tab:

\begin{itemize}
\item "Only one sequence is allowed at a time. Try again", if there is more than one sequence in the file;
\item "Insert a valid sequence, including a header", if the file is empty or does not contain a descriptive line;
\end{itemize}

\(^{10}\) Only .fasta, .fsa and .fna file extensions are accepted.
• “Your sequence already exists in the PlasmidClassifier database. Click in the Details tab to obtain detailed results and download your sequence’s classification report” (self-explanatory, see Figure 13B);

• “Valid sequence”, followed by the content of the file, if it does not exist in the database.

If the existence of the sequence is verified without uploading it, one of two outputs is generated:

• “Your sequence is not present in the PlasmidClassifier database. Upload it and start the classification process in the Details tab”, or

• “Your sequence already exists in the PlasmidClassifier database. Click in the Details tab to obtain detailed results and download your sequence’s classification report” (see Figure 13A).

If the sequence is uploaded and it does not exist in the Bacterial Plasmids Database, a Start Classification Process button appears in the Details tab and the classification process can be started (see Figure 14). This button is only shown if a valid sequence is submitted and, when clicked, it disappears and disables the submission of new files (until the Clear File button is clicked).

Figure 13.: Output in the Sequence tab for existing sequence (A) without or (B) by uploading it.
The classification process takes, approximately, 15 minutes to be completed. This is due to the size of the BLAST databases, as 4301 sequences will be analyzed for the tblastn programs (Inc group and type of putative transferability) and 25806 for the PSI-BLAST tool (MOB group) (see section 3.2.3 for details). Since the server will be busy, no other tasks (e.g. generating new plots) can be performed during this process. The user can, however, visit the Home, Definitions and About/Contacts tabs and explore their contents (see section 4.2 for details).

Whether the sequence already exists (meaning that the characterization was carried out as described in section 3.2.3 and the results are returned automatically) or the classification process is finished, a summary about that sequence is displayed, containing information about its GC Content, size, Inc and MOB groups and type of Putative Transferability (see Figure 15).

Below that summary, there is the option of downloading the classification report in a PDF, HTML or Word format (see Figure 15). This report is produced automatically through a parameterized pre-formatted Rmd (rmarkdown) file, using, as mentioned in section 3.3, the rmarkdown and knitr packages. The parameters passed from the Shiny application are the

A progress bar is shown throughout the process, displaying one of the following messages: Creating Nucleotide Database, Running Inc Classification, Creating Translated Protein Database, Running MOB Classification, Running Transferability Classification or Finishing Process.
plasmidInfo and taxonomy tables, the sequence name and the assigned Inc and MOB groups and types of MPF, T4CP and VirB4 (and subsequent type of putative transferability) and the data is generated accordingly.

The Plasmid Classification report includes three chapters: Inc Group, MOB Group and Putative Transferability.

- The Inc Group chapter includes the incompatibility groups to which the user’s sequence was matched. For these groups, the three most common countries and years of isolation, average GC content and plasmid size, environmental source and typical MOB group are included. In this section, the plots about the matched Inc Groups are also illustrated.

- The MOB Group section comprises information about the MOB Groups matched to the user’s sequence, including the number of plasmids classified as that MOB Group in the Bacterial Plasmids Database, the three most common entries for their year of isolation and geography, the main environmental source, average GC content and size, the two most frequent phyla, the typical type of putative transferability and the most repeated Inc Group.

- Finally, the Putative Transferability section consists of an explanation of the possible putative transferability results and, for the user’s matched result, a table with the plasmid’s MOB Group, type of Mating Pair Formation (Mpf), Type 4 Coupling Protein (T4CP), VirB4 and Putative Transferability is available, as well as the information mentioned for the MOB Groups.

Figure 15.: Sequence details (A) without and (B) after Classification process
For a sequence not previously existing in the database, i.e., that suffered classification in the PlasmidClassifier app, the final feature of the Plasmid Classification panel is the option of adding it to the database (see Figure 15B). If the user selects Yes, a new – previously hidden – tab (Submission) is opened, showing a form for the user to fill (see Figure 16). The data submitted by the user is then added to a downloadable table denominated seqRepository.csv – available in the PlasmidClassifier application in the Sequence Repository tab – containing the following columns: Accession, Author, Email, Plasmid, Type, TaxonomyID, Strain, Size, GC Content, Country of Isolation, Collection Date, Source, Comments, Inc Group, MOB Group, MPF, T4CP, VirB4, Putative Transferability and Status.

Figure 16.: Adding a new sequence to the Database
For the sequence to be added, the only mandatory information is the *Accession Number* and the *Submit Sequence* button is only available when this field has been filled. However, if the user’s sequence is not yet available in EMBL or NCBI (data not verified), any sequence name is considered valid, except for an empty string. In order to avoid sequences with the same code, the existence of the accession number is verified in the table and, if it is found, the string ".1" is appended to the name before adding it. Lastly, if the user clicks on the *Submit Sequence* button and the sequence is added, the message *Your sequence was added successfully* is displayed (replacing the *Submit Sequence* button).

Once a sequence is added, it will become visible in the *Sequence Repository* tab, with its status marked as *Pending*. All new sequences will be individually analyzed to evaluate if they are suitable to enter the *Bacterial Plasmids Database*: if they are, their status is updated to *Accepted* (see Figure 17); otherwise, they will be removed from the table.

Finally, if the user wishes to obtain information or classify a new sequence, he/she can click on the *Clear File* button, which removes the uploaded file and, consequently, resets all inputs and outputs, enables the *Verify Sequence* feature and, if shown during that particular session, hides the *Submission* tab.
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Details</th>
<th>Submission</th>
<th>Sequence Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Author</td>
<td>Email</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>author1</td>
<td>email2</td>
<td></td>
</tr>
</tbody>
</table>

Figure 17: Sequences in the seqRepository table
RESULTS AND DISCUSSION

In this chapter, the main results are presented, regarding the classification into Inc and MOB groups and Putative Transferability and the PlasmidClassifier application, with the latter divided into two subsections: (i) showing the generated plots and tables for the full plasmidInfo and taxonomy tables, that is, for when no filters are selected by the user and (ii) explaining the user reports obtained for different matches.

4.1 PLASMID CLASSIFICATION RESULTS

The results obtained for the Inc and MOB group classifications, as well as the Putative Transferability, showed some discrepancies when compared to the outcomes found in Shintani et al. (2015) and Orlek et al. (2017). These inconsistencies may be caused by the database size, as it can influence the produced E-value (Pevsner, 2015). Since an E-value is the probability of a hit being found by chance multiplied by the database size\(^1\), it is implied that a low E-value might be overestimated (too low) for smaller databases and underestimated (too high) for larger databases: for the former, low E-values can be found in unrelated sequences, whereas, for the latter, a high E-value can be produced when the sequences are, in fact, related (Pevsner, 2015).

The classification into MOB groups was accomplished by executing a PSI-BLAST program, based on the criteria described in Orlek et al. (2017), which resulted in 1909 plasmids being attributed a MOB group in the plasmidInfo table. In order to verify the assumption that the E-value is dependent on the database size, Orlek et al. (2017)’s research was reproduced using the proposed parameters and the same results were obtained (data not shown). Hence, it was verified that the database size does, indeed, influence the produced results (Orlek et al. (2017)’s database has approximately half of the Bacterial Plasmids Database’s size – 12582 vs. 25800).

\(^1\) An E-value of \(x\) states that, for a given database, there are \(x\) hits expected to have the same or a higher score than the current hit.
As stated in section 3.2.3, it was initially planned to classify the MOB groups relying on the *tblastn* tool. However, the results using this method were discarded, since only 1224 plasmids were attributed a MOB group, which is noticeably worse than the observed with *PSI-BLAST* (difference of 685 plasmids).

The *tblastn* results were compared to the data provided by Shintani et al. (2015). In the *plasmidInfo* table, 1387 plasmids were assigned a known Inc group. However, some plasmids were attributed numerous GR groups (*NC_010309*, for example), which might entail that the parameters for those specific Inc subgroups are not rigorous enough.

Finally, in terms of Putative Transferability, 2389 plasmids were classified as non-mob, 2 as non-mob, protein export, 1600 as mob, 233 as determined conjugative, 1 as undetermined conjugative and 75 as unknown.

The comparison between the overlapping results between this study and Shintani et al. (2015) and Orlek et al. (2017) are shown in Tables 10 and 11.

Table 10.: Comparison between the results obtained in this study and Shintani et al. (2015) and Orlek et al. (2017)

<table>
<thead>
<tr>
<th>MOB Groups (Orlek)</th>
<th>Inc Groups (Shintani)</th>
<th>Transferability (Shintani)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared Results(^a)</td>
<td>451</td>
<td>4105</td>
</tr>
<tr>
<td>Matches</td>
<td>429 (95.12%)</td>
<td>2971(^b) (72.37%)</td>
</tr>
</tbody>
</table>

\(^a\) Plasmids that exist in the *Bacterial Plasmids Database* and in the author’s results

\(^b\) At least partially matched

Table 11.: Transferability outcomes for the results in common between this study and Shintani et al. (2015)

<table>
<thead>
<tr>
<th></th>
<th>Shintani</th>
<th>This Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>determined conj</td>
<td>407</td>
<td>200</td>
</tr>
<tr>
<td>mob</td>
<td>773</td>
<td>1529</td>
</tr>
<tr>
<td>non-mob</td>
<td>2726</td>
<td>2302</td>
</tr>
<tr>
<td>non-mob, protein export?</td>
<td>156</td>
<td>2</td>
</tr>
<tr>
<td>undetermined conj</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>unknown</td>
<td>—</td>
<td>71</td>
</tr>
</tbody>
</table>

**Total: 4105 plasmids**
4.2 THE PLASMIDCLASSIFIER APPLICATION

This Shiny application has been tested on an 8.1 Windows with a 64-bit operation system and x64-based processor on Mozilla Firefox, Google Chrome and Internet Explorer, in versions 56.0.1 (64-bit), 61.0.3163.100 (64-bit) and 11.0.9600.18817, respectively. Although the differences are minimal, the PlasmidClassifier app shows its optimal behavior in Google Chrome and, thus, this browser should be preferred when running it and is the one shown in the prints throughout this chapter. However, this app has not been tested in any Linux or Apple operating systems, so its performance on those systems is still unknown.

4.2.1 Generated Plots and Tables

As explained in the last chapter, the Lists/Search Tool & Interactive Plots tabsetPanel in the PlasmidClassifier application allows the user to generate plots and select specific data from the plasmidInfo and taxonomy tables by applying varying filters. For example, when no filters are applied the plots depicted in Figures 18 and 19 give the following information:

a) **Distribution by Geographical Source**: in this plot, it becomes noticeable that most of the plasmids have an unknown region of isolation and that the most common countries are the United States of America, China, Japan and Germany.

b) **Distribution by Incompatibility Group**: An evidence of the complexity of the Inc groups is the number of total results in this graph, which, unlike the other plots, totalizing 4300 plasmids, is composed of 6943 results, because this information is not exclusive (there can be more than one Inc group associated with a given plasmid). In this plot, it is evidenced that the majority of the plasmids is not classified according to Incompatibility and that the most common Inc group is IncFII(RepA4), followed by IncZ and pCD1.

c) **Distribution by Lineage**: it is evident that the prevailing phyla are Proteobacteria and Firmicutes (2101 and 1097 plasmids, respectively), being much more numerous than the third most common phylum, Spirochaetes (413 results).

d) **Distribution by MOB Group and Transferability**: although predominantly not classified, the main MOB group found in the Bacterial Plasmids Database is MOBP and the least dominant is MOBH.

e) **Distribution by Source**: although essentially clinical, the plasmids are found in a multitude of environments, with highlight to food products and soil.
f) Distribution by Year: most of the plasmids have not been associated with a specific isolation year, indicating that this information is lacking in both the plasmids' descriptive files and related publications.

g) GC Content by Lineage: This plot shows the affluence of the different phyla according to their GC content. It becomes evident that most plasmids have a typical GC content value ranging from 0.20 to 0.30 and that plasmids with a GC Content above 0.60 are rarely found. Proteobacteria and Firmicutes are the groups with the most varying GC Content and, despite being very common, Spirochaetes is mainly limited to a value between 0.20 and 0.40.

h) GC Content by Source: There are clinical plasmids (including clinical potential and veterinary) for all GC Content percentages. However, these display their maximum abundance at a GC Content value below 60%, especially between 20% and 40%. Inversely, with the exception of plasmids found in food products, which show a similar distribution to the clinical specimens, the environmental plasmids typically have a high GC Content percentage, usually ranging from 40% to 70%.

i) Size by Lineage: Proteobacteria occurs in all sizes, while the other phyla are restricted to a limited range.

j) Size by Source: On one hand, it is observable that there are no clinical plasmids (including clinical potential and veterinary) in the Bacterial Plasmids Database with a size over 500kb or under 1kb (100 base pairs) and are typically small, with an average size ranging from 20kb to 40kb. The environmental plasmids, on the other hand, are present in all sizes, particularly between 100 kb and 200 kb.

From the 4498 complete RefSeq files mentioned in section 3.2, only 4300 plasmids were considered, including only bacterial plasmids (viruses and archaeal and eukaryotic plasmids were removed). Figures 20, 21 and 22 show the generated information when no selections are made, comprising data about the 4300 plasmids and their MOB groups from the plasmidInfo table and 1517 taxonomy ids from the taxonomy table.

4.2.2 Classification Results

In the Attachments, three examples of user reports in PDF format are included: (i) one for when no Inc or MOB groups are found in A.1 and (ii) two for successful classification processes in A.2 (single match) and A.3 (multiple matches). Albeit having some information in common, there are several differences between the three user reports, allowing some deductions to be inferred:
• For no matches, it is evidenced that unclassified incompatibility groups are commonly found in plasmids without a known MOB group and, consequently, classified as non mobilizable (or protein-export) in terms of transferability. Inversely, in A.2 and A.3, two different types of MOB groups were found as the most common entry for the different types of Inc groups and the plasmid was classified as mobilizable or determined conjugative, respectively. This shows that a given Inc group can be associated with a specific and characteristic MOB group, which can help in the validation of its classification;

• Different MOB groups seem to be related to different countries. For example, plasmids classified as MOBF appear to be very common in Japan, which is not true for the MOBP plasmids. This evidence could be helpful in proving that the geographical location of the plasmid could have an influence on its characteristics;

• Independently of the case, the most common phylum is Proteobacteria and the typical environment is clinical\(^2\) (this was expected, since most of the plasmids in the database are associated with those categories);

• Although no conclusions can be drawn about the GC Content percentage, the plasmids associated with a Inc Group seem to have a higher average size than the unclassified plasmids, which suggests that it could be a differentiating factor in plasmid characterization.

From the information listed above, it becomes evident that, when paired with further analysis and validation against multiple results, the depicted plots and summaries provided in the classification report can be very useful in plasmid characterization, providing crucial information about the plasmid sequence in study.

\(^2\) A plasmid from a veterinary source is still considered as from a clinical environment
Figure 18.: Distribution Plots depicted in the PlasmidClassifier app when no filters are selected
Figure 19: Proportion Plots drawn in the PlasmidClassifier app when no filters are selected, using the ggplot2 package.
The PlasmidClassifier Application

Figure 20: First 10 entries of the Plasmid Table available in the Lists/Search and Interactive Plots Panel when no filters are selected.
### Figure 21. First 10 entries of the Taxonomy Table available in the Lists/Search and Interactive Plots panel when no filters are selected

<table>
<thead>
<tr>
<th>TaxonomyID</th>
<th>Organism</th>
<th>Lineage</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Buchneria aphidicola</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>Enterobacteriales</td>
<td>Erwiniaaceae</td>
<td>Buchneria</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Myxococcus fulvus</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Delta proteobacteria</td>
<td>Myxococcales</td>
<td>Cystobacterinae</td>
<td>Myxococcales</td>
<td>Myxococcus</td>
</tr>
<tr>
<td>3</td>
<td>Borrelia burgdorferi</td>
<td>Bacteria</td>
<td>Spirochaetes</td>
<td>Spirochaetes</td>
<td>Borreliaeaceae</td>
<td>Borrelia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Borrelia hermsii</td>
<td>Bacteria</td>
<td>Spirochaetes</td>
<td>Spirochaetes</td>
<td>Borreliaeaceae</td>
<td>Borrelia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Treponema denticola</td>
<td>Bacteria</td>
<td>Spirochaetes</td>
<td>Spirochaetes</td>
<td>Spirochaetesaceae</td>
<td>Treponema</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Leptospirillum ferrooxidans</td>
<td>Bacteria</td>
<td>Nitrospirae</td>
<td>Nitrospirales</td>
<td>Nitrospiraceae</td>
<td>Leptospirillum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Campylobacter coli</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Epsilon proteobacteria</td>
<td>Campylobacteriales</td>
<td>Campylobacteraceae</td>
<td>Campylobacter</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Campylobacter jejuni</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Epsilon proteobacteria</td>
<td>Campylobacteriales</td>
<td>Campylobacteraceae</td>
<td>Campylobacter</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Campylobacter lari</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Epsilon proteobacteria</td>
<td>Campylobacteriales</td>
<td>Campylobacteraceae</td>
<td>Campylobacter</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Helicobacter pylori</td>
<td>Bacteria</td>
<td>Proteobacteria</td>
<td>Epsilon proteobacteria</td>
<td>Campylobacteriales</td>
<td>Helicobacteraceae</td>
<td>Helicobacter</td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 1,517 entries

Previous: 1 2 3 4 5 ... 152 Next

Download Table
Figure 22: First 10 entries of the MOB Table available in the Interactive Plots panel when no filters are selected.

<table>
<thead>
<tr>
<th>Accession</th>
<th>MOB_Group</th>
<th>MPF</th>
<th>T4CP</th>
<th>VirB4</th>
<th>Putative_Transferability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_000808</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000914</td>
<td>MOBQ</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mob</td>
</tr>
<tr>
<td>NC_000923</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000937</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000938</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000948</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000949</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000950</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000951</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
<tr>
<td>NC_000952</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 4,300 entries.
CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

From the critical research of existing tools and development and testing of the PlasmidClassifier application, some conclusions could be drawn.

Through the methods described in section 3.2.3, approximately 72% of the plasmids’ isolation countries are now available and all the collection dates were assigned, although most were assigned a time interval (1655 plasmids) and only a few an exact year of isolation (1645 plasmids), 32.2% and 44.4% were classified according to incompatibility and mobility and only 4.5% still have an unknown isolation source. These results could, however, be further improved by the normalization of the information in GenBank and by adjusting the filtering parameters and database size used in the similarity searches. This is still a meticulous process, which will require carefully studying the response of the BLAST+ programs to the varying filters.

After working with Shiny (and its accompanying packages), used to develop the PlasmidClassifier application, it is evidenced that this is a thoroughly documented and very helpful tool for the creation of webpages and/or applications. It has a plenitude of functionalities, which, when paired with some basic knowledge of HTML and CSS, allow the automation/parameterization of processes which otherwise would be impossible, such as the output of the reactive rmarkdown classification report (see section 3.4.2).

As previously stated in section 2.5, the current tools for plasmid classification are quite restrictive to the clinical environment, focusing on the Enterobacteriaceae family and other clinically related bacterial genera, such as Enterococcus, Staphylococcus and Streptococcus. The PlasmidFinder and pMLST tools (Carattoli et al., 2014) are an example of this limitation, not only for focusing singularly on those specific organisms, but, in pMLST’s case, also for only supporting a typing analysis for five incompatibility groups.
During the development of this thesis, a new online tool for plasmid classification, named PlasmidProfiler, was created (Zetner et al., 2017). Albeit partially sharing some features with the PlasmidClassifier application, this tool relies on Whole Genome Shotgun (WGS) reads, which are genome assemblies of incomplete genomes/chromosomes of prokaryotes or eukaryotes (Clark et al., 2016) and it only classifies plasmids in terms of Incompatibility, not comprising information about the MOB group or the type of Putative Transferability. It also serves a different purpose, with the final goal of visualizing the results as an interactive heat map (Zetner et al., 2017).

The PlasmidClassifier application can, hence, be considered dissimilar from any preexisting tool, showing the prospective of being helpful to any user with the objective of classifying new plasmid sequences independently of their source environment or simply interested on the detailed information about previously classified fully sequenced RefSeq plasmids. For evolutionary/epidemiological study purposes, it is in the best interest of this repository to be continually enriched with complete sequences, which will demand an effort from the scientific community. This web tool is intended to be an unbiased contribution to the organization and analysis of bacterial plasmids, not only restricted to clinical settings and contexts, but including all environmental sources. Even though obtaining complete plasmid sequences can still present challenges, this web tool is an attempt to encourage the effort. A richer and more representative plasmid database, containing the complete replicon sequence and the ecological features, is mandatory to carry proper epidemiological and evolutive analyses.

By applying the implemented methodology (described in chapter 3), the PlasmidClassifier application may, in some cases, replace the heavily time-consuming process of in vitro classification. Despite directed mainly towards complete plasmid sequences, and not scaffolds or incomplete genomic regions (see table 4), this tool is also capable of dealing with incomplete sequences, as long as the classification obtained for these sequences is not viewed as indubitable, since adding a new portion or closing the sequence might generate a different outcome.

Lastly, it should be noted that the PlasmidClassifier application is still under development and has not been evaluated by scientific peer review, reason for which the asserted conclusions cannot be considered final and should be further tested.
5.2 PROSPECT FOR FUTURE WORK

The first stages of the future work for the PlasmidClassifier App have already been outlined. One of the first tasks to complete following this thesis is allowing the user to upload several sequences at a time. This step will consist of:

- Deciding if the multiple sequences should be uploaded by a single file with multiple sequences, multiple files containing just one sequence or both;

- Adapting the python scripts executed in the Shiny App through the system command (see section 4.2).

- Verifying each sequence, that is, checking if it exists in the database;

- Instead of choosing the download type of the resulting user reports (see subsection 4.2), returning a .zip file to the user, containing a PDF, HTML and Word document for each of the sequences inputted, with the file name being the name of the sequence;

- Upon user submission, depending on whether the sequences exist in the Bacterial Plasmids Database or not, creating new tabs for each of the sequences, prompting the user to start the classification process or automatically returning the sequence summary and download possibility.

- For each classified sequence, providing the option of adding it to a new table, with the possibility of being included in the Bacterial Plasmid Database. If the user decides to add that sequence, new tabs will be created and opened in the Plasmid Classification tabsetPanel, in order to fill the form about that plasmid and submit it (see section 4.2).

Another essential task will be monitoring the updates in the NCBI Genome Repository (see section 2.4) and adding/excluding sequences according to updated data (sequences can be excluded due to the reasons stated in section 2.4). Although having the goal of completely automating the process, most of the fields in the table are not standardized in the NCBI Genome Repository, which means that some manual curation will be required before updating the database. Hopefully, with NCBI starting to systematize the Genbank genomic files (see Table 6), the manual curation and necessity to consult the related publications will decrease gradually over time. Two examples of this systemization are the pDESACI plasmids (NC_018066¹-67), whose GenBank entries contain a table with the environmental source, year and geographical region of isolation.

In terms of interaction with the user, a convenient feature would be to notify the user by email any time the Bacterial Plasmids Database is updated or gains a new functionality or if his sequence(s) is/are accepted. To receive the email with the notification of a new update, the user should be, with his consent, added to a mailing list, while the message for an accepted sequence would require no authorization, just inserting the email in the form to submit sequences (see section 4.2).

Aside from the tasks mentioned above, other minor improvements and aesthetic corrections will be applied, such as centering the labels in the `tabsetPanel` titles and creating different URLs for each tab, which would increase the accessibility to the different tools of the application.

By completing the aforementioned tasks, new ones will certainly emerge, since Plasmid-Classifier is an ongoing project designed to be supplemented with new features and updated regularly.
REFERENCES


62


---

REFERENCES 66
Notes:
1 A result of "-" means that your sequence was not matched to any Inc/MOB group.
2 ND, for the country of isolation, indicates that the requested information about the plasmid is not known, since it was available neither in the sequence file nor in the related publications.
3 These results are subjective and depend on the database size. In this case, your sequence was compared to 4300 plasmids. There is no assurance that the obtained results would be the same for a database with a different size, since the smaller the database, the larger the overestimation of the results (that is, plasmids are assigned a higher identity or query coverage than expected).
4 If your sequence is incomplete, the results cannot be considered conclusive, as they may differ for the complete sequence. Ideally, a complete sequence should be submitted.

Inc Group

For a short summary about plasmid incompatibility, please visit the Definitions tab in the app.

Your sequence had similarities with the following Inc Group(s):
- 
In the plasmid database, there are:
  -> 2913 unclassified plasmids (79.791%)

For this/these group(s), the three most common countries in which it/they exist(s) are:
  -> ND, USA, China

The three most common entries in the Collection_Date field are:

The average GC_Content and Plasmid Size are, respectively:
  -> 0.421; 61880.3

The main environmental source is:
  -> clinical

The most seen MOB Group is:
  -> -
Unknown Inc Group

Distribution by Geography

- Australia (1.68%)
- China (8.14%)
- France (3.19%)
- Germany (3.3%)
- Japan (7.42%)
- ND (29.56%)
- Other Countries (27.46%)
- Spain (1.79%)
- UK (2.06%)
- USA (15.41%)

Distribution by Lineage

- Actinobacteria (4.67%)
- Bacteroidetes (2.13%)
- Chlamydiae (2.23%)
- Cyanobacteria (4.67%)
- Deinococcus-Thermus (1.37%)
- Firmicutes (24.65%)
- Others (1.72%)
- Proteobacteria (42.26%)
- Spirochaetes (14.18%)
- Tenericutes (2.13%)

Distribution by MOB Group

- - (63.47%)
- MOBC (3.36%)
- MOBF (4.74%)
- MOBH (0.96%)
- MOBP (12.77%)
- MOBQ (7.24%)
- MOBV (7.45%)

Distribution by Transferability

- determined conjugative (0.51%)
- mob (34.57%)
- non-mob (63.44%)
- non-mob, protein export (0.03%)
- undetermined conjugative (0.03%)
- unknown (1.41%)

Distribution by Source

- clinical (26.47%)
- clinical potential (7.38%)
- environmental (2.95%)
- environmental-air (0.17%)
- environmental-animal (5.84%)
- environmental-estuarine (0.48%)
- environmental-food (12.02%)
- environmental-freshwater (5.73%)
- environmental-human (2.47%)
- environmental-plant and algae (7.93%)
- environmental-soil (10.09%)
- environmental-wastewaters (2.3%)
- ND (4.46%)

GC Content by Lineage

Plasmid GC Content

- 0.20-0.30
- 0.30-0.40
- 0.40-0.50
- 0.50-0.60
- 0.60-0.70
- 0.70-0.80

Number of Plasmids

GC Content by Source

Plasmid GC Content

- 0.20-0.30
- 0.30-0.40
- 0.40-0.50
- 0.50-0.60
- 0.60-0.70
- 0.70-0.80

Number of Plasmids
For more detailed results and interactive downloadable plots similar to those above, please visit the Interactive Plots tab in the app, which allows you to filter tables according to any MOB Group, as well as the source environment, GC Content percentage, size (kb) and country/region of origin.

**MOB Group**

Your sequence had similarities with the following MOB Group(s): -

In the plasmid database, there are 2391 - plasmids (55.605%). The three most common countries in which this result appears are: ND, USA, China. The three most common entries for the collection date are <2008, <2007, <2011, the main environmental source for that group is clinical and the average GC Content and size are 0.4143863 and 62633.598, respectively. Lastly, the two most frequent phyla are Proteobacteria and Firmicutes, with 931 and 605 entries, respectively and the typical type of putative transferability for the - group is non-mob (see definition below) and the most repeated Inc Group is -.

For more detailed results and interactive downloadable plots similar to those above, please visit the Interactive Plots tab in the app, which allows you to filter tables according to any MOB Group, as well as the source environment, GC Content percentage, size (kb) and country/region of origin.

**Putative Transferability**

In terms of transferability, a plasmid can belong to one of 6 categories:

For a plasmid without an assigned MOB Group (Without relaxase):
- **Non-mob**: non transmissible plasmids.
- **Non-mob, protein export?**: plasmids that contain T4CP, VirB4, MPF (or any two of the three), being potentially involved in protein exportation.

*For a plasmid with an assigned MOB Group (With relaxase):*
- **Mob**: mobilizable plasmids that code for a relaxase, but are lacking VirB4 and MPF.
- **Determined conj**: Conjugative plasmids that code for a known type of T4SS, plus relaxase and T4CP.
- **Undetermined conj**: Conjugative plasmids that contain a relaxase, T4CP and VirB4 but an unknown MPF.
- **Unknown**: plasmids that do not fall into any of the other categories.

For your sequence, the following table was obtained:

<table>
<thead>
<tr>
<th>Accession</th>
<th>MOB_Group</th>
<th>MPF</th>
<th>T4CP</th>
<th>VirB4</th>
<th>Putative_Transferability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2122</td>
<td>NC_014163</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>non-mob</td>
</tr>
</tbody>
</table>

In the plasmid database, there are **2389** plasmids classified as **non-mob** (55.56 %). The most common MOB group associated with this result is - and the most repeated Inc Group is -.

For the non-mob plasmids, the most common phylum is **Proteobacteria**, the main environmental source is **clinical**, the three most seen countries are **ND, USA, China**, the three most common entries for the year of isolation are <2008, <2007, <2011. Finally, the average Size and GC Content percentage are, respectively, 62622.16 and 0.4143.

To create downloadable interactive plots, please visit the **Interactive Plots** tab in the app, which allows you to filter tables according to any type of putative transferability, as well combining it with the source environment, GC Content percentage, size (kb) and country/region of origin.
Plasmid Classification Report
NC_002636

Notes:
1 A result of “-” means that your sequence was not matched to any Inc/MOB group.
2 ND, for the country of isolation, indicates that the requested information about the plasmid is not known, since it was available neither in the sequence file nor in the related publications.
3 These results are subjective and depend on the database size. In this case, your sequence was compared to 4300 plasmids. There is no assurance that the obtained results would be the same for a database with a different size, since the smaller the database, the larger the overestimation of the results (that is, plasmids are assigned a higher identity or query coverage than expected).
4 If your sequence is incomplete, the results cannot be considered conclusive, as they may differ for the complete sequence. Ideally, a complete sequence should be submitted.

Inc Group

For a short summary about plasmid incompatibility, please visit the Definitions tab in the app.

Your sequence had similarities with the following Inc Group(s):
IncQ (=IncP-4)

In the plasmid database, there are:
- 43 IncQ (=IncP-4) plasmids (1%)
For this/these group(s), the three most common countries in which it/they exist(s) are:
- ND, Japan, Australia
The three most common entries in the Collection_Date field are:
The average GC_Content and Plasmid Size are, respectively:
- 0.552; 74523.3
The main environmental source is:
- veterinary
The most seen MOB Group is:
- MOBP
MOB Group

Your sequence had similarities with the following MOB Group(s): **MOBQ**

In the plasmid database, there are 416 MOBQ plasmids (9.674%). The three most common countries in which this result appears are: ND, Japan, USA. The three most common entries for the collection date are <2006, <2009, <2011, the main environmental source for that group is **clinical** and the average GC Content and size are 0.4975082 and 141367.346, respectively. Lastly, the two most frequent phyla are **Proteobacteria and Firmicutes**, with 286 and 93 entries, respectively and the typical type of putative transferability for the MOBQ group is **mob** *see definition below* and the most repeated Inc Group is -.

For more detailed results and interactive downloadable plots similar to those above, please visit the Interactive Plots tab in the app, which allows you to filter tables according to any MOB Group, as well as the source environment, GC Content percentage, size (kb) and country/region of origin.

Putative Transferability

In terms of transferability, a plasmid can belong to one of 6 categories:

*For a plasmid without an assigned MOB Group (Without relaxase):*
- **Non-mob**: non transmissible plasmids.
- **Non-mob, protein export?**: plasmids that contain T4CP, VirB4, MPF (or any two of the three), being potentially involved in protein exportation.

*For a plasmid with an assigned MOB Group (With relaxase):*
- **Mob**: mobilizable plasmids that code for a relaxase, but are lacking VirB4 and MPF.
- **Determined conj**: Conjugative plasmids that code for a known type of T4SS, plus relaxase and T4CP.
- **Undetermined conj**: Conjugative plasmids that contain a relaxase, T4CP and VirB4 but an unknown MPF.
- **Unknown**: plasmids that do not fall into any of the other categories.

For your sequence, the following table was obtained:

<table>
<thead>
<tr>
<th>Accession</th>
<th>MOB_Group</th>
<th>MPF</th>
<th>T4CP</th>
<th>VirB4</th>
<th>Putative_Transferability</th>
</tr>
</thead>
<tbody>
<tr>
<td>192</td>
<td>NC_002636</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>mob</td>
</tr>
</tbody>
</table>

In the plasmid database, there are **1600** plasmids classified as **mob (37.21 %)**. The most common MOB group associated with this result is **MOBP** and the most repeated Inc Group is `-`. For the mob plasmids, the most common phylum is **Proteobacteria**, the main environmental source is **clinical**, the three most seen countries are **ND, USA, Japan**, the three most common entries for the year of isolation are `<2012, <2011, <2009`. Finally, the average Size and GC Content percentage are, respectively, 88666.24 and 0.4612.

To create downloadable interactive plots, please visit the *Interactive Plots* tab in the app, which allows you to filter tables according to any type of putative transferability, as well combining it with the source environment, GC Content percentage, size (kb) and country/region of origin.
Plasmid Classification Report
NC_002122

Notes:
1 A result of “-” means that your sequence was not matched to any Inc/MOB group.
2 ND, for the country of isolation, indicates that the requested information about the plasmid is not known, since it was available neither in the sequence file nor in the related publications.
3 These results are subjective and depend on the database size. In this case, your sequence was compared to 4300 plasmids. There is no assurance that the obtained results would be the same for a database with a different size, since the smaller the database, the larger the overestimation of the results (that is, plasmids are assigned a higher identity or query coverage than expected).
4 If your sequence is incomplete, the results cannot be considered conclusive, as they may differ for the complete sequence. Ideally, a complete sequence should be submitted.

Inc Group

For a short summary about plasmid incompatibility, please visit the Definitions tab in the app.

Your sequence had similarities with the following Inc Group(s):
IncB/O, IncFII(RepA1), IncFII(RepA4), IncI, IncK

In the plasmid database, there are:
- 92 IncB/O plasmids (2.14%)
- 92 IncFII(RepA1) plasmids (2.14%)
- 219 IncFII(RepA4) plasmids (5.093%)
- 92 IncI plasmids (2.14%)
- 92 IncK plasmids (2.14%)

For this/these group(s), the three most common countries in which it/they exist(s) are:
- USA, ND, Switzerland
- USA, ND, Switzerland
- ND, USA, China
- USA, ND, Switzerland
- USA, ND, Switzerland

The three most common entries in the Collection_Date field are:
- <2014, 2013, 2010
- <2014, 2013, 2010
- <2014, 2013, 2010
- <2014, 2013, 2010

The average GC_Content and Plasmid Size are, respectively:
- 0.499; 99495.9
- 0.499; 99495.9
- 0.498; 106737.5
- 0.499; 99495.9
- 0.499; 99495.9

The main environmental source is:
The most seen MOB Group is:

- MOBP
- MOBP
- MOBP
- MOBP
- MOBP

**IncB/O**

- Distribution by Geography:
  - ND (23.91%)
  - Other Countries (39.13%)
  - Switzerland (6.52%)
  - UK (5.43%)
  - USA (25%)

- Distribution by Lineage:
  - Proteobacteria (100%)

- Distribution by MOB Group:
  - MOBP (78.26%)
  - MOBC (1.09%)
  - MOBF (9.78%)
  - Other (10.87%)

- Distribution by Transferability:
  - determined conjugative (79.35%)
  - mob (8.7%)
  - non-mob (10.87%)
  - unknown (1.09%)

- Distribution by Source:
  - clinical (46.74%)
  - clinical potential (11.96%)
  - environmental (1.09%)
  - environmental-animal (6.52%)
  - environmental-food (1.09%)
  - environmental-human (5.43%)
  - environmental-plant and algae (2.17%)
  - environmental-soil (1.09%)
  - ND (5.43%)
  - veterinary (18.48%)

- Distribution by Year:
  - Others (50%)
  - <2011 (6.52%)
  - <2012 (4.35%)
  - <2013 (5.43%)
  - <2014 (8.7%)
  - 2002 (4.35%)
  - 2010 (7.61%)
  - 2011 (4.35%)
  - 2013 (8.7%)
  - Others (50%)
IncFII(RepA1)

**Distribution by Geography**
- ND (23.91%)
- Other Countries (39.13%)
- Switzerland (6.52%)
- UK (5.43%)
- USA (25%)

**Distribution by Lineage**
- Proteobacteria (100%)

**Distribution by MOB Group**
- determined conjugative (79.35%)
- mob (8.7%)
- non-mob (10.87%)
- unknown (1.09%)

**Distribution by Transferability**
- determined conjugative (79.35%)
- mob (8.7%)
- non-mob (10.87%)
- unknown (1.09%)

**Distribution by Source**
- clinical (46.74%)
- clinical potential (11.96%)
- environmental (1.09%)
- environmental-animal (6.52%)
- environmental-food (1.09%)
- environmental-human (5.43%)
- environmental-plant and algae (2.17%)
- environmental-soil (1.09%)
- ND (5.43%)
- veterinary (18.48%)

**GC Content by Lineage**
- Proteobacteria

**GC Content by Source**
- clinical
- clinical potential
- environmental
- environmental-animal
- environmental-food
- environmental-human
- environmental-plant and algae
- environmental-soil
- ND
- veterinary

**Number of Plasmids**
- 0 to 10:
  - 0.40−0.50
  - 0.50−0.60
- 10 to 20:
  - 0.40−0.50
  - 0.50−0.60
- 20 to 30:
  - 0.40−0.50
  - 0.50−0.60
- 30 to 40:
  - 0.40−0.50
  - 0.50−0.60
- 40 to 50:
  - 0.40−0.50
  - 0.50−0.60

**Number of Plasmids by Year**
- <2011 (6.52%)
- <2012 (4.35%)
- <2013 (5.43%)
- <2014 (8.7%)
- 2002 (4.35%)
- 2010 (7.61%)
- 2011 (4.35%)
- 2013 (8.7%)
- Others (50%)
**GC Content by Lineage**

Plasmid GC Content by Lineage:
- **Proteobacteria**

**GC Content by Source**

Plasmid GC Content by Source:
- **clinical**
- **clinical potential**
- **environmental**
- **environmental-animal**
- **environmental-estuarine**
- **environmental-food**
- **environmental-human**
- **environmental-plant**
- **environmental-soil**
- **environmental-wastewater**
- **ND**
- **veterinary**

**Size by Lineage**

Plasmid Size (kb) by Lineage:
- **Proteobacteria**

**Size by Source**

Plasmid Size (kb) by Source:
For more detailed results and interactive downloadable plots similar to those above, please visit the Interactive Plots tab in the app, which allows you to filter tables according to any MOB Group, as well as the source environment, GC Content percentage, size (kb) and country/region of origin.
MOB Group

Your sequence had similarities with the following MOB Group(s): MOBP

In the plasmid database, there are 597 MOBP plasmids (13.884%). The three most common countries in which this result appears are: ND, USA, Japan. The three most common entries for the collection date are <2012, <2011, <2010, the main environmental source for that group is clinical and the average GC Content and size are 0.4680511 and 59837.491, respectively. Lastly, the two most frequent phyla are Proteobacteria and Firmicutes, with 399 and 110 entries, respectively and the typical type of putative transferability for the MOBP group is mob (see definition below) and the most repeated Inc Group is -.

For more detailed results and interactive downloadable plots similar to those above, please visit the Interactive Plots tab in the app, which allows you to filter tables according to any MOB Group, as well as the source environment, GC Content percentage, size (kb) and country/region of origin.

Putative Transferability

In terms of transferability, a plasmid can belong to one of 6 categories:

For a plasmid without an assigned MOB Group (Without relaxase):
- Non-mob: non transmissible plasmids.
- Non-mob, protein export?: plasmids that contain T4CP, VirB4, MPF (or any two of the three), being potentially involved in protein exportation.

For a plasmid with an assigned MOB Group (With relaxase):
- Mob: mobilizable plasmids that code for a relaxase, but are lacking VirB4 and MPF.
- Determined conj: Conjugative plasmids that code for a known type of T4SS, plus relaxase and T4CP.
- Undetermined conj: Conjugative plasmids that contain a relaxase, T4CP and VirB4 but an unknown MPF.
- Unknown: plasmids that do not fall into any of the other categories.

For your sequence, the following table was obtained:

<table>
<thead>
<tr>
<th>Accession</th>
<th>MOB_Group</th>
<th>MPF</th>
<th>T4CP</th>
<th>VirB4</th>
<th>Putative_Transferability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_002122</td>
<td>MOBP</td>
<td>MPF_I</td>
<td>T4CP_I</td>
<td>VirB4_I</td>
<td>determined conjugative</td>
</tr>
</tbody>
</table>

In the plasmid database, there are 233 plasmids classified as determined conjugative (5.42 %). The most common MOB group associated with this result is MOBF and the most repeated Inc Group is IncFII(RepA4).

For the determined conjugative plasmids, the most common phylum is Proteobacteria, the main environmental source is clinical, the three most seen countries are ND, USA, China, the three most common entries for the year of isolation are 2010, <2013, 2013. Finally, the average Size and GC Content percentage are, respectively, 132601.71 and 0.5165.

To create downloadable interactive plots, please visit the Interactive Plots tab in the app, which allows you to filter tables according to any type of putative transferability, as well combining it with the source environment, GC Content percentage, size (kb) and country/region of origin.