Investigation into plasmid replication and partitioning systems in proteobacteria

Nayoung Kim

PhD

Department of Biology

University of York

September 2012

Abstract

Thanks to the development of numerous sequencing projects, a large volume of publicly available bacterial genomic information has accumulated. Relatively little, however, has been published regarding plasmid backbone systems based on public genomes. More specifically, a comprehensive analysis of plasmid replication and partitioning systems that would investigate their distribution and evolutionary history across bacteria is lacking. This thesis firstly developed a database for the plasmid backbone systems in proteobacteria. Using the database as a platform, eight major families of plasmid replication systems and five major families of plasmid partitioning systems, both of which are abundant in proteobacteria, were identified and a phylogenetic analysis for each family was performed. For the replication systems in proteobacteria, it was found that most plasmids do not show a wide host range, especially over class or phylum level, except for those that are already well known as broad host range plasmids, which, nevertheless, have a variety of replication systems. Plasmids, however, have been shown to move at least between related hosts, particularly within an order or class level. Regarding partitioning systems, four discrete ParA types of Type I class in partitioning modules were characterized according to their partner protein ParB. Based on the phylogenetic results, partitioning systems are also restricted to the class level. The members of each type do not seem to move outside the class level, except in the case of broad host range plasmids. This thesis has suggested that Rep initiators can be a good marker for classification. In particular, their phylogeny might be a reliable indicator of incompatibility between plasmids. Partitioning modules are a factor in the analysis, but it has been verified that in some cases the Rep and Par cannot identify the incompatibility groups. Through the case study of Rep and Par systems in 72 *Rhizobium leguminosarum* strains, this thesis demonstrated that although there is no movement of large plasmids between species, there are some cases indicating the possible horizontal transfer for relatively small plasmids between species. Finally, it was observed that there are cases of movement between strains, which might act as a vehicle for specific accessory genes, or might reflect a variety of intracellular recombination.

Table of Contents

List of Figures

Chapter 6

Appendix 1

Appendix 2

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Peter Young. It has been an honour working with him and this Ph.D. thesis would not have been realized had it not been for his time and ideas. His enthusiasm kept me focused on continuing my research as a biologist in the future. I would also like to give many thanks to my thesis advisory panel, Prof. Neil Bruce and Dr. Peter Ashton for their sound advice and meaningful questions during each meeting, which made my work more exciting and significant.

The members of the J1 group have contributed to my personal and professional development at the University of York. I am especially grateful to Peter Harrison and Ryan Lower who included me in their 'chromid' publication in 2009. I would also like to give many thanks to the past and present members of J1 who have helped me through the past four years: Jaikoo Lee, Jasper Green, Muhammad Shakeel, Ganesh Lad, Nitin Kumar, Kailin Hui, Piyachat Udomwong and the numerous visiting and undergraduate students who came through the lab.

Lastly, I would like to thank my family and friends for all their love and encouragement. To my parents, Dr Kyuhyun Kim and Mrs Dongwoo Kim, I owe many thanks for raising me and giving me their unconditional support in all of my pursuits. I would also like to give many thanks to my brother Dr. Hyungsuk Kim who always supports me. I am also grateful for the many Korean friends whom I have met in York, especially my wonderful flatmates: Sookyung Lee, Yoojung Kim and Yoonji Jung for their understanding and faithful support, as well as their patience. I must also thank my loyal friends in Korea for their continuous support: Se-eun Lim, Jinhee Kim, Hyeyoung Jang, Sunghee Park, Hyunjin Jung, Sanghyeb Lee, Yerim Huh, Daniel Oh and Soyoung Lee.

Nayoung K. D. Kim

Author's declaration

All work presented in this thesis is my own unless otherwise stated. Part of the work described in Chapter 3 and 4 has been published in the following research article:

Harrison et al. "Introducing the bacterial 'chromid': not a chromosome, not a plasmid." Trends in Microbiology**18**(4): 141-148

Chapter 1. Introduction

Recent developments in sequencing technology not only reduce the cost of sequencing projects, but also make available thousands of genome sequences in a relatively short time. Although the sequence data published so far might have been biased because many genomes are studied for their potential medical relevance, much more wide-ranging research has become possible now thanks to the technology, which has had a great effect on evolutionary microbiology. The availability of multiple sequenced isolates of one species through a fast and accurate method has also made it possible to map out those genomes in ways not previously possible, which has generated a variety of concepts describing the genomes.

'Pan-genome', which was introduced by Tettelin et al. [1] is one of the popularly used terms these days. In the study of 8 strains of *Streptococcus agalactiae*, they defined the pan-genome consisting of three parts: i) core genes that are shared by all the strains, ii) dispensable genes that shared by some but not all the strains, and iii) strain-specific gene that are present in only one strain and absent from all the others. Young et al. [2] also pointed out the distinction between 'core' and 'accessory' genes with a historical review of those terms. In their study, the core genes are essential, present in every genome and have a higher G+C composition, while the accessory genes have a long-term relationship with a bacterial species but provide adaptation to specific niches, and do not always show core-like composition. According to the research based on multiple isolates of *Escherichia coli* and *Staphylococcus aureus*, the genes that are not conserved in all strains of the same species comprise approximately 20~30% of each genome [3], which is not a small portion of the whole genome.

The research into core and accessory (dispensable) genomes has gained much more attention from biologists, particularly because the accessory component of

the genomes is significant for solving the evolutionary questions, in terms of bacterial adaptation. Generally such genes are located in mobile genetic elements (MGEs). Although a variety of recent comparative analyses have contributed to understanding of the MGEs, there is still much more room to improve our understanding of the accessory genomes in bacteria. This introductory chapter firstly starts to introduce MGEs as important agents of horizontal gene transfer (HGT), and secondly also describes the 'plasmid', one of the important classes of MGE and the main theme of this study. Finally, aims and objectives will be set out to open up the subsequent chapters in this thesis.

1.1 Mobile genetic elements and plasmids

1.1.1 Advent of mobile genetic elements

Mobile Genetic Elements refer to any segments of DNA that are able to move within cell (intracellular mobility), or between different cells (intercellular mobility) [4, 5]. In the 1940s, Barbarda McClintock firstly described 'jumping genes', which are later termed as 'Transposable Elements (TEs)'. When she studied the maize genome, she thought that some genes seemed unstable and they might change their 'position within a single cell. Her research is meaningful because it is suggested that genomes of an organism are not stationary, but possible to be altered freely. TEsare considered as one class of MGEs, along with plasmids, (conjugative) transposons, integrons, genomic islands and bacteriophages.

The discovery of TE not only won McClintock the Nobel Prize, but also allowed MGEs to be studied as important agents in bacteria in the next half century, which led to effects on a variety of fields. The fact that MGEs usually confer many beneficial traits to their host including antibiotic resistance, detoxifying elements and enzymes for secondary metabolism, is significant for humans, because some of those traits affect serious problems in the clinical field, such as infections and disease. On the other hands, MGEs play a crucial role in the plasticity of the genome, which specifically allows prokaryotes to adjust to new environmental

niches. This makes them a key element in understanding ecology.

1.1.2 A variety of mobile genetic elements

1.1.2.1 Plasmids

Plasmids are self-replicating DNA molecules existing in cells as extra-chromosomal replicons. Plasmids were first identified by the American geneticist Joshua Lederberg [6]. He identified the F-factor, which allowed recombination in *Escherichia coli* by promoting conjugation. In 1952, Lederberg suggested the term "plasmid" to refer to "extranuclear" chromosomes, specifically for the R-factor (that was identified by Frederic Griffith in 1928) and the F-factor. Since then, plasmid biology has fast become one of the major research fields in biology.

Plasmids occur in prokaryotes and sometimes in eukaryotic organisms such as yeast *Saccharomyces cerevisiae*. The most studied plasmids are circular double-stranded DNA, some of which are able to insert themselves into chromosomes; linear plasmids, however, exist as well. The size of plasmids varies, generally between 1 kb to 500 kb, although many larger plasmids exist that are over 1 Mb. It has been known that most HGT has taken place by mid-size plasmids (30-300 kb) in general. The copy number of plasmids also varies from 1 in a single cell to thousands of copies.

Plasmids contribute to a flexible gene pool, resulting from the highly frequent gene acquisition and loss [7]. It has become known that plasmids often carry a variety of antibiotic resistance genes, which might be harmful for humans. Recently, however, a large number of plasmids that do not negatively affect human beings have also been discovered thanks to the proliferation of genome sequencing projects. A variety of elements captured by plasmid backbones are able to increase genetic diversity, which provides their hosts with additional functions in order to adapt in different environmental niches [8, 9]. Many naturally occurring plasmids have evolved as an integral part of the bacterial genomes, serving as a helper to their hosts [10].

1.1.2.2. Other mobile genetic elements

Bacteriophages Bacteriophages provide the transfer vessel and the mechanism for the packaging and delivery of genetic information by transduction [11]. They are often classified into two types: the virulent class, which always kills the host cell, and the temperate class, which may kill it but can alternatively integrate with the chromosome and reside in the host cell [12]. Phages are distributed very widely from the soil to the intestines of animals. The size of bacteriophage genomes ranges between 3 and 500kb [13].

IS elements IS elements contains only inverted repeats and a gene coding for transposase (**Figure 1.1**), which catalyzes the cutting and resealing of DNA that occurs in transposition and recognition sites [14]. The elements are relatively short and genetically compact, encoding no functions (phenotypes) other than those involved in their mobility, but they can affect on gene inactivation, expression or arrangement [15]. They are widely distributed in both eukaryotic and prokaryotic genomes and their range in size is from 0.8 to 2.5 kb.

Transposons Bacterial transposons generally include both IS elements and composite transposons, however, composite transposons are typically considered as 'transposons in practice. Transposons are the elements that can jump into different chromosomal localizations [16], which indicates that they move from one site to another on the same chromosome or to other chromosomes or plasmids. For this, they were firstly called 'jumping genes' as seen in section 1.1.1. Transposons contains two copies of the same IS elements and one or several genes functioning variously such as antibiotic resistance genes.

Genomic Islands Genomic Islands are defined as large linear chromosomal regions (10 – 200kb) that are part of the flexible gene pool, carrying one or more genes that can increase the adaptability and versatility of bacteria [17]. GIs are frequently associated with *tRNA* genes and contain mobility or conjugation genes that code for the integrases or transposases required for chromosomal integration and excision [18]. GIs are able to code various functions including symbiosis or pathogenesis, and can be divided into several sub-classes, such as pathogenicity islands (PAIs) or antibiotic resistance islands.

Integrons Integrons refer to a genetic system that allows bacteria to capture and express gene cassettes [19]. Typically, they are composed of three elements: an *intI*

gene, a recombination site *attI*, and a promoter. *intI* encodes an integrase that catalyses the incorporation or excision of gene cassettes by site-specific recombination. A promoter is responsible for the expression of inserted gene cassettes. Integrons exist in many forms that differ in number and identity of captured genes. As integrons can be located in transposons and in conjugative plasmids, when they catalyze movement of host genes into themselves, the overall result is movement of genes into transposons and conjugative plasmids [9].

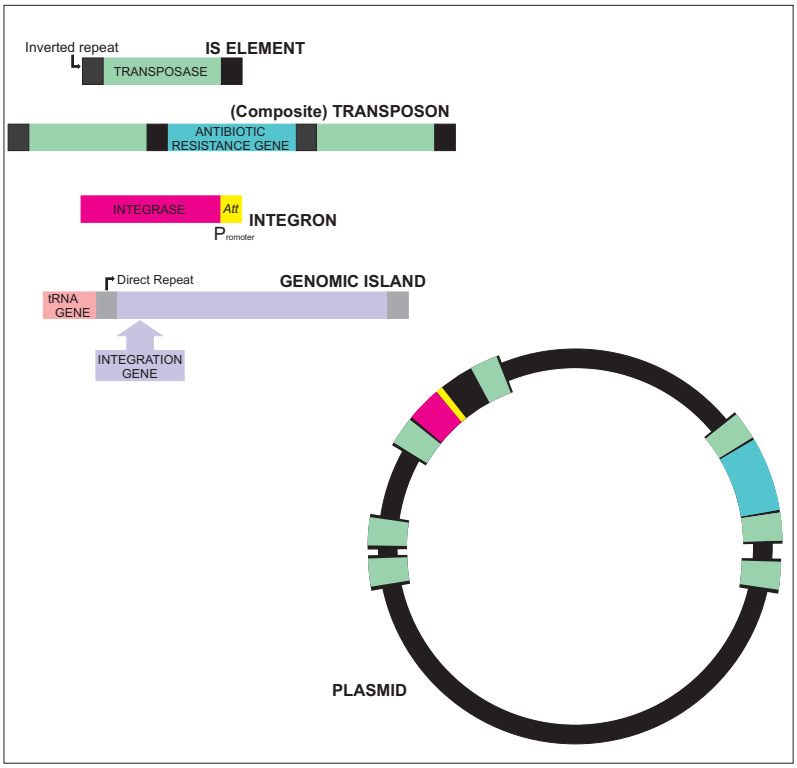


Figure 1.1 A variety of mobile genetic elements in prokaryotes

Mobile genetic elements (MGEs) refer to any segments of DNA that are able to move within cell or between different cells. MGEs include: Insertion sequence (IS) elements containing inverted repeats and a gene coding for transposases, transposons including both IS elements and composite transposons, genomic islands (GIs) containing mobility or conjugation genes that code for the integrases or transposases, and Integrons being composed of an *intI* gene, a recombination site *attI*, and a promoter.

1.1.3 Mobile genetic elements as agents of horizontal gene transfer

1.1.3.1 Horizontal gene transfer in prokaryotes

A gene, or a set of genes, located in prokaryotes or eukaryotes can be altered by a variety of phenomena. Such phenomena are considered significant because they are responsible for the huge diversity extant not only within species, but also across entire prokaryotic or eukaryotic populations. Biologists have discovered that genes are able to change via point mutation, insertions, and deletions, affecting one or a number of nucleotides [20]. On some occasions genes are changed via events such as segmental duplication, interstitial deletions, or chromosomal translocations [21-23]. Numerous past researches in biological evolution relatively more seemed to focus on how populations responded to mutations, what mechanisms were involved in this process, or how the mutation proceeded within the cell, and so on.

However, it has been argued that genes can also be transferred between different organisms [6, 24]. This is referred to as Horizontal Gene Transfer (HGT), also known as Lateral Gene Transfer (LGT). HGT is defined as the mechanism through which one organism acquires genetic materials and information from another species. This phenomenon contrasts to the vertical gene transfer of genetic information from parent to offspring. Evolution by mutation can be slow relative to acquisition by HGT. It has been discovered that during bacterial evolution, the ability to adapt to new environments most often ensues from the acquisition of new genes through horizontal transfer, rather than by the alteration of gene functions through the numerous mutations mentioned above [25]. HGT allows the recipient to build on its unique, pre-existing adaptations and invade a new niche or improve its performance in its current environment [26]. It is estimated that between 1.6 and 32.6 per cent of the genes of each prokaryotic genome has been acquired through horizontal gene transfer [27].

Ongoing research in the study of HGT has revealed that HGT could occur in all domains and in all directions [28]: from Bacteria to Archaea [29], from Archaea to

Bacteria [30], from Archaea to Eukarya [31], from Bacteria to Eukarya [32], from Eukarya to Bacteria [33] and even within Eukarya [34]. Horizontally transferred genes confer a variety of beneficial, as well as occasionally negative, effects on their host genomes. Although it was previously believed that prokaryotic HGT occurred more often than eukaryotic HGT, the detection of the latter is gradually beginning to be reported.

1.1.3.2 Main mechanisms of horizontal gene transfer

Understanding the mechanisms of HGT is crucial to understanding of microbial evolution. Generally, HGT in prokaryotes can be the result of three mechanisms: transduction, transformation, and conjugation. In the first place, transduction is a process mediated by bacterial viruses (bacteriophages). As the transfer of DNA by transduction requires that the donor and recipient share cell surface receptors for phage binding, this mechanism is usually limited to closely related bacteria. The size of the phage head, temperature and pH are also related to the phenomenon [35, 36].

Secondly, transformation can be defined as the uptake or integration of naked DNA and usually mediates in the exchange of any part of a chromosome. This process is most common in bacteria as opposed to eukaryotes, and only short DNA fragments are typically exchanged. Natural transformation occurs when cells enter into a transient physiological state called 'competent'. Naked DNA from extracellular environment can be chromosomal DNA, plasmid DNA, or viral DNA and is derived from dead prokaryotic, eukaryotic cells or viral particles [37]. The size of uptake DNA can range from a few hundred nucleotides to more than 55,000 nts.

Finally, the conjugative mechanism mediates the exchange of mobile genetic elements. Conjugation is believed to be the most important mechanism responsible for short-term bacterial adaptation [9], as it can transfer genetic material even between phylogenetically unrelated organisms [36, 38]. Conjugation is the process of DNA transfer from a donor to a recipient cell through cell-to-cell contact. The DNA transferred by conjugation can include conjugative plasmids, conjugative transposons, or integrative and conjugative elements (ICEs).

1.2 A dive into the plasmids

1.2.1 Plasmid backbone modules

1.2.1.1 Replication systems

A replicon is a region of DNA that is replicated from a single origin of replicion. It could potentially be a hole bacterial chromosome or plasmid. A minimal replicon comprises just the essential elemnts for replication, i.e. the Rep systems (**Figure 1.2**). Regarding plasmids, the first essential gene existing on plasmids is a *rep* gene, which is responsible for the initiation of plasmid replication. The role of the *rep* gene is to ensure that the replication is balanced with the host cell growth cycle. Failure to be so might burden the host cell [39]. Consequently, the *rep* gene is often coupled with the *cop* gene, in order to control both replication and copy number. More details will be described in chapter 3.

1.2.1.2 Maintenance and stability systems

In addition to controlling replication and copy number, low-copy-number plasmids in particular need additional modules for their maintenance and stability [39, 40]. There are several mechanisms that have been linked with plasmid stability. Multimer resolution system (*mrs*) is one of the better-studied systems for plasmid maintenance. Many plasmids contain a site-specific recombinase or resolvase in order to enable multimer resolution. This is important because the plasmid multimer can interfere with the appropriate segregation of plasmids into daughter cells [40, 41], which decreases their stability. Post-Segregational Killing (PSK) systems or addiction systems [42] are also effective methods to get rid of potential intercellular competition and, thus, contribute to plasmid maintenance (see more details in chapter 4). The *hok/sok* systems of plasmid R1 is a well-known example of a PSK system and prevents the translation of toxin mRNA, which generates the Hok protein that kills the host cell [43]. Finally, active partitioning systems are the

most significant mechanisms for plasmid stability. Partitioning systems are required in order to ensure that plasmids are actively moved into the proper position prior to cell division. This process is controlled by two coupled partitioning genes (*par*). Plasmid partitioning systems can be classified according to the cytoskeletal components that they encode [39]. These will be discussed in more detail in chapter 4.

1.2.1.3 Transfer systems

Modules affecting plasmid propagation are *tra* genes, which consist of DNA transfer and replication (Dtr) and Mating Pair Formation (MPF) [9]. It has been shown that self-transmissible plasmids are able to transfer unaided when they contain both Dtr and MPF components, while mobilizable plasmids are only able to transfer in the presence of a self-transmissible plasmid when they possess Dtr components [44]. A recent study [45] performed a comprehensive analysis of the transfer systems, which is very useful to this study. We will discuss this in chapters 5 and 7.

1.2.2 Plasmid accessory modules

One of the reasons why plasmids were studied extensively in the first place was their ability to convey genes that can sometimes have a considerable effect on human life and bacterial evolution [46]. Accessory modules generally include genes that confer specific phenotypic characteristics to the host. The earliest described accessory function of plasmids was antibiotic resistance. After a few decades, bacterial genome sequencing accelerated the discovery of other types of accessory modules carried by plasmids [47]. The operative elements are indeed diverse, including symbiosis genes that contribute to nitrogen fixation, virulence factors that might facilitate colonization of eukaryotic cells, or even mechanisms for detoxification of heavy metal substances (**Table 1.1**). In most cases, these modules have significant differences in their G+C content from that of the plasmid

backbone [9].

Table 1.1 Example of accessory modules on plasmids

Types	Pathogenicity	Resistance	Metabolism
Example	Toxins, colonization factors	Antibiotic resistance, metal resistance	Nitrogen fixation, photosynthesis, xenobiotic compounds

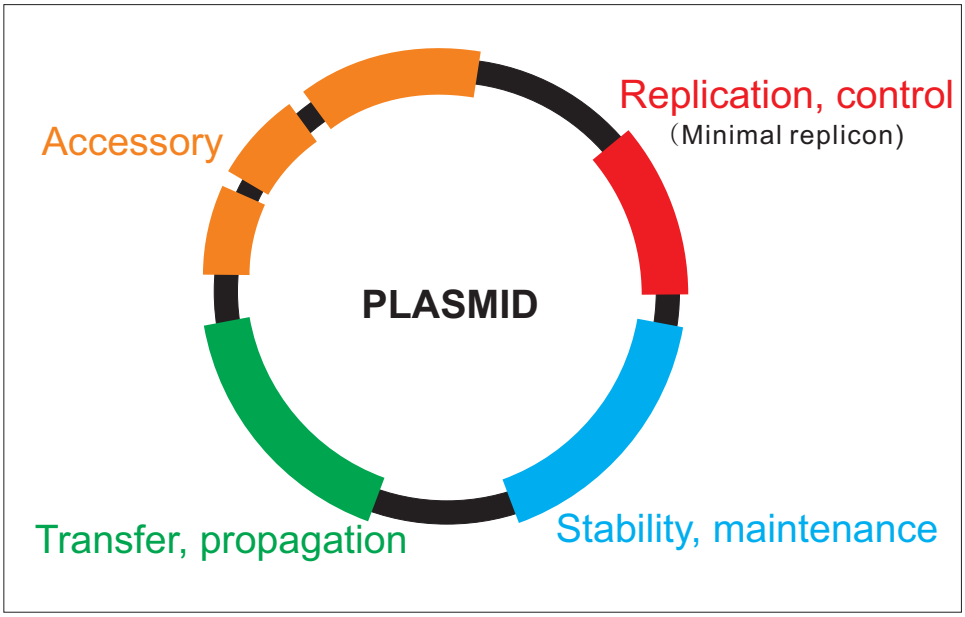


Figure 1.2 The organization of a plasmid consisting of backbone and accessory modules

Modules can be divided into four parts: Replication and control (also defined as a 'minimal replicon'), stability and maintenance, transfer and propagation and accessory modules. Taken from Norman et al. [48].

1.3 Aims and objectives of the thesis

In section 1.1 and 1.2, we have introduced basic knowledge about MGEs and plasmid biology. The overall aim of this thesis is to investigate the main backbone

systems of plasmids, and in particular, the plasmid replication and partitioning systems in proteobacteria. More specifically, the study will focus on the following questions:

I. Would it be possible to collect homologous genes involved in plasmid backbone functions effectively and to make a database for the genes?
II. How can the genes collected above be classified, and be used as an indicator of classification of the plasmids?
III. How are the gene families distributed, and what are their host range of each family,
IV. How can we infer the evolutionary history of the systems based on their phylogenetic tree and distribution?
V. What can we learn from a case study to research one of the typical plasmid backbone systems in proteobacterial plasmids?

To answer these mission statements, each chapter has specific objectives described below.

As a method chapter, we will first develop a database (chapter 2) based on the amino acid sequences responsible for the two systems, where homologous genes will be stored into discrete types. In this chapter, we aim to

- Constitute a comprehensive research of the NCBI complete genome database, and construct an in-house database for two systems and a web-server.
- Make the database to be (a) a research tool to browse patterns in the taxonomic distribution and phylogeny of various components and, (b) an annotation tool for newly sequenced bacterial genomes, providing reliable information to biologists.

In chapter 3, we will investigate the plasmid replication systems in proteobacterial plasmids in detail. We will

- Identify and classify gene families involved in replication systems and propose a classification scheme. Review the general genetic organization of the main families of Rep proteins that we have defined.

- Investigate diversity and distribution of each family across proteobacteria.
- Perform a phylogenetic analysis of each type in order to research evolutionary history.

In chapter 4, among the several mechanisms identified for plasmid maintenance, we will focus on the plasmid partitioning systems. Following the same steps as in chapter 3, we aim to

- Identify and classify gene families involved in partitioning systems.
- Investigate diversity and distribution of each family across proteobacteria.
- Perform a phylogenetic analysis of each type in order to research evolutionary history.

Based on previous research regarding the genetic organization of the two systems and their phylogenetic analysis, the plasmid replication and partitioning systems appear to be closely related, which might give us a hint for the plasmids' evolutionary history. In chapter 5, we will

- Compare the distribution and phylogenies of each family of two systems presented in previous chapters (chapters 3 and 4)
- Identify the specific elements of the systems that have possibly evolved together, or have resulted in recombination events, or horizontal gene transfer.

In chapter 6, we will

- Perform a case study in order to investigate the distribution of one of the typical replication and partitioning systems in bacteria, the RepABC operon.
- Search and map our 454 reads of 72 *Rhizobium leguminosarum* strains from the Wentworth College in York against a list of reference genomes in order to find replication and partitioning regions.
- Construct a phylogenetic tree, based on the alignment of all the RepABC operons, in order to research the general sequence variation within each type.
- Look into "magnified" phylogenies within each type to investigate the amount of strain-to-strain spread of the plasmids.

Finally, we will close our thesis with general discussion. The main contribution to understanding plasmid biology will be discussed and how this study could be used for further research.

Chapter 2. A database for the replication and partitioning systems in proteobacterial plasmids

Over the last a couple of decades, various experimental studies have investigated the accessory modules in bacterial plasmids. The development of genome sequencing technology has clearly made it much easier to retrieve the target sequence information of specific modules; however, no database has been developed so far to investigate plasmid backbone systems comprehensively. In this chapter, we will introduce a new in-house database and web-interface that focuses on plasmid replication and partitioning systems. By constructing a database to investigate backbone modules in proteobacterial plasmids, this chapter aims to the comprehensive research of plasmid backbone systems in proteobacteria based on the NCBI complete genomes. This will form the foundation for the phylogenetic analysis of the systems described in chapter 3 and 4.

2. 1 Background

2.1.1 Genome sequencing projects

Genome sequencing projects are constantly identifying complete sequences of specific organisms, belonging from bacteria to human. There are numerous sequencing labs performing next generation sequencing around the world. For example, Illumina Hiseq can currently (2012) provide 120GB per 27 hours on a single sequencer and there is no doubt that innovation will continue to spring up as time goes by. As full genome sequencing projects provide raw data, the number of repositories for storing that data is also exponentially growing [49]. The successful increase of sequencing data seems to be very promising in solving

clinical problems that human face with. Increasing the quantity of biological information, however, does not guarantee the quality of the ensuing analysis. Indeed, it has become more challenging to manage and interpret the data. Comparative genomics is one method employed to gain insight into many aspects of the evolution of modern species based on the ever-increasing mass of biological data. There are more ways, however, to handle the data and analyse them effectively. There is a definite need to develop more tools that extract information and summarize it.

2.1.2 Mobile genetic elements databases

In MGE biology, there have been many attempts to effectively manage and analyze MGEs based on their own database, as a tremendous number of MGE sequences have been published recently. In particular, it has been made clear that genes in MGEs are responsible for a variety of interesting phenotypes and contribute to bacterial HGT, to which researchers have increasing been paying more attention in the last few decades. The databases developed so far have advantages and disadvantages depending on their particular purpose. What follows is a review of some of the most comprehensive databases targeting MGEs (**Table 2.1**).

Regarding plasmids, the Plasmid Genome Database (PGD) [50], was developed in the early 2000s. The PGD includes a list of fully sequenced plasmids and the structural maps of each of these. Although the PGD only contained 500 plasmid genomes at first, (and this number is very small compared to today's data), this was one of the first attempts to manage plasmid sequences. Recently, this database has been incorporated into ACLAME (see below). Also, although not used anymore, the Database of Plasmid Replicons (DPR) by Osborn (web-based, but unpublished and no longer available) was one of the early attempts to provide a (non-comprehensive) list of plasmid sequences, their sequence alignments and phylogenies. The information it stored, however, was very limited, as it was constructed before genome sequencing technology was made widely available.

Other approaches to investigate MGEs include Islandpath [51], a web-based interface to display island-associated features. Its unique characteristic is that it is

not only a collection of sequences, but is also able to create visualizations for the analysis of genomic islands. The downside of Islandpath is that it only concentrates on pathogenicity islands, rendering the information shown very limited. For other MGEs, ISfinder (Insertion Sequence Finder) [52] is a database that includes bacterial Insertion Sequences. It provides individual files containing general features of ISs (name, size, origin, etc.) as well as DNA and protein sequences. ISfinder interacts with other databases, which were also incorporated into ACLAME.

ACLAME (A CLAssification of Mobile genetic Elements) [4] is the most recent MGE database that includes proteins coded by mobile elements. The goal of this database is to identify functional modules in MGEs and to classify the available newly sequenced ISs, transposons, plasmids, viruses and phages. At present, it provides a general classification and nomenclature (of mostly plasmids and phages) by clustering proteins using all-against-all BLAST and the Markov algorithm. This is a mostly on-going project contributing to the research of MGEs. The way to classify protein sequences in order to define the families of the sequences, however, is based on automatic BLAST, which might not give enough accuracy required for the study of evolution. Moreover, too many families with only a small number of protein members are included, which are difficult to extract meaningful information by the researchers.

In addition to individual databases for MGEs, there are several well-known projects that provide information on MGEs. These are all databases for classifying proteins on sequences genomes, which include certain protein families that are involved in MGEs functions. COG (Clusters of Orthologous Groups of proteins) [53], composed by all-against-all BLAST alignments of protein sequences from microbial genomes, attempts to classify proteins by encoding complete genomes phylogenetically. Of course, various genes on MGEs are included therein. NCBI Protein Clusters [54] use similar concepts as COG in order to classify proteins. Pfam [55] also contains MGE information through the use of HMM based algorithms. The advantage of this type of database is that it is continuously updated, with newly sequenced genomes. As all of the above, however, are not developed to annotate MGEs, they might provide incorrect information on account of their automatic pipeline for core gene functions (see further below).

Table 2.1 Web resources and databases mentioned in this thesis

Database	Sequences	Link	Reference
PGD	Plasmids	No longer available	Molbak et al. [56]
DPG	Plasmids	No longer available	Unpublished
IslandPath	Pathogenicity islands	http://www.pathogenomics.sfu.ca/islandpath/	Hsiao et al. [57]
ISfinder	IS elements	http://www-is.biotoul.fr/	Siguier et al. [58]
ACLAME	Plasmids and phages	http://aclame.ulb.ac.be/	Leplae et al. [59]
COG	All	http://www.ncbi.nlm.nih.gov/COG/	Tatusov et al. [60]
NCBI Protein Cluster	All	http://www.ncbi.nlm.nih.gov/proteinclusters	Klimke et al. [61]
Pfam	All	http://pfam.sanger.ac.uk/	Punta et al. [55]

2.1.3 Mobile genetic elements databases and their difficulties

Despite the bulk of previous literature on MGE evolution and diversity, it is a fact that relatively little attention has been paid to them, and in particular MGE bioinformatics, by the research community so far [5]. This may be the result of the small number of independent genome projects in the past or from a lack of interest in bacterial accessory genomes. Having said that, recent attempts in the last twenty years, such as the construction of databases for different types of MGEs mentioned above and the development of annotation tools for MGEs based on next-generation sequencing projects, constitute significant contributions to MGE biology.

Additional work is required, however, towards a more comprehensive research into MGEs. Frost et al. [5] have highlighted two main challenges in the study of MGEs. Firstly, they point out poor annotation of MGEs, which becomes particularly apparent when the research of MGEs is part of a whole genome project. For example, in the case of databases that were made for gene identification, the

properties of MGEs, such as the very different GC content and codon preferences from chromosomal genes, may result in the construction of ineffective training sets, thus inhibiting accurate predictions [5]. Most automated gene predictions are appropriate for chromosomal sequences and not for MGEs. This is the reason why MGE annotation has to be performed manually. As a result, automatic annotation tools for MGEs are still far from sufficient, which consequently makes it hard to handle the ever-increasing genomic information from the next generation sequencing technology.

Frost et al. have also argued for the need to establish standard formats for MGEs, including the nomenclature and ontology that will be used by all biologists. A confusing nomenclature makes it hard to classify and investigate MGEs, particularly since there is no universally agreed system for naturally occurring plasmids and transposons, as well as no central nomenclature authority for naming newly discovered MGEs. Despite the efforts to overcome these difficulties through ISfinder and ACLAME, it is still challenging to link and apply them to standard databases for MGEs. Well organized chromosomal and MGE databases will unequivocally generate more powerful research tools for the more efficient study of genomics.

2.2 Aims and objectives

In this chapter, a database for plasmid backbone systems in proteobacteria, which form the foundation for investigation in chapters 3, 4, and 5, will be constructed. As all the details of each family of plasmid backbone systems will be discussed in detail in the individual chapters, here we will demonstrate the purpose of the database, and the methods that we have used to collect data that are stored in the database. We will also present the statistics that originate from the data and describe how they will be presented through the web-interface. What follows constitutes a discussion of the three main objectives of the database.

2.2.1 Concentration on key backbone genes

The database concentrates on key backbone genes in plasmids. As mentioned in chapter 1, the genes on plasmids can be divided into two types: backbone and accessory genes. Accessory genes confer certain phenotypes on the host bacterium, such as symbiosis, antibiotic resistance, production of toxin compounds, virulence, and so on. These elements usually enable host prokaryotes to succeed within specific environmental niches. Conversely, backbone genes are related to plasmid replication, maintenance, stability and mobility. Examples include replication, partitioning and conjugative transfer genes. In this study, we aim to contribute to the analysis of plasmid backbone genes, and in particular, plasmid replication and partitioning systems.

2.2.2 Comprehensive coverage of bacterial genomes

The main point of the database is to provide information on the general distribution of notable gene families based on all proteobacterial genomes and is not based on specific genus or family. Our database is built to conduct searches gene by gene, to yield homologous sequences from all complete genomes. It should be noted that the sequences used in this work were based on complete bacterial genomes (including plasmid genomes), which means some plasmids were not included as they are not part of a complete genome, although their backbone genes were deposited in NCBI. Moreover, the research performed in this thesis was based on the sequences published up to October 2011. We expect that a comprehensive research tool will allow us to study the diversity of plasmid backbone genes and possibly map the pathways of HGT. With full alignments, the phylogenetic analysis of the gene families can be investigated comprehensively.

2.2.3 Consistent annotation and future update

As mentioned previously, it is frequently the case that misleading annotations of MGEs including plasmids are encountered, which can cause confusion and hinder progress in the field of MGE research. For instance, we found that genes that are annotated as 'cobyrinic acid *a,c*-diamide synthase' in GenBank clearly have some relationship with partitioning proteins. No information, however, is available as to whether these belong to a large number of ParA proteins. We believe that our database can contribute to an accurate analysis of our target genes. For this purpose, we developed HMM profiles (see next section), which are also important for the future update of the database.

2.3 Data acquisition, structure of a database, and general statistics of data in the database

2.3.1 Database design

The database is implemented in a MySQL relational database running on a UNIX server and a Web Apache server. Perl DBI modules were used for queries and we also designed a web interface using the CGI modules of Perl. So far, there are five tables in the database and more will be added according to the expansion of various functions. The tables created in the database are explained in **Table 2.2** with more details about each inserted in **Table 2.2 (a)** to **(e)**.

Table 2.2 A list of tables created in the database

Details of each table are indicated from (a) to (e).

Tables	Contents
GENE	Basic information of each gene. Linked to GENE_FAMILY and GICONVERT.
GENE_FAMILY	Family IDs allocated to the gene investigated in this study. Possible to change or update the family type.
FAMILY	Collection of major families in proteobacteria.
GICONVERT	In order to obtain the replicon type and name, a gene ID is converted to a replicon accession number. Linked to GENE and P_LIST.
P_LIST	All taxonomic information based on NCBI, which is needed for studying distribution of gene families. Linked to GICONVERT.

(a) GENE

Field	Type	NULL	Entry example
geneacc	varchar(20)	NO	YP_471934.1
annotation	varchar(200)	YES	Plasmid replication protein RepCb
sequence	varchar(1000)	YES	ESGSVTTPFGRRP…MALVRTNSPIRGKTG
giacc	int(11)	NO	86360044

(b) GENE_FAMILY

Field	Type	NULL	Entry example
famannot	int(10)	NO	1
geneacc	varchar(20)	NO	YP_471934.1

(c) FAMILY

Field	Type	NULL	Entry example
famannot	int(10)	NO	1
famtype	varchar(20)	NO	Replication
famname	varchar(255)	NO	RepC (of RepABC)

(d) GI_CONVERT

Field	Type	NULL	Entry example
geneid	int(11)	NO	86360044
nugi	int(11)	NO	86359881
nucacc	varchar(20)	NO	NC_007763

(e) P_LIST

Field	Type	NULL	Entry example
taxaid	int(11)	No	435
nucacc	varchar(20)	No	NC_001275
refacc_A	int(11)	No	10955174
refacc_B	varchar(255)	No	AF110140
species	varchar(255)	No	Acetobacter aceti
replicon	varchar(255)	No	Plasmid pAC5
division	varchar(255)	No	Alphaproteobacteria

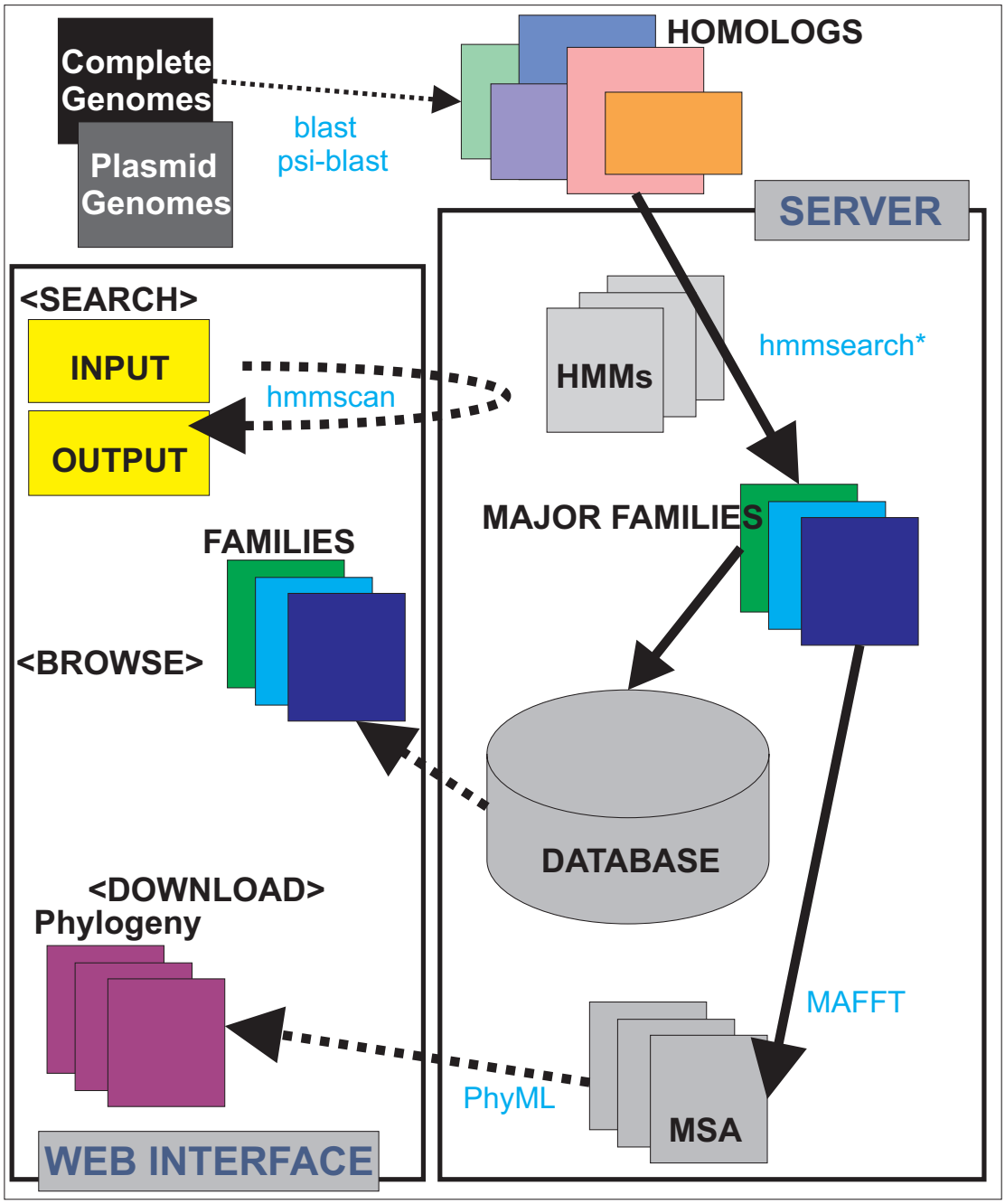


Figure 2.1 Data flow and filtering steps for the protein families defined

Homologous sequences are searched against NCBI complete genomes (including plasmid genomes) using blast and psi-blast. In-house HMMs filters the homologs again using hmmsearch (*), which are then collected in the database. Major families can be accessible by users through <Browse> menu, and users also can search their sequences using <Search> menu. The sequences of each family are aligned by MAFFT and used to generate phylogenetic trees by PhyML. All HMMs and phylogenies are stored as flat file for users to be accessed through <Download> menu.

2.3.3 Retrieval and identification of homologs

A database was developed based on complete genomes from NCBI, including plasmid genomes in 2009, in order to analyze the distribution of replication and partitioning system genes among all sequenced microbial genomes. In order to collect homolog sets of each type of gene, we ran blast and psi-blast using the representative well known genes as a seed (confirmed experimentally) of each family. Based on experience, 3-4 iterations were ideal for collecting the most likely homologs. In addition, we also used an HMM-based search using our in-house HMMs (see section 2.3.5), if more than one family was found per query sequence by psi-blast. For example, if one gene was retrieved twice by different types of families, then the gene was allocated to one family alone, depending on the result of hmmsearch with a better e-value based on the HMMs (**Figure 2.1** and see more in section 2.3.5).

2.3.4 Phylogenetic tree constructions

For each group of homologous protein sequences, we constructed an alignment using Multiple Alignment using Fast Fourier Transform (MAFFT) [62] with default parameters. MAFFT's G-INS-I algorithm uses an iterative refinement method. MAFFT is rapid and has performed accurately in algorithm comparisons. Then, the alignment was used to construct phylogenetic trees. Maximum-likelihood phylogenies were constructed using PhyML (version 3.0) [63]. For the automatic generation of phylogenetic trees, we use the LG model in PhyML. To assess validation in the data, 100 bootstrap replicates were generated from the data set. Details of the option in PhyML can be obtained from the web interface. Finally, TreeView [64] was used to visualize the phylogenetic trees. All the phylogenies can be accessed via our web-interface. In every tree (chapter 3, 4, and 5), thick lines indicate that each bootstrap value is over 70%. All original trees can be found in appendices at the end of the thesis.

2.3.5 Hidden markov models (HMMs) construction

Profile HMMs are the most popular modeling concept for searching conserved motifs in protein and are also used to classify new sequences into families based on domain architecture [65]. Profile HMMs are based on the attempt to find the most likely explanation for the observed variable. They model a family of sequences that is derived from a multiple alignment and then capture position-specific information concerning the level of residue conservation and the likelihood of each residue occurring at that position. Various studies have shown that profile-based methods are much more effective in detecting homologs than pairwise methods [65, 66], although there is limitation that profile-based methods are much slower.

One of the most important objectives of this project is to provide a reliable annotation for the identification of newly sequenced plasmid backbone systems. We have constructed profile HMMs for each family of genes in plasmid replication and partitioning systems. In this project, HMM profiles were modeled and calibrated using HMMER Version 2.3.2 [66]. **Figure 2.2** is one example of how the in-house HMMs classify protein sequences into different families having common motifs. If the query sequence is screened by HMMs, the profile HMM with the lowest e-value allocates the sequence into one of the matched families. If two HMMs give a matching result, only one HMM allocates it to the matching family. For example, if the sequences from SEQ1 to SEQ3 scan against each HMMs developed, SEQ1 matches to the IncC HMM with a relatively high e-value (1.6e-250), while it also matches to the ParA HMM (1e-97). In this case, the profile having lowest e-value allocates the sequence into the IncC family. This concept is used in the 'search' menu, as well in the web-interface, in order to classify newly sequenced genes into each family of Rep and Par systems.

(a) SEQ1, SEQ2, and SEQ3 against the IncC HMM

E-Value	score	bias	Sequence
1.6e-250	820.4	10.0	SEQ1 (gi|68500027|gb|AAY97970.1|)
1.2e-08	20.3	4.8	SEQ2 (gi|289823764|ref|ZP_06543376.1|)
1.2e-08	20.3	0.1	SEQ3 (gi|355533046|gb|EHH02388.1|)

(b) SEQ1, SEQ2, and SEQ3 against the ParA HMM

E-Value	score	bias	Sequence
4.4e-212	693.6	5.0	SEQ3 (gi|355533046|gb|EHH02388.1|)
1e-97	314.9	0.1	SEQ1 (gi|68500027|gb|AAY97970.1|)
1.7e-13	36.0	2.4	SEQ2 (gi|289823764|ref|ZP_06543376.1|)

Figure 2.2 Results of HMMSEARCH indicating how the HMMs determine the best category

(a) SEQ1, SEQ2, and SEQ3 against the IncC HMM, (b) SEQ1, SEQ2, and SEQ3 against the ParA HMM

2.3.6 General statistics of the database in this study

2.3.6.1 Plasmid replication systems

Table 2.3 shows the numbers of Rep sequences involved in replication systems in proteobacterial replicons. This actually includes many homologs not only of plasmids but also of chromids and chromosomes. Some families are only found in a specific division of proteobacteria, while others in more than one. For example, RepC families are only found in alphaproteobacterial replicons and RepFIIA, RepFIA, and RepFIB are mostly found in gammaproteobacterial replicons, while other are found in diverse divisions of proteobacteria. More details regarding distribution and phylogenetic analysis are described in chapter 3.

Table 2.3 Replication systems in proteobacterial plasmids and chromids

Family Accession	Replication protein	Division of proteobacteria				Total
		α	β	γ	others	
1	RepC	141	0	0	0	141
2	RepA-like	110	43	23	0	176
3	RepB-like	7	59	10	2	78
4	TrfA	1	32	26	1	60
5	RepA	32	25	15	31	103
7	RepFIA	0	0	114	0	114
8	RepFIB	0	0	106	0	106
9	RepFIIA	0	0	140	0	140

2.3.6.2 Plasmid partitioning systems

While all plasmids need replication genes in order to replicate by themselves, not all plasmids have partitioning systems. Unlike the low copy-number plasmids, which mostly possess the active partitioning systems, some plasmids have developed alternatives for their maintenance and stability, such as multimer resolution or PSK systems (see chapter 1.2.1). Therefore, the research of partitioning systems might not cover all the plasmids in proteobacteria generally. In this study, we have mainly collected data based on experimentally well-discovered partitioning systems As indicated in **Table 2.4 (a)** and **(b)**, the numbers

of NTPase and DNA-binding proteins are not same. There are several reasons for this, i) many ParA sequences are found twice in one replicon due to duplication, ii) there is an in-frame translation site in some plasmids' *par* genes, which generates more than one Par protein, or iii) there are many plasmids that lack binding proteins in their partitioning system. Also, it should be noted that we have not included ParG type that is a partner of ParF, because there is no sequence homology between ParGs. Further details for each family can be found in chapter 4.

Table2. 4 Partitioning systems in proteobacterial plasmids

(a) NTPases

Family Accession	Replication system	Division of proteobacteria				Total
		α	β	γ	others	
11	ParA	168	21	142	0	331
12	IncC	7	22	28	1	58
15	Short ParA	0	60	1	0	61
14	ParF	2	0	11	0	13
13	ParM	1	2	104	0	107

(b) DNA-binding proteins

Family Accession	Replication system	Division of proteobacteria				Total
		α	β	γ	others	
21	ParB	161	42	81	0	284
22	KorB	3	17	21	0	41
23	ParB-like	0	59	0	0	59
24	ParR	0	0	61	0	61

2.3.6.3 Example of replicons in this study

Table 2.5 shows several examples of plasmids (or chromids) investigated in this study that have Rep and Par systems. In general, most plasmids have one Rep initiator each for replication, and one NTPase and one DNA binding protein for partitioning (eg. pRL12, pCTX-M3). Some plasmids, however, have multiple Rep

proteins in one plasmid (more details in chapter 3). Examples include pECL_A and pPNAP01. Several plasmids actually have more than 3 Rep proteins (e.g. pAPEC-O20ColV), but only one seems to be active. Several plasmids also have two types of partitioning coupled proteins (e.g. pECL_A), or some have two copies of one coupled operon (e.g. pOU7519).

Table 2.5 Example replicons and their Rep and Par systems in this study

* : Accession number in the database.

Proteob-acteria	Species	Plasmid	Replication system (accession)*	Partitioning system (accession)*
α	*Acidiphilium multivorum* AIU301	pACMV1	RepC (PBD01) RepA (PBD05)	ParA (PBD11), ParB (PBD21)
α	*Agrobacterium vitis* S4	pTiS4	2 RepC (PBD01)	2 ParA (PBD11), ParB(PBD21)
α	*Ketogulonicigenium vulgare* Y25	pYP	RepA-like (PBD02) RepB-like (PBD03)	2 ParA PBD11), 2 ParB (PBD21)
α	*Rhizobium leguminosarum* 3841	pRL12 (Chromid 1)	RepC (PBD01)	ParA (PBD11), ParB (PBD21)
β	*Polaromonas naphthalenivorans* CJ2	pPNAP01	RepA-like (PBD02) RepB-like (PBD03) TrfA (PBD04)	Short ParA (PBD15), ParB-like (PBD23)
β	*Ralstonia pickettii* 12D	pRL12D01	RepB-like (PBD03) RepA (PBD05)	Short ParA (PBD15), ParB-like (PBD23)
Y	*Citrobacter freundii*	pCTX-M3	RepFIIA (PBD09)	ParM (PBD13), ParR (PBD24)
Y	*Coxiella burnetii*	QpDV	RepB-like (PBD03)	ParA (PBD11), ParB (PBD21)
Y	*Enterobacter cloacae* sub.cloacae ATCC 13047	pECL_A	RepFIIA (PBD09) RepFIB(PBD08)	ParA (PBD11), ParB (PBD21), ParM (PBD13), ParR (PBD24)
Y	*Escherichia coli*	pAPEC-O2-ColV	RepFIIA (PBD09) RepFIB(PBD08) RepFIA(PBD07)	ParA (PBD11), ParB (PBD21)
Y	*Salmonella enterica subsp. enterica serovar Choleraesuis*	pOU7519	2 RepFIB(PBD08)	2 ParA (PBD11), ParB (PBD21)

2.4 Availability of database and user interface functionality

The database is presented through a web interface programmed in PHP, CGI, both of which realize requests to the MySQL database. An intuitively simple menu and search engine make it possible for users to navigate each family of genes comfortably (http://bioplasmid.godohosting.com).

2.4.1 Browser

A 'browse' tab is available, which allows users to obtain all homologous genes in one table. This means that our database can be used as a research tool to explore patterns in the taxonomic distribution and phylogeny of plasmid backbone genes. **Figure 2.4** shows the web interface with its two-step process to reach a table for each type of homolog set. The user selects the module that they want to research. The results consist of the table of homologs of each type. All (or selected) homologs can be downloaded as a fasta-formatted file (unaligned or aligned). Also, there is a sorting option allowing users to arrange a table if they want.

2.4.2 Search

A 'Search' menu is available, which can be used as an annotation tool for amino acid sequences whose gene functions are unknown. Users submit a sequence (query) and the server returns authoritative annotation for those components that belong to families in the database. For creating a search, both uploading sequences and just pasting them in the text area are availabl. One of the programs in the HMMER package, hmmsearch, is used. A default value is set but users can also specify the e-value of hmmsearch in case the sequences are diverged (default 1e-10). The output follows a similar format as that of BLAST type results, which indicate the families to which a protein query belongs. Once a match has been made, the user can easily download either the raw fasta file of the cluster or the MSA file to create a tree incorporating the new sequence(s).

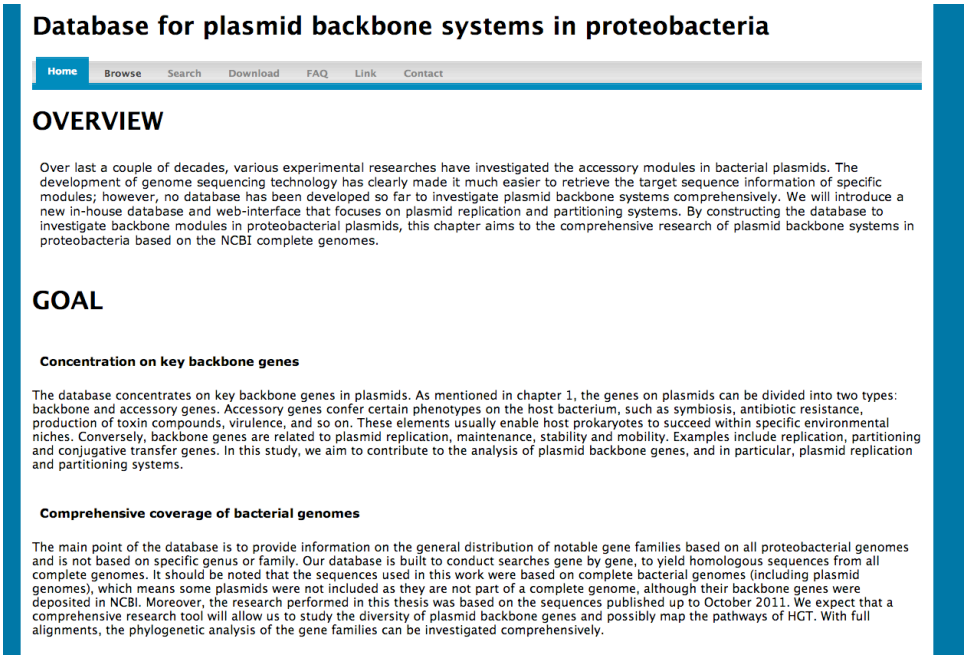
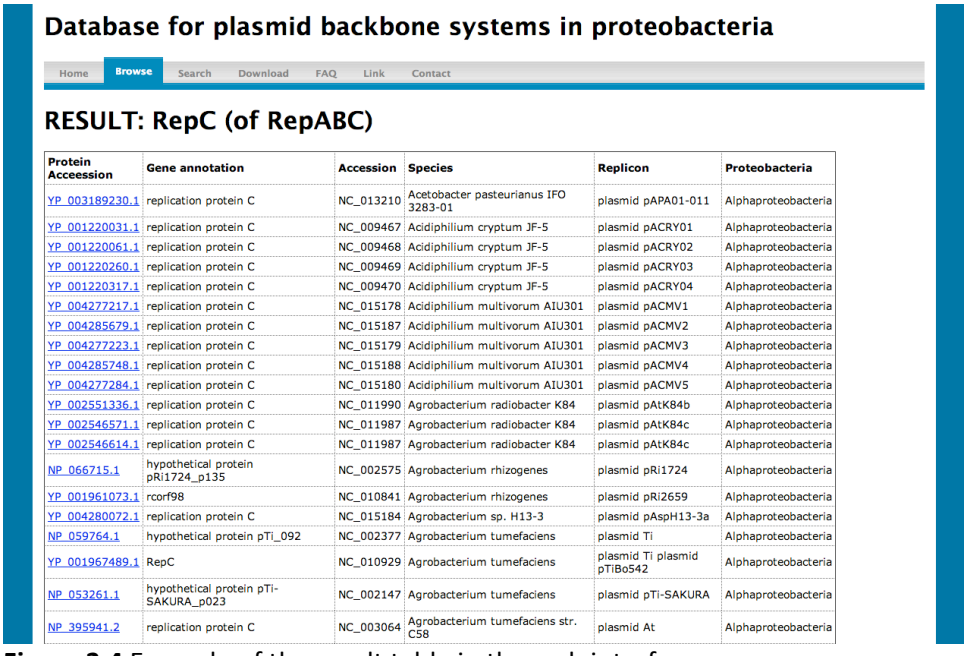


Figure 2.3 Web interface (http://bioplasmid.godohosting.com)

There are three main menus: browse, search and download. 'Browse' has a couple of steps in order to reach a result table. Users can select the module that they want to search.



Protein Accession	Gene annotation	Accession	Species	Replicon	Proteobacteria
YP_003189230.1	replication protein C	NC_013210	Acetobacter pasteurianus IFO 3283-01	plasmid pAPA01-011	Alphaproteobacteria
YP_001220031.1	replication protein C	NC_009467	Acidiphilium cryptum JF-5	plasmid pACRY01	Alphaproteobacteria
YP_001220061.1	replication protein C	NC_009468	Acidiphilium cryptum JF-5	plasmid pACRY02	Alphaproteobacteria
YP_001220260.1	replication protein C	NC_009469	Acidiphilium cryptum JF-5	plasmid pACRY03	Alphaproteobacteria
YP_001220317.1	replication protein C	NC_009470	Acidiphilium cryptum JF-5	plasmid pACRY04	Alphaproteobacteria
YP_004277217.1	replication protein C	NC_015178	Acidiphilium multivorum AIU301	plasmid pACMV1	Alphaproteobacteria
YP_004285679.1	replication protein C	NC_015187	Acidiphilium multivorum AIU301	plasmid pACMV2	Alphaproteobacteria
YP_004277223.1	replication protein C	NC_015179	Acidiphilium multivorum AIU301	plasmid pACMV3	Alphaproteobacteria
YP_004285748.1	replication protein C	NC_015188	Acidiphilium multivorum AIU301	plasmid pACMV4	Alphaproteobacteria
YP_004277284.1	replication protein C	NC_015180	Acidiphilium multivorum AIU301	plasmid pACMV5	Alphaproteobacteria
YP_002551336.1	replication protein C	NC_011990	Agrobacterium radiobacter K84	plasmid pAtK84b	Alphaproteobacteria
YP_002546571.1	replication protein C	NC_011987	Agrobacterium radiobacter K84	plasmid pAtK84c	Alphaproteobacteria
YP_002546614.1	replication protein C	NC_011987	Agrobacterium radiobacter K84	plasmid pAtK84c	Alphaproteobacteria
NP_066715.1	hypothetical protein pRi1724_p135	NC_002575	Agrobacterium rhizogenes	plasmid pRi1724	Alphaproteobacteria
YP_001961073.1	rcorf98	NC_010841	Agrobacterium rhizogenes	plasmid pRi2659	Alphaproteobacteria
YP_004280072.1	replication protein C	NC_015184	Agrobacterium sp. H13-3	plasmid pAspH13-3a	Alphaproteobacteria
NP_059764.1	hypothetical protein pTi_092	NC_002377	Agrobacterium tumefaciens	plasmid Ti	Alphaproteobacteria
YP_001967489.1	RepC	NC_010929	Agrobacterium tumefaciens	plasmid Ti plasmid pTiBo542	Alphaproteobacteria
NP_053261.1	hypothetical protein pTi-SAKURA_p023	NC_002147	Agrobacterium tumefaciens	plasmid pTi-SAKURA	Alphaproteobacteria
NP_395941.2	replication protein C	NC_003064	Agrobacterium tumefaciens str. C58	plasmid At	Alphaproteobacteria

Figure 2.4 Example of the result table in the web interface

The result table includes the accession number, name of species, gene annotation from NCBI, gene length, taxa ID, name of replicons, and information of whether it is on plasmids or chromosomes.

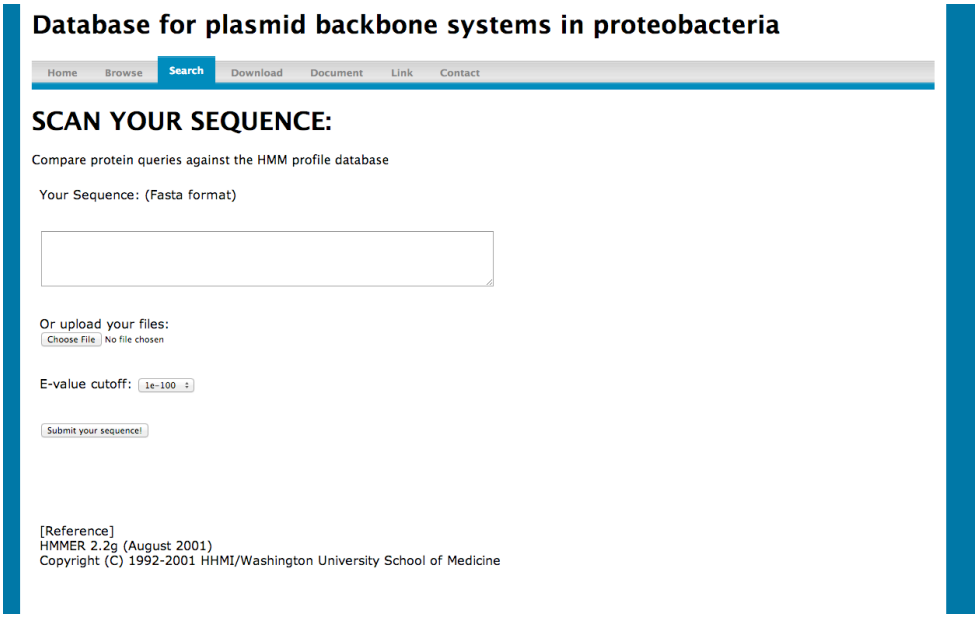


Figure 2.5 Search menu in the web interface

Users can submit their chosen sequence either as a raw text or a file (fasta-formatted).

2.5 Discussion

2.5.1 Contribution of this study

A variety of past attempts to develop databases and analysis tools have proven to be only partly successful in providing a better understanding of the genomes of multiple species. They have offered researchers authoritative repositories and curated data as a method of managing the massively increasing amount of biological data [49]. For the last few decades, research on MGEs has been accelerated because of their potential ability to confer specific phenotypes. MGE bioinformatics are alsoincluded in the current trends in biology, which have result in numerous related databases and tools being created as simple repositories or visualization/analysis tools.

To our knowledge, this study is the first work to contribute to the storage of plasmid backbone systems, including replication and partitioning modules and

their comprehensive analysis. We have collected protein sequences involved in plasmid backbone functions from public genomes and categorized them into distinct gene families. 1933 protein sequences (17 gene families) in total were analysed, and the database can be used to study distribution, host range, and phylogenetic relationship. Therefore, it is significant to note that this database is not just a tool for the collection of numerical data, but also includes biological meanings. Moreover, the web-interface, which allows access to the gene families, sequences, and phylogenies through the 'Browse' menu, provides authoritative annotation for unknown raw data.

2.5.2 Limitation and future perspectives

2.5.2.1 Interact with other databases and analysis for better understanding

A most promising field for future work would be the study of plasmid backbone and accessory modules in their totality, although the handling of the different type of gene families distributed across bacteria requires a lot of efforts. Despite the publication of a variety of individual works that are concentrated on specific genes (mostly accessory genes), the way data was collected In each is different, while the research itself is largely constrained to specific species or genera, which makes it difficult to bring together the ideas.

This study, however, might be a first step to integrate information at least for plasmid backbone modules. For example, Smillie and his colleagues [45] reviewed the mobility of conjugative plasmids comprehensively. In particular, they stated that their in-house database was generated for the sake of analysis. By integrating the results of our replication, partitioning modules and their transfer modules, we might gain interesting ideas about the distribution and diversity of plasmid backbone systems in general. Furthermore, more studies investigating specific genes in IS elements and genomic islands might be useful because they could tell a different story regarding plasmid backbone genes, which could indicate recent movement between species.

2.5.2.2 Automatic update

One of the main drawbacks of an individually developed database and analysis is the difficulty to maintain and update the data regularly and continuously. According to NCBI, complete genomes of bacteria are updated (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) every day, and plasmid genomes (ftp://ftp.ncbi.nih.gov/genomes/Plasmids/) every week. Therefore, gene families identified in this study also should be re-evaluated, through the addition of newly sequenced genomes. Moreover, phylogenetic trees should be updated. It is not easy, however, to collect all the homologous sequence by psi-blast (sometimes manually) every time the complete genomes of NCBI are updated, and classify them by MCL clustering, as many MGE databases now have been adopted. The process of generating phylogenetic trees whenever new families are generated is also computationally expensive in terms of time and memory. One way to overcome this problem would be by using our in-house HMMs, which can update members of the gene families relatively fast. If individual HMMs for MGEs are constructed reliably, it is much faster to collect homologous genes than doing MCL clustering after the blast search. It is still problematic, however, to make phylogenetic analysis time-efficient, as other web services for phylogenetic analysis have not achieved. Hence, this study should be considered as a mere stepping-stone for further research in this direction, in order to solve these problems.

2.5.2.3 Why are not all plasmid sequences included?

There are three main reasons why not all plasmid replicon sequences are included in the analysis. Firstly, many plasmids have been sequenced and deposited in NCBI, but some of these are not part of a complete genome sequence. In this case, we did not include the replicon information. Secondly, the sequences deposited after October 2011 were not included in this analysis. The research was based on the sequences stored in NCBI up to the end of October 2011. Finally, if rep or par systems of certain plasmids do not belong to the major families considered in this analysis, then they are not included, because we only explored the main families of each system in this thesis.

Chapter 3. Diversity of plasmid replication systems in proteobacteria

The self-replicating process is one of the significant defining characteristics of plasmids, which differentiates them from other mobile genetic elements, such as phages, transposons and genomic islands. A common region is responsible for replication and its control, i.e. the origin of replication site (*ori*) and the *rep* gene encoding the replication initiator. In this chapter, we investigate the plasmid replication systems in proteobacteria. Based on the discrete types of the replication initiators (Rep proteins) in the database, we firstly review the general genetic organization of the main families of Rep proteins. We then study diversity and distribution of each family, and perform a phylogenetic analysis in order to research their evolutionary history.

3.1 General introduction

3.1.1 Plasmid replication and its control

The process of self-replication in an autonomous way is one of the significant characteristics that defines plasmids and differentiates them from other mobile genetic elements, such as phages, transposons and genomic islands. During plasmid replication, it is firstly important to control their own replication rate. Excessive copy numbers can burden the host cell, while too low copy numbers might result in disappearance in the next generation. In general, most controlling systems are involved in the initiation of leading strand synthesis of DNA. The genes for plasmid replication and control are usually clustered together in what is called the basic replicon (**Figure 3.1**). There are various ways to regulate the process, but it is mainly antisense RNA-mediated inhibition and iteron binding of replication initiation proteinthat are involved in controlling [67].

In addition to the control systems, the plasmid replication mechanisms have been extensively researched in the past decades. Commonly, plasmid replication requires two basic elements. The origin of replication site (*ori*), harboured within several hundred base pairs, is the first element that is a *cis*-acting DNA. It supports the autonomous replication of the plasmids and is the region where DNA strands are melted to initiate the process [68]. The second element is the plasmid specific gene that encodes the replication initiation protein (usually called 'Rep') that binds to *ori* (**Figure 3.2a)**. Although plasmid replication systems demonstrate much variety, they are generally divided into three types (**Figure 3.3**): theta type, rolling-circle type, and strand displacement mechanisms [68].

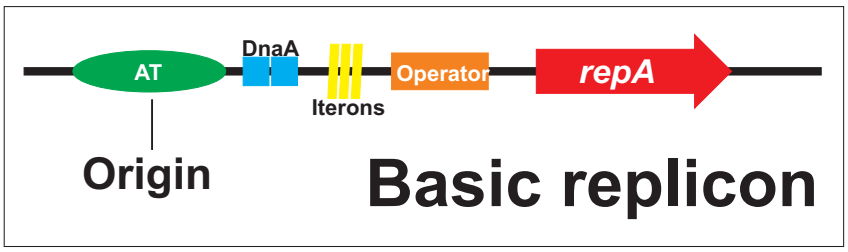


Figure 3. 1 The genetic organization of a typical minimal replicon

A basic replicon (also known as a minimal replicon) consists of the origin of replication site (*ori*) and a *repA* gene (represented as a red arrow), which encodes the Rep initiator. The green oval is an AT-rich region. Yellow lines are iterons and the orange box is an operator site. Blue boxes are sequences that DnaA proteins are bound to.

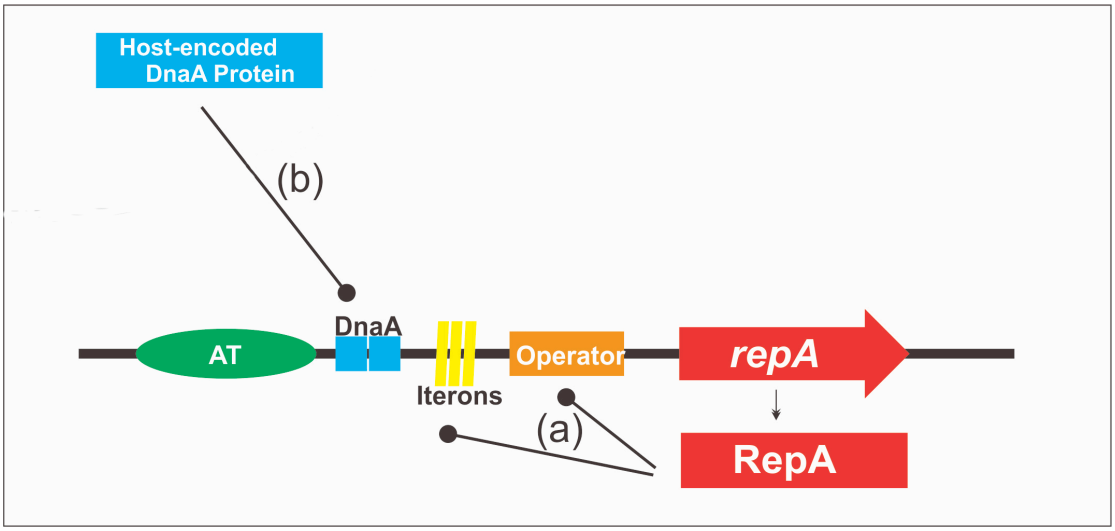


Figure 3.2 Basic elements required for plasmid replication

(a) Plasmid initiator Rep protein. Different plasmids encode different Rep proteins (red box), but generally DNA-binding protein recognizes iterons near origin and autoregulates *rep* gene expression; (b) DnaA protein. Host-encoded DnaA protein binds to DnaA boxes. RepA-DnaA-DNA nucleoprotein complex promotes stand separation at AT rich regions, which contain origin.

3.1.1.1 Theta replication

Plasmids using the theta replication mechanism open double strands of DNA at the *ori* region. In *ori,* an array of ~20 bp repeats (called 'iterons') plays an important role in the initiation and control of replication [69]. The *rep* gene encodes the Rep protein that initiates the replication event by recognizing the origin of replication site on a replicon and creating a theta structure (resembling the Greek letter theta

'θ'), which acts as an intermediate during replication. Rep exists in both monomer and dimer forms, but only monomers bind specifically to iterons and serve as initiators, while dimers are inactive during iteron binding [70]. In this process, plasmids can be uni- or bi-directionally replicated. Replication stops when the fork returns to the origin (uni-) or when two replication forks meet on the other side of the plasmids (bi-directional).

The theta mechanism is the most common form of DNA replication, especially in gram-negative bacteria [69]. Although there are several common regions related to the replication process, which include iterons, regions, such as AT-rich sites or DNA boxes, can vary between replicons. Moreover, after initiation, several host-encoded proteins begin to get involved in the process, which usually renders plasmids dependent on one or a few closely related hosts. For instance, DnaA, which is a protein required for the initiation of replication of the bacterial chromosome, often promotes the making of Rep-DnaA-DNA complex for melting the strands (**Figure 3.2b**). In addition, the host-encoded DnaB helicase has a role in initiating complex formation. Finally, DnaG as a primase and DNA polymerase III in the initiation of synthesis in both leading and lagging strands are important for completing replication reactions [71].

3.1.1.2 Strand displacement replication

Strand Displacement Replication is a mechanism used by a small portion of plasmids. Three proteins are involved in this mechanism: RepA (helicase), RepB (primase) and RepC (replication initiator) [72]. The RepC initiator firstly binds to the iterons and then promotes open complex, by bending of the DNA. RepC also recruits the RepA helicase, which separates the DNA and exposes two *ssi* sites, *ssiA* and *ssiB*. RepB is a primer that prioritizes continuous DNA synthesis in opposite directions by host-encoded DNA polymerase III. As these three proteins do not need help from host-encoded replication proteins, such as DnaA, DnaB or DnaC, they are relatively easy to replicate in a broad range of hosts[68]. IncQ plasmids, including prototype RSF1010, are amongst the best examples of the Strand Displacement Replication mechanism.

3.1.1.3 Rolling-circle mechanism

Some plasmids replicate using a rolling-circle replication (RCR) method. This usually involves small plasmids (<10kb). Originally, RCR plasmids were identified in Gram-positive bacteria, but were later also found in Gram-negative bacteria as well. RCR plasmids usually carry antibiotic or heavy metal resistance genes, but some are also cryptic, which means that they do not encode any obvious phenotypes. Many RCR plasmids have a broad host range, while research has documented strong evidence regarding horizontal gene transfer.

There are three elements involved in the RCR mechanism: a replication initiation protein, a double-strand origin (called *dso*) and a single-strand origin (called *sso*) [73]. The Rep protein of rolling-circle plasmids binds to the double stranded *ori* (*dso*) site. This binding then distorts the DNA and nicks it in the *dso*. The nicks leave a 3' hydroxyl end that serves as a primer for leading strand synthesis. Using the unnicked strand as a template, replication proceeds around the circular DNA molecule. The 5' end is displaced and forms a tail of single-stranded DNA that extends from the circle. Plasmids using RCR generally need host-encoded proteins. For example, a host-encoded helicase opens double stranded DNA to expose *sso,* while DNA polymerase displaces the original leading strand using the 3' hydroxyl end at the nick. The host-encoded proteins are also necessary in order to generate a second copy by initiating replication at the *sso* on the displaced strand and in order to synthesize the lagging strand [73].

3.1.2 Classification of bacterial plasmids

The classification of all kinds of organisms has been practiced since the dawn of the biological sciences. Well-defined classification is a good foundation upon which to investigate the distribution and diversity of the organisms themselves, the relationships between them, as well as to discover their origin in terms of evolution. This concept is equally helpful for plasmids. As more plasmids became identified, a lot of attempts were made to catalogue them in different schemes. The traditional plasmid classification into F, Col, R, etc. began almost 30 years ago.

Initially the most popular way to classify bacterial plasmids has been the 'incompatibility group' [74]. If two plasmids contain the same origin of replication system, they cannot coexist in the same cell. Thus, one of the plasmids becomes segregated. This phenomenon is known as 'incompatibility'. Over 30 groups of plasmids have been defined on the basis of *inc* properties so far [75]. In particular, plasmids of gram-negative bacteria have been mostly categorized based on incompatibility, which is also an important tool for tracing antimicrobial resistance plasmids [75, 76]. Although this method is relatively conducive to defining interesting plasmids, incompatibility grouping can be problematic when dealing with the exponentially increasing amount of newly sequenced data. This is because analysis should be conducted every time a new sequence is published and such experimental work is time-consuming and laborious [75].

In 1998, Couturier et al. demonstrated a classification method based on hybridization with 19 DNA probes that matched different plasmids. It is, however, difficult to be applied to a large number of strains and also very time-consuming (69). Later, Carattoli et al. [77] suggested PCR-based detection, which traces plasmids conferring drug resistance. Their target plasmids, however, are very limited. On the other hand, ongoing research suggests that the mobility of plasmids should be considered as an element for classification [45, 78]. Using plasmid relaxases and the related transfer systems, these scholars have tried to group conjugative plasmids into 6 families. Although their idea is important, they can only study a part of the whole family of bacterial plasmids, as many do not actually possess transfer systems.

In order to establish a valid classification scheme, it is important for the criteria used to be based on genetic traits that are commonly present [75]. As mentioned above, the reaction of initiator Rep would be different between plasmids. The initiation of replication in plasmids, however, mainly depends on the origin of replication and replication initiation proteins, although the plasmids also require additional host-encoded genes[69]. This makes it a good yardstick for classifying plasmids. Moreover, the replication protein is an essential element in defining incompatibility groups, so much so that the concept of incompatibility has to be encompassed within this scheme [79], despite the fact that there are still some

unresolved issues regarding the plasmids having multiple replication sites s(see section below).

3.1.3 Host range of plasmids

Prior to the separate close investigation of each type of replication system in plasmids, the concept of 'host range of plasmids' should be considered carefully. In general, plasmid host range refers to the range of hosts that a plasmid is able to replicate. The concept originates from the fact that some plasmids can be maintained in a couple of related bacterial hosts exhibiting a 'narrow' host range, while others can transfer and replicate in distantly related bacteria indicating a 'broad' host range.

The terms 'narrow' and 'broad' host ranges, however, can be interpreted in different ways. Firstly, the host range can be 'long-term' [80] where plasmids can not only replicate, but also should be maintained stable in a long-term basis. This is a very strict concept. Secondly, 'replication host range' is normally considered as the range where a plasmid is just able to replicate. This is different from the 'transfer host range' [81, 82] where the plasmid can be transferred by conjugation. This is the third concept. On the other hand, the host range can be interpreted also as 'observed host range' [81, 83], where a plasmid is actually found in different niches. Finally, Suzuki et al. [79] designated the category 'evolutionary host range' where plasmids would have replicated many times during their evolutionary history. The actual process that they have been evolved is unknown, but presumably the evolutionary host range is narrower than the replication range.

In this study, the term 'host range' will be closely related to 'observed host range' and 'replication host range'. To put it succinctly, a lot of plasmid data used here originates from the completed genome projects, which include diverse 'naturally occurring' plasmids, rather than plasmids used as vectors. Therefore, we can say that the host range of this study is the range in which plasmids are actually found, unless there is a note in the NCBI entry that they come from 'lab hosts'. The following sections are divided into three categories: i) replication systems in

plasmids showing naturally narrow host ranges, ii) same as i) but having multiple replicons and iii) replication systems in broad host range.

3.1.4 Chapter objectives

The primary purpose of this chapter is to investigate the plasmid replication systems in bacteria based on public genomes. More specifically, our investigation of replication systems will focus on plasmids in proteobacteria. The main target protein in studying the replication system is the replication initiator Rep. We will study not only its functional significance, but also its role as one of the major factors according to which we can classify plasmids both bioinformatically and phylogenetically. The Rep protein is worthy of closer investigation, as Rep proteins are encoded in most plasmids and generally share a common function. Based on the database we have established in chapter 2, the aims of the present chapter are to:

1. Identify and classify gene families involved in plasmid replication, review them and propose a classification scheme. No recent comprehensive review of plasmid replication systems has taken place, despite the fact that biological data have been published massively. We will investigate the genetic aspect of each replication type in proteobacteria and demonstrate how it can be used as a classification scheme.
2. Investigate the diversity and distribution of plasmid replication systems across bacteria. One of the most important contributions that we hope to make in this study is to examine how replication systems are distributed across the bacteria by using publicly available genomes. It is envisaged that this study will provide insights on how each replication system is related to the various incompatibility groups and in particular to the host range that the plasmids appear naturally. We expect that we will also be able to infer the host ranges of other incompatibility groups, which have been researched in less detail so far by using *in silico* methods.
3. Construct the phylogenetic tree of each family and analyze them in terms of

the evolutionary history of each replication system. We aim to investigate whether there is possible recent traffic between specific species groups and study the evolutionary relationships between different types of replication systems based on their phylogenies.

3.2 Overview of plasmid replication systems

As described in chapter 2, we have downloaded complete genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) and plasmid genomes (ftp://ftp.ncbi.nih.gov/genomes/Plasmids/) from NCBI and have identified 8 main families involved in plasmid replication systems. This is because we now know that plasmid replication systems are also detected in the second and third chromosomes or large plasmids, which are known as 'chromids' [84]. The families we defined include 943 initiator proteins in total, not only of plasmids, but also of chromids.

In this chapter, we investigated 8 main families, which are mostly abundant in proteobacterial plasmids and are defined as a major group here. Before further discussion, it should be noted that there is the possibility that source data might be biased. Smillie et al. [45] have pointed out in their paper that much research has focused on gammaproteobacteria and Firmicutes mostly because of their antibiotic resistance among human-pathogenic bacteria. This has resulted in a relatively high percentage of information about these bacteria. Until October 2011, 2543 plasmid genomes have been published, 1112 of which plasmids originated from proteobacteria (43%) and 667 were gammaproteobacterial plasmids (60%). Therefore, one should be aware that the number of specific divisions of bacteria might not reflect that which exists naturally. Similarly, the actual number of Reps closely related to the replicons from any proteobacteria might not necessarily reflect actual prevalence in replication systems.

Table 3.2 indicates the coverage of replicons in this study. More than 50% of alpha- and betaproteobacterial plasmids can be categorized according to these families.

Conversely, the main families only include 44% of gammaproteobacterial plasmids. It appears that many plasmids have initiators that are not included in the major groups. The thesis assumes that there are many different families having a small number of initiators, particularly the plasmids in the gamma-division of proteobacteria. One should also note that the plasmid replication systems of delta- and epsilonproteobacterial plasmids are not included in the main families, since their genome data is limited at the moment, which means many species in these divisions are very similar to each other.

On the other hand, some plasmids possess more than one replication initiator. Firstly, this may be because there is an in-frame translation initiation or start codon, which produces two proteins through one gene. Examples include many *trfA*s, encoding short and long TrfA proteins [85, 86]. Secondly, some plasmids have two or more replication systems in one plasmid. Numerous IncF or IncH plasmids belong to this category [87, 88]. Finally, some plasmids have two initiators, but one of these may correspond to a non-functional replicon as a result of insertion sequences, or other MGEs, which render it inactive - or it may not be translated.

Table 3.1 Eight main families of replication initiator proteins studied in this study

Family accession	Common name	Alternate name	Example (Replicon/Host)
001	RepC (of RepABC)	-	▪ pRL7-12, pRLG01-05 / *Rhizobium leguminosarum* ▪ Chromid / *Brucella sp., Agrobacterium sp., Sinorhizobium sp.*
002	RepA-like	RepA, RepB	▪ pRSPA02 / *Rhodobacter sphaeroides*
003	RepB-like	-	▪ Chromid 1, 2 / *Burkholderiasp.* ▪ Chromid / *Rhodobacter sp.* ▪ pQpDG / *Coxiella sp.*
004	RepFIIA	RepA	▪ NR1, R100, pB171 / *Escherichia coli* ▪ pLeu-Sg, pLeu-Dn / *Buchnera sp.* ▪ R64, ColIb-P9 / *Samonella enterica* ▪ R721 / *Escherichia coli* ▪ pCTX-M3 / *Citrobacter freundii*
005	RepFIA	RepE	▪ R46 / *Escherichia coli* ▪ pHCM1, R27, Plasmid F / *Salmonella enterica*
006	RepFIB	RepHIA, RepHIB, RepB	▪ R478 / *Serratia marcescens* ▪ R27 / *Salmonella enterica* ▪ pKPN3, pKPN4 / *Klebsiella pneumoniae*
007	TrfA	-	▪ R751 / *Enterobacter aerogenes* ▪ pB4 / *Pseudomonas sp*. ▪ pKJK5 / *Escherichia coli*
008	RepA	-	▪ pIE321, pMAK3 / *Salmonella enterica* ▪ IncW plasmid / *Providencia rettgeri*

Table 3.2 Coverage of replicons from proteobacteria based on main families

Divisions of proteobacterias	Total plasmid Reps	Reps in the main families of this study	Reps not belonging to the main families	Percentage covered
α	252	148	104	58%
β	116	68	48	58%
γ	667	299	368	44%
δ	15	5	10	33%
ε	62	0	62	0%
Total	**1112**	**520**	**592**	**48%**

3.3 Results

3.3.1 RepC family

The first that we have investigated is the RepC family known as a replication initiator in RepABC replicons. The RepABC replicons are one of the most extensively studied groups in bacterial plasmids, regarding their functions and evolution. The replication region of RepABCreplicons comprises three genes, *repA*, *repB*, and *repC*, which always appear in the same order (**Figure 3.3)**: *repA* is upstream, *repB* is next to *repA,* and *repC* is downstream of this operon [89]. Note that *repA* and *repB* are involved in the replication process, particularly repression, but they are mainly responsible for the segregation process. They are also frequently called *parA* and *parB*, and are discussed in detail in chapter 4. *repC* encodes the RepC protein, which functions as the plasmid replication initiator by binding to the origin of replication.

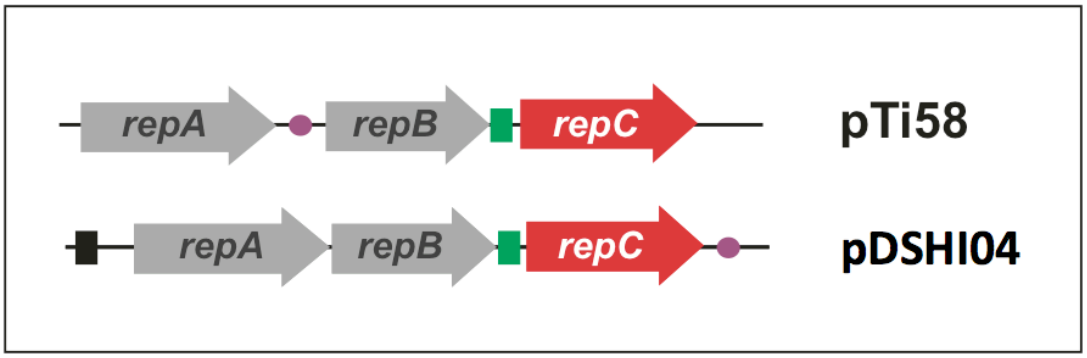


Figure 3.3 The genetic organizations of replication regions in RepABC replicons

pTi58 from *Agrobactrium tumefaciens* [90], and pDSHI04 from *Dinoroseobacter shibae* DFL 12 [91]. Purple circles indicate *parS* sites; green boxes are antisense RNA genes; black boxes are promoters.

RepABC replication regions are mainly found in bacterial plasmids. They also exist, however, in 2nd and 3rd chromosomes or megaplasmids, as briefly mentioned in section 3.2. Harrison et al. [84] argued that if a chromosome or megaplasmid possesses a plasmid replication system rather than the chromosome replication system, and they also have core genes that are usually located in chromosomes, showing chromosome-like genetic composition, they should be defined as a 'chromid'. They also suggested that chromids might originate from plasmids, incorporating other core genes from the main chromosomes. All known chromids of alphaproteobacteriahave RepABC replication systems. In addition to the ones defined as 'extrachromosomes (2nd or 3rd chromosomes)' or 'megaplasmids', many large plasmids also satisfy the criteria that Harrison et al. set up. Examples include pRL12, pRL11 of *Rhizobium leguminosarum* bv. viciae 3841, p42f, p42e of *Rhizobium etli* CFN42, and pSymB of *Sinorhizobium meliloti* 1021.

The RepC family of RepABC replicons is widely distributed, but is restricted only to alphaproteobacteria, mostly in the *Rhizobiales*, such as *Rhizobiaceae (Rhizobium, Sinorhizobium, Agrobacterium), Brucellaceae (Brucella, Ochrobactrum), Bradyrhizobiaceae (Oligotropha, Bradyrhizobium) and Methylobacteraceae*. It is interesting to note that most plasmids in *Rhizobiales* usually contain *repABC* replicons [exceptions include pTAR from *Agrobacterium tumefaciens* LBA4301 [92] or pSB102 from *Sinorhizobium meliloti* [93]], while *Rhodobacterales* and *Rhodospirillales* do not necessarily. Plasmids from those families have the RepC

family of replication systems, including *Rhodobacteraceae (Rhodobacter, Dinoroseobacter)* or *Acetobacteraceae (Gluconobacter)*, but a variety of other types of replication systems exist in comparison to *Rhizobiales*, which are discussed in the next section.

We have constructed a phylogenetic tree based on the RepC sequences across bacteria (**Figure 3.4**) and there are a couple of noticeable facts in the phylogeny. Firstly, each clade is particularly important in classifying incompatibility groups in several clades, that is, generally only one plasmid of each strain is located in each clade, which might reflect actual incompatibility groups. For example, when looking at the clade denoted by a grey shaded box, there is no strain having two plasmids in the same clade. They are relatively well-resolved clades in this family. We are able to assume that each clade represents an incompatibility group, which might be used as a marker for the classification of plasmids.

However, there are several exceptions (denoted by red circles) such as pRL12 and pRL9 of *Rhizobium leguminosarum* bv. viciae 3841, pR132501 and pR132504 of *Rhizobium leguminosarum* bv. trifolii WSM1325, pOANT01 and pOANT03 of *Ochrobactrum anthropi* ATCC 49188, pATS4a and pATS4b of *Agrobacterium vitis* S4. Cevallos et al. [89] suggested that this might be due to gene duplication. Young et al. [2] have suggested that divergence of RepC is not significant in terms of compatibility. The divergence of RepA and RepB in their phylogenetic tree shows clear incompatibility groups (see chapter 4 for details). This suggests that RepAB modules might be more important for incompatibility, which is interesting because it has been known that replication-based incompatibility is usually more stringent than segregation-based incompatibility [69, 74, 75, 94, 95]. Because the replication initiator protein has a close relationship with partitioning modules, it is important to investigate the evolution of plasmids based on both systems. In chapter 5, we discuss a possible evolutionary history in *repABC* replicons in detail.

In the phylogeny, most clades are conserved at the order level of bacterial species, which indicates that RepABC replicons show possible movement mostly within the order level. Although Slater el al. [96] argued that they are conserved at the family level of bacteria, conservation at the order level is more general based on the exception of *Ochrobactrum* and *Brucella* species in the upper clade. In the lower

clades, several plasmids from non-*Rhizobiales* such as *Rhodobacterales* and *Rhodospirillales* are also well-resolved mostly within the order level.

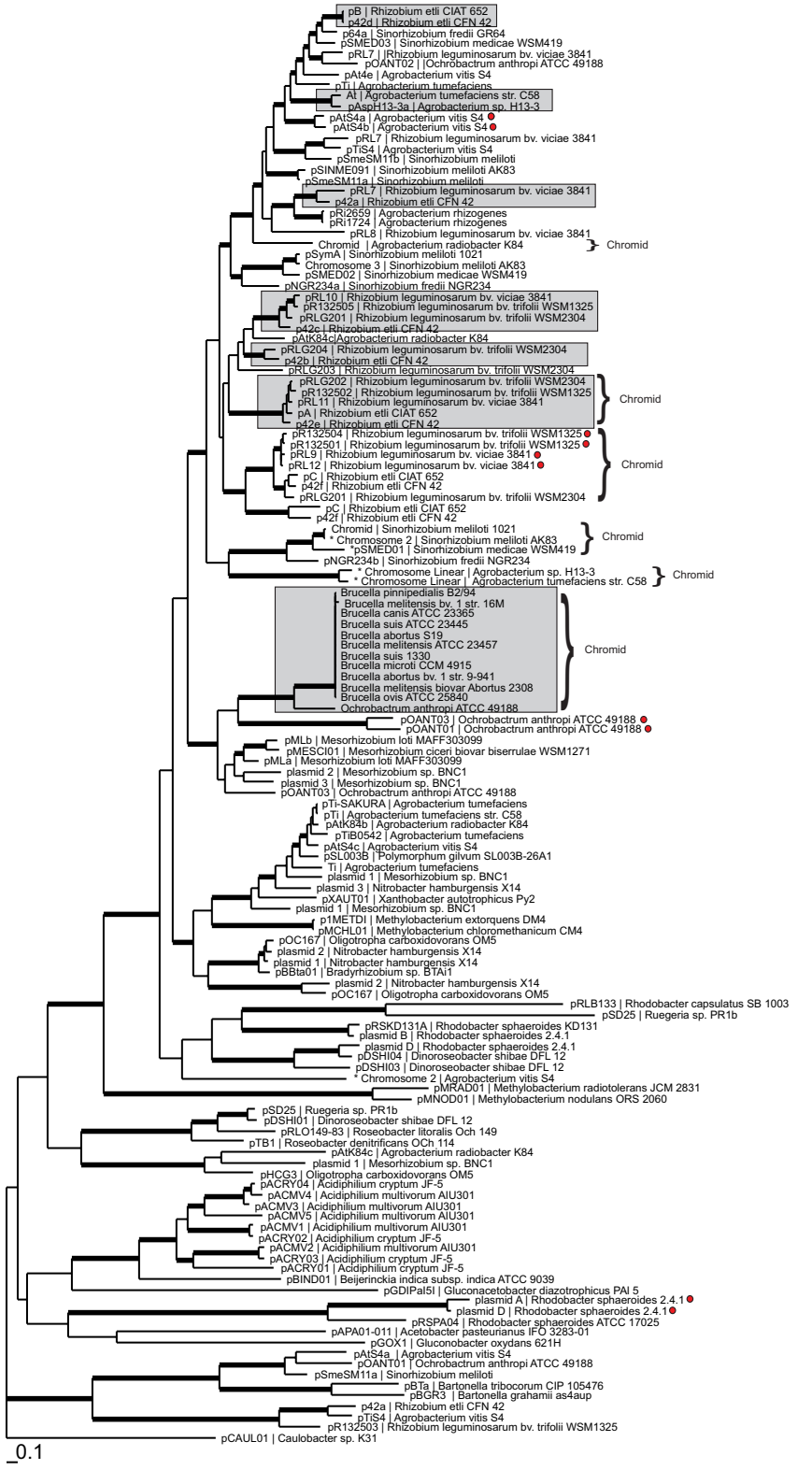


Figure 3.4 The phylogenetic tree of RepC initiators

Unrooted Maximum Likelihood tree. If each clade corresponds to specific incompatibility group, it is indicated by grey shaded boxes. Red circles are initiators in different replicons in the same strain. The chromids of all *Brucella* were originally described as second chromosomes. See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.3.2 RepA-like family

As mentioned above, the replication initiators of some alphaproteobacterial plasmids such as the ones from *Rhodobacterales* vary, in contrast to those of *Rhizobiales.* One of the initiators is called the RepA-like protein (also called just 'RepA' in NCBI Protein Clusters). The members in the RepA-like family are relatively less studied than RepC of RepABC replicons, however, there are very interesting things in this family. RepA-like initiators are generally 280-330 aa in size, and distribution is much wider than RepC, since they are found in alpha-, beta- and gammaproteobacterial plasmids.

Basically, it seems that the gene encoding RepA-like protein and functioning as a Rep system is located adjacent to partitioning modules (**Figure 3.5)**. Unlike RepABC, the direction of RepA-like proteins translated varies. 171 homologous sequences of this family have been found and **Figure 3.6** is the phylogenetic tree of those. Among a variety of RepA-like proteins, the clade denoted by '*' consists of the plasmids possessing RepA-like, ParA and ParB as Rep and Par system. Examples include pDSHI02 from *Dinoroseobacter shibae* DFL 12, pSD20 from *Ruegeria sp.* PR1b, pRSPA02 from *Rhodobacter sphaeroides* ATCC 17025, pRSPH01 from *Rhodobacter sphaeroides* ATCC 17029, etc.

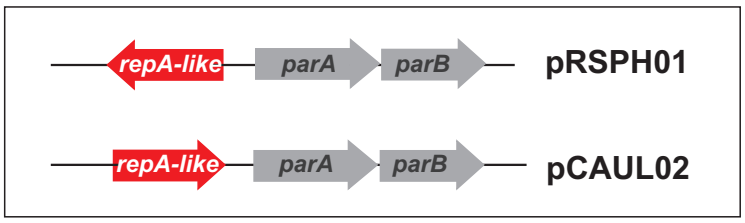


Figure 3.5 The genetic organization of plasmids having RepA-like initiators

pRSPH01 from *Rhodobacter sphaeroides* ATCC 17029 and pCAUL02 from *Caulobacter sp.* K31,

In addition to the RepA-like proteins in these plasmids from *Rhodobacterales*, the distribution of other homologs is interesting because a large number (116) of the RepA-like sequences have actually been found in chromosomes. It is not clear why so many plasmids replication initiators are in chromosomes, and whether they actually function or not. There is a possibility that they have an ancient history that has been incorporated into chromosomes. They might not actually function in the chromosomes. Presumably, these RepA-like initiators could be from integrated plasmids or prophages, because there are *tra* regions, or phage integrase (site specific recombinases), which are just adjacent to Rep initiators. In this case, the partitioning systems that are supposed to be next to Rep do not exist.

Most clades are conserved in the class level in general, but Rep from beta- and gammaproteobacterial plasmids are often mixed in one clade, which indicate relatively close evolutionary relationships between beta- and gammaproteobacterial plasmids, rather than alphaproteobacterial plasmids.

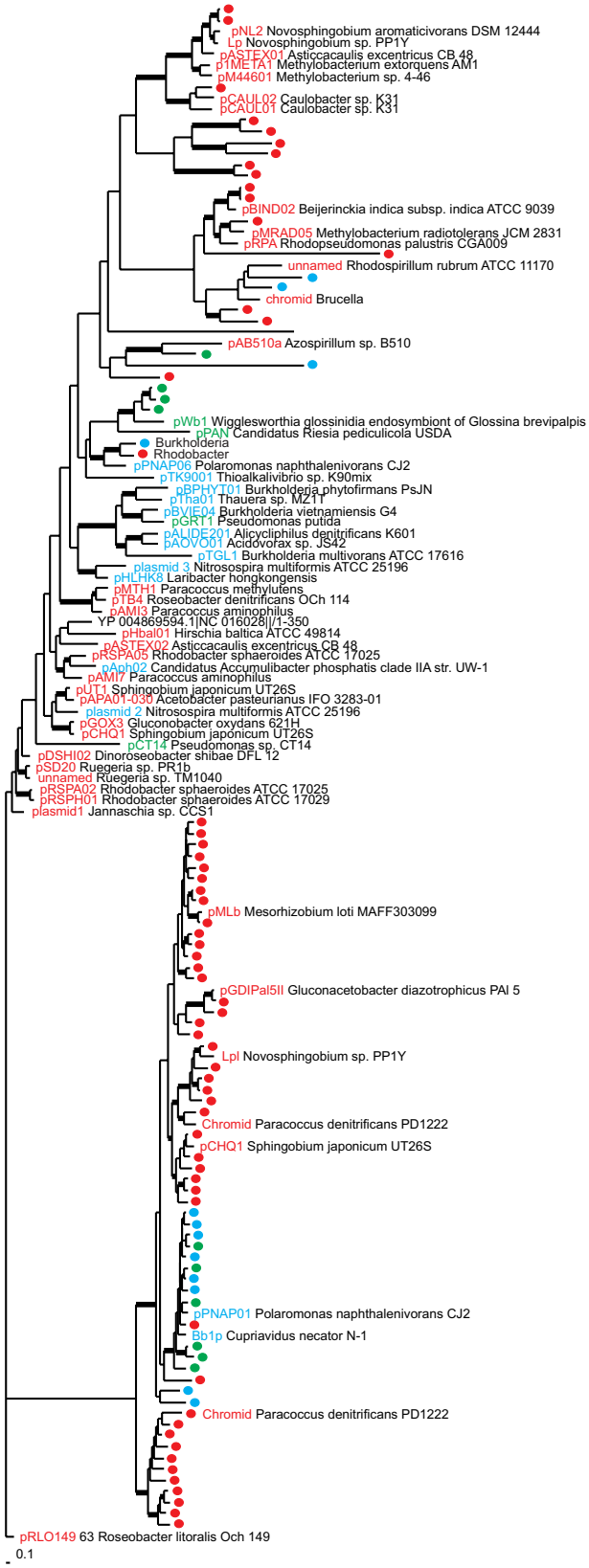


Figure 3.6 The phylogenetic tree of RepA-like initiators

Unrooted Maximum Likelihood tree. Alphaproteobacteria in red, betaproteobacteria in blue, and gammaproteobacteria in green. Red circle is a homolog in the chromosome; otherwise a name of a plasmid is indicated. See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.3.3 RepB-like family

Members of this family are also frequently present in many narrow host range plasmids, particularly in chromids from *Burkholderia*. This type of Rep initiator family has been referred to with various names; the RepB initiator protein in NCBI Protein Clusters, Rep_3 (PF01051) in pfam, RepB-like initiators in Petersen et al. [91] etc. We will follow Petersen et al. in this thesis. It should be noted that this is different from the RepB proteins in RepABCreplicons. RepB-like initiators are mostly located adjacent to partitioning systems (**Figure 3.7)**, similarly to RepA-like and RepABC replication systems; however, their direction and evolution appear to be distinct from those (this is discussed in more detail in chapters 4 and 5).

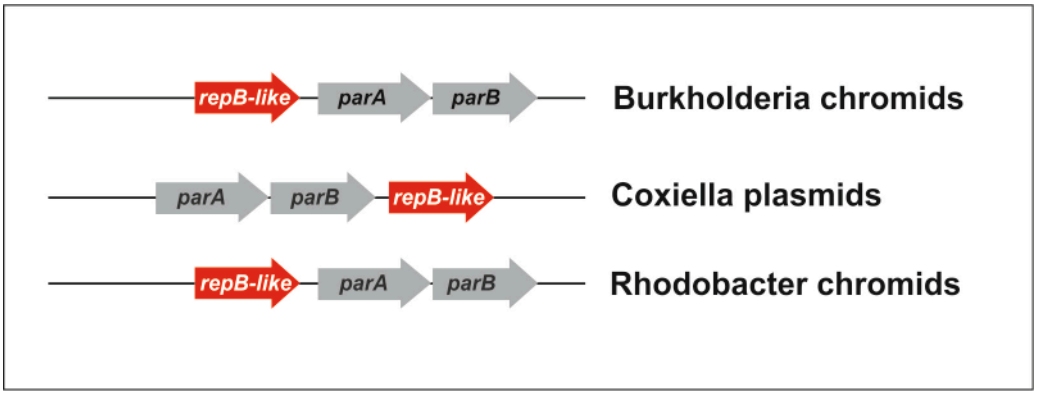


Figure 3.7 The genetic organization of replication regions RepB-like initiators

Species from three different classes of proteobacteria (e.g. *Rhodobacter* from α-, *Burkholderia* from β-, and *Coxiella* from ɣ-proteobacteria) are shown. Rep encoded by the repB-like in red, Par encoded by parA and parB in grey.

Unlike RepABC, distribution of the RepB-like is not restricted to alphaproteobacteria. **Figure 3.8** illustrates the phylogenetic tree of RepB-like initiator homologs. Although species distribution of RepB-like initiator proteins is wide in alpha-, beta-, gammaproteobacteria, it does not seem to have any recent movement outside the order level of species. More specifically, there are four well-resolved clades. Firstly, 1st and 2nd chromids of *Burkholderiales* are shown on the upper clade (clade A and clade B, respectively). The two clades are conserved at

the genus level, showing two separate origins at the genus level [84]. These clades are located with small plasmids of the same order, such as plasmids from *Burkholderia, Ralstonia, Polaromonas*, etc., which are not well separated.

In the bottom clade, there are two well-separated clades. Clade D consists of plasmids in *Legionellales* including *Coxiella* and *Legionella*. This is particularly interesting as they are the only gammaproteobacterial plasmids that have this type of replication initiator. Based on the fact that the bootstrap supporting clade C and clade D is very low, we cannot conclude that clade C is closely related to clade D. Clade C contains various plasmids mostly from *Rhodobacterale*s such as *Rhodobacter, Dinoroseobacter* strains [97]. Chromids of *Rhodobacter* species are also included in this clade. In addition, there are a couple of *Burkholderia* plasmids (e.g. pBVIE05, pBPHY01) having homologous replication proteins, even though the bootstrapping values are not high. Overall, small plasmids are more diverse and there are fewer cases of closely related plasmids from different strains.

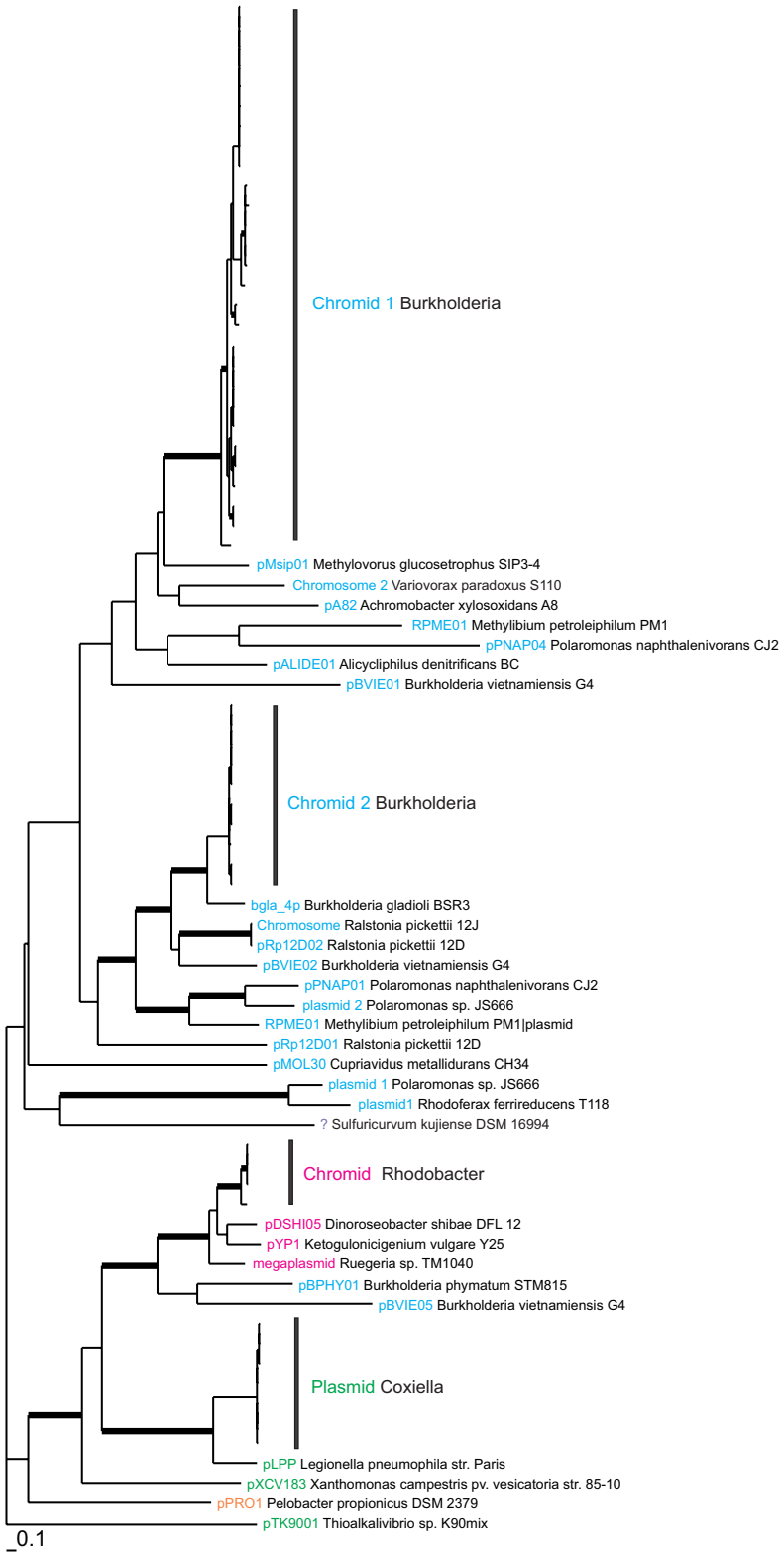


Figure 3.8 The phylogenetic tree of RepB-like sequences

Unrooted Maximum Likelihood tree. 28 Rep proteins of Burkholderia chromid 1, 10 Rep proteins of Burkholderia chromid 2, 4 Rep proteins of Rhodobacter chromid, 7 Rep proteins of *Coxiella* are contained in each of black bars. Alphaproteobacteria in pink, betaproteobacteria in blue, and gammaproteobacteria in green. See appendix 1 for details of methods and the full tree (displaying sequence accession numbesr).

3.3.4 RepFIA Rep protein family

Although most small plasmids in general contain a single replication region, some large plasmids do contain multiple replication sites that are dispersed on plasmids, or expressed as different forms of a replication initiator. Well-known examples include F plasmids, pTB19 from *Bacillus*, pAMa1 from Streptococcus, etc. It is not clear how these two replication origins interact to start initiate replication primarily on one origin, but obviously that only one is active *in vivo*. The presence of two replication sites makes the classification based on the replication systems more difficult. It remains still unclear why these multiple minimal replicons exist in a single plasmid, but many previous works have suggested that having multiple replicons may contribute to overcoming restrictions related to replicating in narrow hosts, allowing the plasmid to locate in unrelated hosts [98, 99]. pGSH500 from *Klebsiella pneumononas* has been shown to be a good example of this as its host range actually extended to two different ranges by having two replicons [99]. However, it should also be an important fact that many plasmids having multiple replicons still remain narrow in host range, including F plasmid [100]. It has been rarely reported that naturally occurring composite replicons have an extended host range [101]. F plasmid is an IncF plasmid, and this is limited in host range to the order *Enterobacteriales*. Their replication relies on both self- and host-encoded factors, and they are usually low copy number plasmids, over 100kb. Many plasmids of the IncF group have been shown to possess more than one basic replicon. It should be noted that different types of Rep initiators (proteins) such as RepFIA Rep protein, RepFIB Rep protein that are found in the plasmids having multiple replication regions are shortly indicated as 'RepFIA', 'RepFIB', 'RepFIIA' in this chapter. RepFIA, RepFIB have been found in various combinations in this group and, depending on the third replicon, there are generally divided into two groups, IncFI and IncFII having RepFIC and RepFIIA, respectively [102]. For example, in plasmids F and p307, there are RepFIA, RepFIB and RepFIC while in pB171 there are RepFIA, RepFIB and RepFIIA (**Figure 3.9**). By contrast, plasmids R1 and R100 of *E.coli* contain only one functional FII replicon, both having neither RepFIA nor RepFIB. It has been demonstrated that FII replicons are free to diverge when associated with FIA and FIB replicons since they do not participate in the

initiation of replication of the plasmid, generating new compatible variants that can be used to overcome the incompatibility barrier with incoming IncF plasmids [102].

RepFIA is the primary replicon controlling replication of the F plasmid containing both unidirectional and bidirectional origins of replication, *oriV* and *oriS*, respectively. We have found 144 homologs across bacteria. **Figure 3.10** presents the phylogenetic tree of RepFIA homologs, which consists of three big clades. Firstly, the bottom clade consists of two groups, one is typical FI including plasmid F (NP_061424), and the other is RepFIA-like group including R27 (NP_058391) and pHCM1 (NP_569349). The latter group belongs to the IncH incompatibility group, having multiple replication regions, such as RepHIA, RepH2A or RepHIB (**Figure 3.11**). Gabant et al. [103] suggested that these replicons might have a secondary role in replication, unlike RepFIA in IncFI plasmids. There is a well-resolved clade, which consists of the IncN incompatibility group including prototype plasmids R46. It is interesting that the host range of naturally occurring IncN plasmids might be mostly restricted to gammaproteobacteria, particularly in *Enterobacterales (Salmonella, Escherichia,* and *Klebsiella)*. Various studies [104, 105] proved that IncN plasmids are able to transfer and replicate in other divisions of proteobacteria, but based on public genomes obtained so far, it does not appear to be frequent for the plasmids to actually show broad host ranges in a natural condition. Suzuki et al. [79] also suggested that the host range of IncN plasmids based on the genomic signature is limited. More specifically, their range is much narrower than IncP's.

There are more homologous sequences of the RepFIA family in the upper clades, including *Pasteurellales, Pseudomonadales, Chromatiales, Alteromonadales, Enterobacteriales, Burkholderiales* and these are distributed in beta- or gammaproteobacteria. Most clades are not clearly resolved and some homologs of RepFIA are also found in gram-positive bacteria including *Bacillus* or *Geobacillus*.

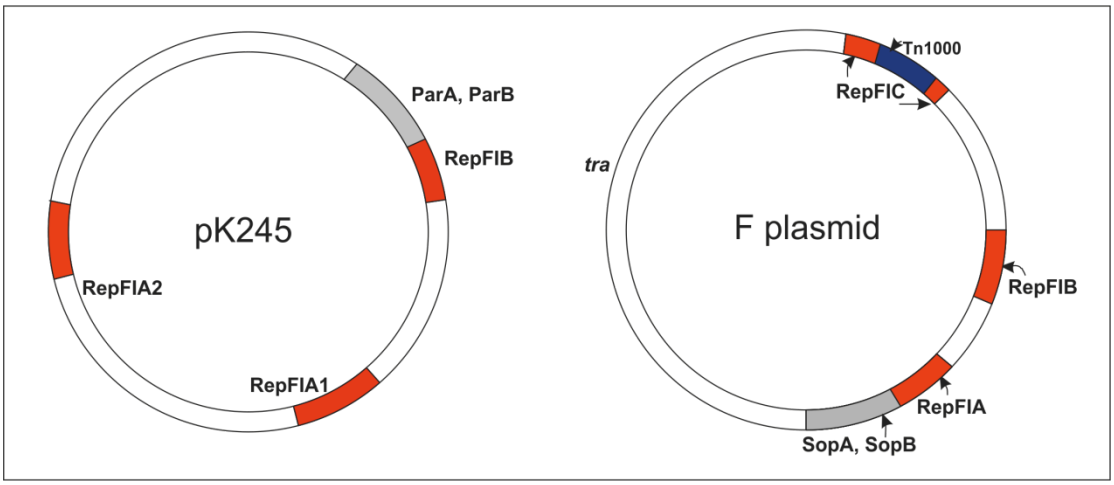


Figure 3.9 Schematic map of plasmids possessing RepFIA Rep initiators

Plasmid pK245 having RepFIA1, RepFIA2, and RepFIB [106] and F plasmid having RepFIA and RepFIB [102]. Rep in red, Par in grey, and Tn in blue.

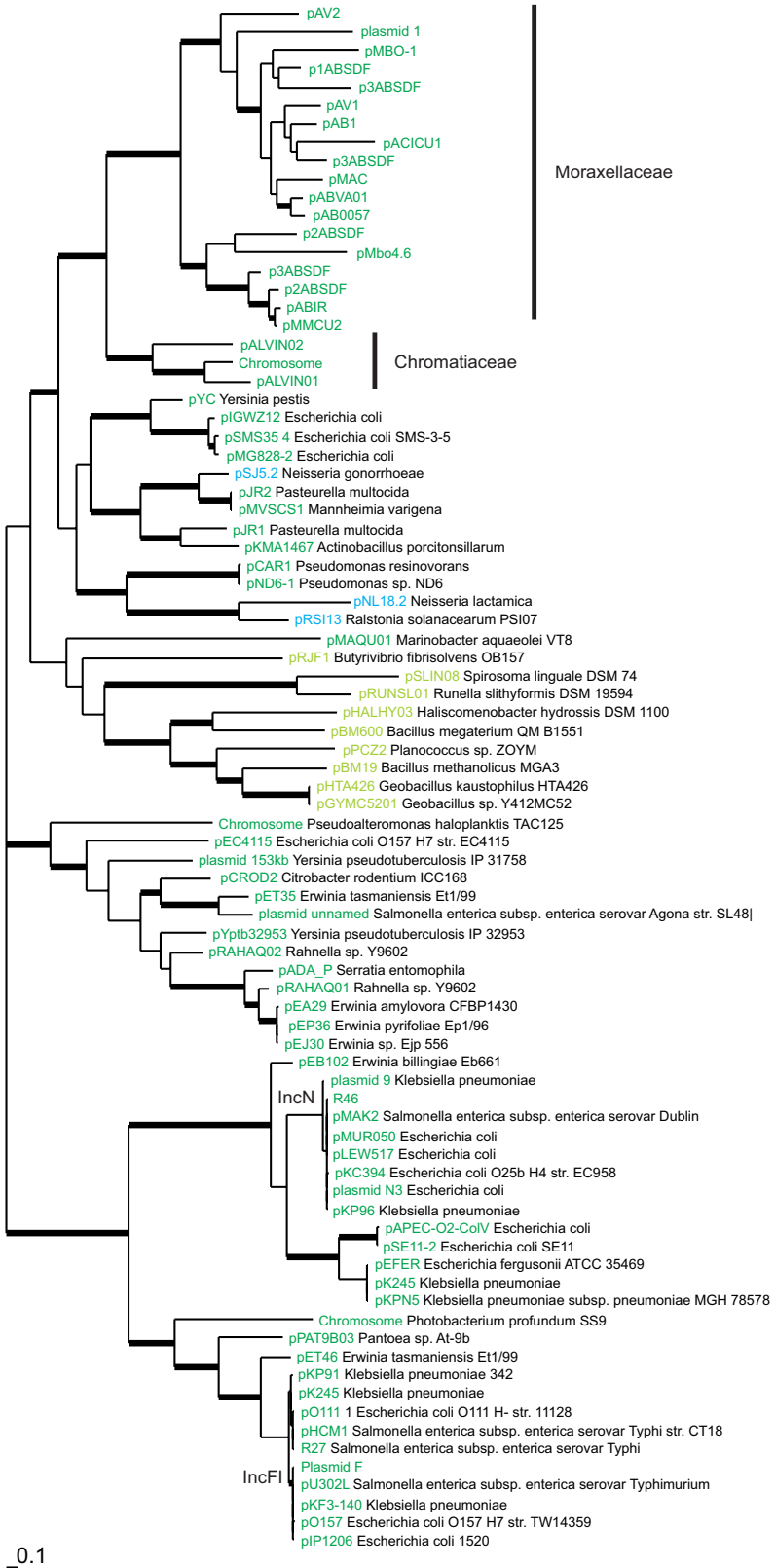


Figure 3.10 The phylogenetic tree of RepFIA Rep initiator sequences

Unrooted Maximum Likelihood tree. Betaproteobacteria in blue, gammaproteobacteria in green, and non-proteobacteria in light green. Well known incompatibility group IncN and IncFI are shown in the tree. See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.3.5 RepFIB Rep protein family

In addition to RepFIA, RepFIB replicons are also frequently found in multiple replicons. As briefly discussed in section 3.3.2.1, it has been known that RepFIB can sustain the plasmid replication in the absence of RepFIA [107]. All RepFIB homologs are present in the plasmids from gammaproteobacteria, and they include some known incompatibility group. Firstly, it is interesting that RepFIB has high homology with Rep in IncHI incompatibility groups. The IncHI plasmid is a large (>150kb) conjugative plasmid, exhibiting a thermosensitive transfer mechanism [16]. They can be divided into different incompatibility groups, HI1, HI2, and HI3, and RepHIA and RepHIB are specific for IncH1 and two proteins can efficiently replicate the entire plasmids [103]. The RepFIB family includes both RepHIA and RepHIB. RepHIA and RepHIB are similar in organization but sequences are 34.9% similar [103].

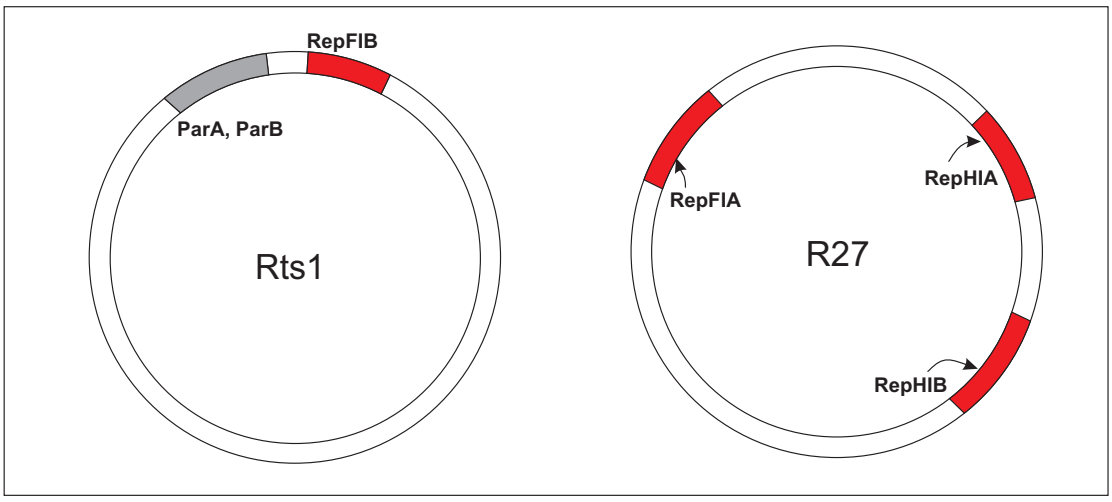


Figure 3.11 Schematic map of plasmids possessing RepFIB Rep initiators Rts1 [108] and R27 [109]. Rep in red and Par in Grey.

As discussed before, many IncHI plasmids including R27 and pHCM1 have a RepFIA-like replicon in addition to RepHIA or RepHIB. The RepFIA-like sequence of IncH is highly similar to RepFIA in IncF plasmids in a same clade of **Figure 3.10**, which appear to be incompatible. However, RepFIB of F plasmids and RepHIA, RepHIB of IncH plasmids are different. **Figure 3.12** is the phylogenetic tree based

on RepFIB homologs. The upper clade contains RepFIB in F plasmids and pB171, and each group regarding RepHIB, RepHI1A and RepHI2A is distributed in different clades with high bootstrapping number. Gabant et al. [103] have explained that there is a one way incompatibility, that is when IncHI plasmids go into a host having F plasmid, the latter is segregated, while when F plasmids go into the one having IncHI plasmids, they are compatible. This phylogeny might explain that because the main replication system of F plasmid is RepFIA, they cannot survive if IncHI plasmid comes, but IncHI plasmids can survive even if an F plasmid comes because RepHIA or RepHIB is located in a different clade phylogenetically.

In addition to IncFI and IncHI, there is also the IncT incompatibility group including prototype plasmid, Rts1, from *Proteus vulgaris*, which also appears to be related to the plasmid of *Pantoea vagans* C9-1.

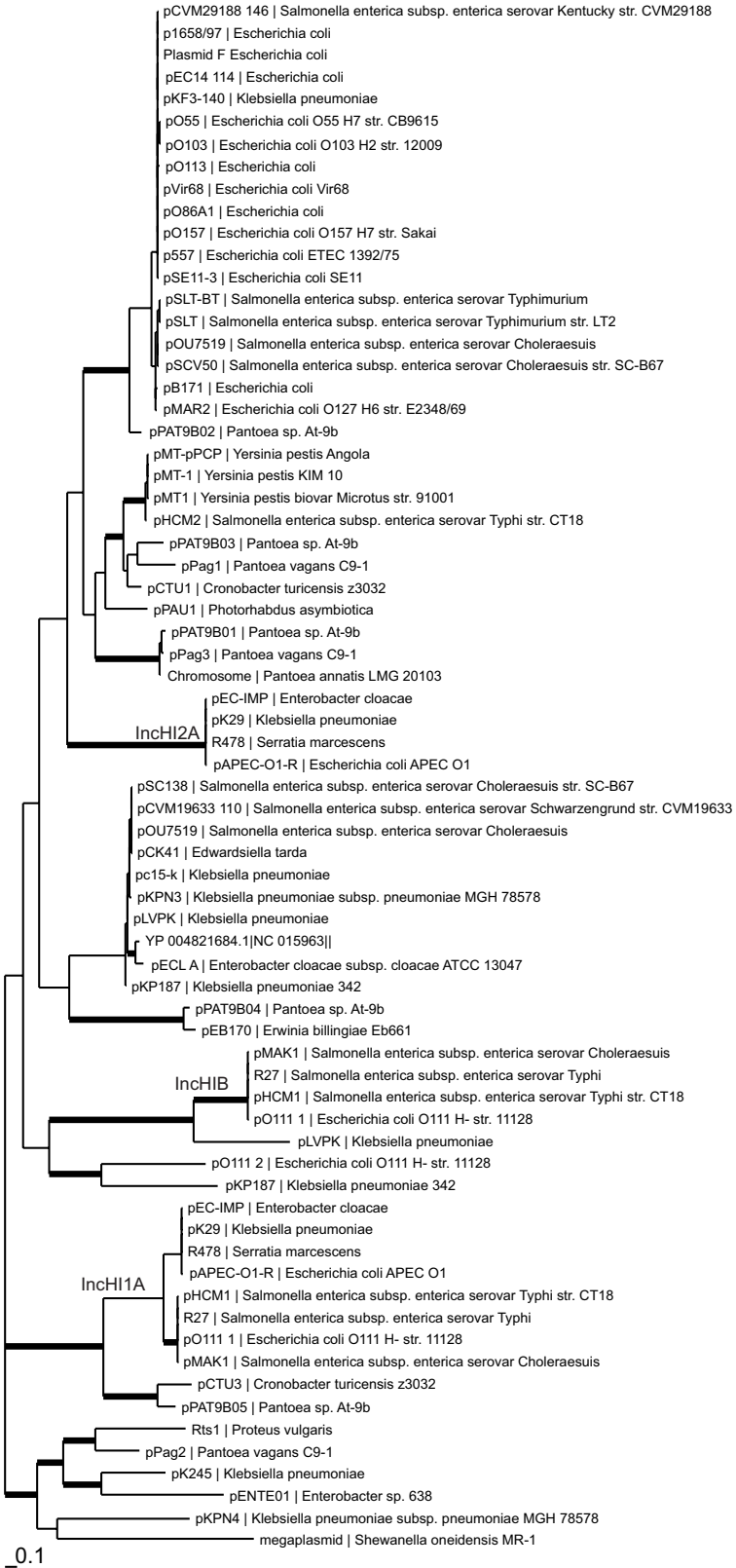


Figure 3.12 The phylogenetic tree of RepFIB Rep initiatorsequences

Unrooted Maximum Likelihood tree. The plasmids defined as IncHI1A, IncHI2A and HIB are indicated in each clade. See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.3.6 RepFIIA Rep protein family

We have found 136 sequences of the RepFIIA family in plasmids. Many belong to the IncFII plasmids, including the prototype R100. Circular maps of example plasmids having RepFIIA are shown in **Figure 3.13**. **Figure 3.14** illustrates the phylogenetic tree of the RepFIIA sequences. It shows three clear clades, A, B and C. The upper clade is separated into two sub-clades. The first clade is composed of well-known IncFII plasmids including pB171, NR1 and R100, and their sequence variation is very low. Hosts of those are mostly *E.coli* but can also include *Shigella, Salmonella* and *Citrobacter*. Subclade The second clade can be divided into three small clades, which are resolved at a genus level. In the second clade of the upper clade, there are pKPN3 and pKPN4 in *Klebsiella pneumoniae* in the same clade, and this is against the compatibility rule. **Figure 3.13** explains this by showing that RepFIBs of the two plasmids are located in separate clades. The main replication module of pKPN3 and pKPN4 is RepFIB.

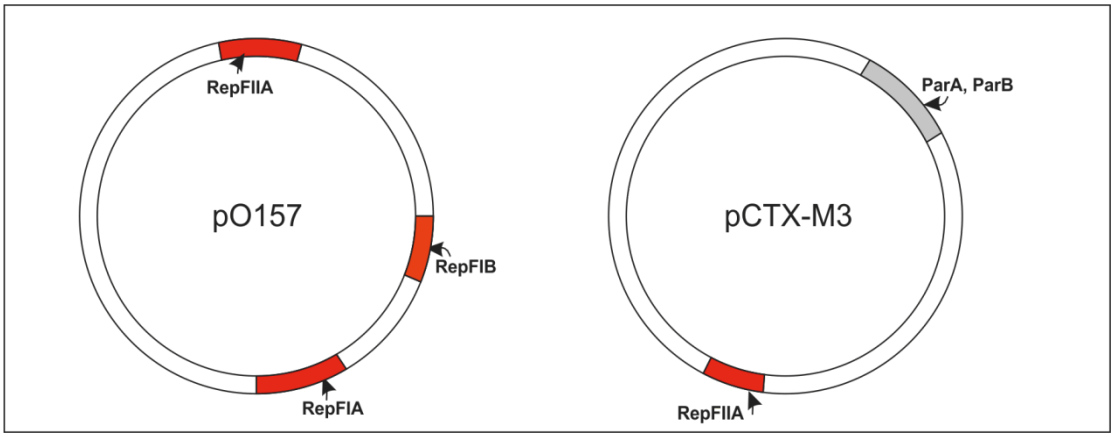


Figure 3.13 Schematic map of plasmids possessing RepFIIA Rep initiators

pO157 [110] from *E.coli* and pCTX-M3 [111] from *Citrobacter freundii*. Rep in red and Par in grey.

In the middle clade, RepFIBs of plasmids hosted on *Buchnera* and *Sodalis* are located, which are also resolved at a genus level. Osborn et al. [87] called this group as an 'extended family of RepA proteins from IncFII-related replicons', which also appears to have a long related history. The lower clade contains two

known incompatibility groups: IncIα-1, 2 including Collb-P9, R721 and IncL/M including pCTX-M3. There are two replication initiators of plasmid pCoo in subclades A-1 and C-1. Froehlich et al. [112] showed that pCoo is formed by the recombination of two independent replicons because the two pCoo origins are separated by long (1953 bp) direct repeats comprising recent IS100 insertions. It appears that one of two replicons is silent during replication [112].

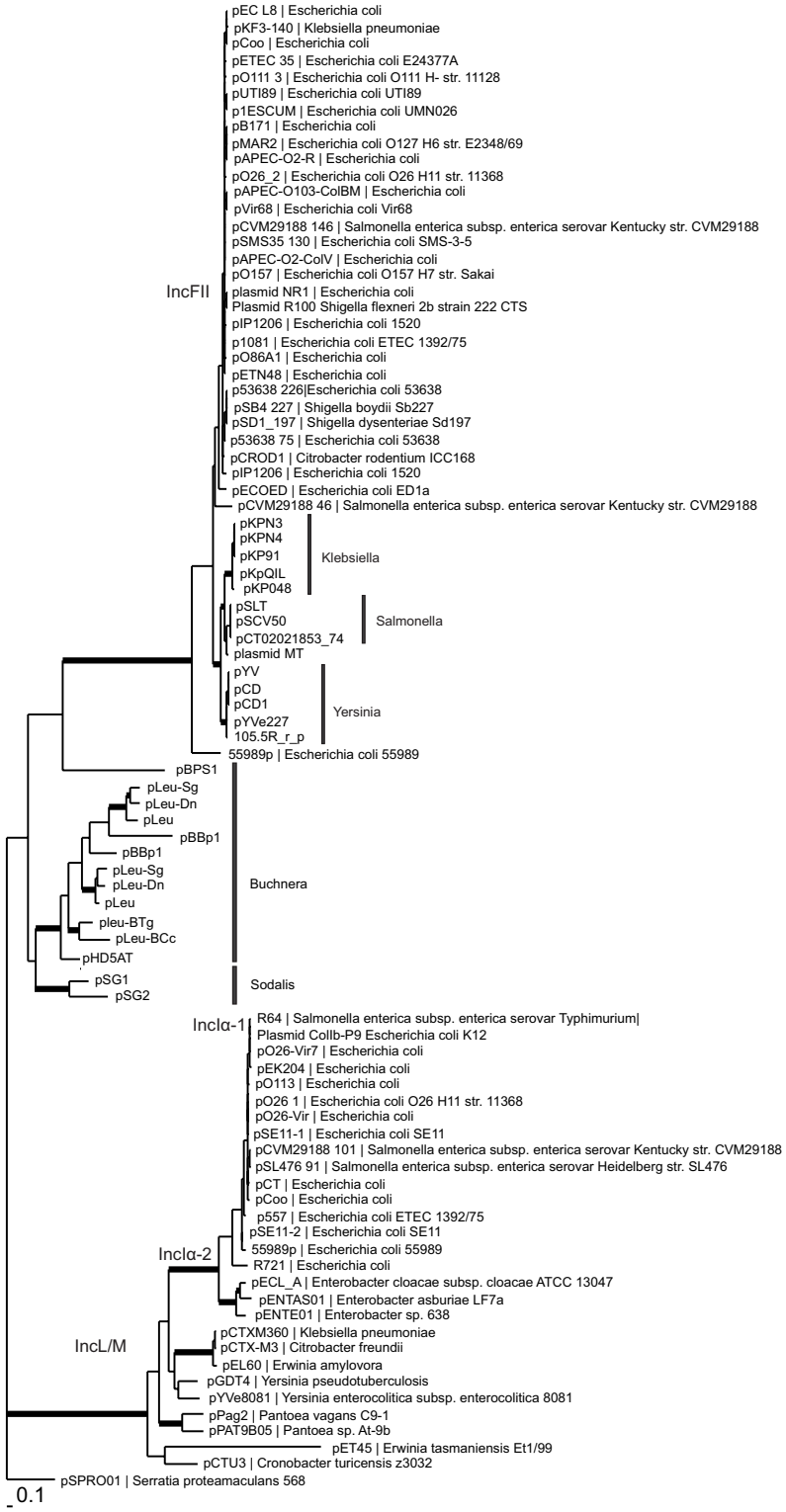


Figure 3.14 The phylogenetic tree of RepFIIA initiator sequences

Unrooted Maximum Likelihood tree. Well known incompatibility groups (Inclα, IncL/M, IncFII) are indicated in different clades. Only genus name is shown if all plasmids are from the same genus in the clades (by bars). See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.3.7 TrfA family

This section discusses broad host range (BHR) replication initiators. BHR plasmids exist in a wide range of bacterial hosts because they have an ability to replicate and transfer, and are therefore often called “promiscuous” plasmids. Initially, it was suggested that IncP plasmids were able to exist in not only *Enterobacteriaceae* but also *Pseudomonas* species. However, more recent work has demonstrated that many plasmids are able to replicate and transfer in other subclasses of proteobacteria, including gamma- and betaproteobacteria. Meyer et el. [113] and Guiney et al. [82] proposed that broad and narrow host range are related to their replication systems, particularly, the restriction site in the replication systems, which means BHR plasmids have lower numbers of restriction sites in replication genes than narrow host range (NHR) plasmids. Therefore, BHR plasmids overcome the restriction barrier of the host cells more easily.

The majority of plasmids harboring the *trfA*-related replication system are plasmids belonging to IncP group. IncP plasmids are one of the most promiscuous groups of plasmids, as they are able to transfer and maintain within various divisions of bacteria in different geographic locations. IncP plasmids have been extensively studied as they often carry phenotypic genes such as antibiotic resistance or catabolic pathways. It has been known that their ability to transfer and replicate enables them to proliferate in gram-positive bacteria, cyanobacteria, or even yeast [80, 114, 115].

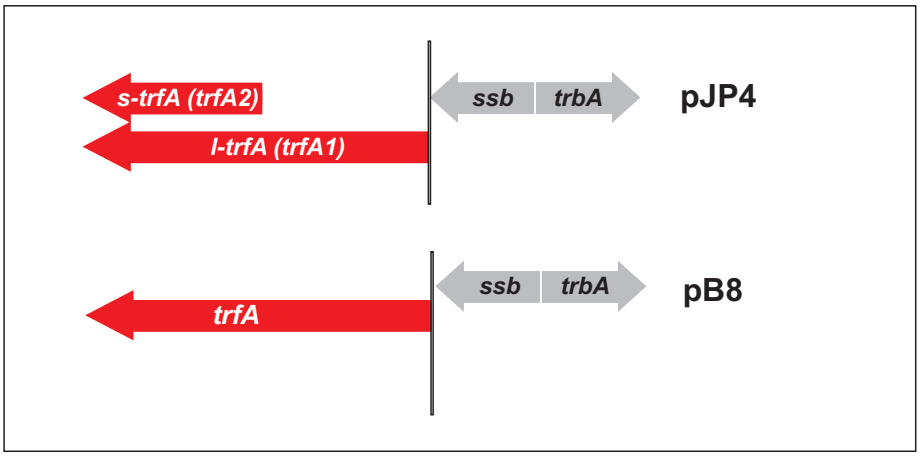


Figure 3.15 The genetic organization of plasmids having TrfA initiators

pJP4 from *Ralstonia eutropha* JMP134 [116] and pB8 from *Pseudomonas sp.* B13 GFP1 [117]. Rep in red and the genes adjacent to Rep in grey. *trfA* of pJP4 contains two translational starts, generating TrfA1 and TrfA2. *ssb* encodes a putative single-stranded DNA binding protein and *trbA* encodes a conjugal transfer protein

Many researchers have been trying to classify IncP plasmids (also known as IncP-1 in the Pseudomonas classification) based on their phylogenetic analysis and recently, division into 5 subgroups (α to ε) has been suggested. This includes; RK2 (IncP-α), R751 (IncP-β), respectively [82], pQKH54 (IncP-γ) [118], pEST4001 [119], and pKJK5 (IncP-ε) [120]. Stenger and Lee [86] have recently suggested a γ-expanded group, which is a possible new group in IncP plasmids. Norberg et al. [121] also mentioned in their research that the ζ-subgroup including pMCBF1 was formed as a novel clade in the phylogenetic tree of IncP plasmids. The list of naturally occurring IncP plasmids is growing continuously, so are the subgroups.

Based on the prototype of IncP plasmids, RP4 and R751, the replication is initiated at a single *cis*-acting origin, *oriV* and the plasmids mostly rely on a theta replication mode [122]. *trfA* (**Figure 3.15**), just adjacent to *oriV*, encodes the TrfA protein, which is used as not only a positive initiator of replication, but also as a negative regulator of plasmid copy number. Actually, TrfA is the only protein required for replication that plasmids encode [123]. Interestingly, a trans-acting gene *trfA* sometimes encodes two proteins, TrfA1 (long TrfA, known as TrfA44) and TrfA2 (short TrfA, known as TrfA33) because there are separate in-frame translation

start sites [124, 125]. Normally IncP plasmids belonging to α, β, and ε subgroups encode two TrfA proteins while the plasmids involved in subgroup γ and δ just encode only one TrfA. It has been known that the short TrfA, which is highly conserved, is sufficient for plasmid replication in many hosts, including *E. coli* and *Pseudomonas putida.* Although the exact role of the long TrfA has been unclear, early research revealed that it appears to be involved in replication and maintenance, particularly in *Pseudomonas aeruginosa* [124, 126, 127]. Jiang et al. [128] have shown that long TrfA seems to load and activate the DnaB helicase of *P. aeruginosa* or *P. putida in vitro* in the absence of the DnaA protein. Short TrfA protein, however, requires DnaA protein to load and activate the helicase of *P. putida* and requires DnaA plus DnaC to load the helicase of *E. coli.* Furthermore, recent research by Yano et al. [129] argued that TrfA promotes transformation efficiency and plasmid copy number, although it does not seem to affect long-term plasmid persistence.

Figure 3.16 illustrates a phylogenetic tree constructed based on the amino acid sequences of the TrfA family (Note that we have chosen long TrfA protein sequences). The distribution of the TrfA family is very wide as we expected, from well-known gammaproteobacteria such as *Klebella, Salmonella*, and *E.coli*, to betaproteobacteria including *Ralstonia* and *Burkholderia*. Interestingly, plasmids found in *Sphingomonas* in alphaproteobacteria are also members of this family. Although IncP plasmids have been experimentally proved to be able to transfer and replicate in any division of proteobacteria, they have not been frequently detected in alphaproteobacteria [80, 130, 131].

As shown in the upper clade, IncP plasmids can be divided into 5 groups from α- to ε -sub division as previous research supports, and all clades are well resolved with high bootstrap values. In addition, there are other clades having plasmids of TrfA-related replication systems and most of these are found in betaproteobacteria. Those replication systems have clear conserved domains, which indicate a possible related history with TrfA proteins. However, the partitioning systems are short ParA-ParB couples rather than IncC type proteins (not shown), indicating their partitioning systems might have evolved (Chapters 4 and 5 provide a more detailed discussion on this matter).

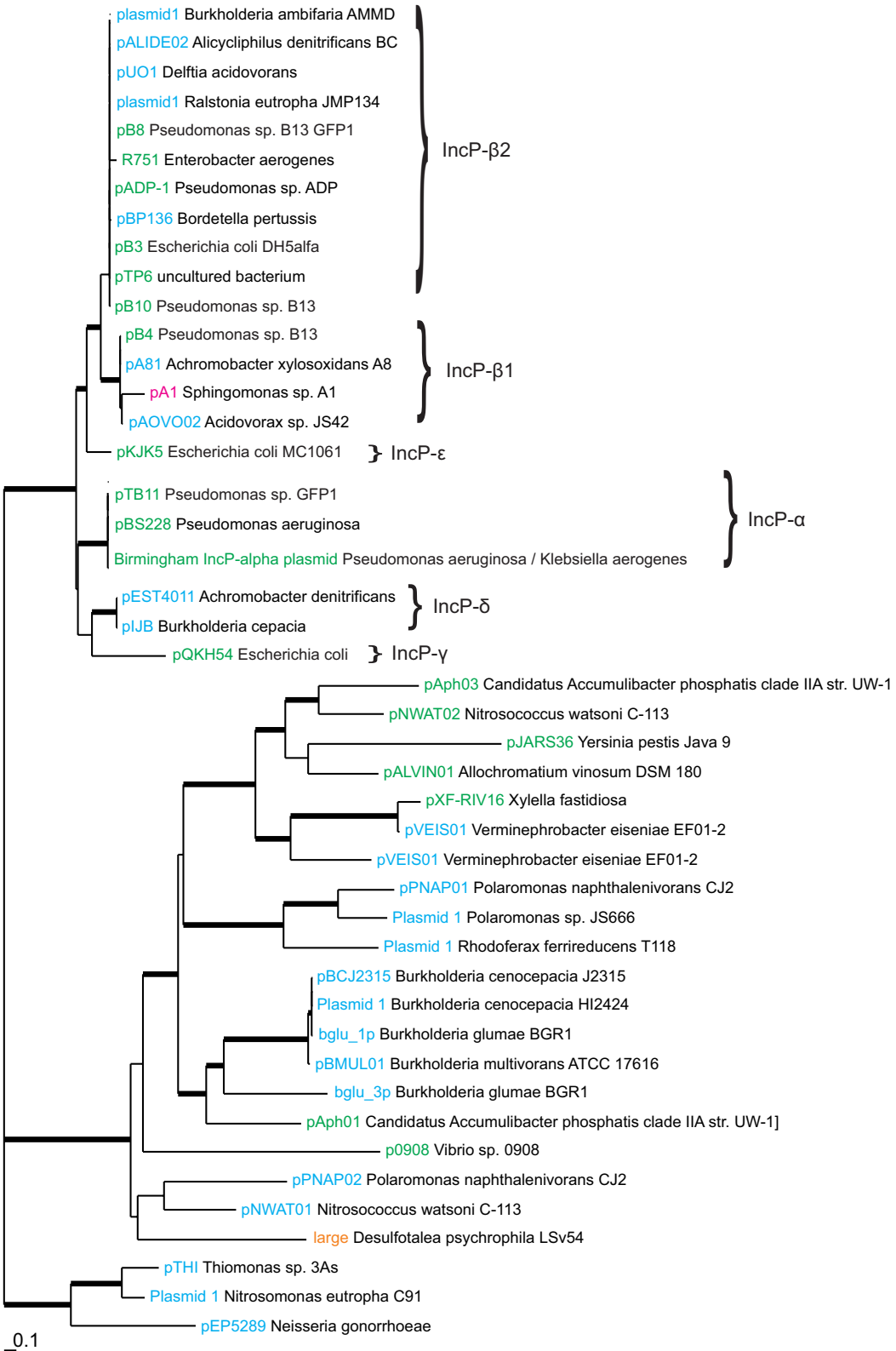


Figure 3.16 The phylogenetic tree of TrfA sequences

Unrooted Maximum Likelihood tree. Alphaproteobacteria in pink, betaproteobacteria in blue, gammaproteobacteria in green, and deltaproteobacteria in orange. Five divisions of IncP plasmids are shown in the upper clade. See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.3.8 RepA family

The replication region in this section is generally located in IncW plasmids, and commonly called RepA. IncW plasmids show the smallest size of conjugative plasmids among bacteria so far, and their copy numbers are generally low [132]. The IncW plasmid is also a well-studied group in bacteria because of interesting characteristicsin terms of the genes they confer, and their host range. They have homology in replication and transfer regions but carry a variety of different antibiotic resistance genes. They have also received the researchers' attention because some IncW plasmids have the ability to inhibit tumor induction by *Agrobacterium tumefaciens* [133]. Hence, they were frequently used as a vector for cloning and analysis of Ti plasmids.

Watanabe et al. [134] firstly isolated the IncW plasmid, pSa, from *Shigella* in Japan and R388 from *Escherichia coli* [135], R7K from *Providencia rettgeri* [136] were later also isolated. Most of IncW plasmids were isolated from proteobacteria, but they have been experimentally proved to be able to transfer and stabilize in various bacterial hosts, therefore, IncW plasmids are considered BHR plasmids. However, Bradley et al. [137] confirmed that they could only conjugate when they mate on solid surface, rather than in liquid.

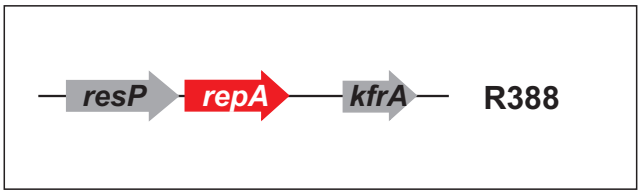


Figure 3.17 The genetic organization of the replication region having the RepA initiator

The region shown is part of the R388 plasmid [83]. *resP* encodes ResP providing a resolution activity for plasmids dimers. *kfrA* is a DNA binding protein related to plasmid stability. Rep in red, other genes related to replication in grey.

Based on the prototype IncW plasmid, pSa and R388, the replication region is known as generally consisting of *oriV* and two ORFs to be translated (**Figure 3.18**). The replication starts from the origin of replication site, *oriV*. One of the two

operons, *resP*, encodes ResP (218aa), which provides a resolution activity for plasmids dimers resulting from replication. *resP* is coupled to gene *repA*, which codes a replication initiator protein (323aa) [138]. The replication initiation protein, RepA, does not show sequence similarity to that of other broad host range plasmids.

By comparison with the replication genes of other plasmids, RepA of IncW shows an extremely wide distribution in various species from gram-negative to gram-positive bacteria. **Figure 3.19** illustrates the phylogenetic tree of RepA homologs of IncW plasmids showing diverse spread across bacteria. In general, various plasmids replicate by RepA homologs hosting in various kinds of proteobacteria such as *Salmonella enterica, Providencia rettgeri,* (gammaproteoacteira), *Polymorphum gilvum* (alphaproteobacteria), *Pelobacter propionicus* (epsilonproteobacteria), *Burkholderia pseudomallei, Burkholderia glumae, Ralstonia solanacearum* (betaproteobacteria), and so forth. In particular, some chromids and plasmids of β-proteobacteria, including Cupriavidus and Ralstonia, have this type of replicase. Apart from proteobacteria, RepA homologues are also distributed in Actinobacteria such as *Bifidobacterium longum, Kineococcus radiotolerans* or Acidobacteria such as *Grabulicella tundricola*.

There is clade B that is composed of the prototype IncW plasmid pIE321 of *Salmonella enterica* subsp. *Enteria* serovar Dublin in the middle section, which is the identical sequence of the prototype IncW plasmids R388 and pSa. Several close homologs are distributed in plasmids and chromosomes in beta- or gammaproteobacteria though they are not well resolved in the phylogeny. Above this clade, clade A consists of pIPO2T and pSB102, and pTer331, which were recently proposed as the PromA group [139]. It is interesting because their partitioning system is related to those of IncP plasmids, but replication modules are clearly involved in the IncW type replication system (This is discussed in more detail in chapters 4 and 5.).

Moreover, many gram-positive bacteria also have plasmids replicated by this type of *repA* homolog, such as *Rhodothermus marinus* and *Corynebacterium striatum*. As shown in the tree, plasmids from closely related genera do not cluster together suggesting a history of horizontal gene transfer between distant bacterial species. In the upper clade, although the level of homology is quite low, some

gammaproteobacterial plasmids (from *Xanthomonas*) and alphaproteobacterial plasmids (from *Acetobacter*) also have homologs of RepA replicase, which might indicate a possibility of related evolutionary history.

Figure 3.18 The phylogenetic tree of RepA sequences

Unrooted Maximum Likelihood tree. Alphaproteobacteria in pink, betaproteobacteria in blue, gammaproteobacteria in green, and deltaproteobacteria in orange. Other plasmids or chromosomes not belonging to proteobacteria are in black. Recently proposed plasmids as the PromA group are indicated as the clade A and IncW prototype plasmids are in the clade B. See appendix 1 for details of methods and the full tree (displaying sequence accession numbers).

3.4 Discussion

3.4.1 Contribution of this chapter

Thanks to the development of these numerous sequencing projects a large volume of publicly available genomic information has accumulated from genome projects, and so have plasmid genomes. However, relatively little has been done with them, although the phenotypes that plasmids confer such as antibiotic resistance and symbiotic ability have been studied partially. The heyday of experimental plasmid research was in the 1970s and 1980s.

However, additional effort is required to deal with large amounts of genomic data, which should be explained by biological meanings. In particular, there has not been comprehensive analysis in order to investigate plasmid backbone systems based on the modules involved. This is the first research to investigate a variety of plasmid replication systems using public genomes from NCBI, to identify major families based on the Rep initiators, and to classify plasmids comprehensively in this regard. Although some previous researches such as pfam, NCBI protein clusters, or ACLAME, etc. suggest protein clusters of the replication systems, it still remains to investigate each family in detail, particularly based on the phylogenetic analysis, host range, and distribution.

This study generally covers major groups of plasmid replication systems in proteobacteria. We analysed 8 families: RepC, RepA-like, RepB-like, RepFIA, RepFIB, RepFIIA, TrfA, and RepA. In detail, the families cover over 50% of alpha- and betaproteobacterial plasmids, and 44% of gammaproteobacterial plasmids. We did not look into delta-, and epsilonproteobacteria because their sequence information is relatively limited compared to other divisions of proteobacteria.

3.4.2 Plasmid replication systems in proteobacteria

3.4.2.1 Plasmid movement is active in bacteria?

As we have shown, there does not seem to be frequent movement of plasmids in different proteobacterial divisions, based on the analysis of Rep, except for broad

host range IncP plasmids and IncW plasmids. The fact that plasmids show mosaic structure composed of a diverse set of genetic elements indicates that movement through plasmids may be very extensive. However, it does not seem very common that many plasmids actually show environmentally broad host range, especially over class or phylum level. As seen in section 3.3.2.1, BHR plasmids IncN do not show a wide range of distribution in a natural environment, even if it has been experimentally proved that they are easily transferred to other hosts in a laboratory environment.

Nevertheless, plasmids do move at least in related hosts, particularly within a order or class level. We have illustrated the closely related history of RepCs in *Rhizobiales, Rhodobacterales*, etc. in **Figure 3.4**. Moreover, there are several cases where two plasmids that are found in unrelated niches are contained in the same clade, which indicates a common origin (**Figure 3.6** and **Figure 3.8**). Obviously Reps of BHR plasmids are distributed across different divisions of bacteria (**Figure 3.16** and **Figure 3.18**).

3.4.2.2 How many replication initiators are there in bacteria?

In this chapter, we have covered nearly half of the proteobacterial plasmids (520 among 1112 plasmids) based on the 8 major families. These major families have been selected because their members are abundant in the various hosts, and are also among the relatively well-characterized families. However, it is clear that there are more families in proteobacteria that can be categorized in addition to the main families. The number of members in one family might not be particularly large at the moment, but it is still worth looking into them in detail. For instance, Petersen et al. [140] demonstrated that DnaA-like replication systems in marine bacteria can be categorized as a novel system, because their homology is not related to Reps from other plasmids in the same species. Although it is not necessary to categorize all remaining initiators into a number of small groups, more sequencing work published in the future will reveal a new family.

In addition to plasmids from proteobacteria, there are more plasmids such as plasmids from Firmicutes, which are gram-positive bacteria. Moreover, IncQ plasmids are also well studied BHR plasmids although there are not many in

sequenced genomes, and as such, they are not included in this study. The mechanism of these plasmids is mostly the RCR mechanism, or strand displacement mechanism, which is different from those that are more common in proteobacteria (mainly theta mechanism). However, most of those plasmids also have Replication initiators, so that they can also be characterized based on the same process of our research. Replication systems are significant for classification of plasmids because they are constantly present except for few cases (this is discussed in more detail below).

3.4.2.3 Is the replication system a good method to classify plasmids?

We have highlighted the significance of the classification of plasmids in section 3.1.2. In general, it appears that Rep initiators can be used as a good marker for classification. Particularly, phylogenetic analysis might be a possible indicator of the actual incompatibility of plasmids. For example, RepC is a good marker for classification of alphaproteobacterial plasmids. Based on the fact that most plasmids from alphaproteobacteria in **Figure 3.4** were isolated naturally, each clade is a good source to indicate actual incompatibility. In case of RepA-like, RepB-like initiators, it appears that unrelated plasmids are in the same families. However, their initiators obviously have sequence homology with each other, and clades in the phylogeny indicate meaningful information. There are some exceptions that two plasmid systems in the same strain are actually similar to each other (red circles in **Figure 3.4**), which is against the incompatibility rule. However, this suggests that the classification might be improved by considering partition systems (see chapter 4).

We have discussed a complicated situation in the plasmids having multiple sites of replication (section 3.2.2), which smakes it difficult to classify the plasmids based on the Rep regions. Some phylogenetic results might be supporting their incompatibility groups, but obviously further experimental results should be produced in this direction.

3.4.3 Limitation of this work and future direction

As described above, there are several facts that should be considered carefully in this work. First of all, the possibility that data are biased. Our initial aim was to investigate general distribution of plasmid Rep systems using a public database (NCBI). Based on the assumption that there are a substantial number of plasmid genomes (2511 plasmids up to Oct 2011), we defined our major groups based on Reps that are abundant in bacterial plasmids. However, it is also true that they might be biased because each division of plasmids has not been equally sequenced. In particular, there have been various researches that look into gammaproteobacterial plasmids because of their antibiotic resistance. Relatively more plasmids were a target of research, and presumably more plasmids must have been sequenced. Therefore, there is no guarantee that the major groups that we defined are naturally more abundant than others. It is clear, however, that NGS technology will continue to generate large amounts of information, which will shed light on more general ideas. That is why we have developed HMMs in order to screen and add more members to the main families continuously (chapter 2).

Secondly, we have demonstrated that many plasmid replication systems seem to have evolved to work with partitioning systems, particularly in the case of large conjugative plasmids. These are frequently located adjacent to each other, while sometimes the partitioning systems act as a regulator. Previous work has revealed several cases regarding their possible relationships. This is why we are going to look at plasmid partitioning systems, in order to obtain a better insight on their evolutionary history (chapter 4).

Finally, there have been a couple of recent efforts to deal with increasing plasmid data *in silico,* which would be useful for this study. A representative of these efforts is the research from Smillie et al. [45]. They suggest a method to classify plasmid mobility automatically based on VirB4- and T4SS-based searches. This analysis covers plasmids of proteobacteria. Suzuki et al. [79] also used public genomes in order to investigate plasmids' genomic signature and their evolutionary host range. Several recent papers have examined plasmids in terms of a specific genus, family or incompatibility group, including Loftie-Eaton and Rawlings et al. [72]. It

is expected that more efforts would be followed to study plasmids backbone systems, in which *tra* would be one of the future direction, for instance.

Chapter 4. Diversity of plasmid partitioning systems in proteobacteria

In addition to the plasmid replication systems discussed in chapter 3, the correct partitioning of two plasmids into daughter cells after replication is an also essential maintenance process. This is called the active partitioning (segregation) process, and is particularly significant for low-copy number plasmids and for chromosomes. Failure to perform the partition may result in instability of the host bacterium or in reduce fitness [141]. In the last couple of decades, plasmid partitioning systems have been studied extensively in terms of the process of the systems, genes involved in them, etc. However, *in silico* analysis still deserves more attention because no comprehensive analysis of the systems has been described so far. This chapter aims to investigate the partitioning systems in proteobacteria.

4.1.1 Plasmid partitioning systems

There are several mechanisms to ensure the accurate stability and maintenance especially in large low-copy number plasmids (see chapter 1). Best examples include multimer resolution systems, copy number control, post-segregational killing systems, and active partitioning systems [142-145]. The active partitioning systems are the most extensively studied amongst the examples, which ensure that plasmids are separated to each daughter cell, in short called the 'Par'system.

The Parsystem usually requires two plasmid-encoded *trans*-acting partition proteins and one *cis*-acting DNA sequence[146]. The two partitioning genes are organized in tandem as an operon and are transcribed from a common transcriptional promoter [145]. The upstream gene encodes an ATPase called 'ParA', while the one in the downstream encodes a DNA-binding factor called 'ParB'. Generally, ParB interacts directly with the '*parS*' site, which is normally composed of direct or inverted iterated sequences [145, 147, 148] that also serve as the incompatibility determinant in a number of plasmids [149-152].

Hayes and Barilla [40] suggested the model for the process of active partitioning systems in plasmids (**Figure 4.1**). After the plasmids are replicated successfully, they are first aligned in the mid-point of the cells (1). ParB (red squares) binds to *parS* (yellow circles) forming a 'segrosome', which begins to recruit ParA proteins (blue circles in **Figure 4.1**) (2). The ParAs polymerize toward the end of the cells with help from the ATP, pushing the two plasmids apart (3), which results in the separation of the plasmids into two daughter cells (4). Following this successful distribution, the ParAs become disassembled (5).

Depending on the type of ATPase ParA involved, active partitioning systems are divided into several distinct groups (**Table 4.1**). So far, previous research has shown that there are three main types: Type I Walker-type ATPase ParA and its partner ParB, (which is later divided into two subtypes ParA Ia and ParA Ib)[145], Type II Actin-like ParM and ParR[153], and Type III TubZ and TubR [39, 143].

Simpson et al. [154] suggested a single gene responsible for segregation in *Staphylococcus aureus*, which is often classified as a fourth type of partitioning system. This classification is based on the experimental results, rather than *in silico* analysis. The following sections explain in detail regarding each group of plasmid partitioning systems that are studied before.

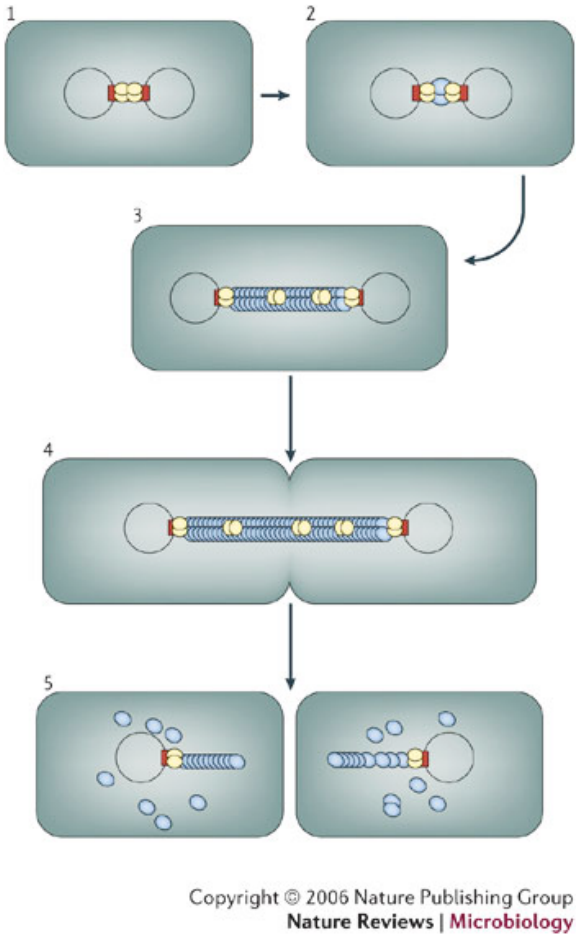


Figure 4.1 Speculative model for the plasmid partitioning process

Replicated plasmids are first aligned in the mid-point of the cells. ParB (red square) binds to parS (yellow circle) and they begin to recruit ParA proteins (blue circles) The ParAs polymerize toward the end of the cells, pushing the two plasmids apart. The ParAs become disassembled. Taken from Hayes and Barilla [40].

Table 4.1 Discrete types of plasmid partitioning systems

Based on the ATPase involved and related references. Types I to III have been characterized extensively so far. Type IV has been suggested by Simpson et al. (2003) to be a single gene for partitioning (*).

Type	Cytoskeletal element, DNA-binding protein	Altenate name	References
I	Walker-box ATPase ParA and ParB	SopA and SopB	Mori et al. [155]
		IncC and KorB	Bechhofer and Figurski [156]
		Soj and Spo0J	Mysliwiec et al. [157] Ogasawara and Yoshikawa [158]
		RepA and RepB	Turner et al. [159]
		ParF and ParG	Hayes and Barrila [40]
II	Actin-like ATPase ParM, ParR	ParM and ParR	Moller- Jensen et al. [160] Mohl and Gober [161]
III	Tubulin-like GTPase TubZ, TubR	-	Larsen et al. [162]
IV*	Single gene *par*	-	Simpson et al. [154]

4.1.1.1 Type I Walker-type ATPase

Type I *par* operon consists of a *parA* gene encoding ParA that has Walker box ATPase motifs and a *parB* gene encoding the DNA-binding protein ParB. This is further divided into two types depending on the site of the Par proteins, the localization of the partition site, and the mechanism of transcriptional regulation. ParA of Type Ia is constituted normally of ~300 to ~400 amino acids located upstream of the partition operon and acts not only as an ATPas*e*, but also as a repressor of the operon. ParB comprises of ~320 to ~360 amino acids acting as a scaffold for segrosome assembly. ParB binds to centromere *parS* to make a segrosome, which recruits a ParA protein [145].

The type Ib family of partitioning systems includes ParA homologs that are called ParF (they are sometimes known as short ParAs, see section 4.3.1.3 for more details) and DNA binding proteins, usually called ParG. ParFis half the size (~200 amino acids) of the prototypical ParA protein encoded by the plasmid P1 (Type Ia), which does not have a helix-turn-helix (HTH) motif in the N-terminal region.*parG*

encodes a small protein of 80-90 amino acids having no homology with other ParB-like proteins. ParG is dimeric and interacts with ParF and *parC (parS)* [145].

4.1.1.2 Type II Actin family

The ParMRC locus was originally isolated from the large, low-copy number, multiple-antibiotic resistant plasmid R1 from *E.coli* [152, 163]. ParM, also known as StbA, is an actin-like ATPase [160] that consists of about ~300 to ~350 amino acids, and ParR (known as StbB) is a DNA-binding protein, which interacts with the centromere-like *parC* DNA region [144, 164]. ParMs are relatively unstable with ATP, although the ParR-parCcomplex helps the filaments not to be disassembled by capping them [144, 165], which results in the separation of the plasmid copies to the opposite poles [153, 160].

4.1.1.3 Type III TubZR family

TubZ and TubR proteins encoded by the pBtozis plasmid from *Bacillus thuringiensis* have been identified and suggested as the third type of partitioning systems. TubZ, which is homologous to FtsZ, has a strong GTPase activity. It comprises an operon with a predicted small DNA-binding protein TubR, as well as a putative centromere region, *tubC.* TubR binds these centromere sites having four repeat sites in the plasmid and recruits TubZ, which forms a filament to segregate plasmids to the cell pole [166] by treadmilling elongation [162].

4.1.1.4 Type IV

Although its existence is still controversial, there is a distinct gene responsible for active partitioning systems, which can be categorized as Type IV. This type of gene has no homology with previously identified types of partitioning genes. Simpson et al. [154] identified the Type IV system from *Staphylococcus aureus.* In this system, only a single protein-encoding gene is needed for segregation, which means that partitioning is accomplished in the absence of ATPase activity. This is significantly

different from the other types of partitioning systems. The gene is located upstream of the *rep* gene, while it appears that plasmid maintenance is independent from plasmid replication [154].

4.1.2 Chromosomal partitioning systems: Soj / Spo0J coupled proteins

Although the majority of the coupled genes *parA* and *parB* for the active partitioning systems are located in low-copy number plasmids, there are many such homologous genes in bacterial chromosomes. Studies of sporulation in *B. subtilis* were important in understanding the mechanism of action of the partitioning systems in chromosomes [167]. When the bacterium has a difficulty with environmental stress, it tries to keep DNA away from the stress until it is safe [168]. For this reason, the bacterium forms 'spore' and this developmental process is called 'sporulation'. After replication, a cell is divided asymmetrically (**Figure 4.2**), which generates two parts: a spore and a mother cell compartment. The smaller forespore at first possess 30% of chromosomes, but both the spore and mother compartment have complete chromosomes eventually when the forespore becomes a spore. In this process, two types of proteins are involved in: one is for localization of chromosomes after replication, and the other is for transferring remaining DNAs [169].

Here two proteins involved in localization are highly homologous with plasmid partitioning systems: Soj and Spo0J. Spo0J and Soj are related to early acting sporulation genes. In vegetative growth, *oriC* is condensed by Spo0J. After onset of sporulation, Spo0j binds to *oriC* forming a nucleoprotein couple. Like ParA, Soj drives the *oriC*-Spo0J complex pole-ward. Studies have shown that Soj and Spo0J couple can function in plasmid partitioning as well, particularly they are significant for chromosome organization and condensation [170].

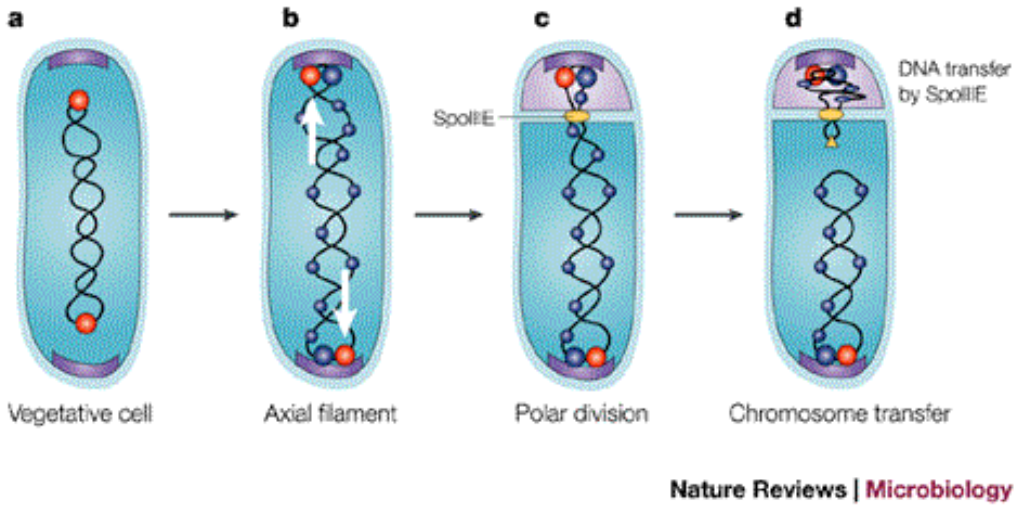


Figure 4.2 The morphological stages of sporulation in *Bacillus*

a) Spo0J protein binds to *oriC* of the chromosome. The oriC/Spo0J complexes (red) are positioned at about the one-quarter and three-quarter positions along the length of the cell. b) The *oriC* regions move towards the cell poles and are bound to the poles by Soj (not shown) and RacA (blue), and the polar anchor protein, DivIVA (purple). c) Asymmetric division results in trapping of the one-third of the chromosome that is attached to the *oriC* region in the prespore compartment. d) SpoIIIE (yellow) is recruited to the leading edge of the septum, completing the segregation of the prespore chromosome. Taken from Errington [170].

4.1.3 Chapter objectives

The main objective of this chapter is to investigate the plasmid partitioning systems in proteobacteria based on publicly available genomes. As we have seen in chapter 3, to date no research has described plasmid backbone systems comprehensively, especially using the *in silico* method, even though a large volume of genomic information is easily accessible. It is for this reason that we are going to investigate the plasmid partitioning systems across proteobacteria. The main targets of this investigation are two coupled proteins involved in active partitioning systems, ParA and ParB (ParM and ParR). In summary, we will:

1. Identify and classify gene families in plasmid partitioning systems and propose a classification scheme. Previous experimental results in order to figure out the mechanisms of partitioning systems have revealed, in particular, a variety of genes involved in the partitioning process. No clear classification, however, has been established so far for a large number of homologous sequences of Par proteins. Based on the main families we identified in chapter 2, we will first investigate the most abundant families in proteobacterial plasmids. Establishing the main families by classifying the genes would be useful not only for the arrangement of currently identified genes, but also for the future research of novel system types.

2. Research the diversity and distribution of plasmid partitioning systems in proteobacteria. Based on our current knowledge of the partitioning systems from experimental results (**Table 4.1**), we seek to examine which system types are most widely distributed in proteobacteria. This study will also investigate whether there is any specific feature of distribution for each family and whether any evidence exists that would shed light on the incompatibility groups of replicons. The patterns of plasmid diversity and host range that emerge from the data will be discussed.

3. Construct the phylogenetic tree of each family. The purpose of this investigation would be to see if there is any evolutionary relationship between different types of families. This can also be used to compare two different phylogenies of replication and partitioning systems and infer any recombination events between them. The analysis based on the two systems will be studied mainly in chapter 5.

4.2 Overview of plasmid partitioning systems

In this section, we have basically followed the same steps as in chapter 3 for the plasmid replication systems in proteobacteria. We used the data from the database

in chapter 2, which we obtained from publicly available genomes up to late October 2011, using blast, psi-blast, and our in-house HMMs. We have identified major groups that are distributed in proteobacteria. The data for the analysis include not only all the plasmids, but also chromids in bacteria, as per the Rep analysis in chapter 3.

In this chapter, we investigated 5 main families mostly abundant so far in proteobacterial plasmids, including one family that is reported previously based on the experimental result (see section 4.3.1.3). As mentioned in chapter 3, the actual number of Pars might not necessarily reflect actual prevalence in partitioning systems (see section 3.2) because the research is based on the published complete genomes to date. **Table 4.2** presents an overview of the classification of partitioning systems used in this chapter. We have classed Type I divisions into four parts, based on the size of ParA proteins and their partner proteins. The main phylogenetic analysis conducted in this chapter is based on ParA proteins. We have, however, analysed a coupled ParB protein when their sequences are homologous and long enough to make a multiple alignment, which results in the generation of ParB, KorB, and ParR phylogenies.

Table 4.3 indicates the coverage of replicons in this study. 46% and 44% of partitioning systems in alpha- and betaproteobacterial plasmids can be categorized according to the families in this study. Note that there are many more ParA homologs stored in the database (chapter 2) because many ParA homologs actually are located in chromids in addition to plasmids, particularly in case of betaproteobacteria. **Table 4.3** only shows how many plasmids, excluding chromids, having the Par systems were dealt with in this study.

Around one third of gammaproteobacterial plasmids have the Par systems that are categorized in this chapter. Again, there are more sequences in the database, since many gammaproteobacterial plasmids have multiple partitioning systems in one replicon. Nevertheless, there are many gammaproteobacterial plasmids that are not included in this study. This is because i) there are a small number of homologous sequences in each famly, and ii) many plasmids actually do not have active partitioning systems, as they are too small, or they have alternate systems

for maintenance. **Figure 4.3** shows the distribution of ParA homologs in different divisions of proteobacteria based on the families.

Table 4.2 Overview of active partitioning systems in proteobacteria classified in this study

Common name	Alternate name	Example hosts in proteobacteria
(Long) ParA-ParB	SopA-SopB, RepA-RepB	*Agrobacterium tumefaciens* *Burkholderia phymatum* *Escherichia coli* *Rhizobium leguminosarum* *Yersinia pestis*
IncC-KorB		*Acidovorax sp.* *Pseudomonas aeruginosa* *Sphingomonas sp.*
Short ParA-ParB		*Polaromonas naphthalenivorans* *Ralstonia solanacearum*
ParF-ParG		*Salmonella enterica*
ParM-ParR	StbA-StbB	*Escherichia coli* *Serratia marcescens*

Table 4.3 Coverage of plasmids in this chapter

Division of proteobacteria	Total number of plasmids	Plasmids which include ParAsbelonging to the main families	Plasmids which include ParAs not belonging to the main families
α	250	116	134
β	117	51	66
γ	667	223	444
Total	1034	390	644

4.3 Results

4.3.1 ParA family

There are a variety of Walker-type ParA homologs that exist across proteobacterial plasmids. As we have seen in section 4.1, they are often defined by different names, such as ParA, RepA, SopA, IncC, short ParA, ParF, according to the motifs they are composed of, the length of the protein sequences, and even the species in which the partitioning region is located. The homologous sequences have common motifs, such as the one helix-turn-helix region, A, A', and B, C motifs (**Figure 4.3**), except that short ParA, IncC, ParF do not havea HTH motif [141]. We have divided these Walker-type ParA homologs into four types, in order to construct efficient phylogenetic trees for the analysis of evolution: i) ParA (also called SopA in gammaproteobacteria), ii) IncC, iii) Short ParA, and iv) ParF.

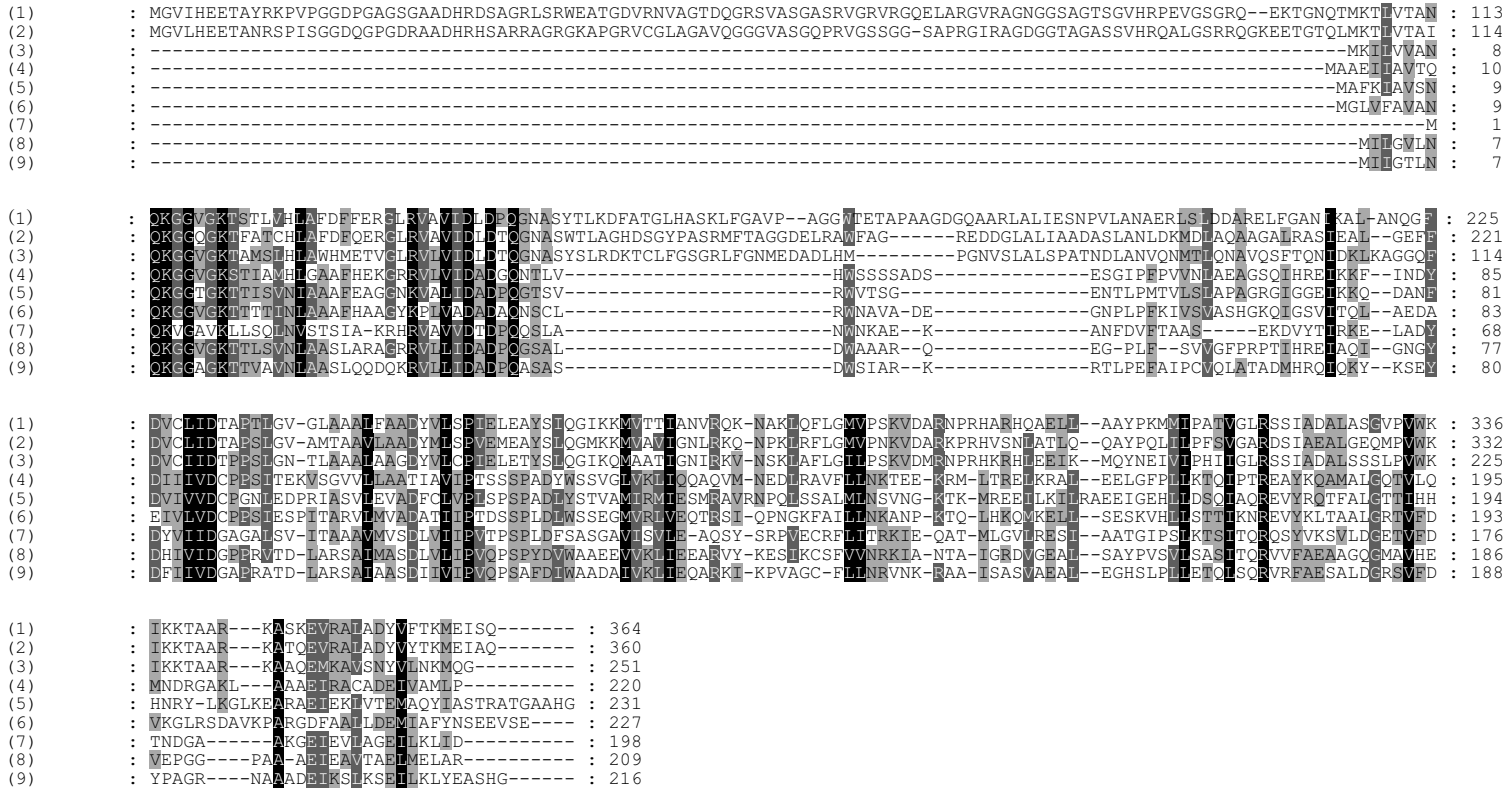


Figure 4.3 Alignment of IncC, ParF and short ParA

IncC: (1) YP_112422 of Birmingham IncP plasmid, (2) YP_974131 of *Acidovorax sp*. JS42, (3) YP_003600460 of *Neisseria gonorrhoeae*, short ParA: (4) YP_336660 of *Burkholderia pseudomallei* 1710b, (5) YP_625664 of *Burkholderia cenocepacia* AU1054, (6) YP_004362629 of *Burkholderia gladioli* BSR3, ParF: (7) YP_003212752 of *Cronobacter turicensis* z3032, (8) YP_004277303 of *Acidiphilium multivorum* AIU301, (9) YP_003422618 of *Zymomonas mobilis* subsp. Mobilis ZM4.

4.3.1.1 ParA-ParB family

The first sub-family we defined is the ParA-ParB family (often also called SopA-SopB in gammaproteobacterial plasmids). The amino acid sequences of ParA in this type are about 400 aa, having A, A', B and C motifs in them. The distribution of this group is wide from alpha to the gamma division of proteobacteria, but the majority of the group is ParA (RepA) in RepABC replicons, which are entirely in alphaproteobacteria. As seen in chapter 3, ParA, ParB (also known as RepA, RepB) and RepC are always located adjacently in the same order. Interestingly the distribution of ParA (also ParB) is different from that of RepC, in terms of the fact that other divisions of proteobacteria include ParAB homologs, while only alphaproteobacterial plasmids possess RepCs. Several betaproteobacterial plasmids possess this type of partitioning modules, such as *Burkholderia*, *Cupriavidus*, and *Ralstonia*. In addition, plasmids from various species in gammaproteobacteria showcase this family, such as *Enterobacter*, *Escherichia*, *Salmonella*, etc.

Figure 4.4 is a phylogenetic tree of ParA homologs. Note that we have deleted over 98% of similar sequences (see section 2.3.4) because there are numerous ParA homologs distributed across proteobacterial chromids and plasmids that have nearly identical sequences of partitioning systems. This means, therefore, that there were originally more sequences in the homolog set (see our website and appendix 2 for a full list of plasmids) and that is why some plasmids in rhizobia are missing in the tree. Nevertheless, many ParAs (RepAs) of RepABC replicons are shown abundantly in the phylogeny. As shown on the upper clade in the tree, the pattern of phylogeny is basically similar to RepC in RepABC (see **Figure 3.5** and chapter 5 for more details).

In chapter 3, we have examined the incompatibility groups according to the clades where the plasmids are located. In the replication-based phylogeny (**Figure 3.5**), some clades have not provided clear information on incompatibility. This is because some of the replication regions of two different plasmids are basically similar with each other, which results in two plasmids located in one clade in the tree. In other words, it was not sufficient to explain in this case that replication modules can be used as a marker for incompatibility. In the case of partitioning

sequences, however, there exist clear incompatibility groups (such as clades denoted by '*'). Therefore, this supports the idea that the incompatibility of Rhizobial RepABC replicons might rely more on partitioning systems, than replication ones [2].

On the phylogenetic tree there are more clades at the bottom, which contain the sequences adjacent to the RepB-like or RepA-like replication system in the plasmids, such as chromids from the family *Rhodobacteraceae* including *Rhodobacter sphaeroides* 2.4.1, *Dinoroseobacter shibae* DFL 12, etc., which do not belong to RepABC replicons. It is interesting to note that different replication systems often have the same partitioning systems. More details will be presented in chapter 5.

In addition to the big clade consisting of alphaproteobacterial plasmids, there are many from gammaproteobacterial plasmids (coloured in green). Firstly, it is noticeable that chromids and several plasmids from the Genus *Vibrio* have the ParAB (SopAB)-type partitioning system. Although the replication family of *Vibrio* is totally unrelated to other gammaproteobacterial plasmids, their partitioning regions might be evolutionarily related, sharing the same origin. Moreover, among gammaproteobacterial plasmids, plasmids from *Coxiella* have ParAB-type partitioning systems, while their replication system is based on the RepB-like replication system, as seen in chapter 3. The translation direction of the three genes is actually similar to that of RepABC replications. Other clades (Clade F) composed of gammaproteobacteria seem to be a mixture of different incompatibility groups, such as IncF, IncH, IncI, etc. As we have seen in chapter 3, plasmids in these groups are mostly the plasmids that have multiple replication types. Like multiple replication regions, multiple partitioning regions often exist, not only in ParAB (SopAB) but also in ParMR (see section 4.3.2).

The clade consisting of the plasmids coloured in blue contains several betaproteobacterial plasmids, including megaplasmids from *Ralstonia*, or 2nd chromids and small plasmids from *Burkholderia*. This is interesting because most 1st and 2nd chromids and some plasmids contain short ParA-ParB coupled proteins (see section 4.3.1.3). This presumably indicates that the origin of the plasmids in this clade of this particular tree is different from the 1st and 2nd chromids and from other plasmids of the same species.

We also constructed the phylogenetic tree of ParB homologues that are found together with ParA homologs in **Figure 4.4** (**Figure 4.5**). The pattern of the tree is basically similar, which shows that many clades are conserved in the order level of the species. The pattern inside the clades, however, is rather different, indicating that possible recombination events have taken place in the operons. Moreover, the variation of ParB sequences seems faster than ParAs.

Figure 4.4 The phylogenetic tree of ParA proteins

Unrooted Maximum Likelihood tree. Alphaproteobacteria in red, betaproteobacteria in blue, gammaproteobacteria in green. See appendix 2 for details of methods and the full tree (displaying sequence accession numbers).

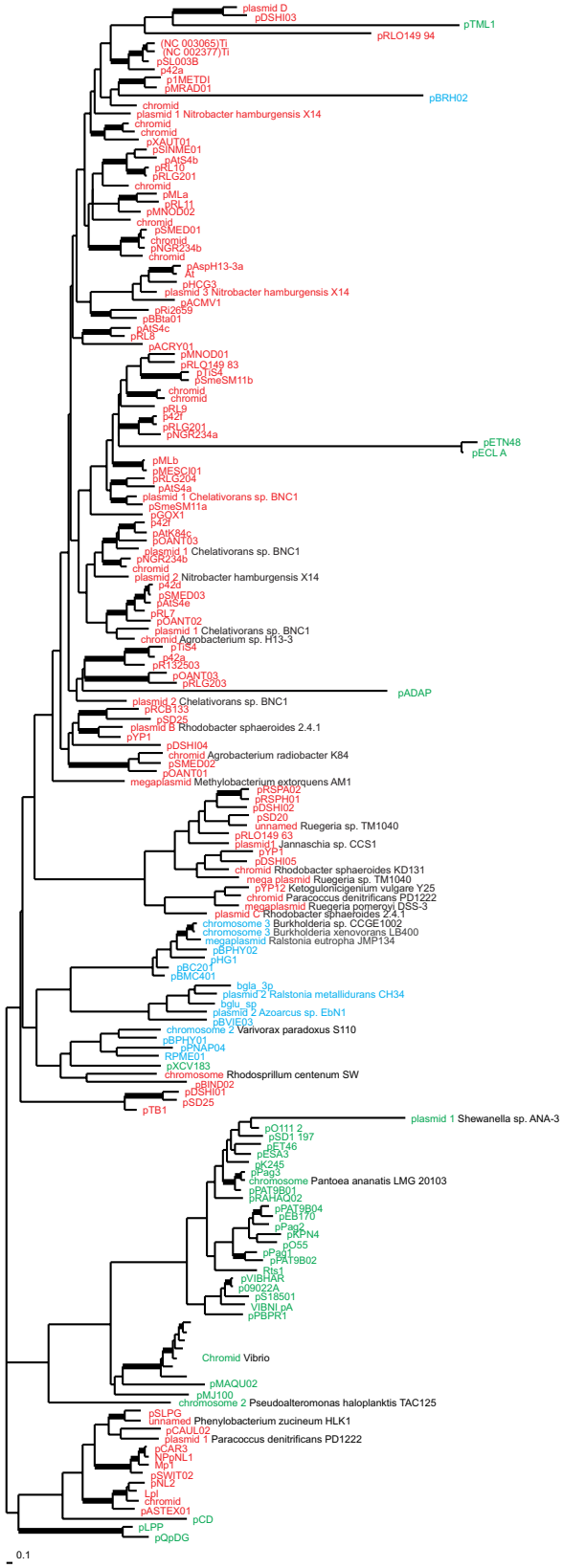


Figure 4.5 The phylogenetic tree of ParB proteins

Unrooted Maximum Likelihood tree. Alphaproteobacteria in red, betaproteobacteria in blue, gammaproteobacteria in green.See appendix 2 for details of methods and the full tree (displaying sequence accession numbers).

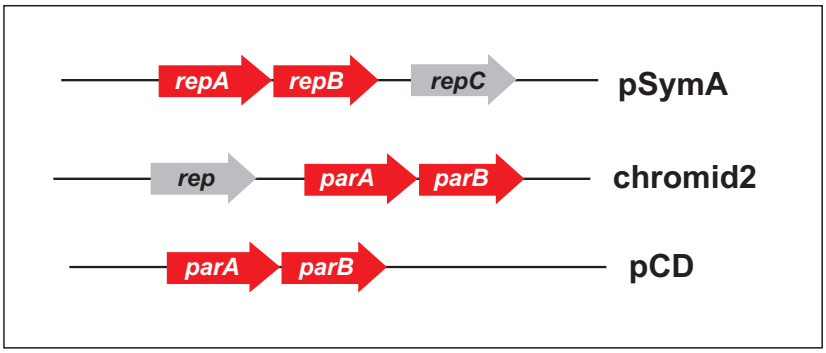


Figure 4.6 The genetic organization of partitioning regions having the ParA family

repA and *repB* are functionally similar to *parA* and *parB*. Partitioning proteins in red, and replicaton initiators in grey. Plasmids from three different divisions of proteobacteria having *parA* and *parB* are presented in the figure. pSymA from *Sinorhizobium meliloti* 1021, chromid 2 from *Burkholderia* species, and pCD from *Yersinia pestis* Java 9.

4.3.1.2 IncC-KorB family

IncC and KorB are the main partitioning modules in IncP plasmids. The coupled genes are located adjacent to several related regulator genes (grey arrows in **Figure 4.7**). In general, the gene *incC* encodes both a longer protein IncC1 and a shorter IncC2 [171], like TrfA (see chapter 3), because there is an inside translation frame. Studies have shown that IncC2 is sufficient for the partitioning activity; however, IncC1 does have a role in regulating polymerization and aiding depolymerisation [172].

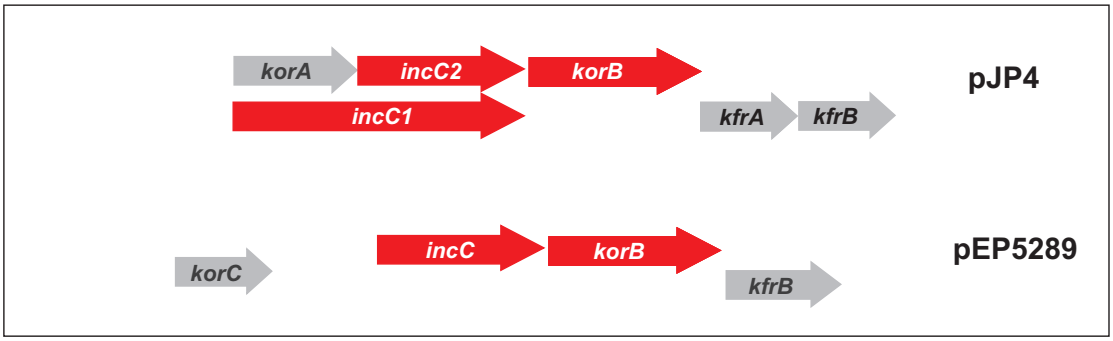


Figure 4.7 The genetic organization of partitioning regions in plasmids having the IncC-KorB system

Partitioning genes in red, and adjacent genes in grey. While *incC1* and *incC2* proteins share the same C-terminal sequences, KorA is read in a different frame to IncC1 and does not share the N-terminal sequences with the larger protein. *korC* encodes a transcriptional repressor, KorC. KfrABC complex consisting of *kfrA, kfrB* and *kfrC* is an important component in the organization and control of the plasmid clusters that seem to form segrosomes in bacterial cells. pJP4 from *Ralstonia eutropha* JMP134 and pEP5289 from *Neisseria gonorrhoreae*.

As we expected based on the TrfA distribution in chapter 3, the distribution of IncC and KorB is very wide, from alpha- to gammaproteobacterial plasmids. This is because IncP plasmids are promiscuous and are able to move around different species freely. **Figure 4.8** presents the phylogenetic tree of IncC homologs. Note that based on the sequences from the family we identified, we have only included IncC1 protein sequences rather than IncC2s in those cases where there are two *incC* translated proteins. The main big clade A is mostly composed of IncC sequences of IncP plasmids, including alpha to epsilon division, which reflect exactly the TrfA phylogeny that we have constructed in chapter 3. Obviously, IncC/KorB coupled proteins are essential for the IncP plasmids in ensuring their partitioning process.

There are plasmids, however, that do not appear in the TrfA phylogeny. Examples include plasmid_59kb from *Yersinia pseudotuberculosis* and plasmid 2 from *Aromatoleum aromaticum EbN1* (denoted by closed squares in **Figure 4.8**). It is interesting that these plasmids do not have TrfA homologs for their replication systems, which might indicate that their partitioning systems might be incorporated from other plasmids. On the other hand, although not defined as IncP plasmids, there are some that are included in both TrfA and IncC phylogenies, such

as pP9014 from *Photobacterium damselae* subsp. piscicida or pEP5289 from *Neisseria gonorrhoeae* (also found in the KorB phylogeny in **Figure 4.9**), which means they might be IncP plasmids. These are denoted by closed circles in the **Figure 4.8**.

One also notices that there is a clade B, which consists of ParA of IncW plasmids with a well-separated clade. As we have seen in chapter 3, RepA of IncW appeared as a very distinct family (see section 3.3.3.2) from the TrfA family of IncP plasmids (see section 3.3.3.1). The partitioning systems of IncW plasmids, however, are highly similar with the ones of IncP plasmids, which might indicate their related evolutionary origin in terms of partitioning systems, in comparison with their replication systems.

We have also constructed the phylogeny for KorB amino acid sequences, in order to track the relationships between two coupled proteins (**Figure 4.9**). The pattern of the two phylogenies is highly similar as expected: there are five clear clades for the alpha to epsilon sub-families of IncP plasmids in clade A. The sub-clades inside clade A, however, are slightly different in that the same grouping is not observed inside the IncP-β group. In more detail, KorB sequences of pB10 and pJP4 actually belong to IncP-β2 and not IncP-β1 in the IncC phylogeny, because many KorB sequences are basically similar to each other, although the clades having IncC sequences show a very distinct grouping.

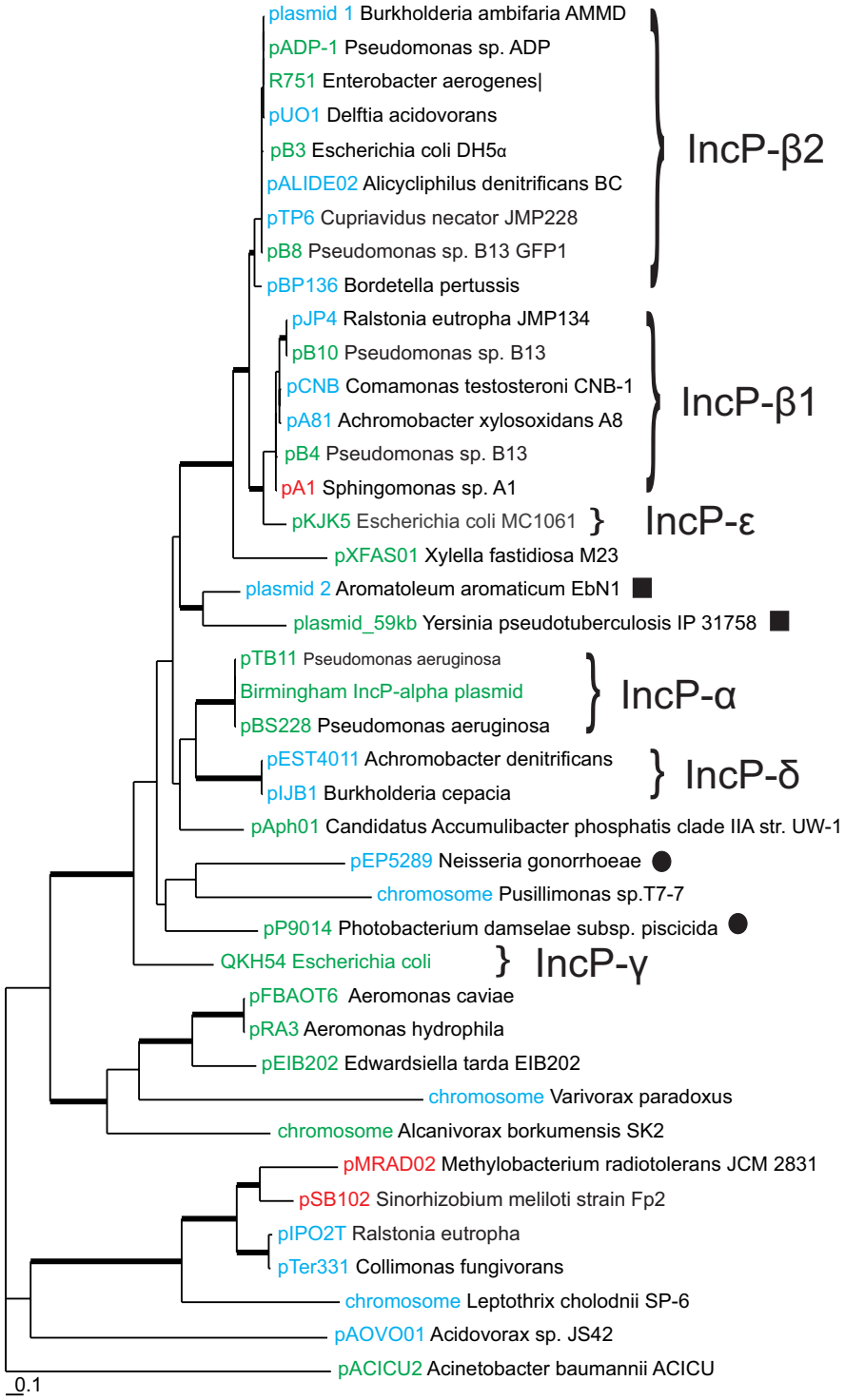


Figure 4.8 The phylogenetic tree of IncC proteins

Unrooted Maximum Likelihood tree. Alphaproteobacteria in red, betaproteobacteria in blue, gammaproteobacteria in green. The IncP groupings are shown in large-lettering. The plasmids not having TrfA homologs for their replication systems, but possessing IncC homologs for their partitioning systems are denoted by squares. The sequences that are included in both TrfA and IncC phylogenies are denoted by circles, although they are not defined as IncP plasmids. See appendix 2 for details of methods and the full tree (displaying sequence accession numbers).

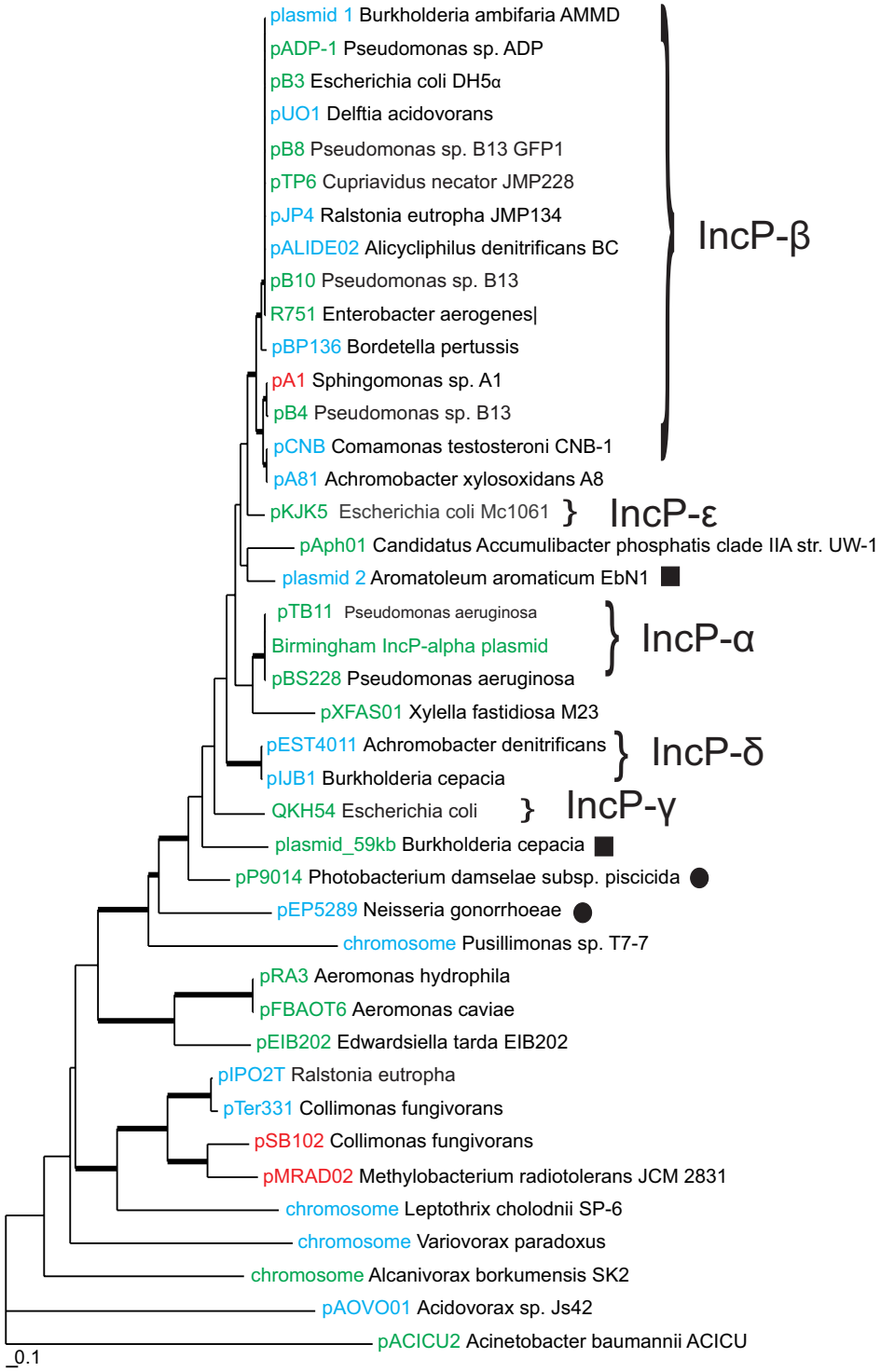


Figure 4.9 The phylogenetic tree of KorB proteins

Unrooted Maximum Likelihood tree. Alphaproteobacteria in red, betaproteobacteria in blue, gammaproteobacteria in green. The IncP groupings are shown in large-lettering.The plasmids not having TrfA homologs for their replication systems, but possessing IncC homologs for their partitioning systems are denoted by squares. The sequences that are included in both TrfA and IncC phylogenies are denoted by circles, although they are not defined as IncP plasmids. See appendix 2 for details of methods and the full tree (displaying sequence accession numbers).

4.3.1.3 Short ParA-ParB and ParF-ParG family

Among the homologous sequences that do not belong to the long ParA or to IncC, there are two subtypes of families showing similarity in size and functions, which are called short ParA and ParF. Neither type has HTH in the N-terminal region of the gene. As they share a high sequence similarity and a nearly similar size, both types of sequences can be collected by BLAST with a high e-value. There is a clear difference, however, between them; their partner protein for partitioning is very different. Short ParA is mostly adjacent to ParB, which counts 300aa. Next to ParF, however, we find ParG, with ~80 aa (**Figure 4.10**). Interestingly, ParB being adjacent to ParA and ParG being adjacent to ParF do not show any sequence homology between them, although ParG is known as a functional analogue of ParB [145].

Thus, we have divided the families into two sub-types: ParF-ParG and short ParA-ParB according to their partner proteins, ParG and ParB. The distribution of the short ParA and ParB families is mostly in betaproteobacterial plasmids. In particular, it is the 1st chromid and 2nd chromid of the *Burkholderia* species and some plasmids of the *Ralstonia* that have this type of partitioning systems. **Figure 4.11** shows a phylogenetic tree based on both short ParA and ParF sequences. It is clear that the clade being composed of ParF sequences forms a unique clade (underlined) as the sequences are separated well. The main distribution of ParF is in gammaproteobacteria; mostly in *E. coli, Salmonella* or *Enterobacter*, but the number of ParF homologs (with ParG) is limited, so it is difficult to define their distribution. The only exceptions are the ParF sequences from some alphaproteobacteria (pACMV6 from *Acidiphilium multivorum* AIU301 and pZZM401 from *Zymomonas mobilis* subsp. Mobilis ZM4), but these appear more diverged than the others. It is possible that the ParF-ParG type might have evolved twice independently.

On examining the long ParA-ParB and short ParA-ParB family, it was assumed that the short ParA was a cut-down version of the long ParA because both lacked HTH sequences in the N-terminal region. If this is true, their partner protein ParB might have an evolutionary relationship with them. In order to confirm this, we have constructed ParB homologues that were adjacent to both long and short ParA. A phylogenetic tree of two types of ParB (not shown) indicates that ParB sequences

with short ParA form a well separated clade (denoted by *). Therefore, we might conclude that they have been diverged for a long historical period.

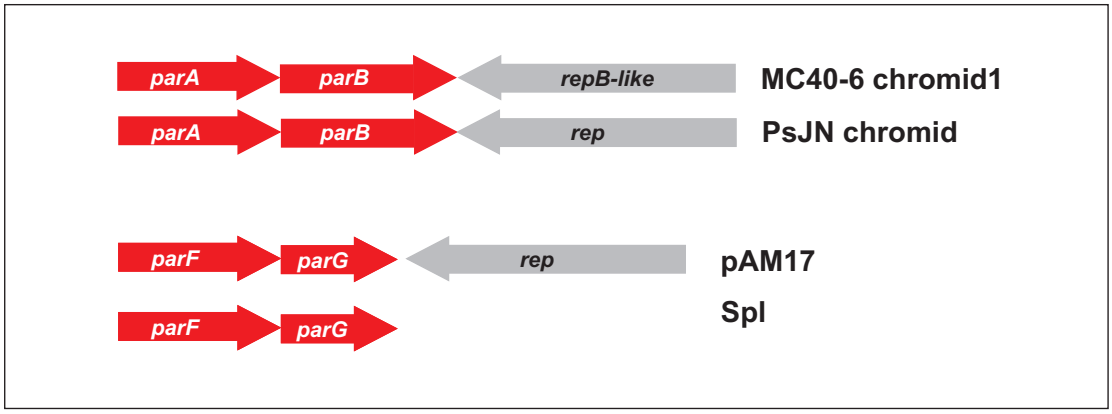


Figure 4.10 The genetic map of plasmids having short ParA-ParB and ParF-ParG

Short ParA-ParB family, chromid 1 from *Burkholderia ambifaria* MC40-6 and chromid from *Burkholderia phytofirmans* PsJN. ParF-ParG family, pOLA52 from *Escherichia coli* and pOU1115 from *Salmonella enterica* subsp. enterica serovar Dublin.

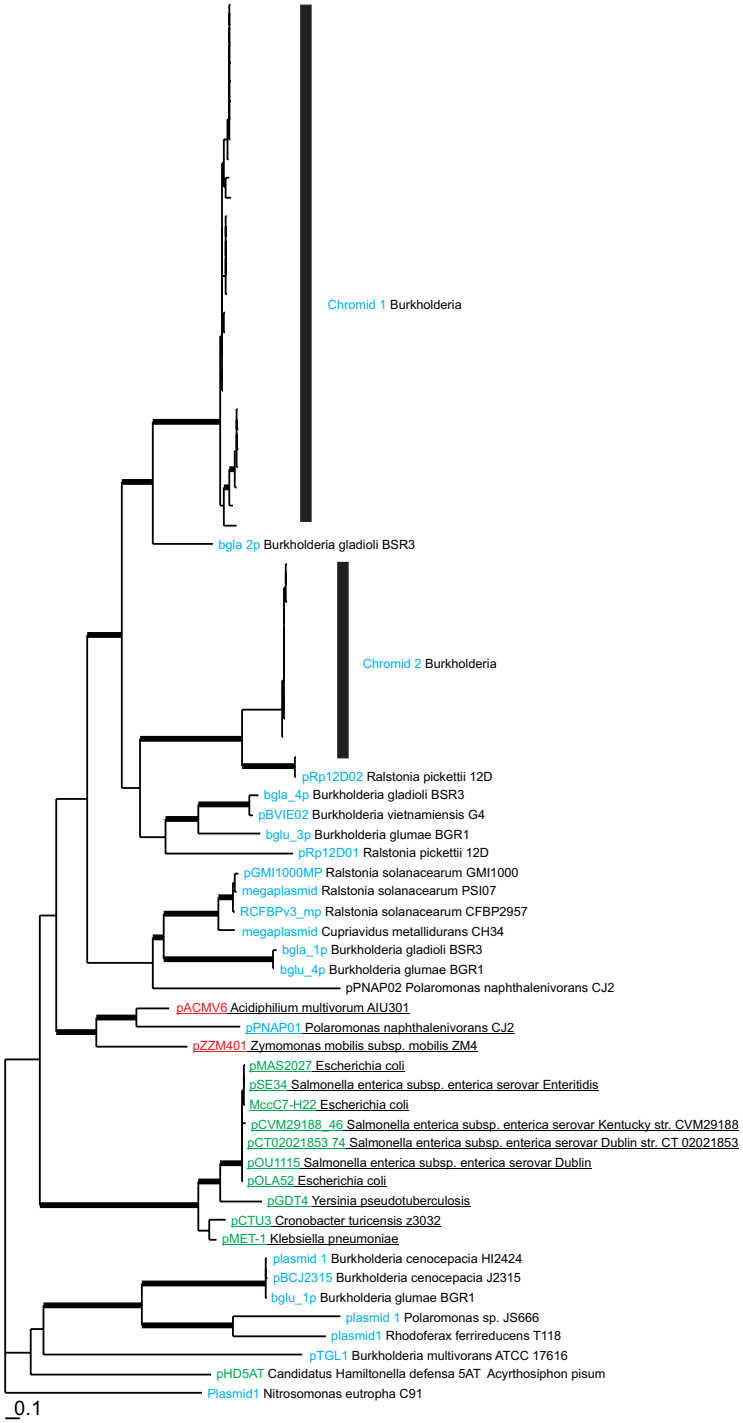


Figure 4.11 The phylogenetic tree of all homologous sequences in the family type of short ParA and ParF proteins

Unrooted Maximum Likelihood tree. 28 sequences of chromid 1 and 11 sequences of chromid 2 from *Burkholderia* are contained in the clades indicated by the black bars. Alphaproteobacterial plasmids in red, betaproteobacterial plasmids in blue, and gammaproteonacterial plasmids in green. Species underlined have ParF homologs as a partitioning system. See appendix 2 for details of methods and the full tree (displaying sequence accession numbers).

4.3.2 ParM-ParR family

ParM-ParR family is generally known as the Par system of plasmid R1, which is basically equivalent to Stb of plasmid R100, encoding two *trans*-acting proteins ParM and ParR, and one *cis*-acting region *parC* [143, 164]. Homologs search shows that the distribution of ParM family is mostly restricted to gammaproteobacteria. Among the members in the family, only two coupled proteins from *Burkholderia vietnamiensis* G4 and *Achromobacter xylosoxidans* A8 have been found, which belong to betaproteobacteria.

Figure 4.12 is a phylogenetic tree of ParM homologs. It is firstly interesting that there are two ParM sequences from betaproteobacterial plasmids. There are in a very strong branch separating this clade from the rest, but the sequences in the clade are highly diverged from each other and their relationships are poorly resolved. This does not necessarily imply that they do not have a common origin. The rest of members in the family generally belong to *Esherichia, Salmonella, Shigella, Klebsiella,* etc. Although there are many plasmids possessing this family for partitioning, the homologs of the ParM family does not seem variable in comparison with other partitioning types such as ParA, IncC, etc., as they basically have nearly similar sequences. On the other hand, many homologs have been found in chromosomes, and they form several distinct clades.

In contrast to the ParM homologs that commonly show homology in the core region structurally, their partner protein ParR does not seem to show high homology among them, which means that they are highly diverged [143]. ParM-ParR coupled proteins do look similar with ParF-ParG proteins in terms of structure, the genetic organization and the length of the sequences; however, there is no sequence similarity between two families, which indicates that they might have evolved independently [143].

Figure 4.12 The phylogenetic tree of ParM proteins

Unrooted Maximum Likelihood tree. 23 sequences of chromosome from *Escherichia coli* are contained in three clades indicated by the black bars. Betaproteobacteria in blue, otherwise all is from gammaproteobacteria. See appendix 2 for details of methods and the full tree (displaying sequence accession numbers).

4.4 Discussion

4.4.1 Contribution

Like plasmid replication systems, plasmid partitioning systems are significant collections of genes found in the plasmid backbone, especially for low-copy number plasmids in proteobacteria. Thanks to a variety of genome projects and the experimental results of individual gene functions, several types of plasmid partitioning systems have been characterized and the mechanisms of each type have been studied. No comprehensive research to investigate the distribution of each type of partitioning system has been attempted before, however, and the available evidence was not enough to study evolutionary history based on the phylogenetic analysis of each family. Thus, we have studied in this chapter five main types of partitioning systems in proteobacteria.

Table 4.4 ParA homologs presented in different databases

N/A: There is no cluster including the ParM family.

Family	COG	ACLAME	Pfam	Protein cluster in NCBI
ParA	COG1192D	all6	PF10609	CLSK2465197 CLSK514544 CLSK2343020 CLSK 865073
IncC	COG1192D	all6	PF01656	CLSK687405
Short ParA	COG1192D	all6	PF10609	CLSK2505399
ParF	COG1192D	all6	PF01656	CLSK909785 CLSK909784
ParM	N/A	all578	PF06406	CLSK861782 CLSK861711

As mentioned briefly in chapter 2, several previous researches have investigated plasmid partitioning modules. **Table 4.4** shows a short list of modules that are stored in individual databases or web interfaces. Individual studies have their pros and cons. For example, while Pfam has relatively accurate homologues because of HMMs (motif)-based search, a large number of sequences are not conducive to inferring their evolutionary history. This is because, unlike Rep, many Par systems share the same motifs, so that each family in Pfam ends up having a tremendous quantity of members. ACLAME was developed to investigate MGEs extensively (including plasmids), which might provide more accurate annotation; however, it still offers just a collection of protein sequences. Protein Cluster in NCBI has an advantage for inferring evolution because it considers the preceding and following genes alongside the gene in question. The search for homologs, however, is based on Blast results, which contain the risk of losing many informative homologues, resulting in too low a number of members in one family. Based on the pros and cons of previous research, this study has contributed by examining the evolutionary history of plasmid partitioning systems, firstly by considering neighbor genes and domains related to interesting genes, and secondly by performing an analysis on each phylogenetic tree of the families defined.

4.4.2 General questions of plasmid partitioning systems in proteobacteria

4.4.2.1 Patterns of plasmid diversity and host range

Best current knowledge of plasmid partitioning system is that there are three major types involved in the systems: Walker A cytoskeletal P-loop ATPases, actin-like ATPase, tubulin-like ATPase. In proteobacteria, a majority of partitioning proteins belongs to Walker type ATPase. Although these proteins share the similar domains, they can be divided to distinct types based on their partner proteins. Therefore we have identified four discrete types in the Type I class and one type in the Type II class, which are most abundant in proteobacterial plasmids.

Generally partitioning systems are within the class level. This is shown very strongly in the plasmids having long ParA sequences such as RepABC replicons from alphaproteobacteria, and several plasmids from beta-, gammaproteobacteria. Plasmids possessing short ParA, ParF and ParM type sequences also show their distribution is restricted to the class level of proteobacteria in the phylogenies. In addition, host range also shows similar patterns. The members of each type do not seem to dramatically move outside the class level, except for the case of broad host range plasmids. The only exception is the IncC-KorB system of broad host range plasmids, which shows wide distribution across different division of proteobacteria.

4.4.2.2 How many families are there in partitioning systems in proteobacteria?

In this chapter, we have investigated five families of plasmid partitioning systems that we have identified in proteobacteria. In comparison with the replication systems in proteobacterial plasmids, major families of partitioning modules appear more broadly used across bacteria. As we have seen in **Table 4.2**, there are five major families covering 390 proteobacterial plasmids, which is comparable to the eight Rep families covering 520 plasmids we have investigated in chapter 3. While the ParM family is relatively restricted to gammaproteobacteria, other families are distributed in different divisions of proteobacteria. An IncC-KorB couple, in particular is encountered very broadly, while other families are found in different divisions of proteobacteria. Although we have investigated widely used and experimentally studied partitioning systems in this study, there are more types of Par systems having a small number of members yet.

4.4.2.3 Would it be possible to use the partitioning systems to classify plasmids?

As we have discussed in chapter 3, both replication and partitioning systems in bacteria are considered as an important indicator for possibly classifying plasmids. As we have shown in chapter 3, the RepABC replicons are not easy to be classified by Rep systems, because the Rep systems of some plasmids actually have nearly similar Rep sequences. As such, plasmids showing different incompatibility can be shown to belong to the same clade in the phylogenetic tree (see **Figure 3.6**). The

Par system, however, actually can play a role in the classification of plasmids, as shown in **Figure 4.5**. Thus, in the case of RepABC replicons, the Par system is significant in classifying plasmids. It appears, however, that not all Par systems are efficient for plasmid classification. Firstly, it is not true that all plasmids have Par modules, which might leave out numerous replicons in bacterial classification. Also, based on **Figure 4.12**, the multiple numbers of partitioning modules in one plasmid alone can make it hard to classify them effectively.

4.4.2.4 Evolution of partitioning systems in proteobacteria

As seen above, ParA, IncC, Short ParA and ParF share sequence homology. In order to answer the question of which proteobacterial class each type originated in, we have constructed phylogenetic trees using all the sequences from four types and make the tree rooted by ParF sequences (**Figure 4.14**). In the tree, the clade containing short ParA (red bar) seems recent plasmids than other clades that are mostly well separated with high bootstrapping values in a long history.

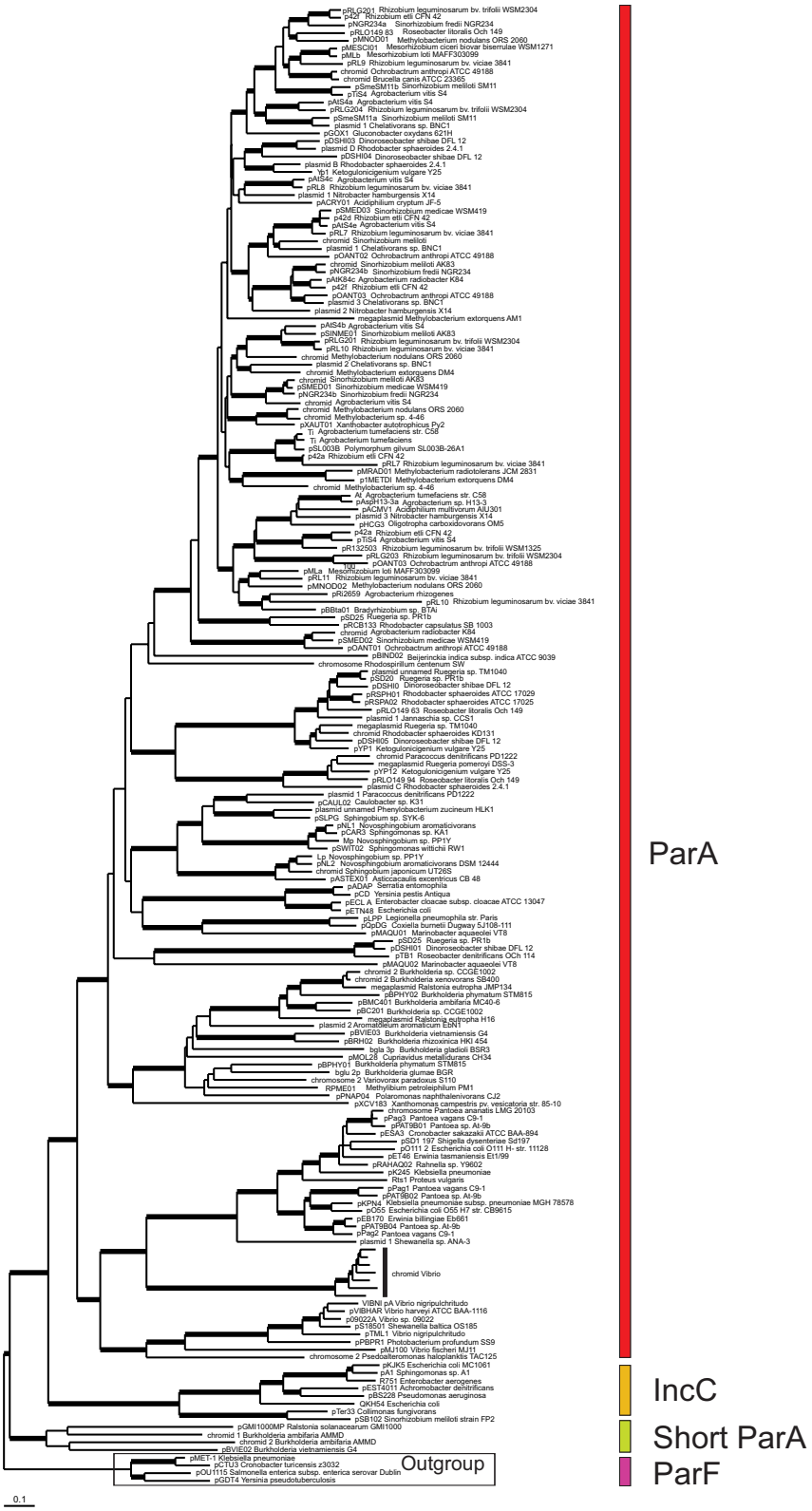


Figure 4.13 The phylogenetic tree of all Type I Par homologs

Rooted by the sequences from ParF family. The clade including ParF homologs was used as the out-group. The coloured vertical bars represent different partitioning systems (ParA in red, IncC in orange, Short ParA in green, and ParF in purple).

4.3.3 Limitations of this work and future directions

As mentioned in sections 4.3.1.3 and 4.3.1.4, it has been difficult to characterize efficiently some families in this chapter, because 4 out of 5 of them actually share the same motifs in their sequences. Therefore, some sequences are very ambiguous in terms of their categorization with each family if they are too evolutionary diverged. That is why, at first, we considered both sequence similarity methods, such as BLAST, and motif-based methods, such as Psi-blast and HMMs. Moreover, we have defined the families based on two coupled proteins for partitioning. This gives us more insights on each family, such as the characterization of the short ParA and the ParF families. A lot of sequences, however, remain that do not belong to any family based on either sequence similarity or domains.

Further studies will take place in the following chapter. In chapter 5, we will study both replication and partitioning systems, in order to investigate their evolutionary relationships. As we have seen in chapters 3 and 4, many plasmids have both systems located very closely and even seemingly affecting each other. The comparison of the distribution of both systems with the phylogenetic results might widen our understanding of the systems themselves and even of plasmids as well. We will study both systems across an individual division of proteobacteria.

Chapter 5. Evolutionary history of plasmid backbone systems in proteobacteria

In chapters 3 and 4, we have investigated plasmid replication and partitioning systems, respectively, in proteobacteria. Previous research has revealed that the two systems are often located adjacently, they are functionally related to each other, and might have evolved together [91]. A comparative assessment of the evolutionary relationship of the two systems is able to shed light on how these plasmids have evolved over time. In particular, both systems are part of the essential backbone that exists in most plasmids, which makes them significant factors in discerning plasmids' evolutionary history. In this chapter, therefore, we will compare the distribution and phylogenetic trees of the two systems examined in chapters 3 and 4 and will investigate the typical or unique patterns of plasmid diversity and host range that emerge from the data.

5.1 Review of plasmid backbone systems

5.1.1 Why are they worth investigating?

Both replication and partitioning systems are essential for self-replication and for the stable maintenance of plasmids [173]. Most plasmids, especially large low-copy number plasmids, in general, have to possess these two systems, while they additionally confer other specific phenotypes based on different environmental niches. As we have seen in chapters 3 and 4, they often exist close to each other and are involved in the specific regulation of their operons. Due to their localization and phylogenetic relationships, many previous studies [97] have argued that they might share in a similar evolutionary history, which would be a key in investigating the diversity of plasmids and their host range.

In 2009, Petersen et al. [91] published their research regarding a novel plasmid replication type in *Rhodobacterales* and attempted to set up an evolutionary scenario for the origin of replication modules based on phylogenetic comparisons of ParA and ParB partitioning modules. Using 40 completely sequenced genomes, they have identified a novel plasmid type, DnaA-like replicon, which does not have any homology with the RepA-like or the RepB-like replicon type, which are main Rep types of *Rhodobacterales.* They have also argued that there are at least 18 compatible plasmids in *Rhodobacterales* based on the phylogenies of Replication types, including 9 groups of RepABC replicons [91]. This study is one of the most important breakthroughs towards tracking the evolutionary background of replicons in different isolates, although it is restricted to alphaproteobacteria, particularly *Rhodobacterales*.

5.1.2 Review of plasmid replication and partitioning systems in proteobacteria

In chapter 3, we have explored the replication modules in proteobacteria. Among 8 families, some replicon types show very wide distribution across species, including TrfA-, RepA- (of IncW) type replication systems. Basically, they are not specifically restricted to any division of proteobacteria, but are present in a wide range of bacteria. On the other hand, the RepC family is only found in alphaproteobacteria. RepA-like and RepB-like sequences are distributed in different divisions of proteobacteria. It seems, however, that they have been conserved with a long history within the class level. RepFIIA-, RepFIA- and RepFIB-type systems are mostly distributed in gammaproteobacteria, with the exception of several replicons in *Bacillus*.

In chapter 4, we have researched the active partitioning systems in proteobacteria, which are also significant factors for plasmids' maintenance. Based on the coupled partitioning modules (ParA-, ParB-type or both), we have defined 5 major families. Of the 5 families, 4 have similar motifs including A, A', B and C. Most ParA homologs belong to RepABC replicons from alphaproteobacteria, while several homologs are also found in beta- and gammaproteobacteria. Short ParA and ParF

show a similar length to each other, even if their partner proteins (ParB and ParG) and their distribution are significantly different. Short ParA-ParB coupled proteins are dominant in betaproteobacteria, while ParF-ParG proteins are mostly found in gammaproteobacteria.

5.1.3 Chapter objectives

As discussed briefly in chapters 3 and 4, many replication and partitioning systems are often located adjacently, work together and might affect each other in replicons. A comparison of the distribution of the two systems and a phylogenetic analysis might reveal the evolutionary history of the systems and might also allow us to infer the evolution of replicons, including chromids and plasmids. In this chapter, therefore, we aim to do the following:

i) Firstly, we will look into the general features of the two systems and shed light on their relationship based on the comparison of each type of Rep and Par systems. From the results of chapters 3 and 4, we aim to show which Par systems are combined with the individual Rep system. In particular, in this chapter, we will investigate different divisions of proteobacteria, from alpha to gamma, and study the general patterns of Rep and Par systems.

ii) Secondly, we will investigate phylogenetic trees of both Rep and Par systems, and try to track plasmids' evolution. Rep systems from chapter 4 will be mapped on the phylogenies of Par systems from chapter 3, in order to compare the patterns of the two systems and to see their evolutionary history.

5.2 Results

5.2.1 Distribution and general patterns of Rep and Par systems

5.2.1.1 Alphaproteobacteria

The alphaproteobacteria consist of various families of bacteria, which include phototrophic genera, carbon metabolizing genera, symbionts, and even pathogens. Up until October 2011, completely sequenced genomes involved the following orders: *Caulobacterales (Asticcacaulis, Caulobacter, Phenylobacterium*), *Rhizobiales (Agrobacterium, Bartonella, Beijerinckia, Bradyrhizobium, Chelativorans, Mesorhizobium, Methylobacterium, Nitrobacter, Ochrobacterum, Oligotropha, Sinorhizobium, Rhizobium, Xanthobacter), Rickettsiales (Rickettsia), Rhodobacterales (Dinoroseobacter, Hirschia, Hannaschia, Ketogulonicigenium, Paracoccus, Rhodobacter, Roseobacter, Ruegeria), Rhodospirillales (Acetobacter, Acidiphilium, Azospirillum, Gluconacetobacter, Gluconobacter, Magnetospirillum),*and *Sphingomonadales (Novosphingobium, Sphingobium, Sphingomonas, Sphingopyxis, Zymomona*s) [174].

Many species from alphaproteobacteria have not only chromosomes but also plasmids and chromids in them [84]. Most replicons basically have their own Rep proteins. It is important to note, however, that some small plasmids do not have Rep proteins but only have origins of replication, and other transposes, resolvase, or mobilization genes. Examples include pAC5 (NC_001275) from *Acetobacter aceti.* These replicons might use (borrow) Rep proteins from other replicons, mostly the ones from chromosomes; otherwise they would not replicate and would be lost. In general, among all the plasmids in complete genomes of proteobacteria, most of the replication systems in alpha division belong to the major families we have defined. There are still, however, unique Rep systems and uncharacterised Rep initiators, which confirm the diversity of plasmids.

In *Rhizobiales*, most replicons replicate based on RepABC operons, with several exceptions. For instance, the Rep and Par systems of pAgK84 (NC_011994) from *Agrobacterium radiobacter* K84, pSINME02 (NC_015592) from *Sinorhizobium meliloti* AK83 and pRM1132f from *Sinorhizobium meliloti* 1132 do not show any

homology with RepABC operons. Most plasmids from *Methylobacterium* do not have RepABCs. A couple of plasmids from *Ochrobactrum*, such as pW240 (NC_010917) and pOANT04 (NC_004965) might have unique Rep systems. On the other hand, pSB102 (NC_003122) from *Sinorhizobium meliloti* strain FP2 and pMRAD02 of *Methylobacterium radiotelarans* JCM 2831 have RepA of IncW-like plasmids and IncC/KorB of IncP plasmids as seen in the previous section.

In *Rhodobacterales*, ParAB partitioning systems are commonly found, but Rep systems show a variety. Petersen et al. [140] pointed out various combinations of Rep and Par systems in this order, such as RepB-like, RepA-like, RepC-like and DnaA-like replication systems with ParAB partitioning systems. We found out that there also exist very small plasmids having no Rep or Par, such as pMG160 (NC_004527) from *Rhodobacter blasticus*.

There are not many complete genomes in *Sphingomonadales*, but generally, both Rep and Par systems vary. As seen in section 3.3.7, pA1 from *Sphingomonas* sp. A1 interestingly possesses IncP Rep and Par systems, namely TrfA, IncC/KorB [86]. This is clearly implying a recent transfer of this IncP beta "broad host-range" plasmid type to alphaproteobacteria from beta or gamma, where it is much more common. In *Zymomonas*, replicons do not show any homology with major families, indicating that there are unique Rep and Par systems in this species.

Table 5.1 is a list of alphaproteobacterial plasmids and their Rep and Par systems in this study. If either Rep or Par of plasmids does not belong to the major groups that we identified, or if plasmids actually do not have Rep or Par proteins, we did not include them in this list. In general, the majority of combination of Rep and Par is RepABC as expected. There are also several plasmids that possess RepA-like, RepB-like system with ParAB. The combination of RepA (of IncW) and ParAB exist, and a couple of plasmids (e. g. pACMV1, pGOX1) do have RepC with RepA, which might indicate later acquisition of the RepA system.

Table 5.1 List of alphaproteobacterial plasmids and their Rep and Par systems identified in this study

RepC indicated by C, ParA by A, ParB by B, RepA (not ParA, but Rep of IncW plasmids) by W, Short ParA by S, IncC by I, KorB by K. Double or triple alphabet was shown if there are multiple Rep or Par proteins.

Accession	Species	Plasmid	Initiator	NTPase	Binding protein
NC_009467	*Acidiphilium cryptum* JF-5	pACRY01	C	A	B
NC_015178	*Acidiphilium multivorum* AIU301	pACMV1	CW	A	B
NC_011990	*Agrobacterium radiobacter* K84	pAtK84b	C	A	B
NC_011987	*Agrobacterium radiobacter* K84	pAtK84c	CC	A	B
NC_002575	*Agrobacterium rhizogenes*	pRi1724	C	A	B
NC_010841	*Agrobacterium rhizogenes*	pRi2659	C	A	B
NC_015184	*Agrobacterium sp.* H13-3	pAspH13-3a	C	A	B
NC_002377	*Agrobacterium tumefaciens*	Ti	C	A	B
NC_010929	*Agrobacterium tumefaciens*	pTiBo542	C	A	B
NC_002147	*Agrobacterium tumefaciens*	pTi-SAKURA	C	A	B
NC_003064	*Agrobacterium tumefaciens* str. C58	At	C	A	B
NC_003065	*Agrobacterium tumefaciens* str. C58	Ti	C	A	B
NC_011986	*Agrobacterium vitis* S4	pAtS4a	CC	A	B
NC_011991	*Agrobacterium vitis* S4	pAtS4b	C	A	B
NC_011984	*Agrobacterium vitis* S4	pAtS4c	C	A	B
NC_011981	*Agrobacterium vitis* S4	pAtS4e	C	A	B
NC_011982	*Agrobacterium vitis* S4	pTiS4	CC	AA	BB
NC_014818	*Asticcacaulis excentricus* CB 48	pASTEX01	A	A	B
NC_009475	*Bradyrhizobium sp.* BTAi1	pBBta01	C	A	B
NC_010333	*Caulobacter sp*. K31	pCAUL02	A	A	B

NC_008242	*Chelativorans sp*. BNC1	plasmid 1	CCC	A	BB
NC_008243	*Chelativorans sp.* BNC1	plasmid 2	C	A	B
NC_008244	*Chelativorans sp.* BNC1	plasmid 3	C	A	B
NC_009955	*Dinoroseobacter shibae* DFL 12	pDSHI01	C	A	B
NC_009956	*Dinoroseobacter shibae* DFL 12	pDSHI02	A	A	BB
NC_009957	*Dinoroseobacter shibae* DFL 12	pDSHI03	C	A	B
NC_009958	*Dinoroseobacter shibae* DFL 12	pDSHI04	C	A	B
NC_009959	*Dinoroseobacter shibae* DFL 12	pDSHI05	B	A	B
NC_006672	*Gluconobacter oxydans* 621H	pGOX1	CW	A	B
NC_007801	*Jannaschia sp.* CCS1	plasmid1	A	A	B
NC_014621	*Ketogulonicigenium vulgare* Y25	pYP1	AB	AAS	BB
NC_014918	*Mesorhizobium ciceri* biovar biserrulae WSM1271	pMESCI01	C	A	B
NC_002679	*Mesorhizobium loti* MAFF303099	pMLa	C	A	B
NC_002682	*Mesorhizobium loti* MAFF303099	pMLb	CA	A	B
NC_011758	*Methylobacterium chloromethanicum* CM4	pMCHL01	C	A	B
NC_012811	*Methylobacterium extorquens* AM1	megaplasmid	M	A	B
NC_012807	*Methylobacterium extorquens* AM1	p1META1	CA	A	B
NC_011892	*Methylobacterium nodulans* ORS 2060	pMNOD01	C	A	B
NC_010510	*Methylobacterium radiotolerans* JCM 2831	pMRAD01	C	A	B
NC_010509	*Methylobacterium radiotolerans* JCM 2831	pMRAD02	W	I	K
NC_007959	*Nitrobacter hamburgensis* X14	plasmid 1	C	AS	B
NC_007960	*Nitrobacter hamburgensis* X14	plasmid 2	CC	A	B
NC_007961	*Nitrobacter hamburgensis* X14	plasmid 3	C	A	B
NC_002033	*Novosphingobium aromaticivorans*	pNL1	W	A	B
NC_009426	*Novosphingobium aromaticivorans* DSM 12444	pNL1	W	A	B

NC_009427	*Novosphingobium aromaticivorans* DSM 12444	pNL2	A	A	B
NC_015579	*Novosphingobium sp.* PP1Y	Lpl	AA	A	B
NC_009669	*Ochrobactrum anthropi* ATCC 49188	pOANT01	CC	A	B
NC_009670	*Ochrobactrum anthropi ATCC* 49188	pOANT02	C	A	B
NC_009671	*Ochrobactrum anthropi ATCC* 49188	pOANT03	CC	AA	BB
NC_005873	*Oligotropha carboxidovorans* OM5	pHCG3	C	A	B
NC_015689	*Oligotropha carboxidovorans* OM5	pHCG3	C	A	B
NC_015685	*Oligotropha carboxidovorans* OM5	pOC167	C	A	B
NC_003122	*Sinorhizobium meliloti* strain FP2	pSB102	W	I	K
NC_007762	*Rhizobium etli* CFN 42	p42a	CC	AA	BB
NC_007763	*Rhizobium etli* CFN 42	p42b	C	A	B
NC_007764	*Rhizobium etli* CFN 42	p42c	C	A	B
NC_007765	*Rhizobium etli* CFN 42	p42e	C	A	B
NC_007766	*Rhizobium etli* CFN 42	p42f	CC	AA	BB
NC_004041	*Rhizobium etli* CFN 42	symbiotic plasmid p42d	C	A	B
NC_010998	*Rhizobium etli* CIAT 652	pA	C	A	B
NC_010996	*Rhizobium etli* CIAT 652	pB	C	A	B
NC_010997	*Rhizobium etli* CIAT 652	pC	CC	AA	BB
NC_012848	*Rhizobium leguminosarum* bv. trifolii WSM1325	pR132501	C	A	B
NC_012858	*Rhizobium leguminosarum* bv. trifolii WSM1325	pR132502	C	A	B
NC_012853	*Rhizobium leguminosarum* bv. trifolii WSM1325	pR132503	C	A	B
NC_012852	*Rhizobium leguminosarum* bv. trifolii WSM1325	pR132504	C	A	B
NC_012854	*Rhizobium leguminosarum* bv. trifolii WSM1325	pR132505	C	A	B
NC_011368	*Rhizobium leguminosarum* bv. trifolii WSM2304	pRLG201	CC	AA	BBB
NC_011366	*Rhizobium leguminosarum* bv. trifolii WSM2304	pRLG202	C	A	B

NC_011370	*Rhizobium leguminosarum* bv. trifolii WSM2304	pRLG203	C	A	B
NC_011371	*Rhizobium leguminosarum* bv. trifolii WSM2304	pRLG204	C	A	B
NC_008381	*Rhizobium leguminosarum* bv. viciae 3841	pRL10	C	AA	B
NC_008384	*Rhizobium leguminosarum* bv. viciae 3841	pRL11	C	A	B
NC_008378	*Rhizobium leguminosarum* bv. viciae 3841	pRL12	C	A	B
NC_008382	*Rhizobium leguminosarum* bv. viciae 3841	pRL7	CCC	AAA	BB
NC_008383	*Rhizobium leguminosarum* bv. viciae 3841	pRL8	C	A	B
NC_008379	*Rhizobium leguminosarum* bv. viciae 3841	pRL9	C	A	B
NC_014035	*Rhodobacter capsulatus* SB 1003	pRCB133	C	A	B
NC_009429	*Rhodobacter sphaeroides* ATCC 17025	pRSPA01	B	A	B
NC_009430	*Rhodobacter sphaeroides* ATCC 17025	pRSPA02	A	A	B
NC_009040	*Rhodobacter sphaeroides* ATCC 17029	pRSPH01	A	A	B
NC_011962	*Rhodobacter sphaeroides* KD131	pRSKD131A	C	A	B
NC_011960	*Rhodobacter sphaeroides* KD131	pRSKD131B	A	A	B
NC_009007	*Rhodobacter sphaeroides* 2.4.1	plasmid A	CA	A	B
NC_007488	*Rhodobacter sphaeroides* 2.4.1	plasmid B	C	A	B
NC_007490	*Rhodobacter sphaeroides* 2.4.1	plasmid D	CC	A	B
NC_008386	*Roseobacter denitrificans* OCh 114	pTB1	C	A	B
NC_008387	*Roseobacter denitrificans* OCh 114	pTB2	A	A	B
NC_015729	*Roseobacter litoralis* Och 149	pRLO149_63	A	A	B
NC_015728	*Roseobacter litoralis* Och 149	pRLO149_83	C	A	B
NC_006569	*Ruegeria pomeroyi* DSS-3	megaplasmid	B	A	B
NC_004929	*Ruegeria sp.* PR1b	pSD20	A	A	B
NC_004574	*Ruegeria sp.* PR1b	pSD25	CC	AA	B
NC_008042	*Ruegeria sp.* TM1040	plasmid unnamed	A	A	B
NC_015742	*Sinorhizobium fredii* GR64	p64a	C	A	B

NC_000914	*Sinorhizobium fredii* NGR234	pNGR234a	C	A	B
NC_012586	*Sinorhizobium fredii* NGR234	pNGR234b	CC	AA	BB
NC_009620	*Sinorhizobium medicae* WSM419	pSMED01	C	AA	BB
NC_009621	*Sinorhizobium medicae* WSM419	pSMED02	C	A	B
NC_009622	*Sinorhizobium medicae* WSM419	pSMED03	C	A	B
NC_013545	*Sinorhizobium meliloti*	pSmeSM11a	CC	A	B
NC_003037	*Sinorhizobium meliloti* 1021	pSymA	C	A	B
NC_003078	*Sinorhizobium meliloti* 1021	pSymB	C	AA	BB
NC_015597	*Sinorhizobium meliloti* AK83	pSINME01	C	A	B
NC_010865	*Sinorhizobium meliloti* SM11	pSmeSM11b	C	A	B
NC_014007	*Sphingobium japonicum* UT26S	pCHQ1	AA	A	B
NC_007353	*Sphingomonas sp*. A1	pA1	T	I	K
NC_008308	*Sphingomonas sp*. KA1	pCAR3	W	A	B
NC_009717	*Xanthobacter autotrophicus* Py2	pXAUT01	C	A	B

5.2.1.2 Betaproteobacteria

According to NCBI, up until October 2011, complete genomes in betaproteobacteria have mostly been obtained from three order levels: *Burkholderiales (Achromobacter, Acidovorax, Bordertella, Burkholderia, Collimonas, Comamonas, Cupriavidus, Laribacter, Methylibium, Methylovorus, Polaromonas, Ralstonia, Rhodoferax, Verminephrobacter), Neisseriales (Neisseria),* and *Nitrosomonadales (Nitrosomonas, Nitrosospira)*[174]. Among them, the most abundant Rep and Par systems identified in this thesis were found in the *Burkholderiales* order.

Table 5.2 is a list of plasmids of betaproteobacteria possessing Rep and Par in this study. In comparison with Rep and Par in alphaproteobacteria, there are various combinations of the systems. The majority of plasmids, particularly in the *Burkholderia* genus, have the short ParA-ParB Par system, but their Rep can be RepB-like, RepA-like, TrfA-like (40% similarity with TrfA), or RepA (of IncW, with 40% similarity with RepA). There are several uncharacterised Par systems (not shown in **Table 5.2**), such as the ones in pBVIE04 (which contains a pseudogene next to ParA, which might be non-functioning), pBVIE05 from *Burkhholderia vietnamiensis* G4, and pBPHYT01 from *Burkholderia phytofirmans* PsJN. In addition, Rep and Par of BHR IncP plasmids can be found frequently in betaproteobacterial plasmids.

Figure 5.1 shows one example of the diversity of different replication and partitioning systems within one species. There are five replicons from *Burkholderia glumae* BGR1. 4 replicons possess short ParA-ParB coupled proteins as a partitioning system, but there are 4 different Rep systems: RepB-like, TrfA-like, RepO-like (unique Reps adjacent to *parAB* of *Burkholderia*), and RepA-like. This is a remarkable diversity in comparison with the replicons in *Rhizobium*. *Rhizobium* species possess only RepABC replicons, although their sequences are different enough to make the plasmids compatible. In particular, in the case of the partitioning system of *B. glumae* BGR1, the long ParB (associated with long ParA) does show close homology with ParB in *Vibrio*, rather than the ParB adjacent to short ParA in other plasmids of *B. glumae*, which might indicate that the short ParA and long ParA in *Burkholderia* have different evolutionary histories.

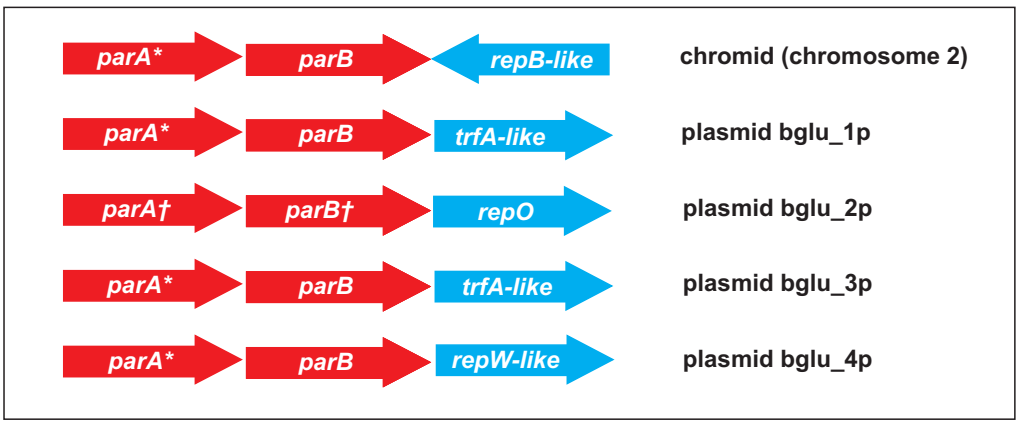


Figure 5.1 The genetic organization of replication and partitioning regions of 5 replicons from *Burkholderia glumae* BGR1

Arrows in red are the partitioning systems and those in blue are the replication system. A long or short version of *parA* and *parB* are indicated as '†' or '*', respectively.

Table 5.2 List of betaproteobacterial plasmids and their Rep and Par systems identified in this study

RepB-like indicated by B, RepA-like by A, TrfA by T, Rep with long ParAB in betaproteobacteria by C, Short ParA by S, ParB with Short ParA by Q, IncC by I, KorB by K, Unique group (that is not identified as a major group) by U, one distantly homologous with TrfA by TrfA-like, RepA by W, one distantly homologous with RepA of IncW plasmids by W-like, ParM by M.

Accession	Species	Plasmids	Initiator	NTPase	Binding protein
NC_013193	Candidatus *Accumulibacter phosphatis* clade IIA str. UW-1	pAph01	T	I	KQ
NC_005793	*Achromobacter denitrificans*	pEST4011	T	I	K
NC_006830	*Achromobacter xylosoxidans* A8	pA81	T	I	K
NC_014641	*Achromobacter xylosoxidans* A8	pA81	T	I	K
NC_008766	*Acidovorax sp.* JS42	pAOVO02	T	I	K
NC_008459	*Bordetella pertussis*	pBP136	T	I	K
NC_008385	*Burkholderia ambifaria* AMMD	plasmid 1	T	I	K
NC_010553	*Burkholderia ambifaria* MC40-6	pBMC401	C	A	B
NC_008545	*Burkholderia cenocepacia* HI2424	plasmid 1	T	S	Q
NC_011003	*Burkholderia cenocepacia* J2315	pBCJ2315	T	S	Q
NC_013666	*Burkholderia cepacia*	pIJB1	TT	I	K
NC_015382	*Burkholderia gladioli* BSR3	bgla_1p	TW	S	Q
NC_015377	*Burkholderia gladioli* BSR3	bgla_2p	U	S	Q
NC_015378	*Burkholderia gladioli* BSR3	bgla_3p	CTW	A	B
NC_015383	*Burkholderia gladioli* BSR3	bgla_4p	BW	S	Q
NC_012723	*Burkholderia glumae* BGR1	bglu_1p	T-like	S	Q
NC_012718	*Burkholderia glumae* BGR1	bglu_2p	C	A	B
NC_012720	*Burkholderia glumae* BGR1	bglu_3p	T-like	S	Q
NC_012725	*Burkholderia glumae* BGR1	bglu_4p	B	S	Q

NC_010070	*Burkholderia multivorans* ATCC 17616	pBMUL01	AT	S	Q
NC_010802	*Burkholderia multivorans* ATCC 17616	pTGL1	A	S	Q
NC_010625	*Burkholderia phymatum* STM815	pBPHY01	CB	A	B
NC_010627	*Burkholderia phymatum* STM815	pBPHY02	C	A	B
NC_014723	*Burkholderia rhizoxinica* HKI 454	pBRH02	C	A	B
NC_014120	*Burkholderia sp*. CCGE1002	pBC201	C	A	B
NC_009230	*Burkholderia vietnamiensis* G4	pBVIE01	B	A-like	B-like
NC_009227	*Burkholderia vietnamiensis* G4	pBVIE02	B	S	Q
NC_009229	*Burkholderia vietnamiensis* G4	pBVIE03	CM	A	B
NC_010332	*Collimonas fungivorans*	pTer331	W	I	K
NC_010935	*Comamonas testosteroni* CNB-1	pCNB	T	I	K
NC_007974	*Cupriavidus metallidurans* CH34	megaplasmid	W	S	Q
NC_006525	*Cupriavidus metallidurans* CH34	pMOL28	C	A	B
NC_007972	*Cupriavidus metallidurans* CH34	pMOL28	C	A	B
NC_015727	*Cupriavidus necator* N-1	BB1p	CA	A	B
NC_015724	*Cupriavidus necator* N-1	BB2p	C	A	B
NC_010529	*Cupriavidus taiwanensis*	pRALTA	C	A	B
NC_005088	*Delftia acidovorans*	pUO1	T	II	K
NC_008826	*Methylibium petroleiphilum* PM1	RPME01	CB	A	B
NC_014105	*Neisseria gonorrhoeae*	pEP5289	T	I	K
NC_008341	*Nitrosomonas eutropha* C91	Plasmid1	T-like	S	Q
NC_003213	*Ralstonia eutropha*	pIPO2T	W	I	K
NC_008757	*Polaromonas naphthalenivorans* CJ2	pPNAP01	ABT	S	Q
NC_008758	*Polaromonas naphthalenivorans* CJ2	pPNAP02	T	S	Q
NC_008760	*Polaromonas naphthalenivorans* CJ2	pPNAP04	CB	A	B
NC_007949	Polaromonas sp. JS666	plasmid 1	BT	S	Q

NC_005241	*Ralstonia eutropha* H16	megaplasmid pHG1	C	A	B
NC_007337	*Ralstonia eutropha* JMP134	plasmid 1	T	I	K
NC_007336	*Ralstonia eutropha* JMP134	megaplasmid	C	A	B
NC_005912	*Ralstonia eutropha* JMP134	pJP4	TT	I	K
NC_012855	*Ralstonia pickettii* 12D	pRp12D01	BW	S	Q
NC_014309	*Ralstonia solanacearum* CFBP2957	RCFBPv3_mp	W	S	Q
NC_003296	*Ralstonia solanacearum* GMI1000	pGMI1000MP	W	S	Q
NC_014310	*Ralstonia solanacearum* PSI07	Megaplasmid	W	S	Q
NC_007901	*Rhodoferax ferrireducens* T118	plasmid1	BTT	S	Q

5.2.1.3 Gammaproteobacteria

As mentioned in chapter 3 and chapter 4, inferring the evolutionary history of gammaproteobacterial plasmids is not easy, because many of them possess more than one replication and/or partitioning system [103]. Moreover, it is difficult to differentiate which Rep or Par operons actually function in the cell. Based on the distribution and phylogenetic trees of plasmids, however, it appears that the mobility of plasmids is not that wide-ranging, but largely restricted within the class level, because Rep and Par systems of specific replicons in gammaproteobacteria do not frequently show up in other divisions of proteobacteria. Presumably their Rep and Par origin is much more complex than other divisions of proteobacteria, and has a long history of Rep and Par system incorporation or exchange over time.

The reason why the analysis is made more complicated is that the combination of Rep and Par systems appears random. Although exceptions include one of the majority patterns, TrfA-IncC/KorB, a main backbone of BHR IncP plasmids, the combination of other replicons especially from *Escherichia*, *Klebsiella, Pseudomonas, Salmonella*, etc. show various combination of Rep and Par systems. They do not show necessarily similar patterns, which might indicate that they are more flexible than alpha- or betaproteobacteria and this might be the result of exchanging the two systems often.

For example, *Vibrio* species have a very unique Rep, which is called RctB [175], but their partitioning systems seem similar to Par of other gammaproteobacterial plasmids, which might indicate a same origin. On the other hand, there is a diversity of Rep and Par systems in *E. coli*, one of the most sequenced bacterial species. The most abundant ones are RepB-ParA-ParB, RepFIIA-ParA-ParB, RepFIA-ParA-ParB, and RepFIA-ParM-ParR, but the combination of the two systems is not consistent. Many members of the ParM-ParR family are detected in this species and are also found in *Klebsiella* and a few in *Salmonella*. As seen in chapter 4, the ParM-ParR family is mostly restricted to gammaproteobacteria, but the Rep systems working with ParM-ParR do not seem to have any typical pattern.

In *Yersinia*, Rep and Par systems show a relatively universal pattern in comparison with other gammaprotebacterial plasmids. The most abundant pattern is RepFIIA-

ParA-ParB or RepFIB-ParA-ParB, which is a main system in pCD-type and pMT-type plasmids respectively. There does exist, however, a unique type of Rep and Par combination, which has not been investigated experimentally. A full list of Rep and Par systems in this study is in **Table 5.3**.

Table 5.3 List of gammaproteobacterial plasmids and their Rep and Par systems identified in this study

TrfA indicated by T, IncC by I, KorB by K, RepFIIA by V, RepFIA by Y, RepFIB by X, ParF by F, ParG by G, SopA/ParA by A, SopB/ParB in Par by B, RepB-like in Rep by B, ParM by M, ParR by R.

Accession	Species	Plasmids	Initiator	NTPase	Binding protein
NC_001621	Birmingham IncP-alpha plasmid	Birmingham IncP-alpha plasmid	TT	II	K
NC_004464	*Citrobacter freundii*	pCTX-M3	V	M	R
NC_002131	*Coxiella burnetii*	QpDV	B	A	B
NC_002118	*Coxiella burnetii*	QpH1	B	A	B
NC_011526	*Coxiella burnetii* CbuK_Q154	pQpRS_K_Q154	B	A	B
NC_009726	*Coxiella burnetii* Dugway 5J108-111	pQpDG	B	A	B
NC_010258	*Coxiella burnetii* MSU Goat Q177	QpRS	B	A	B
NC_010115	*Coxiella burnetii* RSA 331	QpH1	B	A	B
NC_004704	*Coxiella burnetii* RSA 493	pQpH1	B	A	B
NC_013283	*Cronobacter turicensis* z3032	pCTU1	Y	A	B
NC_013285	*Cronobacter turicensis* z3032	pCTU3	VY	F	G
NC_014725	*Edwardsiella tarda*	pCK41	Y	A	B
NC_001735	*Enterobacter aerogenes*	R751	TT	I	K
NC_012555	*Enterobacter cloacae*	pEC-IMP	YY	M	R
NC_012556	*Enterobacter cloacae*	pEC-IMPQ	YY	M	R
NC_014107	*Enterobacter cloacae* subsp. cloacae ATCC 13047	pECL_A	VY	AM	BR
NC_009425	*Enterobacter sp*. 638	pENTE01	VY	A	B
NC_005706	*Erwinia amylovora*	pEA29	X	F	G
NC_005246	*Erwinia amylovora*	pEL60	V	M	R
NC_013972	*Erwinia amylovora* ATCC 49946	plasmid 1	X	F	G

NC_013957	*Erwinia amylovora* CFBP1430	pEA29	X	F	G
NC_004445	*Erwinia pyrifoliae* Ep1/96	pEP36	X	F	G
NC_004834	*Erwinia sp.* Ejp 556	pEJ30	X	F	G
NC_010699	*Erwinia tasmaniensis* Et1/99	pET45	V	M	R
NC_010693	*Erwinia tasmaniensis* Et1/99	pET46	X	A	B
NC_009133	*Escherichia coli*	NR1	V	M	R
NC_004998	*Escherichia coli*	p1658/97	VY	A	B
NC_011980	*Escherichia coli*	pAPEC-1	VY	A	B
NC_011964	*Escherichia coli*	pAPEC-O103-ColBM	VY	M	R
NC_007675	*Escherichia coli*	pAPEC-O2-ColV	VXY	A	B
NC_006671	*Escherichia coli*	pAPEC-O2-R	V	M	R
NC_002142	*Escherichia coli*	pB171	VY	M	R
NC_005327	*Escherichia coli*	pC15-1a	V	M	R
NC_007635	*Escherichia coli*	pCoo	VV	MM	RR
NC_013175	*Escherichia coli*	pEC14_114	VY	M	R
NC_014384	*Escherichia coli*	pEC_L8	VX	A	B
NC_013122	*Escherichia coli*	pEK499	VX	A	B
NC_013121	*Escherichia coli*	pEK516	V	M	R
NC_014615	*Escherichia coli*	pETN48	VY	A	B
NC_010862	*Escherichia coli*	pMAR7	VY	M	R
NC_007365	*Escherichia coli*	pO113	VY	M	R
NC_011812	*Escherichia coli*	pO26-L	V	M	R
NC_012487	*Escherichia coli*	pO26-Vir	VV	M	R
NC_008460	*Escherichia coli*	pO86A1	VY	M	R
NC_009602	*Escherichia coli*	pSFO157	VXYY	A	B
NC_010409	*Escherichia coli*	pVM01	VY	A	B

NC_010558	*Escherichia coli* 1520	pIP1206	VVXY	A	B
NC_010719	*Escherichia coli* 53638	p53638_226	V	AM	BR
NC_010720	*Escherichia coli* 53638	p53638_75	V	M	R
NC_011752	*Escherichia coli* 55989	55989p	VY	M	R
NC_009837	*Escherichia coli* APEC O1	pAPEC-O1-ColBM	VY	A	B
NC_009838	*Escherichia coli* APEC O1	pAPEC-O1-R	YY	M	R
NC_006352	*Escherichia coli* DH5?	pTB11	T	I	K
NC_009787	*Escherichia coli* E24377A	pETEC_35	V	M	R
NC_009788	*Escherichia coli* E24377A	pETEC_73	V	M	R
NC_009790	*Escherichia coli* E24377A	pETEC_74	V	A	B
NC_009786	*Escherichia coli* E24377A	pETEC_80	V	M	R
NC_011754	*Escherichia coli* ED1a	pECOED	V	M	R
NC_014232	*Escherichia coli* ETEC 1392/75	p1081	V	M	R
NC_014233	*Escherichia coli* ETEC 1392/75	p557	VY	A	B
NC_014234	*Escherichia coli* ETEC 1392/75	p746	V	M	R
NC_007680	*Escherichia colii* J53	pTP6	TT	II	K
NC_013354	*Escherichia coli* O103:H2 str. 12009	pO103	Y	A	B
NC_013365	*Escherichia coli* O111:H- str. 11128	pO111_1	XYY	M	R
NC_013370	*Escherichia coli* O111:H- str. 11128	pO111_2	Y	A	B
NC_013366	*Escherichia coli* O111:H- str. 11128	pO111_3	V	M	R
NC_011603	*Escherichia coli* O127:H6 str. E2348/69	pMAR2	VY	M	R
NC_011350	*Escherichia coli* O157:H7 str. EC4115	pO157	VY	A	B
NC_007414	*Escherichia coli* O157:H7 str. EDL933	pO157	VXY	A	B
NC_002128	*Escherichia coli* O157:H7 str. Sakai	pO157	VY	A	B
NC_013010	*Escherichia coli* O157:H7 str. TW14359	pO157	VX	A	B
NC_013728	*Escherichia coli* O26:H-	pO26-CRL	VY	M	R
NC_013369	*Escherichia coli* O26:H11 str. 11368	pO26_1	V	M	R

NC_013362	*Escherichia coli* O26:H11 str. 11368	pO26_2	V	M	R
NC_013942	*Escherichia coli* O55:H7 str. CB9615	pO55	Y	A	B
NC_011747	*Escherichia coli* S88	pECOS88	V	A	B
NC_011419	*Escherichia coli* SE11	pSE11-1	V	AM	BR
NC_010488	*Escherichia coli* SMS-3-5	pSMS35_130	V	A	B
NC_011749	*Escherichia coli* UMN026	p1ESCUM	VY	M	R
NC_007941	*Escherichia coli* UTI89	pUTI89	VY	M	R
NC_012944	*Escherichia coli* Vir68	pVir68	VXXY	A	B
NC_011743	*Escherichia fergusonii* ATCC 35469	pEFER	X	A	B
NC_008272	*Escherichia coli* MC1061	pKJK5	T	I	K
NC_008055	*Escherichia coli*	QKH54	T	I	K
NC_006388	*Escherichia coli* DH5alfa	pB3	T	I	K
NC_002483	*Escherichia coli* K-12 strain CR63	plasmid F	XY	A	B
NC_002122	*Escherichia coli* strain K-12	ColIb-P9	V	M	R
NC_012752	Candidatus *Hamiltonella defensa* 5AT	pHD5AT	V	F	G
NC_011641	*Klebsiella pneumoniae*	pCTXM360	V	M	R
NC_010886	*Klebsiella pneumoniae*	pK245	XXY	A	B
NC_010870	*Klebsiella pneumoniae*	pK29	YY	M	R
NC_013951	*Klebsiella pneumoniae*	pKF3-140	VXY	A	B
NC_013542	*Klebsiella pneumoniae*	pKF3-70	V	M	R
NC_013950	*Klebsiella pneumoniae*	pKF3-94	VY	M	R
NC_014312	*Klebsiella pneumoniae*	pKP048	VX	AM	BR
NC_014016	*Klebsiella pneumoniae*	pKpQIL	VY	M	R
NC_005249	*Klebsiella pneumoniae*	pLVPK	YY	A	B
NC_010726	*Klebsiella pneumoniae*	pMET-1		F	G
NC_015154	*Klebsiella pneumoniae*	pc15-k	VY	AM	BR
NC_011282	*Klebsiella pneumoniae* 342	pKP187	YY	A	B

NC_011281	*Klebsiella pneumoniae* 342	pKP91	VX	A	B
NC_009649	*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578	pKPN3	VY	A	B
NC_009650	*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578	pKPN4	VY	AM	BR
NC_009651	*Klebsiella pneumoniae* subsp. pneumoniae MGH 78578	pKPN5	XX	AM	BR
NC_006625	*Klebsiella pneumoniae* subsp. pneumoniae NTUH-K2044	pK2044	YY	A	B
NC_006365	*Legionella pneumophila* str. Paris	pLPP	B	A	B
NC_008738	*Marinobacter aquaeolei* VT8	pMAQU01	X	AM	BR
NC_014838	*Pantoea sp.* At-9b	pPAT9B01	Y	A	B
NC_014839	*Pantoea sp.* At-9b	pPAT9B02	Y	A	B
NC_014840	*Pantoea sp.* At-9b	pPAT9B03	XY	F	G
NC_014841	*Pantoea sp.* At-9b	pPAT9B04	Y	A	B
NC_014561	*Pantoea vagans* C9-1	pPag1	Y	A	B
NC_014563	*Pantoea vagans* C9-1	pPag2	VY	A	B
NC_014258	*Pantoea vagans* C9-1	pPag3	Y	A	B
NC_003905	*Proteus vulgaris*	Rts1	Y	A	B
NC_008357	*Pseudomonas aeruginosa*	pBS228	T	I	K
NC_004956	*Pseudomonas sp.* ADP	pADP-1	T	I	K
NC_004840	*Pseudomonas sp.* B13	pB10	T	I	K
NC_003430	*Pseudomonas sp.* B13	pB4	TT	I	K
NC_007502	*Pseudomonas sp.* B13 GFP1	pB8	T	I	K
NC_015062	*Rahnella sp*. Y9602	pRAHAQ01	X	AFM	BR
NC_015063	*Rahnella sp*. Y9602	pRAHAQ02	X	AF	BG
NC_007208	*Salmonella enterica*	pOU1113	V	A	B

NC_009981	*Salmonella enterica* subsp. enterica serovar Choleraesuis	pMAK1	YY	M	R
NC_010119	*Salmonella enterica* subsp. enterica serovar Choleraesuis	pOU7519	YY	AA	B
NC_006856	*Salmonella enterica* subsp. enterica serovar Choleraesuis str. SC-B67	pSC138	Y	M	R
NC_006855	*Salmonella enterica* subsp. enterica serovar Choleraesuis str. SC-B67	pSCV50	VY	A	B
NC_010422	*Salmonella enterica* subsp. enterica serovar Dublin	pOU1115	V	F	G
NC_011204	*Salmonella enterica* subsp. enterica serovar Dublin str. CT_02021853	pCT02021853_74	V	F	G
NC_011081	*Salmonella enterica* subsp. enterica serovar Heidelberg str. SL476	pSL476_91	V	M	R
NC_011077	*Salmonella enterica* subsp. enterica serovar Kentucky str. CVM29188	pCVM29188_101	V	M	R
NC_011076	*Salmonella enterica* subsp. enterica serovar Kentucky str. CVM29188	pCVM29188_146	VY	A	B
NC_011078	*Salmonella enterica* subsp. enterica serovar Kentucky str. CVM29188	pCVM29188_46	V	F	G
NC_012124	*Salmonella enterica* subsp. enterica serovar Paratyphi C strain RKS4594	pSPCV	VY	A	B
NC_011092	*Salmonella enterica* subsp. enterica serovar Schwarzengrund str. CVM19633	pCVM19633_110	Y	A	B
NC_002305	*Salmonella enterica* subsp. enterica serovar Typhi	R27	XYY	M	R
NC_003384	*Salmonella enterica* subsp. enterica serovar Typhi str. CT18	pHCM1	XYY	M	R
NC_005014	*Salmonella enterica* subsp. enterica serovar	R64	V	M	R

	Typhimurium				
NC_013437	*Salmonella enterica* subsp. enterica serovar Typhimurium	pSLT-BT	VY	A	B
NC_006816	*Salmonella enterica* subsp. enterica serovar Typhimurium	pU302L	XY	A	B
NC_002523	*Serratia entomophila*	pADAP	X	A	B
NC_005211	*Serratia marcescens*	R478	YY	M	R
NC_004349	*Shewanella oneidensis* MR-1	megaplasmid	Y	A	B
NC_010660	*Shigella boydii* CDC 3083-94	pBS512_211	V	M	R
NC_007608	*Shigella boydii* Sb227	pSB4_227	V	A	B
NC_007607	*Shigella dysenteriae* Sd197	pSD1_197	VV	AMM	BRR
NC_004851	*Shigella flexneri* 2a str. 301	pCP301	V	AM	BR
NC_002134	*Shigella flexneri* 2b strain 222	R100	V	M	R
NC_002698	*Shigella flexneri* 5a	pWR501	V	AM	BR
NC_013727	*Shigella sonnei*	pEG356	V	M	R
NC_007385	*Shigella sonnei* Ss046	pSS_046	V	A	B
NC_010579	*Xylella fastidiosa* M23	pXFAS01	W	I	K
NC_004564	*Yersinia enterocolitica*	pYVa127/90	V	A	B
NC_005017	*Yersinia enterocolitica*	pYVe8081	V	A	B
NC_008791	*Yersinia enterocolitica* subsp. enterocolitica 8081	pYVe8081	V	A	B
NC_015475	*Yersinia enterocolitica* subsp. palearctica 105.5R(r)	105.5R(r)p	V	A	B
NC_006323	*Yersinia pestis*	pG8786	VY	A	B
NC_010157	*Yersinia pestis* Angola	new_pCD	V	A	B
NC_010158	*Yersinia pestis* Angola	pMT-pPCP	Y	A	pseudo
NC_008122	*Yersinia pestis* Antiqua	pCD	V	A	B
NC_008120	*Yersinia pestis* Antiqua	pMT	Y	A	B
NC_009595	*Yersinia pestis* CA88-4125	pCD1	V	A	B

NC_009596	*Yersinia pestis* CA88-4125	pMT1	Y	A	B
NC_003131	*Yersinia pestis* CO92	pCD1	V	A	B
NC_003134	*Yersinia pestis* CO92	pMT1	Y	A	B
NC_015056	*Yersinia pestis* Java 9	pCD	V	A	B
NC_004836	*Yersinia pestis* KIM 10	pCD1	V	A	B
NC_004839	*Yersinia pestis* KIM 10	pCD1	V	A	B
NC_004838	*Yersinia pestis* KIM 10	pMT-1	Y	A	B
NC_004835	*Yersinia pestis* KIM 10	pMT1	Y	A	B
NC_008118	*Yersinia pestis* Nepal516	pMT	Y	A	B
NC_009377	*Yersinia pestis* Pestoides F	CD	V	A	B
NC_009378	*Yersinia pestis* Pestoides F	MT	VY	A	B
NC_014017	*Yersinia pestis* Z176003	pCD1	V	A	B
NC_014022	*Yersinia pestis* Z176003	pMT1	Y	A	B
NC_005813	*Yersinia pestis* biovar Microtus str. 91001	pCD1	V	A	B
NC_005815	*Yersinia pestis* biovar Microtus str. 91001	pMT1	Y	A	B
NC_011759	*Yersinia pseudotuberculosis*	pGDT4	V	F	G
NC_009705	*Yersinia pseudotuberculosis* IP 31758	plasmid_153kb	X	M	R
NC_006153	*Yersinia pseudotuberculosis* IP 32953	pYV	VX	A	B
NC_010635	*Yersinia pseudotuberculosis* PB1/+	pYPTS01	V	A	B

5.2.2 Phylogenies of Rep and Par systems

5.2.2.1 (Long) ParA system with various Rep systems

As mentioned in chapter 3, Rep systems are largely conserved at the order level of bacterial species. On the other hand, many replicons that we have investigated in proteobacteria ensure partitioning of daughter cells using Type I ParA ATPase. The pattern is not just restricted to the Proteobacteria, but is generally found in a wide range of bacteria. **Figure 5.2** is a rooted tree from chapter 4. It depicts a phylogeny based on all the Type I ParA sequences of partitioning systems. Here, we have indicated replication systems on the tree in order to compare the patterns of the two systems. Most partitioning systems occur in association with the Rep systems that we have identified, although there are several exceptions including the DnaA-family and unknown Rep in *Vibrio* and *Shewanella* plasmids. Each conserved clade has a specific Rep family.

5.2.2.2 IncC/KorB with TrfA

As shown briefly in chapter 4, most patterns of distribution regarding replication and partitioning systems in broad host range bacteria are almost similar to each other (**Figure 5.3**). This is because IncP plasmids commonly have the TrfA-type replication system and the IncC/KorB-type partitioning systems (see section 3.3.7 in chapter 3 and section 4.3.1.2 in chapter 4). As seen in **Figures 4.10** and **4.11**, the patterns of the clades are similar. Inside, the topologies of the two systems also mirror each other. A few exceptions have been found, for example the Par system of plasmid 2 from *Aromatomeum aromaticum* EbN1 and plasmid 59kb from *Yersinia pseudotuberculosis* IP 31758 are closely related to IncP plasmids, but their Rep systems seem different (undefined Rep in this study). Presumably, their Par system might have originated from IncP plasmids, which is why it has a different evolutionary origin from their Rep systems.

Moreover, a group of IncC sequences, which is relatively diverged from the groups defined as IncP plasmids, is related to a different type of replication system. This includes IncW plasmids, such as pSB102, pMRAD02, pIPO2T, pTer331, etc. A possible explanation for this is that the partitioning systems of IncW plasmids

might have been incorporated from plasmids of the IncP type a long time ago. In particular, pSB102 (NC_003122) from *Sinorhizobium meliloti* strain FP2 and pMRAD02 of *Methylobacterium radiotolerans* JCM 2831 has RepA of IncW-like plasmids and IncC/KorB of IncP plasmids as seen in the previous section. They are specifically related to two betaproteobacterial plasmids in both Rep and Par, which might indicate a very old inter-class transfer (**Figure 5.3**).

5.2.2.3 Short ParA, ParF and related Rep systems

Figure 5.4 is a phylogenetic tree of Short ParA and ParF from chapter 4, and we have indicated Rep systems in the tree. Basically, each distinct clade in the partitioning phylogeny has their Rep system. Exceptions include firstly bgla_1p from *Burkholderia gladioli* BSR3, and bglu_4p from *Burkholderia glumae* BGR1, which have TrfA-like and RepA systems, and RepB-like systems respectively. There is another interesting clade having ParFs of two alphaproteobacterial plasmids, pACMV6 from *Acidiphilium multivorans* AIU301, pZZM401 from *Zymomonas mobilis* subsp. Mobilis ZM4, and short ParA of one betaproteobacterial plasmid pPNAP01 from *Polaromonas naphthalenivorans* CJ2. Three Rep sequences exist in pPNAP01, which are RepA-like, RepB-like and TrfA-like proteins. However, none of the three sequences is homologous to Rep of pACMV6 and pZZM401, which indicates that a recombination event might have happened between Rep and Par systems during their evolution.

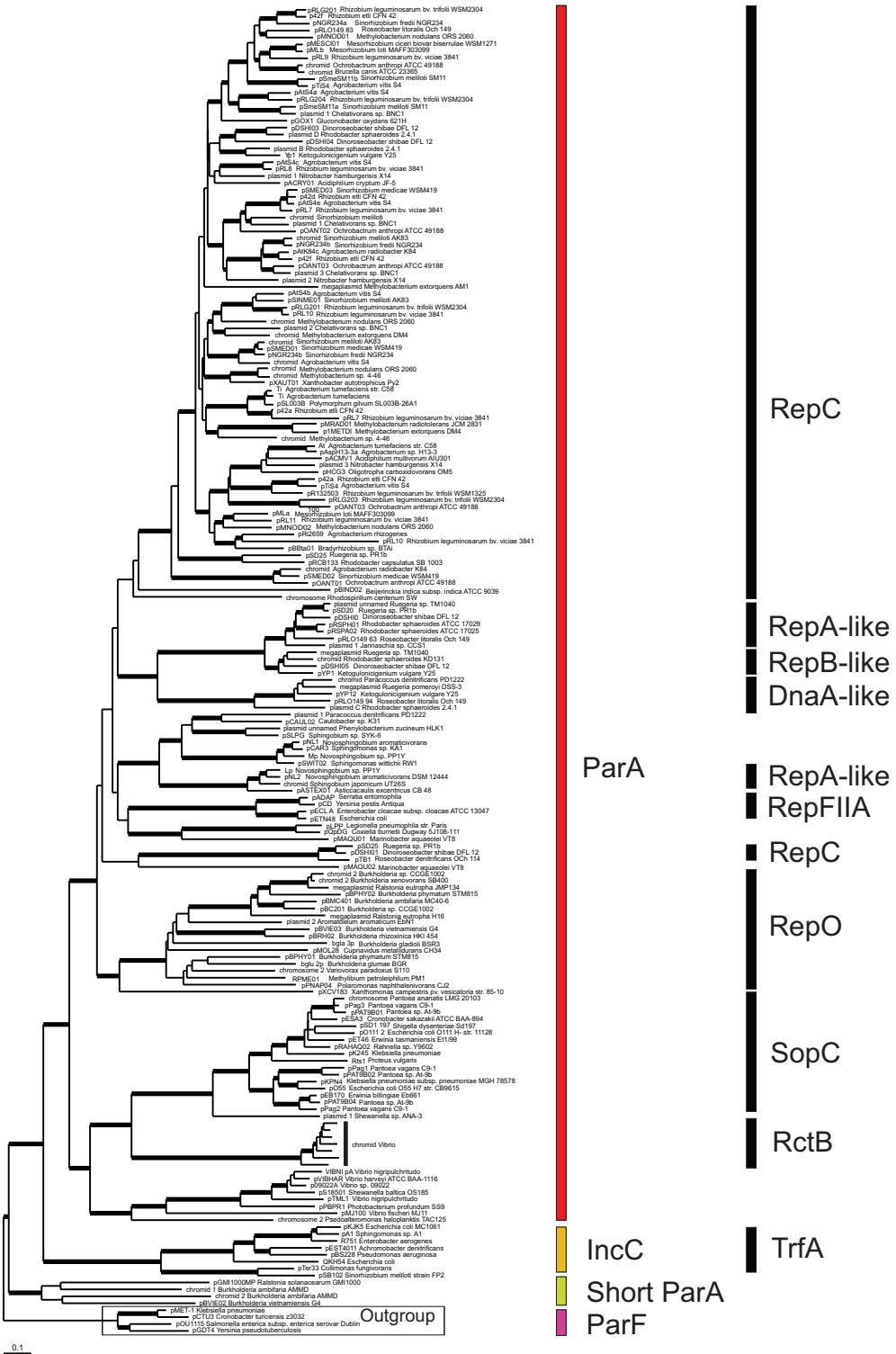


Figure 5.2 The phylogenetic tree of ParA homologs with different Rep systems

Taken from chapter 4. Different types of replication systems investigated in chapter 3 are presented. The clade including ParF homologs was used as the out-group. The coloured vertical bars represent different partitioning systems (ParA in red, IncC in orange, Short ParA in green, and ParF in purple) and the black bars indicate different replication systems.

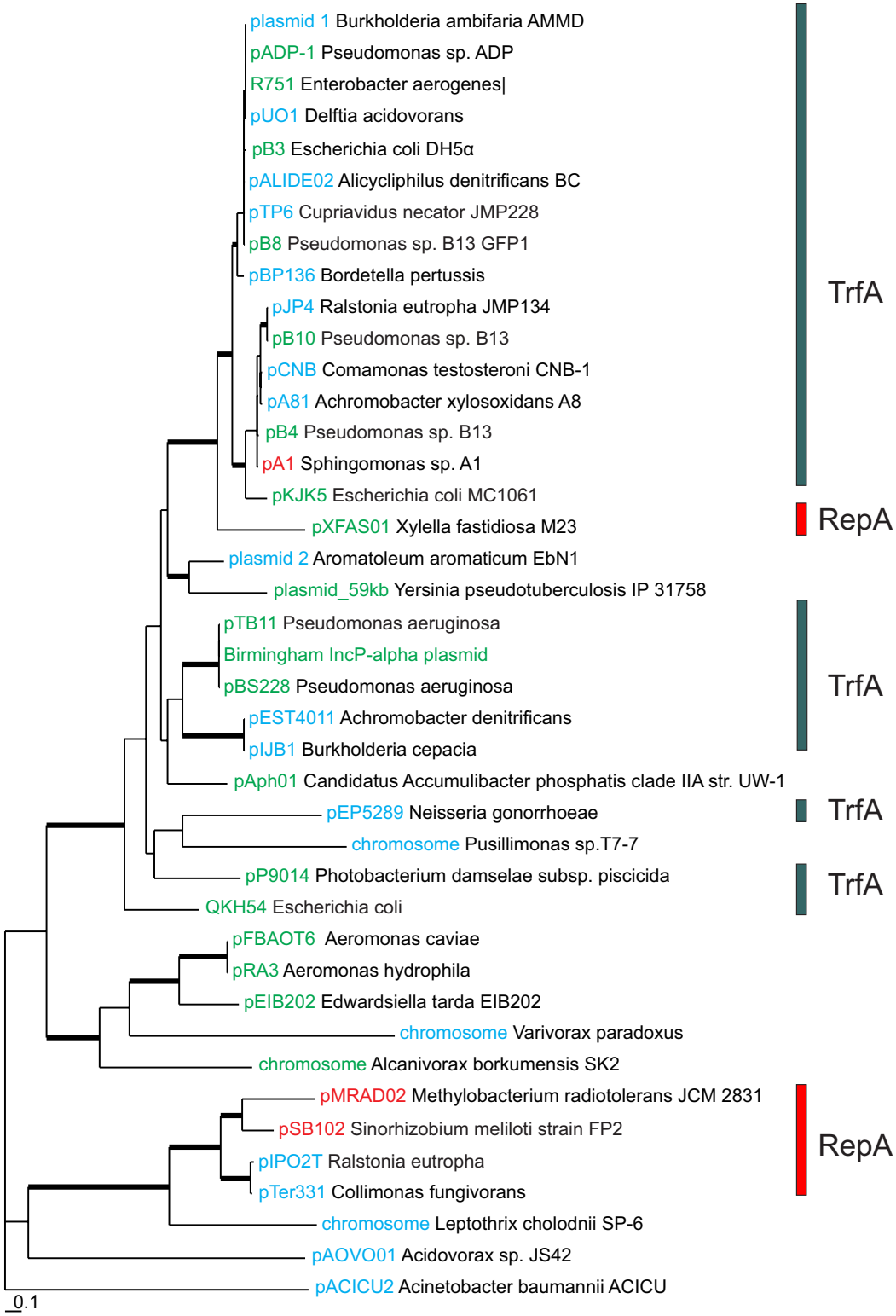


Figure 5.3 The phylogenetic tree of IncC homologs

Taken from chapter 4. Different types of replication systems investigated in chapter 3 are presented in boxes. Green vertical bars represent plasmids possessing TrfA initiators and red bars indicate plasmids having RepA initiators.

!

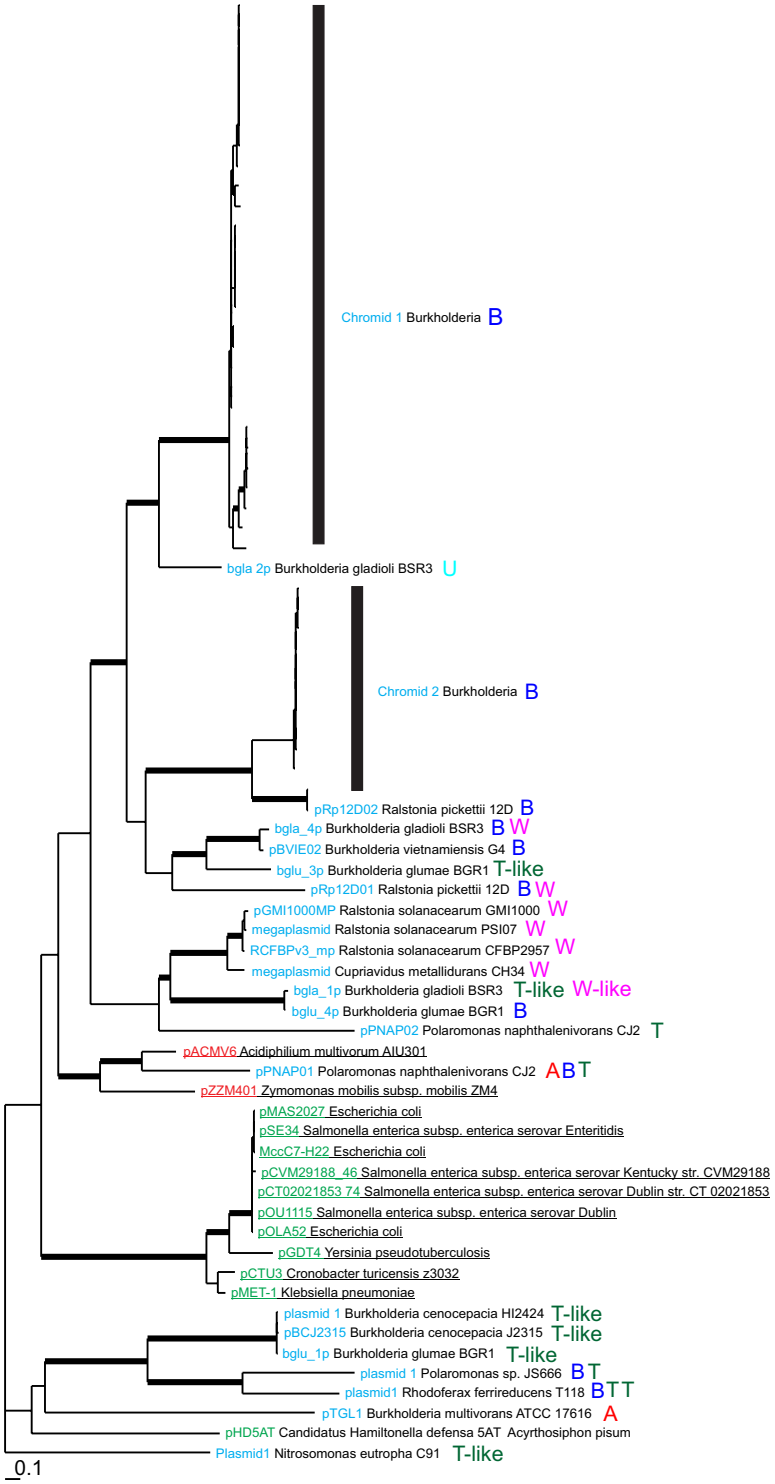


Figure 5.4 The phylogenetic tree of short ParA and ParF homologs

Taken from chapter 4. Different types of replication systems investigated in chapter 3 are presented. Indicated RepB-like by B, A by RepA-like, RepA by W, TrfA by T. Note that some plasmids have multiple replication initiators, which are indicated by multiple letters (e.g. BTT).

5.3 Discussion

5.4.1 Contribution of this study and general questions

In chapters 3 and 4, we have investigated plasmid replication and partitioning systems, respectively, in proteobacteria. Since both systems are essential components, in particular for many low-copy number plasmids and chromids, the research of individual systems of plasmids would help to understand their evolutionary history. This chapter mainly aimed to study general patterns of Rep and Par systems in terms of the relationships between them and distribution of individual divisions of proteobacteria. In addition, we have studied the evolutionary history of Rep and Par systems by comparing their phylogenies.

5.4.1.1 Patterns of Rep or Par systems

Among plasmids in proteobacteria, several do not have the Rep initiator protein. This is curious because the definition of plasmids is that they can replicate by themselves. This can be explained, however, when it becomes clear that plasmids may use Rep borrowed from their host machinery [176]. On the other hand, it may be that in most cases the genes for replication have simply not been characterized yet, so that we just do not know which sequences are translated to the Rep proteins.

In addition, a significant number of replicons do not have Par. Unlike Rep, many replicons actually do not need Par systems if the replicons are not low-copy number plasmids. Alternatively, if they have other systems, such as post-segregational killing or multimer resolution systems, plasmids do not necessarily require Par systems [43, 173, 177, 178]. Therefore, although most large replicons (including plasmids and chromids) do have replication and maintenance systems at the same time [89, 90], especially when they need to be partitioned very carefully, we have found many replicons that simply do not have Par systems (chapter 4).

Nevertheless, many replicons do have both systems, and it was observed that there are many typical patterns. RepABC is one of the best studied examples showing a clear pattern, having three genes as an operon. Many replicons in

alphaproteobacteria have this RepABC pattern. SopABC (indicated by 'ParAB' and 'Rep' in this study), in addition, is also located in many replicons with a typical pattern, particularly the ones from gammaproteobacteria. BHR IncP plasmids, in general, have the TrfA, IncC and KorB combination, which is also very frequent. On the other hand, there are many cases in which one ParAB system is associated with different Rep; for instance, many ParABs have a combination with either RepA-like or RepB-like systems, and the examples were seen in alpha-, beta-, and gammaproteobacteria. In addition, several IncW plasmids have the RepA system with IncC and KorB homologs or short ParA and ParB. In gammaproteobacteria, plasmids having the ParAB system can have RepFIA, RepFIIA, and RepFIB, and this does not seem to have any typical pattern, based on the replicons that we have studied.

5.4.1.2 Good for investigating evolution and classification of plasmids?

In chapters 3 and 4, we have continually asked whether the Rep and/or Par systems of plasmids can be a good indicator for researching plasmids' evolution. Basically, these two systems are a significant source in this direction. Rep systems can be a good indicator for tracking the evolution of plasmids. In addition to this, the phylogenies of Par can shed light on recombination events. As discussed in sections 3.4.2 and 4.4.2, it would be a good criterion for specific types of plasmids to be classified based on two systems. Firstly, RepABC replicons are definitely one of the best examples involved in this, because of the possible co-evolution of Rep and Par systems. BHR plasmids are also appropriate as a model system for research on evolution and classification. As seen in section 5.2.2, however, it is very tricky to determine whether multireplicons that have more than two Rep systems, or Par systems, or both, can be classified according to this criterion.

5.4.1.3 Evolution of Rep, Par systems, and plasmids

As seen above, there have been several attempts to investigate diversity and evolutionary history of plasmids using Rep and Par systems. Cevallos et al. [89] have reviewed the RepABC systems in detail, including regulatory motifs, conserved domains in the systems, plasmid copy number control, etc., but their

Neighbour Joining phylogenies based on alphaproteobacteria have not provided comprehensive information. Castillo-Ramírez et al. [179] also published their idea in terms of HGT in alphaproteobacteria. Their codon adaptation index (CAI) analysis was interesting although they also have provided limited taxonomic range of the host bacteria, which offers the phylogenies consisting of a small number of RepABC replicons in *Rhizobiales.* There is a more recent paper by Mazur et al. [180] that also has RepABC trees. However, their study is less broad in scope, concerning just one species. As mentioned in the introductory section, Petersen et al. [97] and two earlier publications [91, 140] have also provided interesting ideas regarding the evolutionary history of Rep and Par systems. However, their analysis was only based on *Rhodobacterales*, although they suggested using the scheme to classify the plasmids and track the evolutionary history of plasmids in other bacterial orders.

We, therefore, believe that this study is meaningful because it provides comprehensive information regarding Rep and Par systems across proteobacterial replicons, based on the study of distribution, and the phylogenetic analysis of the two systems. Certainly, there are missing Rep and Par systems in this study, particularly for gammaproteobacterial plasmids. However, the framework established in this chapter can readily be expanded by including new families as more examples become available.

5.4.2 Future directions

5.4.2.1 Plasmids transfer systems

As mentioned in chapter 1, plasmid backbone systems can be categorized into three parts. The first two, replication and partitioning, were investigated in chapters 3 and 4, while the third is transfer. As reviewed in detail in chapter 1, research on plasmid transfer systems might give a different perspective on plasmid evolution. It is especially valuable for large conjugative plasmids, because the analysis of transfer systems might indicate the ability to move to other species, rather than replicate and contribute to maintenance as is the case with Rep and

Par systems. Smillie et al. [45] have developed a database storing all homologous genes involved in transfer systems. Their analysis, which involves searching for homologs and generating phylogenies, is largely similar to the analysis in this thesis. In future, research based on all three systems might be a good basis to investigate the evolution of plasmids in general.

5.4.2.2 More families definitely needed, but how does it work?

As seen in chapters 3 and 4, the families that were analysed in this thesis focused on the most abundant Rep and Par systems in proteobacteria available in public genomes from NCBI. This means that because of shortage of information, we might have missed out on some families that are not abundant but might be important in terms of understanding the evolution of bacterial plasmids. Therefore, eventually it is important to collect Rep and Par systems in their totality.

It is not easy, however, to define all the families based on the published genomes so far. As many partitioning genes share the same motifs (see chapter 4), it is difficult to assign automatically homologous sets to individual families. This is one of the disadvantages of gene-by-gene search analysis. Moreover, the published genomes are still not enough to develop effective phylogenetic trees to investigate the evolution of plasmids. Therefore, the efforts to not only collect published genomes, but also screen individual Rep and Par families should be intensified.

Chapter 6. Investigation into the *repABC* replicons of 72 *Rhizobium leguminosarum* strains

In this chapter, we perform a case study to investigate the distribution of one of the typical replication and partitioning systems in bacteria, the *repABC* operon. We analyse nucleotide sequences of 72 *Rhizobium leguminosarum* strains from Wentworth College in York. The GS Reference Mapper and our in-house HMMs pipeline is used to search and map 454 reads against a list of reference genomes. A phylogenetic tree based on the alignment of all the *repABC*operons will be firstly constructed to research general sequence variation within each type. We will also study "magnified" phylogenies within each type, in order to investigate the amount of strain-to-strain spread of the plasmids. We aim to detect whether there is any frequent movement between strains based on the phylogenies of each plasmid type.

6.1 Background

6.1.1 Into the world of rhizobia

There has been a variety of research of nitrogen-fixing bacteria. Rhizobia are the best known among these bacteria. Depending on the legumes that the bacterium is associated with, rhizobia include several species, such as *Sinorhizobium meliloti* (the symbiont of alfafa) and *Bradyrhizobium japonicum* (soybean). *Rhizobium leguminosarum* bv. Viciae 3841 [2], which is the reference genome of this study, was isolated from a nodule on pea (*Pisum sativum*) in England. This was entirely sequenced and published in 2006 [2].

In this study, we have investigated *repABC* operons in 72 strains of *R. leguminosarum* in total. This project was started in 2007 and is ongoing. The strains were isolated from a meter squared near Wentworth College in the University of York. Of the 72 strains, 36 strains are biovar trifolii (called TRX_n, where 'n' is a strain number), while the rest of the 36 strains are biovar viciae (called VSX_n).

6.1.2 The phylogenetic tree of core genomes on 72 *Rhizobium leguminosarum* strains

When Young et al. [2] published their research on *R. leguminosarum* 3841, they have importantly pointed out the concept of two distinct components: a 'core' and an 'accessory' genome (**Figure 6.1**). The 'core' genome includes all the genes for essential functions in bacteria, shows higher G+C value and is located mostly in chromosomes. Normally, these are evolutionarily conserved. On the other hand, the 'accessory' component is known to be dispensable and is mostly located in plasmids and chromosomal islands sporadically. It has a low G+C value and is usually not evolutionarily conserved.

In the case of *R. leguminosarum* 3841, there is one chromosome, two chromids, and four plasmids, which are approximately 7.75 Mb in total. The chromosome (5.05 Mb) is the largest replicon containing a variety of genes that are essential for living. Most genes have high GC content and dinucleotide relative abundance (DRA) is fairly uniform [2]. Conversely, others show a variety of forms; two large chromids such as pRL12 and pRL11, two medium plasmids such as a symbiotic plasmid pRL10 and pRL9, and two relatively smaller plasmids pRL7 and pRL8, which are transferable by conjugation. Although most essential genes are located in chromosomes, plasmids do confer significant genes including symbiosis, ABC transporters, cell division proteins, etc [2].

An interesting feature that the plasmids in *R. leguminosarum* 3841 commonly share is that they all possess a related locus, called the *repABC* operon, which encodes a replication and partitioning system. Three genes are involved in *repABC* operons: *repA* and *repB* are responsible for active partitioning systems (normally

called *parA* and *parB*) and *repC* is a replication initiator protein. These genes are mostly located in the same order and a couple of intergenic sequences are located in the operons, which seem to play a role in the control of replication [89, 181].

Harrison et al. [84] published the definition of 'chromids'. Of the seven replicons in *R. leguminosarum* 3841, pRL12 and pRL11 are defined as chromids. Chromids are plasmids that have become chromosome-like through the acquisition of core genes, but they do have plasmid replication systems (*repABC* operons), which make them exhibit different features from normal chromosomes. The concept of 'core genes' is also presented in this article [84]. In short, the core genes are defined by conserved sets, which offer the fundamental processes, such as protein synthesis and information processing, in addition to covering a wide phylogenetic range.

Nitin Kumar (personal communication) has created a phylogenetic tree based on 305 core genes of the 72 Wentworth *R. leguminosarum* strains with the reference genome *R. leguminosarum* 3841 (**Figure 6.2**). The tree can be divided into five cryptic species presented from A to E. A big group, composed only of 52 strains is presented on the right and is called the cryptic species C. The cryptic species B on the left side includes *R. leguminosarum* 3841 and 12 more strains. The cryptic species A has only one strain and the cryptic species D and E contain 4 and 3 strains respectively. Generally, it does not appear that the core genome of the two biovars is phylogenetically distinct from one another. **Table 6.1** contains a complete list of each cryptic group.

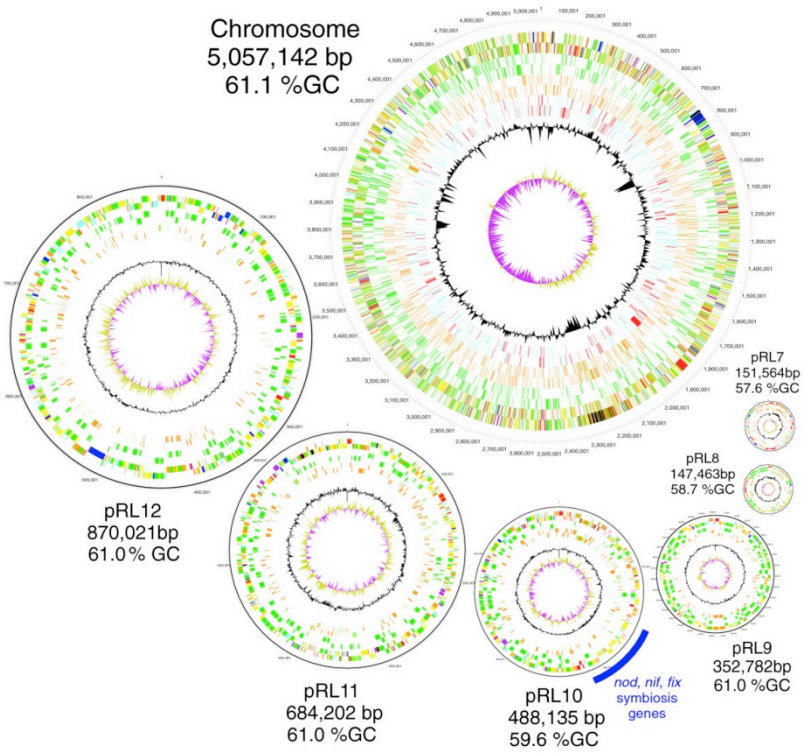


Figure 6.1 Core and accessory genomes in the species *Rhizobium leguminosarum* 3841

The genome comprises one chromosome, two chromids (pRL12, pRL11), and four plasmids (pRL10, 9, 8, 7). The chromids and plasmids are presented at the same relative scale, and the chromosome at one-fourth of that scale. The figure is taken from Young et al. [2] and the colour scheme and the structure of the circles in the figure are explained in (2).

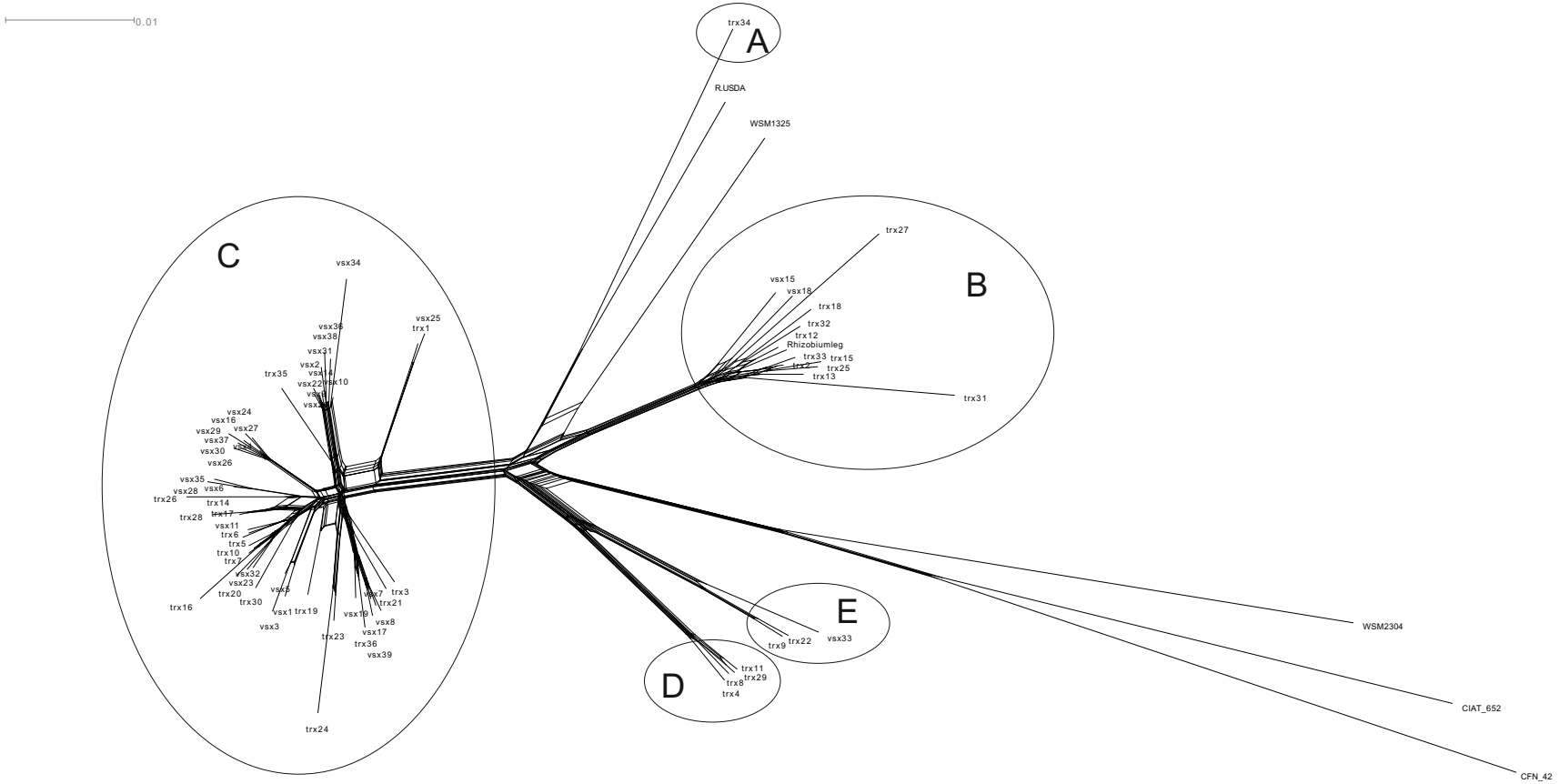


Figure 6.2 Neighbour-net phylogeny of the core genome of 72 *Rhizobium leguminosarum* strains

Nucleotide sequences of core genes from 72 *R. leguminosarum* strains and other related genomes (*R. leguminosarum* 3841, *R. leguminosarum*WSM 2304, *R. etli*CIAT 652 and *R. etli* CFN 42) were concatenated and used to construct the tree. Five different cryptic species in the 72 strains are indicated as A to E. (Nitin Kumar, personal communication).

Table 6.1 List of five cryptic species in 72 *Rhizobium leguminosarum* strains based on the core genes

Type	Biovar trifolii(TRX)	Biovar viciae(VSX)	(TRX/VSX) Total
A	34	none	(1/0) 1
B	2 12 13 15 18 25 27 31 32 33	15 18	(10/2) 12
C	1 3 5 6 7 10 14 16 17 19 20 21 23 24 26 28 30 35 36	1 2 3 4 5 6 7 8 9 10 11 14 16 17 19 21 22 23 24 25 26 27 28 29 30 31 32 34 35 36 37 38 39	(19/33) 52
D	4 8 11 29	None	(4/0) 4
E	9 22	33	(2/1) 3

6.1.3 Chapter objectives

The object of this chapter is to investigate replication and partitioning systems (*repABC* operons) in *R. leguminosarum* strains. This project is part of the research conducted by the Young Group at the University of York. Overall, the objective is different from that pursued in Chapters 2 and 3. While we have investigated the replication and partitioning systems based on publicly available genomes in order to gain an understanding of the general distribution and classification across whole bacteria, in this chapter we have investigated one specific species, *R. leguminosarum*. Therefore, we have focused more on distribution within species, rather than across bacteria. Based on the reads and contigs assembled from the 454 sequencing system, we would like to:

I. Obtain a list of all *repABC* types in the 72 TRX and VSX strains, including the 7 particular types (pRL12, pRL11, pRL10, pRL9, pRL8, pRL7a, pRL7b) that are represented in *R. leguminosarum* 3841 (the reference strain) and all new ones, if any exist. We explore the distribution of plasmids across the 5 cryptic species.
II. Construct a phylogenetic tree based on the alignment of all the *repABC* operons in the strains. This would show general sequence variation within each type of *repABC* region. Moreover, it would help with the classification of discrete types that might represent possible examples of recombination between types.
III. Generate a 'magnified' version of phylogenetic trees within each type of *repABC*. The phylogenies of each type will demonstrate the amount of strain-to-strain spread of the plasmids and will be used to compare the distinctive features that the movement of larger or smaller plasmids might have.

6.2 Materials and methods

6.2.1 GS reference mapper

In Chapter 2, we have shown that each clade of the phylogenetic tree (**Figure 3.3-10**) generally indicates for the most part the actual incompatibility groups. It is important to observe that the sequences within a clade are not able to coexist in a bacterial host since the plasmids, whose sequences are nearly the same, have to compete with each other. Based on the incompatibility groups indicated by the phylogenetic tree, therefore, we were able to make our own list of reference genomes by selecting one representative from each incompatibility group. The list is shown in **Table 6.2**. In this table, ' * ' denotes the representatives from each group.

In this study, we have used the GS Reference Mapper from the 454 sequencing system software (version 2.6) by Roche. The list of the representatives above was used as the reference. In short, the GS Mapper aligns the sequencing reads against the reference genomes with or without associated annotation and produces the output including contigs, sequence alignments and basic statistical information, such as the percentage of matching reads. The GS Reference Mapping system is easy to use via a Graphical User Interface and generates files that are ready for further analysis. Default parameters have been chosen in this chapter.

Using GS mapper and blast search against all reads and contigs, we have obtained a list of the *repABC* systems in Wentworth strains, and investigated their distribution across stains. It should be noted that this was done based on assemblies of each strain separately. However, it is possible that there are additional *rep* systems that have been missed in this study. The sequencing coverage was low (average about 2x, but as low as 1x for some strains), so some genes may not be represented in the contigs. Because this study only included *repABC* operons when all three genes are present in the genome, and are long enough to be aligned, some contigs that only have a part of the *rep* operon have been excluded. For example, based on some assemblies of closely related strains in cryptic species, most belong to our familiar types of *rep* and *par* groups in general. However, there are several contigs that haveonly one gene, which we have not

included in this study. Moreover, if there are contigs that have either *repA* or *repB* missing, they have been excluded as well.

6.2.2 Hidden markov models pipeline

Although the GS Reference Mapper searches for the best alignment to the reference genome, there is the possible limitation that the mapper might give insufficient information, especially when the reference genome would not cover potential operons that have a distant homology with other known genomes. This means that the GS Reference mapper might not detect if one of the *repABC* operons of the reference genome is divergent from currently known *repABC* operons.

Therefore, we have also made a pipeline using our own Hidden Markov Models (HMMs) of *repABC* operons in order to screen possible missing *repABC* replicons. This pipeline searches against all the contigs based on both amino acids (after translation) and nucleotides. 'hmmsearch' (default parameters), which is one type of software in HMMER (version 3.0) [66], generates the output indicating the alignment of matching motifs based on the HMMs. Thus, we can obtain the alignment of motifs, which the GS Reference Mapper did not detect. Through this application, we are able to screen thoroughly all the contigs.

Table 6.2 List of *repABC* replicons used when mapping reads by GS Reference Mapper

Group	Species	Accession	Plasmid	RepA	RepB	RepC
Chromid I	*Rhizobium etli* CFN 42	NC_007766.1	p42f	YP_472832.1	YP_472831.1	YP_472830.1
	Rhizobium etli CIAT 652	NC_010997.1	pC	YP_001985035.1	YP_001985034.1	YP_001985033.1
	Rhizobium leguminosarum bv. trifolii WSM1325	NC_012848.1	pR132501	YP_002973362.1	YP_002973363.1	YP_002973364.1
	Rhizobium leguminosarum bv. trifolii WSM2304	NC_011368.1	pRLG201	YP_002279381.1	YP_002279382.1	YP_002279383.1
*	*Rhizobium leguminosarum* bv. viciae 3841	NC_008378.1	pRL12	YP_764518.1	YP_764519.1	YP_764520.1
Chromid II	*Rhizobium etli* CFN 42	NC_007765.1	p42e	YP_472619.1	YP_472620.1	YP_472621.1
	Rhizobium etli CIAT 652	NC_010998.1	pA	YP_001985988.1	YP_001985989.1	YP_001985990.1
	Rhizobium leguminosarum bv. trifolii WSM1325	NC_012858.1	pR132502	YP_002985141.1	YP_002985140.1	YP_002985139.1
	Rhizobium leguminosarum bv. trifolii WSM2304	NC_011366.1	pRLG202	YP_002278287.1	YP_002278286.1	YP_002278285.1
*	*Rhizobium leguminosarum* bv. viciae 3841	NC_008384.1	pRL11	YP_771034.1	YP_771035.1	YP_771036.1
III	*Rhizobium etli* CFN 42	NC_007764.1	p42c	YP_472164.1	YP_472165.1	YP_472166.1
	Rhizobium leguminosarum bv. trifolii WSM1325	NC_012854.1	pR132505	YP_002979312.1	YP_002979311.1	YP_002979310.1
	Rhizobium leguminosarum bv. trifolii WSM2304	NC_011368.1	pRLG201	YP_002279471.1	YP_002279472.1	YP_002279473.1
*	*Rhizobium leguminosarum* bv. viciae 3841	NC_008381.1	pRL10	YP_770304.1	YP_770305.1	YP_770306.1
IV *	*Rhizobium leguminosarum* bv. viciae 3841	NC_008379.1	pRL9	YP_765299.1	YP_765300.1	YP_765301.1
	Rhizobium etli CFN 42	NC_007763.1	p42b	YP_471932.1	YP_471933.1	YP_471934.1
V	*Rhizobium leguminosarum*	NC_008383.1	pRL8	YP_770902.1	YP_770903.1	YP_770904.1

*	bv. viciae 3841					
VI	*Rhizobium etli* CFN 42	NC_004041.2	symbiotic plasmid p42d	NP_660042.2	NP_660041.1	NP_660040.1
*	*Rhizobium etli* CIAT 652	NC_010996.1	pB	YP_001984631.1	YP_001984632.1	YP_001984633.1
VII *	*Rhizobium leguminosarum* bv. viciae 3841	NC_008382.1	pRL7 (pRL7a)	YP_770746.1	YP_770747.1	YP_770748.1
	Sinorhizobium medicae WSM419	NC_009622.1	pSMED03	YP_001314978.1	YP_001314979.1	YP_001314980.1
	SInorhizobium fredii GR64		p64a			Yp_004716839
	Ochrobactrum anthropi ATCC 49188	NC_009670.1	pOANT02	YP_001373236.1	YP_001373237.1	YP_001373147.1
VIII *	*Rhizobium leguminosarum* bv. viciae 3841	NC_008382.1	pRL7 (pRL7b)	YP_770825.1	YP_770826.1	YP_770827.1
	Agrobacterium vitis S4	NC_011982.1	pTiS4	YP_002539981.1	YP_002539982.1	YP_002539983.1
IX *	*Sinorhizobium meliloti* SM11	NC_010865.1	pSmeSM11b	YP_001965500.1	YP_001965501.1	YP_001965502.1
X *	*Rhizobium etli* CFN 42	NC_007763.1	p42b	YP_471932.1	YP_471933.1	YP_471934.1
	Rhizobium leguminosarum bv. trifolii WSM2304	NC_011371.1	pRLG204	YP_002284253.1	YP_002284252.1	YP_002284251.1
XI *	*Rhizobium etli* CFN 42	NC_007762.1	p42a	YP_471629.1	YP_471630.1	YP_471631.1
XII *	*Rhizobium leguminosarum* bv. trifolii WSM1325	NC_012853.1	pR132503	YP_002978848.1	YP_002978847.1	YP_002978846.1
XIII *	*Rhizobium etli* CFN 42	NC_007762.1	p42a	YP_471769.1	YP_471770.1	YP_471771.1
	Rhizobium leguminosarum 3841		pRL7 noB			Yp_770781
XIV *	*Rhizobium leguminosarum* bv. trifolii WSM2304	NC_011370.1	pRLG203	YP_002283933.1	YP_002283934.1	YP_002283935.1
XV	*Agrobacterium radiobacter* K84	NC_011987.1	pAtK84c	YP_002546569.1	YP_002546570.1	YP_002546571.1

*							
XVI *	*Rhizobium etli* CIAT 652		NC_010997.1	pC	YP_001985607.1	YP_001985608.1	YP_001985609.1
	Rhizobium etli CFN 42		NC_007766.1	p42f	YP_473186.1	YP_473187.1	YP_473188.1

6.2.3 Phylogenetic analysis

6.2.3.1 All the *repABC* replicons of 72 *Rhizobium leguminosarum* strains

Based on the nucleotide sequences produced by the GS Reference mapper and our own in-house HMMs of *repABC* in section 6.2.2, we have aligned *repABC* sequences detected in 72 strains using MAFFT (version 6.717b) [62].The alignment extended from the start codon of *repA* to the stop codon of *repC* and included the intergenic regions. Seaview (version 4.2.4) [182] was used to produce a phylogenetic tree based on a Neighbour-joining method, with 1000 bootstrapping replicates performed. All the resulting trees were displayed using Treeview (version 1.6.5) [64].

6.2.3.2 Magnified version of phylogenies within each plasmid type

For the magnified version of phylogenetic trees, the same process was carried out with all the *repABC* replicons. Moreover, a maximum likelihood (ML) phylogenetic tree was also produced. The GTR substitution matrix was used for calculating the nucleotide substitutions. A discrete-gamma distribution with four categories was used to account for the variable substitution rates between sites. The gamma distribution parameter was estimated by PHYML [63]. A BIONJ distance tree was used as the starting tree, which was refined by the maximum likelihood algorithm. The robustness of the tree was determined by bootstrapping using 100 repetitions. Treeview was again used for visualizing the trees.

6.3 Results

6.3.1 Distribution of *repABC* replicons

Based on the methods in the previous section 6.2, we have detected 314 candidate sequences of *repABC* replicons in total. It should be noted that we do not include contigs having just one or two of the genes in the operon. For example, two contigs which includedjust fragments of *repB* were excluded because we required that all three genes be full-length. Moreover, one contig having *repA* and *repC*, but not *repB*, was also excluded for the same reason. Only contigs that were long enough to include at least part of all three genes were used in this chapter.

First of all, we explored the general distribution of *repABC* replicons of the strains. All 36 TRX and 36 VSX strains commonly possess operons closely related to those of pRL12 and pRL11, the large plasmids (or chromids) of *R. leguminosarum* 3841. Note that we do not know that these are on similarly large plasmids in every strain, but know that they are rep systems similar to those of pRL12 and pRL11. Therefore we will call the plasmids pRL12-type and pRL11-type plasmids.pRL10-type plasmids exist in most strains, except in 4, 8, 9, 11, 22, 24, 29 of TRX and 29, 35 of VSX. pR130503-type plasmids were also found in all strains except in 3, 12, 18, 31, 34 of TRX and 1, 3, 5, 15, 18 of VSX. A relatively small number of pRL9-, pRL8-, pRL7a-, pRL7b- and pRL1-type operons were detected. In particular only TRX18 has a sequence to link the whole A, B, and C genes of a pRL8-type plasmid, although the mapped sequences are still shorter than *repABC* of pRL8because the contig is incomplete. ThirteenpRL9-, 8 pRL7a-, 7 pRL7b and 16 pRL1-type plasmids were found. The distribution of each type of *repABC* operon is listed in **Table 6.3** and shown in **Figure 6.3**. **Table 6.4** lists each of the 72 strains, indicating which plasmid type has been detected.

As mentioned above, the core genome phylogeny had categorised the 72 strains into 5 cryptic species. When mapping the list presented in **Table 6.3**, it is very interesting to note that the distribution of some types of plasmids is biased towards specific cryptic species. For example, pRL7a-, pRL7b-, pRL8- and pRL9-type plasmids are mostly distributed in the cryptic species B (**Figure 6.3**). These plasmid replication systems are presumably a characteristic component of

genomes in the cryptic species B, which includes the reference strain 3841 that is the source of the original plasmids that define all these types. Unlike 3841, though, the majority of strains in cryptic species B also have a pRL1-type system.

The cryptic species C is different from B. The cryptic species C is composed of 52 strains and is the dominant population in the 72 strains. With the exception of the VSX11 strain that contains pRL7a- and pRL7b-type plasmids, no strains in this group have pRL9-, pRL8-, pRL7a- and pRL7b-type plasmids, while most strains have pRL12- pRL11-, pRL10- and pR132503-type plasmids only, except for VSX1 and 5, which also have pRL1-type. Strains VSX1, 3 and 5 do not contain pR132503-type plasmids, while TRX24 lacks pRL10-type.

Although the cryptic species D and E include a small number of the strains, there is a common rule observed that they have pRL12-, pRL11- and pR132503-type plasmids, but lack pRL10-type. The majority also has pRL1-type, while the TRX11 strain has a pRL7a-type plasmid as well. The sole representative of cryptic species A, TRX34, has pRL12-like, pRL11-like, pRL10-like and pRL9-like plasmids.

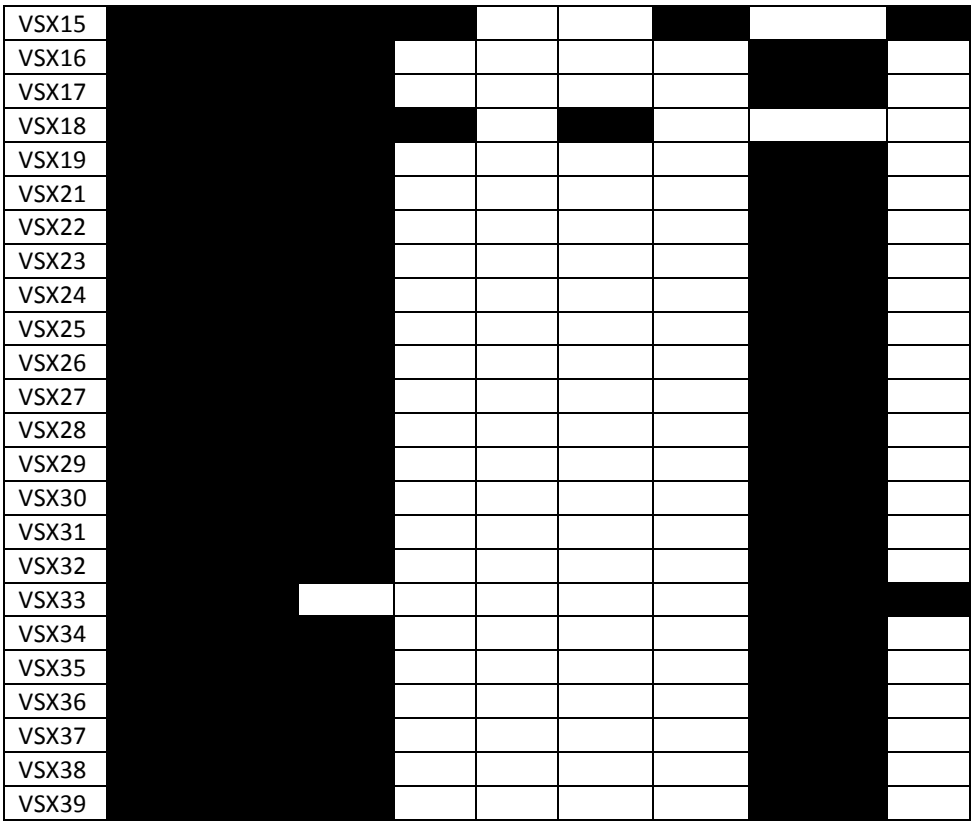
Table 6.3 List of *repABC* replicons in 72 *Rhizobium leguminosarum* strains

Type	Biovar trifolii (TRX)	Biovar viciae (VSX)	TRX/VSX
pRL12	All strains	All strains	36/36
pRL11	All strains	All strains	36/36
pRL10	All except 4, 8, 9, 11, 22, 24, 29	All except 33	29/35
pRL9	2, 12, 13, 15, 18, 25, 27, 31, 32, 33, 34	15, 18	11/2
pRL8	18	None	1/0
pRL7a	2, 11, 13, 15, 25, 33	11, 18	6/2
pRL7b	2, 13, 15, 25, 33	11, 15	5/2
pR132503	All except 3, 12, 18, 31, 34	All except 1, 3, 5, 15, 18	31/31
pRL1	2, 9, 11, 13, 15, 22, 25, 27, 29, 31, 32, 33	1, 5, 15, 33	12/4
Total	167	147	314

Table 6.4 List of 72 *Rhizobium leguminosarum* strains and their type of replicons detected

Present in black and absent in white.

Strains	pRL12	pRL11	pRL10	pRL9	pRL8	pRL7a	pRL7b	pR132503	pRL1
TRX1	■	■	■					■	
TRX2	■	■	■	■		■	■	■	■
TRX3	■	■	■						
TRX4	■	■						■	
TRX5	■	■	■					■	
TRX6	■	■	■					■	
TRX7	■	■	■					■	
TRX8	■	■						■	
TRX9	■	■						■	■
TRX10	■	■	■					■	
TRX11	■	■				■		■	■
TRX12	■	■	■	■					
TRX13	■	■	■	■		■	■	■	■
TRX14	■	■	■					■	
TRX15	■	■	■	■		■	■	■	■
TRX16	■	■	■					■	
TRX17	■	■	■					■	
TRX18	■	■	■	■	■				
TRX19	■	■	■					■	
TRX20	■	■	■					■	
TRX21	■	■	■					■	
TRX22	■	■						■	■
TRX23	■	■	■					■	
TRX24	■	■						■	
TRX25	■	■	■	■		■	■	■	■
TRX26	■	■	■					■	
TRX27	■	■	■	■				■	■
TRX28	■	■	■					■	
TRX29	■	■						■	■
TRX30	■	■	■					■	
TRX31	■	■	■	■					■
TRX32	■	■	■	■				■	■
TRX33	■	■	■	■		■	■	■	■
TRX34	■	■	■	■					
TRX35	■	■	■					■	
TRX36	■	■	■					■	
VSX1	■	■	■						■
VSX2	■	■	■					■	
VSX3	■	■	■						
VSX4	■	■	■					■	
VSX5	■	■	■						■
VSX6	■	■	■					■	
VSX7	■	■	■					■	
VSX8	■	■	■					■	
VSX9	■	■	■					■	
VSX10	■	■	■					■	
VSX11	■	■	■			■	■	■	
VSX14	■	■	■					■	


VSX15
VSX16
VSX17
VSX18
VSX19
VSX21
VSX22
VSX23
VSX24
VSX25
VSX26
VSX27
VSX28
VSX29
VSX30
VSX31
VSX32
VSX33
VSX34
VSX35
VSX36
VSX37
VSX38
VSX39

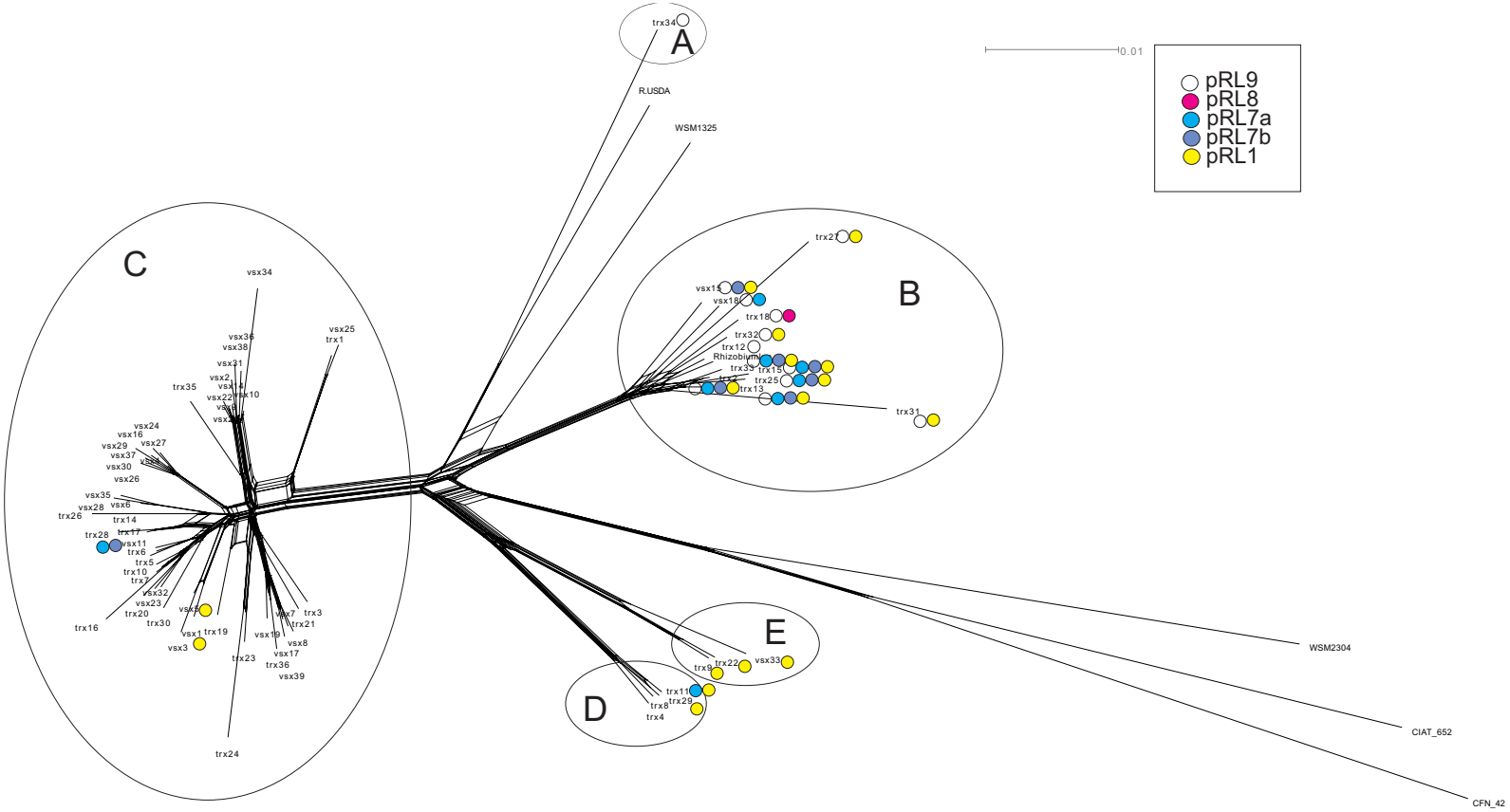


Figure 6.3 Mapping different types of replication and partitioning regions from 72 strains on the core genome tree

Nucleotide sequences of core genes of 72 *R. leguminosarum* strains and other related genomes (*R. leguminosarum* 3841, *R. leguminosarum*WSM 2304, *R. etli*CIAT 652 and *R. etli* CFN 42) were concatenated and used to construct the tree. Five different cryptic species are indicated as A to E. Replication systems (pRL9, pRL8, pRL7a, pRL7b, and pRL1) are mapped on the tree as different coloured circles (white, pink, light blue, blue, and yellow, respectively). (Nitin Kumar, personal communication).

6.3.2 Phylogeny based on the alignment of all the *repABC* operons

We have constructed a phylogenetic tree based on all the *repABC* replicons in 72 *R. leguminosarum* strains, in order to provide a means of classification into discrete types (**Figure 6.4**). The 322 sequences of replication systems that we found (297 sequences in 72 strains and 8 reference sequences of *R. leguminosarum* bv. viciae 3841 and *R. leguminosarum* bv. trifoliiWSM1325, and pRL1) were split into nine groups according to the phylogeny: pRL12-, pRL11-, pRL10-, pRL8-, pRL9-, pRL7a-, pRL7b-, pR132503- and pRL1-type. Most clades are clearly distinguishable.

Observing the clades closely in **Figure 6.4**, pRL12- and pRL11-type replication regions occurring in all 72 stains in particular exhibit very clear clades with little divergence. It appears that they are more conserved than other replication regions. On the other hand, pR132503- and pRL10-types show that their variation is possibly larger than pRL12- and pRL11-type plasmids, although they also exhibit distinctive clades. A relatively small number of pRL9-, pRL8-, pRL7a-, pRL7b- and pRL1-type replicons also demonstrate resolved clades in the phylogenetic tree. Although not many strains have pRL9-type plasmids, the clade is close to (though distinct from) that of the pRL12-type plasmids. This is because the replication initiator gene *repC* of pRL9 is nearly the same as that of pRL12, while the coupled partitioning genes *repA* and *repB* (normally called *parA* and *parB*) are phylogenetically distant, as already discussed in Chapter 3. The partitioning genes of plasmids seem to be enough to overcome incompatibility. We will look into individual phylogenies of each type of replication regions in the next section 6.3.3.

In the case of pRL7a- and pRL7b-type plasmids, two distinct clades have been shown in the phylogenetic tree. Young et al. [2] suggested that pRL7 of *R. leguminosarum* 3841 has ‘extra’ *repABC* replicons. Among the 72 strains, TRX02, TRX13, TRX15, TRX25, TRX33 and VSX11 seem to have both pRL7a- and pRL7b-type replicons, which follow the same pattern as the *R. leguminosarum* 3841 plasmids. It is not clear which replication regions among those are actually functioning. Further research is required in order to clarify this.

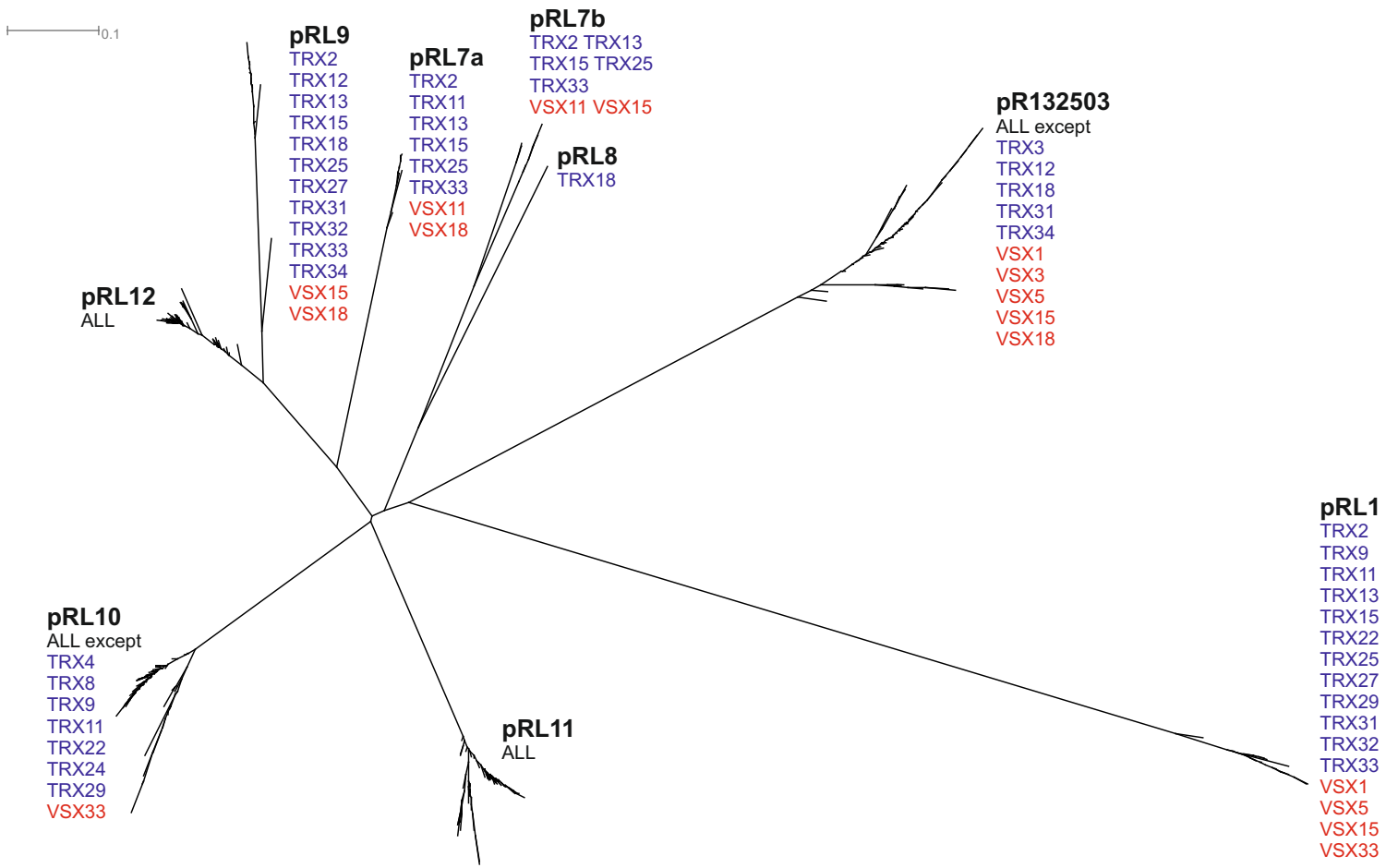


Figure 6.4 The phylogenetic tree of *repABC* replicons in 72 *Rhizobium leguminosarum* strains and 8 replicons from reference genomes. Concatenated nucleotide sequences of Rep and Par regions of 72 strains and *R. leguminosarum* 3841 were used to construct this tree.

6.3.3 The phylogenetic trees of each plasmid type in 72 *Rhizobium leguminosarum* strains

In order to acquire more details on each type of *repABC* replicons in the 72 *R. leguminosarum* strains, each alignment of pRL12-, pRL11-, pRL10-, pRL9- and pR132505-type was reconstructed. The alignments were used to build Maximum Likelihood (ML) phylogenetic trees (**Figures 6.5-10**). Only bootstrapping numbers of over 70% were indicated. These trees are important because different patterns of phylogenetic trees can be detected if there is any horizontal gene transfer between species, compared to the phylogeny of the core genome (**Figure 6.2**) and that of each plasmid type. Therefore, this analysis would shed light on the amount of strain-to-strain spread of the plasmids. Note that the following trees are presented in the order of the relatively large plasmids of *R. leguminosarum* reference genomes (pRL12, pRL11, pRL10, pR132503) firstly, and then other smaller plasmids (pRL9, pRL7, pRL1).

6.3.3.1 Phylogenetic analysis of two large plasmids

First of all, we have constructed the phylogenetic trees for *repABC* regions of pRL12- and pRL11-type plasmids (**Figures 6.5 and 6.6**). These two replicons are known as 'chromids' and possess not only core genes commonly found in chromosomes, but also the plasmid replication system. Comparing the two phylogenies to the core gene tree (**Figure 6.2**) respectively, the chromid phylogenies mirror that of the core genome. All strains of TRX and VSX have pRL12- and pRL11-type plasmids (36 each) and these replication sequences are clearly divided into five groups, which are the same as those of the core genome phylogeny. Similar to the core genome tree, one big clade exists on the left side (the cryptic species C) and three clear groups (the cryptic species B, D and E) show well-resolved clusters on the right side. The TRX34 strain made up one separate group (the cryptic species A). The fact that the core genome and the plasmid trees are congruent indicates that there is very limited movement, at least between species, in the case of chromids. The corresponding history of the core genomes and the chromid trees means, therefore, that the replicons have not moved between the cryptic species as shown in the phylogeny.

6.3.3.2 Phylogenetic analysis for other plasmids

Similarly, the phylogenetic tree based on all replication sequences of pR132503-type plasmids (**Figure 6.8**) and of pRL9-type plasmids [**Figure 6.9 (a)**] is clearly divided into five groups and two groups, respectively, which are the same as those of the core genome phylogeny. The replicons, therefore, also have not moved between the cryptic species frequently. Note that the cryptic species C, D do not contain pRL9-types of plasmids.

On the other hand, the phylogenetic tree of pRL10-type plasmids has yielded different observations. **Figure 6.7** contains the phylogenetic tree of pRL10-type plasmids. Note that the cryptic species D and E are not present in this tree because both D and E do not contain these types of plasmids. 64 sequences were used to construct this tree, as TRX4, TRX8, TRX9, TRX11, TRX22, TRX24, TRX29, VSX15 and VSX18 do not have pRL10-type plasmids. Most sequences were grouped into clearly distinct clades, including the cryptic species A, B and C. As shown in the phylogeny, there are two separate B clades. A possible explanation might be the replacement of this replicator by a related but diverged version in the ancestor of one of the clades, though this would have to have been a long time ago, since the clades are both diverse. The grouping pattern of the tree is the same with that of other plasmids, such as pRL12- and pRL11-types. There is one exception, however, which is the strain VSX15 in the cryptic species C.

This strain contains five replication systems in total, which have pRL12-, pRL11-, pRL10-, pRL9- and pRL7a-type respectively. All systems with the exception of the pRL10-type, are grouped into the cryptic species B, just like in the core genome phylogeny, with only the pRL10-type replication region of VSX15 belonging to the cryptic species C. In the pRL10 phylogeny, VSX15 is clustered with TRX26 and the bootstrapping value is also high (99%), which seems reliable. This is interesting because it is a clear example of a plasmid moving between the cryptic species. pRL10 is known to confer symbiosis genes, such as *nod, nif,* and *fix*[2]. However, there is no evidence that symbiosis genes are carried on the pRL10-like plasmid in any of the Wentworth strains, and in case of TRX06 (which has thehighest coverage among all the strains), they are on the p132503-type plasmid.

The phylogenetic tree of pRL1 (**Figure 6.10**) is also very interesting. The plasmid tree clearly does not mirror that of the core genome tree. For example, VSX15 in

the cryptic species B shares its origin with VSX5 in the cryptic species C with a bootstrap of 85%. Moreover, although the strains VSX1 and VSX5 belong to C, they are located in a separated clade in the plasmid tree. Therefore, the movement of pRL1 is rather freer than that of other large plasmids mentioned above.

6.3.3.3 Investigation into the movement within cryptic species

Based on the grouping pattern of 5 different plasmids, we have seen that there is very limited movement between species. However, the inner topology of each clade is quite different between the core genome and the plasmid trees. Although many bootstrapping values are low, resulting in some of the clades not being resolved well, the phylogenetic tree shows in many cases the clearly distinct topologies. For example, the replication regions of VSX31 and VSX35 are closely associated (bootstrapping value is 89%) in the pRL12-type plasmid tree (**Figure 6.5**), but they are located in a different clade in the core genome tree. The plasmids of VSX15, TRX25 and TRX13 within group B also exhibit a different pattern with the respective group in the core genome tree. In addition, in the case of pRL11, VSX17 and TRX17, they compose a clade with a high bootstrapping value (97%), but they are located at a distance in the core genome tree. Other phylogenies for different types of replicons (**Figure 6.6-10**) also indicate that the inner topology is not congruent with the core genome tree. Although further research would be required, the incongruence of inner topology might imply the possibility of free movement inside the species groups. Previous research related to this will be reviewed in the next section 6.4.

6.3.3.4 Concerted evolution of *repC* in two plasmid types

In order to gain more insight on the evolutionary history of pRL12- and pRL9-type plasmids of 72 *R. leguminosarum* strains, we have aligned both the sequences of *repAB* and the sequences of *repC* for pRL12- and pRL9-type plasmids. The alignments were used to make separate phylogenetic trees (**Figure 6.11**). (A) is a phylogenetic tree of *repC* and (B) is a phylogenetic tree of *repA* and *repB* (the alignment is based on concatenation of the two partitioning genes). The two trees clearly show a different pattern; each *repC* of the strains having both pRL12- and

pRL9-type plasmids is tightly linked within a clade (all bootstrapping values are above 90%), but *repAB* of pRL12- and pRL9-type plasmids is located in a separated clade. More details are in section 6.4.2.5.

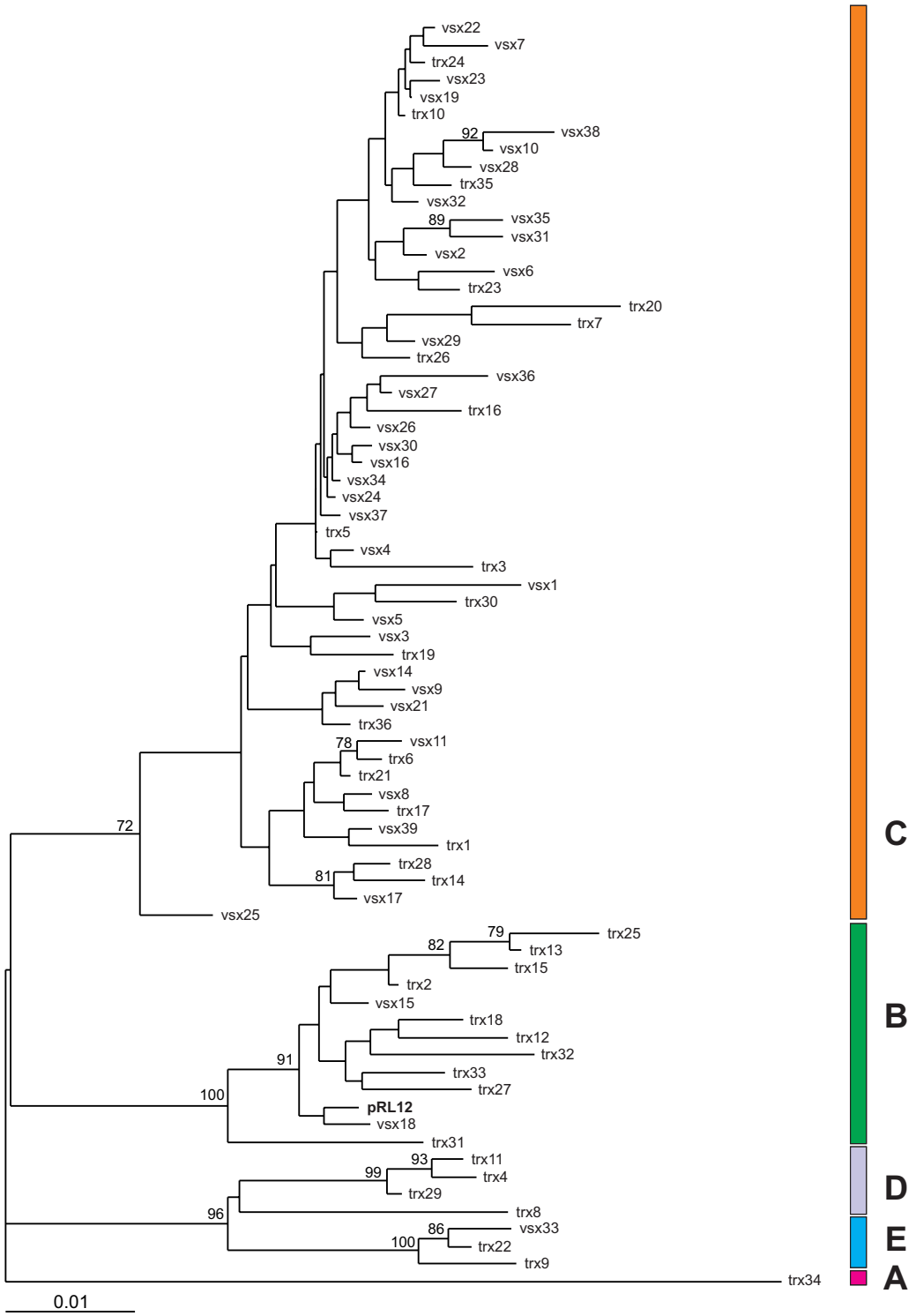


Figure 6.5 The phylogenetic tree for *repABC* regions of pRL12-type plasmids of 72 *Rhizobium leguminosarum* strains

The tree was constructed by the Maximum Likelihood (ML) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pRL12 (bold) is from *R. leguminosarum* 3841. Coloured bars indicate sequences from cryptic species A-E.

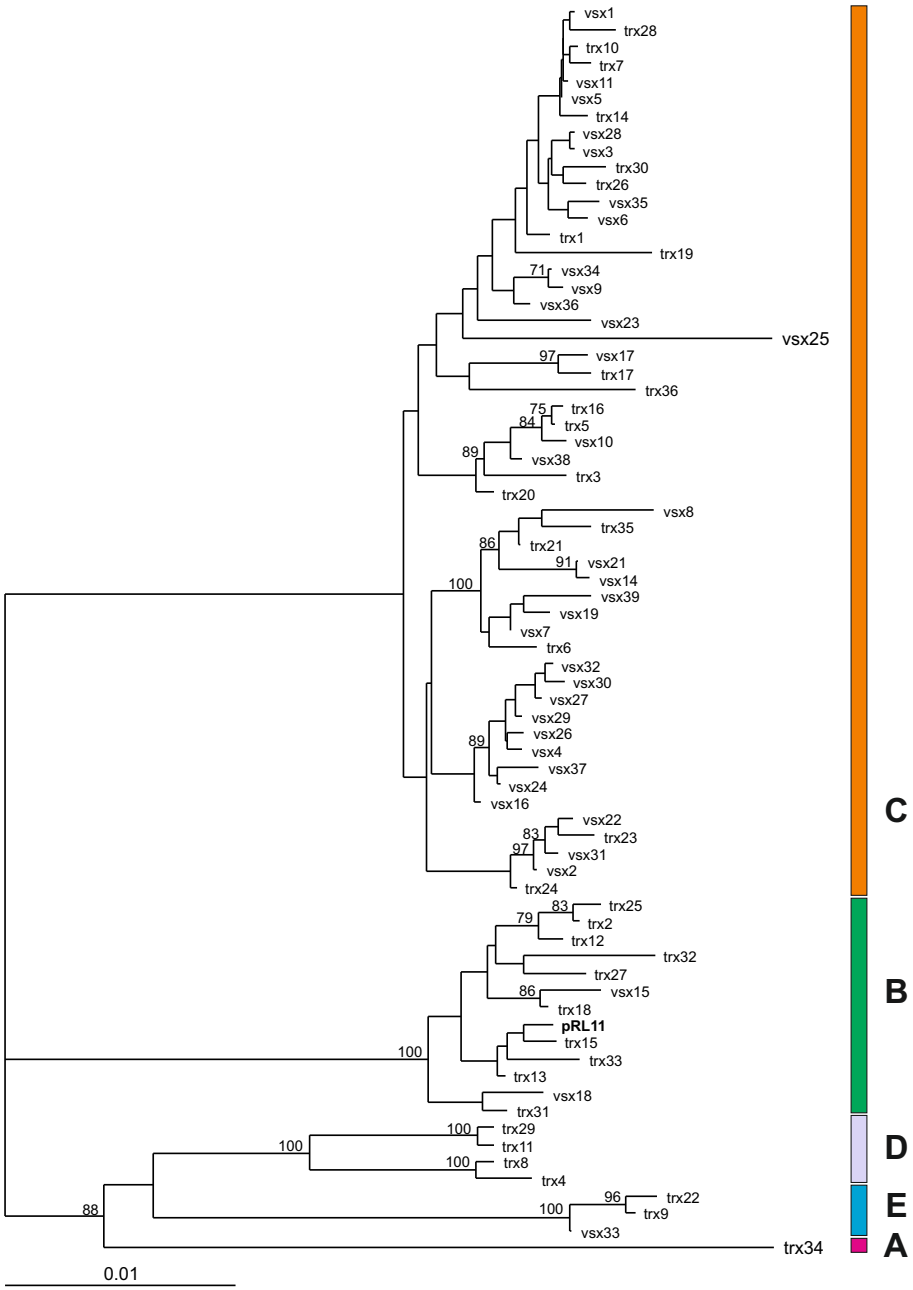


Figure 6.6 The phylogenetic trees for *repABC* regions of pRL11-type plasmids of 72 *Rhizobium leguminosarum* strains

The tree was constructed by the Maximum Likelihood (ML) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pRL11 (bold) is from *R. leguminosarum* 3841. Coloured bars indicate sequences from cryptic species A-E.

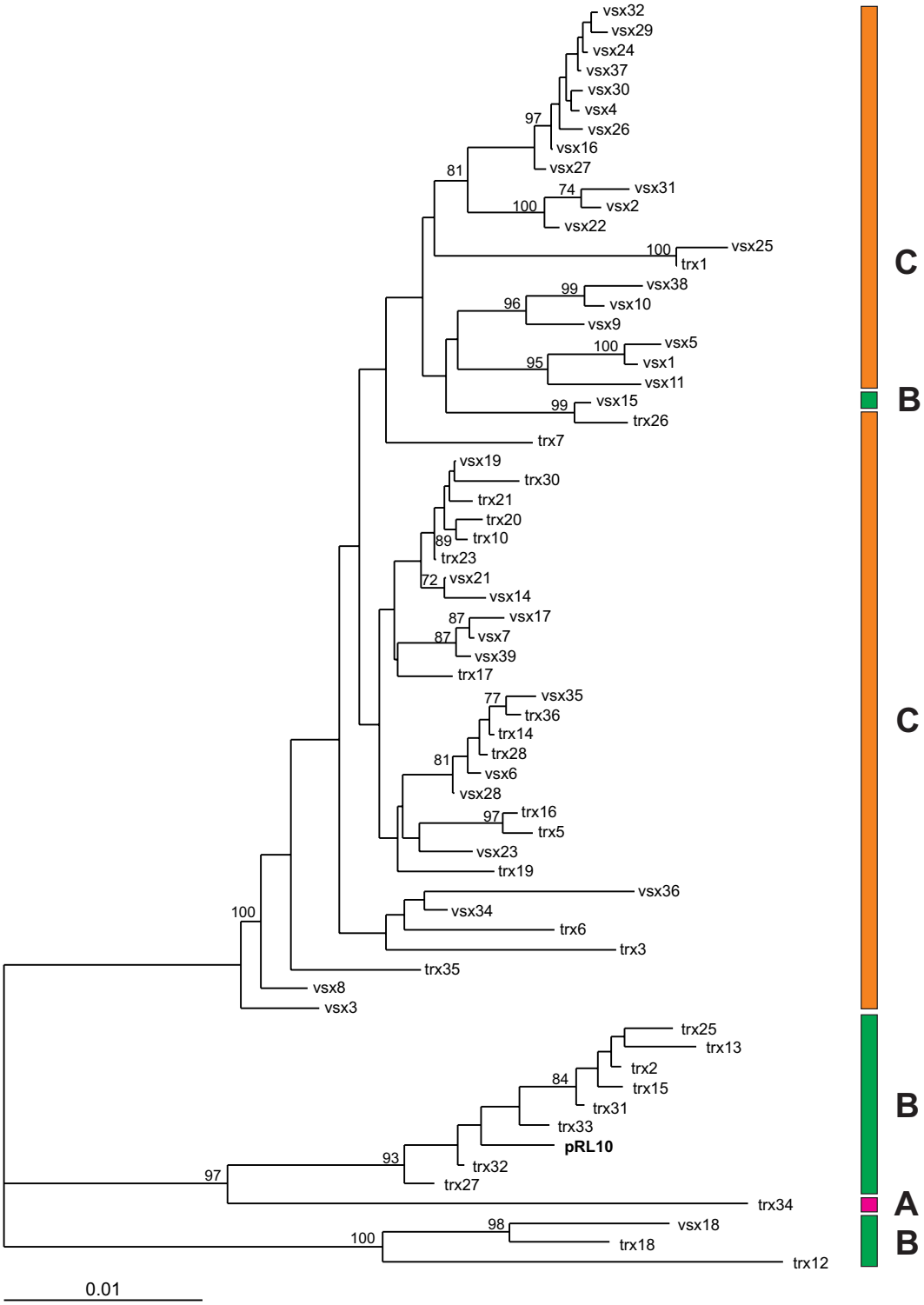


Figure 6.7 The phylogenetic trees for *repABC* regions of pRL10-type plasmids of 72 *Rhizobium leguminosarum* strains

The tree was constructed by the Maximum Likelihood (ML) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pRL10 (bold) is from *R. leguminosarum* 3841. Coloured bars indicate sequences from cryptic species A-C.

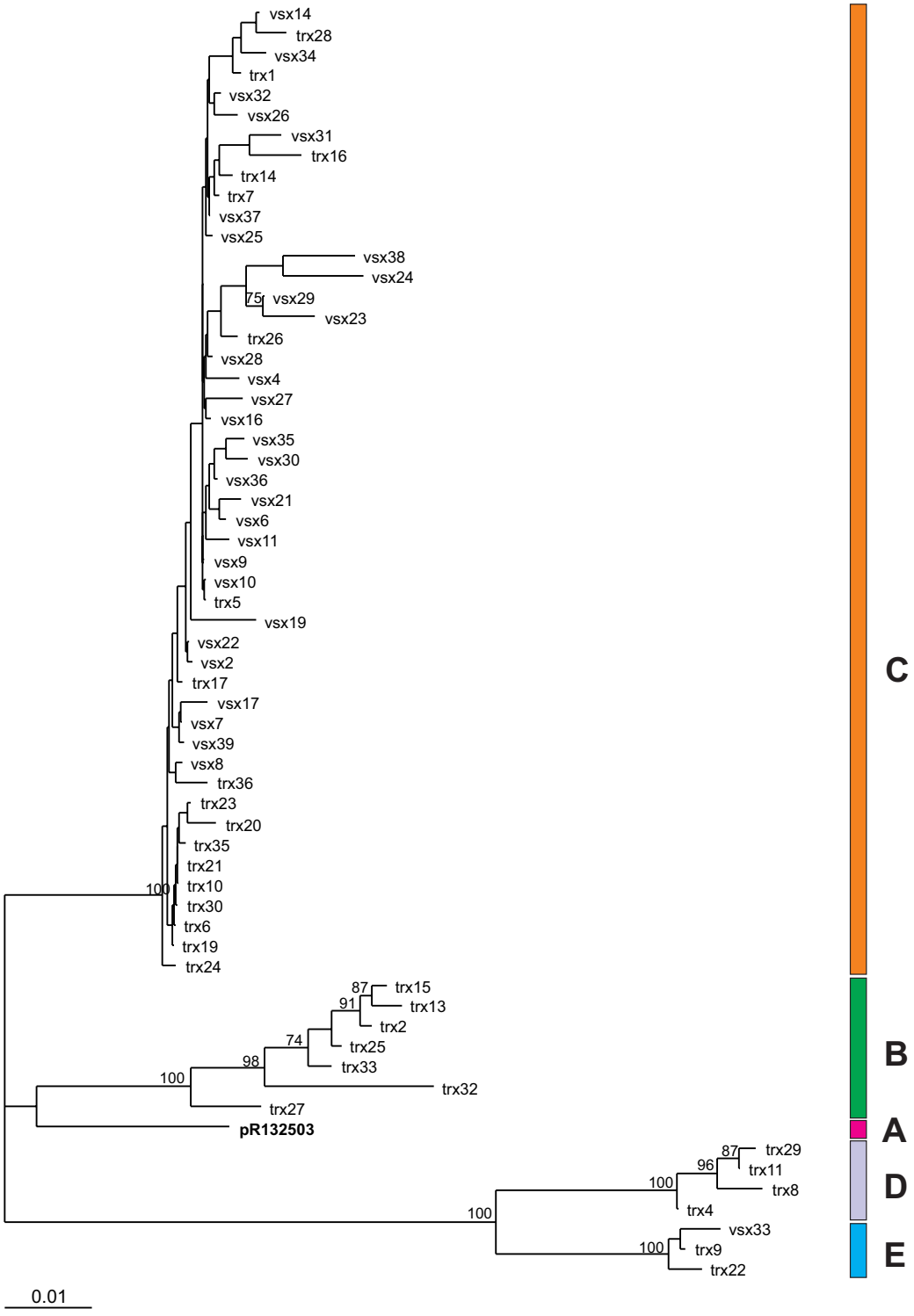


Figure 6.8 The phylogenetic trees for *repABC* regions of pR132503-type plasmids of 72 *Rhizobium leguminosarum* strains

The tree was constructed by the Maximum Likelihood (ML) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pR132503 (bold) is from *R. leguminosarum* bv. trifolii WSM1325. Coloured bars indicate sequences from cryptic species A-E.

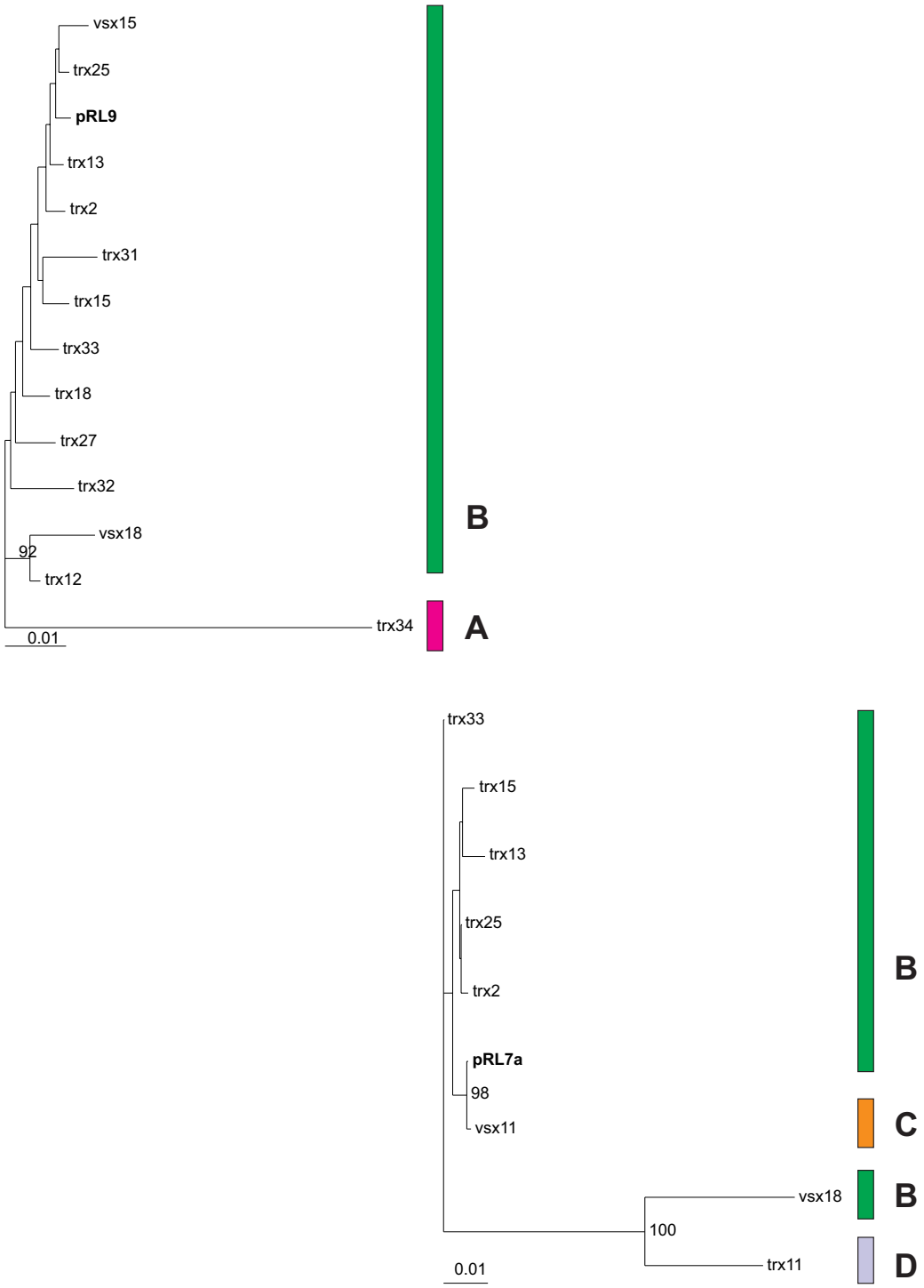


Figure 6.9 The phylogenetic trees for *repABC* regions of (a) pRL9-type plasmids and (b) pRL7a-type plasmids of 72 *Rhizobium leguminosarum* strains

The tree was constructed by the Maximum Likelihood (ML) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pRL9 and pRL7a (bold) is from *R. leguminosarum* 3841. Coloured bars indicate sequences from cryptic species A-D.

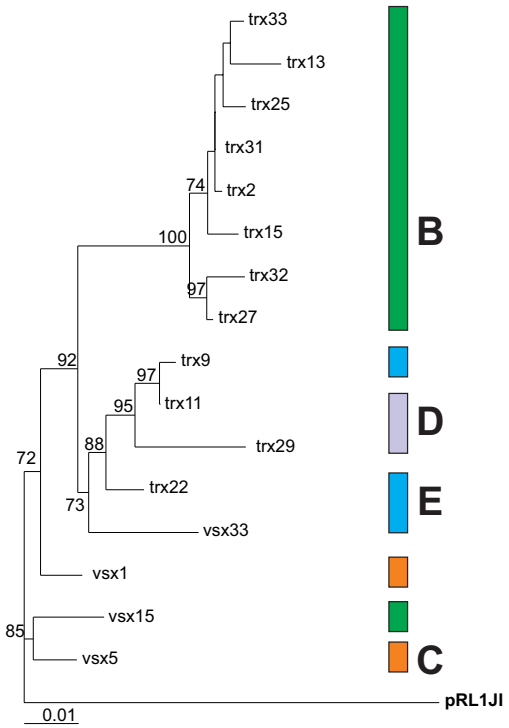


Figure 6.10 The phylogenetic trees for *repABC* regions of pRL1-type plasmids of 72 *Rhizobium leguminosarum* strains

The tree was constructed by the Maximum Likelihood (ML) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pRL1JI (bold) is from *R. leguminosarum*. Coloured bars indicate sequences from cryptic species B, D, and E.

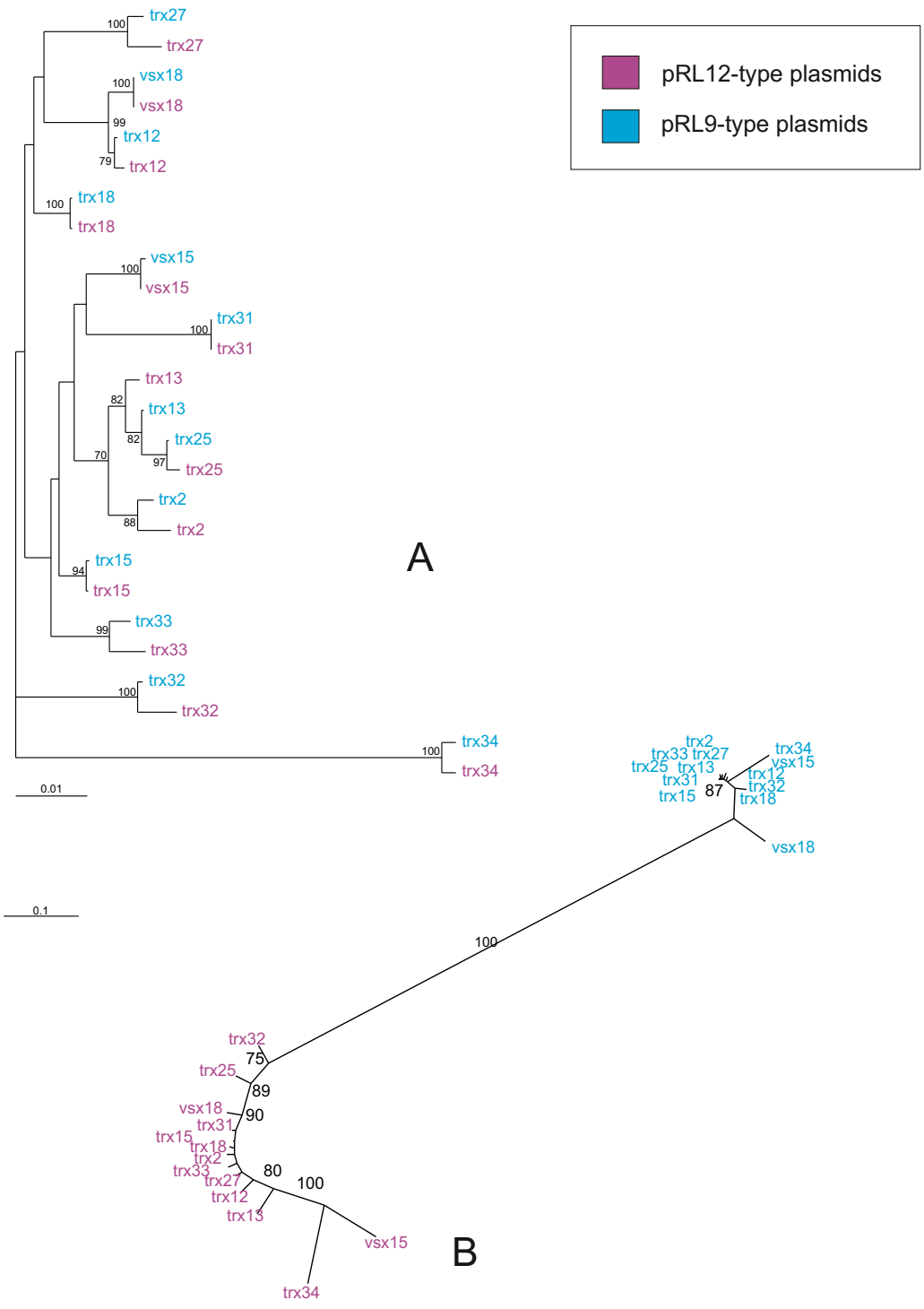


Figure 6.11 Comparison of the *repC* phylogeny (A) with the *repAB* phylogeny (B)

The tree was constructed by the Neighbour-Joining (NJ) method based on concatenated nucleotide sequences. Bootstrap percentages over 70% are indicated. Plasmid pRL12 (bold) is from *R. leguminosarum* 3841. pRL12-type plasmids are shown in purple and pRL9-type in blue.

6.4 Discussion

6.4.1 Summary

In this chapter, we have analyzed the *repABC* operons distributed across 72 *R. leguminosarum* strains. Nine phylogenetically distinct groups of *repABC* replicons were found among the strains. *repABC* regions of large plasmids, such as pRL12- and pRL11-type replicons show less variation, while those of smaller plasmids show more divergence. There are two types of *repABC* operons that are not present in the reference genome *R. leguminosarum* 3841; one is the pR132503-type, close to that of the plasmid pR132503 in *R. leguminosarum* WSM1325, and the other is the pRL1-type. Both demonstrate phylogenetically distant relationships with the plasmids of *R. leguminosarum* 3841 (**Figure 6.4**). No clear movement of two large replicons (pRL12- and pRL11-type plasmids, called chromids) was observed based on their congruence with the core genome tree. However, possible movement between the species groups in the phylogenetic tree of pRL10- and pRL1-type plasmids was detected, with the addition of a variety of possible cases between strains. Replicon sequences might move between strains by conjugation of entire plasmid sequences or the transfer of partial sequences containing functional or non-functional replicon regions.

6.4.2 Are plasmids active vehicles in Rhizobia?

6.4.2.1 Are large plasmids not able to move freely?

Plasmids are one of the representatives of mobile genetic elements (MGEs). In particular, broad host range plasmids, such as IncP, IncW and IncQ, are known to be able to replicate and transfer across bacterial cells freely. Therefore, it was considered a possibility that some of the plasmids in rhizobia would also be able to move frequently. However, based on the phylogenetic tree of the core genome and the 'magnified' version trees of each type, no clear movement between cryptic species was observed for large plasmids in *R. leguminosarum* strains.

Harrison et al. [84] have examined chromid replication systems. In their paper, they have argued that all chromids within a genus have closely related replication

and partitioning genes. Our data are consistent with this. As we have seen here, chromids' replication and partitioning systems are conserved with a long history and, moreover, are not acting like vehicles between species.

6.4.2.2 Do other plasmids show free movement?

Despite the absence of recognizable movement in large plasmids, other plasmids seem more flexible in acting as vehicles. One of the pRL10-type repliconshas been shown to be transferred from the cryptic species B to C based on the incongruence of the core genome tree (**Figure 6.2**) and the replicon tree (**Figure 6.7**). Although this is the only actual movement observed for pRL10-type plasmids between different species groups in the 72 strains, this is arguably an important example of transfer. In addition, we have seen clear incongruent phylogenies between pRL1-type plasmids and the core genome. Therefore, we can assume that some small plasmids might have the capacity to be transferred.

Surprisingly, there were not many observed cases of the movement of small plasmids such as pRL7- or pRL8-type plasmids. One explanation could be that in rhizobia, other MGEs, such as transposons or genomic islands, might be responsible for transfer between the species groups, rather than plasmids.

6.4.2.3 Are plasmids transferred from other groups?

There were some cases observed that exhibited interesting patterns. In **Figure 6.3**, pRL9-type plasmids are dominant in the cryptic species B. TRX34 in the cryptic species A, however, also possesses a pRL9-type plasmid. In order to establish whether there is a possibility that this type is transferred from B, we have constructed a phylogeny based on the sequences of pRL9-type plasmids [**Figure 6.9 (A)**]. It was shown that the *repABC* operons of TRX34 are distinct from those of other strains in the cryptic species B. Therefore, it could not have been transferred recently from B.

Conversely, it is interesting that VSX11 in the cryptic species C seems to have acquired the pRL7a-like replicon very recently from the cryptic species B, as VSX11 is located in the upper clade, closely related to plasmids in the cryptic species B in the phylogeny [**Figure 6.9 (B)**]. Moreover, VSX18 in species B is located in the same clade as TRX11 in species D, with 100% bootstrapping value, which

might indicate thatthe pRL7a-type plasmid of VSX18 does not have the same origin as other pRL7a plasmids in species B.

6.4.2.4 Is movement between strains possible?

In comparison to the movement between species, relatively more cases of transfer from strain to strain have been found. From large plasmids, such as pRL12- and pRL11-type, to small ones like pRL7-type, a variety of plasmids have been observed with the ability to move from other strains based on the incongruent topology of specific clades in the phylogenetic trees.

Although some examples of movement between strains based on the phylogenetic trees (**Figures 6.5-10**) have been identified, it is true that some clades of those phylogenetic trees are not very well resolved. This is because the nucleotide sequences of all the *repABC* operons are apparently similar or nearly so with each other. Therefore, it is difficult to get a clearer idea of recombination events within species. Nevertheless, the analysis of the movement between strains is still significant since it might be used as a vehicle for specific accessory genes.

6.4.2.5 Duplication and recombination of *repABC* operons?

In chapter 3 and section 6.3.2 in this chapter, we have seen that pRL12 and pRL9 plasmids of *R. leguminosarum* 3841 exhibit a unique feature. In the phylogenetic tree of *repC* (**Figure 3.4** in Chapter 3), usually only one plasmid of each strain is located in each clade, which might reflect actual incompatibility groups. The *repC* sequence of pR12 and pRL9 was one of the exceptions in the phylogenetic tree.

We have aligned both the sequences of *repAB* and the sequences of *repC* for pRL12- and pRL9-type plasmids, in order to gain more insight on the evolutionary history of pRL12- and pRL9-type plasmids of 72 *R. leguminosarum* strains (**Figure 6.11**). The two trees present a different pattern; each *repC* of the strains that possess both pRL12- and pRL9-type plasmids is tightly linked within a clade, but *repAB* of pRL12- and pRL9-type plasmids is located in a separated clade.

Previous research has shed some light on the evolutionary history of *repABC* operons and it is clear that *repA, repB* and *repC* do not have the same background [96]. When Slater et al. compared the *repA, repB* and *repC* phylogenies for their *Agrobacterium* genomes, they found that the phylogeny of

repC lacks congruence with the trees *repA* and *repB,* while *repA* and *repB* trees had consistent phylogenies with each other. This suggests that, in general, there are various duplication and recombination events occurring between *repAB* and *repC* genes. We have also provided clear evidence that the partitioning coupled genes *repAB* show different phylogenetic patterns than replication gene *repC*. As indicated in **Figure 6.11 (A)**, *repC*s of pRL12- and pRL9-type plasmids are highly similar, while *repAB*s of two replicons are not (**Figure 6.11 (B)**). Therefore, we propose that in the ancestor of the pRL9-like plasmids, the original *repC* was replaced with a copy of the pRL12-like *repC* by homologous recombination. After this, the Rep systems in different lineages started to diverge. Although they diverged from their orthologs in other strains, the *repC* genes in the pRL9-like and pRL12-like plasmids within a single cell remained similar as a result of repeated gene conversion that "corrected" one against the other, transferring the same mutations to both of them. This is the same mechanism that keeps the different ribosomal RNA genes identical (or nearly so) within a genome, while allowing species to diverge . Gene conversion is necessary to explain why the pRL9-like and pRL12-like *repC* genes in a strain are always each other's closest relatives. This also brings the idea that the partitioning systems are more important than the replication system for plasmid incompatibility in *repABC*-type plasmids.

Although the concerted evolution of *repC* in pRL9 and pRL12 might seem like an oddity, our research shows that this phenomenon has actually arisen multiple times in *repABC* plasmids. Similar pairs of almost identical RepC sequences are found in pR132501 and pR132504 of *R. leguminosarum* WSM1325, pMLa and pMLb of *Mesorhizobium loti* MAFF 303099, pl2 and pl3 of *Mesorhizobium sp.* BNC1, pOANT01 and pOANT03 of *Ochrobactrum anthropi* ATCC 49188, and pATS4a and pATS4b of *Agrobacterium vitis* **(Figure 3.5)**. With the exception of the sequences in *R. leguminosarum* WSM1325, which are very similar to those in 3841, each of these pairs is in a very different part of the RepC phylogeny, implying multiple independent origins.

6.5 Conclusion

We have investigated *repABC* operons of 72 *R. leguminosarum* strains isolated from Wentworth College in 2007. In total, 314 *repABC* replicons were found and these are divided into 8 groups phylogenetically. There were two types that did not belong to the plasmids of *R. leguminosarum* 3841. Overall, the phylogenetic trees of

each type of *repABC* replicon mirror the core genome tree of the species and this is particularly true for the large plasmids, pRL12- and pRL11-type, indicating no movement for them between species. Some examples, however, were found for the possible horizontal transfer of other plasmid types between species. On the other hand, there are many cases of movement between strains within species, which might be used as a vehicle for specific accessory genes.

Chapter 7. Conclusion and perspectives: where are we now and where are we going?

The main purpose of this thesis was to analyze plasmid replication and partitioning systems in proteobacteria in a comprehensive way, based on the publicly available genomes. In chapter 2, we developed a database, which collected the families that are involved in the two systems. In chapters 3 and 4, we investigated the distribution and host range of each family and performed their phylogenetic analysis. In chapter 5, we discussed the patterns of plasmid diversity in general, based on the relationships betweentwo systems. As a case study, we analyzed the RepABC replicon, one of the best-known replication and partitioning systems in plasmids, in chapter 6, in order to study the distribution of plasmids within a small population. In this final chapter, we will firstly review the reason why this analysis is significant for plasmid biology and the current status of the research in plasmid backbone systems. We will then demonstrate the contribution and the limitations of this research. Based on the knowledge we have gained through this study, we will finally propose some future directions of analysis that would further the investigation of the evolutionary history of plasmids.

7.1 Why is this thesis significant for plasmid biology?

The main functional modules involved in plasmids can be divided into two parts, plasmid backbone systems and plasmid accessory systems [9, 173, 183]. Replication and partitioning systems are two of the most essential systems in plasmid backbones, because plasmids need to replicate by themselves using their own mechanisms and ensure their partitioning after replication, in order to ensure

that they are propagated to both daughters when the cell divides [39, 69, 145, 184]. Copy number control allows them to achieve this while minimising the burden of plasmid carriage and hence the selection against the plasmid [185]. Previous research (e.g. Mikesell et al [186], Bennett [187], etc.) has concentrated more on the study of accessory modules, rather than backbone modules, because of the many medical applications of accessory modules. There has, however, been an increase in research on the backbone systems in plasmids, since they can provide important insights in the diversity and the evolutionary history of plasmids.

This thesis analyses plasmid replication and partitioning systems in proteobacteria. The constructed database that stores the defined families of target sequences functions as a basis for biologists to research a variety of topics, from the diversity of plasmid backbone systems to that of plasmids themselves. In particular, research in terms of the distribution and host range of each family defined here is significant for the investigation of plasmid incompatibility, as well as speculation of their future host range. This is also helpful for the classification of various plasmids, and our analysis can provide a tool for finding and classifying the replication and partitioning genes for newly sequenced genomes as well.

7.2 Contribution of the thesis

7.2.1 A database for plasmid backbone systems in proteobacteria

One of the aims of this thesis was to develop a database that would function as a basis for the research of plasmid backbone systems. This involved the selection and collection of target sequences, their storage according to the families that we have defined, and the design of a web site where all the information would be available. To our knowledge, this study is the first work to contribute to the storage of plasmid backbone systems, including replication and partitioning modules and their comprehensive analysis.

In order to achieve this, target modules, namely replication initiator proteins in plasmid replication systems (normally called Rep) and partitioning coupled

proteins (called Par), were selected. Homologous genes were collected based on the result of psi-blast and our in-house HMMs and were stored in a MySQL database. All families and their members can be accessed and downloaded via a website called 'Database for plasmid backbone systems in proteobacteria' (http://bioplasmid.godohosting.com). The phylogenetic trees of each family are also available on the site. Although there is room for improving the database, such as implementing automatic updates of the database, as well as increasing the number of modules involved in the two systems, the database and the website aid the classification of plasmids, and highlight their incompatibility and the research for the diversity of the two systems in proteobacteria, which are discussed in chapters 3 and 4.

7.2.2 Distribution and host range of plasmid backbone systems in proteobacteria

7.2.2.1 Diversity of plasmid replication systems

One of the main aims of this thesis was to research the distribution and host range of plasmid backbone systems in proteobacteria. Tracking the evolutionary history of plasmids' movement based on the backbone systems, therefore, can be useful for the research of plasmids' diversity. Based on the backbone systems, it does not seem to be very common, however, for plasmids to exhibit an environmentally broad host range, especially over class or phylum level. Frequent movement of plasmids in different proteobacterial divisions, based on the analysis of Rep, has not been detected, except for broad host range IncP plasmids and IncW plasmids. BHR plasmids IncN do not show a wide range of distribution in a natural environment, even though some research conducted in a lab environment reported that they could be transferred from other hosts [79, 104, 188]. Plasmids, however, do move at least between related hosts, particularly within an order or class level, (e.g. in the case of RepCs in *Rhizobiales, Rhodobacterales*, etc.) Moreover, there are several cases where two plasmids, found in unrelated niches, are contained in the same clade in the phylogeny, which indicates a common origin.

7.2.2.2 More families across proteobacteria

8 major families of plasmid replication systems in proteobacteria were investigated in total. It is clear that, in addition to the main families, there are more families in proteobacteria that can be categorized. The number of members in one family might not be particularly large at the moment, but they are still worth investigating in more detail, as shown by the example of DnaA-like replication systems in marine bacteria that Petersen et al. [140] have characterized and categorized as a novel system. More sequencing work published in the future will reveal the new families.

7.2.2.3 Classification of plasmids based on the replication systems

We have highlighted the significance of the classification of plasmids and we have concluded that Rep initiators can be used as a reliable marker for classification, in comparison with previous criteria [45, 74, 77, 78]. In particular, phylogenetic analysis might be a possible indicator of the actual incompatibility. There are some exceptions where two replication systems in the same strain are verysimilar to one another. We have suggested, however, that the classification might be improved by considering partition systems [2]. In the case of multireplicon plasmids, however, it is not easy to know which one is the actual replication site, which makes it difficult to classify the plasmids based on the Rep regions. The clades in the phylogenetic analysis might be supporting their incompatibility groups, but further experiments should be conducted in order to back up the results based on the Rep systems.

7.2.2.4 Diversity of plasmid partitioning systems

Based on our current knowledge of plasmid partitioning systems, they are normally classified in three major categories: Walker A cytoskeletal P-loop ATPases, actin-like ATPase and tubulin-like ATPase [143, 145, 153, 160]. In proteobacteria, we studied that the majority of partitioning proteins belongs to Walker type ATPase. Although these proteins share similar domains, we suggested that they could be divided into distinct types according to their partner proteins.

Therefore, we have identified four discrete types in the Type I class and one type in the Type II class, which are most abundant in proteobacterial plasmids. Based on the results, partitioning systems are restricted to the class level. This is shown very strongly in plasmids having long ParA sequences, such as RepABC replicons from alphaproteobacteria, and several plasmids from beta- and gammaproteobacteria. Plasmids possessing short ParA, ParF and ParM type sequences also show that their distribution is restricted to the class level of proteobacteria in the phylogenies. The members of each type do not seem to move outside of the class level excessively, except in the case of broad host range plasmids, such as the IncC-KorB system of broad host range plasmids, which manifest a wide distribution across different divisions of proteobacteria.

7.2.2.5 Classification of plasmids based on their partitioning systems

Both replication and partitioning systems in bacteria are considered as an important indicator for classifying plasmids. Particularly in the case of the RepABC replicons, which are not easy to classify by Rep systems, the Par system actually does play a role in the classification of plasmids. It appears, however, that not all Par systems are suitable for plasmid classification, because of the fact that not all plasmids have Par modules, which might leave out numerous replicons in bacterial classification. Further, the presence of multiple partitioning modules in one plasmid can make it hard to classify the plasmid effectively, although the partitioning modules can be categorized.

7.2.3 Learning from the case study of RepABC replicons in the same species

We have investigated one type of the well known replication and partitioning systems in proteobacteria in chapter 6, the RepABC-type replicon. The RepABC-type replicon is extensively studied because they have three genes involved in the replication and partitioning systems, always placed in the same order and exhibiting evidence that they might have evolved together [91]. Based on the RepABC replicons in 72 strains of *R. leguminosarum*, the phylogenetic tree of the

core genome and the 'magnified' version trees of each type have revealed that no clear movement was observed for large plasmids between the cryptic species. Despite the absence of recognizable movement among large plasmids, relatively smaller plasmids seem more flexible in acting as vehicles, such as the pRL10-type plasmids showing evidence oftransfer from species group B to group C and the pRL1-type plasmids showing clear incongruent phylogenies with the core genomes. Therefore, we have assumed that some small plasmids might have the capacity to be transferred.

In comparison to the movement between species, relatively more cases of transfer from strain to strain have been found. From large plasmids, such as pRL12- and pRL11-type, to small ones like pRL7-type, a variety of plasmids have been observed with the ability to move from other strains based on the incongruent topology of specific clades in the phylogenetic trees. We have also provided clear evidence that the partitioning coupled genes *repAB* show different phylogenetic patterns than replication gene *repC*. *repC*s of pRL12- and pRL9-type plasmids are highly similar, while *repAB*s of the two replicons are not, which might indicate that following the replication of two replicons, gene conversion might take place, making them assume the discrete types of *repAB*s. This also suggests that partitioning systems are more important than replication systems for plasmid incompatibility in rhizobia [2].

7.3 Final remarks on the evolution of plasmid backbone systems

By performing a comprehensive analysis of the plasmid replication and partitioning systems in proteobacteria, we have examined the distribution, host range and possible relationships of the two systems. The development of the database provides a good and convenient platform for their research. The phylogenetic analysis of both systems also offers more detailed insights of the systems from an evolutionary perspective. Previous research has been based on the 'vertical' analysis of the individual bacterial genome, particularly within the

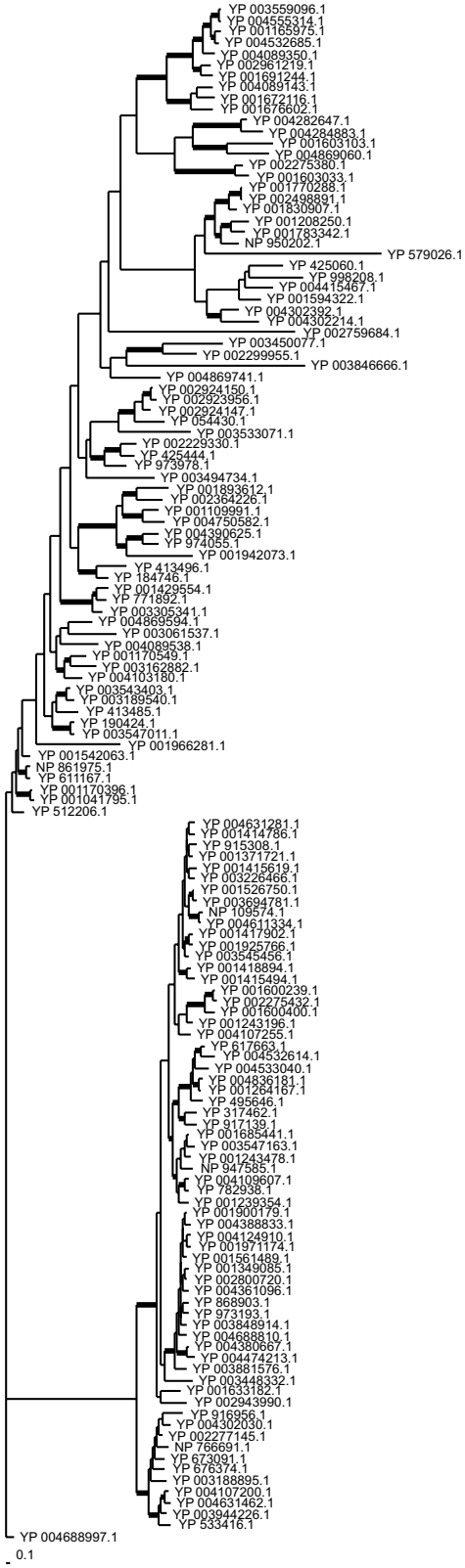
species level. This is also significant, because it contributes to the identification of genes involved in the genome, which was a first step for identifying the entire genome. It does not, however, give effective information in terms of the plasmids' diversity, which differentiates our analysis from previous work. Clearly, more work is needed to improve the present research, such as implementing the automatic update of the database andexpanding research into more families of the two systems. Moreover, integration of the work with that of other groups who study different backbone systems, such as those involved in motility [45], would be very interesting, because it would provide a differentperspective onthe evolutionary history of plasmids.

Appendix 1

Original phylogenetic trees presented in chapter 3. Amino acid sequences were aligned using MAFFT and phylogenies were constructed using PhyML. If there were many homologs consisting of highly similar sequences, over 98% similar sequences were deleted and the fact was indicated in the text below the figures.

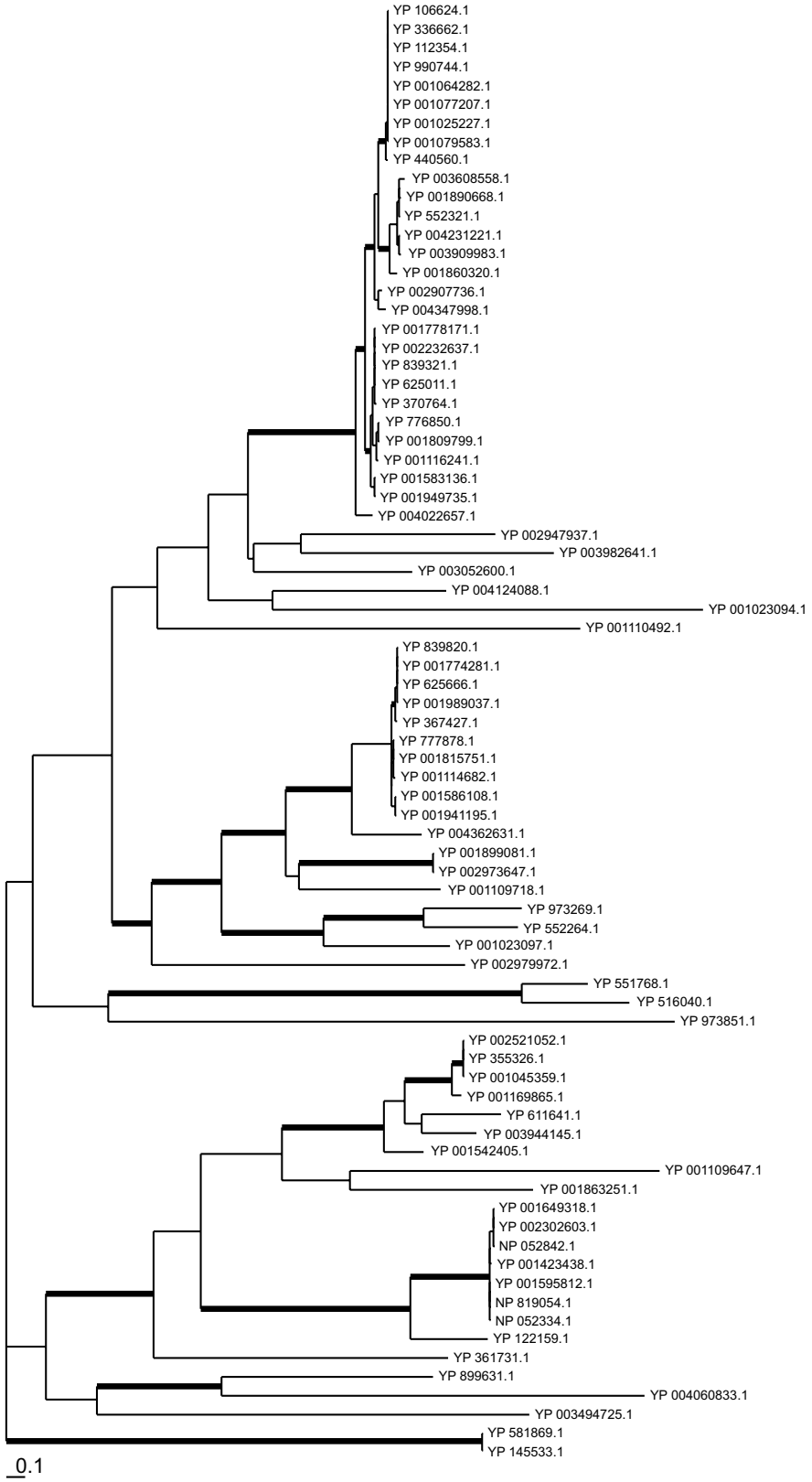
Appendix 1.1 Original tree of Figure 3.4

Shown is the original phylogenetic tree based on the RepC protein sequences used to produce the final tree provided in Figure 3.5 in Chapter 3.



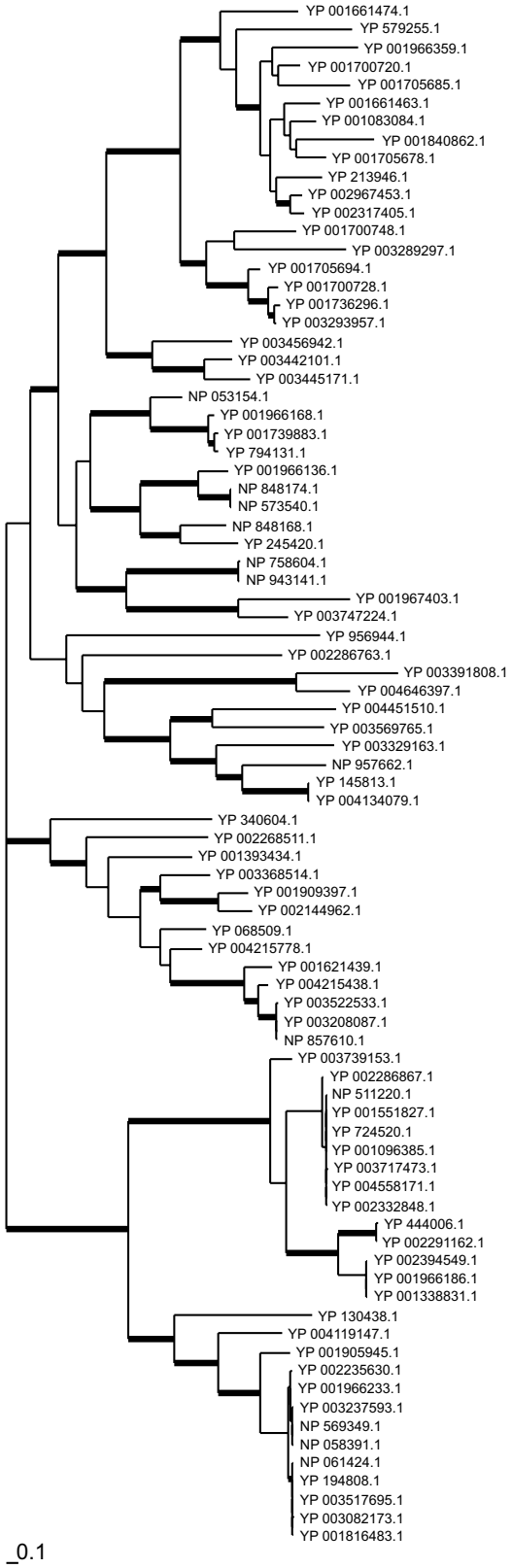
Appendix 1.2 Original tree of Figure 3.6

Shown is the original phylogenetic tree based on the RepA-like protein sequences used to produce the final tree provided in Figure 3.7 in Chapter 3.



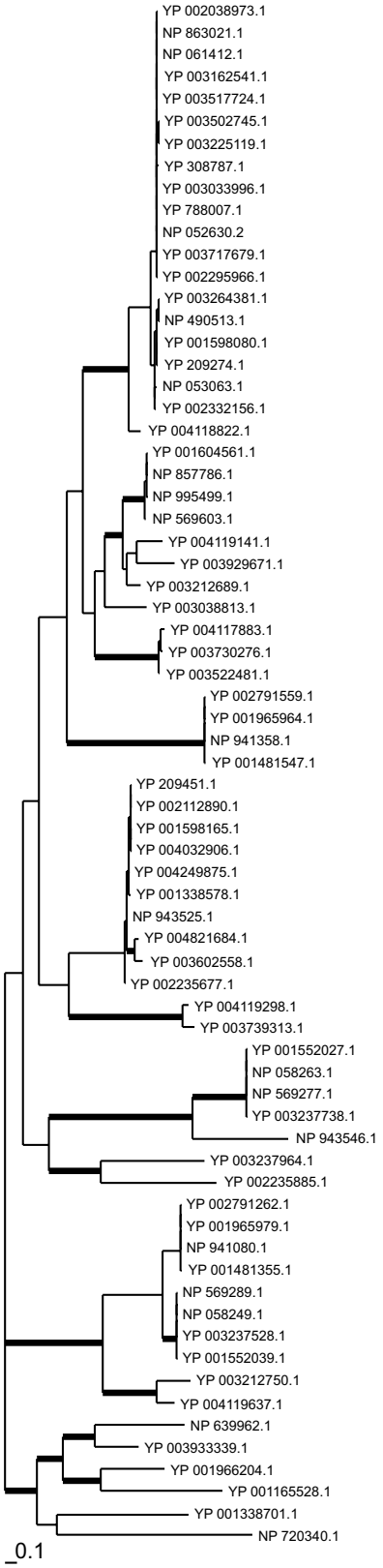
Appendix 1.3 Original tree of Figure 3.8

Shown is the original phylogenetic tree based on the RepB-like protein sequences used to produce the final tree provided in Figure 3.9 in Chapter 3.



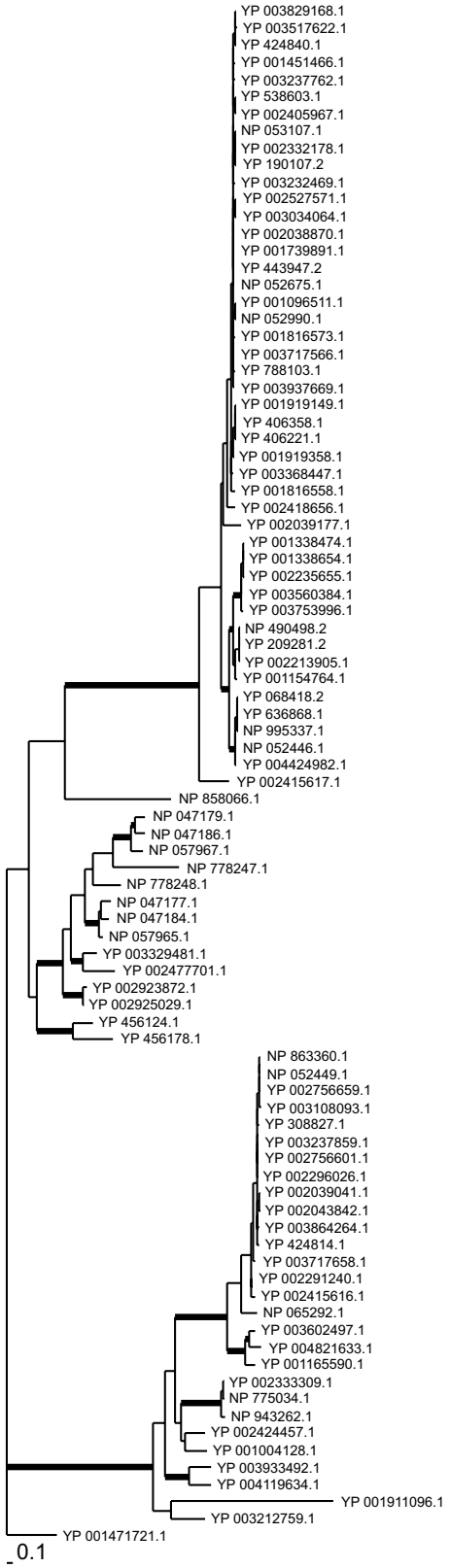
Appendix 1.4 Original tree of Figure 3.10

Shown is the original phylogenetic tree based on the RepFIA protein sequences used to produce the final tree provided in Figure 3.11 in Chapter 3.



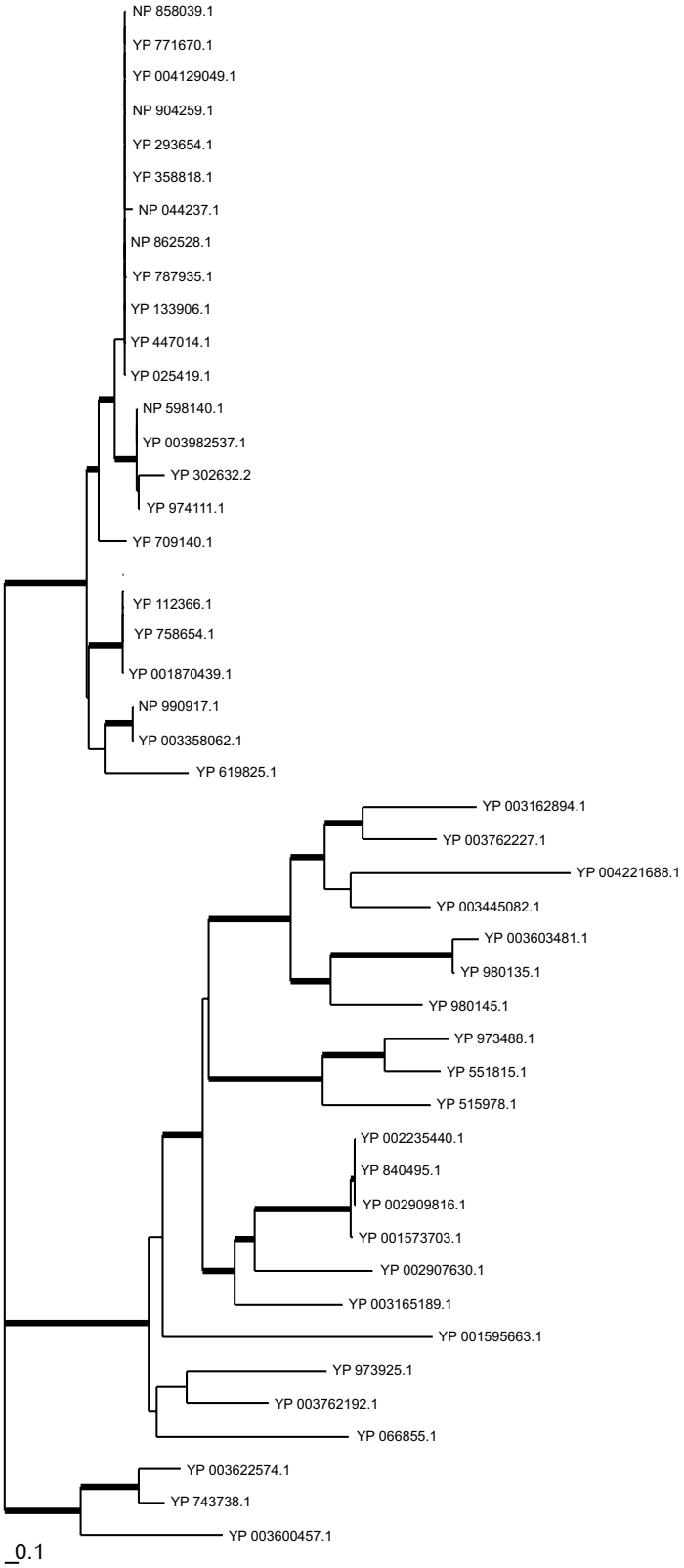
Appendix 1.5 Original tree of Figure 3.12

Shown is the original phylogenetic tree based on the RepFIB protein sequences used to produce the final tree provided in Figure 3.13 in Chapter 3.



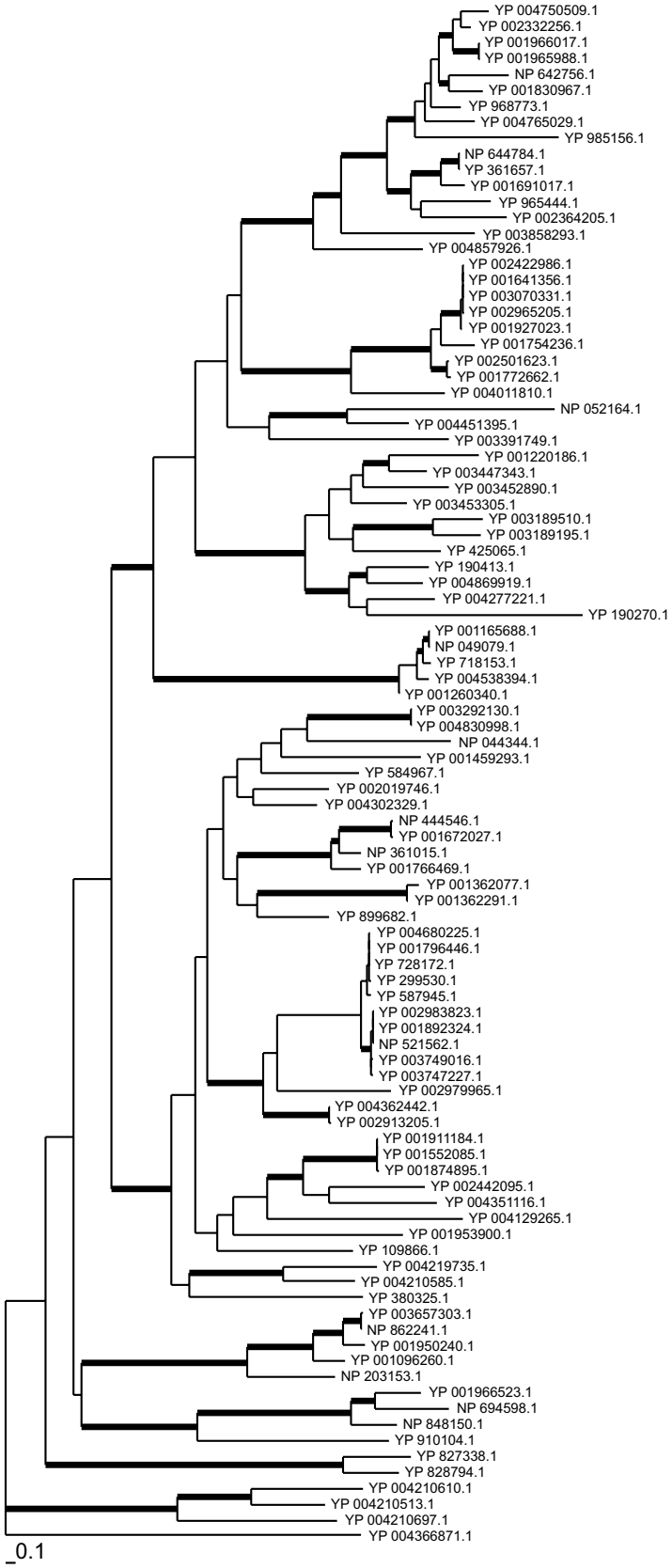
Appendix 1.6 Original tree of Figure 3.14

Shown is the original phylogenetic tree based on the RepFIIA protein sequences used to produce the final tree provided in Figure 3.15 in Chapter 3.



Appendix 1.7 Original tree of Figure 3.16

Shown is the original phylogenetic tree based on the TrfA protein sequences used to produce the final tree provided in Figure 3.17 in Chapter 3.

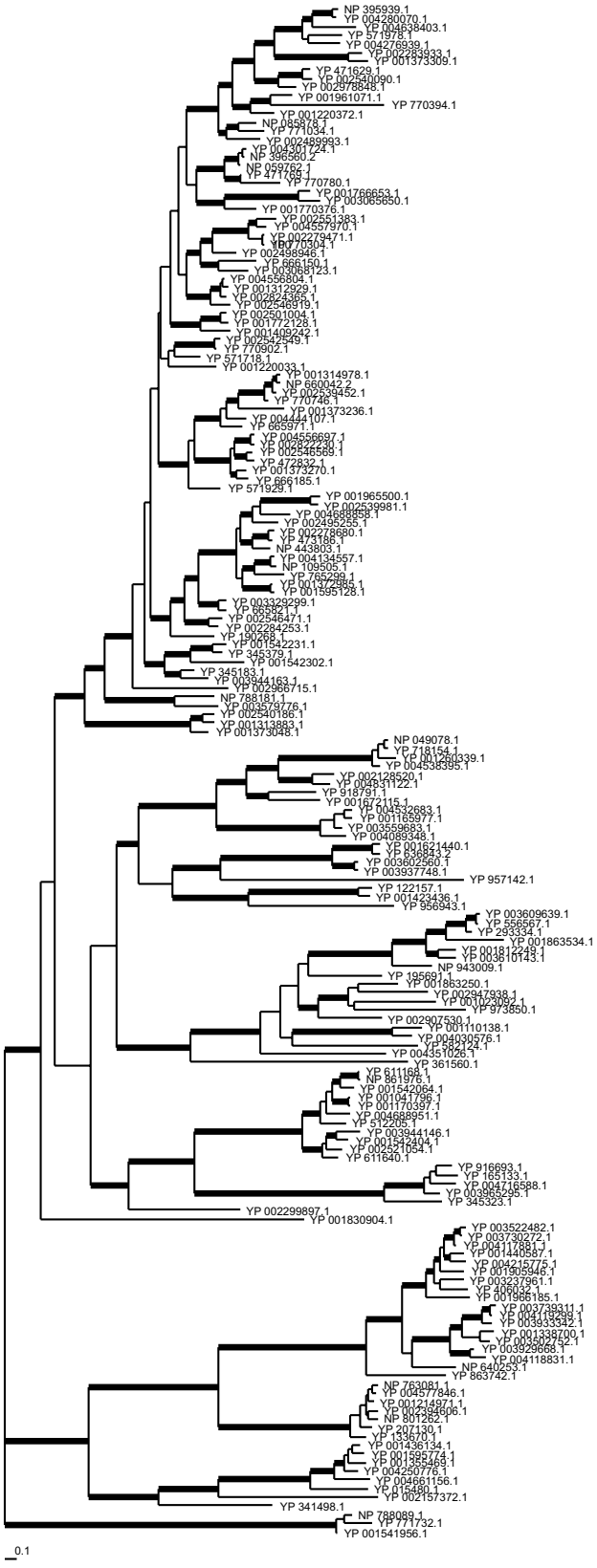


Appendix 1.8 Original tree of Figure 3.18

Shown is the original phylogenetic tree based on the RepA protein sequences used to produce the final tree provided in Figure 3.19 in Chapter 3.

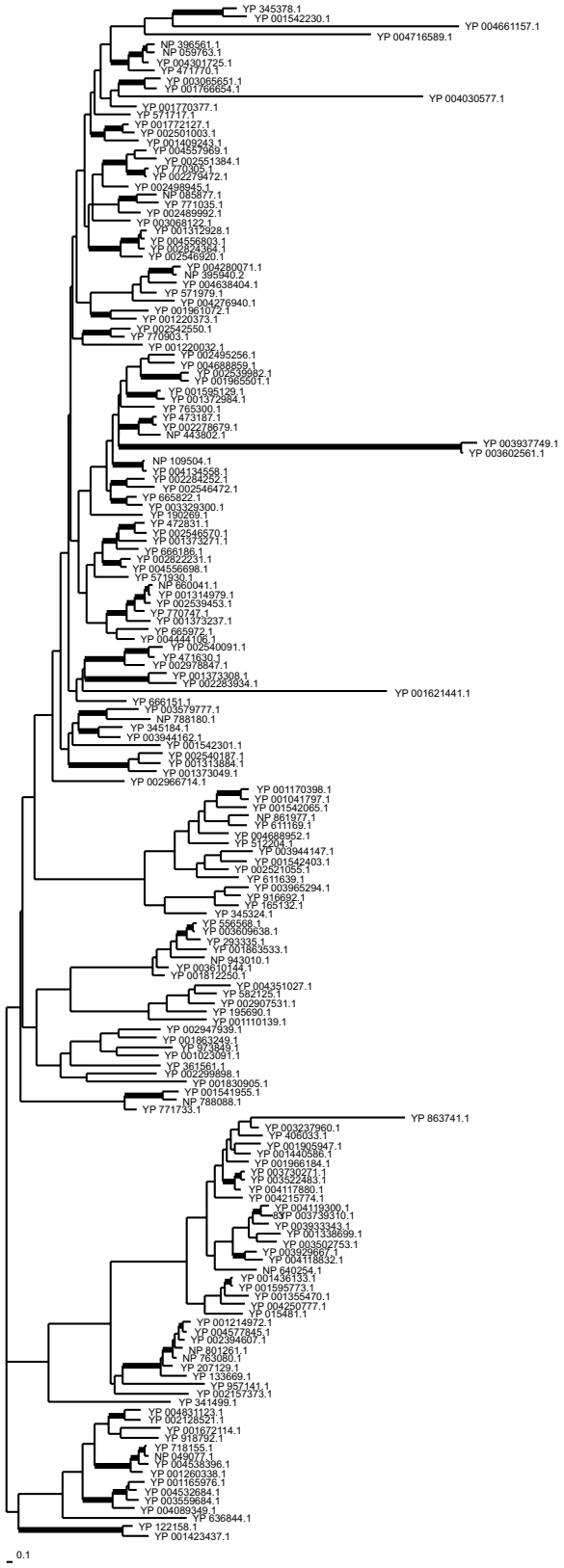
Appendix 2

Original phylogenetic trees presented in chapter 4. Amino acid sequences were aligned using MAFFT and phylogenies were constructed using PhyML. If there were many homologs consisting of highly similar sequences, over 98% similar sequences were deleted and the fact was indicated in the text below the figures.



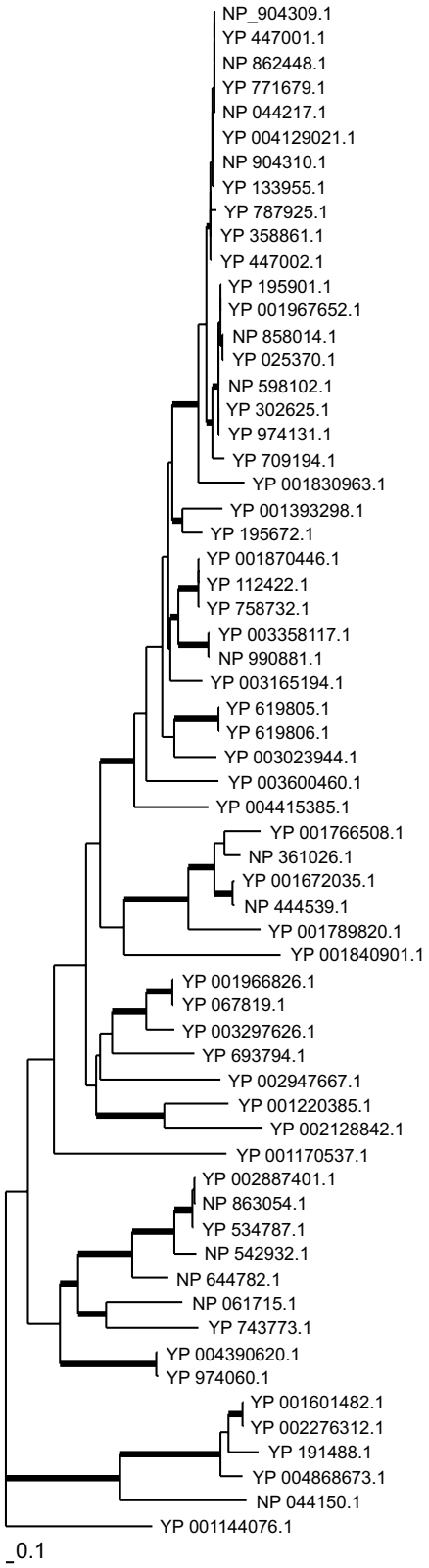
Appendix 2.1 Original tree of Figure 4.4

Shown is the original phylogenetic tree based on the ParA protein sequences used to produce the final tree provided in Figure 4.4 in Chapter 4



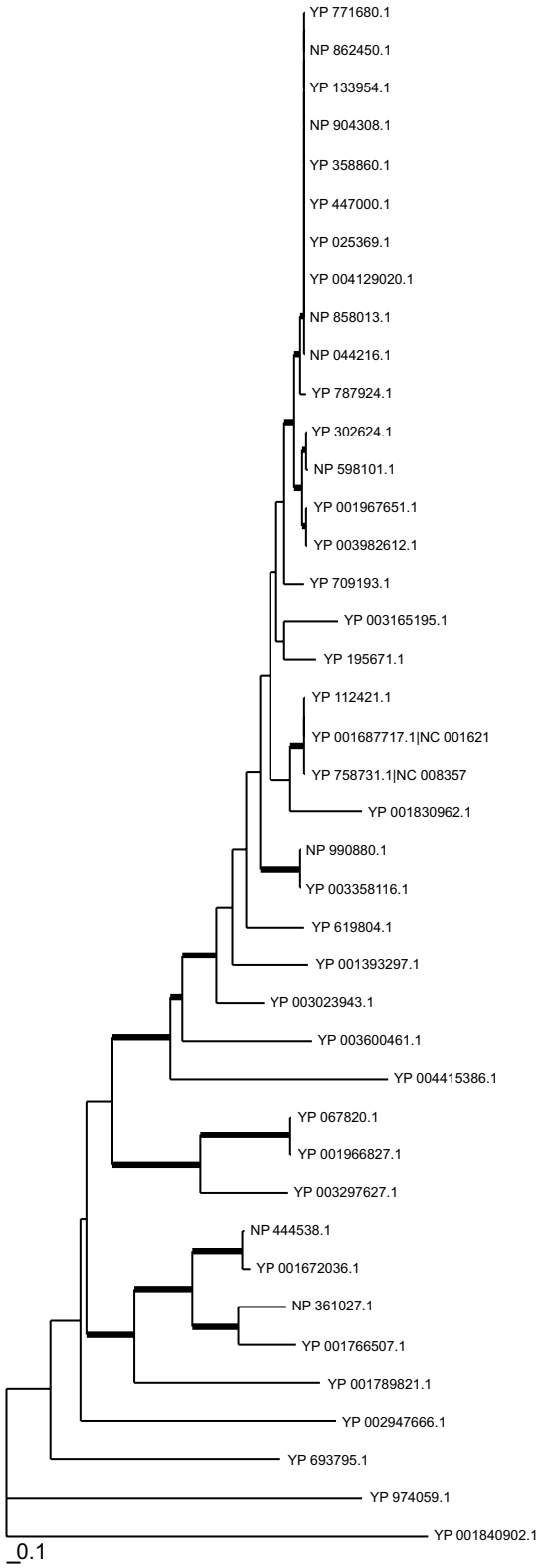
Appendix 2.2 Original tree of Figure 4.5

Shown is the original phylogenetic tree based on the ParB protein sequences used to produce the final tree provided in Figure 4.5 in Chapter 4.



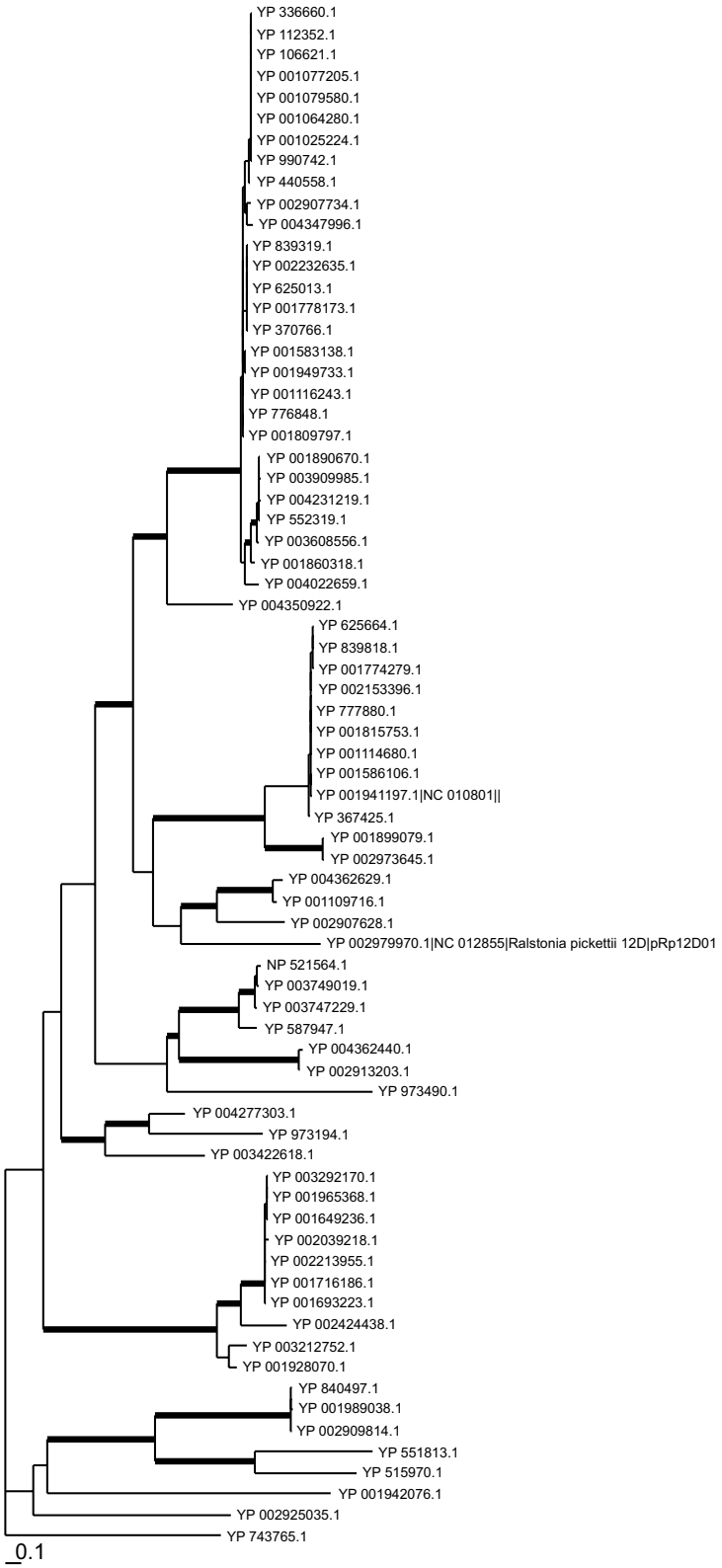
Appendix 2.3 Original tree of Figure 4.8

Shown is the original phylogenetic tree based on the IncC protein sequences used to produce the final tree provided in Figure 4.8 in Chapter 4.



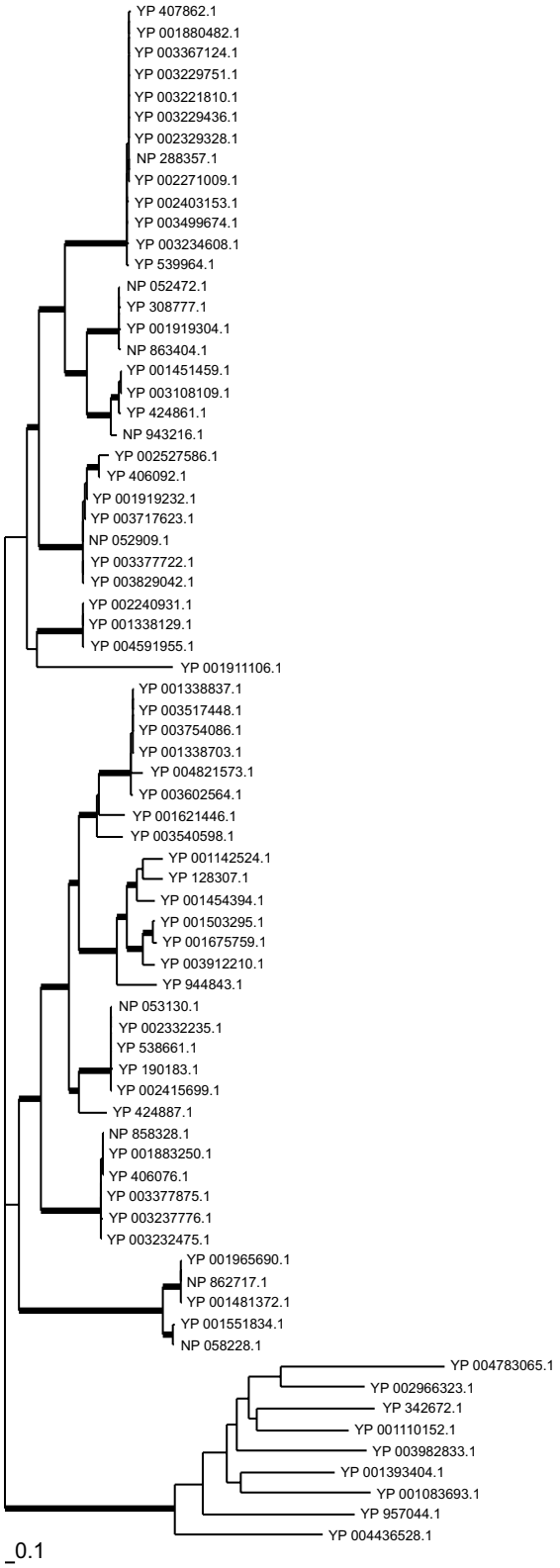
Appendix 2.4 Original tree of Figure 4.9

Shown is the original phylogenetic tree based on the KorB protein sequences used to produce the final tree provided in Figure 4.9 in Chapter 4.



Appendix 2.5 Original tree of Figure 4.11

Shown is the original phylogenetic tree based on the short ParA and ParF protein sequences used to produce the final tree provided in Figure 4.11 in Chapter 4.



Appendix 2.6 Original tree of Figure 4.12

Shown is the original phylogenetic tree based on the ParM protein sequences used to produce the final tree provided in Figure 4.12 in Chapter 4.

Appendix 3

A CD includes all programming codes, web interface files and reference sequence files.

3.1 Source codes for collection of homologous genes, and storing them into database

fill_hmm_res.pl	Parsing a hmmer result and generating a file for accession numbers
fill_gene.pl	Storing data collected into database
make_fasta.py	Calling fasta-formatted files of each family from NCBI
fill_ plist	Storing data into database (Table: p_list)

3.2 Database (table and data, see chapter 2)

p_list.sql	sql fie for table 'p_list'
giconvert.sql	sql fie for table 'giconvert'
gene_family.sql	sql fie for table 'gene_family'
gene.sql	sql fie for table 'gene'
family.sql	sql fie for table 'family'

3.3 Source codes for Web interface

about.html	html file for short introduction
browse_step.html	html file for menu 'browse'
browse_result.cgi	CGI file for the result of 'browse'
search_step1.html	html file for menu 'search'
search_result.cgi	CGI file for the result of 'search'
download.html	html file for downloading data
faq.html	html file for FAQ section
link.html	html file for useful links
Index.html	main page of the web interface

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci* 2005, **102**(39):13950-13955.
2. Young JP, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, Wexler M, Curson AR, Todd JD, Poole PS *et al*: **The genome of *Rhizobium leguminosarum* has recognizable core and accessory components**. *Genome Biology* 2006, **7**(4):R34.
3. Muzzi A, Masignani V, Rappuoli R: **The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials**. *Drug Discovery Today* 2007, **12**(11-12):429-439.
4. Leplae R, Aline Hebrant, Wodak SJ, Toussaint A: **ACLAME: A CLAssification of Mobile genetic Elements**. *Nucleic Acids Research* 2004, **32**(suppl 1):D45-D49.
5. Frost LS, Leplae R, Summers AO, Toussaint A: **Mobile genetic elements: the agents of open source evolution**. *Nature Reviews Microbiology* 2005, **3**(9):722-732.
6. Zinder ND, Lederberg J: **Genetic exchange in *Salmonella***. *Journal of Bacteriology* 1952, **64**(679-699).
7. Heuer H, Abdo Z, Smalla K: **Patchy distribution of flexible genetic elements in bacterial populations mediates robustness to environmental uncertainty**. *FEMS Microbiology Reviews* 2008, **65**(3):361-371.
8. Souza V, Eguiarte LE: **Bacteria gone native vs. bacteria gone awry?: Plasmidic transfer and bacterial evolution**. *Proc Natl Acad Sci* 1997, **94**:5501-5503.
9. Norman A, Hansen LH, Sørensen SJ: **Conjugative plasmids: vessels of the communal gene pool**. *Phil Trans R Soc* 2009, **364**:2275-2289.
10. Paulsson J: **Multileveled selection on plasmid replication**. *Genetics* 2002, **161**:1373-1384.
11. Miller RV, Day MJ: **Microbial evolution: gene establishment, survival, and exchange**: ASM Press; 2004.
12. Doskocil J, Forstova J, Stokrova J: **Temperate and virulent forms of phage theta attacking *Bacillus licheniformis***. *Molec gen Genet* 1978, **160**:311-317.
13. Abedon ST, Kuhl SJ, Blasdel BG, Kutter EM: **Phage treatment of human infections**. *Bacteriophage* 2011, **1**(2):66-85.
14. Mahillon J, Léonard C, Chandler M: **IS elemtns as constituents of bacterial genomes**. *Res Microbiol* 1999, **150**:675-687.
15. Kothapalli S, Nair S, Alokam S, Pang T, Khakhria R, Woodward D, Johnson W, Stocker BA, Sanderson KE, Liu SL: **Diversity of genome structure in**

***Salmonella enterica* serovar Typhi populations**. *Journal of Bacteriology* 2005, **187**(8):2638-2650.
16. Salyers AA, Shoemaker NB, Stevens AM, Li LY: **Conjugative transposons: an unusual and diverse set of integrated gene transfer elements**. *Microbiological Reviews* 1995, **59**(4):579-590.
17. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW: **Genomic islands: tools of bacterial horizontal gene transfer and evolution**. *FEMS Microbiology Reviews* 2009, **33**(2):376-393.
18. Langille MGI, Hsiao WWL, Brinkman FSL: **Detecting genomic islands using bioinformatics approaches**. *Nature Reviews, Microbiology* 2010, **8**:373-382.
19. Hall RM, Collis CM: **Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination**. *Molecular Microbiology* 2006, **15**(4):593-600.
20. Freese EB: **Transitions and transversions induced by depurinating agents**. *Genetics* 1961, **47**:540-545.
21. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: organization and impact within the current human genome project assembly**. *Genome Research* 2001, **11**:1005-1017.
22. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T *et al*: **Lineage-specific gene duplication and loss in human and great ape evolution**. *PLoS Biology* 2004, **2**(7):937-954.
23. Fraser JA, Huang JC, Pukkila-Worley R, Alspaugh JA, Mitchell TG, Heitman J: **Chromosomal translocation and segmental duplication in *Cryptococcus neoformans***. *Eukaryotic Cell* 2005, **4**(2):401-406.
24. Tatum EL, Lederberg J: **Gene recombination in the bacterium *Escherichia Coli***. *Journal of Bacteriology* 1947, **53**(6):673-684.
25. Maiden MCJ: **Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria**. *Clinical Infectious Diseases* 1998, **27**(suppl 1):S12-20.
26. Cohan FM, Koeppel AF: **The origins of ecological diversity in prokaryotes**. *Current Biology* 2008, **18**:R1024-R1034.
27. Koonin EV: **Horizontal gene transfer: the path to maturity**. *Molecular Microbiology* 2003, **50**(3):725-727.
28. Boto L: **Horizontal gene transfer in evolution: facts and challenges**. *Proc R Soc B* 2010, **277**:819-827.
29. Rest JS, Mindell DP: **Retroids in Archaea: Phylogeny and lateral origins**. *Mol Biol Evol* 2003, **20**(7):1134-1142.
30. Gophna U, Charlebois RL, Doolittle WF: **Have archaeal genes contributed to bacterial virulence?** *Trends in Microbiology* 2004, **12**(5):213-219.
31. Andersson JO: **Lateral gene transfer in eukaryotes**. *Cellular and Molecular Life Sciences* 2005, **62**:1182-1197.
32. Watkins RF, Gray MW: **The frequency of eubacterium-to-eukaryote lateral gene transfers shows significant cross-taxa variation within Amoebozoa**. *Journal of Molecular Evolution* 2006, **63**:801-814.
33. Guljamow A, Jenke-Kodama H, Saumweber H, Quillardet P, Frangeul L, Castets AM, Bouchier C, Marsac NTd, Dittmann E: **Horizontal gene**

transfer of two cytoskeletal elements from a eukaryote to a *s*. *Current Biology* 2007, **17**(17):R757-R759.
34. Nedelcu AM, Miles IH, Fagir AM, Karol K: **Adaptive eukaryote-to-eukaryote lateral gene transfer: stress-related genes of algal origin in the closest unicellular relatives of animals**. *Journal of Evolutionary Biology* 2008, **21**(6):1852-1860.
35. Bergh O, Borsheim KY, Bratbak G, Heldal M: **High abundance of viruses found in aquatic environments**. *Nature* 1989, **340**:467-468.
36. Dröge M, Pühler A, Selbitschka W: **Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies**. *Biology and Fertility of Soils* 1999, **29**(3):221-245.
37. Chen I, Dubnau D: **DNA uptake during bacterial transformation**. *Nature Reviews, Microbiology* 2004, **2**:241-249.
38. Lilley AK, Fry JC, Day MJ, Bailey MJ: ***In situ* transfer of an exogenously isolated plasmid between *Pseudomonas* spp. in sugar beet rhizosphere**. *Microbiology* 1994, **140**(1):27-33.
39. Ebersbach G, Gerdes K: **Plasmid segegation mechanisms**. *Annual Review Genetics* 2005, **39**:453-479.
40. Hayes F, Barilla D: **The bacterial segrosome: a dynamic nucleoprotein machine for DNA trafficking and segregation**. *Nature Reviews Microbiology* 2006, **4**(2):133-143.
41. Shih YL, Rothfield L: **The bacterial cytoskeleton**. *Microbiology and Molecular Biology Reviews* 2006, **70**:729-754.
42. Thisted T, S.Sorensen N, H.Wagner EG, Gerdes K: **Mechanism of post-segregational killing: Sok antisense RNA interacts with Hok mRNA via its 5'-end singlestranded leader and competes with the 3'-end of Hok *mRNA* for binding to the *mok* translational initiation region**. *The EMBO Journal* 1994, **13**(8):1960-1968.
43. Thisted T, Gerdes K: **Mechanism of post-segregational killing by the *hok/sok* system of plasmid R1: Sok antisense RNA regulates *hok* gene expression indirectly through the overlapping *mok* gene**. *Journal of Molecular Biology* 1992, **223**(1):41-54.
44. Ding H, Hynes MF: **Plasmid transfer systems in the rhizobia**. *Canadian Journal of Microbiology* 2009, **55**(8):917-927.
45. Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EPC, Cruz Fdl: **Mobility of Plasmids**. *Mircobiology and Molecular Biology Reviews* 2010, **74**(3):434-452.
46. Jackson RW, Vinatzer B, Arnold DL, Dorus S, Murillo J: **The influence of the accessory genome on bacterial pathogen evolution**. *Mobile Genetic Elements* 2011, **1**(1):55-65.
47. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F: **The bacterial pan-genome: a new paradigm in microbiology**. *International Microbiology* 2010, **13**:45-57.
48. Norman A, Hansen LH, Sorensen SJ: **Conjugative plasmids: vessels of the communal gene pool**. *Phil Trans R Soc* 2009, **364**(1527):2275-2289.
49. Lathe WC, Williams JM, Mangan ME, Karolchik D: **Genomic data resources: challenges and promises**. *Nature Education* 2008, **1**(3).

50. Mølbak L, Tett A, Ussery DW, Wall K, Turner S, Bailey M, Field D: **The plasmid genome database**. *Microbiology Commnet* 2003, **149**(11):3043-3045.
51. Hsiao W, Wan I, Jones SJ, Brinkman FSL: **IslandPath: aiding detection of genomic islands in prokaryotes**. *Bioinformatics* 2003, **19**(3):418-420.
52. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M: **ISfinder: the reference centre for bacterial insertion sequences**. *Nucleic Acids Research* 2005, **34**(suppl 1):D32-D36.
53. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**(41):1-14.
54. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S *et al*: **The National Center for Biotechnology Information's Protein Clusters Database**. 2008, *Nucleic Acids Research*(37)D216-D223.
55. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J *et al*: **The Pfam protein families database**. *Nucleic Acids Research* 2012, **40**(D1):D290-D301.
56. Molbak L, Tett A, Ussery DW, Wall K, Turner S, Bailey M, Field D: **The plasmid genome database**. *Microbiology* 2003, **149**(Pt 11):3043-3045.
57. Hsiao W, Wan I, Jones SJ, Brinkman FS: **IslandPath: aiding detection of genomic islands in prokaryotes**. *Bioinformatics* 2003, **19**(3):418-420.
58. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M: **ISfinder: the reference centre for bacterial insertion sequences**. *Nucleic Acids Research* 2006, **34**(Database issue):D32-36.
59. Leplae R, Hebrant A, Wodak SJ, Toussaint A: **ACLAME: a CLAssification of Mobile genetic Elements**. *Nucleic Acids Research* 2004, **32**(Database issue):D45-49.
60. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.
61. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S *et al*: **The National Center for Biotechnology Information's Protein Clusters Database**. *Nucleic Acids Research* 2009, **37**(Database issue):D216-223.
62. Katoh K, Kuma K-i, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic Acids Research* 2005, **33**(2):511-518.
63. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0.** *Systematic Biology* 2010, **59**(3):307-321.
64. Page RDM: **TREEVIEW: An application to display phylogenetic trees on personal computers**. *Computer Applications in the Biosciences* 1996, **12**:357-358.
65. Eddy SR: **Accelerated profile HMM searches**. *PLoS Computational Biology* 2011, **7**(10):1-16.

66. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching**. *Nucleic Acids Research* 2011, **Web Server Issue**(39):W29-W37.
67. Girlich D, Poirel L, Schluter A, Nordmann P: **TLA-2, a novel Ambler class A expanded-spectrum beta-lactamase**. *Antimicrobial Agents and Chemotherapy* 2005, **49**(11):4767-4770.
68. Solar GD, Giraldo R, Ruiz-echevarria MJ, Espinosa M, Diaz-orejas R: **Replication and control of circular bacterial plasmids**. *Microbiology and Molecular Biology Reviews* 1998, **62**(2):434-464.
69. Kues U, Stahl U: **Replication of plasmids in gram-negative bacteria**. *Microbiological Reviews* 1989, **53**(4):491-516.
70. Ishiai M, Wada C, Kawasaki Y, Yura T: **Replication initiator protein RepE of mini-F plasmid: functional differentiation between monomers (initiator) and dimers (autogenous repressor)**. *Proc of the Natl Acad Sci* 1994, **91**(9):3839-3843.
71. Masai H, Arai K: ***Escherichia coli dnaT* gene function is required for pBR322 plasmid replication but not for R1 plasmid replication**. *Journal of Bacteriology* 1989, **171**(6):2975-2980.
72. Loftie-Eaton W, Rawlings DE: **Diversity, biology and evolution of IncQ-family plasmids**. *Plasmid* 2012, **67**(1):15-34.
73. Kramer MG, Khan SA, Espinosa M: **Plasmid rolling circle replication: identification of the RNA polymerase-directed primer RNA and requirement for DNA polymerase I for lagging strand synthesis**. *The EMBO journal* 1997, **16**(18):5784-5795.
74. Novick RP: **Plasmid Incompatibility**. *Microbiological Reviews* 1987, **51**(4):381-395.
75. Couturier M, Bex F, Bergquist PL, Maas WK: **Identification and classification of bacterial plasmids** *Microbiological Reviews* 1988, **52**(3):375-395.
76. Datta N, Hedges RW: **Compatibility groups among fi - R factors**. *Nature* 1971, **234**(5326):222-223.
77. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ: **Identification of plasmids by PCR-based replicon typing**. *Journal of Microbiological Methods* 2005, **63**(3):219-228.
78. Garcillan-Barcia MP, Francia MV, de la Cruz F: **The diversity of conjugative relaxases and its application in plasmid classification**. *FEMS Microbiology Reviews* 2009, **33**(3):657-687.
79. Suzuki H, Yano H, Brown CJ, Top EM: **Predicting plasmid promiscuity based on genomic signature**. *Journal of Bacteriology* 2010, **192**(22):6045-6055.
80. De Gelder L, Ponciano JM, Joyce P, Top EM: **Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship**. *Microbiology* 2007, **153**(Pt 2):452-463.
81. Smith CA, Thomas CM: **Narrow-host-range IncP plasmid pHH502-1 lacks a complete IncP replication system**. *Journal of General Microbiology* 1987, **133**(8):2247-2252.
82. Guiney DG, Jr.: **Promiscuous transfer of drug resistance in gram-negative bacteria**. *The Journal of Infectious Diseases* 1984, **149**(3):320-329.

83. Fernandez-Lopez R, Garcillan-Barcia MP, Revilla C, Lazaro M, Vielva L, Cruz Fdl: **Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution**. *FEMS Microbiology Reviews* 2006, **30**:942-966.
84. Harrison PW, Lower RP, Kim NK, Young JP: **Introducing the bacterial 'chromid': not a chromosome, not a plasmid**. *Trends in Microbiology* 2010, **18**(4):141-148.
85. Durland RH, Helinski DR: **Replication of the broad-host-range plasmid RK2: direct measurement of intracellular concentrations of the essential TrfA replication proteins and their effect on plasmid copy number**. *Journal of Bacteriology* 1990, **172**(7):3849-3858.
86. Stenger DC, Lee MW: **Phylogeny of Replication Initiator Protein TrfA Reveals a Highly Divergent Clade of Incompatibility Group P1 Plasmids**. *Applied and Environmental Microbiology* 2011, **77**(7):2522-2526.
87. Osborn AM, da Silva Tatley FM, Steyn LM, Pickup RW, Saunders JR: **Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons**. *Microbiology* 2000, **146 (Pt 9)**:2267-2275.
88. Taylor DE, Hedges RW, Bergquist PL: **Molecular homology and incompatibility relationships between F and IncH1 plasmids**. *Journal of General Microbiology* 1985, **131**(6):1523-1530.
89. Cevallos MA, Cervantes-Rivera Rn, Gutiérrez-Ríos RMa: **The *repABC* plasmid family**. *Plasmid* 2008, **60**(1):19-37.
90. MacLellan SR, Zaheer R, Sartor AL, MacLean AM, Finan TM: **Identification of a megaplasmid centromere reveals genetic structural diversity within the *repABC* family of basic replicons**. *Mol Microbiol* 2006, **59**(5):1559-1575.
91. Petersen J, Brinkmann H, Pradella S: **Diversity and evolution of *repABC* type plasmids in *Rhodobacterales***. *Environmental Microbiology* 2009, **11**(10):2627-2638.
92. Gallie DR, Kado CI: **Minimal region necessary for autonomous replication of pTAR**. *Journal of Bacteriology* 1988, **170**(7):3170-3176.
93. Schneiker S, Keller M, Droge M, Lanka E, Puhler A, Selbitschka W: **The genetic organization and evolution of the broad host range mercury resistance plasmid pSB102 isolated from a microbial population residing in the rhizosphere of alfalfa**. *Nucleic Acids Research* 2001, **29**(24):5169-5181.
94. Scott JR: **Regulation of plasmid replication**. *Microbiological Reviews* 1984, **48**(1):1-23.
95. Timmis KN, Andres I, Slocombe PM, Synenki RM: **Plasmid-encoded functions involved in the replication and inheritance of antibiotic-resistance plasmid R6-5**. *Cold Spring Harbor Symposia on Quantitative Biology* 1979, **43 Pt 1**:105-110.
96. Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ, Banta L, Dickerman AW, Paulsen I *et al*: **Genome sequences of three *agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria**. *Journal of Bacteriology* 2009, **191**(8):2501-2511.

97. Petersen J: **Phylogeny and compatibility: plasmid classification in the genomics era**. *Archives of Microbiology* 2011, **193**(5):313-321.
98. Osborne D, Frost L, Tobal K, Liu Yin JA: **Elevated levels of WT1 transcripts in bone marrow harvests are associated with a high relapse risk in patients autografted for acute myeloid leukaemia**. *Bone Marrow Transplantation* 2005, **36**(1):67-70.
99. da Silva-Tatley FM, Steyn LM: **Characterization of a replicon of the moderately promiscuous plasmid, pGSH5000, with features of both the mini-replicon of pCU1 and the ori-2 of F**. *Molecular Microbiology* 1993, **7**(5):805-823.
100. Kline BC: **A review of mini-F plasmid maintenance**. *Plasmid* 1985, **14**(1):1-16.
101. Pagotto F, Dillon JA: **Multiple origins and replication proteins influence biological properties of beta-lactamase-producing plasmids from *Neisseria gonorrhoeae***. *Journal of Bacteriology* 2001, **183**(19):5472-5481.
102. Saadi S, Maas WK, Hill DF, Bergquist PL: **Nucleotide sequence analysis of RepFIC, a basic replicon present in IncFI plasmids P307 and F, and its relation to the RepA replicon of IncFII plasmids**. *Journal of Bacteriology* 1987, **169**(5):1836-1846.
103. Gabant P, Chahdi AO, Couturier M: **Nucleotide sequence and replication characteristics of RepHI1B: a replicon specific to the IncHI1 plasmids**. *Plasmid* 1994, **31**(2):111-120.
104. Gotz A, Pukall R, Smit E, Tietze E, Prager R, Tschape H, van Elsas JD, Smalla K: **Detection and characterization of broad-host-range plasmids in environmental bacteria by PCR**. *Applied and Environmental Microbiology* 1996, **62**(7):2621-2628.
105. Novais A, Canton R, Moreira R, Peixe L, Baquero F, Coque TM: **Emergence and dissemination of *Enterobacteriaceae* isolates producing CTX-M-1-like enzymes in Spain are associated with IncFII (CTX-M-15) and broad-host-range (CTX-M-1, -3, and -32) plasmids**. *Antimicrobial Agents and Chemotherapy* 2007, **51**(2):796-799.
106. Chen YT, Shu HY, Li LH, Liao TL, Wu KM, Shiau YR, Yan JJ, Su IJ, Tsai SF, Lauderdale TL: **Complete nucleotide sequence of pK245, a 98-kilobase plasmid conferring quinolone resistance and extended-spectrum-beta-lactamase activity in a clinical *Klebsiella pneumoniae* isolate**. *Antimicrobial Agents and Chemotherapy* 2006, **50**(11):3861-3866.
107. Gibbs MD, Spiers AJ, Bergquist PL: **RepFIB: a basic replicon of large plasmids**. *Plasmid* 1993, **29**(3):165-179.
108. Murata T, Ohnishi M, Ara T, Kaneko J, Han CG, Li YF, Takashima K, Nojima H, Nakayama K, Kaji A *et al*: **Complete nucleotide sequence of plasmid Rts1: implications for evolution of large plasmid genomes**. *Journal of Bacteriology* 2002, **184**(12):3194-3202.
109. Sherburne CK, Lawley TD, Gilmour MW, Blattner FR, Burland V, Grotbeck E, Rose DJ, Taylor DE: **The complete DNA sequence and analysis of R27, a large IncHI plasmid from *Salmonella typhi* that is temperature sensitive for transfer**. *Nucleic Acids Research* 2000, **28**(10):2177-2186.
110. Burland V, Shao Y, Perna NT, Plunkett G, Sofia HJ, Blattner FR: **The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157:H7**. *Nucleic Acids Research* 1998, **26**(18):4196-4204.

111. Golebiewski M, Kern-Zdanowicz I, Zienkiewicz M, Adamczyk M, Zylinska J, Baraniak A, Gniadkowski M, Bardowski J, Ceglowski P: **Complete nucleotide sequence of the pCTX-M3 plasmid and its involvement in spread of the extended-spectrum beta-lactamase gene blaCTX-M-3**. *Antimicrobial Agents and Chemotherapy* 2007, **51**(11):3789-3795.
112. Froehlich B, Parkhill J, Sanders M, Quail MA, Scott JR: **The pCoo plasmid of enterotoxigenic *Escherichia coli* is a mosaic cointegrate**. *Journal of Bacteriology* 2005, **187**(18):6509-6516.
113. Meyer R, Figurski D, Helinski DR: **Physical and genetic studies with restriction endonucleases on the broad host-range plasmid RK2**. *Molecular & General Genetics : MGG* 1977, **152**(3):129-135.
114. Lanka E, Furste JP, Yakobson E, Guiney DG: **Conserved regions at the DNA primase locus of IncP alpha and IncP beta plasmids**. *Plasmid* 1985, **14**(3):217-223.
115. Adamczyk M, Jagura-Burdzy G: **Spread and survival of promiscuous IncP-1 plasmids**. *Acta Biochimica Polonica* 2003, **50**(2):425-453.
116. Trefault N, Iglesia RDl, Molina AM, Manzano M, Ledger T, Pérez-Pantoja D, Sánchez MA, Stuardo M, González B: **Genetic organization of the catabolic plasmid pJP4 from *Ralstonia eutropha* JMP134 (pJP4) reveals mechanisms of adaptation to chloroaromatic pollutants and evolution of specialized chloroaromatic degradation pathways**. *Environmental Microbiology* 2004, **6**(7):655–668.
117. Schluter A, Heuer H, Szczepanowski R, Poler SM, Schneiker S, Puhler A, Top EM: **Plasmid pB8 is closely related to the prototype IncP-1beta plasmid R751 but transfers poorly to *Escherichia coli* and carries a new transposon encoding a small multidrug resistance efflux protein**. *Plasmid* 2005, **54**(2):135-148.
118. Haines AS, Cheung M, Thomas CM: **Evidence that IncG (IncP-6) and IncU plasmids form a single incompatibility group**. *Plasmid* 2006, **55**(3):210-215.
119. Vedler E, Vahter M, Heinaru A: **The completely sequenced plasmid pEST4011 contains a novel IncP1 backbone and a catabolic transposon harboring tfd genes for 2,4-dichlorophenoxyacetic acid degradation**. *Journal of Bacteriology* 2004, **186**(21):7161-7174.
120. Schluter M, Stieltjes B, Hahn HK, Rexilius J, Konrad-verse O, Peitgen HO: **Detection of tumour infiltration in axonal fibre bundles using diffusion tensor imaging**. *The International Journal of Medical Robotics Computer Assisted Surgery : MRCAS* 2005, **1**(3):80-86.
121. Norberg P, Bergstrom M, Jethava V, Dubhashi D, Hermansson M: **The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination**. *Nature Communications* 2011, **2**:268.
122. Smith CA, Pinkney M, Guiney DG, Thomas CM: **The ancestral IncP replication system consisted of contiguous *oriV* and *trfA* segments as deduced from a comparison of the nucleotide sequences of diverse IncP plasmids**. *Journal of General Microbiology* 1993, **139**(8):1761-1766.
123. Karunakaran P, Blatny JM, Ertesvag H, Valla S: **Species-dependent phenotypes of replication-temperature-sensitive *trfA* mutants of plasmid RK2: a codon-neutral base substitution stimulates**

temperature sensitivity by leading to reduced levels of *trfA* expression. *Journal of Bacteriology* 1998, **180**(15):3793-3798.
124. Shingler V, Thomas CM: **Analysis of the *trfA* region of broad host-range plasmid RK2 by transposon mutagenesis and identification of polypeptide products**. *Journal of Molecular Biology* 1984, **175**(3):229-249.
125. Smith CA, Shingler V, Thomas CM: **The *trfA* and *trfB* promoter regions of broad host range plasmid RK2 share common potential regulatory sequences**. *Nucleic Acids Research* 1984, **12**(8):3619-3630.
126. Durland RH, Toukdarian A, Fang F, Helinski DR: **Mutations in the *trfA* replication gene of the broad-host-range plasmid RK2 result in elevated plasmid copy numbers**. *Journal of Bacteriology* 1990, **172**(7):3859-3867.
127. Fang FC, Helinski DR: **Broad-host-range properties of plasmid RK2: importance of overlapping genes encoding the plasmid replication initiation protein TrfA**. *Journal of Bacteriology* 1991, **173**(18):5861-5868.
128. Jiang Y, Pacek M, Helinski DR, Konieczny I, Toukdarian A: **A multifunctional plasmid-encoded replication initiation protein both recruits and positions an active helicase at the replication origin**. *Proc Natl Acad Sci* 2003, **100**(15):8692-8697.
129. Yano H, Deckert GE, Rogers LM, Top EM: **Roles of long and short replication initiation proteins in the fate of IncP-1 plasmids**. *Journal of Bacteriology* 2012, **194**(6):1533-1543.
130. Tennstedt T, Szczepanowski R, Krahn I, Puhler A, Schluter A: **Sequence of the 68,869 bp IncP-1alpha plasmid pTB11 from a waste-water treatment plant reveals a highly conserved backbone, a Tn402-like integron and other transposable elements**. *Plasmid* 2005, **53**(3):218-238.
131. Diaz A, Lacks SA, Lopez P: **Multiple roles for DNA polymerase I in establishment and replication of the promiscuous plasmid pLS1**. *Molecular Microbiology* 1994, **14**(4):773-783.
132. Tait RC, Close TJ, Rodriguez RL, Kado CI: **Isolation of the origin of replication of the IncW-group plasmid pSa**. *Gene* 1982, **20**(1):39-49.
133. Loper JE, Kado CI: **Host range conferred by the virulence-specifying plasmid of *Agrobacterium tumefaciens***. *Journal of Bacteriology* 1979, **139**(2):591-596.
134. Watanabe T, Furuse C, Sakaizumi S: **Transduction of various R factors by phage P1 in *Escherichia coli* and by phage P22 in *Salmonella typhimurium***. *Journal of Bacteriology* 1968, **96**(5):1791-1795.
135. Datta N: **Prevalence of extrachromosomal drug resistance. R Factors in *Escherichia coli***. *Annals of the New York Academy of Sciences* 1971, **182**:59-64.
136. Coetzee JN, Datta N, Hedges RW: **R factors from *Proteus rettgeri***. *Journal of General Microbiology* 1972, **72**(3):543-552.
137. Bradley DE, Taylor DE, Cohen DR: **Specification of surface mating systems among conjugative drug resistance plasmids in *Escherichia coli* K-12**. *Journal of Bacteriology* 1980, **143**(3):1466-1470.
138. Okumura MS, Kado CI: **The region essential for efficient autonomous replication of pSa in *Escherichia coli***. *Molecular & General Genetics : MGG* 1992, **235**(1):55-63.

139. Van der Auwera GA, Krol JE, Suzuki H, Foster B, Van Houdt R, Brown CJ, Mergeay M, Top EM: **Plasmids captured in *C. metallidurans* CH34: defining the PromA family of broad-host-range plasmids**. *Antonie van Leeuwenhoek* 2009, **96**(2):193-204.
140. Petersen J, Brinkmann H, Berger M, Brinkhoff T, Pauker O, Pradella S: **Origin and evolution of a novel DnaA-like plasmid replication type in *Rhodobacterales***. *Molecular Biology and Evolution* 2011, **28**(3):1229-1240.
141. Thomas CM: **The horizontal gene pool: bacterial plasmids and gene spread**. Amsterdam Harwood Academic Publishers 2000.
142. Nordstrom K, Austin SJ: **Mechanisms that contribute to the stable segregation of plasmids**. *Annual Review of genetics* 1989, **23**:37-69.
143. Gerdes K, Moller-Jensen J, Bugge Jensen R: **Plasmid and chromosome partitioning: surprises from phylogeny**. *Mol Microbiol* 2000, **37**(3):455-466.
144. Moller-Jensen J, Jensen RB, Gerdes K: **Plasmid and chromosome segregation in prokaryotes**. *Trends in Microbiology* 2000, **8**(7):313-320.
145. Hayes F, Barilla D: **The bacterial segrosome: a dynamic nucleoprotein machine for DNA trafficking and segregation**. *Nature Reviews, Microbiology* 2006, **4**:133-143.
146. Hale TL, Sansonetti PJ, Schad PA, Austin S, Formal SB: **Characterization of virulence plasmids and plasmid-associated outer membrane proteins in *Shigella flexneri, Shigella sonnei,* and *Escherichia coli***. *Infection and Immunity* 1983, **40**(1):340-350.
147. Surtees JA, Funnell BE: **Plasmid and chromosome traffic control: how ParA and ParB drive partition**. *Current Topics in Developmental Biology* 2003, **56**:145-180.
148. Fothergill TJ, Barilla D, Hayes F: **Protein diversity confers specificity in plasmid segregation**. *Journal of Bacteriology* 2005, **187**(8):2651-2661.
149. Austin S, Nordstrom K: **Partition-mediated incompatibility of bacterial plasmids**. *Cell* 1990, **60**(3):351-354.
150. Kalnin K, Stegalkina S, Yarmolinsky M: **pTAR-encoded proteins in plasmid partitioning**. *Journal of Bacteriology* 2000, **182**(7):1889-1894.
151. Bouet JY, Rech J, Egloff S, Biek DP, Lane D: **Probing plasmid partition with centromere-based incompatibility**. *Molecular Microbiology* 2005, **55**(2):511-525.
152. Bouet JY, Nordstrom K, Lane D: **Plasmid partition and incompatibility--the focus shifts**. *Mol Microbiol* 2007, **65**(6):1405-1414.
153. Salje J, Gayathri P, Lowe J: **The ParMRC system: molecular mechanisms of plasmid segregation by actin-like filaments**. *Nature Reviews Microbiology* 2010, **8**(10):683-692.
154. Simpson AE, Skurray RA, Firth N: **A single gene on the staphylococcal multiresistance plasmid pSK1 encodes a novel partitioning system**. *Journal of Bacteriology* 2003, **185**(7):2143-2152.
155. Caggiula AR, Epstein LH, Perkins KA, Saylor S: **Different methods of assessing nicotine-induced antinociception may engage different neural mechanisms**. *Psychopharmacology* 1995, **122**(3):301-306.
156. Siddique A, Figurski DH: **Different phenotypes of Walker-like A box mutants of ParA homolog IncC of broad-host-range IncP plasmids**. *Plasmid* 2012, **68**(2):93-104.

157. Rowe BP, Saylor DL, Speth RC, Absher DR: **Angiotensin-(1-7) binding at angiotensin II receptors in the rat brain**. *Regulatory Peptides* 1995, **56**(2-3):139-146.
158. Crimmins DL, Saylor M, Rush J, Thoma RS: **Facile, *in situ* matrix-assisted laser desorption ionization-mass spectrometry analysis and assignment of disulfide pairings in heteropeptide molecules**. *Analytical Biochemistry* 1995, **226**(2):355-361.
159. Turner SL, Rigottier-Gois L, Power RS, Amarger N, Young JP: **Diversity of *repC* plasmid-replication sequences in *Rhizobium leguminosarum***. *Microbiology* 1996, **142 (Pt 7)**:1705-1713.
160. Moller-Jensen J, Borch J, Dam M, Jensen RB, Roepstorff P, Gerdes K: **Bacterial mitosis: ParM of plasmid R1 moves plasmid DNA by an actin-like insertional polymerization mechanism**. *Molecular Cell* 2003, **12**(6):1477-1487.
161. Mohl DA, Easter J, Jr., Gober JW: **The chromosome partitioning protein, ParB, is required for cytokinesis in *Caulobacter crescentus***. *Molecular Microbiology* 2001, **42**(3):741-755.
162. Larsen RA, Cusumano C, Fujioka A, Lim-Fong G, Patterson P, Pogliano J: **Treadmilling of a prokaryotic tubulin-like protein, TubZ, required for plasmid stability in *Bacillus thuringiensis***. *Genes & Development* 2007, **21**(11):1340-1352.
163. Nordstrom K, Molin S, Aagaard-Hansen H: **Partitioning of plasmid R1 in *Escherichia coli*. II. Incompatibility properties of the partitioning system**. *Plasmid* 1980, **4**(3):332-339.
164. Dam M, Gerdes K: **Partitioning of plasmid R1. Ten direct repeats flanking the *parA* promoter constitute a centromere-like partition site *parC*, that expresses incompatibility**. *Journal of Molecular Biology* 1994, **236**(5):1289-1298.
165. Westergren A, Norberg E, Hagell P: **Diagnostic performance of the Minimal Eating Observation and Nutrition Form - Version II (MEONF-II) and Nutritional Risk Screening 2002 (NRS 2002) among hospital inpatients - a cross-sectional study**. *BMC Nursing* 2011, **10**:24.
166. Aylett CH, Wang Q, Michie KA, Amos LA, Lowe J: **Filament structure of bacterial tubulin homologue TubZ**. *Proc Natl Acad Sci 2*010, **107**(46):19766-19771.
167. Errington J: ***Bacillus subtilis* sporulation: regulation of gene expression and control of morphogenesis**. *Microbiological Reviews* 1993, **57**(1):1-33.
168. Seavers PR, Lewis RJ, Brannigan JA, Wilkinson JA: **Crystallization and preliminary X-ray analysis of the sporulation factor SpoIIAA in its native and phosphorylated forms**. *Acta Crystallographica Section D, Biological Crystallography* 2001, **57**(Pt 2):292-295.
169. Stragier P, Losick R: **Molecular genetics of sporulation in *Bacillus subtilis***. *Annual Review of Genetics* 1996, **30**:297-241.
170. Errington J, Murray H, Wu LJ: **Diversity and redundancy in bacterial chromosome segregation mechanisms**. *Phil Trans R Soc Biological sciences* 2005, **360**(1455):497-505.
171. Siddique A, Figurski DH: **The active partition gene *incC* of IncP plasmids is required for stable maintenance in a broad range of hosts**. *Journal of Bacteriology* 2002, **184**(6):1788-1793.

172. Batt SM, Bingle LE, Dafforn TR, Thomas CM: **Bacterial genome partitioning: N-terminal domain of IncC protein encoded by broad-host-range plasmid RK2 modulates oligomerisation and DNA binding**. *Journal of Molecular Biology* 2009, **385**(5):1361-1374.
173. Thomas CM: **Paradigms of plasmid organization**. *Molecular Microbiology* 2000, **37**(3):485-491.
174. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2011, **39**(Database issue):D38-51.
175. Egan ES, Duigou S, Waldor MK: **Autorepression of RctB, an initiator of *Vibrio cholerae* chromosome II replication**. *Journal of Bacteriology* 2006, **188**(2):789-793.
176. Diaz-Lopez T, Lages-Gonzalo M, Serrano-Lopez A, Alfonso C, Rivas G, Diaz-Orejas R, Giraldo R: **Structural changes in RepA, a plasmid replication initiator, upon binding to origin DNA**. *The Journal of Biological Chemistry* 2003, **278**(20):18606-18616.
177. Gerdes K, Thisted T, Martinussen J: **Mechanism of post-segregational killing by the *hok/sok* system of plasmid R1: *sok* antisense RNA regulates formation of a *hok* mRNA species correlated with killing of plasmid-free cells**. *Molecular Microbiology* 1990, **4**(11):1807-1818.
178. LeBard RJ, Jensen SO, Arnaiz IA, Skurray RA, Firth N: **A multimer resolution system contributes to segregational stability of the prototypical staphylococcal conjugative multiresistance plasmid pSK41**. *FEMS Microbiology Letters* 2008, **284**(1):58-67.
179. Castillo-Ramírez S, Vázquez-Castellanos JF, González Vc, Cevallos MA: **Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in Alphaproteobacteria: the *repABC* operon**. *BMC Genomics* 2009, **10**:536-547.
180. Mazur A, Majewska B, Stasiak G, Wielbo J, Skorupska A: ***repABC*-based replication systems of *Rhizobium leguminosarum* bv. trifolii TA1 plasmids: incompatibility and evolutionary analyses**. *Plasmid* 2011, **66**(2):53-66.
181. Cervantes-Rivera R, Pedraza-López F, Pérez-Segura G, Cevallos MA: **The replication origin of a *repABC* plasmid**. *BMC Microbiology* 2011, **11**(158).
182. Gouy M, Guindon S, Gascuel O: **SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building**. *Molecular Biology and Evolution* 2010, **27**(2):221-224.
183. Toussaint A, Merlin C: **Mobile elements as a combination of functional modules**. *Plasmid* 2002, **47**(1):26-35.
184. Chu SF, Shu HY, Lin LC, Chen MY, Tsay SS, Lin GH: **Characterization of a rolling-circle replication plasmid from *Thermus aquaticus* NTU103**. *Plasmid* 2006, **56**(1):46-52.
185. Friehs K: **Plasmid copy number and plasmid stability**. *Advances in Biochemical Engineering/Biotechnology* 2004, **86**:47-82.
186. Mikesell P, Ivins BE, Ristroph JD, Dreier TM: **Evidence for plasmid-mediated toxin production in *Bacillus anthracis***. *Infection and Immunity* 1983, **39**(1):371-376.

187. Bennett PM: **Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria**. *British Journal of Pharmacology* 2008, **153 Suppl 1**:S347-357.
188. Brown AMC, Willetts NS: **A Physical and Genetic Map of the IncN Plasmid R46**. *Plasmid* 1980, **5**:188-201.