Are bacterial species really ecotypes?

Nitin Kumar

PhD

University of York

Biology

December 2013

Abstract

Defining bacterial species is still a debatable topic among bacteriologists. It has become clear that, periodic selection and recombination are two main drivers of bacterial species.

Here, we are curious to study the diversity and structure of a local population. For which, we studied a population that was comprised of two symbiovars of *R. leguminosarum*. The draft genomes of 72 isolates (36 *viciae* & 36 *trifolii*) from a square meter of soil were sequenced by Roche 454 sequencing and compared with the published genome of *Rlv* 3841. Chapter Two employs 305 core genes and genomic analysis to demonstrate the existence of five phenotypically indistinguishable (cryptic) genospecies in a local population. Most of the cryptic genospecies include both *viciae* and *trifolii* strains: the genospecies do not reflect the symbiovars. Chapter Three demonstrates that recombination plays a major role in shaping the chromosome of *R. leguminosarum*. Moreover, it demonstrates the preference of intra- genospecies recombination highlighting the occurrence of genetic isolation between genospecies that allows them to be represented as biological species. Chapter Four demonstrates the presence of core genes on different plasmids. The phylogenetic structure of *Rlv* 3841 replicons resembles the structure of core genes phylogeny indicating lack of genes transfer between genospecies in each replicon. However, the phylogenetic networks suggest horizontal transfer of *nod* genes that allow species members to have different host specificity. Chapter Five displays the genetic diversity present between two major genospecies (B and C) of *R. leguminosarum.*

Overall, our results provide direct evidence that core genes and genomic analysis such as ANI should be used to define bacterial species. Moreover, the host specific symbiotic genes are normal accessory genes that have no significant role in the demarcation of bacterial species.

Table of Contents

Abstract ..... **I**

Table of Contents ..... **II**

List of Figures ..... **VI**

List of Tables ..... **X**

Acknowledgements ..... **XIII**

Author's Declaration ..... **XIV**

Chapter 1. Introduction **1**

1.1. Bacterial genomes **2**

1.1.1. Chromosomes 2

1.1.2. Chromids 3

1.1.3. Plasmids 4

1.2. Gene exchange **4**

1.3. Bacterial species **8**

1.3.1. The Polyphasic concept 8

1.3.2. Role of the multilocus sequencing approach 9

1.3.3. Core Genome Hypothesis (CGH) 11

1.3.4. Revolution in the Recombination model 12

1.3.5. The Ecotype concept: 14

1.3.6. Pangenome 15

1.4. Rhizobia and their taxonomy **16**

1.4.1. Symbiovars 17

1.4.2. *R. leguminosarum* species complex 17

1.4.3. Nodulation (*nod*) genes 18

1.5. Aims and objectives **19**

Chapter 2. Phylogenomic analysis reveals cryptic genospecies in a local population of *Rhizobium leguminosarum*. **21**

2.1 Abstract **21**

2.2 Introduction **22**

2.2.1 Objectives 26
2.3 Material and methods 27
2.3.1 Sequence data 27
2.3.2 GS De Novo Assembler 29
2.3.3 GS Reference Mapper 29
2.3.4 Multiple sequence alignment 30
2.3.5 Phylogenetic analysis 30
2.3.6 Average Nucleotide Identity analysis 30
2.3.7 Computational Resources 31
2.4 Results 32
2.4.1 Genome sequencing 32
2.4.2 Phylogenetic analysis of local population of *R. leguminosarum* 35
2.4.3 Global Phylogeny of the *R. leguminosarum* species complex 38
2.4.4 Average Nucleotide Identity analysis 41
2.4.5 Phylogenetic analysis of Swedish and Scottish strains 43
2.5 Discussion 45
2.5.1 Cryptic genospecies in a local population 45
2.5.2 Phylogenetic structure of the *R. leguminosarum* species complex 46
2.5.3 Cosmopolitan nature of cryptic genospecies 46

Chapter 3. Recombination and population structure of cryptic genospecies of *R. leguminosarum* 48
3.1 Abstract 48
3.2 Introduction 49
3.2.1 Objectives 51
3.3 Material and Methods 52
3.4 Results 56
3.4.1 Phylogenetic structure based on reliable core gene 56
3.4.2 Phylogenetic incongruence and Intragenic recombination 56
3.4.3 Intra- and Inter- genospecies recombination 70
3.5 Discussion 78

Chapter 4. Dominating influence of five genospecies on the composition and phylogeny of the accessory genome of *R. leguminosarum* 80
4.1 Abstract 80

4.2 Introduction 81
4.2.1 Objectives 85
4.3 Materials and Methods 86
4.3.1 Construction of presence/absence matrix based on *Rlv* 3841 replicons 86
4.3.2 Construction of phylogenetic networks based on reference replicons 86
4.3.3 Construction of phylogenetic networks based on nod genes 87
4.3.4 Population specific genes 89
4.4 Results 90
4.4.1 Presence/Absence analysis: 90
4.4.2 Phylogenetic analysis of reference replicons 96
4.4.3 Phylogenetic analysis of nodulation (*nod*) genes 102
4.4.4 Population specific genes 106
4.5 Discussion 109
4.5.1 Presence/absence matrices 109
4.5.2 Phylogenetic analysis of reference replicons 110
4.5.3 Phylogenetic analysis of nodulation (*nod*) genes 111
4.5.4 Population specific genes 112

Chapter 5. Comparative genomics of two major genospecies of *R. leguminosarum* 114

5.1 Abstract 114
5.2 Introduction 115
5.2.1 Objectives 116
5.3 Material and Methods 117
5.4 Results 118
5.4.1 De novo assembly of TRX_6 118
5.4.2 Reference-based assembly of the TRX_6 genome with 3841 118
5.4.3 Comparative analysis of TRX_6 with 3841 using CGView and MUMmer 119
5.4.4 Correct assembly of TRX_6 chromosomal related scaffolds using CONTIGuator 129
5.4.5 Automatic annotation of TRX_6 genome 131
5.5 Discussion 134

Chapter 6. General Discussion 137

6.1 Synopsis **138**
6.2 Direction for future research **141**
6.3 Conclusions **142**
Appendix I **144**
Appendix II **163**
Appendix III **165**
References **174**

List of Figures

Chapter 1: Introduction

Figure 1.1 | Mechanisms of gene transfer between bacteria.. 7

Figure 1.2 | Two fuzzy species (A and B) in *Neisseria*.....10

Figure 1.3 | Distribution of selected rhizobia (bold font) in different classes of proteobacteria (α and β classes) in 16S rDNA sequence phylogeny. 16

Chapter 2: Phylogenomic analysis reveals cryptic genospecies in a local population of *Rhizobium leguminosarum*.

Figure 2.1 | Chromosomal network showing divergence among *S. medicae* strains.....23

Figure 2.2 | Genomic structure of *Rlv* 3841 24

Figure 2.3 | Sequence coverage of 36 *viciae* (VSX) strains.....32

Figure 2.4 | Sequence coverage of 36 *trifolii* (TRX) strains.. 33

Figure 2.5 | Sequence coverage of 4 type strains. 34

Figure 2.6 | Sequence coverage of 4 Swedish (VCS_1,TPS_1,VCS_2, VCS_3,VCS_4, VCS_5, TPS_5) and 2 Scottish (VCS_6, TPS_6) strains. 35

Figure 2.7 | Neighbour-net phylogeny of *R. leguminosarum* strains based on concatenated alignment of 305 core genes.....36

Figure 2.8 | Neighbour-net phylogeny of *R. leguminosarum* species complex based on concatenated alignment of 305 core genes.....39

Figure 2.9 | Maximum Likelihood tree of *R. leguminosarum* strains based on concatenated alignment of 305 core genes.....40

Figure 2.10 | Neighbour-net phylogeny of *R. leguminosarum* strains based on concatenated alignment of 305 core genes.....44

Chapter 3: Recombination and population structure of cryptic genospecies of *R. leguminosarum*.

Figure 3.1 | Maximum Likelihood tree based on 100-gene alignment showing the position of 75 *R. leguminosarum* strains. 58

Figure 3.2 | Heatmap showing SH test results of 49 core genes that are highly recombinant.. 62

Figure 3.3 | Heatmap showing the hierarchical clustering (row and column) of SH test results of 49 core genes to obtain similarity in patterns of P values. 63

Figure 3.4 | The Maximum Likelihood tree of RL2957 66

Figure 3.5 | The Maximum Likelihood tree of RL0377 67

Figure 3.6 | The Maximum Likelihood tree of RL2381 68

Figure 3.7 | The Maximum Likelihood tree of RL1551 69

Figure 3.8 | Results of Clanistic analysis. 71

Figure 3.9 | ΔK-values for different K; suggesting K = 5 as the most likely structure population according to Evanno et al. (2005) 74

Figure 3.10 | Bar graph showing results of Structure analysis with K = 5.. 75

Figure 3.11 | Bar graph showing results of Structure analysis with K = 6.. 76

Figure 3.12 | Bar graph showing results of Structure analysis with K = 7.. 77

Chapter 4: Dominating influence of five genospecies on the composition and phylogeny of the accessory genome of *R. leguminosarum*.

Figure 4.1 | Neighbour-Nets showing divergence among *S. medicae* strains for each replicon.....82

Figure 4.2 | The phylogenetic tree based on all RepABC replicons in 72 *R. leguminosarum* strains. 84

Figure 4.3 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 chromosomal genes.. 92

Figure 4.4 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 chromid genes (A: pRL12, B: pRL11).. 93

Figure 4.5 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 large plasmid genes (A: pRL10, B: pRL9).. 94

Figure 4.6 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 small plasmid genes (A: pRL8, B: pRL7).. 95

Figure 4.7 | Phylogenetic network obtained from the chromosomal genes of *Rlv* 3841 for 72 *R. leguminosarum* strains.. 98

Figure 4.8 | Phylogenetic network obtained from the two chromid genes (A: pRL12, B: pRL11) of *Rlv* 3841 for 72 *R. leguminosarum* strains. 99

Figure 4.9 | Phylogenetic network obtained from two large plasmids (A: pRL10, B: pRL9) of *Rlv* 3841 for 72 *R. leguminosarum* strains.. 100

Figure 4.10 | Phylogenetic network obtained from two small plasmids (A: pRL8, B: pRL7) of *Rlv* 3841 for 72 *R. leguminosarum* strains.. 101

Figure 4.11 | Phylogenetic network obtained from 11 *nod* genes of *Rlv* 3841 and *Rlt* WSM1325. 104

Figure 4.12 | Phylogenetic network obtained from 11 *nod* genes of *Rlv* 3841 for 36 *R. leguminosarum viciae* strains.. 105

Figure 4.13 | Phylogenetic network obtained from 11 *nod* genes of *Rlt* WSM1325 for 36 *R. leguminosarum trifolii* strains.. 105

Figure 4.14 | The direct relationship between the number of observed specfic genes and sequence coverage in each of 72 *R. leguminosarum* strains. 107

Figure 4.15 | Heatmap of population specific gene presence (blue) and absence (white) in 72 *R. leguminosarum* strains and absent in *Rlv* 3841.. 108

Chapter 5: Comparative genomics of two major genospecies of *R. leguminosarum*.

Figure 5.1 | The circular maps of Scaffolds 1 (A) and 2 (B) of TRX_6.. 121

Figure 5.2 | The circular maps of Scaffolds 3 (A) and 4 (B) of TRX_6.. 122

Figure 5.3 | The circular maps of Scaffolds 5 (A) and 6 (B) of TRX_6.. 123

Figure 5.4 | The circular maps of Scaffolds 7 (A) and 8 (B) of TRX_6.. 124

Figure 5.5 | The circular maps of Scaffolds 9, 10 and 11 of TRX_6.. 125

Figure 5.6 | The Nucmer plots of TRX_6 scaffolds and 3841 replicons highlighting synteny relationship between them.. 128

Figure 5.7 | The Nucmer plots of TRX_6 and 3841 replicons highlighting synteny relationship between them and correct ordering of chromosomal related scaffolds.....130

Figure 5.8 | The functional subsystems present in TRX_6 determined by the RAST server. 132

Figure 5.9 | The twelve *nod* genes of TRX_6 (symbiovar trifolii) are located in scaffold 7. 133

Appendix II

Figure II.I: Maximum Likelihood tree based on 100-gene alignment showing the position of 75 *R. leguminosarum* strains.. 164

Appendix III

Figure III.I: Phylogenetic network obtained from a small plasmid (pRL7) of *Rlv* 3841 for 72 *R. leguminosarum* strains. 166

List of Tables

Chapter 2: Phylogenomic analysis reveals cryptic genospecies in a local population of *Rhizobium leguminosarum*.

Table 2.1 | Fully sequenced genomes of *Rhizobium* genus used in this study other than *Rlv* 3841 28

Table 2.2 | Type strains sequenced in this study 28

Table 2.3 | Other *R. leguminosarum* strains sequenced in this study 29

Table 2.4 | List of strains in the five clusters in 72 *R. leguminosarum* based on the 305 core genes 37

Table 2.5 | Average nucleotide identity (ANIm) analysis of selected 72 *R. leguminosarum* strains. 42

Table 2.6 | Average nucleotide identity (ANIm) analysis of Swedish strains, TRX_34 (genospecies A) and USDA 2370^{T} 43

Chapter 3: Recombination and population structure of cryptic genospecies of *R. leguminosarum*.

Table 3.1 | Models of nucleotide substitution for each of 100 core genes 53

Table 3.2 | Results from SH tests and PHI tests for 100 core genes... 59

Table 3.3 | Results of Clanistic analysis of five genospecies (gs) 70

Table 3.4 | Results of ClonalFrame analysis 72

Chapter 4: Dominating influence of five genospecies on the composition and phylogeny of the accessory genome of *R. leguminosarum*.

Table 4.1 | The 13 *nod* genes and their position in pRL10 of *Rlv* 3841 88

Table 4.2 | The 11 *nod* genes and their position in pR132501 of *Rlt* WSM1325. 88

Table 4.3 | Bvs (Symbiovar *viciae* specific) genes and their location in pRL8. 91

Table 4.4 | Genes responsible for one of the conjugative systems of pRL7. 97

Table 4.5 | Specific genes present in each genospecies (B-E) 107

Chapter 5: Comparative genomics of two major genospecies of *R. leguminosarum*.

Table 5.1 | The TRX_6 scaffolds with their length and related number of contigs. 118

Table 5.2 | Percentage coverage of 3841 replicons with TRX_6 genome. 119

Table 5.3 | Description of genes of pRL10, pRL11 and pRL12 that are allocated in the homologous region of scaffolds 3, 4, 5 and 7. 126

Table 5.4 | Eleven scaffolds of TRX_6 are arranged on the basis of their similarity with 3841 replicons. 127

Appendix I

Table I.I | The 305 core genes held by all chromid-possessing bacteria. 145

Appendix III

Table III.I | The genospecies B specific-island in Chromid (pRL12). 167

Table III.II | The first genospecies B specific-island in large plasmid (pRL9). 170

Table III.III | The second genospecies B specific-island in pRL9.. 171

Table III.IV | The third genospecies B specific-island in pRL9.. 172

Table III.V | The fourth genospecies B specific-island in pRL9.. 173

Acknowledgements

First of all, I would like to thank my supervisor Prof. Peter Young for his support and guidance throughout my doctorate. I am grateful for his motivation, patience and precious time that he devoted for this PhD thesis. Besides my advisor, I would like to acknowledge my thesis advisory panel, Dr. Peter Ashton and Dr. Gavin H. Thomas for their scientific advice and questions during each meeting.

Next, I must thank all the past and present members of my lab (J1). It has been a pleasure to work in such a culturally diverse group. I would like to express my deep gratitude and respect to Stuart Priest for IT support. Also thanks to the Radhika V. Sreedhar Memorial Scholarship Fund that provided further support during my research.

Finally, I would like to thank my family, relatives and friends for their encouragement and interest in my work. To my parents, Dr. Devender Kumar and Dr. Kiran, I am indebted to you for all your unconditional adoration and support throughout my life. I also express my deep thanks to my brother (Dr. Saurabh Kumar), sister in law (Dr. Shikha) and nephew (Mr. Vedant Kumar) for their laughter and support.

Author's Declaration

I hereby declare and confirm that the results presented in this thesis are my own original work and have not been submitted elsewhere for examination. This work has contributed to the following publications:

KUMAR, N., LAD, G., GIUNTINI, E., KAYE, M. E., UDOMWONG, P., SHAMSANI, N. J., YOUNG, J. P. & BAILLY, X. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biol,* 5.

Chapter 1. Introduction

"Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups" – Mayr, 1942: 120.

This widely accepted species definition works well for eukaryotic organisms such as humans that exchange genetic material to reproduce, but does not define the species of asexual organisms such as bacteria. For the last thirty years, the standard concept for delineating a bacterial population into species is the phenotypic and genotypic similarity shared by members of the same species. This pragmatic "polyphasic" concept has been applied by many studies, but is not necessarily compatible with either of two major theoretical concepts: ecological and recombination. In the "ecological concept", bacterial species consist of ecologically distinct populations driven by periodic selection (Atwood et al., 1951), whereas the cohesive force of recombination maintains species structure in the "recombination concept", mirroring Mayr's biological species concept. Although the recombination model is useful to define species in *Helicobacter pylori* (Falush et al., 2003), *Neisseria* (Hanage et al., 2005), *Wolbachia* (Ellegaard et al., 2013), the ecological concept has attracted the attention of bacteriologists because sequence clusters display strong correlations with ecological species in, for example, recent studies of *Agrobacterium tumefaciens* (Lassalle et al., 2011) and *Bacillus* (Connor et al., 2010).

Shapiro et al. (2012) demonstrated the occurrence of a selective sweep, with no recombination, in a recently diverged sympatric subpopulation of the bacterium *Vibrio cyclotrophicus*. Both ecological differentiation and genetic isolation (recombination barriers between species) has been reported for this recent speciation. Another interesting study of recent speciation (Cadillo-Quiroz et al., 2012) involves the sympatric population of the thermoacidophilic archaeon *Sulpholobus islandicus*, which has recently diverged into two incipient species that are maintained by ecological differentiation alone and lack interspecies genetic barriers.

Here, we propose that sequence clusters of sympatric isolates may have no relationship with historically defined ecological species. Also, recombination may play a significant role in the delineation of bacterial species that will accords with Mayr's definition of species.

The study begins with this introductory chapter that will introduce the distinctive features of bacterial genomes, gene transfer (homologous recombination and horizontal gene transfer), major species concepts, and *Rhizobium leguminosarum* as a model species. Subsequent chapters will include comprehensive analysis of coherent genomic clusters (genospecies) that are present in a local population.

1.1. Bacterial genomes

Fleischmann et al. (1995) sequenced the first bacterial genome (*Haemophilus influenzae*) and since then multiple bacterial genomes have been sequenced and observed. Modern developments in sequencing technologies and computational biology have greatly facilitated genome sequencing. Among the bacterial genomes sequenced so far, genome sizes range from 112 Kb for *Nasuia deltocephalinicola* (Bennett and Moran, 2013), an obligate bacterial symbiont that lives in phloem-feeding insects, to 13.66 Mb for *Ktedonobacter racemifer* (Chang et al., 2011a), a heterotrophic soil bacterium. The genetic information in bacteria is carried on chromosomes, chromids, and plasmids.

1.1.1. Chromosomes

Unlike linear eukaryotic chromosomes, the typical bacterial chromosome is a circular replicon with no free ends that is not separated by a nuclear membrane. The bacterial chromosome harbors a set of essential (core) genes encoding essential metabolic and informational functions. The majority of bacterial genes is protein-coding and syntenically conserved, but might be located at different positions or replicons (Bentley and Parkhill, 2004).

The genome size and GC (percentage of guanine-cytosine) content in bacteria are correlated (Bentley and Parkhill, 2004; Nishida, 2012b; Wu et al., 2012), as high GC content tends to be associated with larger genomes and low GC content (AT-richness) tends to be associated with smaller genomes (Bentley and Parkhill, 2004; Nishida, 2012a). There are two possible explanations for this hypothesis. First, the fact that synthesis of GTP and CTP nucleotides requires more energy than ATP and UTP, which could drive bacteria with small genome to shift toward AT-richness under limited resources (Rocha and Danchin, 2002). Second, absence of DNA repair genes in small genomes makes them more AT-rich (Moran, 2002). However, we must note that results reported in (Bentley and Parkhill, 2004; Nishida, 2012a) is a correlation found in most, but not all, bacteria such as *Candidatus Hodgkinia cicadicola* (144kb; 58.4% GC; McCutcheon et al., 2009) and *Candidatus Tremblaya princeps* (139kb; 58.8%; Lopez-Madrigal et al., 2011) have GC-rich genome. The GC content in the chromosome is somewhat higher than that of the accompanying replicons. In order to identify the initial and end points of a chromosome, a measure known as a GC ((G-C)/(G+C)) skew is implemented (Bentley and Parkhill, 2004). A positive value of GC skew reflects the leading strand, whereas a negative value reflects lagging strands, so these values switch at the site of an origin or terminus of replication.

Generally, a bacterial genome includes other replicons (chromids and plasmids) in addition to the chromosome, but some bacteria carry only a chromosome, for example, *Streptococcus pneumoniae* (Tettelin et al., 2001) and *Porphyromonas gingivalis* (Nelson et al., 2003).

1.1.2. Chromids

By performing systematic analysis, Harrison et al. (2010) suggested a convenient name for replicons that were previously known as ‘megaplasmids’ or ‘secondary chromosomes’. The authors used three main criteria to define chromids and discriminate them from the main chromosome and plasmids: A. Chromids should be based on plasmid type replication and maintenance systems. B. Nucleotide composition of chromids should be similar to the main chromosome. C. Chromids should carry a set of essential (core) genes that are conserved in the chromosome of other species. The

first chromid was observed in the genome of alphaproteobacterium *Rhodobacter sphaeroides* 2.4.1 (Suwanto and Kaplan, 1989a, Suwanto and Kaplan, 1989b), but described as a "secondary chromosome". Since then, many genomes have been shown to have these characteristics. The size and composition of chromids varies in different bacteria. At the time of writing, the maximum number of chromids belongs to the *Azospirillum lipoferum* genome that harbors five chromids (Wisniewski-Dye et al., 2011). Another interesting property of chromids is the presence of genes that are conserved in a particular genus, for example, Harrison et al. (2010) suggested that the chromids of all sequenced *Burkholderia* genomes have a set of conserved genes that are not conserved in the chromids of other bacteria.

1.1.3. Plasmids

Plasmids are the smallest functional replicons that can replicate independently from the primary chromosome (Frost et al., 2005). Generally, plasmids have low GC content and carry non-conserved 'accessory' genes that are critical ecological determinants (Prosser et al., 2007) and provide genomic flexibility to acquire advantageous properties such as virulence and multi-drug resistance. Due to HGT, bacterial strains often show variation in number and size of plasmid. However, some conserved genes, known as 'backbone genes', are also located in these replicons such as the replication system (*repABC*), which is ubiquitous in all plasmid possessing alpha-proteobacteria (Cevallos et al., 2008).

1.2. Gene exchange

The most common form of reproduction in bacteria is binary fission, but bacterial genomes can also exchange genetic material in ways that may play a role analogous to that of sexual reproduction in eukaryotes. Unlike eukaryotes, bacterial genetic exchange is unidirectional from donor to recipient cell. There are two forms of gene exchange:

1. Homologous recombination: This primarily occurs between strains of the same species that possess high nucleotide identity (Redfield, 2001). Instead of the

whole chromosome, a fragment of an advantageous part or gene is migrated from donor to recipient. Because this process involves transfer and integration of homologous DNA, core genes (shared by all isolates) are more affected by homologous recombination than accessory (variable) genes. Homologous recombination can play a crucial role in adaptation of some bacteria, for example, *Streptococcus pneumoniae* (Donati et al., 2010) and *Helicobacter pylori* (Falush et al., 2003).

2. Horizontal Gene Transfer (HGT): HGT is a process by which a bacterium can acquire non-homologous genetic material (DNA) from the same or different species and is often known as lateral gene transfer (Goldenfeld and Woese, 2007). This mechanism is the primary method that helps the host to acquire beneficial genes that encode specific properties such as antibiotic resistance or virulence (Furuya and Lowy, 2006) and is a key factor in bacterial evolution. Because this process involves transfer of nonhomologous DNA, accessory genes (present in some isolates) are more affected by HGT than core (conserved) genes that are generally maintained by homologous recombination. Analysis of different species (Narra and Ochman, 2006) has revealed that there is no clear relationship between the degree of HGT and homologous recombination by which genes within a species can be re-assorted. Moreover, the genetic diversity introduced by HGT is different from the one produced by point mutation that involves alteration of existing genes. Acquisition of novel genetic material by HGT results in exploration of a new ecological niche. In contrast, accumulation of point mutations results in the modification of existing genes that may lead to niche expansion (Lawrence, 1999; Ochman et al., 2000). For example, Lawrence and Ochman (1998) provided evidence that phenotypic differentiation of two sister species, Escherichia coli and Salmonella enterica, can be explained by HGT, and not by point mutation. However, models of amelioration (Lawrence and Ochman, 1997) estimated that similar amount of variation were introduced through HGT and through point mutation.

These two key elements (homologous recombination and HGT) of gene transfer can be achieved by three different mechanisms: transformation, transduction, and conjugation.

Transformation (Figure 1.1a) allows the recipient cell to acquire naked DNA fragments and its genetic traits directly from the environment. The main source of naked DNA is dead cells. Integrated DNA can be used for genetic diversity (new metabolic function, traits such as virulence, antibiotic resistance), DNA repair or as a source of energy (Chen and Dubnau, 2004). The bacteria following the transformation procedure are considered naturally competent bacteria, for example, the *Haemophilus influenzae* and *Neisseria* species.

Transduction (Figure 1.1b) involves the movement of DNA fragments from a donor bacterium to a recipient bacterium with the use of bacterial virus or bacteriophage (Frost et al., 2005). It occurs when the bacteriophage contains fragments of bacterial DNA accidentally incorporated into its own DNA. Once this phage DNA is transferred into another bacterium, foreign DNA fragments integrate with the recipient bacterial genome. Both transformation and transduction require integration of DNA, hence HGT through these processes is limited to isolates of the same species.

Conjugation (Figure 1.1c) is the only process that requires cell-to-cell contact (Frost et al., 2005). In this process, a living donor bacterium transfers the plasmid or transposons (jumping genes) into the living recipient bacterium through a tube like structure. The efficiency of transferring plasmid between different species makes conjugation the primary reason for HGT in distantly related species. Conjugation can only be achieved by those plasmids that harbor conjugative or transfer (*tra*) genes responsible for a stable mating pair with the recipient genome and *oriT* (origin of transfer) sequences that initiate the transfer. These plasmids are known as conjugative plasmids and are self-transmissible, but plasmids that lack *tra* genes and consist of *oriT* sequences are known as mobilizable plasmids and are transferred with the help of self-transmissible conjugative plasmids.

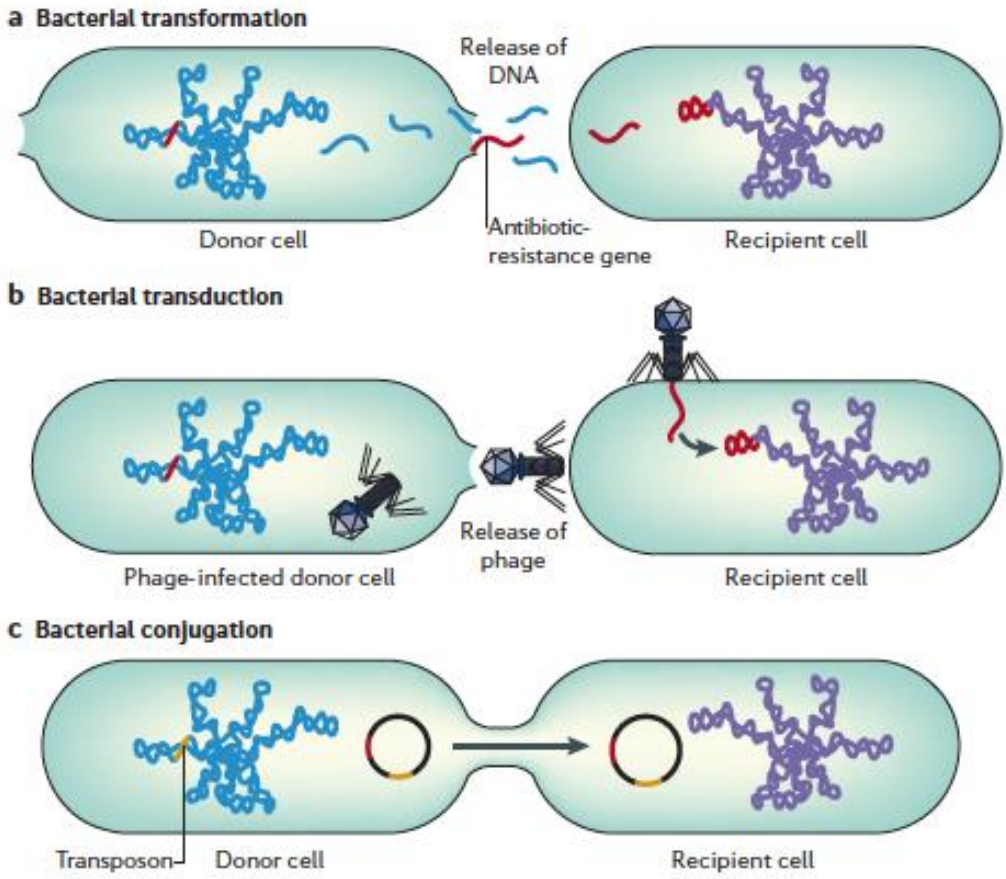


Figure 1.1 | Mechanisms of gene transfer between bacteria. (a) Transformation is the direct uptake of free DNA fragments. (b) Transduction involves fragment transfer through bacteriophage. (c) Conjugation: gene transfer by physical contact (Furuya and Lowy, 2006).

1.3. Bacterial species

Here, we outline some of the major concepts of bacterial species delineation.

1.3.1. The Polyphasic concept

For decades, bacteriologists have suggested a combination of phenotypic, genotypic and phylogenetic information, also known as the polyphasic taxonomy approach (Colwell, 1970; Stackebrandt et al., 2002; Vandamme et al., 1996), to discriminate bacterial species. According to this enhanced version of the numerical taxonomy approach (Sneath and Sokal, 1973), species members must have common phenotypic properties (morphological, biochemical), more than 70% DNA-DNA hybridization (DDH) (Wayne et al., 1987; Grimont, 1988) and similarity (>97%) in 16S ribosomal RNA (rRNA) sequences (Olsen and Woese, 1993; Stackebrandt and Goebel, 1994). Although DDH is a time-consuming process that is not suitable for unculturable bacteria and has an empirical cutoff (>70%), bacteriologists consider it the gold standard for delineating microbial species. Furthermore, the results of studies of new multilocus or whole genome data correspond with DDH.

Cost-effective sequencing and the conserved nature of 16S rRNA sequences (Stackebrandt and Goebel, 1994) facilitate demarcating bacterial species in culturable as well as unculturable bacteria. However, strains with more than 97% 16S rRNA sequences identity must satisfy the 70% DDH criterion to get allocation of the same species (Gevers et al., 2005). The major drawback of 16S rRNA sequences is the lack of resolution in large populations of closely related strains (Gevers et al., 2005; Hanage et al., 2006). Alternatively, Konstantinidis and Tiedje (2005) proposed a new method of calculating average nucleotide identity (ANI), which is more robust than 16S rRNA sequences identity. ANI produces genetic coherent groups by performing pairwise genome comparison of all shared orthologs between two strains. A threshold of 95-96% ANI (Auch et al., 2010; Goris et al., 2007; Konstantinidis and Tiedje, 2005; Richter and Rossello-Mora, 2009) corresponds to the classic threshold of 70% DDH.

1.3.2. Role of the multilocus sequencing approach

Any single bacterial gene, including the 16S rRNA gene, is subject to recombination (Boucher et al., 2004; Gogarten et al., 2002; Hanage et al., 2006). Thus, the multiple genes approach is a perfect alternative for 16S rRNA genes because it involves slowly-evolving housekeeping genes (not slower than 16S rRNA genes) or loci and concatenation of multiple housekeeping genes buffers the effects of recombination. Based on this hypothesis, Maiden et al. (1998) proposed multilocus sequence typing (MLST) that categorize the bacterial strains at the infraspecific level based on the pairwise allelic mismatches of housekeeping genes (usually 7). To classify the strains of similiar species, MLST was modified to multilocus sequence analysis (MLSA; Gevers et al., 2005). This extended version employs phylogenetic procedures based on the nucleotide sequences of multiple housekeeping genes instead of the number of allelic mismatches.

Overall, a simple MLSA involves the construction of a phylogenetic tree based on concatenated sequences of housekeeping or core genes that must be able to delineate the species in a genus. The robust method of MLSA has been supported by many species studies (Achtman and Wagner, 2008; Bishop et al., 2009).

Surprisingly, Hanage et al. (2005) observed ‘fuzzy species’ in the MLSA of genus *Neisseria* (Figure 1.2). These fuzzy species were composed of strains from different predefined *Neisseria* species. However, the remaining three predefined *Neisseria* species were fully resolved and consisted of strains of related species. Moreover, phylogenetic trees based on a single core gene were unable to resolve any of the five species. Afterwards, clusters with strains of different traditionally named species were also observed in many other bacteria such as *Bacillus cereus* (Priest et al., 2004) and *Helicobacter pylori* (Linz et al., 2007).

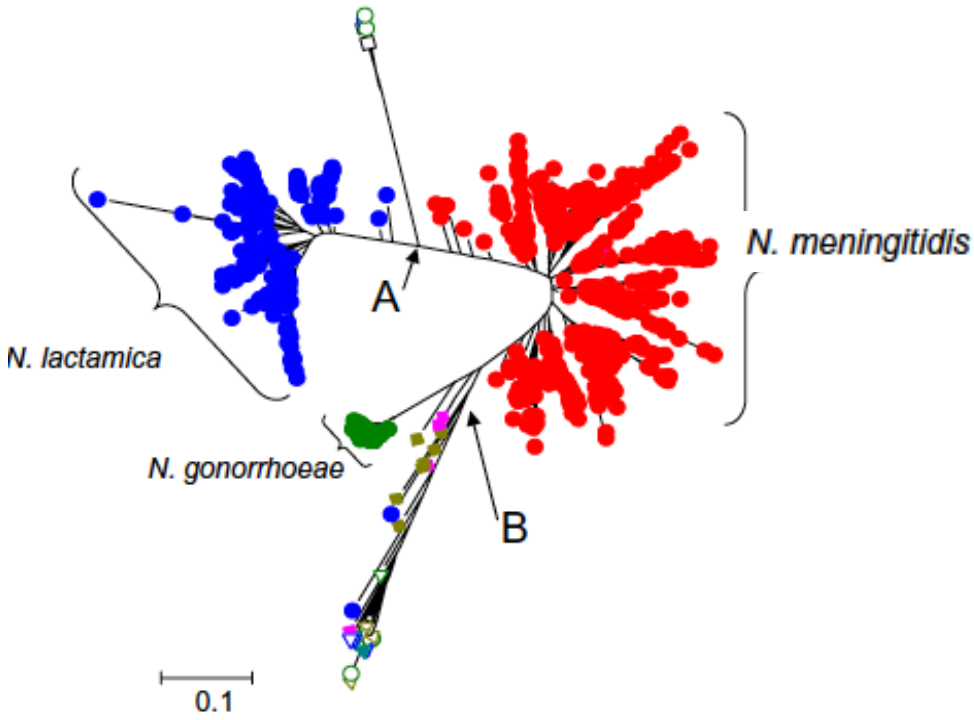


Figure 1.2 | Two fuzzy species (A and B) in *Neisseria*. Bayesian tree based on seven housekeeping genes showing three clusters of predefined species clusters (*N.meningitidis*: red, *N.lactamica*: blue and *N.gonorrhoeae*: green) and two distinct clusters (A and B) composed of different species strains in *Neisseria* genus (Bennett et al., 2007; Hanage et al., 2005).

1.3.3. Core Genome Hypothesis (CGH)

Lan and Reeves (1996) proposed a concept by refining the biological species concept for bacteria suggested by Dykhuizen and Green (1991). According to CGH, the bacterial genome is composed of two different sets of genes: core genes and accessory genes.

Core genes: These are housekeeping genes that encode essential functions and each member of a species possesses these genes. They are shared by the members of the same species via recombination and rarely transfer between species. Core gene divergence coevolves with species divergence and hinders the recombination between species. The characteristic of preferring intraspecific recombination in core genes is considered analogous to Mayr's definition, thus, core genes could be used as powerful phylogenetic markers to demarcate bacterial species.

Accessory genes: Apart from the core genes, another set of genes in the bacterial genome is accessory (non-essential) genes that may or may not be present in a member of a species. They provide genomic plasticity to the genome to survive in different ecological niches or enhance bacterial properties such as virulence or toxicity. They are chiefly driven by HGT within or between different species. Accessory genes have lower GC content than core genes. The prime locations of these genes are genomic islands or plasmids. Comparison of phylogenies of accessory genes with core genes phylogenies can lead to the identification of recombination via conflicting phylogenies. For example, the accessory genome of two non inter-recombining populations of *V. cyclotrophicus* was shaped by HGT (Shapiro et al., 2012), as the phylogeny of the accessory genome contradicted the species phylogeny.

Using massive sequencing and bioinformatics tools, bacteriologists have applied the idea of core genes to define species in different bacteria. Recently, Chan et al. (2012) delineated the species of genus *Acinetobacter* using a phylogenetic tree based on 127 core genes and ANI analysis. The results reflected the traditional classification. However, highly conserved 16s RNA gene sequences were not able to resolve the accepted species of this genus. Recently, another tool, specI (Mende et al., 2013), was

published to delineate bacterial species based on forty universal genes (Ciccarelli et al., 2006 and Sorek et al., 2007) that are conserved in all three domains. We observed that these universal genes are included in the 305 core genes conserved in chromid-possessing bacteria that were identified by Harrison et al. (2010) and are used in this study.

CGH was illustrated in a sympatric population of *Campylobacter* (Lefebure et al., 2010) that consisted of strains from two closely related species of this bacterium. The core genome of both species was shaped by recombination, but was almost free from interspecies recombination. However, Sheppard et al. (2008) showed the evidences of gene exchange between these two closely related species that lead to the process of 'despeciation'.

1.3.4. Revolution in the Recombination model

For a long time, the role of recombination (Dykhuizen and Green, 1991; Lan and Reeves, 1996; Majewski et al., 2000; Roberts and Cohan, 1993; Zawadzki et al., 1995) was suggested for defining bacterial species. However, Fraser et al. (2007) proposed the most promising and robust measures that displayed a relationship between rate of recombination and genetic diversity in which the population was considered free from ecological or any other barriers. Robust analyses suggested the random occurrence of recombination throughout the genome and categorized species into two categories: 'clonal' and 'sexual' species. Clonal species are the unstable result of a lower recombination rate than mutation rate in a population. On the other hand, sexual species are the outcome of an equal or higher recombination rate than mutation rate in a population. Closely related sexual species are characterized by high recombination rate within members of the same species, compared to between those species. Two years after creating this model, the same authors (Fraser et al., 2009) suggested that mechanisms such as geographical or ecological differentiation are responsible for maintaining the reduced rate of recombination between two 'sexual' species. Thus, the combined approach of 'sexual' species and ecological properties defining these species should be used to define populations. This model might shed some light on the fuzzy species of recombinogenic *Neisseria* (Hanage et al., 2005) or other genera.

Since then, many population studies have identified biological species using this concept. For example, Keymer and Boehm (2011) found recombination as a unifying force in the population of *Vibrio cholerae*, which are primarily known to be clonal species. Moreover, some specific genetic tools have been designed to calculate the recombination rate between recipient and donor. Some popular tools like ClonalFrame (Didelot and Falush, 2007), ClonalOrigin (Didelot et al., 2010) and Structure (Pritchard et al., 2000) have been used to reveal the dominating nature of interspecies recombination in a comparative analysis (Doroghazi and Buckley, 2010) of recombination rates (intra- and interspecies) between several pre-defined ecospecies of *Streptomyces*. Similarly, Didelot et al. (2011) revealed genetic isolation between five incipient species present in *Salmonella enterica* subspecies that are traditionally ecologically distinct.

Further, comparative analyses including maximum likelihood, ClonalFrame and Structure have been used to identify bacterial 'introgression' - gene transfer between identified species that can occur between closely or distantly related species. For example, Vernikos et al. (2007) utilized whole genome phylogenetic analysis to identify horizontally transferred genes in the *Salmonella* lineage that were acquired from prophages, but these genes were absent in the sister lineage *Escherichia coli* and their common ancestor. Similarly, Sheppard et al. (2013) identified incidences of introgression between the closely related species *Campylobacter coli* and *Campylobacter jejuni*. Initially to investigate the introgressed genes, unusual characteristics of base compositions including GC content, and synonymous codon usage were used, since recently acquired genetic material may reflect the base compositions of the donor genome while older insertions may ameliorate with time to reflect the base compositions of the recipient genome (Lawrence and Ochman, 1997). In contrast to the comparative analyses, gene flow based on the base composition approach may miss the real number of HGT events, since new insertions from closely related species may not have unusual base compositions or ancient insertions may completely resembles the recipient genomes due to the amelioration process (Koski et al., 2001).

1.3.5. The Ecotype concept:

Another theoretical species concept (Cohan and Perry, 2007; Cohan, 2002; Cohan, 2001) is based on ecological and evolutionary properties. According to this, bacterial species can be categorized into multiple ecotypes. An ecotype is a distinct population within the bacterium that shares the same ecological niche. Rather than rare recombination, a cohesive force of periodic selection maintains the genetic similarity between members of the same ecotype. This cohesive force, first described by Atwood et al. (1951) and later supported by extensive experimental studies (Koch, 1974; Levin, 1981), occurs when a rare beneficial mutation or gene arises in a population of asexual organisms. Selection on this gene leads to replacement of the whole population by the clonal descendants of this favored mutant. Even if there is rare recombination, exchange of the adaptive mutant between two populations of different genetic background will be restricted, and periodic selection will purge the genetic diversity within the population. In addition to the beneficial mutation, the entire genome of the mutant cell will be driven through the population by genetic hitchhiking. At some point, in the future, the population of the favored mutant will be replaced by another advanced adaptive mutant, and so on.

In the ecotype model (Cohan, 2005; Cohan, 2002), different ecotypes evolve independently by their own private periodic selection events and do not hamper the evolution of each other. These distinct properties make ecotypes analogous to the eukaryotic species. Unlike eukaryotic species, the recombination force between ecotypes is inadequate for preventing their ecological divergence. Therefore, a bacterial species is more like a genus that involves multiple ecotypes.

Different names have been given to ecotypes by bacteriologists based on their ecological properties, for example, biovars (biochemical variants) and serovars (antigenic distinct population). In phylogenetic analysis, the distinct monophyletic groups represent the ecotypes (Cohan, 2005; Ward et al., 2008) and these clusters have been widely used to identify the unknown ecotypes present in a particular species or a relative ecotype of a particular strain. For example, Haley et al. (2010) classified two novel ecospecies of the *Vibrio* genus that were misclassified as two strains of *V.*

cholerae. These two novel species are the core suppliers of genomic islands, including virulence, to the other pathogenic *Vibrio* species. Similarly, Lassalle et al. (2011) observed ecospecies in *Agrobacterium tumefaciens* species complex using DNA microarray technique.

Surprisingly, monophyletic groups for different ecospecies were not observed in the whole genome sequences of populations (Zwick et al., 2012) of predefined species of *Bacillus cereus* sensu lato species. These populations were clustered into several clades composed of strains of different species, which were maintained by low inter- and high intra-species recombination.

1.3.6. Pangenome

Tettelin et al. (2005) suggested the concept of the 'Pan-genome' during the comparison of 8 strains of *Streptococcus agalactiae*. According to them, bacterial pan-genomes can be divided into three sections: A. Core genes that are ubiquitous among all strains of a species. B. Dispensable genes that are shared by a subset of species isolates. C. Strain-specific genes that are carried by a single strain only. Pan-genomic differences highlight the lifestyle of the organism and can be used to classify species in two categories: open and closed pan-genomes. An open pan-genome is characterized by the increase in dispensable or strain specific genes with the number of new sequenced strains and is primarily observed in the bacteria adapted to multiple niches or frequent gene exchangers like *Propionibacterium acnes* (Tomida et al., 2013), *Streptococcus agalactiae* (Tettelin et al., 2005), *Streptococcus pneumoniae* (Donati et al., 2010), and *Haemophilus influenzae* (Hogg et al., 2007). On the other hand, species like *Bacillus anthracis* (Medini et al., 2005), *Campylobacter coli and C. jejuni* (Lefebure et al., 2010) with restricted niche and genetic transfer have closed pan-genomes in which genome size will remain constant after the introduction of new strain.

1.4. Rhizobia and their taxonomy

Diazotrophs are bacteria that have a distinct property of fixing atmospheric nitrogen. They are the biological producers of ammonia products from nitrogen and can be classified into two main categories: free-living and symbiotic. The major contributors to the terrestrial nitrogen cycle (Wagner, 2012) are rhizobia, which are symbiotic diazotrophs that form root nodules on leguminous plants like peas, clovers, and beans and exchange ammonia for energy products with legumes. This legume-rhizobium symbiosis is characterized by a high degree of host specificity.

From the point of view of taxonomy (Figure 1.3), most rhizobia are in the five prime genera (*Rhizobium*, *Sinorhizobium*, *Mesorhizobium*, *Azorhizobium* and *Bradyrhizobium*) of the alpha-proteobacteria group (Masson-Boivin et al., 2009). Initially, all legume-symbionts were included in the genus *Rhizobium*, but several taxonomic studies have revealed the presence of different genera. For example, *Bradyrhizobium*, comprising slow growing species, was the second genus classified by Jordan (1982).

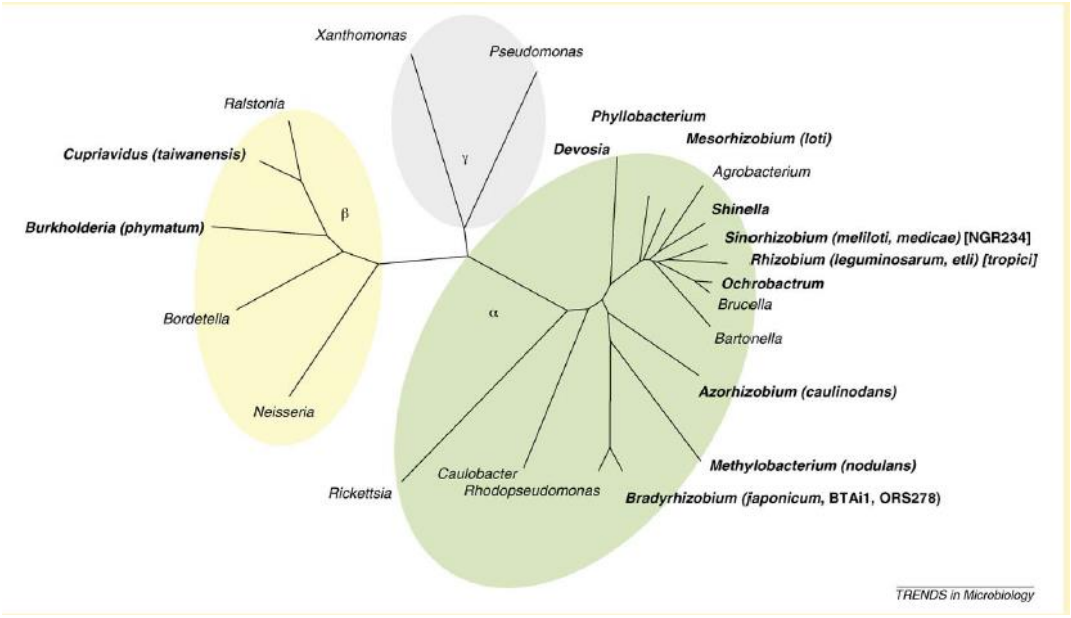


Figure 1.3 | Distribution of selected rhizobia (bold font) in different classes of proteobacteria (α and β classes) in 16S rDNA sequence phylogeny (Masson-Boivin et al., 2009).

1.4.1. Symbiovars

Based on legume (host) specific genes, rhizobial species are classified into different symbiovars (Rogel et al., 2011), which were previously known as biovars. Symbiovars of a rhizobium species are distinct populations with the same symbiotic properties that differ from other populations. Because legume specific genes are capable of migrating from one species to another species or genus, the same symbiovars with diverged homologous legume-specific genes may exist in multiple species. Generally, symbiovars are characterized by rhizobia in the alpha-proteobacteria group. Possession of a different numbers of symbiovars by a species reflects its adaptive nature in different niches (nodules), for example, *R. leguminosarum* has three symbiovars: *viciae* (nodulating vetches, peas, and lentils), *trifolii* (nodulating clover), and *phaseoli* (nodulating common bean), whereas its close relative, *R. etli*, with two symbiovars, interact with the common bean (shared symbiovar *phaseoli* of *R. leguminosarum*) and with *Mimosa affinis* (symbiovar *mimosae*).

1.4.2. *R. leguminosarum* species complex

The type species of the genus *Rhizobium* is *R. leguminosarum* and the type strain of this species (USDA 2370^{T}) belongs to symbiovar *viciae*. Species studies based on *Rhizobium* (Ramirez-Bahena et al., 2008; Tian et al., 2008) have suggested the presence of a *R. leguminosarum* species complex composed of *R. leguminosarum* and its close relatives *R. etli, R. pisi, R. fabae and R. phaseoli*. The complete publicly available sequenced genomes in this complex are *Rhizobium leguminosarum* symbiovar *viciae* (*Rlv*) 3841 (Young et al., 2006), *R. leguminosarum* symbiovar *trifolii* (WSM1325: Reeve et al., 2010a and WSM2304: Reeve et al., 2010b), *R. etli* ($CFN42^{T}$: Gonzalez et al., 2006) and *R. elti* CIAT652 (Gonzalez et al., 2010), which is now classified as *R. phaseoli* (Lopez-Guerrero et al., 2012).

1.4.3. Nodulation (*nod*) genes

Among legume-specific elements, nodulation factors (lipochito-oligosaccharides) and related genes (*nod* genes) initiate the formation of nodules (Spaink et al., 1991). The genomic structure of *nod* genes explains the property of host specificity in symbiovars of rhizobia (Relic et al., 1994; Roche et al., 1996; Spaink, 1994; van Brussel et al., 1990). Here, we discuss the genomic structure of *nod* genes in *R. leguminosarum*, which can be divided into three categories: (A) A regulatory gene, *nodD*, which interacts with flavonoids excreted by plants and activates the other *nod* genes. (B) The common *nod* operon (*nodABCIJ*), which is shared by all rhizobia. The *nodABC* genes are involved in the biosynthesis of nodulation factor. The *nodIJ* genes (Spaink et al., 1995) have a role in secretion system of the nodulation factor. (C) The host specific genes (*nodFELMNTO)* are major determinants of legume specificity, for example, *nodE* gene (Bloemberg et al., 1995) play a dominant role in determining the host range of symbiovars *trifolii* and *viciae*. In addition, other specific *nod* genes have also been observed, such as the *nodX* gene that mediates an O-acetylation of the nodulation factor (Davis et al., 1988; Firmin et al., 1993) and is generally found in symbiovar *trifolii* strains, but not in *viciae* strains except *Rlv* TOM (Davis et al., 1988).

Generally, symbiosis-associated genes involving *nod* genes are located on plasmids or within symbiosis islands (*Bradyrhizobium*) reflecting their ability to travel from one genus to another by HGT. Some studies have demonstrated the evolution of *nod* genes driven by homologous recombination or HGT (Bailly et al., 2007; Moulin et al., 2004). However, *nod* genes are not always necessary for nodulation, for example, Giraud et al. (2007) observed a symbiotic relationship in photosynthetic *Bradyrhizobium* strains that lack *nod* genes.

1.5. Aims and objectives

The main objective of this thesis is to investigate the nature of bacterial species in a closely related population of *R. leguminosarum* using robust bioinformatics tools. Because *R. leguminosarum* symbiovar *trifolii* and *viciae* are more closely related to each other than to the remaining symbiovar *phaseoli*, display extreme host specialization (Jordan, 1984; Rogel et al., 2011; Young, 1996) and frequently occur together, they provide an ideal pair for understanding the relationship between the genetics and ecology of the bacterial species. The main objectives of this study are:

Chapter 2: Phylogenomic analysis reveals cryptic genospecies in a local population of *Rhizobium leguminosarum*.

- Determine the genetic diversity present in the population of *R. leguminosarum* (symbiovars *viciae* and *trifolii*) using core genes phylogeny.
- Investigate the location of *R. leguminosarum* population in the genus *Rhizobium*.
- Determine the genospecies of *R. leguminosarum* and compare them with ecotypes.
- Investigate the cosmopolitan nature of localized genospecies.

Chapter 3: Recombination and population structure of cryptic genospecies of *R. leguminosarum*.

- Determine the phylogeny of 100 core genes that are represented in our data for all strains.
- Investigate the role of recombination in the evolution of core genes.
- Investigate the role of genetic isolation that maintains the structure of five genospecies.

Chapter 4: Dominating influence of five genospecies on the composition and phylogeny of the accessory genome of *R. leguminosarum.*

- Investigate the distribution of genes of reference genome in five genospecies.

- Investigate the consistency of reference-based chromosome, chromids and plasmid phylogenies.
- Investigate the phylogenetic structure based on *nod* (host specific) genes.
- Determine the population specific genes that are absent in reference genome.

Chapter 5: Comparative genomics of two major genospecies of *R. leguminosarum*.

- Investigate the genomic structure of genospecies C (the biggest cluster) based on a candidate strain.
- Determine the genomic differences and similarities between genospecies C and reference related genospecies B.

Chapter 2. Phylogenomic analysis reveals cryptic genospecies in a local population of *Rhizobium leguminosarum*.

2.1 Abstract

Rhizobium leguminosarum is a nitrogen-fixing bacterium that lives in soil and in the root nodules of leguminous plants. Here, we analyzed 72 strains of *R. leguminosarum* (36 of symbiovar *viciae* and 36 of symbiovar *trifolii*) using the fully sequenced genome of strain 3841 as reference. Phylogenetic analysis based on 305 core genes divided this population into five discrete clusters and each cluster is characterized by the co-existence of symbiovars. Average Nucleotide Identity (ANI) analysis, a pairwise genome comparative analysis, indicates that these five phylogenetic clusters are sufficiently diverged to be regarded as species (genospecies). This chapter presents comparative analysis of genospecies (based on core genes) and ecotypes (based on ecological niche) in which ecotypes are not reflected by genospecies. Moreover, these genospecies were not confined to their isolated site (York, UK).

2.2 Introduction

The co-speciation between symbiotic bacteria and their host is a historical separation that provides a physical barrier to gene flow between bacteria associated with different hosts (Funk et al., 2000). This physical barrier can be a major ecological adaptation with no other ecologically related property that allows host specific bacteria to diverge into discrete clusters and corresponds with ecotypes (Cohan, 2006; Normand et al., 1996; Smith et al., 2006). Similarly, symbiovars should diverge into discrete clusters because they definitely have important differences in at least one aspect of their niche (host-specificity), and members of a symbiovar will meet each other on their host, and hence more often than they will meet other symbiovars.

In this study, we investigated phylogenetic diversity in a local population of *R. leguminosarum* strains by using a reference genome, that of strain 3841. This population was isolated from a square metre near Wentworth College at the University of York, UK in 2007 by a former postdoc, Xavier Bailly. It includes 36 strains of symbiovar *trifolii* (represented as TRX_n, n is a strain number) isolated from *Trifolium repens* and 36 strains of symbiovar *viciae* (VSX_n) isolated from *Vicia sativa.* Draft genomes of this population were obtained using 454 sequencing.

The seventy-two *R. leguminosarum* strains shared their location with a population of *Sinorhizobium medicae* strains. In a previous publication, we compared draft genomes of twelve strains from this population (MLX_1-MLX_12) with a fully sequenced genome of *S. medicae* WSM 419 (Bailly et al., 2011). Phylogenetic analysis of the chromosome (Figure 2.1) resulted in star-like phylogenies, indicating lower sequence divergence on the chromosome than other replicons.

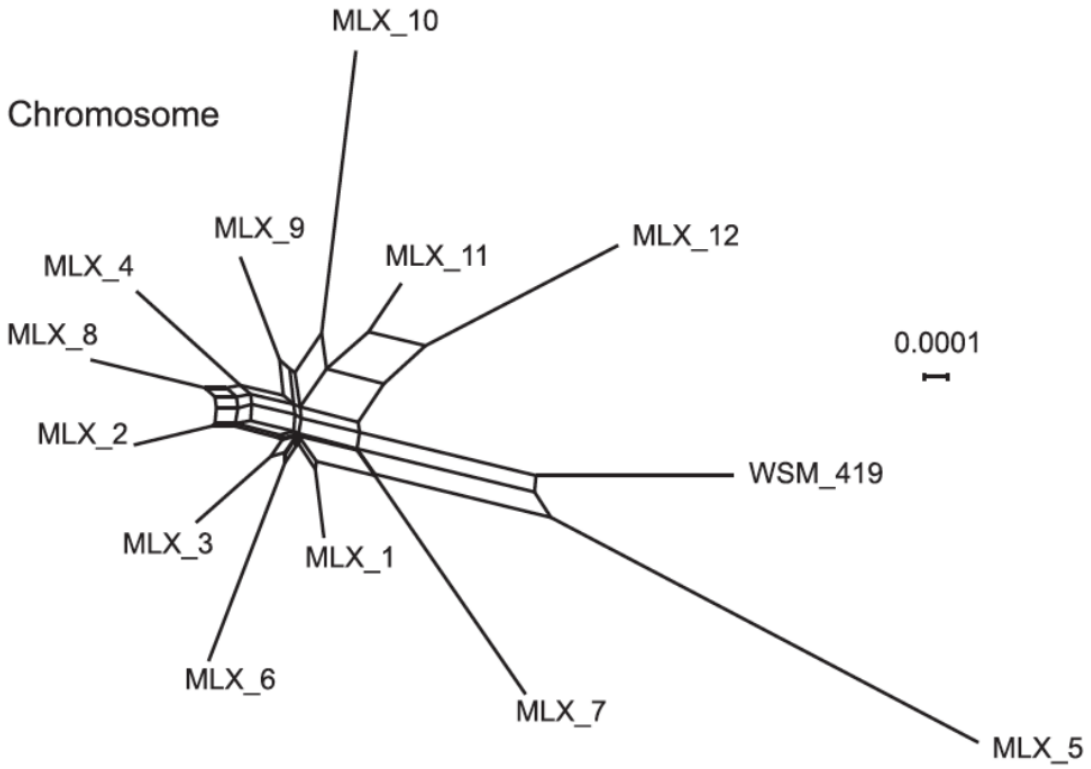


Figure 2.1 | Chromosomal network showing divergence among *S. medicae* strains (MLX_1-MLX_12) using reference genome of WSM 419 (Taken from Bailly et al., 2011)

Rhizobium leguminosarum symbiovar *viciae* (*Rlv*) 3841 that is taken as a reference genome in this study was isolated from a nodule on a pea (*Pisum sativum*) in Norfolk, England and fully sequenced in 2006 (Young et al., 2006). *Rlv* 3841 has a genome of 7.75 Mb (Figure 2.2) that is distributed into one circular chromosome (5.05 Mb) and six circular plasmids (pRL12-pRL7). Like other bacteria, its genome has two components: a conserved 'core' genome and variable accessory genome. Although most of the essential genes are present on the chromosome, plasmids have most of the functional genes such as the symbiosis genes on pRL10 only, ABC transporters on each plasmid, cell division proteins, etc. Harrison et al. (2010) suggested that plasmids pRL12 and pRL11 should be considered as two chromids of *Rlv* 3841. They also provided a list of 305 genes that are conserved in all chromid-possessing bacteria, and this set of genes is used in the present study.

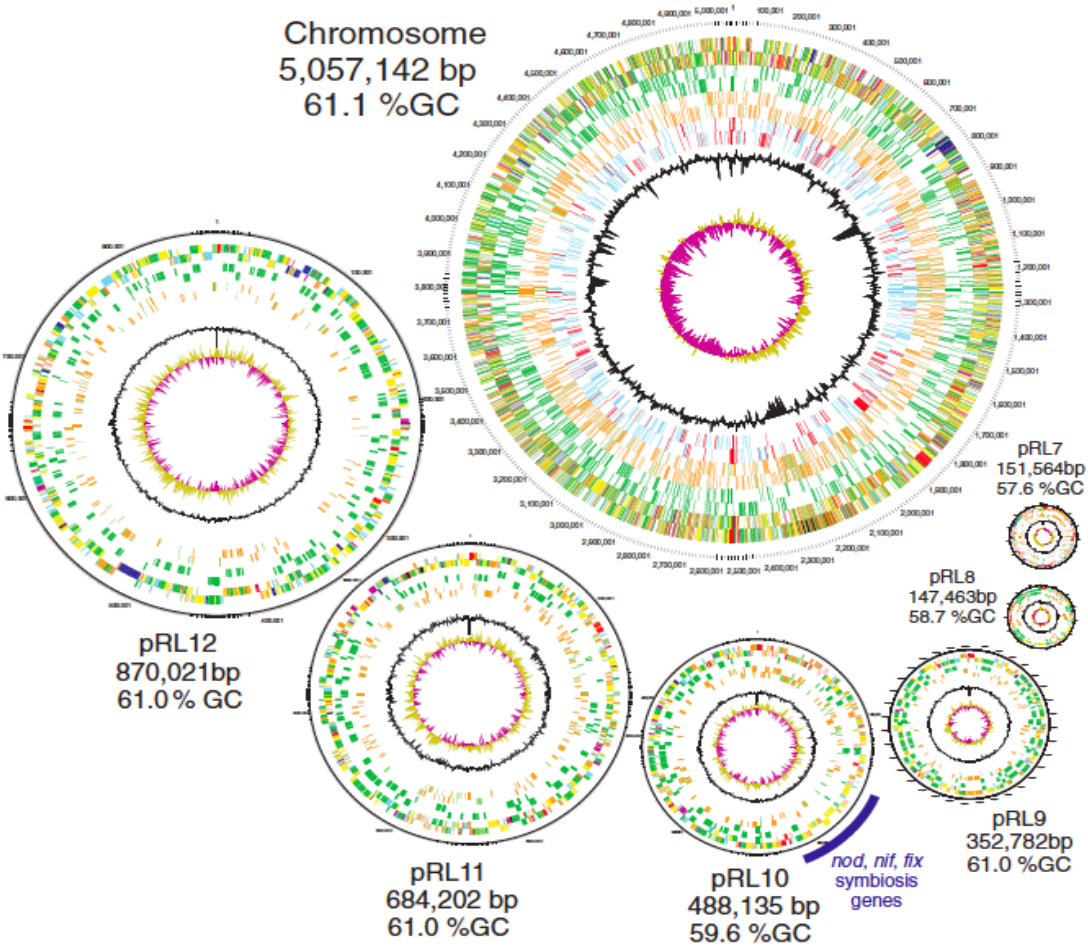


Figure 2.2 | Genomic structure of *Rlv* 3841: one chromosome and six plasmids (Taken from Young et al., 2006)

A taxonomic revision (Ramirez-Bahena et al., 2008) based on DNA-DNA hybridization experiments, phenotypic characteristics and phylogenetic analysis suggested true type strains of *R. leguminosarum* (USDA 2370^T), *R. pisi* (DSM 30132^T) and *R. phaseoli* (ATCC 14482^T). USDA 2370^T and DSM 30132^T were isolated from the same host: *Pisum sativum*. ATCC 14482^T was isolated from *Phaseolus vulgaris* nodules. Tian et al. (2008) defined *R. fabae* CCBAU 33202^T as the type strain of *R. fabae*.

For the rapid and accurate delineation of bacterial species, traditional methods (Gevers et al., 2005) such as DNA-DNA hybridization experiments (time-consuming and expensive), 16S rRNA sequences (not reliable for closely related strains) and MLST (based on seven core genes) are currently being replaced by the high resolution phylogenetic analysis based on several core genes and robust measures of Average Nucleotide Identity (Chan et al., 2012). ANI (Goris et al., 2007; Konstantinidis and Tiedje, 2005; Richter and Rossello-Mora, 2009) calculates the genomic similarity by performing pairwise comparison of all shared genes between two genomes, hence it is limited to the sequenced genomes.

Apart from ecological specialization, geographical isolation is another mechanism by which bacterial species can be organized into dinstinct clusters. Whitaker (2006) discussed several studies that suggested the occurrence of geographical isolation in both local and global populations, while a pattern of ecological isolation was predominant in local populations. Moreover, some bacterial species are free from geographical barriers; for example, Linz et al. (2007) observed two distinct African ancestral populations of *H. pylori*. In *R. leguminosarum*, a few Chinese strains were clustered (Tian et al., 2010) with Western strains in the phylogenetic and Structure analysis based on three chromosomal housekeeping genes, indicating lack of geographical barriers.

2.2.1 Objectives

The main objectives of this chapter are:

A. *De novo* assembly of all unpublished *Rhizobium* genomes used in this study.
B. Phylogenetic analysis of *R. leguminosarum* strains from a local population, based on universal core genes.
C. Phylogenetic analysis of the *R. leguminosarum* species complex based on the same universal core genes.
D. ANI analysis of strains from a local population of *R. leguminosarum*.
E. Phylogenetic analysis of *R. leguminosarum* strains from different locations.

2.3 Material and methods

2.3.1 Sequence data

We obtained a list of 305 core genes (Appendix table I.I) based on the 3841 genome from Dr. Peter Harrison (Harrison et al., 2010) by personal communication. Fully sequenced *Rhizobium* genomes (Table 2.1) used in this study were downloaded from the National Center for Biotechnology Information (NCBI). BLASTn with cutoff e value of 1e-05 and a Perl script with Bioperl (Stajich et al., 2002) module Bio::SearchIO were used to extract genetic information of 305 core genes from downloaded *Rhizobium* genomes.

The type strains of different *Rhizobium* species used in this study are summarized in Table 2.2. In addition to the seventy-two *R. leguminosarum*, other *R. leguminosarum* strains obtained from Sweden (Dr Kerstin Huss-Danell) and Scotland (Dr Euan James) are summarized in Tables 2.3. The strains isolated from different places in Sweden include 5 strains of symbiovar *viciae* (represented as VCS_n) isolated from *Vicia cracca* and 2 strains of symbiovar *trifolii* (TPS_n) isolated from *Trifolium pratense*. The Scottish strains include one strain of symbiovar *viciae* (VCS_6) and symbiovar *trifolii* (TPS_6) isolated from *Vicia cracca* and *Trifolium pratense* respectively.

Table 2.1 | Fully sequenced genomes of *Rhizobium* genus used in this study other than *Rlv* 3841

Strain	Replicon	Size (Mb)	Accession Number
Rhizobium etli CFN42 (Gonzalez et al., 2006)	Chromosome	6.53	NC_007761
	p42a		NC_007762
	p42b		NC_007763
	p42c		NC_007764
	p42d		NC_004041
	p42e		NC_007765
	p42f		NC_007766
Rhizobium etli CIAT 652 (Gonzalez et al., 2010)	Chromosome	6.44	NC_010994
	pA		NC_010998
	pB		NC_010996
	pC		NC_010997
Rhizobium leguminosarum symbiovar *trifolii* WSM1325 (Reeve et al., 2010a)	Chromosome	7.41	NC_012850
	pR132501		NC_012848
	pR132502		NC_012858
	pR132503		NC_012853
	pR132504		NC_012852
	pR132505		NC_012854
Rhizobium leguminosarum symbiovar *trifolii* WSM2304 (Reeve et al., 2010b)	Chromosome	6.87	NC_011369
	pRLG201		NC_011368
	pRLG202		NC_011366
	pRLG203		NC_011370
	pRLG204		NC_011371

Table 2.2 | Type strains sequenced in this study

Strain	Host	Libraries
Rhizobium leguminosarum USDA 2370^T	*Pisum sativum*	Single reads
Rhizobium pisi DSM 30132^T	*Pisum sativum*	Paired ends
Rhizobium fabae CCBAU 33202^T	*Vicia faba*	Paired ends
Rhizobium phaseoli ATCC 14482^T	*Phaseolus vulgaris*	Paired ends

Table 2.3 | Other *R. leguminosarum* strains sequenced in this study

Strains	Host	Location
KHDVB 646.3 (VCS_1)	*Vicia cracca*	Lappland, Sorsele, Ammarnäs meadow
KHDVB 717.3 (TPS_1)	*Trifolium pratense*	Lappland, Sorsele, Ammarnäs meadow
KHDVB 902.1 (VCS_2)	*Vicia cracca*	Lappland, Sorsele, Kraddsele meadow
OYAVB 169.1 (VCS_3)	*Vicia cracca*	Västerbotten, Umeå, Ängersjö edge of field
OYAVB 296.5 (VCS_4)	*Vicia cracca*	Västerbotten, Umeå, Ängersjö edge of field
OYAVB 349.6 (VCS_5)	*Vicia cracca*	Västerbotten, Umeå, Ålidhem roadside
OYAVB 371.3 (TPS_5)	*Trifolium pratense*	Västerbotten, Umeå, Ålidhem roadside
S 273.16(VCS_6)	*Vicia cracca*	Tayport, Shanwell Farm farm track
S 272.1 (TPS_6)	*Trifolium pratense*	Tayport, Shanwell Farm farm track

2.3.2 GS De Novo Assembler

We used a command line (runAssembly) option with 90% sequence identity and 40-bp minimum overlap as parameter to perform *de novo* assembly of all the unpublished *Rhizobium* genomes (seventy-two Wentworth *R. leguminosarum*, plus those listed in Tables 2.2 and 2.3).

2.3.3 GS Reference Mapper

We used a command line (runMapper) option with 90% sequence identity and 40-bp minimum overlap as parameter to perform reference-based assembly of all the unpublished *Rhizobium* genomes using 305 core genes as the reference genes. Nucleotide information of 305 core genes was extracted from every draft genome using extractSequence.pl (http://seqanswers.com/forums/showthread.php?t=9498), and the extracted information was merged with their respective genes present in fully sequenced studied *Rhizobium* genomes.

2.3.4 Multiple sequence alignment

Each of the 305 files was aligned at nucleotide level by MUSCLE (multiple sequence comparison by log-expectation) created by Edgar, 2004 that was run locally on the University of York Biology Linux grid. MUSCLE is an open source bioinformatics tool used for multiple alignments and is meant to be a better alignment tool than ClustalW2 (Larkin et al., 2007) or T-Coffee (Notredame et al., 2000) for speed and accuracy. Each alignment file was checked and gaps were added for strains that had no reads for a given gene. The final results of FASTA alignments were concatenated by strains to form a 305 core genes alignment using Galaxy (Goecks et al., 2010).

2.3.5 Phylogenetic analysis

Phylogenies were constructed using either neighbour-net or maximum likelihood methods. All neighbour-nets were generated using the uncorrected p distances function of SplitsTree version 4.11 (Huson and Bryant, 2006). All maximum likelihood analyses were performed by FastTree (Price et al., 2010) with settings: -gamma –gtr (most reliable and general model), run locally on the University of York Biology Linux grid.

2.3.6 Average Nucleotide Identity analysis

Average Nucleotide Identity (ANI) was calculated using the JSpecies package (Richter and Rossello-Mora, 2009). ANI can be calculated by two methods: BLAST (ANIb) and MUMmer (ANIm). The ANIm is designed to compare large DNA segments with high accuracy and in less time than ANIb, therefore, ANIm was used in this study. The cutoff for percent similarity between two genomes is 96% and is close to the DDH threshold value of 70% (Konstantinidis and Tiedje, 2005). This method was applied to representative strains that were selected based on coverage and to include at least one member from each of the five clusters (A-E) and subclusters present in the biggest cluster (cluster C) of the maximum likelihood tree based on 305 core genes.

2.3.7 Computational Resources

The whole procedure was run on an Apple MacBook Pro Intel Core 2 Duo 2.4 GHz CPU with 4GB 667 MHz RAM running Mac OS X 10.8.4.

Grid computing was used for computationally intensive jobs by using the local University of York Biology Linux grid, comprising of 27 quad core machines each with 2GB RAM and three dedicated clusters each with 4 CPUs and 8GB RAM, controlled by Sun Microsystems Grid engine 6.1u2. Perl version 5.12.4 and R version 2.15.2 were used in this and subsequent chapters.

2.4 Results

We analyzed 90 *Rhizobium* strains that belong to different species of the *R. leguminosarum* species complex.

2.4.1 Genome sequencing

First, we performed a *de novo* assembly of 72 *R. leguminosarum* draft genomes (Figure 2.3 and 2.4). Of 36 VSX strains, VSX_7 (6.13 fold) and VSX_3 (0.70 fold) have the highest and lowest sequence coverage respectively (Figure 2.3). Of 36 TRX strains, TRX_6 (10.06 fold) and TRX_31 (0.68 fold) have the highest and lowest sequence coverage respectively (Figure 2.4). TRX_6 has the highest amount of sequence coverage in this study. Therefore, this strain was selected for the additional genomic analyses described in Chapter 5.

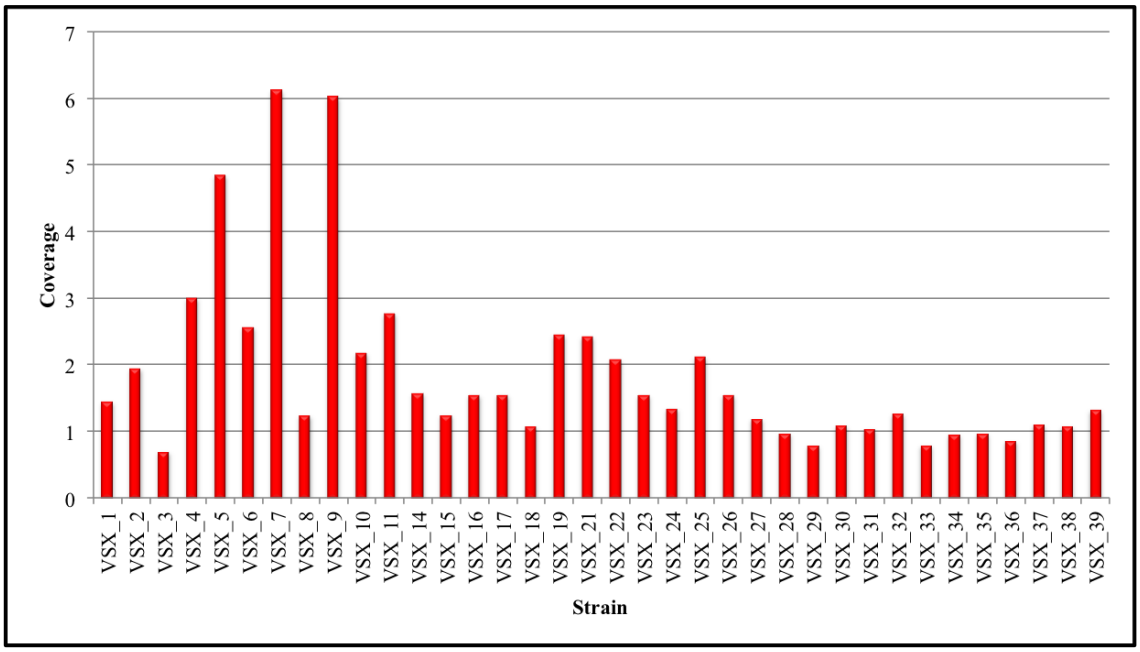


Figure 2.3 | Sequence coverage of 36 *viciae* (VSX) strains. X-axis represents strain number and Y-axis shows coverage of each strain.

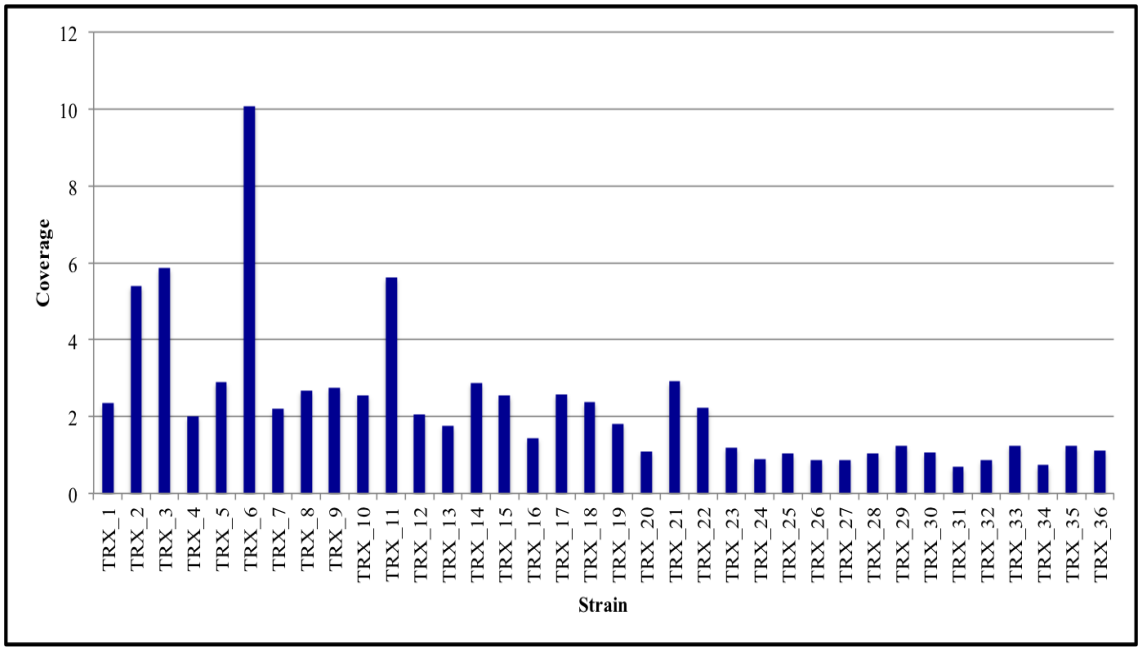


Figure 2.4 | Sequence coverage of 36 *trifolii* (TRX) strains. X-axis represents strain number and Y-axis shows sequence coverage of each strain.

The *de novo* assembly was performed for the type strains of USDA 2370^{T}, *R. pisi* DSM 30132^{T}, *R. fabae* CCBAU 33202^{T}, *R. phaseoli* ATCC 14482^{T} (Figure 2.5). The sequence coverage was in the range of 3.9–5 fold.

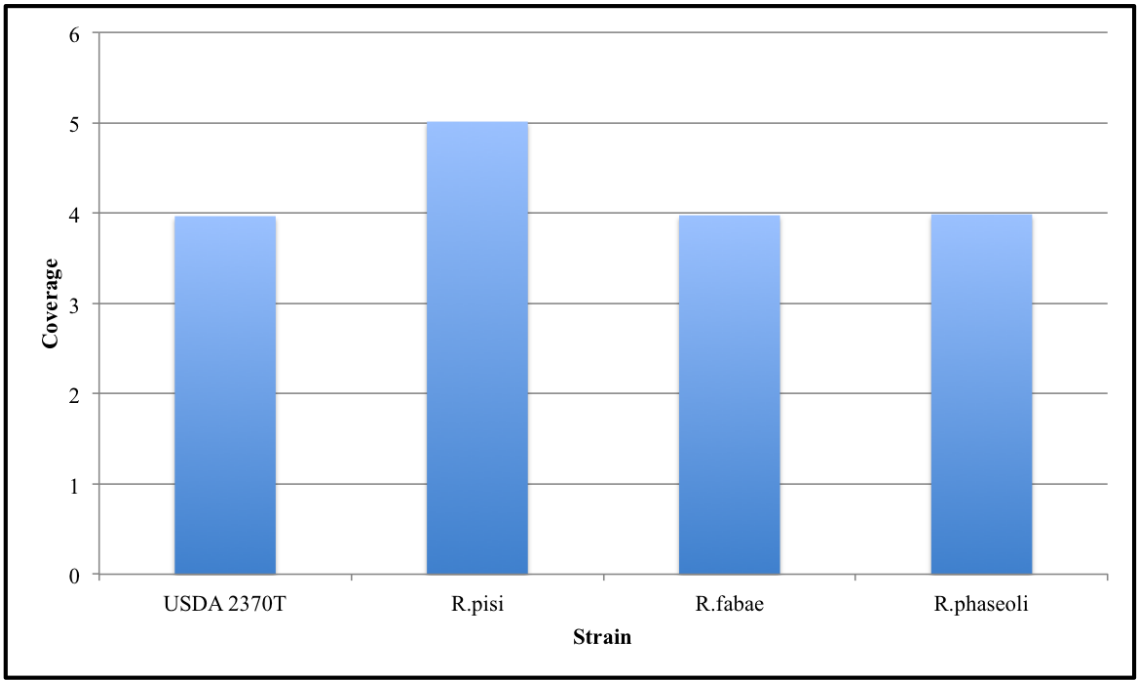


Figure 2.5 | Sequence coverage of 4 type strains. X-axis represents strain number and Y-axis shows sequence coverage of each strain.

The Swedish and Scottish strains were assembled (Figure 2.6) and sequence coverage was revealed in the range of 1.53 and 4.87 fold. Out of the Swedish strains, TPS_5 has the highest (4.87), whereas VCS_3 and VCS_5 have the lowest (1.53) coverage. TPS_6 and VCS_6 (Scottish strains) have sequence coverage of 2.3 fold and 2.25 fold respectively.

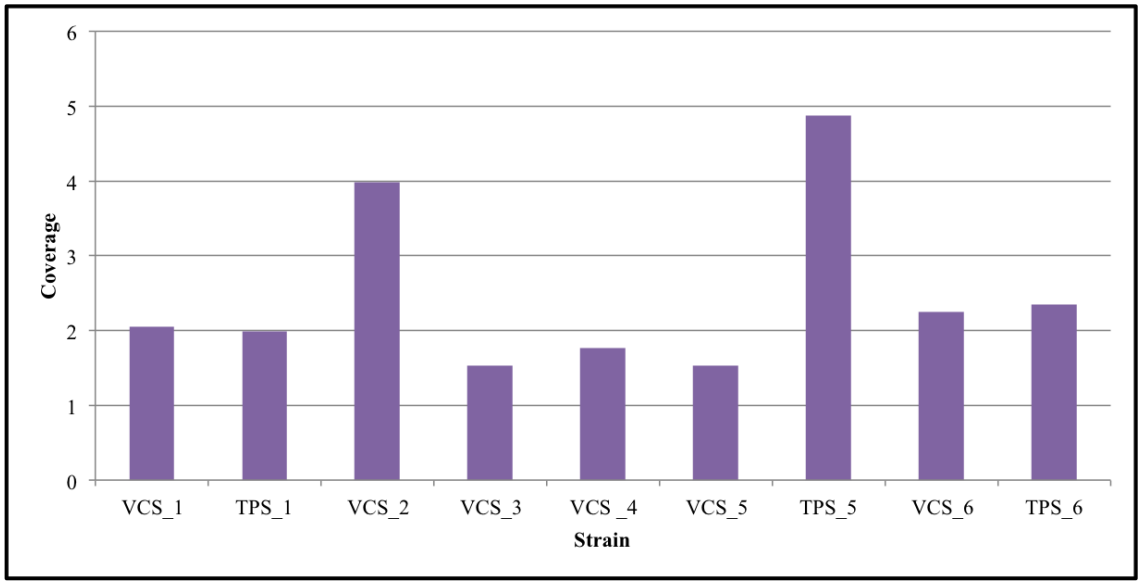


Figure 2.6 | Sequence coverage of 4 Swedish (VCS_1,TPS_1,VCS_2, VCS_3,VCS_4, VCS_5, TPS_5) and 2 Scottish (VCS_6, TPS_6) strains. X-axis represents strain number and Y-axis shows sequence coverage of each strain.

2.4.2 Phylogenetic analysis of local population of *R. leguminosarum*

A phylogenetic network of 72 *R. leguminosarum* strains was inferred based on concatenated alignment of 305 core genes using *Rlv* 3841 as a reference genome. This robust network (Figure 2.7 & Table 2.4) clearly separated this population into five discrete clusters: cluster A (purple ring) with 1 strain (TRX_34), cluster B (salmon ring) comprises 13 strains including *Rlv* 3841, cluster C (green ring) consisting of a maximum number of strains (52), cluster D (cyan ring) with 4 strains and cluster E (dark red ring) consisting of 3 strains. Cluster D is the only cluster that is niche specific. Clusters B, C and E consist of a population of both symbiovars (*viciae* and *trifolii)* of *R. leguminosarum*.

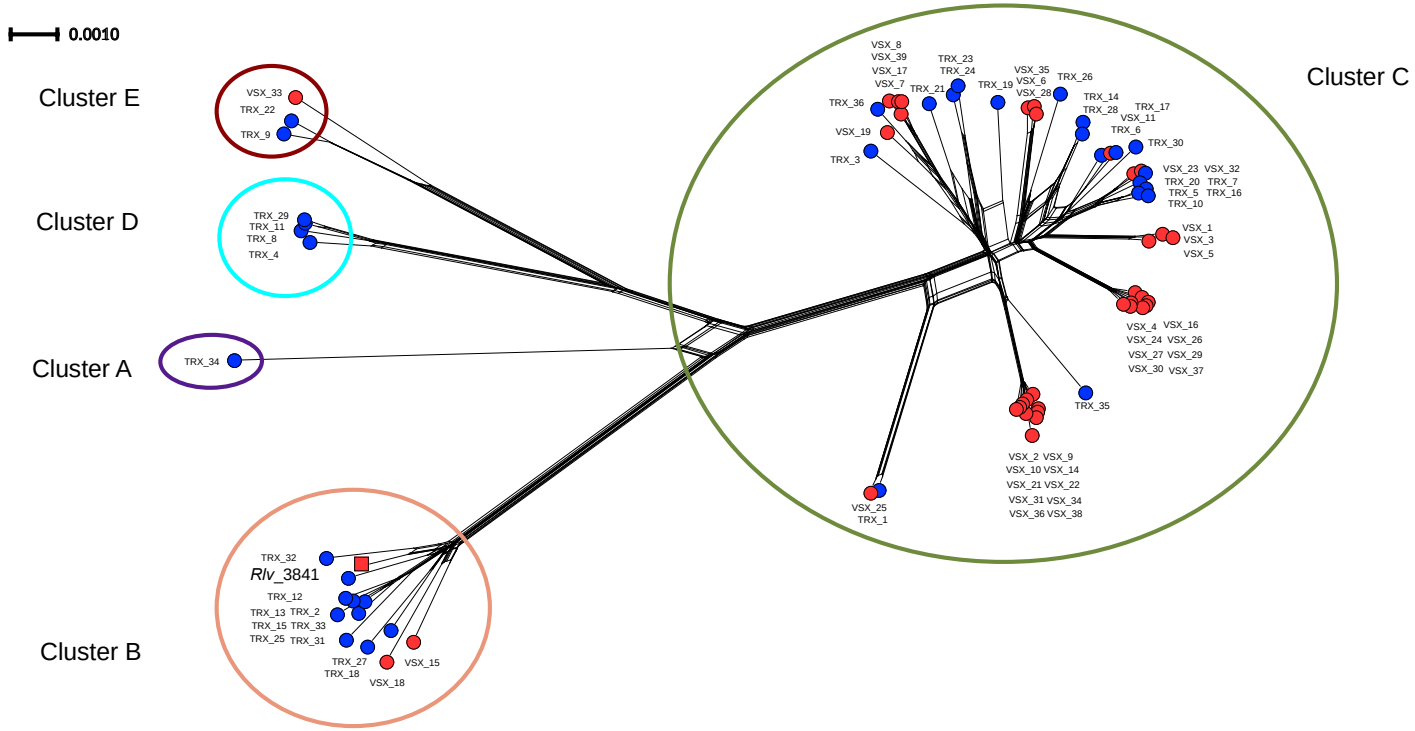


Figure 2.7 | Neighbour-net phylogeny of *R. leguminosarum* strains based on concatenated alignment of 305 core genes. Strains of two symbiovars are shown with blue (TRX) and red circles (VSX). *Rlv* 3841 is shown by the red square.

Table 2.4 | List of strains of two symbiovars (*trifolii*: TRX_ and *viciae*: VSX_) of *R. leguminosarum* that are classified in the five clusters based on the 305 core genes

Clusters	**TRX_**	**VSX_**
A	34	none
B	2 12 13 15 18 25 27 31 32 33	15 18
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
D	4 8 11 29	none
E	9 22	33

2.4.3 Global Phylogeny of the *R. leguminosarum* species complex

Figure 2.8 displays the phylogenetic network that indicates the positions of other rhizobium strains (USDA 2370^T, *R. pisi* DSM 30132^T, *R. fabae* CCBAU 33202^T, *R. phaseoli* ATCC 14482^T, *Rlt* WSM1325, *Rlt* WSM2304, *R. etli* CIAT 652 and *R. etli* CFN42^T) relative to 73 *R. leguminosarum* strains analyzed above. *Rlt* WSM1325 and USDA 2370^T were closer to cluster A (TRX_34) than any other clusters. *R. etli* CIAT 652 shared high similarity *R. phaseoli* ATCC 14482^T rather than with *R. etli* CFN42^T, which is consistent with Lopez-Guerrero et al. (2012). *R. pisi* DSM 30132^T and *R. fabae* CCBAU 33202^T were closely related to each other.

A Maximum Likelihood tree (Figure 2.9) based on the same 305 core genes confirms the phylogenetic results of 75 *R. leguminosarum* strains (72 *R. leguminosarum*, USDA 2370^T, *Rlt* WSM1325 and *Rlv* 3841) and clarifies the sub-clusters present in Cluster B and C.

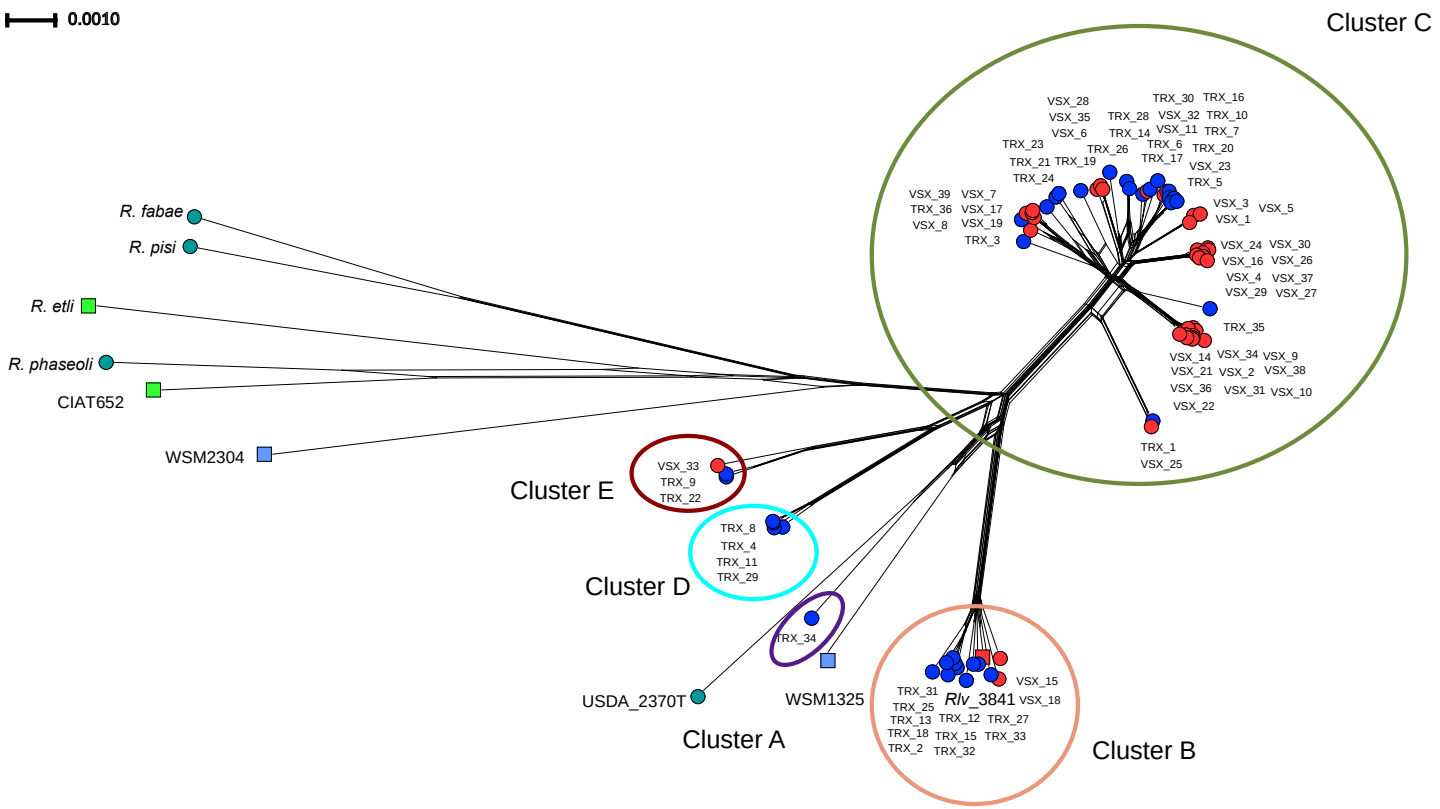


Figure 2.8 | Neighbour-net phylogeny of *R. leguminosarum* species complex based on concatenated alignment of 305 core genes. The local population of *R. leguminosarum* is shown in the blue (TRX) and red circle (VSX). USDA 2370^{T}, *R. pisi* DSM 30132^{T}, *R. fabae* CCBAU 33202^{T}, *R. phaseoli* ATCC 14482^{T} are in dark green circles. *Rlt* WSM1325 and *Rlt* WSM2304 are displayed by blue-green rectangle. *R. etli* CIAT 652 and *R. etli* CFN42^{T}are represented by green rectangle. *Rlv* 3841 is shown by the red square.

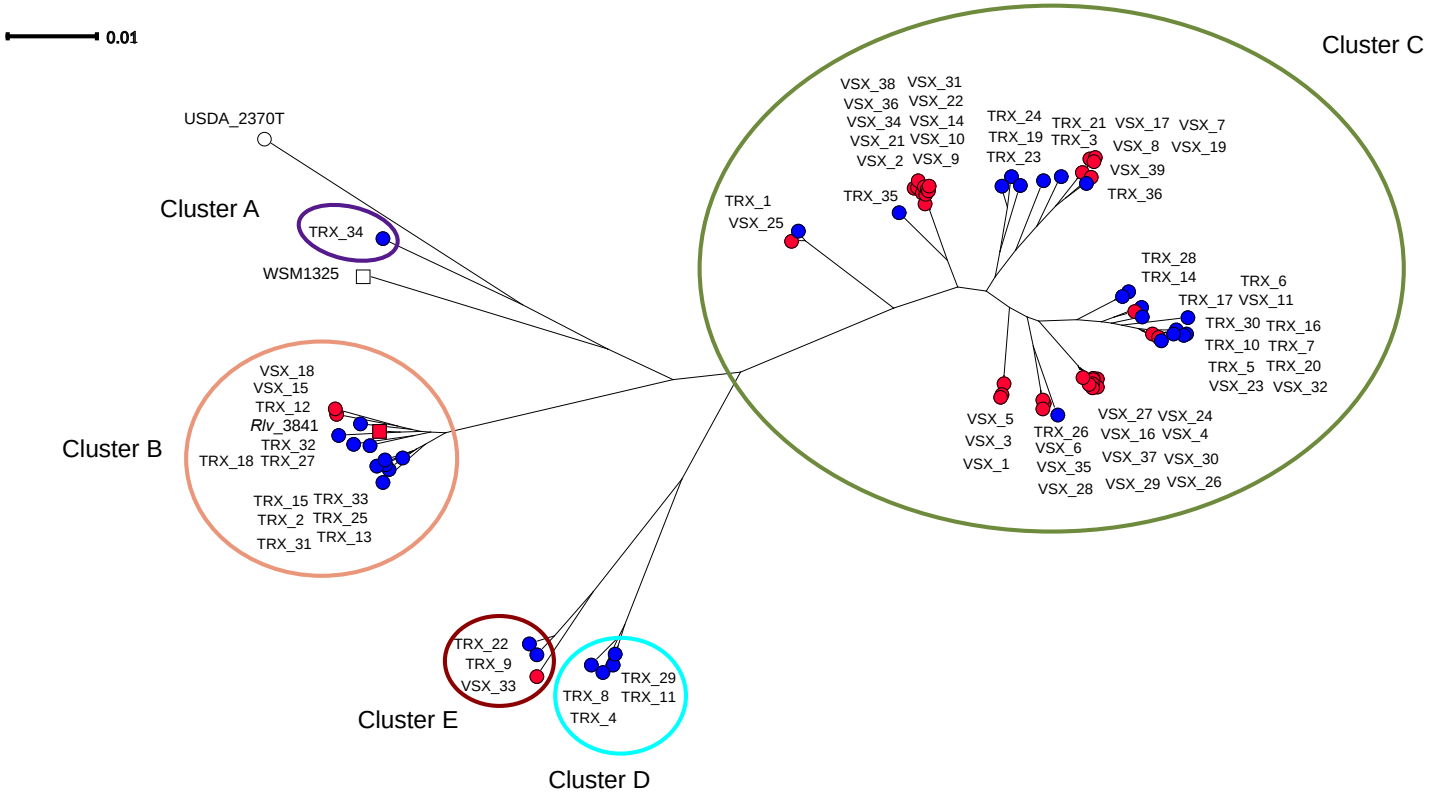


Figure 2.9 | Maximum Likelihood tree of *R. leguminosarum* strains based on concatenated alignment of 305 core genes. Local population of *R. leguminosarum* is shown in the blue (TRX) and red circle (VSX). WSM1325 with white rectangle represents *Rlt* WSM1325. USDA 2370^{T} is shown in white circle. *Rlv* 3841 is shown by the red square.

2.4.4 Average Nucleotide Identity analysis

As an alternative to the traditional DDH method (Gevers et al., 2005; Wayne et al., 1987), a robust technique based on genomic similarity known as Average Nucleotide Identity (ANI) has been proposed to delineate bacterial species. In this technique, two genomes sharing ANI > 95-96% belong to the same species. Otherwise they belong to different species (Auch et al., 2010; Goris et al., 2007; Richter and Rossello-Mora, 2009).

Primarily, ANIm was performed on the selected strains of the local population, the percentage of average nucleotide identity ranges from 93.2% to 99.2% (Table 2.5). The ANIm between the members of Cluster C (TRX_1, TRX_3, TRX_6, TRX_23, VSX_4, VSX_5, VSX_7, VSX_9, VSX_25 and VSX_35) was in the range of 97-99%. The ANIm between the members of Cluster B (*Rlv* 3841 and TRX_2) was 98.5%. The ANIm between the members of Cluster D (TRX_4 and TRX_8) was 98.8%. The ANIm between the members of Cluster E (TRX_9 and TRX_22) was 99.1%. The ANIm between members of different clusters was less than 96%; for example, the ANIm between cluster B and any other cluster (A, C, D and E) is less than 94%. Based on these data, these clusters can be termed genospecies (A-E) because they are based on core genome phylogeny and genomic analysis (ANIm).

ANI was also performed on the strains of *R. leguminosarum* species complex. The ANIm between USDA 2370^T and TRX_34 (genospecies A) was 96.01% (Table 2.6). The ANIm between *Rlt* WSM1325 and TRX_34 (genospecies A) was 94.92%, but *Rlt* WSM1325 shared less than 94% identity with other genospecies. These data suggested that *Rlt* WSM1325 and USDA 2370^T are the members of genospecies A from other geographic locations.

The *R. etli* CIAT 652 and *R. phaseoli* ATCC 14482^T shared a high ANI value of 97.47% that supported Lopez-Guerrero et al. (2012) for classifying the former strain as *R. phaseoli*. The high ANIm (97.69%.) between *R. pisi* DSM 30132^T and *R. fabae* CCBAU 33202^T suggested that they belong to single species.

Table 2.5 | Average nucleotide identity (ANIm) analysis of selected 72 *R. leguminosarum* strains. These strains were selected on the basis of genomic coverage and at least two strains from each cluster, where available. ANI values are in percentages. Black: Values above 96%, Red: Values below 96%

		A	B	B	C	C	C	C	C	C	C	C	C	C	D	D	E	E
Species	**Strains**	**TRX34**	**Rlv_3841**	**TRX2**	**TRX1**	**TRX3**	**TRX6**	**TRX23**	**VSX4**	**VSX5**	**VSX7**	**VSX9**	**VSX25**	**VSX35**	**TRX4**	**TRX11**	**TRX9**	**TRX22**
A	**TRX34**	---	94.35	94.56	94.36	94.35	94.23	94.54	94.46	94.33	94.21	94.33	94.34	94.62	94.68	94.48	94.27	94.26
B	**3841**	94.36	---	98.55	93.99	93.87	93.74	94.08	94	93.9	93.87	93.89	93.99	94.31	93.63	93.47	93.22	93.22
B	**TRX2**	94.58	98.51	---	94.21	94.18	94.01	94.23	94.15	94.01	93.99	94.01	94.18	94.41	93.86	93.63	93.45	93.47
C	**TRX1**	94.37	93.98	94.21	---	97.09	97.31	97.48	97.33	97.28	97.34	97.42	99.2	97.38	94.5	94.32	93.87	93.88
C	**TRX3**	94.36	93.86	94.19	97.09	---	97.7	97.84	97.78	97.71	98.16	97.7	97.07	97.79	94.47	94.28	93.8	93.83
C	**TRX6**	94.26	93.74	94.03	97.34	97.72	---	98.21	98.55	98.37	98.03	97.91	97.22	98.35	94.34	94.21	93.71	93.75
C	**TRX23**	94.54	94.06	94.22	97.48	97.82	98.15	---	98.02	97.94	98.15	97.9	97.31	98.05	94.53	94.46	93.99	94.01
C	**VSX4**	94.47	93.99	94.15	97.35	97.79	98.54	98.02	---	98.32	98.02	98	97.3	98.38	94.54	94.44	93.96	93.97
C	**VSX5**	94.35	93.89	94.02	97.3	97.72	98.36	97.96	98.33	---	97.97	97.91	97.22	98.26	94.32	94.22	93.77	93.75
C	**VSX7**	94.22	93.85	94	97.37	98.17	98.02	98.18	98.03	97.97	---	97.91	97.26	98.07	94.36	94.24	93.69	93.72
C	**VSX9**	94.34	93.88	94.02	97.44	97.71	97.89	97.92	98	97.9	97.9	---	97.35	97.92	94.35	94.23	93.69	93.74
C	**VSX25**	94.35	93.98	94.17	99.19	97.07	97.21	97.31	97.29	97.22	97.24	97.35	---	97.31	94.4	94.29	93.81	93.85
C	**VSX35**	94.62	94.31	94.41	97.38	97.77	98.31	98.04	98.37	98.24	98.05	97.91	97.3	---	94.7	94.6	94.14	94.1
D	**TRX4**	94.7	93.62	93.87	94.5	94.46	94.32	94.53	94.54	94.32	94.36	94.34	94.41	94.7	---	98.8	95.31	95.3
D	**TRX11**	94.51	93.47	93.62	94.32	94.28	94.2	94.48	94.44	94.21	94.23	94.23	94.3	94.61	98.8	---	95.34	95.34
E	**TRX9**	94.28	93.2	93.46	93.87	93.8	93.7	93.99	93.96	93.76	93.69	93.7	93.8	94.16	95.32	95.33	---	99.17
E	**TRX22**	94.27	93.21	93.47	93.88	93.84	93.75	94.01	93.97	93.75	93.72	93.74	93.86	94.11	95.3	95.34	99.18	---

2.4.5 Phylogenetic analysis of Swedish and Scottish strains

The phylogenetic network (Figure 2.10) indicated the positions of Swedish and Scottish strains related to the five genospecies (A-E) discovered in *R. leguminosarum* strains. Both Scottish strains (VCS_6 and TPS_6) were present in genospecies C. VCS_6 was clustered with VSX_1, VSX_3 and VSX_5, while TPS_6 was clustered with VSX_28, VSX_6, VSX_35 and TRX_26. On the other hand, Swedish strains were scattered in different clusters. TPS_5 and VCS_2 showed high similarity with genospecies D and genospecies E respectively.

Other Swedish strains (VCS_1, VCS_3, VCS_4, VCS_5 and TPS_1) were observed in genospecies A. The 95.04-98.87% ANIm (Table 2.6) between these Swedish strains, USDA_2370^T and TRX_34 (genospecies A) indicated that these Swedish strains belong to genospecies A.

Table 2.6 | Average nucleotide identity (ANIm) analysis of Swedish strains, TRX_34 (genospecies A) and USDA 2370^T (Figure 2.10). Selected Swedish strains were phylogenetically closer to TRX_34 and USDA 2370^T. ANI values are in percentages.

Strains	TRX_34	USDA 2370^T	VCS_1	TPS_1	VCS_3	VCS_4	VCS_5
TRX_34	---	96.13	96.79	97.05	98.09	98.23	96.73
USDA 2370^T	96.01	---	97.12	97.23	95.54	95.04	97.95
VCS_1	96.82	97.39	---	98.37	96.76	96.66	98.87
TPS_1	97.07	97.54	98.38	---	96.63	96.73	98.31
VCS_3	98.1	95.65	96.74	96.62	---	97.8	96.68
VCS_4	98.24	95.13	96.65	96.72	97.8	---	96.65
VCS_5	96.75	98.05	98.82	98.29	96.68	96.66	---

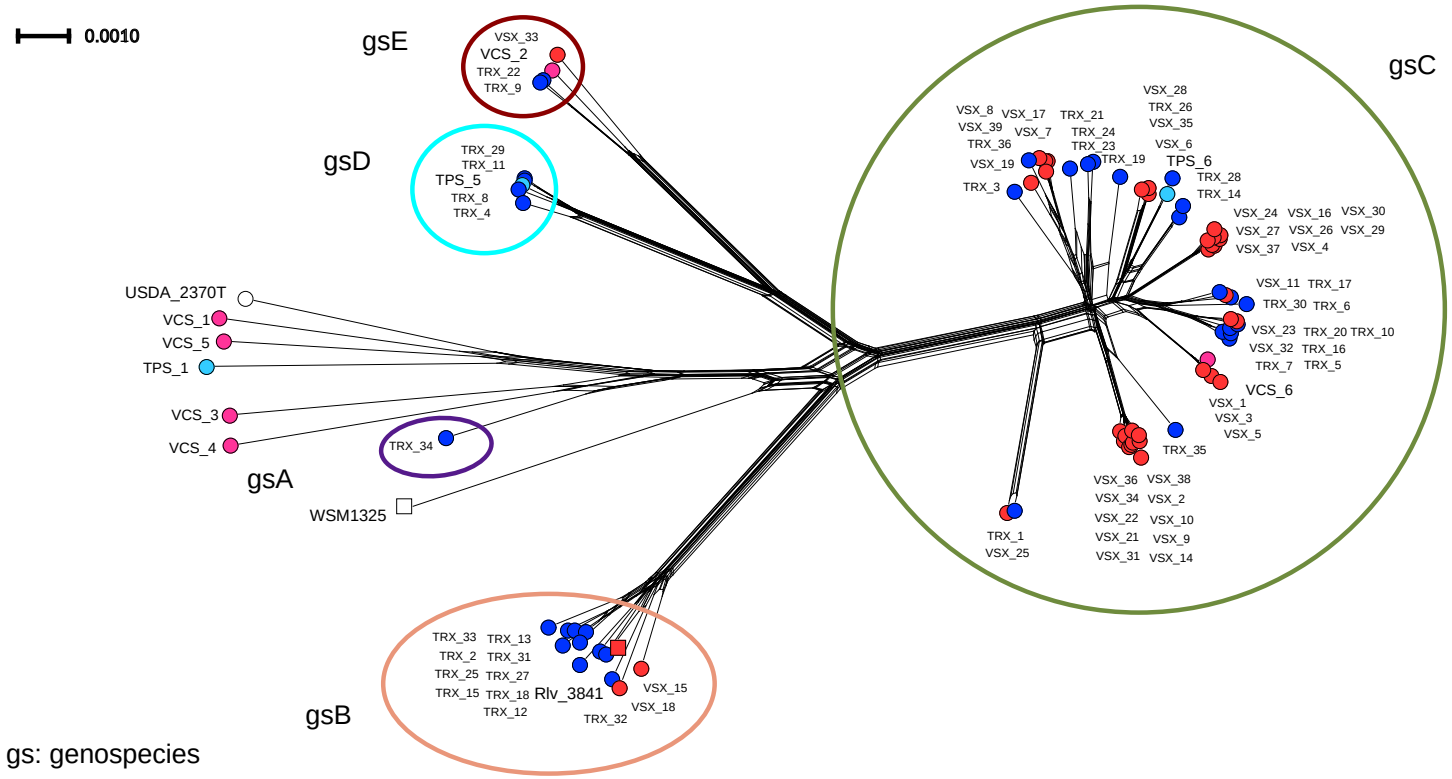


Figure 2.10 | Neighbour-net phylogeny of *R. leguminosarum* strains based on concatenated alignment of 305 core genes. Local population of *R. leguminosarum* is shown in the blue (TRX) and red circle (VSX). Swedish and Scottish strains are represented by dark pink (VCS) and light blue circle (TPS). WSM1325 with white rectangle represents *Rlt* WSM1325. USDA 2370^{T} is shown in white circle. *Rlv* 3841 is shown by the red square.

2.5 Discussion

Both core genome phylogenetic analysis and ANI have shown their ability to delineate the bacterial species (Chan et al., 2012; Scortichini et al., 2013). Although most of the population studies are based on whole genomes or high quality draft genomes, low coverage (averaging around 0.8 X) was sufficient to infer the genomic diversity in draft genomes of *S. medicae* (Bailly et al., 2011). Similarly, low coverage of our dataset (0.68-10 X) was sufficient to infer the taxonomic structure of this population using the methods discussed above. It was also sufficient to perform different genome analysis discussed in other chapters.

2.5.1 Cryptic genospecies in a local population

The phylogenetic analysis based on core genes and ANIm suggests the existence of five genospecies in a population of two symbiovars (36 *trifolii* and 36 *viciae*) of *R. leguminosarum*. These genospecies may be regarded as cryptic genospecies because they are phenotypically indistinguishable, based on our current knowledge of them. Genospecies A contained only one symbiovar *trifolii* strain. Genospecies B contained ten strains of symbiovar *trifolii* and three strains of symbiovar *viciae* including reference genome *Rlv* 3841. Genospecies C comprised thirty-three strains of symbiovar *viciae* and nineteen strains of symbiovar *trifolii*. Genospecies E included one symbiovar *viciae* and two symbiovar *trifolii* strains. A comparison between intra- and inter-genospecies ANI values reveals a gap (≥2%) between them (Table 2.5), so these genospecies are very clearly delineated. These genospecies consist mix strains of two symbiovars, which represent major ecological adaptations in *R. leguminosarum*.

Some previous *Rhizobium* population studies (Santillana et al., 2008; Tian et al., 2010; Vinuesa et al., 2008) have been based on only three housekeeping genes (*recA, glnII* and *atpD*). Recently, the *celC* gene was suggested as a taxonomic marker for *Rhizobium* genomes (Robledo et al., 2011). However, our study provides more robust and reliable results because they are based on 305 housekeeping genes (Harrison et al., 2010) including *recA, glnII, atpD, celC* and 40 universal genes (Ciccarelli et al., 2006 and

Sorek et al., 2007). Phylogenetic diversity shows high levels of sequence divergence in this *R. leguminosarum* population, compared with the population (Figure 2.1) of *S. medicae* at the same location (Bailly et al., 2011).

2.5.2 Phylogenetic structure of the *R. leguminosarum* species complex

In order to observe phylogenetic structure at a broad level, closely related species can be compared with the species of interest (Clarridge, 2004; Fukushima et al., 2002; Tenaillon et al., 2010). The comparison of all five species of the *R. leguminosarum* species complex allowed us to observe the relatedness of our dataset with fully sequenced genomes of this complex. The relatedness of five genospecies with other *R. leguminosarum* (USDA 2370^T, *Rlt* WSM1325 and *Rlt* WSM2304) suggested two additional members of genospecies A: USDA 2370^T (TRX_34: 96.01% ANIm) and *Rlt* WSM1325 (TRX_34: 94.92% ANIm). *Rlt* WSM2304 is a highly diverged strain and may belong to an unknown genospecies of *R. leguminosarum* (less than 90% ANI with genospecies A), or perhaps a new species altogether. On the other hand, ANIm and core gene phylogeny worked well for the other species of this complex. CIAT 652, formerly considered to be *R. etli*, belongs to *R. phaseoli*, which is consistent with Lopez-Guerrero et al. (2012). ANIm value (97.69%) greater than the threshold (96%) between *R. pisi* DSM 30132^T and *R. fabae* CCBAU 33202^T suggests the merger of these species into a single species, for which the name *R. pisi* would take precedence as it was published one month earlier than *R. fabae* (Ramirez-Bahena et al., 2008; Tian et al., 2008).

2.5.3 Cosmopolitan nature of cryptic genospecies

The phylogenetic network (Figure 2.10) based on 305 core genes suggests the existence of these cryptic genospecies in different parts of Sweden and Scotland and not confined to the Wentworth site (York, UK). Scottish strains (VCS_6 and TPS_6) are members of genospecies C. The Swedish strain, TPS_5, is a member of genospecies D. Another Swedish strain, VCS_2, belongs to genospecies E. The rest of the Swedish strains (VCS_1, VCS_3, VCS_4, VCS_5 and TPS_1) are closely clustered and showed high ANI value (>95%) with genospecies A (Table 2.6), thus, they are members of genospecies A. Interestingly, genospecies B (which includes the reference strain *Rlv*

3841) appears to include a lineage of *R. leguminosarum* comprising strains from China, Canada and Spain (Tian et al., 2010). This indicates the cosmopolitan nature of these genospecies, but these results were based on only three housekeeping genes.

Phylogenetic networks (Huson and Bryant, 2006) can be useful to observe reticulate events such as horizontal gene transfer, recombination, etc. and conflicting signals present in the data. The phylogenetic network clearly shows conflicting signals (alternative phylogenetic history) within genospecies and long branches between genospecies indicating less recombination between two genospecies than within genospecies (Figure 2.7). In Chapter 3, we will undertake recombination analysis for 100 core genes (reliable genes) in 75 *R. leguminosarum* strains (72 *R. leguminosarum*, USDA 2370^{T}, *Rlt* WSM1325 and *Rlv* 3841) to infer the role of recombination in shaping the *R. leguminosarum* core genome and to compare inter- and intra-genospecies recombination.

In conclusion, we have observed five cryptic genospecies in the sympatric isolates of two pre-defined symbiovars of *R. leguminosarum*. Normally, ecotypes (Cohan, 2002; Connor et al., 2010; Didelot et al., 2011; Ward et al., 2008) are reflected by distinct species clusters obtained from the genomic analysis of core genes, but these genospecies based on 305 core genes have no relationship with the pre-defined symbiovars, highlighting a lack of ecological barriers between bacterial species. In addition, members from different locations reveal the cosmopolitan nature of these genospecies. The *R. leguminosarum* species can be classified into many cryptic genospecies including the five genospecies observed in this study. We proposed to reclassify *R. pisi* DSM 30132^{T} and *R. fabae* CCBAU 33202^{T} into a new single species.

Chapter 3. Recombination and population structure of cryptic genospecies of *R. leguminosarum*

3.1 Abstract

In the previous chapter, we observed five cryptic genospecies in a predefined species of *R. leguminosarum*. Recombination plays a cohesive force in bacterial species that decreases in strength with an increase in genetic diversity. In this chapter, we investigate the impact of recombination on the evolution of these five genospecies using core genes. First, we analyse the number of core genes that are affected by recombination. Secondly, we observe recombination between and within five genospecies. The results suggest that 89% of core genes were affected by recombination. Finally, we demonstrate that there is a low rate of recombination between genospecies reflecting that some level of genetic isolation exists among genospecies.

3.2 Introduction

Recombination in bacteria is a unidirectional process by which small fragments of DNA are introduced into a recipient cell from a donor cell. Recombination plays an important role in introducing diversity to bacterial strains. Bacterial recombination may be mediated by any of three mechanisms: transformation, transduction and conjugation. Transformation involves uptake of DNA from the environment, transduction involves transfer of DNA from donor to recipient cell by bacteriophages, and conjugation is a process in which DNA transfer occurs between two cells that are in physical contact (Vos, 2009).

Computer simulations (Fraser et al., 2007) showed the relationship between genetic divergence and recombination rate. They showed that change in recombination rate could lead to sexual speciation in bacteria and suggested three properties of bacterial species that arise due to recombination: high rates of recombination in the whole population, more gene transfer within a bacterial species, and reduced rate of gene transfer between two closely related species. Some bacterial studies support this hypothesis, while others do not (Vos and Didelot, 2009), which indicates that this hypothesis is not the sole explanation for all bacterial species.

Bacterial core genomes often show variation in recombination. Numerous studies have revealed the variability of recombination rates in different bacterial species. For example, Didelot et al. (2011) observed the recombination in a population of 114 isolates of *Salmonella enterica* that led to five incipient species within the *S. enterica* subspecies. Likewise, recombination plays a major role in the evolution of *H. pylori* populations (Falush et al., 2003), but not in *Chlamydia trachomatis* (Joseph et al., 2012).

Doroghazi and Buckley (2010) focused on the role of recombination in the evolution of *Streptomyces* species. They found that the rate of intraspecies recombination was more than 100 times that of interspecies recombination, and concluded that conjugation is responsible for gene transfer within these bacterial species.

Variability in recombination rate is also observed in *Rhizobium* species. Acosta et al. (2011) examined the role of recombination in *R. etli* by analysing six draft genomes isolated from different geographical locations and two complete genomes (*R. etli* CFN42 *and R. etli* CIAT652). They concluded that recombination plays a minor role in the evolution of *R. etli* genomes. Moreover, when 12 strains of *Sinorhizobium medicae* were compared using a fully sequenced genome of *S. medicae* WSM419, recombination was found more on the chromid and megaplasmid than on the chromosome (Bailly et al., 2011).

In Chapter 2 we determined the presence of five cryptic genospecies (A-E) in the population of 75 *R. leguminosarum* strains using 305 core genes. This chapter will explore the role of recombination in shaping the core genome of *R. leguminosarum.* For this, we focused on a subset of the 305 core genes for which we have the best coverage. Statistical tests were implemented on these genes to find the number of genes that are strongly affected by recombination. Furthermore, these recombinant genes were scanned to observe horizontal gene transfer events within and between genospecies. Finally, this chapter includes the comparative analysis of recombination rates between and within genospecies using reliable core genes and different bioinformatics tools.

3.2.1 Objectives

The main objectives of this chapter are:

A. Identification of reliable core genes and construction of a Maximum Likelihood tree (100-gene tree) based on these core genes.
B. Identification of core genes that are incongruent with the 100-gene tree.
C. Analysis of each single gene using the Pairwise Homoplasy Index (PHI) test.
D. Identification of core genes that are congruent with other core genes.
E. Clanistic analysis of single gene trees using a matrix based on the 100-gene tree.
F. ClonalFrame and Structure analysis of the core genome.

3.3 Material and Methods

This study was based on 75 strains of *R. leguminosarum* described in Chapter 2. The gsMapper results based on 305 core genes (Appendix table I.I) suggested 100 of these genes that are represented in our data for all strains (minimum one read with minimum length of 100 nucleotides).

FastTree (Price et al., 2010) with settings: -gamma -gtr (most reliable & general model) was run locally on the University of York Biology Linux grid to construct a Maximum Likelihood tree based on the 100-gene alignment. One hundred bootstrap replicates were generated. The tree was visualized using SplitsTree (Huson and Bryant, 2006). An individual maximum likelihood phylogeny of the 100 genes was constructed using PhyML (Guindon et al., 2010) with the best-fit model of nucleotide substitution (Table 3.1) calculated from ModelTest embedded in TOPLAi v2 (Milne et al., 2009).

To compare tree topologies (e.g. single gene trees with 100-gene tree), Shimodaira-Hasegawa (SH) congruence tests implemented in Consel package (Shimodaira and Hasegawa, 2001) were performed ($p < 0.05$: incongruent). Heatmaps for displaying p-values of SH test results were constructed with R package phylcon (Susko et al., 2006). Pairwise Homoplasy Index (PHI) test computed within SplitsTree (Huson and Bryant, 2006) was applied to each of the 100 genes with 5% significance level.

Clanistic analysis was performed using the getDiversity function of the Phangorn (Schliep, 2011) R package to compute perfect clans. ClonalFrame 1.2 (Didelot and Falush, 2007) was applied to our data. Two independent runs of ClonalFrame were performed each consisting of 100,000 MCMC iterations, and the first half was discarded as burn-in. Convergence and mixing of the MCMC were found to be satisfactory by manual comparison of the runs and using Gelman and Rubin's (1992) method implemented in ClonalFrame. Structure v.2.3.4 (Pritchard et al., 2000) was used to identify the hypothetical ancestral populations of our isolates. Initially, ClonalFrame input (concatenated alignment of 100 core genes) was converted into Structure format using xmfa2struct (http://www.xavierdidelot.xtreemhost.com/clonalframe.htm). Four independent runs were performed for a number of populations K ranging from 3 to 9.

For each run, 10^5 burn-in iterations were performed with 10^6 follow-on iterations. Other parameters were used as default. The optimum K value was evaluated by the ΔK method (Evanno et al., 2005). Barplots for Structure results were constructed by R.

Table 3.1 | Models of nucleotide substitution for each of 100 core genes predicted by ModelTest embedded in TOPLAi v2

Locus tag	Functions	Model of nucleotide substitution
RL0012	DNA gyrase subunit B	-a e -c 4 -m GTR
RL0021	tryptophan synthase subunit beta	-a e -c 4 -m TN93
RL0024	FolC bifunctional protein [Includes: folylpolyglutamate synthase (Folylpoly-gamma-glutamate synthetase) (FPGS); dihydrofolate synthase]	-a e -c 4 -v e -m TN93
RL0042	imidazole glycerol phosphate synthase subunit HisF	-a e -c 4 -m HKY
RL0106	30S ribosomal protein S1	-a e -c 4 -m TN93
RL0120	polynucleotide phosphorylase/polyadenylase	-a e -c 4 -m TN93
RL0125	translation initiation factor IF-2	-a e -c 4 -m 012230
RL0127	transcription elongation factor NusA	-a e -c 4 -m TN93
RL0134	DNA polymerase III subunits gamma and tau	-a e -c 4 -m HKY
RL0160	DNA polymerase I	-a e -c 4 -m TN93
RL0161	cell division DNA translocase protein	-a e -c 4 -m GTR -v e
RL0181	transmembrane component of ABC transporter	-a e -c 4 -v e -m HKY
RL0254	GTP-binding protein LepA	-a e -c 4 -m TN93
RL0270	phenylalanyl-tRNA synthetase subunit beta	-a e -c 4 -m TN93
RL0282	exodeoxyribonuclease VII large subunit	-a e -c 4 -m HKY
RL0315	GMP synthase	-a e -c 4 -m HKY
RL0326	4Fe-4S ferredoxin protein	-m F81 -a e -c 4
RL0357	bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase	-m F81 -v e
RL0375	chromosomal replication initiation protein	-a e -c 4 -m HKY
RL0377	coproporphyrinogen III oxidase	-a e -c 4 -m HKY
RL0389	S-adenosylmethionine synthetase	-a e -c 4 -m TN93
RL0394	phosphate starvation-induced protein	-a e -c 4 -m HKY
RL0404	transmembrane mviN virulence factor homologue	-v e -m HKY
RL0406	DNA mismatch repair protein MutS	-a e -c 4 -m HKY
RL0611	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	-a e -c 4 -m TN93
RL0613	histidinol dehydrogenase	-a e -c 4 -m TN93
RL0680	bifunctional preprotein translocase subunit SecD/SecF	-a e -c 4 -m HKY
RL0877	histidyl-tRNA synthetase	-a e -c 4 -m TN93
RL0883	chaperonin GroEL	-a e -c 4 -m TN93
RL0886	bifunctional riboflavin kinase/FMN adenylyltransferase	-a e -c 4 -m HKY
RL0889	isoleucyl-tRNA synthetase	-a e -c 4 -v e -m TN93
RL0892	ribosomal large subunit pseudouridine synthase	-a e -c 4 -m GTR -v e

	B	
RL0910	DNA mismatch repair protein	-a e -c 4 -v e -m 012210
RL0969	23S rRNA (uracil-5-)-methyltransferase	-a e -c 4 -m HKY
RL0973	1-deoxy-D-xylulose-5-phosphate synthase	-a e -c 4 -m HKY
RL1543	cysteinyl-tRNA synthetase	-a e -c 4 -m 012230
RL1546	amidophosphoribosyltransferase	-a e -c 4 -m TN93
RL1548	DNA repair protein RadA	-m F81
RL1551	replicative DNA helicase	-a e -c 4 -m TN93
RL1605	aspartyl-tRNA synthetase	-a e -c 4 -m TN93
RL1620	serine hydroxymethyltransferase	-a e -c 4 -m TN93
RL1621	riboflavin biosynthesis protein	-a e -c 4 -m 012314
RL1723	DNA polymerase III subunit alpha	-a e -c 4 -m TN93
RL1735	DNA topoisomerase I	-a e -c 4 -m TN93
RL1767	DNA-directed RNA polymerase subunit beta'	-a e -c 4 -m TN93
RL1771	elongation factor G	-a e -c 4 -m TN93
RL1777	50S ribosomal protein L2	-a e -c 4 -m HKY
RL2035	valyl-tRNA synthetase	-a e -c 4 -v e -m TN93
RL2041	arginyl-tRNA synthetase	-a e -c 4 -m TN93
RL2048	Sec-independent protein translocase protein	-a e -c 4 -m HKY
RL2049	seryl-tRNA synthetase	-a e -c 4 -m HKY
RL2055	bifunctional preprotein translocase subunit SecD/SecF	-a e -c 4 -m HKY
RL2099	single-stranded-DNA-specific exonuclease	-a e -c 4 -m TN93
RL2288	siroheme synthase	-a e -c 4 -m HKY
RL2381	bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase	-a e -c 4 -m 010020
RL2384	ATP-dependent DNA helicase RecG	-a e -c 4 -m HKY
RL2386	transcription-repair coupling factor	-a e -c 4 -m TN93
RL2392	glutamine synthetase I	-a e -c 4 -v e -m TN93
RL2398	excinuclease ABC subunit A	-a e -c 4 -m 012230
RL2401	DNA gyrase subunit A	-a e -c 4 -v e -m TN93
RL2511	CTP synthetase	-a e -c 4 -m TN93
RL2588	tyrosyl-tRNA synthetase	-a e -c 4 -m TN93
RL2636	alanyl-tRNA synthetase	-a e -c 4 -m TN93
RL2691	ATP-binding component of ABC transporter	-a e -c 4 -v e -m 012210
RL2957	excinuclease ABC subunit B	-a e -c 4 -m HKY
RL3245	acetolactate synthase 3 catalytic subunit	-a e -c 4 -m TN93
RL3276	ATP-dependent DNA helicase	-a e -c 4 -m 012230
RL3301	D-alanine--D-alanine ligase	-a e -c 4 -m HKY
RL3306	UDP-N-acetylmuramate--L-alanine ligase	-a e -c 4 -m GTR
RL3310	phospho-N-acetylmuramoyl-pentapeptide-transferase	-a e -c 4 -m TN93
RL3311	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	-a e -c 4 -m HKY
RL3313	penicillin binding protein	-a e -c 4 -m TN93
RL3402	RNA polymerase sigma factor RpoD	-a e -c 4 -m TN93
RL3408	DNA primase	-a e -c 4 -m TN93
RL3419	carbamoyl phosphate synthase large subunit	-a e -c 4 -m TN93
RL3521	anthranilate synthase	-a e -c 4 -m TN93
RL3768	adenylosuccinate synthetase	-a e -c 4 -m TN93
RL3965	cell division protein FtsH	-a e -c 4 -m TN93

RL3990	Holliday junction DNA helicase RuvB	-a e -c 4 -m HKY
RL4006	transketolase	-a e -c 4 -m TN93
RL4060	pyruvate kinase	-a e -c 4 -v e -m TN93
RL4085	glutamate synthase [NADPH] large chain	-a e -c 4 -m TN93
RL4184	glutamyl-tRNA synthetase	-a e -c 4 -m HKY
RL4279	chaperone ClpB (heat-shock protein)	-a e -c 4 -m TN93
RL4298	preprotein translocase subunit SecA	-a e -c 4 -m TN93
RL4412	primosome assembly protein PriA	-a e -c 4 -v e -m HKY
RL4506	GTP-binding protein TypA/BipA (tyrosine phosphorylated protein A)	-a e -c 4 -m TN93
RL4507	peptidyl-dipeptidase	-a e -c 4 -m 012230
RL4515	argininosuccinate synthase	-a e -c 4 -m TN93
RL4563	transmembrane DNA translocase	-a e -c 4 -m TN93
RL4630	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	-a e -c 4 -m HKY
RL4707	3-isopropylmalate dehydrogenase	-a e -c 4 -m HKY
RL4722	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase	-a e -c 4 -m 012230
RL4732	leucyl-tRNA synthetase	-a e -c 4 -m HKY
RL4736	chromosome partitioning protein	-a e -c 4 -m 012345
RL4739	tRNA modification GTPase TrmE	-a e -c 4 -v e -m 010010
pRL120279	putative exported tail-specific protease precursor	-a e -c 4 -m HKY
pRL120416	putative alanine racemase	-a e -c 4 -m HKY
pRL120642	chaperonin GroEL	-a e -c 4 -v e -m TN93
pRL110033	putative ATP-binding component of ABC transporter	-a e -c 4 -m HKY

3.4 Results

3.4.1 Phylogenetic structure based on reliable core gene

The Maximum Likelihood tree based on concatenated alignment of 100 core genes (Figure 3.1) displayed strong support (Appendix figure II.I), except for a few internal branches, for the same five genospecies (A-E) as discovered in the previous chapter. This tree is termed the 100-gene tree (Figure 3.1) and used for further analyses such as SH tests.

3.4.2 Phylogenetic incongruence and Intragenic recombination

We were curious to know the number of genes that are incongruent with the 100-gene tree. SH test was used to find the incongruent genes by comparing single gene ML trees with the 100-gene tree. SH tests discovered that 63% of the genes were incongruent with the 100-gene tree (Table 3.2). These are shown in blue, while green shows congruent genes.

Meanwhile, PHI test was used to detect intragenic recombination in each of the 100 gene sequences, which indicated that 72% of the genes were recombinant (Table 3.2, with the same color code of SH tests). Based on SH tests and PHI tests, we were surprised that most (89 of 100) of the reliable core genes showed significant evidence of recombination.

Next, we examined 49 putative recombinant genes (in bold in Table 3.2) that failed the SH tests as well as identified by PHI tests. These genes were compared with each other using SH tests to observe the number of genes that are congruent with each other and share the same evolutionary history. SH tests results (p values) are displayed as a heatmap (Figure 3.2). In Figure 3.2, each column and row represent an individual gene marker and gene topology respectively. The diagonal line represents the P value of tree topologies with their corresponding gene markers. Most of the topologies were

incongruent with many gene markers, corresponding to the white color, indicating that these genes have a different evolutionary history. However, there are a few topologies that are congruent with many gene markers. These topologies were detected by performing hierarchical clustering (Figure 3.3) on the same dataset. Figure 3.3 suggested the presence of four topologies (bottom of Figure 3.3) that are, between them, congruent with 30 gene markers.

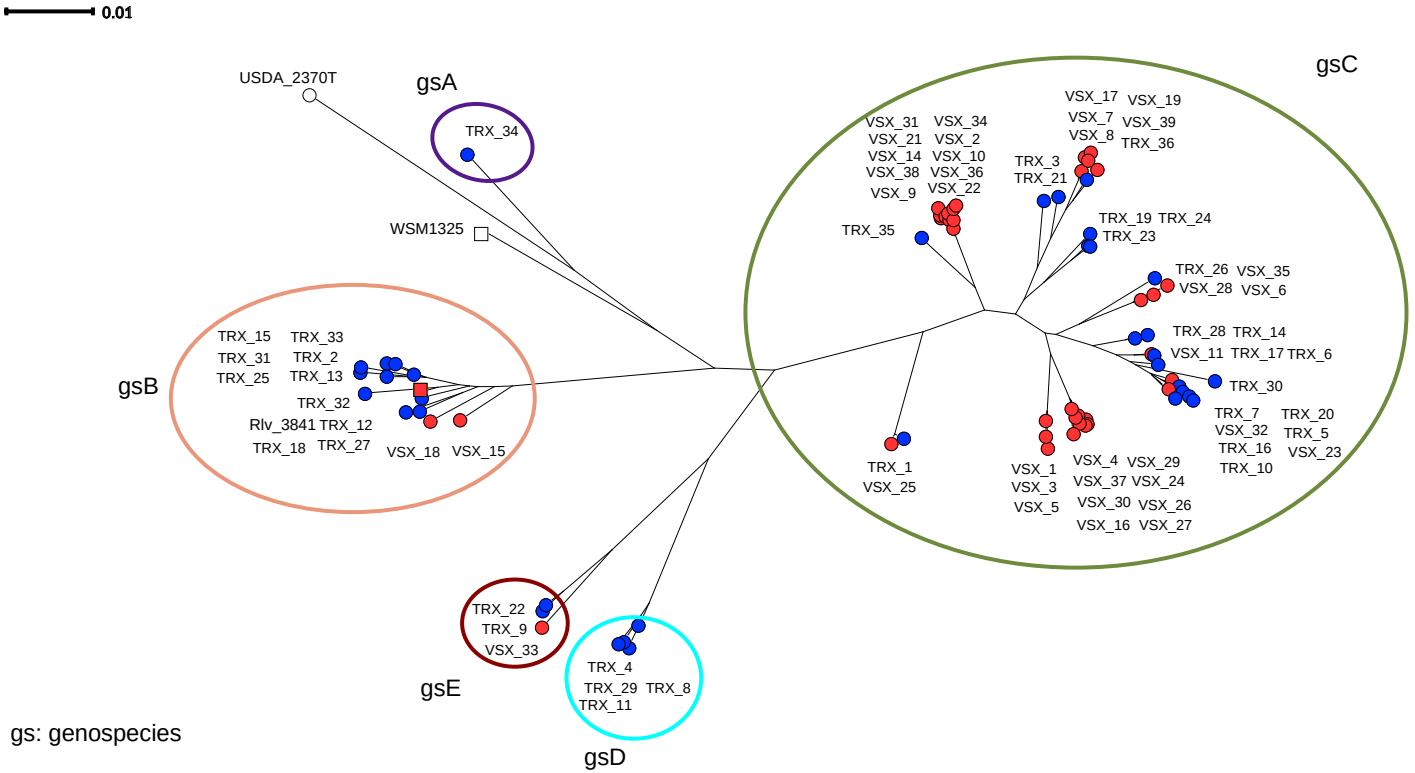


Figure 3.1 | Maximum Likelihood tree based on 100-gene alignment showing the position of 75 *R. leguminosarum* strains (TRX: blue circle & VSX: red circle). WSM1325 (white square) represents *Rlt* WSM1325. USDA_2370^{T} is shown by white circle. *Rlv* 3841 is shown by red square.

Table 3.2 | Results from SH tests and PHI tests for 100 core genes. In both test, green: $p > 0.05$ and blue: $p < 0.05$. There are just 11 genes (highlighted in red) with no significant evidence for recombination. Genes that were incongruent with the 100-gene tree (SH test) and had evidence of intragenic recombination (PHI test) are shown in bold with asterisk.

Locus tag	SH tests	PHI
RL0012*	0.002	6.56E-07
RL0021*	7.00E-05	6.44E-04
RL0024*	0.023	6.21E-09
RL0042	0.266	1.09E-04
RL0106	0.121	3.61E-03
RL0120	0.125	4.69E-01
RL0125*	0	1.25E-36
RL0127	0.037	2.17E-01
RL0134*	0.001	3.02E-15
RL0160*	0	1.46E-04
RL0161	0	2.79E-01
RL0181	0.013	1.52E-01
RL0254	0.006	1.26E-01
RL0270*	0.002	1.00E-03
RL0282*	3.00E-05	2.03E-07
RL0315	0.156	3.64E-11
RL0326*	0	1.79E-03
RL0357*	0	8.89E-23
RL0375	0.42	3.36E-06
RL0377*	0	3.88E-05
RL0389	0.013	6.80E-01
RL0394	0.021	2.02E-01
RL0404	0.47	9.82E-01
RL0406	0.092	1.57E-09
RL0611*	0.021	6.05E-06
RL0613	0	3.27E-01
RL0680*	0.027	3.23E-04
RL0877*	0	1.67E-07
RL0883	0.084	8.74E-05
RL0886	0.143	4.22E-01
RL0889*	2.00E-04	5.35E-03
RL0892	0.118	1.34E-05
RL0910	0.232	7.21E-03
RL0969	0.067	2.04E-03
RL0973	0.079	4.26E-03

RL1543*	0.019	1.21E-02
RL1546*	0	3.62E-06
RL1548	1.00E-04	8.51E-01
RL1551*	0.001	1.82E-24
RL1605*	0	1.61E-10
RL1620	0.096	3.09E-01
RL1621	0.084	9.84E-01
RL1723*	4.00E-04	2.51E-11
RL1735*	0	1.50E-05
RL1767*	0.004	9.48E-05
RL1771	0.006	5.67E-02
RL1777*	0.005	4.28E-02
RL2035	0.238	1.23E-06
RL2041	0.001	2.13E-01
RL2048	0.11	3.57E-03
RL2049	0.393	9.15E-02
RL2055	0.061	8.34E-02
RL2099*	0	4.43E-03
RL2288*	0	4.12E-04
RL2381*	0.002	4.70E-03
RL2384	0.153	1.60E-14
RL2386*	0	8.48E-19
RL2392*	0	0.00E+00
RL2398*	0	3.63E-02
RL2401*	2.00E-05	0.00E+00
RL2511*	1.00E-04	2.86E-04
RL2588	0.332	5.27E-01
RL2636	0.069	1.23E-18
RL2691*	0.021	5.66E-05
RL2957*	0	1.15E-08
RL3245	0.241	2.45E-01
RL3276	0.056	4.12E-16
RL3301*	0	1.00E-02
RL3306*	0.008	1.56E-06
RL3310*	0	2.91E-03
RL3311	0.093	3.07E-02
RL3313*	0.011	2.83E-02
RL3402	0.192	1.34E-03
RL3408	0.2	3.01E-01
RL3419*	0.015	3.00E-37
RL3521	2.00E-04	2.15E-01
RL3768*	0.022	5.62E-04
RL3965	0.094	6.49E-02
RL3990*	0.022	9.04E-03

RL4006	0.322	2.51E-01
RL4060	0.074	7.41E-03
RL4085	0.085	0.00E+00
RL4184	0.418	8.50E-03
RL4279*	2.00E-04	2.89E-09
RL4298	0.277	9.94E-01
RL4412*	0	6.03E-05
RL4506	0.093	1.36E-08
RL4507*	0	2.61E-07
RL4515*	0	1.35E-10
RL4563	0.218	6.16E-02
RL4630	0	5.02E-01
RL4707	0.045	5.63E-01
RL4722	0.157	5.38E-11
RL4732*	4.00E-04	6.19E-07
RL4736	0.222	1.94E-03
RL4739*	0.003	1.07E-03
pRL120279*	0	1.77E-02
pRL120416*	0	4.16E-08
pRL120642	0	1.03E-01
pRL110033*	0.002	6.17E-05

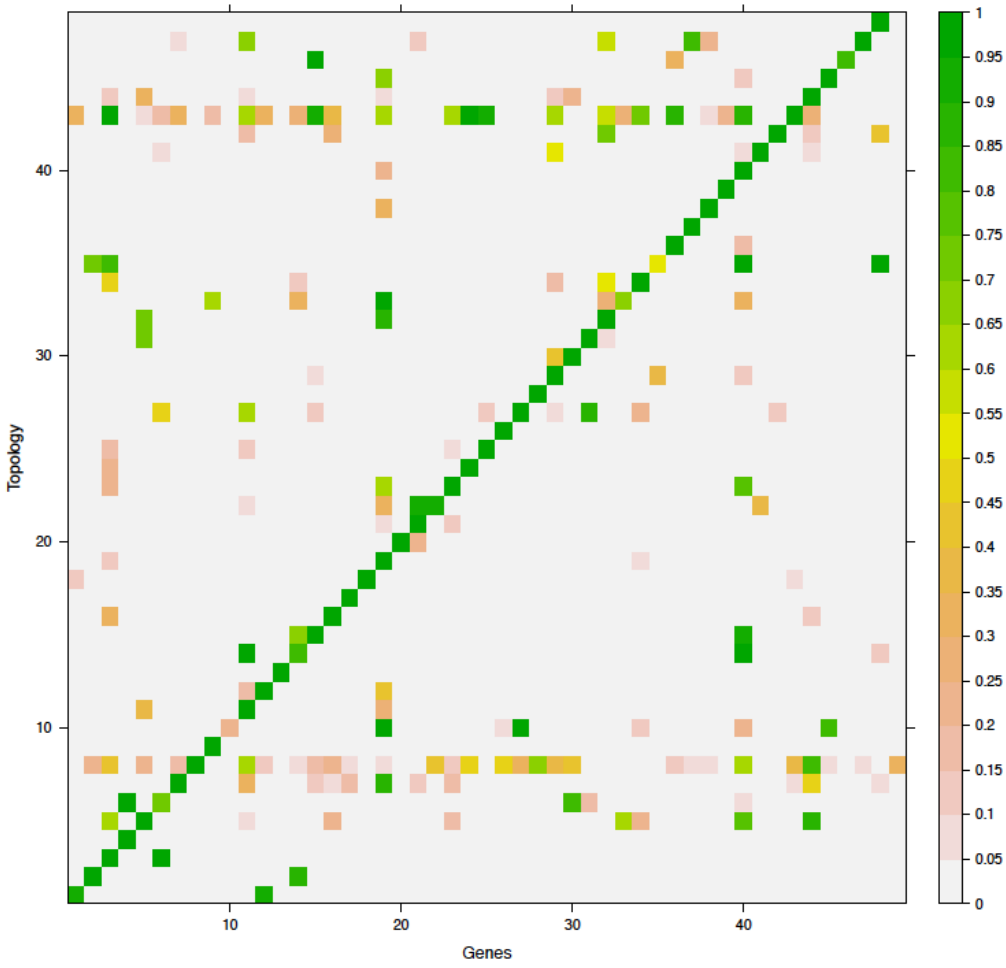


Figure 3.2 | Heatmap showing SH test results of 49 core genes that are highly recombinant. Columns: each gene, and rows: each topology. Significance level: 5%. The cells of the heatmap were colored according to the P values of the SH test (white: $P < .05$, other colours: $P \geq .05$).

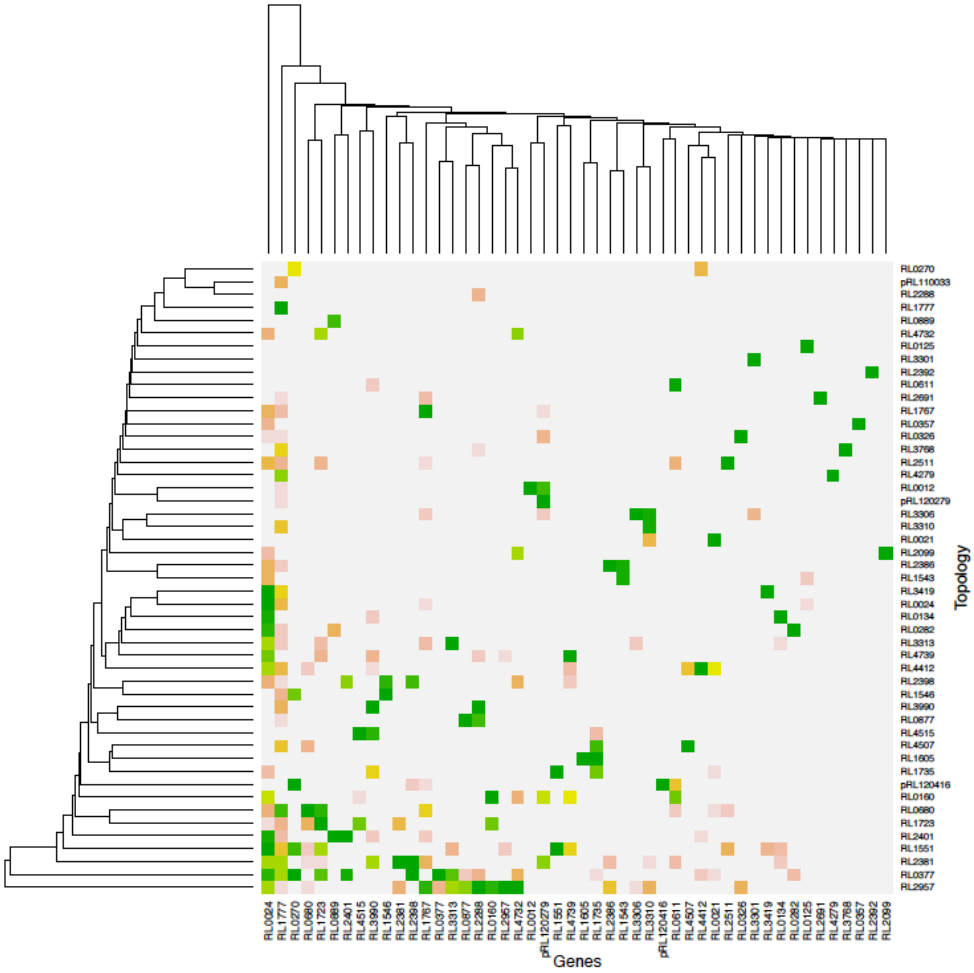


Figure 3.3 | Heatmap showing the hierarchical clustering (row and column) of SH test results of 49 core genes to obtain similarity in patterns of P values. Columns: each gene, and rows: each topology. The same color-coding was used as in Figure 3.2

The four topologies (Figure 3.4–3.7) were mid-rooted using Dendroscope (Huson and Scornavacca, 2012) and examined to identify transfer events in putatively recombinant genes. Reliable branches in each tree were detected using 100 bootstrap replicates. In every tree, strain nodes are colored according to their genospecies (A: purple, B: salmon, C: green, D: cyan and E: dark red) and branches with a bootstrap value greater than 70% are shown in red (compatible with the 100-gene tree in Appendix figure II.I) and blue (incompatible). The maximum number of reliable branches was observed in the RL2957 phylogeny (Figure 3.4). In comparison to the 100-gene tree, these phylogenies showed several reliable examples of horizontal gene transfer that occurred within genospecies (discussed below). The single reliable inter-genospecies event was observed in the RL2381 phylogeny (Figure 3.6).

The RL2957 phylogeny (Figure 3.4) suggested many gene transfer events that occurred within genospecies. For example, a member of genospecies C, TRX_21, was observed in a different sub-cluster (VSX_23, TRX_10, TRX_5, VSX_32, TRX_7, TRX_16, TRX_20) of genospecies C. Also, this sub-cluster was grouped with a different sub-cluster (VSX_7, VSX_8, VSX_39, VSX_17, and TRX_36) of genospecies C.

A strongly supported sub-cluster (TRX_19, TRX_23, TRX_24, TRX_1, TRX_25) of genospecies C was observed with a strong cluster (TRX_18, VSX_18, TRX_27, TRX_12, TRX_33, TRX_2, TRX_25, TRX_15, TRX_13) of genospecies B, but these groups were poorly supported by bootstrap values.

The RL0377 phylogeny (Figure 3.5) was able to delineate the five genospecies (A-E) observed in 100-gene tree, but without any strong bootstrap branch that conflicts with branches of 100-gene tree.

The phylogeny of RL2381 (Figure 3.6) indicated the occurrence of intra-species transfer in genospecies C. The inclusion of the closely related members (VSX_28, VSX_6, VSX_35) and TRX_35 into the sub-cluster comprised of TRX_36, VSX_7, VSX_8, VSX_39, VSX_17 and VSX_19 resulted into the formation of new sub-cluster in this gene tree. Interestingly, this phylogeny also provided an example of inter- genospecies transfer (supported by bootstrap value) that occurred between the genospecies B and genospecies A (TRX_34 and *Rlt* WSM1325).

The phylogeny of RL1551 (Figure 3.7) suggested the evidences of intra- genospecies transfer in the members of genospecies B. For example, a group comprised of VSX_15, VSX_18, TRX_18 and *Rlv* 3841 (reference genome) was strongly supported.

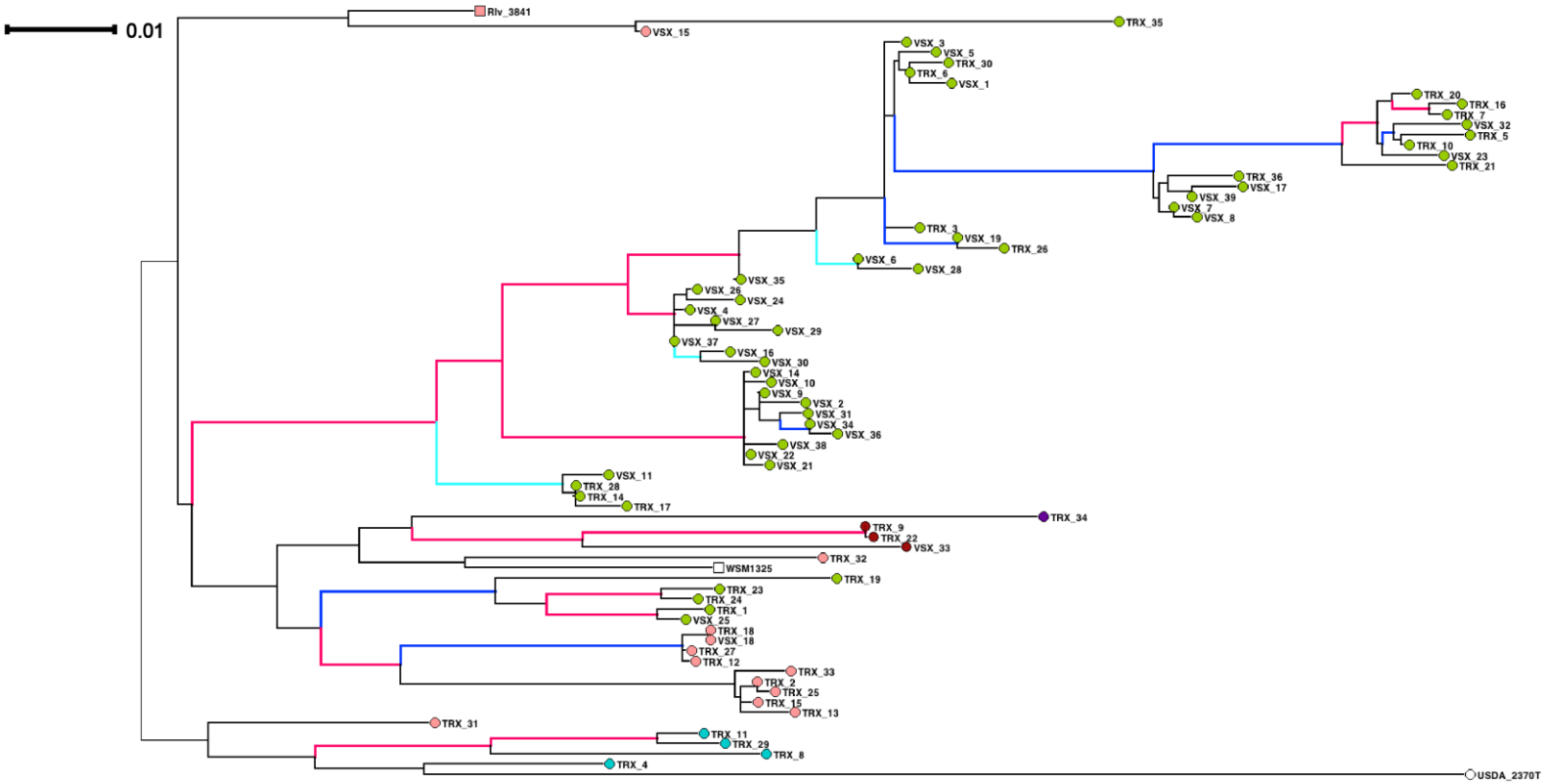


Figure 3.4 | The Maximum Likelihood tree of RL2957. Strain nodes are colored on the basis of their genospecies (A: purple, B: salmon, C: dark green, D: cyan and E: dark red). Square boxes represent full sequenced genomes. Branches with bootstrap values > 70% are colored. Branches congruent with the 100-gene tree are in red. Incongruent branches are in dark blue. Branches in light blue are poorly supported in 100-gene tree.

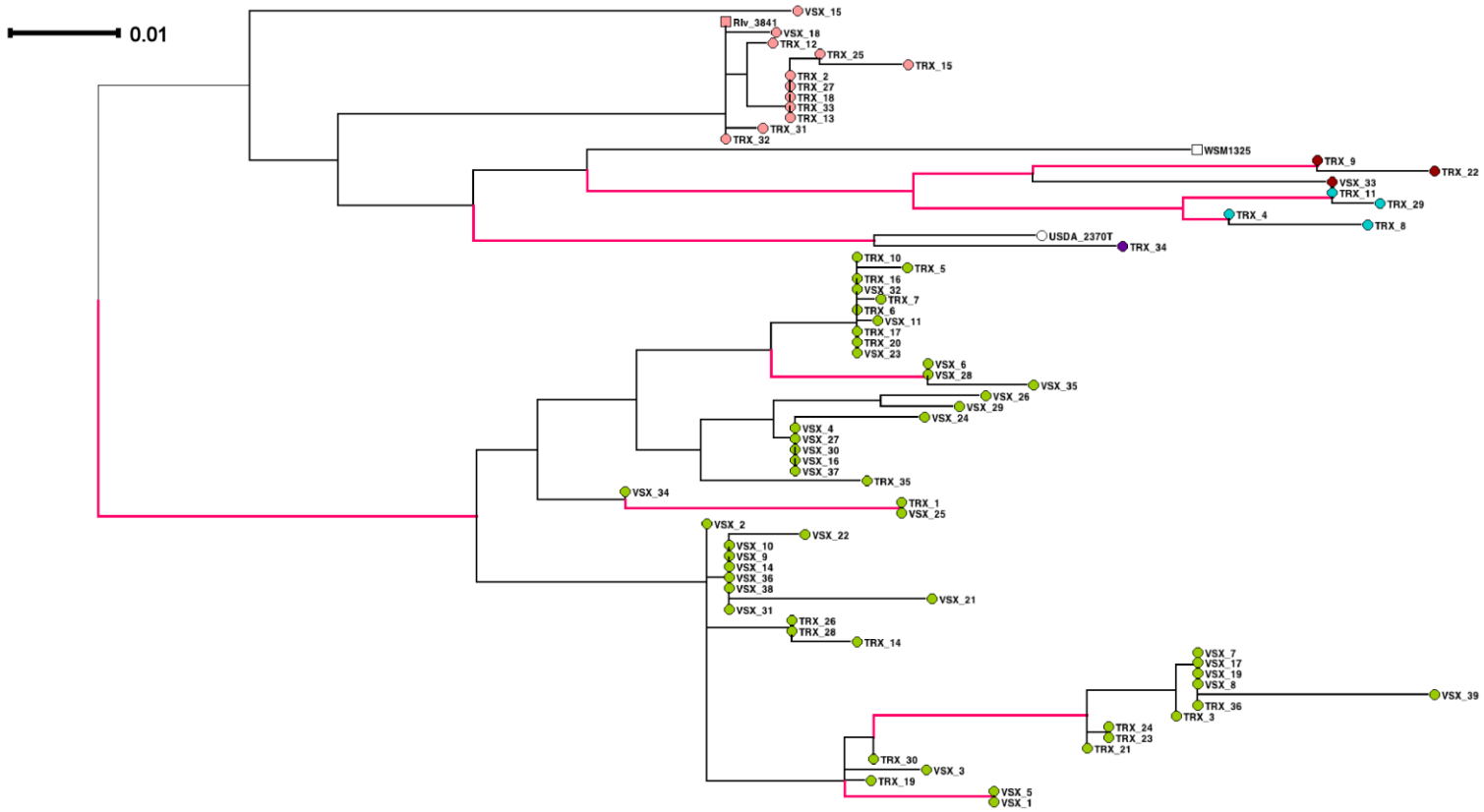


Figure 3.5 | The Maximum Likelihood tree of RL0377. The same color-coding and symbols were used for strains as in Figure 3.4.

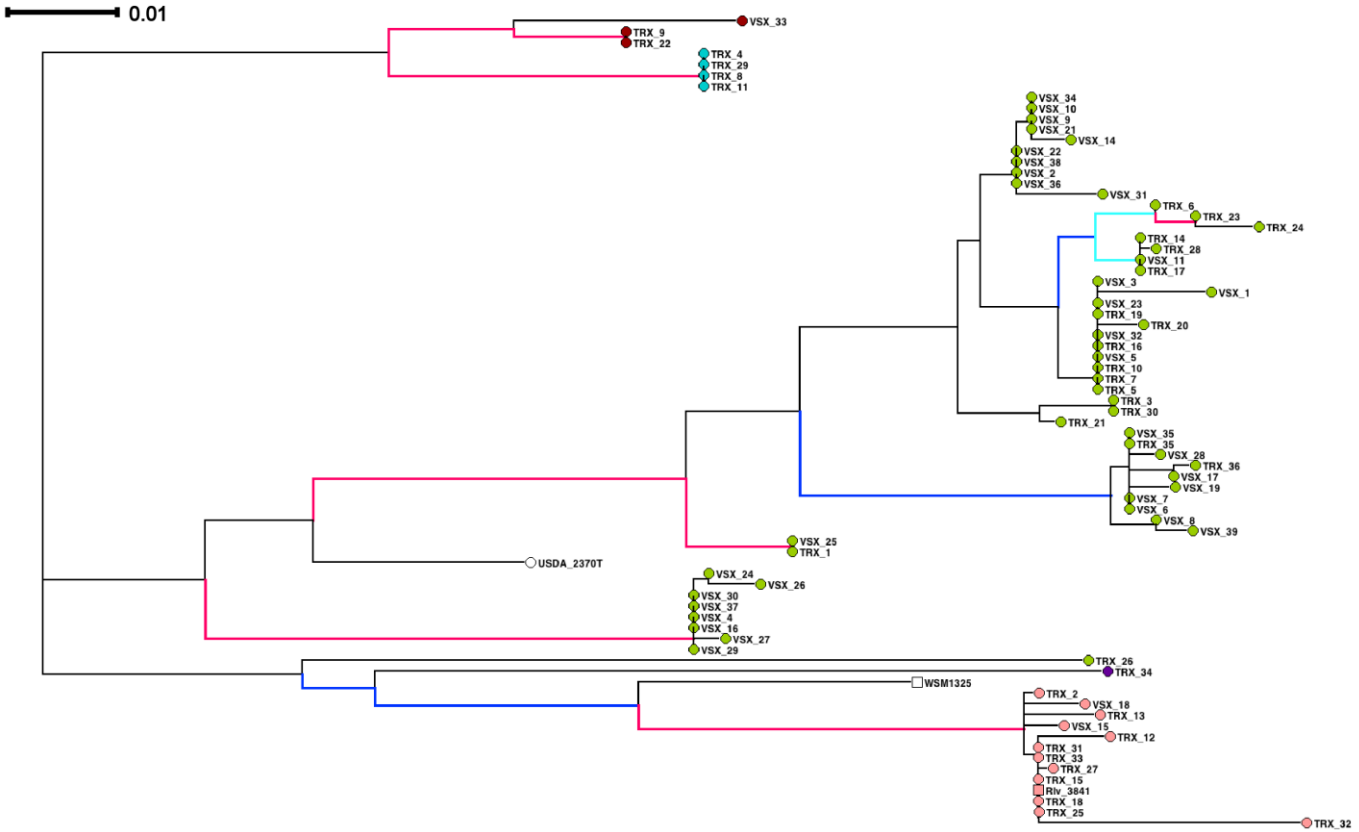


Figure 3.6 | The Maximum Likelihood tree of RL2381. The same color-coding and symbols were used for strains as in Figure 3.4.

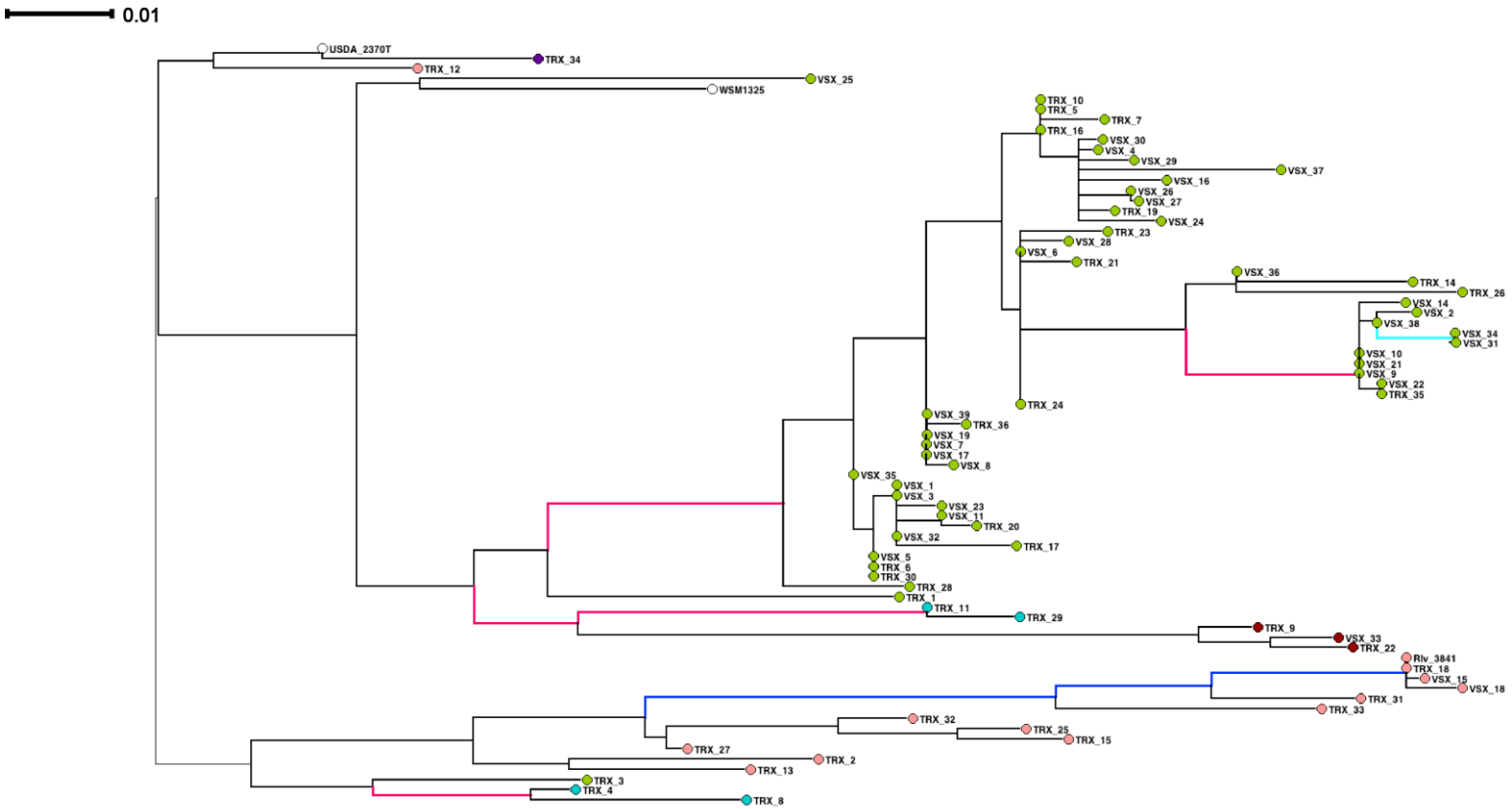


Figure 3.7 | The Maximum Likelihood tree of RL1551. The same color-coding and symbols were used for strains as in Figure 3.4

3.4.3 Intra- and Inter- genospecies recombination

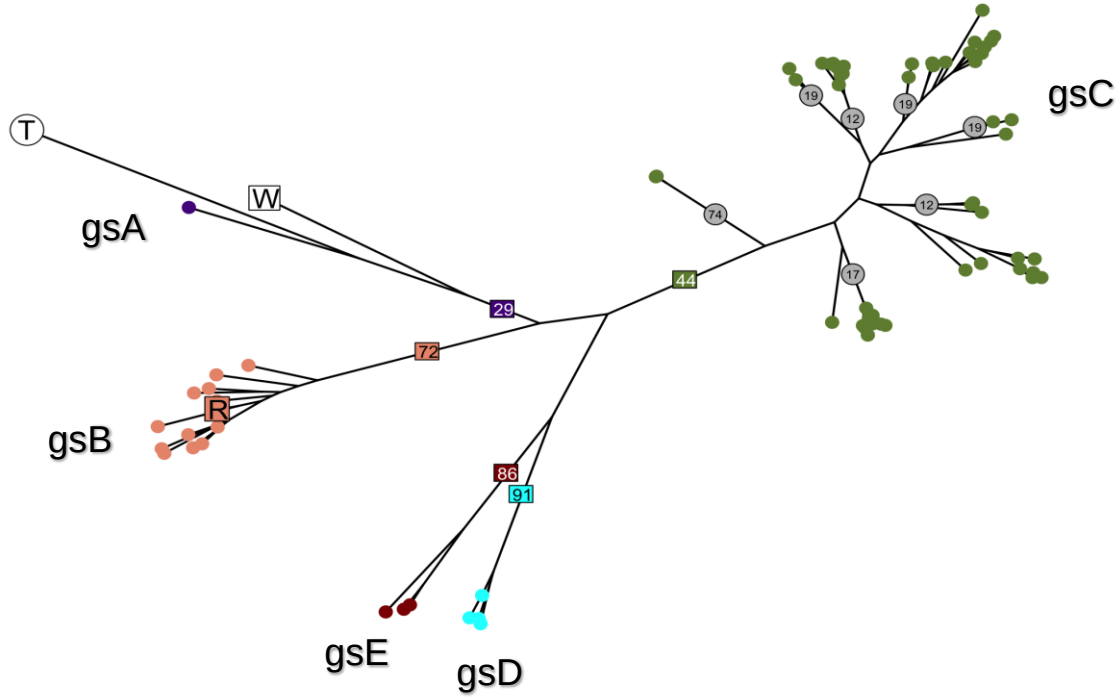
Three different strategies (Clanistic analysis, ClonalFrame and Structure) were used to observe recombination between and within the five genospecies. Results of these three strategies are described below.

3.4.3.1 Clanistic analysis

This analysis calculates the number of occurrences of each of the five genospecies as a clan in the set of 100 single-gene ML trees. Three genospecies (B, D & E) were conserved in most of the gene trees (Table 3.3 and Fig. 3.8). Also, genospecies C was observed in half (approx.) of the gene trees. However, genospecies A was observed only in 29 gene trees suggesting these strains are highly diverged from each other. Moreover, we observed the number of occurrences of sub-clans present in genospecies C. Most of the sub-clans are conserved in fewer than 20 gene trees (Fig. 3.8) except for one sub-clan (TRX_1 and VSX_25), which was conserved in 74 gene trees.

Table 3.3 | Results of Clanistic analysis of five genospecies (gs).

Clans	Strains	No of occurrences in 100 gene trees
gsA	3	29
gsB	13	72
gsC	52	44
gsD	4	86
gsE	3	91



gs: genospecies

Figure 3.8 | Results of Clanistic analysis. Numbers on each branch reflect the number of occurrences of that clan in 100 ML gene trees. *Rlt* WSM1325 (W) is shown by white square and USDA_2370^{T} (T) is shown by white circle. *Rlv* 3841 is shown by salmon square. Strains are colored on the basis of their genospecies (A: purple, B: salmon, C: dark green, D: cyan and E: dark red).

3.4.3.2 ClonalFrame

ClonalFrame identifies clonal relationships among isolates of a population and estimates a quantitative value for mutation and recombination events. ClonalFrame estimates two statistical values (ρ/θ & r/m), where ρ/θ is the ratio of rates at which recombination and mutation occur, whereas r/m calculates the relative impact of recombination relative to mutation. ClonalFrame produced the same five clades of genospecies as found in the 100-gene tree. The mean value of ρ/θ and the r/m ratio (Table 3.4) for the whole population are 1.32 and 5.9 respectively indicating that the population is highly affected by recombination (thresholds described by Vos and Didelot, 2009). To observe recombination within genospecies, we performed ClonalFrame on genospecies C and genospecies B (most of the strains are clustered in these groups). The high mean value of ρ/θ and the r/m value for genospecies C and genospecies B (Table 3.4) reflected that recombination occurred more frequently than mutation within these genospecies.

Table 3.4 | Results of ClonalFrame analysis

	Strains	ρ/θ	r/m
Whole	75	1.32	5.92
Genospecies C	52	0.787	4.29
Genospecies B	13	26.59	102.93

3.4.3.3 Structure

The population structure of *R. leguminosarum* was observed by Structure v.2.3.4. The optimal K value was evaluated by the ΔK method (Evanno et al., 2005). In Figure 3.9, different K (4-9) values are presented on the X-axis, while the Y-axis shows the values of Delta K at each K population. The K value with the highest peak is considered to be the optimal K value (our study: 5). These five hypothetical ancestral populations mirror the genospecies (A-E). A detailed Structure bar graph (Figure 3.10) was constructed in which strains are classified according to their respective genospecies (from left to right: genospecies A-E) to display the proportion of hypothetical ancestral population (different colors) present in each strain. Results indicated that there was little admixture in this population.

Figure 3.10 demonstrates that genospecies A got foreign DNA from all other genospecies (B-E). Genospecies (B and C) got foreign DNA from genospecies A only. Genospecies D and E shared DNA with each other. Both of these species acquired some foreign DNA from genospecies A but not from genospecies B and C. Moreover, some strains were observed that are entirely from one ancestral population for example USDA_2370^{T} (genospecies A), TRX_2 (genospecies B), VSX_7 and VSX_9 both from genospecies C and TRX_11 (genospecies D), suggesting that these strains have never received DNA from another genospecies. The results suggest that the two large genospecies (genospecies B and C) have never exchanged DNA with each other. The K=6 (Figure 3.11) and K=7 (Figure 3.12) analyses continue to support the five strong clusters that echo genospecies (A-E) and the lack of mixture of B and C. They also suggest that the acquired genes in B and C are not actually from A, but from outside the five genospecies we have sampled. The TRX_2 (genospecies B), VSX_9 (genospecies C) and TRX_11 (genospecies D) still belonged to one Structure population.

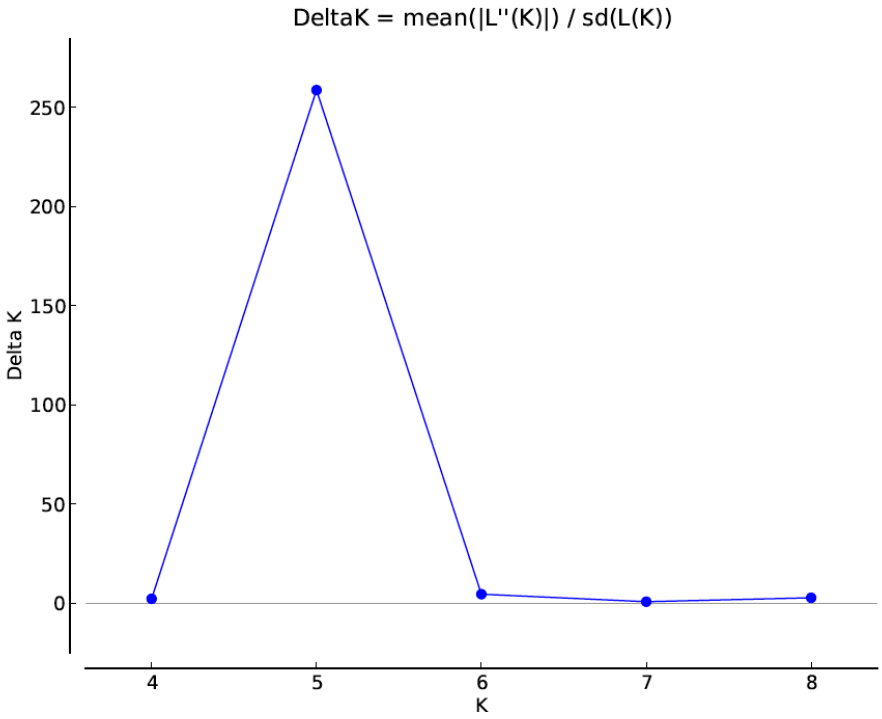


Figure 3.9 | ΔK-values for different K; suggesting K = 5 as the most likely structure population according to Evanno et al. (2005)

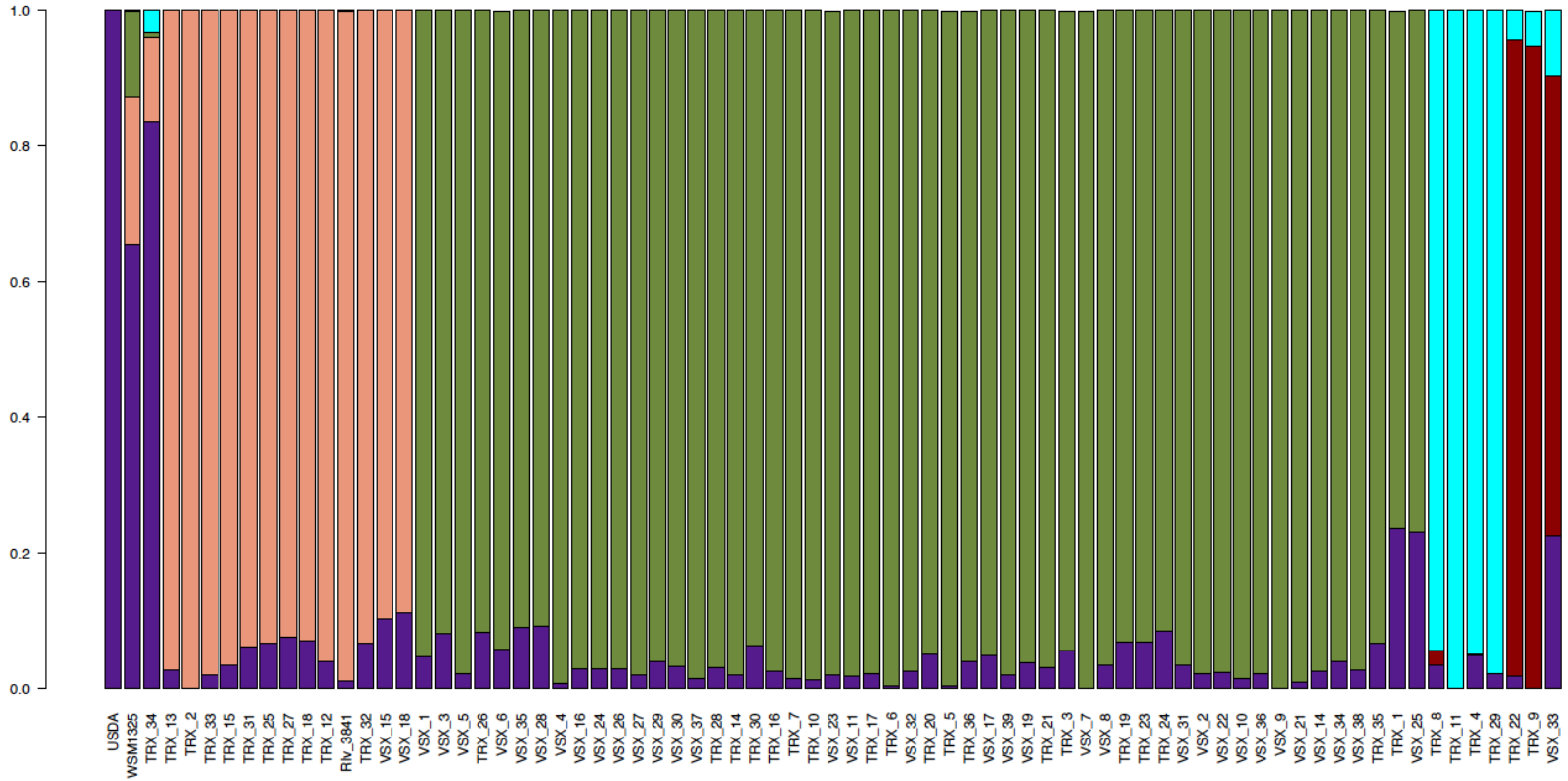


Figure 3.10 | Bar graph showing results of Structure analysis with K = 5: Each vertical bar represents one of the 75 *R. leguminosarum* strains. The coloring of each bar is proportional to the ancestry of each strain from each of the 5 populations (Y-axis: purple, dark salmon, dark green, cyan and dark red respectively).

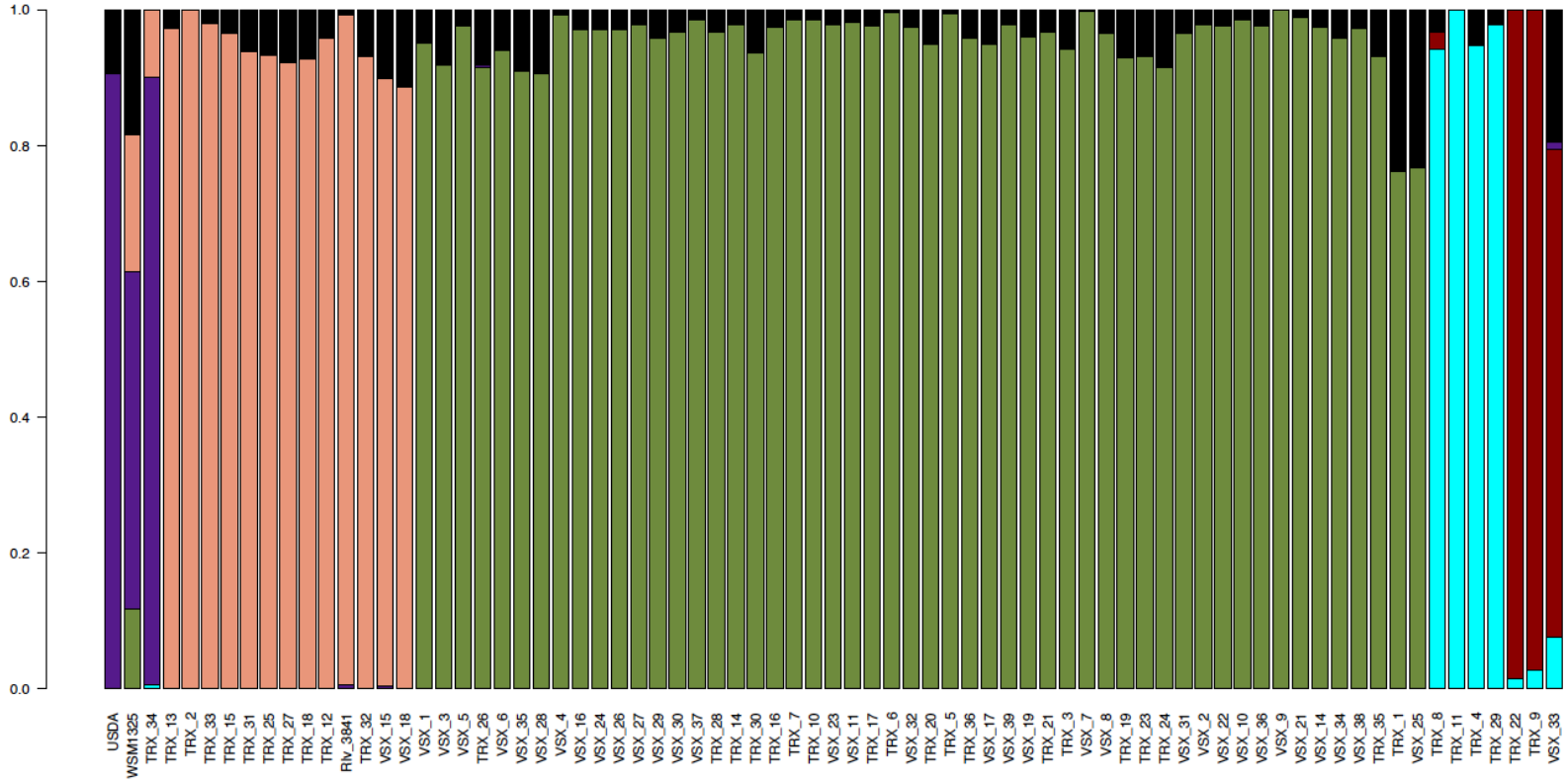


Figure 3.11 | Bar graph showing results of Structure analysis with K = 6: Each vertical bar represents one of the 75 *R. leguminosarum* strains. The coloring of each bar is proportional to the ancestry of each strain from each of the 6 populations (Y-axis: purple, dark salmon, dark green, cyan, dark red and black respectively).

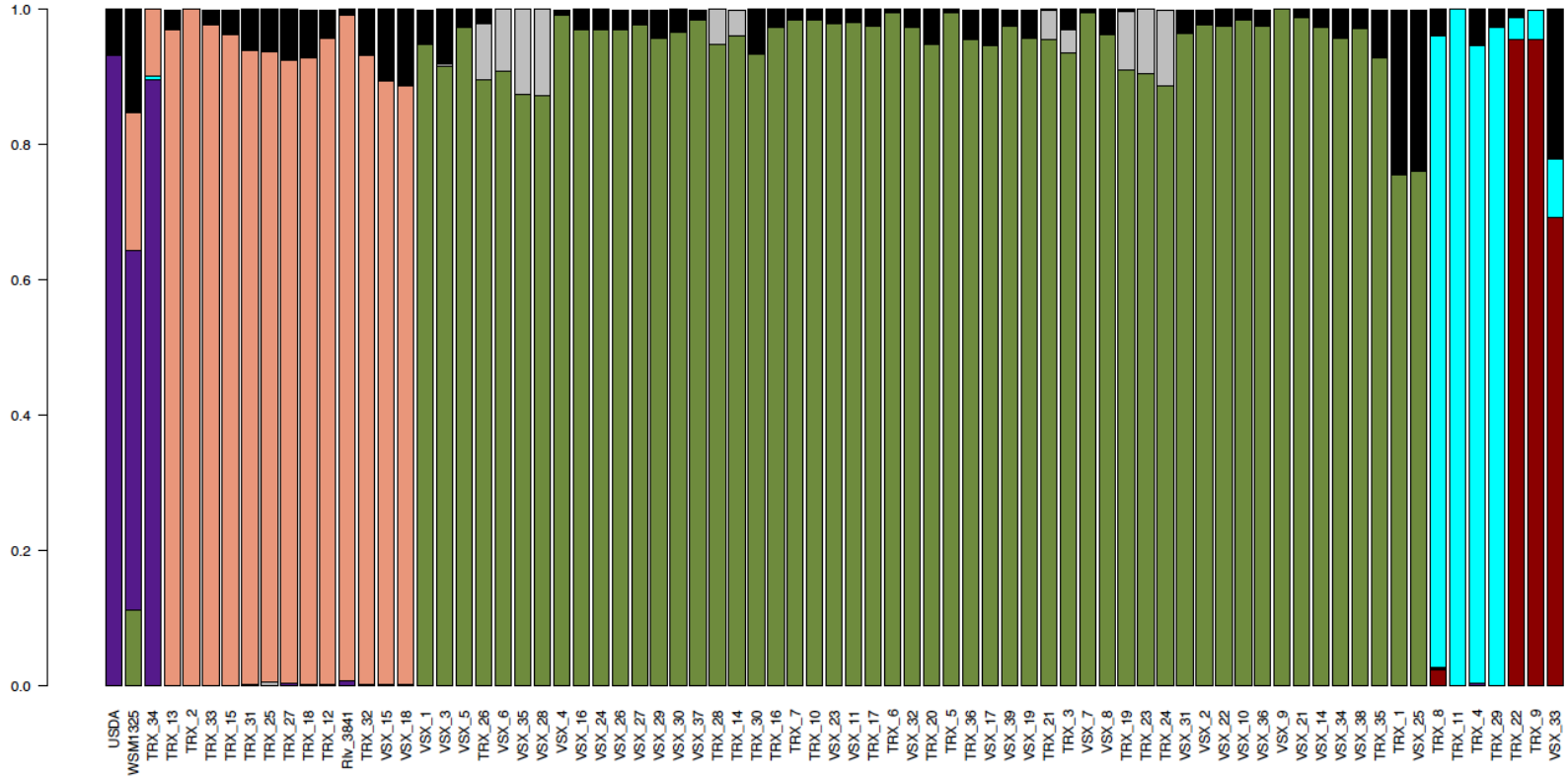


Figure 3.12 | Bar graph showing results of Structure analysis with K = 7: Each vertical bar represents one of the 75 R. leguminosarum strains. The coloring of each bar is proportional to the ancestry of each strain from each of the 7 populations (Y-axis: purple, dark salmon, dark green, cyan, dark red, black and grey respectively).

3.5 Discussion

We present a comprehensive analysis that shows recombination plays an important role in the evolution of the core genome of *R. leguminosarum*.

Previously, core genes are considered to be stable (resistant to recombination) part of bacterial genome but several comparative studies suggested that some bacteria have a stable core genome (Acosta et al., 2011; Joseph et al., 2012; Shi and Falkowski, 2008) and others do not (Cadillo-Quiroz et al., 2012; Didelot et al., 2011). In this study, we observed that majority (89%) of core genes are affected by recombination, suggesting that core genome of *R. leguminosarum* is shaped by recombination. Moreover, these core genes were largely located on the chromosome, reflecting chromosomal recombination. On the other hand, sister species (*R. etli*) of *R. leguminosarum* has a genome of low recombination (Acosta et al., 2011). However, some strains (CIAT 652, BRAZIL 5, 8C-3) used in Acosta et al.'s study (2011) are actually *R. phaseoli* (Lopez-Guerrero et al., 2012).

The conserved nature of genospecies was observed in the phylogenetic results of four topologies (Figure 3.4-3.7) that were congruent with many putative recombinant genes. For example, the ML tree of RL0377 (Figure 3.5) produced same five genospecies (A-E) observed in 100-gene tree (Figure 3.1). Moreover, some sub-clusters of genospecies C are well supported for example sub-cluster comprised of 8 symbiovar *viciae* strains (VSX_24, VSX_26, VSX_30, VSX_37, VSX_4, VSX_16, VSX_27, VSX_29) in Figure 3.6. The possible example of inter-genospecies was observed in the ML tree of RL2381 (Figure 3.6) where genospecies A members (TRX_34 and *Rlt* WSM1325) were located in a cluster that includes members of genospecies B indicating the presence of other genospecies B strains that are not sampled here. These results suggest the preference of intra- genospecies transfer in most of the recombinant genes.

ML analysis, ClonalFrame and Structure strongly suggest the existence of 5 genospecies in our dataset. Bacterial species may be regarded as biological species when recombination becomes the cohesive force in members of the same species and its strength decreases with an increase in genetic divergence (Fraser et al., 2007). Our data

strongly correlates with this hypothesis suggesting these genospecies may be regarded as biological species. Results from Clanistic analysis, ClonalFrame and Structure clearly show that more gene transfer or recombination has occurred within genospecies than between them. For example, ClonalFrame suggested that genospecies B (r/m: 102.93) is more affected by recombination than the whole population (r/m: 5.92) or genospecies C (r/m: 4.29). These results suggest that these genospecies are biological species.

Clanistic analysis is a fast way to compare multiple gene trees and 100-gene tree. Schliep et al. (2011) introduced this method by comparing 6901 unrooted gene trees. However, this method does not provide any statistical tests, so we decided to choose two different statistical methods: ClonalFrame and Structure. Although these tools are time consuming and computationally intensive, they have been used for identifying bacterial species (Cadillo-Quiroz et al., 2012; Didelot et al., 2011).

In conclusion, this study employs phylogenetic and population genetics approach to understand the role of recombination in the genome evolution of *R. leguminosarum* and its five genospecies. Most of the core genes (63 of 100) display phylogenetic incongruence, while 72 of 100 genes are under the influence of intra-genic recombination. These findings confirm that core genome of *R. leguminosarum* is shaped by the force of recombination. Putative recombinant genes exhibit similar five genospecies structure and preferred intra- genospecies transfer. Results of ClonalFrame and Structure analyses support the idea that the rate of homologous recombination declines with sequence divergence (Fraser et al., 2007; Roberts and Cohan, 1993; Zawadzki et al., 1995), so that successful recombination is less likely between than within species. This facilitates understanding of processes that might lead to more recombination within species rather than between species. The most plausible reason for this observation could be the more frequent occurrence of conjugation among members of the same genospecies. This conclusion is strongly supported by the phylogenetic analyses based on each type of *repABC* replicon (Kim, 2012). Kim (2012) noted lack of movement for pRL12- and pRL11-type plasmids between genospecies as they have similar core gene phylogenetic structure. Moreover, the phylogenetic tree of other plasmid-types displayed the preference of intra- genospecies transfer, for example, movement is less likely between than within genospecies for pRL10-type plasmids.

Chapter 4. Dominating influence of five genospecies on the composition and phylogeny of the accessory genome of *R. leguminosarum*

4.1 Abstract

The results of the previous chapters have demonstrated the presence of five cryptic genospecies (A-E) in *R. leguminosarum* based on the core genes. These species exhibit a high level of intra-species recombination that serves as a barrier to gene exchange between these species. This chapter focuses on the genetic diversity present in the accessory (variable) genome of *R. leguminosarum*. The genome of *Rlv* 3841 (genospecies B) was used as reference to explore the diversity of accessory genes in other members of genospecies including genospecies B. The remarkable regions of genetic similarities were localized in the regions of chromosome, chromids and large plasmids. One of the smallest reference plasmids (pRL8) carries a set of five host-specific genes known as Bvs (biovar *viciae* specific) genes, which is absent in strains of the other symbiovar (*trifolii*). The phylogenetic networks based on chromosome, chromids and large plasmids exhibit the same five genospecies observed in the core genes phylogeny. The nodulation (host specific) genes are introduced through horizontal gene transfer in these five genospecies. Finally, the specific accessory genes of the local population that differentiate it from reference genome indicated the adaptive nature of this population.

4.2 Introduction

The accessory genome is the variable component of the bacterial genome that may or may not be present in a bacterial strain. Accessory genes may be located on genomic islands, bacteriophages, transposons, insertion sequences and integrons (Dobrindt et al., 2004), but are most commonly carried on variable sized plasmids (Frost et al., 2005) that carry large amount of DNA. These genes provide genomic flexibility that helps in adaptation, antibiotic resistance, toxin production etc. Both horizontal gene transfer (HGT) and homologous recombination are responsible for shaping the accessory genomes (Vos, 2009). The accessory genome of *Legionella pneumophila* (Gomez-Valero et al., 2011) was maintained by extensive homologous recombination as well as HGT.

Studies based on the accessory genomes shed light on ecological adaptations in bacterial strains. Tenaillon et al. (2010) reviewed the population structure of commensal *Escherichia coli* strains to understand the ecological adaptations that convert useful bacteria into harmful pathogens. Comparative genomics of multiple strains of the *Azospirillum* (Wisniewski-Dye et al., 2012) identified the role of accessory genes in niche adaptations. Accessory genomic study of plant pathogen *Erwinia amylovora* (Mann et al., 2013) revealed host specific genes and metabolic pathways that are present in only one of the *E. amylovora* pathovars.

The diversity in the accessory genomes of rhizobium strains is well studied. Bailly et al. (2011) compared phylogenetic networks of plasmids with chromosomal phylogeny (Figure 4.1). The comparison suggested that these phylogenies do not reflect each other. Moreover, some specific accessory genes such as rhizobitoxine synthesis genes were revealed in 12 strains of *S. medicae* population that are absent in *S. medicae* WSM 419 (reference genome) suggesting that these genes might play a role in ecological adaptation.

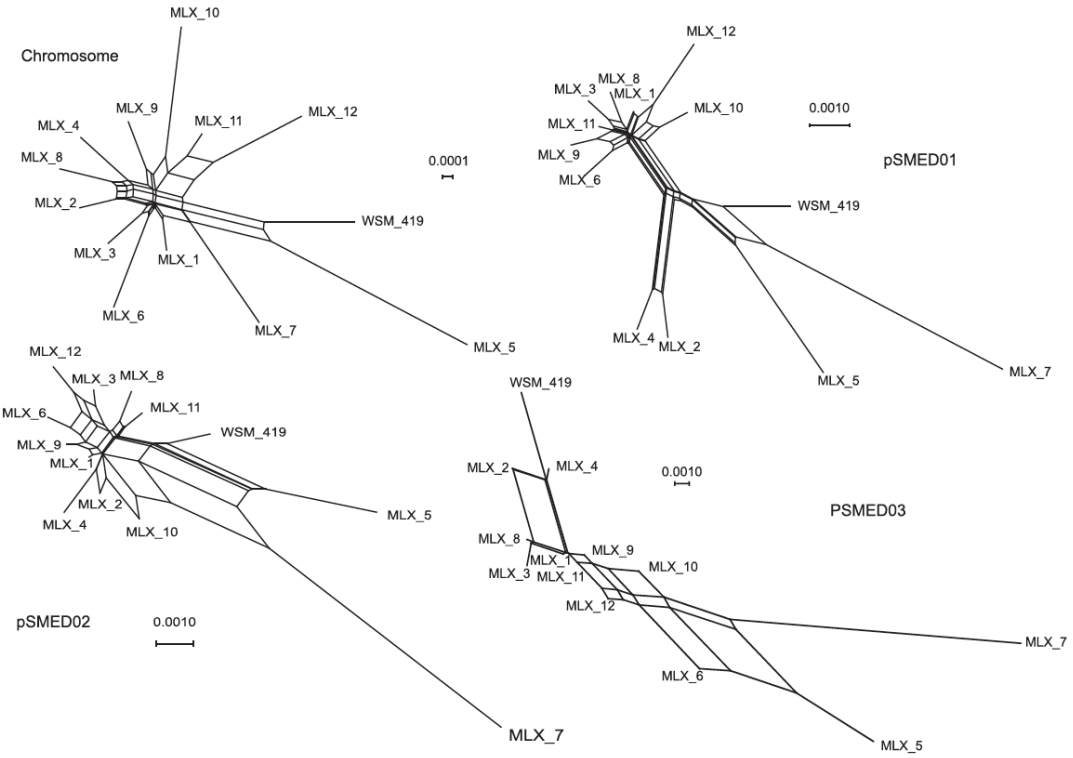


Figure 4.1 | Neighbour-Nets showing divergence among *S. medicae* strains for each replicon (Taken from Bailly et al., 2011)

Sugawara et al. (2013) analyzed genomic diversity in the accessory genome of the genus *Sinorhizobium* by comparing multiple strains of five *Sinorhizobium* genospecies. The results indicated a correlation between symbiotic efficiency and the presence of type IV secretion systems. Moreover, each clone has its own strategy to interact with the host plants and environments.

Almost every rhizobium accessory genome harbors nodulation (*nod*) genes that are essential for nodulating leguminous plants (Rogel et al., 2011). The *nod* genes are host specific and are used to classify symbiovars of *Rhizobium* species. These accessory genes are widely used to observe diversity in the multiple strains of *Rhizobium* species. Alvarez-Martinez et al. (2009) observed the common origin of housekeeping genes and nodulation genes in the population of *R. leguminosarum* isolated from different locations. Interestingly, Chang et al. (2011b) analyzed multiple strains of rhizobia

isolated from different parts of Southern China and found that *nod* genes are transferred vertically in some strains, but some strains acquired these genes through HGT.

The plasmid replication system genes (*repABC*) are another set of interesting genes that are located in the accessory genome of *Rhizobium*. RepA and B proteins are essential for plasmid partitioning, while RepC protein helps in replication. Kim (2012) investigated the distribution of *repABC* genes in 72 *R. leguminosarum* strains (Figure 4.2) and their phylogenetic diversity. The replication system of two chromids (pRL12 and pRL11) of *Rlv* 3841 was present in each of the Wentworth strains.

In chapter 1, we observed five genospecies (A-E) present in *R. leguminosarum* species. A genospecies can include strains of two symbiovars, which represent a major and long-term divergence in ecological adaptation (Jordan, 1984; Young, 1996). Chapter 2 demonstrated that recombination is a major evolutionary force in shaping the core genome evolution and these five genospecies are maintained by recombination. In this chapter, we focus on the dynamic nature of the accessory genome of *R. leguminosarum*. Initially, the distribution of genomic data of *Rlv* 3841 was observed in the local population of *R. leguminosarum* by constructing presence/absence matrices based on the genes of *Rlv* 3841 replicons (chromosome, two chromids and four plasmids). These matrices are a convenient way that can be used to observe the distribution of *Rlv* 3841, host and genospecies specific genes in this dataset. Next, we constructed and compared the phylogenetic networks based on each *Rlv* 3841 replicon. In order to observe the relationship between core and nodulation genes, a phylogenetic network based on the *nod* genes was compared with the core gene phylogeny and further investigated. Finally, we identified accessory genes that are present only in the local population of *R. leguminosarum*, but absent in the reference genome of *Rlv* 3841.

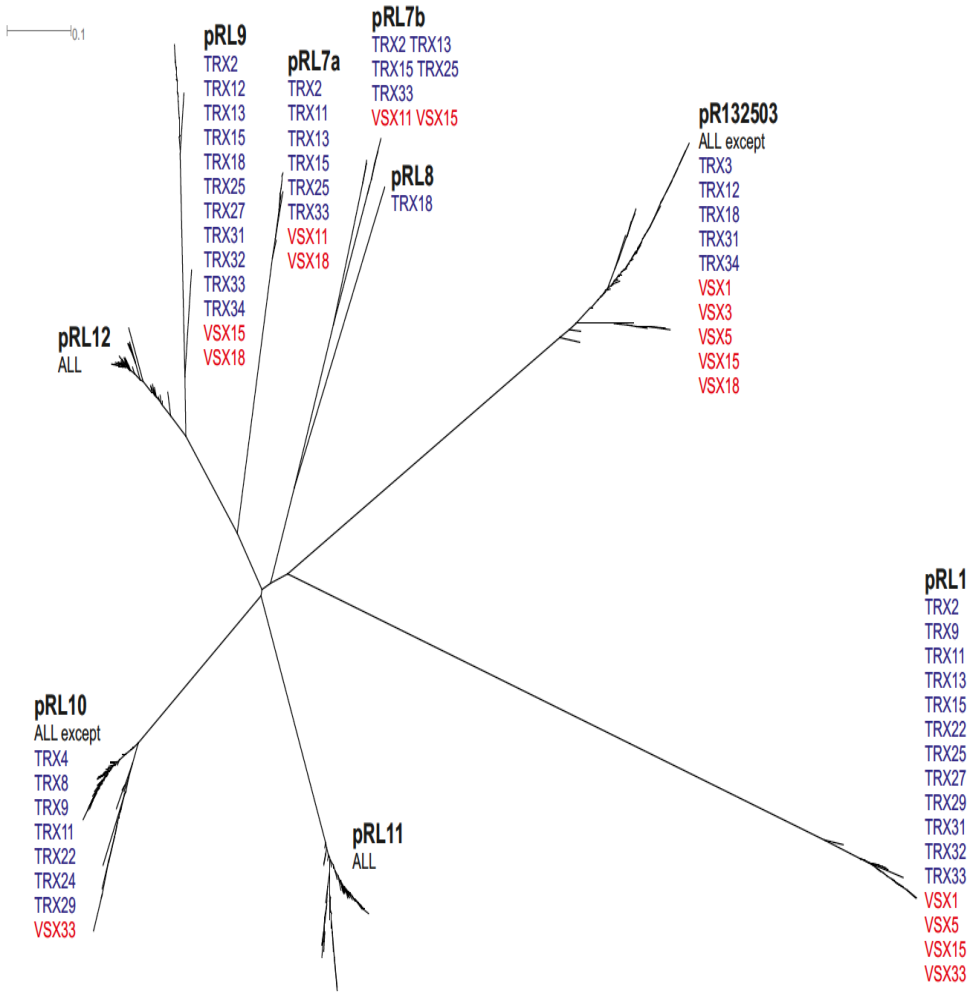


Figure 4.2 | The phylogenetic tree based on all RepABC replicons in 72 *R. leguminosarum* strains (Taken from Kim, 2012). Each terminal cluster represents a distinct plasmid compatibility type, and the isolates that carry that type are indicated.

4.2.1 Objectives

The main objectives of this chapter are:

1. Distribution of *Rlv* 3841 genes in the local population of *R. leguminosarum* strains.
2. Phylogenetic analyses of the *R. leguminosarum* strains based on *Rlv* 3841 replicons (chromosome, chromids and plasmids).
3. Phylogenetic analyses of the local population based on nodulation genes.
4. Identification of specific genes that are presents only in the local population and absent in *Rlv* 3841.

4.3 Materials and Methods

The strains studied in this chapter are: 72 *R. leguminosarum* strains, USDA 2370^{T}, *Rlt* WSM1325 (Reeve et al., 2010a) and a reference genome of *Rlv* 3841 (Young et al., 2006). The file (.fnn) of gene information for each *Rlv* 3841 replicons was downloaded from the National Center for Biotechnology Information (NCBI). These files were concatenated into one file and named the super gene file.

4.3.1 Construction of presence/absence matrix based on *Rlv* 3841 replicons

The Newbler 2.5 software with 90% sequence identity and 40-bp minimum overlap as parameter was used from the command line to perform individual reference-based assembly of *R. leguminosarum* strains against the super gene file (file consisting of genetic information of all *Rlv* 3841 replicons). Linux "grep" command was used to extract information based on *Rlv* 3841 genes from the file 454RefStatus.txt (one of the outputs of Newbler that provides the statistical information on the number of mapping reads to each reference sequence). The extracted data was converted into binary format (1-present & 0-absent) based on at least one unique mapped read by Perl. In order to construct a presence/absence matrix of each *Rlv* 3841 replicon, the binary formatted files were combined on the basis of seven *Rlv* 3841 replicons. These seven presence/absence matrices were displayed as heat maps using R.

4.3.2 Construction of phylogenetic networks based on reference replicons

For this analysis, the genome *Rlt* WSM1325 (Reeve et al., 2010a) and draft genome of USDA 2370^{T} was also included. In order to construct phylogenies, we used the output file: 454AllContigs.fna (contigs larger than 100 bp) to extract the nucleotide information based on *Rlv* 3841 genes from the assembly results using extractSequence.pl (http://seqanswers.com/forums/showthread.php?t=9498), and the

extracted information was merged with their respective genes present in *Rlt* WSM1325. The gene files were merged according to their location in the seven replicons resulting in seven multi FASTA files. MAFFT (Katoh and Standley, 2013) with auto option was implemented to align each of the seven multi FASTA files. The alignment results were visualized as phylogenetic networks using SplitsTree version 4.11 (Huson and Bryant, 2006).

4.3.3 Construction of phylogenetic networks based on nod genes

The genetic information of 13 nodulation genes (Young et al., 2006) of *Rlv* 3841 (Table 4.1) in *viciae* strains was extracted from the above results of reference-based assembly. Similarly, this information was extracted for *trifolii* strains by performing reference-based assembly using *nod* genes (Table 4.2) of pR132501 (*Rlt* WSM1325) as reference strains (Reeve et al., 2010a). The multiple alignment of each individual gene was performed by MUSCLE (Edgar, 2004). Each alignment was checked manually for sequence information. The sequence data for *nodT* and *nodO* genes was available for fewer of the isolates than other *nod* genes, thus they were not used in this study. The remaining *nod* gene alignments were concatenated into a super-alignment. The super-alignment results were visualized as phylogenetic networks using SplitsTree version 4.11 (Huson and Bryant, 2006).

Table 4.1 | The 13 *nod* genes and their position in pRL10 of *Rlv* 3841. Genes with an asterisk were used in the phylogenetic analysis.

Locus tag	Gene symbol	Position	Strand
pRL100175	*nodO*	174948..175802	+
pRL100178	*nodT*	177479..178927	-
pRL100179*	*nodN*	179095..179580	-
pRL100180*	*nodM*	179658..181484	-
pRL100181*	*nodL*	182056..182628	-
pRL100182*	*nodE*	183222..184433	-
pRL100183*	*nodF*	184434..184712	-
pRL100184*	*nodD*	185357..186289	-
pRL100185*	*nodA*	186548..187138	+
pRL100186*	*nodB*	187135..187785	+
pRL100187*	*nodC*	187808..189085	+
pRL100188*	*nodI*	189123..190157	+
pRL100189*	*nodJ*	190161..190940	+

Table 4.2 | The 11 *nod* genes and their position in pR132501 of *Rlt* WSM1325.

Locus tag	Gene symbol	Position	Strand
Rleg_4909	*nodN*	288676..289167	-
Rleg_4910	*nodM*	289247..291073	-
Rleg_4911	*nodL*	291504..292055	-
Rleg_4913	*nodE*	292544..293752	-
Rleg_4914	*nodF*	293752..294030	-
Rleg_4915	*nodD*	294503..295450	-
Rleg_4916	*nodA*	295685..296275	+
Rleg_4917	*nodB*	296272..296910	+
Rleg_4918	*nodC*	296931..298211	+
Rleg_4919	*nodI*	298242..299270	+
Rleg_4920	*nodJ*	299267..300055	+

4.3.4 Population specific genes

The methodology employed to identify population specific genes in Bailly et al. (2011) was used here to find accessory genes that were confined to the seventy-two *R. leguminosarum* strains. The Newbler 2.5 software with 90% sequence identity and 40-bp minimum overlap as parameter was used from the command line to perform a reference-based assembly of all the seventy-two strains taken together against the *Rlv* 3841 genome. From the output file 454ReadStatus.txt, unmapped reads were extracted and assembled *de novo* as described in 2.3.2. Contigs larger than 500 bp were annotated using the RAST server (Aziz et al., 2008) that resulted into 13,252 genes. The presence/absence matrix of these genes was constructed using the same strategy as described above.

4.4 Results

4.4.1 Presence/Absence analysis:

The presence/absence matrices (Figure 4.3-4.6) based on the genes of *Rlv* 3841 replicons (chromosome, 2 chromids and 4 plasmids) displayed the distribution of the accessory genes in the local population of *R. leguminosarum*. In each matrix, genes are represented as columns, while rows represent strains of the population and organized according to their related genospecies (top to bottom: A to E). Genes that are present in individual strains are colored blue, whereas white indicates absent genes. Genomic islands of chromids and plasmids that are described in this study are highlighted in green.

The chromosomal matrix (Figure 4.3) revealed the ubiquitous distribution of chromosomal genes in this population, but some genomic islands were observed in this population or in some individuals. For example, the known genomic island (RL0790-RL0841: G1 in Figure 4.3) in *Rlv* 3841 (Young et al., 2006) was absent except TRX_9 (a member of genospecies E). Other missing genomic island was comprised of around 40 genes (RL2155- RL2195: G2 in Figure 4.3) with unknown functions.

The Chromid matrices (Figure 4.4) displayed the conserved nature of genes of pRL12 (Figure 4.4A) and pRL11 (Figure 4.4B) in this population. The pRL12 matrix (Figure 4.4A) displayed a genomic island, which was mostly carried by the isolates of genospecies B (VSX_18, VSX_15, TRX_33, TRX_32, TRX_31, TRX_27, TRX_25, TRX_18, TRX_15, TRX_13, TRX_12 and TRX_2) and absent in the members of other genospecies. This region harbors many ABC transporter genes (Appendix table III.I). The enrichment of ABC transporter genes in this genomic island was confirmed by using the Pearson's Chi-square test, which was performed by using 269 ABC transporter genes (total) of *Rlv* 3841 genome (13/269 genes, P value < 0.05).

Interestingly, the genetic information of two large plasmids (Figure 4.5) was widely distributed in this population (especially in genospecies B). We observed that the

pRL10 matrix (Figure 4.5A) could be divided into two halves. The first half (around the first 200 genes) reflected the property of accessory genes (sporadically distributed) and the other half showed the property of core genes (highly conserved). The nodulation genes (shown in gold) are located in the accessory part or first half. These genes are highly diverged in *trifolii* strains, so are mostly shown as "absent" in this symbiovar.

The genes of the other large plasmid (Figure 4.5B) were abundant in genospecies B. We observed four genomic islands that are specific to the genospecies B. These islands include many ABC transporter genes (Appendix table III.II-V). Collectively, the enrichment of ABC transporter genes in these genomic islands was confirmed by using the Pearson's Chi-square test, which was performed by using total number of ABC transporter genes in *Rlv* 3841 genome (12/269 genes, P value < 0.05).

The genetic information of the small plasmids (pRL8 and pRL7) was almost absent in this population (Figure 4.6). Interestingly, we found a set of five putative pRL8 genes (Table 4.3) that were present only in symbiovar *viciae* strains (except VSX_18) and absent in all *trifolii* strains (shown in brown in Figure 4.6A), thus these genes are called Bvs (biovar *viciae* specific) genes.

Table 4.3 | Bvs (Symbiovar *viciae* specific) genes and their location in pRL8.

Locus tag	Position	Annotated Function	Strand
pRL80073	76316..77521	putative cysteine desulfurase	-
pRL80074	77684..78604	LysR family transcriptional regulator	-
pRL80075	78784..79167	putative endoribonuclease L-PSP family protein	+
pRL80076	79231..80253	putative aliphatic nitrilase	+
pRL80077	80256..81245	putative molybdenum-binding oxidoreductase	+

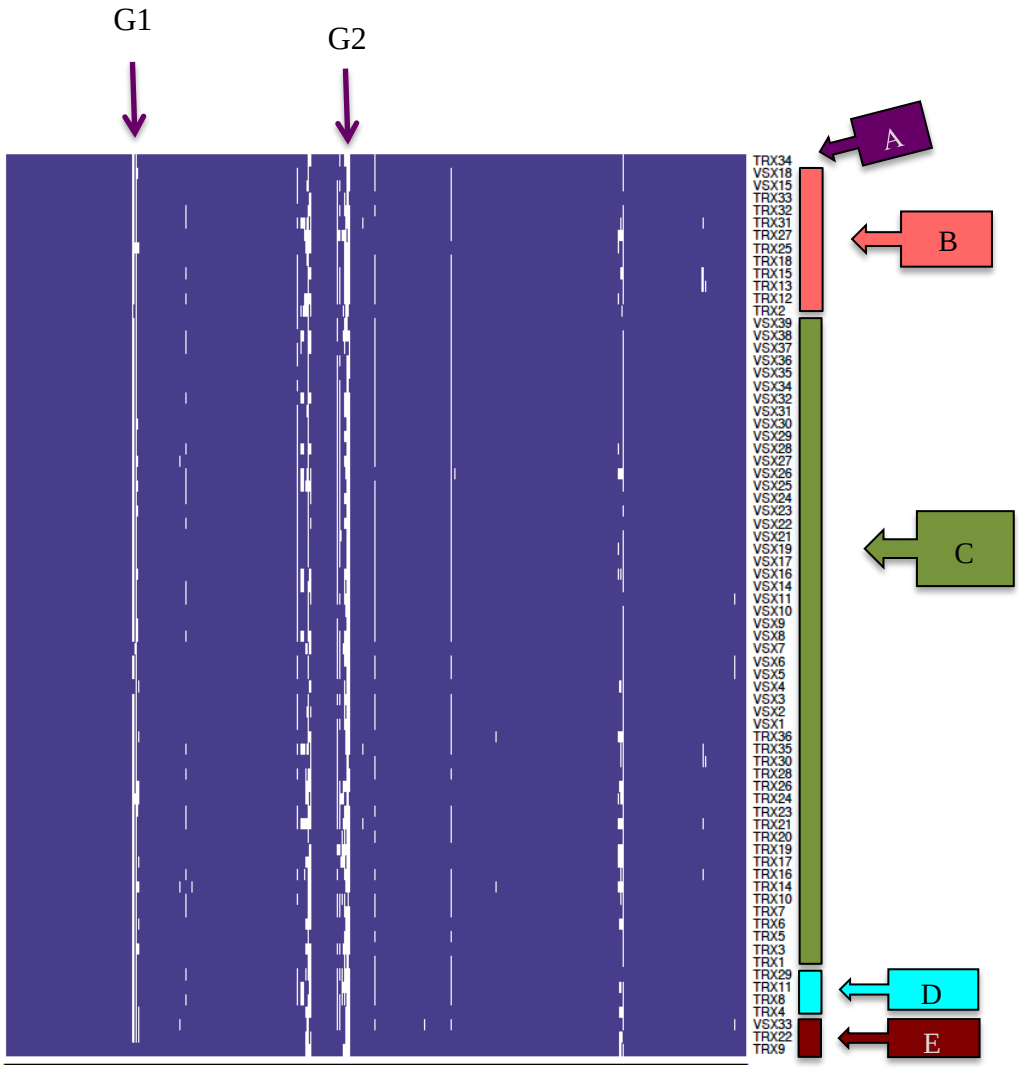
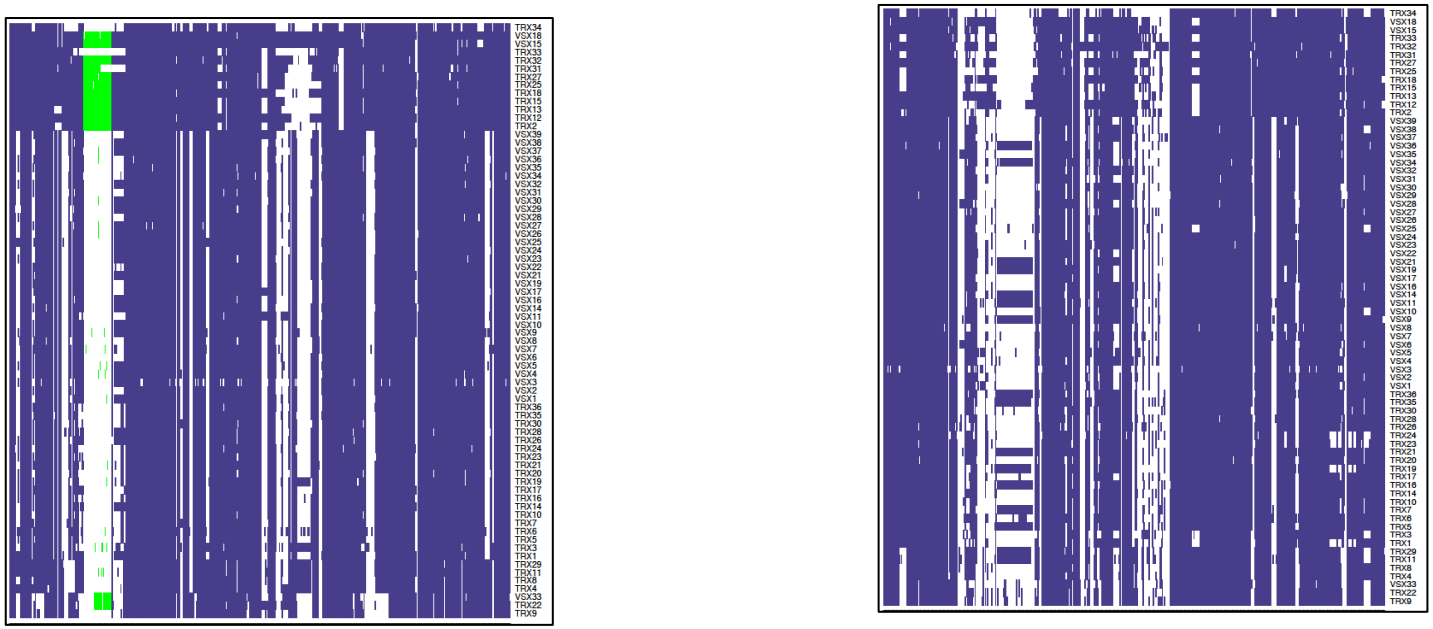
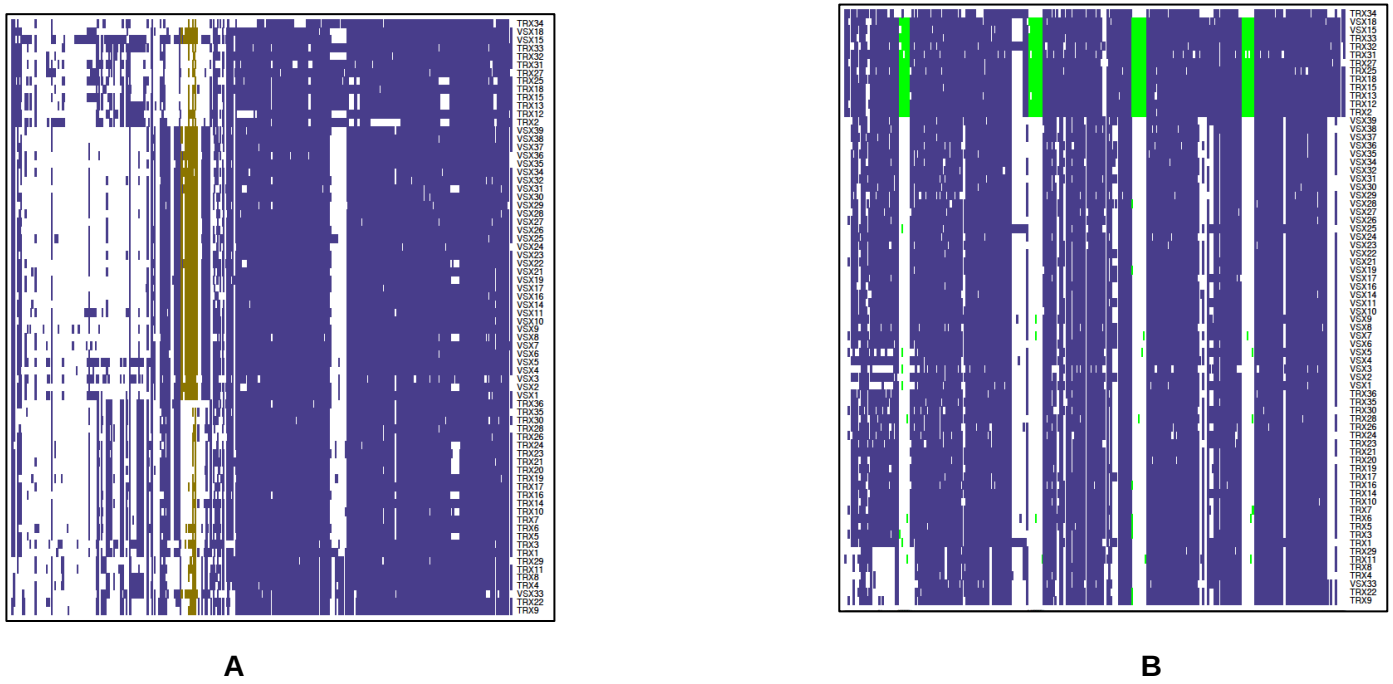


Figure 4.3 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 chromosomal genes. The presence of genes is shown in blue, and absent genes are in white. Rows represent 72 strains, and columns represent chromosome genes that are longer than 100 bp. Strains are arranged according to their respective genospecies (A-E). G1 and G2 represent two missing genomic islands of this population.



A B

Figure 4.4 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 chromid genes (A: pRL12, B: pRL11). In both matrices, the presence of genes is shown in blue, absent genes are in white. Rows represent 72 strains, and columns represent plasmid genes that are longer than 100 bp. Strains are arranged according to their respective genospecies (A-E). Studied genospecies B specific island in pRL12 (Appendix table III.I) is highlighted in green.



A B

Figure 4.5 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 large plasmid genes (A: pRL10, B: pRL9). In both matrices, the presence of genes is shown in blue, absent genes are in white. Rows represent 72 strains, and columns represent plasmid genes that are longer than 100 bp. Strains are arranged according to their respective genospecies (A-E). The nod genes in pRL10 are shown in gold. Studied genospecies B specific islands in pRL9 (Appendix table III.II-V) are highlighted in green.

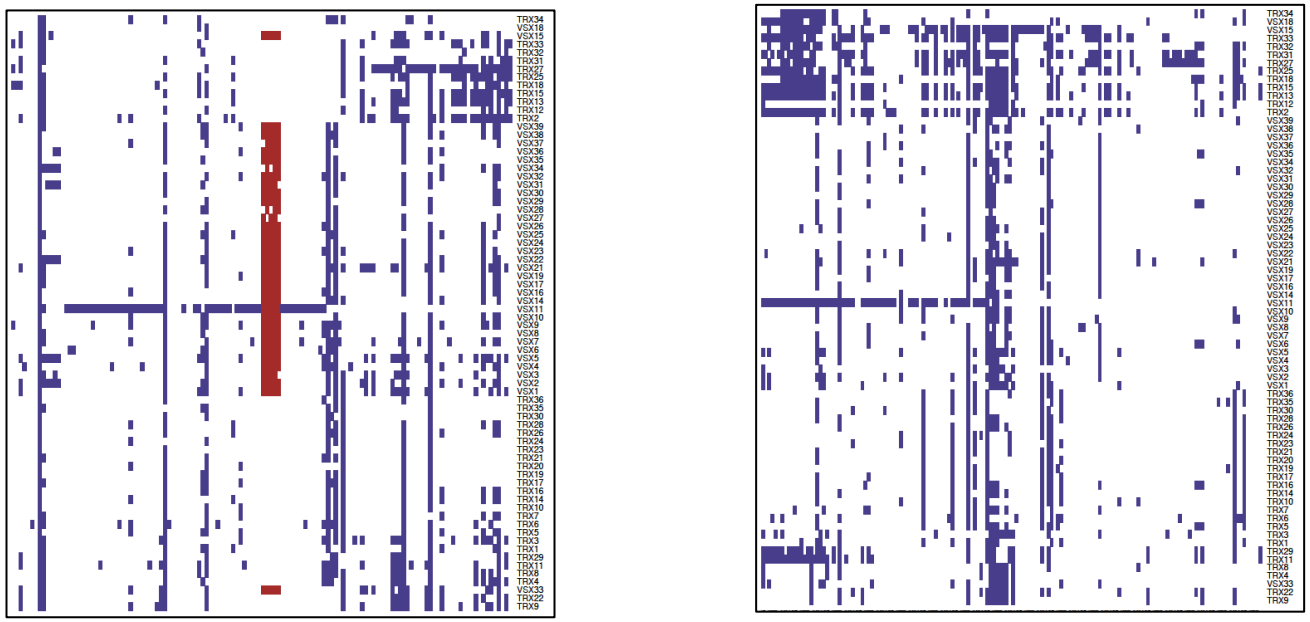


Figure 4.6 | Presence/Absence matrix obtained for 72 *R. leguminosarum* strains using *Rlv* 3841 small plasmid genes (A: pRL8, B: pRL7). In both matrices, the presence of genes is shown in blue, absent genes are in white. Rows represent 72 strains, and columns represent plasmid genes that are longer than 100 bp. Strains are arranged according to their respective genospecies (A-E). The Bvs (symbiovar viciae specific) genes in pRL8 are in brown.

4.4.2 Phylogenetic analysis of reference replicons

In order to investigate the phylogenetic diversity in the local population based on the observed genes, phylogenetic networks based on each replicon were constructed (Figure 4.7 - 4.10). For this analysis, we also included the draft genome of USDA 2370^{T} and the *Rlt* WSM1325 genome to obtain a complete picture, but not for the analysis of pRL8 and pRL7.

The networks obtained from the chromosome (Figure 4.7), chromids (pRL12: Figure 4.8A and pRL11: Figure 4.8B) and large plasmids (pRL10: Figure 4.9A and pRL9: Figure 4.9B) displayed the same five genospecies (A-E) observed in the core genes phylogeny (Figure 3.1). Most of the strains are clustered in two big genospecies (B and C). Small genospecies are in the same position observed in the core genes phylogeny. The reference genome is still located in genospecies B. Both USDA 2370^{T} and *Rlt* WSM1325 are still closely related to the strain of genospecies A (TRX_34). Moreover, deep investigation of each network revealed the conserved nature of sub-clusters of genospecies C discovered in the 100-gene ML tree (Figure 3.1).

The chromosomal network (Figure 4.7) displayed almost all the sub-clusters of genospecies C. For example, a mixed group contained ten *viciae* (VSX_38, VSX_14, VSX_ 22, VSX_2, VSX_21, VSX_9, VSX_10, VSX_31, VSX_36, VSX_34) and one *trifolii* (TRX_35) strains, a *viciae* cluster comprised of three *viciae* isolates: VSX_5, VSX_1, and VSX_3 and a hybrid cluster contained one *trifolii* (TRX_26) with three *viciae* (VSX_28, VSX_26 and VSX_35) members.

Some of these clusters were also observed in the networks of chromids (Figure 4.8A and B) and large plasmids (Figure 4.9A and B). For instance, a *viciae* cluster (VSX_4, VSX_27, VSX_16, VSX_30, VSX_24, VSX_37, VSX_29 and VSX_26) was conserved in all the networks. The cluster comprised of TRX_1 and VSX_25 in the 100-gene ML tree (Figure 3.1) was located at the same position in pRL11 (Figure 4.8B), pRL10 (Figure 4.9 A) and pRL9 (Figure 4.9 B) networks.

The network obtained from the smallest plasmid pRL8 (Figure 4.10 A) was based on less genetic information and was poorly resolved, but it differentiated the population into five genospecies (A-E) except for two strains (reference plasmid and VSX_15) of genospecies B that are clustered with genospecies C. The possible reason of different location of reference plasmid and VSX_15 is the presence of Bvs genes in the symbiovar *viciae* population.

Finally, the phylogenetic network of pRL7 (Figure 4.10 B) was completely different from the core genes phylogeny (Figure 3.1). The network was not fully resolved and was based on limited information, but there are some clear groups that consisted of members of different genospecies such as one group that includes members of genospecies E (TRX_9, VSX_33 and TRX_22), C (TRX_3), B (TRX_12) and D (TRX_8 and TRX_4) were clustered into a single group and genes responsible for this group belong to one of the conjugative systems (Table 4.4) of pRL7. Intriguingly, this network (Figure 4.10 B; Appendix figure III.I) could be divided into two splits based on two symbiovars (*trifolii* and *viciae*) of *R. leguminosarum*. One half consists of all *trifolii* strains except TRX_14 and the other half includes all *viciae* strains except VSX_33 that indicate the presence of host specific genes of both symbiovars on this plasmids.

.

Table 4.4 | Genes responsible for one of the conjugative systems of pRL7.

Locus tag	Gene symbol	Position	Strand
pRL70084	*traD*	68791..69006	-
pRL70085	*traC*	69011..69304	-
pRL70086	*traA*	69560..72886	+
pRL70087	*traF*	72883..73446	+
pRL70088	*traB*	73436..74599	+
pRL70089	*traH*	74616..75236	+
pRL70091	*traI*	75784..76398	-

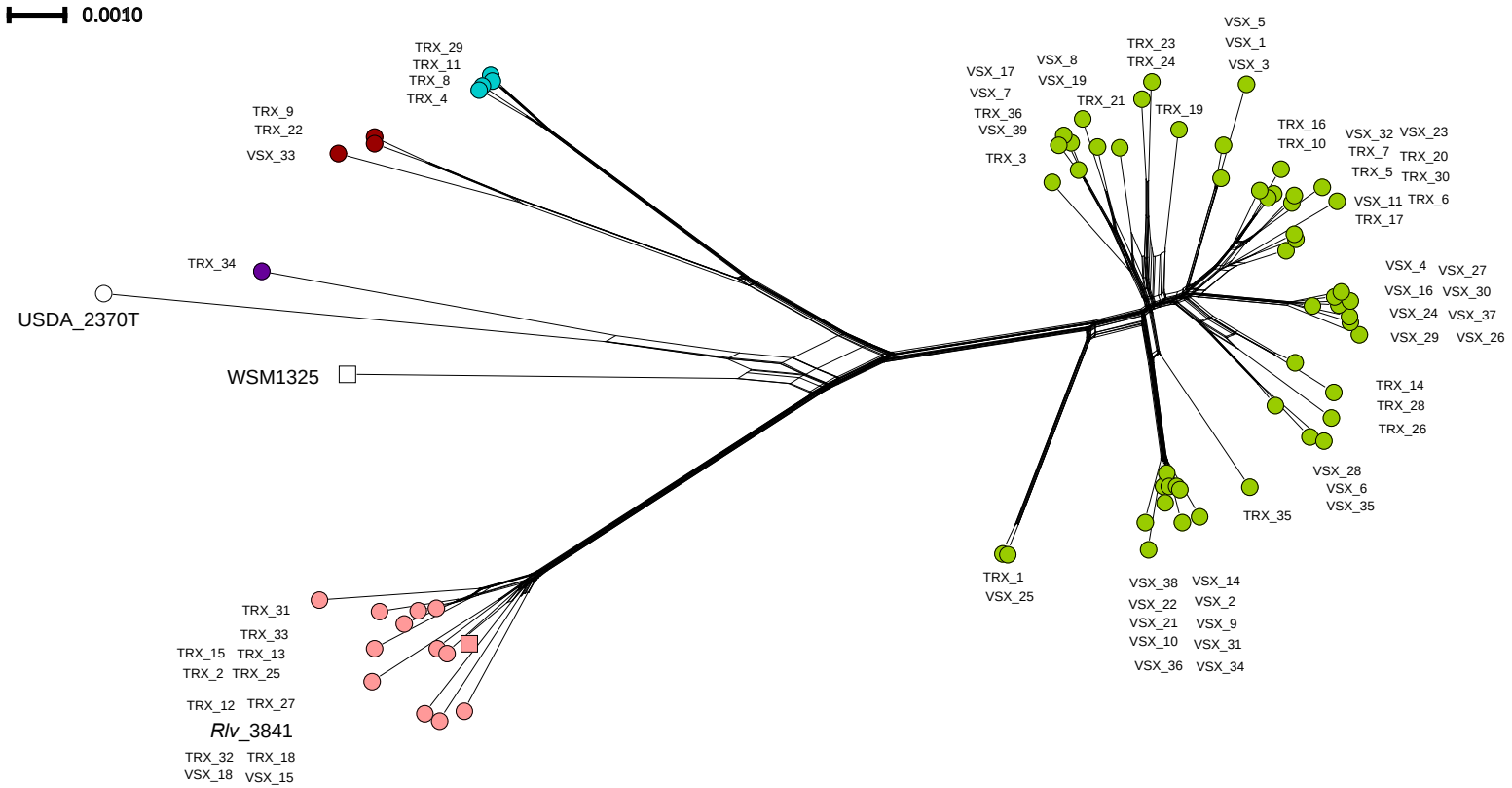


Figure 4.7 | Phylogenetic network obtained from the chromosomal genes of *Rlv* 3841 for 72 *R. leguminosarum* strains. Strain nodes are circled according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). *Rlv* 3841 is shown by the salmon square, *Rlt* WSM1325 in the white square and USDA_2370^{T} in the white circle.

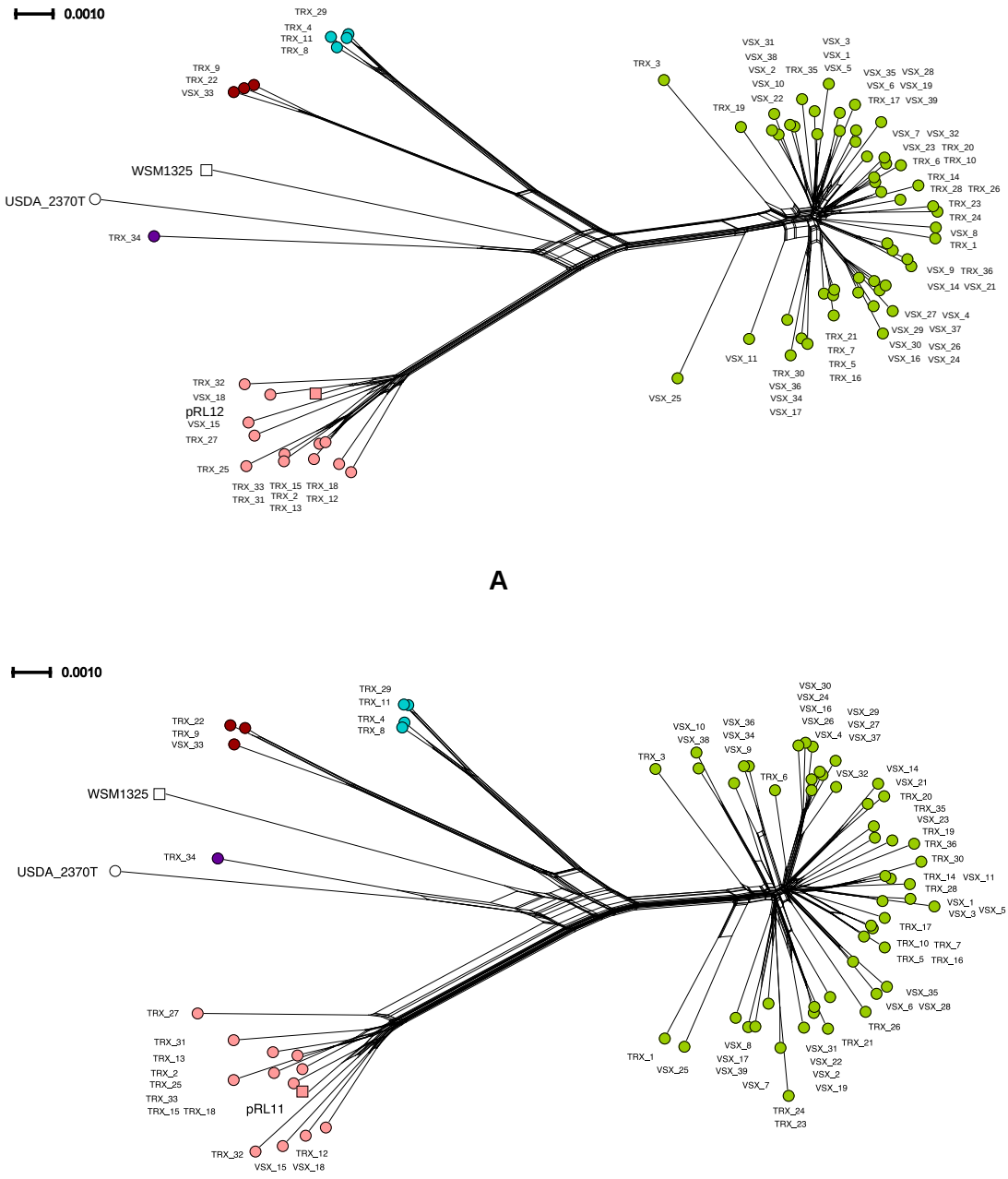
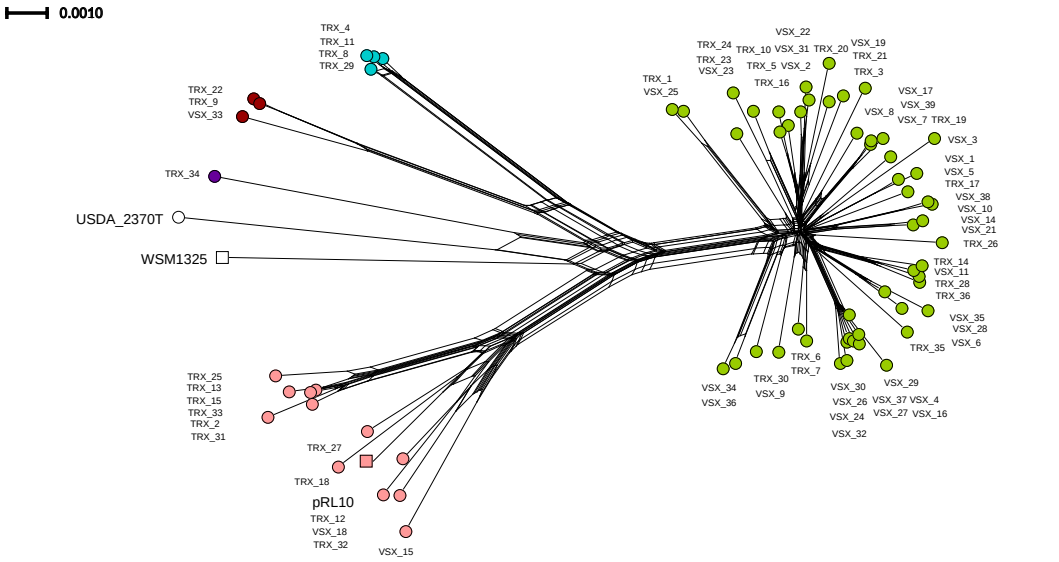
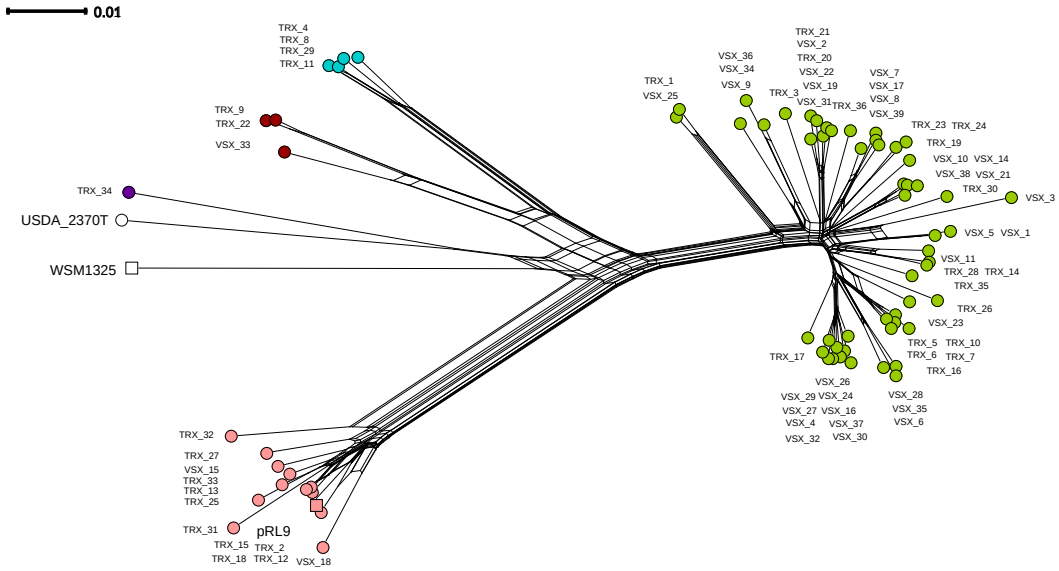


Figure 4.8 | Phylogenetic network obtained from the two chromid genes (A: pRL12, B: pRL11) of *Rlv* 3841 for 72 *R. leguminosarum* strains. Strain nodes are circled according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). *Rlv* 3841 is shown by the salmon square, *Rlt* WSM1325 in the white square and USDA_2370^{T} in the white circle.



A



B

Figure 4.9 | Phylogenetic network obtained from two large plasmids (A: pRL10, B: pRL9) of *Rlv* 3841 for 72 *R. leguminosarum* strains. Strain nodes are circled according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). *Rlv* 3841 is shown by the salmon square, *Rlt* WSM1325 in the white square and USDA_2370^{T} in the white circle.

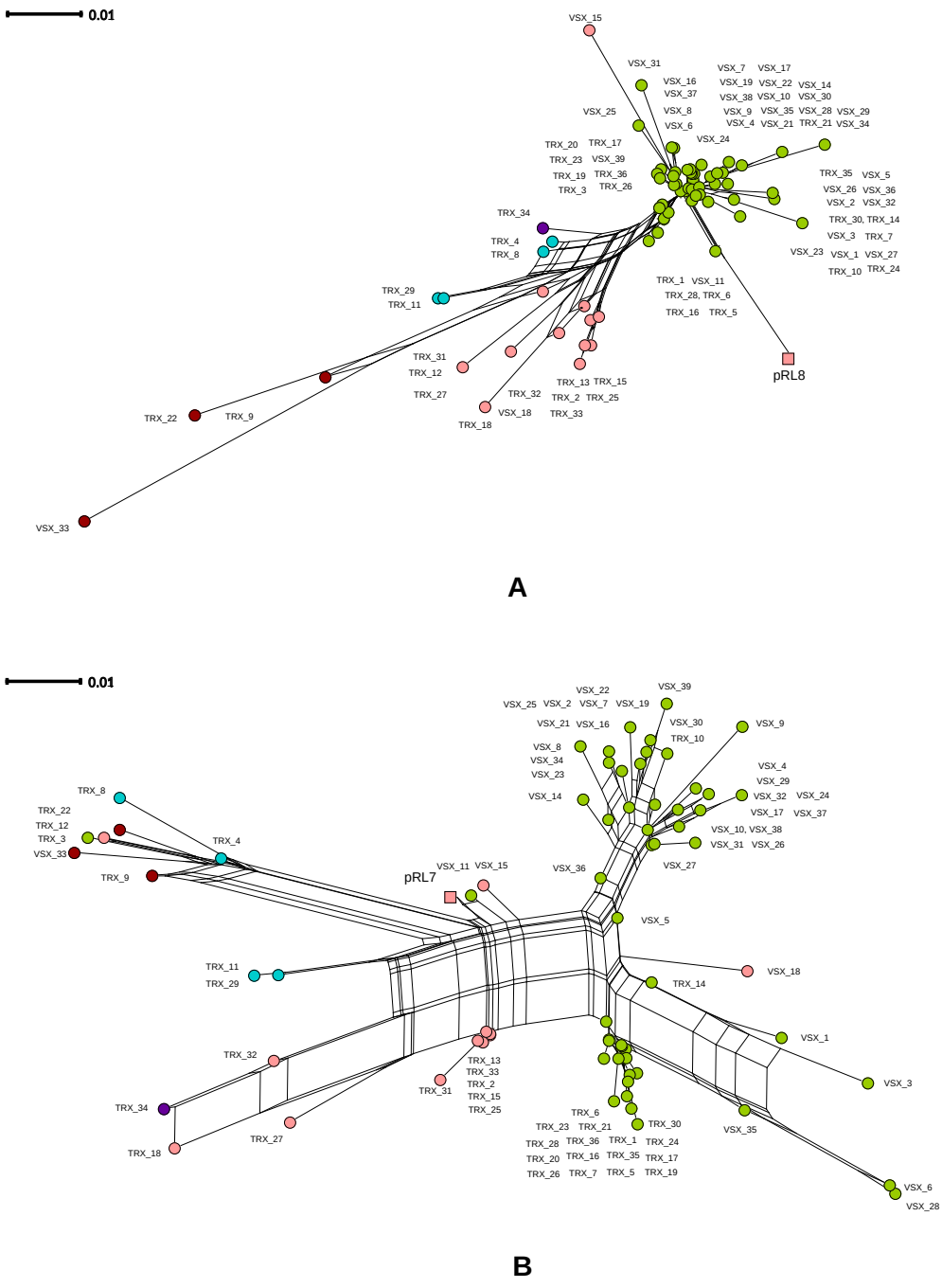


Figure 4.10 | Phylogenetic network obtained from two small plasmids (A: pRL8, B: pRL7) of *Rlv* 3841 for 72 *R. leguminosarum* strains. Strain nodes are circled according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). *Rlv* 3841 is shown by the salmon square.

4.4.3 Phylogenetic analysis of nodulation (*nod*) genes

In the presence/absence matrix of pRL10 (4.5A), the *nod* genes of *Rlv* 3841 were highly conserved in *viciae* strains, but not in *trifolii* strains. Therefore, a phylogenetic network (Figure 4.11) was constructed based on 11 *nod* genes of *Rlv* 3841 (Table 4.1) and *Rlt* WSM1325 (Table 4.2) in which strains are colored according to their genospecies (A: purple, B: salmon, C: dark green, D: cyan and E: dark red). The network (Figure 4.11) can be divided into two groups: *trifolii* group (consisted of all *trifolii* strains) on the left hand side with *Rlt* WSM1325 and *viciae* group (all *viciae* strains) on the right hand side with *Rlv* 3841 (pRL10). The separation between these two groups reflected the genetic diversity of nodulation genes between two symbiovars of *R. leguminosarum*. Moreover, these two groups are composed of several sub groups. In order to explore the mechanism of inheritance in *nod* genes, these two groups (*trifolii* and *viciae*) were compared individually with the core gene phylogeny (Figure 3.1). A detailed comparative analysis is described below:

The phylogenetic network obtained from *Rlv* 3841 *nod* genes differentiates the *viciae* isolates except VSX_11, VSX_15 and VSX_18 into four clear clusters (Figure 4.12). The multiple strains of genospecies C with or without the strains of different genospecies are differentiated into these four groups: 1. A cluster consisted of sub cluster of genospecies C (VSX_34, VSX_22, VSX_2, VSX_31 and VSX_36) and one strain (reference plasmid pRL10) of genospecies B. 2. A cluster contained two strains of a subgroup of genospecies C (VSX_3, VSX_5) and genospecies E strain (VSX_33). 3. A cluster consisted of 4 isolates (VSX_25, VSX_9, VSX_10 and VSX_38) of genospecies C. The hybrid strain, VSX_1 of genospecies C, was located in the middle of two groups (2 and 3). 4. The remaining genospecies C strains were located in the last cluster (poorly resolved)

The phylogenetic network obtained from the *nod* genes of *Rlt* WSM1325 was based on the poor genomic alignment data (two of eleven genes have sequence information for all the isolates) that resulted into many reticulation events, but differentiates the *trifolii* members except *Rlt* WSM1325 (WSM1325) and TRX_17 into five clear groups (Figure 4.13). The strains of genospecies B were clearly differentiated into two groups: 1. A cluster consisted of 4 members (TRX_27, TRX_32, TRX_18 and TRX_12) with the member of genospecies A (TRX_34). 2. A cluster consisted of 6 strains (TRX_31, TRX_25, TRX_13, TRX_2, TRX_33 and TRX_15). Similarly, most of the genospecies C strains were distributed into two major groups (not present in core genes phylogeny) except TRX_17 and TRX_3. The completely resolved group of genospecies C was comprised of four *trifolii* strains (TRX_1, TRX_28, TRX_14 and TRX_26). Finally, the last group consisted of strains from different genospecies: genospecies D (TRX_4, TRX_8, TRX_11 and TRX_29), a strain of genospecies C (TRX_3) and two strains of genospecies E (TRX_9 and TRX_22).

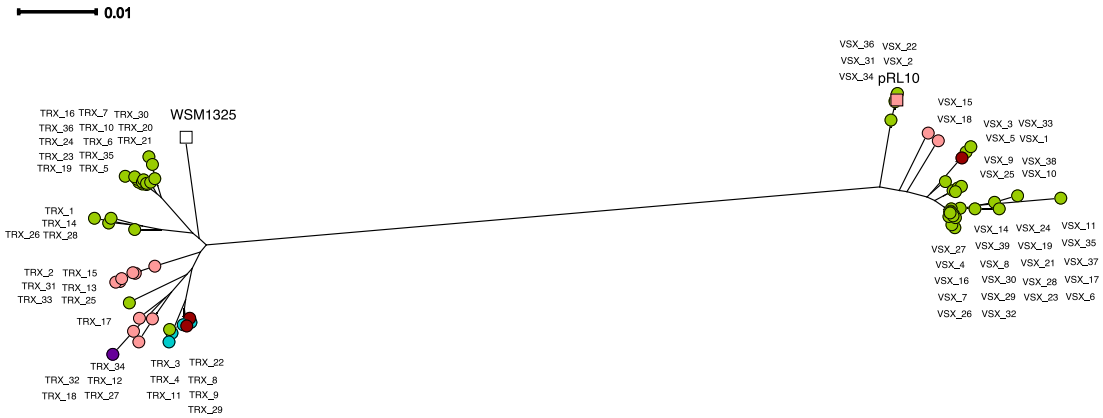


Figure 4.11 | Phylogenetic network obtained from 11 *nod* genes of *Rlv* 3841 and *Rlt* WSM1325 for 36 *R. leguminosarum viciae* and 36 *trifolii* strains respectively. Strain nodes are colored according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). The pRL10 in the salmon square represents *Rlv*. 3841. *Rlt* WSM1325 is shown by white rectangle.

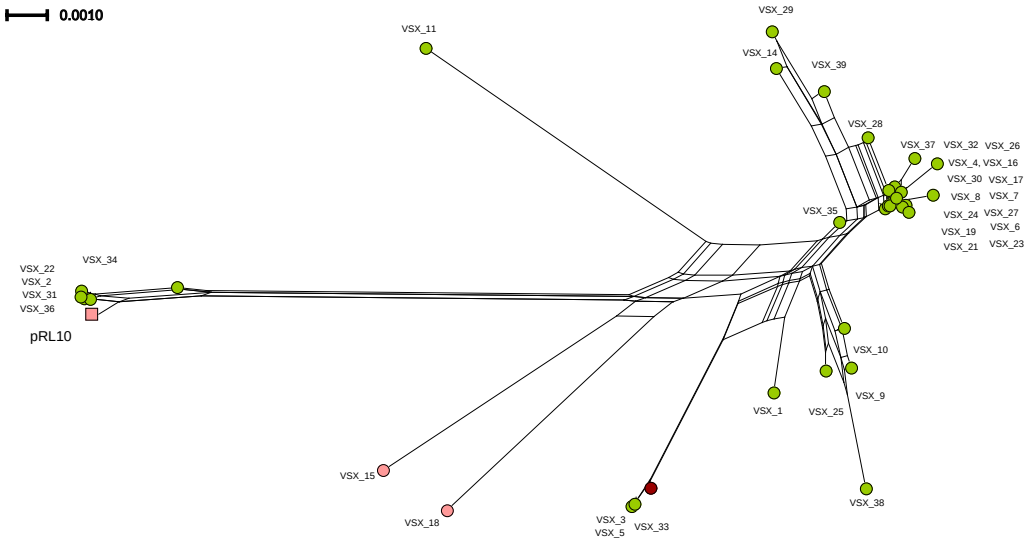


Figure 4.12 | Phylogenetic network obtained from 11 *nod* genes of *Rlv* 3841 for 36 *R. leguminosarum viciae* strains. Strain nodes are colored according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). The pRL10 in the salmon square represents *Rlv*. 3841.

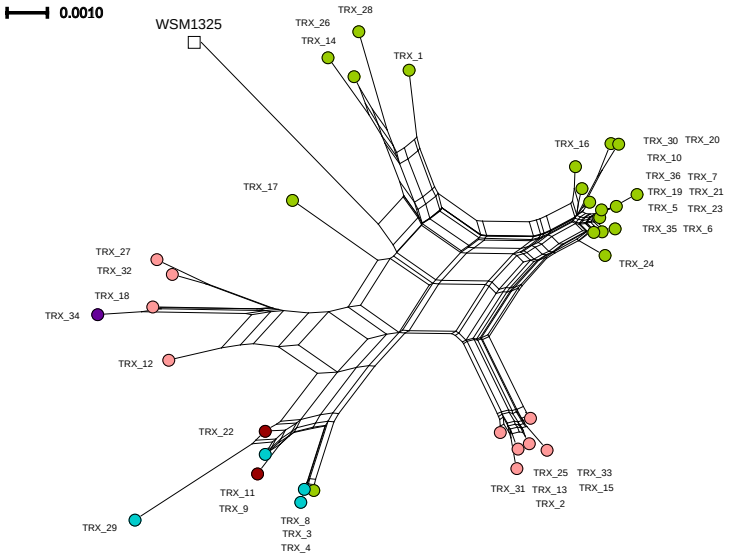


Figure 4.13 | Phylogenetic network obtained from 11 *nod* genes of *Rlt* WSM1325 for 36 *R. leguminosarum trifolii* strains. Strain nodes are colored according to their genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red). *Rlt* WSM1325 is shown by white rectangle.

4.4.4 Population specific genes

We obtained 8802 contigs with a total size of 11,250,877 bp from the *de novo* assembly of 454 reads that are confined to 72 *R. leguminosarum* strains from the Wentworth population. These contigs were automatically annotated by the RAST server (Aziz et al., 2008), which resulted in 13,252 CDS. They showed a close relationship of codon usage with the genes of *Rlt* WSM1325 (score: 521), CIAT 652 (score: 480) and *Rlv*_3841 (score: 478). Genes were functionally categorized under 181 subsystems. The maximum number of features in the subsystem was related to carbohydrate metabolism, amino acids and derivatives.

In order to explore the genetic diversity of 13,252 genes in 72 isolates, a presence/absence matrix (Figure 4.15) was generated in which rows represent strains and columns represent genes. The matrix was clustered on the basis of both rows (strains) and columns (genes). The row clustering resulted in the arrangement of strains according to their genospecies (A-E) except A (TRX_34), which is clustered between the strains of genospecies B. The genes clustering resulted in the grouping of genes according to their distribution. This clustering reflected the genes that are shared by the members of a specific genospecies, two genospecies, all genospecies and genes that are strain specific. For example, genes located on the extreme right (Figure 4.15) were conserved in all the members of genospecies C. We examined the genes that are genospecies specific. The specific genes of genospecies B, D and E are based on conserved genes shared by all the members of the related species, whereas specific genes of genospecies C are those that may or may not be present in a strain of this genospecies. As expected, the maximum number of specific genes is located in the members of genospecies C (Table 4.5) and the genospecies B harbors minimum number of these genes (Table 4.5). The specific genes of genospecies A (one strain) were not considered for this analysis.

The direct relationship of specific genes with the sequence coverage (Figure 4.14) suggested the presence of many other unique genes. The maximum numbers of unique genes were observed in TRX_6, which had highest sequence coverage (10.06 X), while strains of genospecies B (VSX_18: 1.08 X, TRX_27: 0.87 X, TRX_31: 0.68 X,

TRX_32: 0.86 X) and genospecies A (TRX_34: 0.73 X) had carried minimum number of specific genes (bottom left in Figure 4.14).

Table 4.5 | Specific genes present in each genospecies (B-E)

Genospecies	Number of specific genes
B	145
C	4670
D	355
E	367

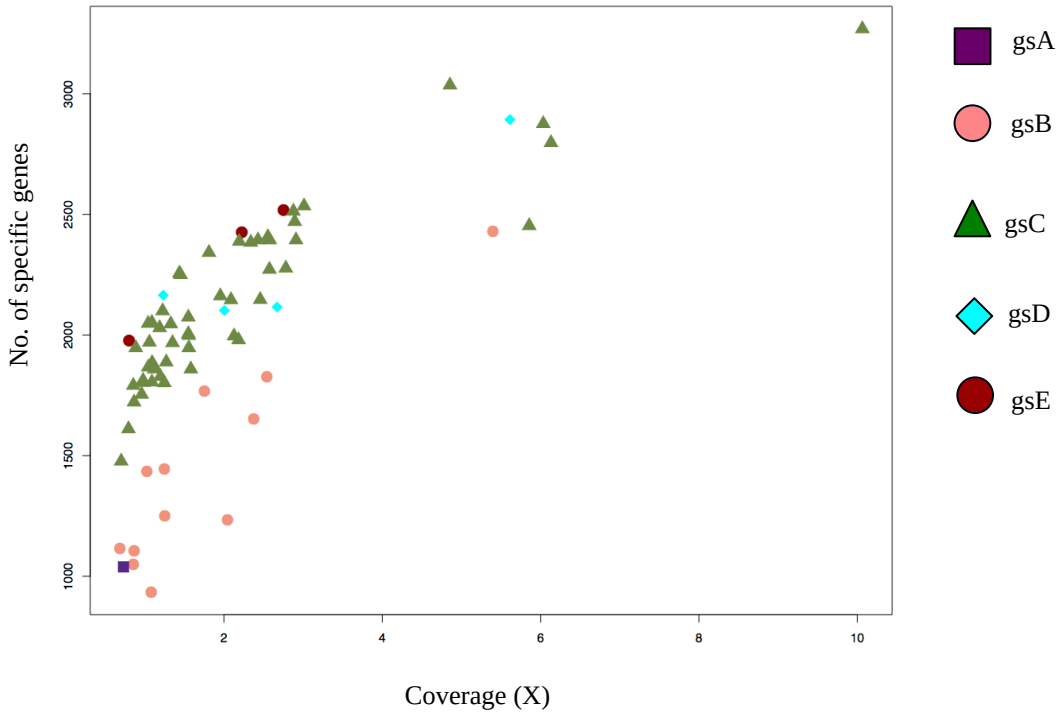


Figure 4.14 | The direct relationship between the number of observed specfic genes and sequence coverage in each of 72 *R. leguminosarum* strains.

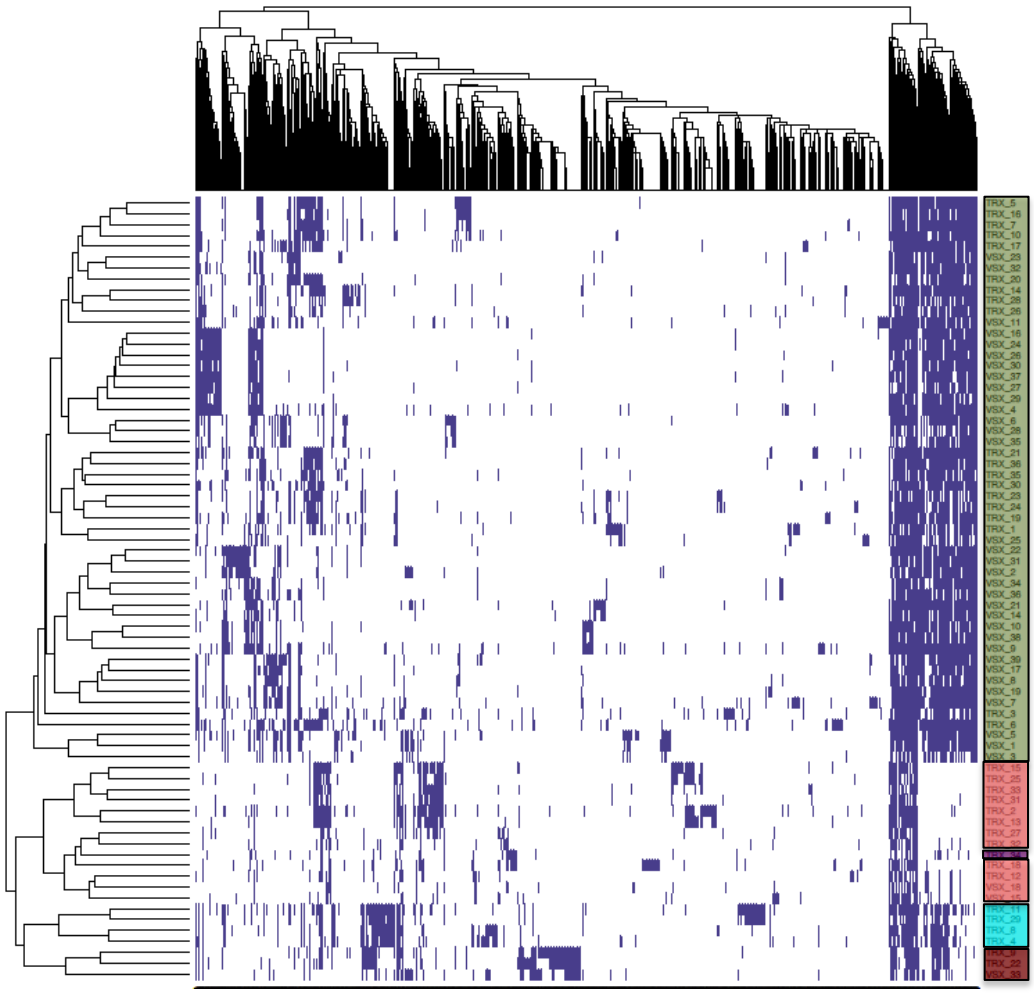


Figure 4.15 | Heatmap of population specific gene presence (blue) and absence (white) in 72 *R. leguminosarum* strains and absent in *Rlv* 3841. Both strains (rows) and genes (columns) were clustered. Strains are highlighted by their related genospecies (A: purple, B: salmon, C: green, D: cyan, E: dark red).

4.5 Discussion

The two standard approaches for genome assembly are: reference based assembly and *de novo* assembly. The reference-based assembly is based on a reference genome from the same organism or a closely related species that allow us to identify and align the reference genes that are present in the assembled genome. A major disadvantage of this approach is that accessory genes of assembled genome that are absent from the reference genome will be ignored. In contrast, *de novo* assembly assembles the genome without the aid of a reference genome. Unlike reference-based assembly, it assembles whole of the genome but is computationally intensive and unable to separate the long tandem repeats (Didelot et al., 2012a). In this chapter, we represented the detailed study of the accessory genome in the members of five genospecies (A-E) of *R. leguminosarum* based on the reference based assemblies (4.3.1-4) using *Rlv* 3841 genome as a reference and a *de novo* assembly (4.3.4).

4.5.1 Presence/absence matrices

The results of presence/absence matrices (Figure 4.3-4.6) suggested that the genes of each replicon (except pRL7 and pRL8) of *Rlv* 3841 are highly conserved in this population. The majority of reference genes were found in the strains of the related genospecies (genospecies B) of the reference genome. The chromosomal genes are highly conserved in this dataset (Figure 4.3), but a few missing genomic islands were observed.

The conserved nature of chromids (Harrison et al., 2010) can be observed in the presence/absence matrices based on the genes of pRL12 and pRL11 replicons (Figure 4.4). The presence of genomic islands in these matrices displayed evidences of HGT in chromid genes. For instance, ABC transporter enriched genomic island (green region) of pRL12 (Figure 4.4A) was located in the members of genospecies B only. Similarly, the genomic island of pRL11 (Figure 4.4B) was conserved in some members of genospecies (C and D), but absent in genospecies B.

Similarly, most of the genes in plasmids pRL10 and pRL9 are widely distributed in this dataset (Figure 4.5A and B). Young et al. (2006) proposed that the genomic region of pRL10 is divided into two compartments. The first compartment is an accessory compartment composed of the first 200 genes, including nodulation genes, while the other compartment consists of core genes. This structure of pRL10 was also reflected in the presence/absence matrix of pRL10: the first compartment of pRL10 is sporadically distributed, whereas the second compartment is highly conserved in this dataset. Also, the matrix shows the conserved nature of *nod* genes in *viciae* strains, which allowed us to conclude that *viciae* strains shared these genes with *Rlv* 3841, while the other symbiovar (*trifolii*) of *R. leguminosarum* harbors a diverged copy of these genes and acquired these genes from a different ancestor.

Although most of the pRL7 and pRL8 genes are absent, the pRL8 presence/absence matrix (Figure 4.6A) helped us to identify five Bvs genes in this replicon that are specifically present in the population of one symbiovar (*viciae*) and absent in other symbiovar (*trifolii*). These genes might play a role in adaptation or have a specific function such as nitrogen fixation, but one that is needed only in certain host plants

4.5.2 Phylogenetic analysis of reference replicons

Because accessory genes are present in a subset of species strains, the chief mechanism that drives these genes is horizontal gene transfer (HGT). This mechanism allows gene exchange between the strains of same or different species, which results in the phylogenetic discordance between genes and species trees (Beiko et al., 2005; Eisen, 2000; Maddison, 1997; Schliep et al., 2011; Tian et al., 2010). We observed that the replicon based phylogenies (except pRL8 and pRL7) showed a similar core gene phylogenetic structure, which indicates that most of the genes in each replicon share the core gene phylogeny. Also, it reflected that there is not much exchange between genospecies. Similarly, Mazur et al. (2011) compared multiple *R. leguminosarum* symbiovar *trifolii* strains in which the phylogenetic structure of core genes was observed in the phylogenies of the chromosome, chromids and plasmids. Mazur et al. (2011) hypothesize the presence of stable genes in each replicon for this phylogenetic similarity.

It is interesting to note that TRX_1 and VSX_25 are outliers in genospecies C in the phylogeny of chromosome (Figure 4.7), pRL11 (Figure 4.8B), pRL10 (Figure 4.9A) and pRL9 (Figure 4.9B). However, these strains were located at different positions in the network of pRL12 (Figure 4.8A). According to the analysis of pRL12-like *repABC* genes (Kim, 2012), the VSX_25 is still an outlier in genospecies, whereas TRX_1 is more typical, similar to VSX_39 (a member of genospecies C). These results suggest that largest chromid (pRL12) was conserved in VSX_25, but TRX_1 acquired this chromid from a more typical genospecies C strain.

The phylogenetic network of pRL7 (Figure 4.10B) is the only plasmid network that showed no similarity with the core gene phylogeny. Surprisingly, it has strong phylogenetic signals that discriminate the two symbiovars (*viciae* and *trifolii*), indicating that a large number of pRL7 genes confer the host specificity. However, weak phylogenetic signals of individual genes were not able to discriminate the two symbiovars. Moreover, this network represents the perfect example of gene transfer between the members of five genospecies in the form of a cluster composed of isolates of genospecies E (TRX_9, VSX_33 and TRX_22), C (TRX_3), B (TRX_12) and D (TRX_8 and TRX_4).

4.5.3 Phylogenetic analysis of nodulation (*nod*) genes

Like other accessory genes, nodulation genes can be transferred vertically (Alvarez-Martinez et al., 2009), horizontally (Laranjo et al., 2008; Tian et al., 2010) or both (Chang et al., 2011b; Menna and Hungria, 2011; Moulin et al., 2004) in different *Rhizobium* species. However, most *nod* studies are based on less genetic information, but this study is based on the 11 *nod* genes of *Rlv* 3841 (pRL10) and *Rlt* WSM1325 (pR132501) producing a more robust analysis. We observed that the five genospecies structure of the core gene phylogeny is disturbed in the nodulation genes network. Moreover, some distinct groups were revealed in the networks of the *viciae* strains (Figure 4.12) and *trifolii* strains (Figure 4.13). These groups showed the evidence of gene transfer events between genospecies such as a group in the *viciae* network (Figure 4.12) composed of 5 strains (VSX_34, VSX_22, VSX_2, VSX_31 and VSX_36) of

genospecies C and one strain (reference plasmid pRL10) of genospecies B. These results suggested the role of HGT in shaping the *nod* genes of this population.

4.5.4 Population specific genes

The identification of genes that are unique to the population can be useful to identify ecologically adaptive genes such as Bailly et al. (2011) found rhizobitoxine synthesis genes. In this population, there are 13,252 specific genes that are absent in the reference genome of *Rlv* 3841. The presence of specific genetic material in each genospecies, shared between genospecies, shared between two strains irrespective of genospecies and strain specific (Figure 4.15) reflects the plasticity of the accessory genome of *R. leguminosarum* and these genes might have an important role in specific ecological adaptations. The accessory genome of *R. leguminosarum* might have a wealth of additional specific genes (genospecies or symbiovars) that have been missed in this study because of limited sequence coverage (Figure 4.14).

Recently, Lassalle et al. (2011) found genospecies specific genes and their associated functions that helped a genospecies of *Agrobacterium tumefaciens* to adapt into a new ecological niche. Interestingly, the unique genes present in the genospecies C, D and E included many ABC transporter genes. These genes were also enriched in the genomic islands of genospecies B that are observed in the presence/absence matrix of pRL12 (Figure 4.4A) and pRL9 (Figure 4.5B). Although functional information of these unique genes is unknown at present, these results provided a clue that these genospecies are phenotypically distinguishable from each other and hence are not cryptic.

In conclusion, this chapter explores the structure and diversity present in the accessory (variable) genome of *R. leguminosarum*. The core genome of the *R. leguminosarum* is not limited to the chromosome and chromids, but extends up to large plasmids. The presence of five Bvs (biovar *viciae* specific) genes located in pRL8 may play a major role in the ecological adaptation. The phylogenetic results of this study demonstrate the presence of five genospecies (A-E) of core gene phylogeny in most of the replicon networks, which describes a low level of gene transfer between genospecies in these replicons. On the other hand, the phylogeny of pRL7 suggests the property of host

specificity that allows pRL7 genes to cross the boundaries of genospecies. The difference between the phylogenetic structure of host specific (*nod*) genes and core genes suggest the occurrence of HGT that allows species members to have different host specificity. The variety of specific genes in this population indicates the presence of different adaptations strategies in *R. leguminosarum.*

Chapter 5. Comparative genomics of two major genospecies of *R. leguminosarum*

5.1 Abstract

Advances in sequencing technologies reveal diversity among strains of bacterial species. In this chapter, we observed genetic diversity in two major genospecies of *R. leguminosarum* by comparing a draft genome of TRX_6 (a member of genospecies C) with the published genome of 3841 (a member of genospecies B). The draft genome of TRX_6 is based on 10.06-fold coverage (89.6 Mb reads), which is assembled into 7.8 Mbp in size and distributed into one chromosome, two chromids and two plasmids. Comparative analyses displayed high similarity in the regions of chromosomes and chromids of 3841 and TRX_6 respectively. One of the TRX_6 plasmids is a unique self-transmissible plasmid with its own replication system. Another plasmid harbors homologous genes of pRL9 and pRL10 (3841 plasmids) with the replication system of pRL10. A well-conserved region of conjugative genes was identified in the chromids as well as plasmids. This comparative analysis reflects the genomic diversity present between two major genospecies of *R. leguminosarum*.

5.2 Introduction

With the aid of sequencing technologies and bioinformatics tools, bacterial genomes can be compared conveniently at the genomic level. Comparative genomics shed light on the genetic variation present in the strains of a bacterial species (Pallen and Wren, 2007) and often used to discover new plasmids or genomic islands in a population. There are numerous bioinformatics tools that aid in two basic steps (sequence alignment and visulization) of comparative genome analysis (Edwards and Holt, 2013). The most common tool for sequence alignment is BLAST followed by MUMmer (Delcher et al., 2002) and Mauve (Darling et al., 2004). Sequence visualization can be represented in linear or circular layouts. Linear layouts can be produced by ACT (Carver et al., 2005), MUMmer (Delcher et al., 2002) and Mauve (Darling et al., 2004) and are useful to identify rearrangements. However, these layouts are not suitable for analyzing multiple datasets. For which, circular layouts are preferred and can be constructed by using CGView (Stothard and Wishart, 2005), Circos (Krzywinski et al., 2009), and BRIG (Alikhan et al., 2011).

The results of the previous chapters revealed five cryptic genospecies (A-E) present in *R. leguminosarum* species that are shaped by recombination. Genospecies B and C are two major genospecies (Figure 2.7) in which most of the strains are clustered. In this chapter, we studied the genetic differences between genospecies B and C for which we chose the draft genome of TRX_6 (strain with highest sequence coverage (Figure 2.4) among all the draft *R. leguminosarum* genomes used in this study) as a representative of genospecies C. The genomic data of TRX_6 was compared with the replicons of 3841 (representative of genospecies B) using different bioinformatics tools. For example, a user-friendly Java application, CGView (Stothard, Wishart 2005), was used to visualize the genomic properties such as GC content, GC skew, annotation, Blast results etc. CONTIGuator (Galardini et al., 2011), a genome-finishing tool, was used for accurate assembly of TRX_6 based on a reference genome of 3841.

5.2.1 Objectives

The main objectives of this chapter are:

A. *De novo* and reference-based assembly of TRX_6 using a reference genome.
B. Comparative analysis of TRX_6 scaffolds with 3841 replicons.
C. Correct assembly of TRX_6 chromosomal related scaffolds using CONTIGuator.
D. Annotation of TRX_6 genome using RAST annotation server.

5.3 Material and Methods

In Chapter 2, we performed the *de novo* assembly of 72 *R. leguminosarum* strains (Figure 2.4 and 2.5) including the TRX_6 strain using Newbler 2.5 (Roche). In this chapter, we performed a reference-based assembly of TRX_6 strains with a reference genome of *Rlv* 3841 using Newbler 2.5 (Roche).

A java application known as CGView (Stothard and Wishart, 2005) was used to visualize the alignments of TRX_6 scaffolds with 3841 replicons (Chromosome and pRL12-pRL7) using local blast (BLASTn) with cut off e value of 1e-05 (most general value) in a circular fashion. This was achieved by modifying the Perl scripts of get_cds.pl and local_blast_client.pl that were available at

http://www.ualberta.ca/~stothard/software.html.

The Perl script of cgview_xml_builder.pl (available in CGView) was used to construct the eleven XML files. Each XML file consisted of information for each scaffold and its blast results with 3841 replicons. The command line of CGView (java –jar cgview.jar -i input.xml –o output.png -f png) was used to produce circular maps of scaffolds from each of the 11 XML files.

MUMmer 3.22 (Delcher et al., 2002) was used to align large the TRX_6 scaffolds (Scaffolds 1-8) with 3841 replicons. The following commands were used:

1. ./nucmer -p output_name reference query
2. ./mummerplot output.delta --layout --filter -t postscript

CONTIGuator 1.1 (Galardini et al., 2011) was used to map the contigs of TRX_6 chromosomal scaffolds (scaffolds that showed high synteny with the *Rlv* 3841 chromosome) against the chromosomal sequence of 3841. The results of CONTIGuator were observed using MUMmer 3.22. The RAST (Rapid Annotations using Subsytems Technology) server provided by Aziz et al. (2008) was used for automatic annotation of the 690 contigs of the TRX_6 assembly.

5.4 Results

5.4.1 De novo assembly of TRX_6

The *de novo* assembly of the TRX_6 strain suggested the presence of eleven scaffolds that comprised 560 contigs (Table 5.1). Scaffold 1 contained the maximum number of contigs (261), while each scaffold (9, 10 and 11) was comprised of one contig only. However, some of the contigs (130) were not assembled in any of the scaffolds.

Table 5.1 | The TRX_6 scaffolds with their length and related number of contigs.

Scaffold	Length (Mbp)	No. of contigs
1	3.240047	261
2	0.927736	66
3	0.784154	48
4	0.714784	44
5	0.648505	32
6	0.574747	27
7	0.510855	54
8	0.423327	25
9	0.003711	1
10	0.002921	1
11	0.002174	1
	7.832961	560

5.4.2 Reference-based assembly of the TRX_6 genome with 3841

The reference-based assembly indicated that 72.55% of the TRX_6 reads were mapped to the 3841 genome, while 27.45% of genomic data was unmapped. The chromosome of 3841 was best covered (Table 5.2) by the TRX_6 reads, whereas the two small plasmids (pRL7 and pRL8) were least covered.

Table 5.2 | Percentage coverage of 3841 replicons with TRX_6 genome.

Replicons of 3841	Percentage coverage
Chromosome	84.77%
pRL12	67.44%
pRL11	70.46%
pRL10	59.92%
pRL9	63.57%
pRL8	3.79%
pRL7	5.60%

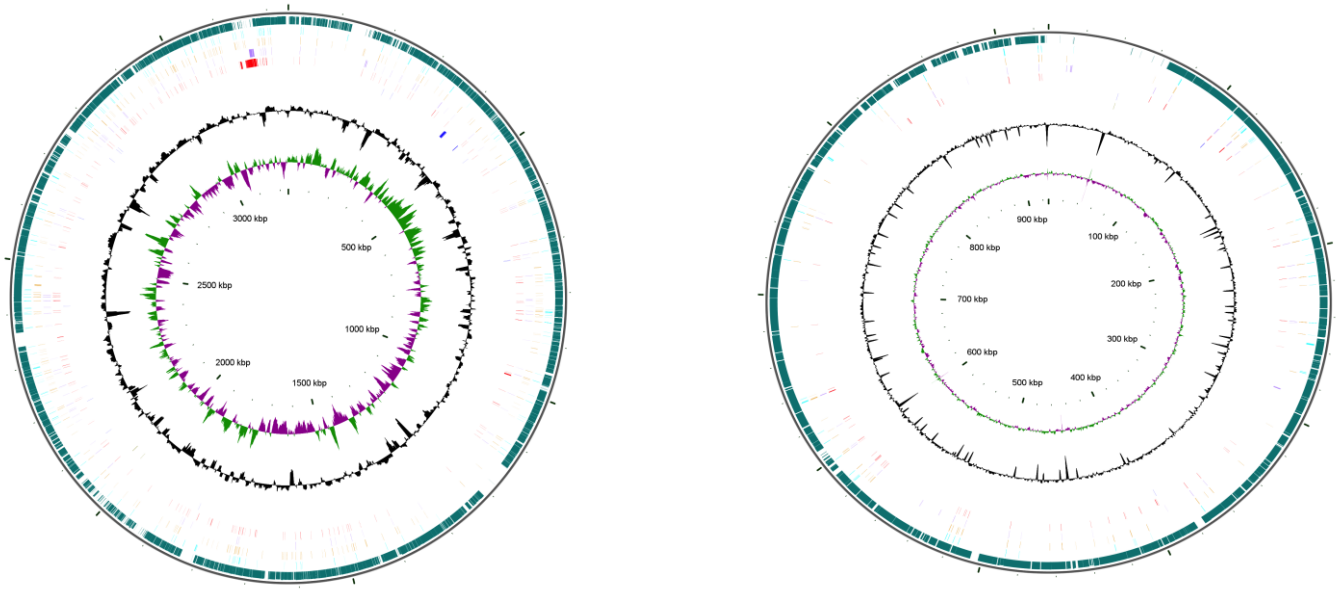
5.4.3 Comparative analysis of TRX_6 with 3841 using CGView and MUMmer

CGView generated eleven circular maps (Figure 5.1-5.5) showing alignment of TRX_6 scaffolds with each of the replicons of 3841. Multiple scaffolds (scaffold 1, 2, 6, 8, 9, 10, 11) shared high similarity with the 3841 chromosome. Scaffold 3 exhibits high similarity with pRL12. Scaffold 4 displayed high resemblance with pRL11. A part of Scaffold 5 displayed similarity with half of the pRL10 genes (236/446), while another part reflects the genes of pRL9 (199/305). Scaffold 7 showed little similarity with any of the *Rlv* 3841 replicons. All the scaffolds were deficient in the genes of pRL7 and pRL8.

Intriguingly, circular maps of scaffolds 3, 4, 5 and 7 reflected a homologous region present in them (black rectangle in Figure 5.2 [A, B], 5.3 [A] and 5.4 [A]). This homologous region displayed high similarity with the region of pRL12, pRL11, and pRL10 that includes genes such as conjugal transfer protein *traA*, putative mobilization

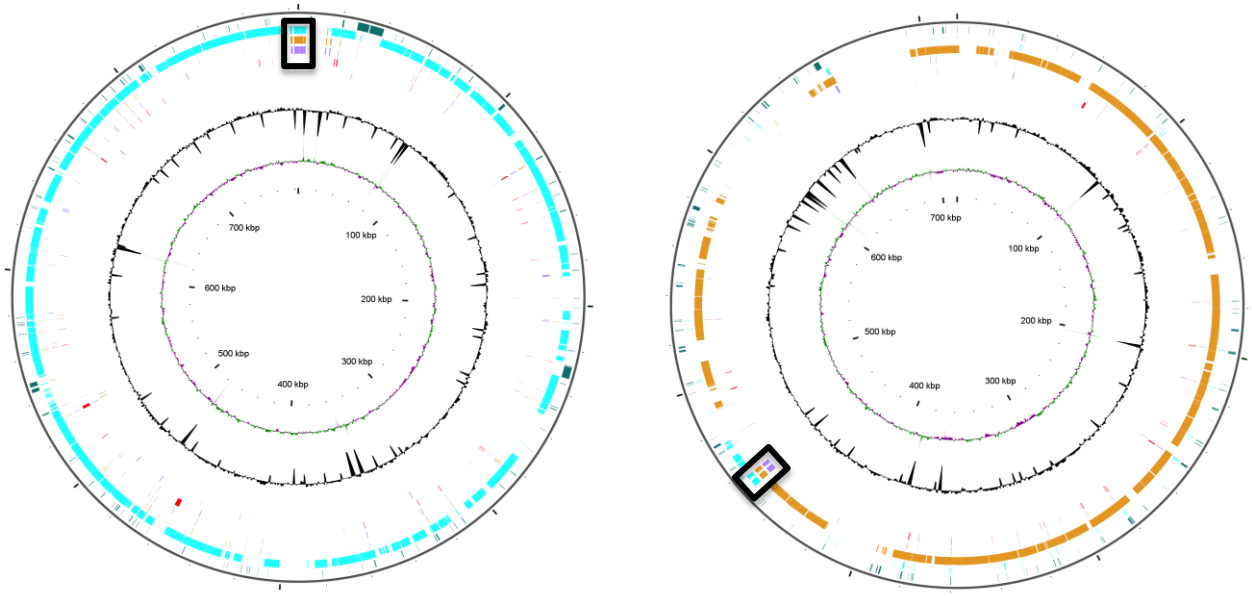
protein, putative transmembrane *traG* transfer-related protein and transcriptional regulators. A detailed description of the genes of pRL12, pRL11 and pRL10 that are allocated to this region is given in Table 5.3. The scaffold 7 carried almost all the essential conjugative genes (*traA* encoding Dtr system *oriT* relaxase, *trbB*, *trbI*, *trbH* and *traG*), and these showed sequence similarity with the conjugative system of the pR132503 (Reeve et al., 2010a) of *Rlt* WSM1325.

The GC content of each scaffold revealed some potential genomic islands, such as various regions of low GC-content in scaffold 7 (Figure 5.4A). The GC skew is not clearly observed in any of the scaffolds.



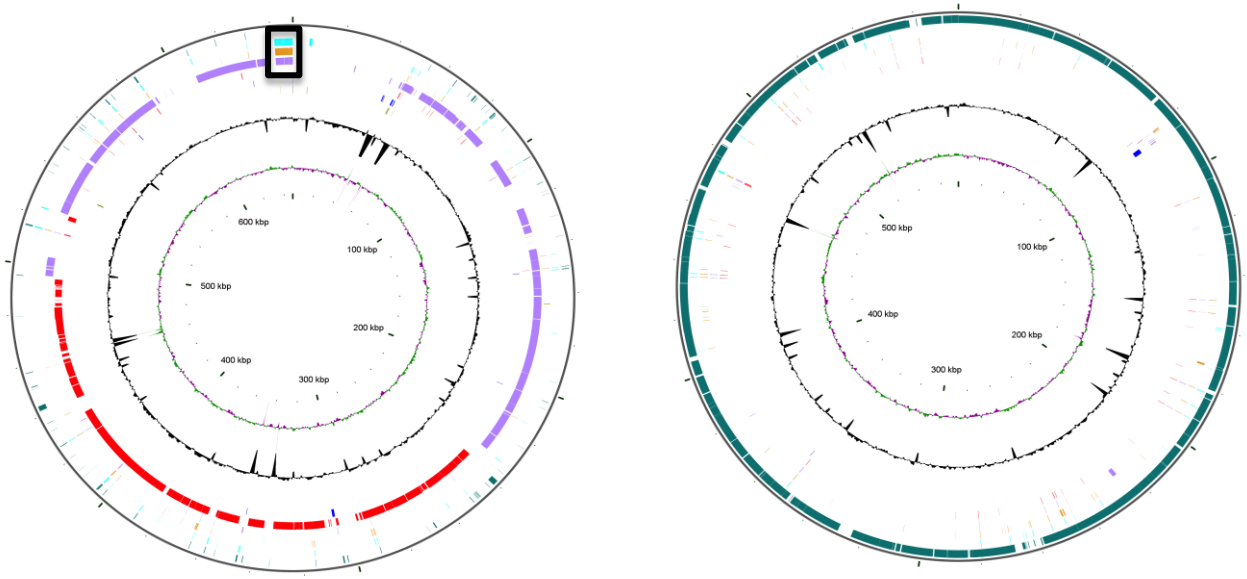
A **B**

Figure 5.1 | The circular maps of Scaffolds 1 (A) and 2 (B) of TRX_6. From outside to centre: Alignments are shown in different ring colours: chromosome (blue-green), pRL12 (cyan), pRL11 (gold), pRL10 (violet), pRL9 (red), pRL8 (blue) and pRL7 (olive green) of 3841 using BLASTn, GC content (black), GC skew (+: green, -: dark purple).



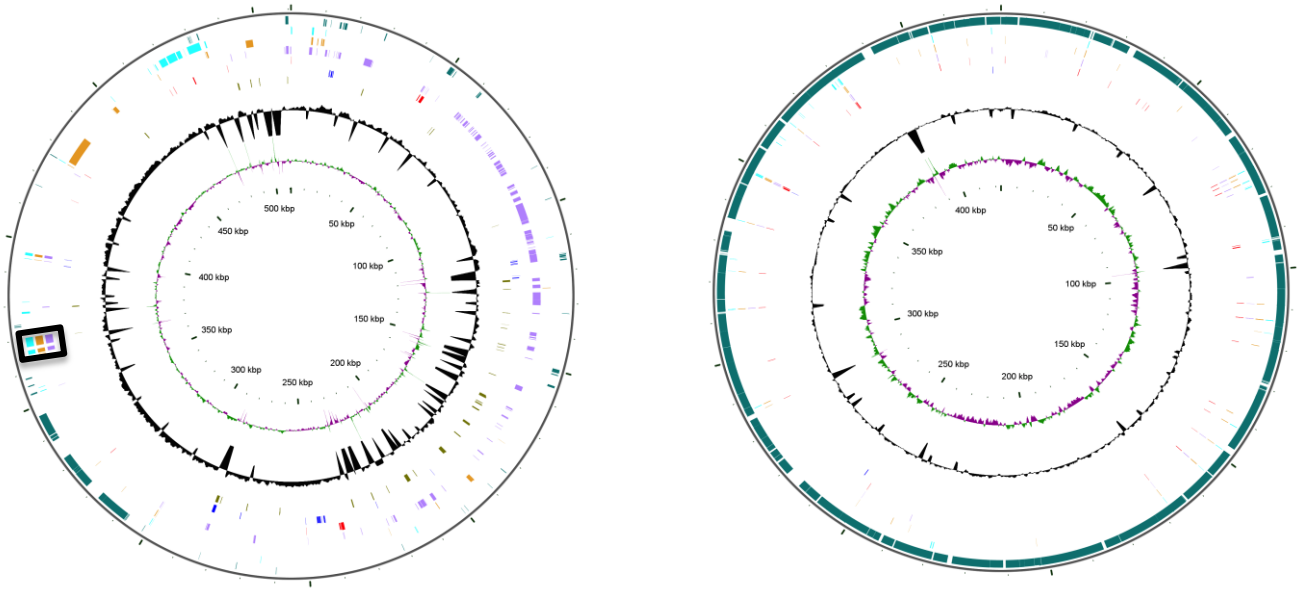
A **B**

Figure 5.2 | The circular maps of Scaffolds 3 (A) and 4 (B) of TRX_6. From outside to centre: Alignments are shown in different ring colours: chromosome (blue-green), pRL12 (cyan), pRL11 (gold), pRL10 (violet), pRL9 (red), pRL8 (blue) and pRL7 (olive green) of 3841 using BLASTn, GC content (black), GC skew (+: green, -: dark purple). The conserved region is shown in the black rectangle in A and B.



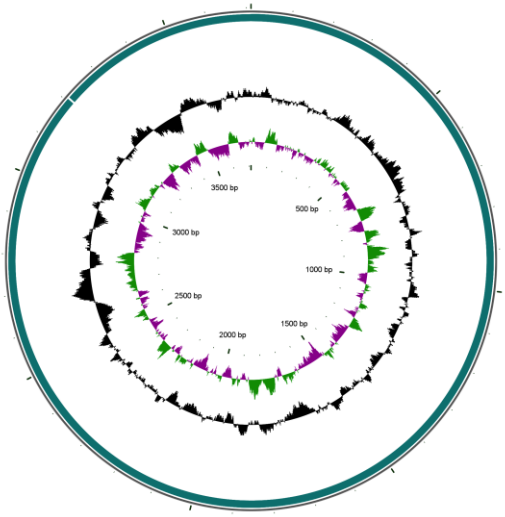
A **B**

Figure 5.3 | The circular maps of Scaffolds 5 (A) and 6 (B) of TRX_6. From outside to centre: Alignments are shown in different ring colours: chromosome (blue-green), pRL12 (cyan), pRL11 (gold), pRL10 (violet), pRL9 (red), pRL8 (blue) and pRL7 (olive green) of 3841 using BLASTn, GC content (black), GC skew (+: green, -: dark purple). The conserved region is shown in the black rectangle in A.

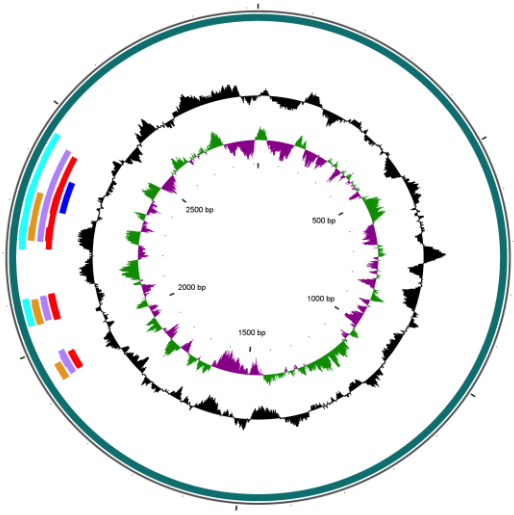


A **B**

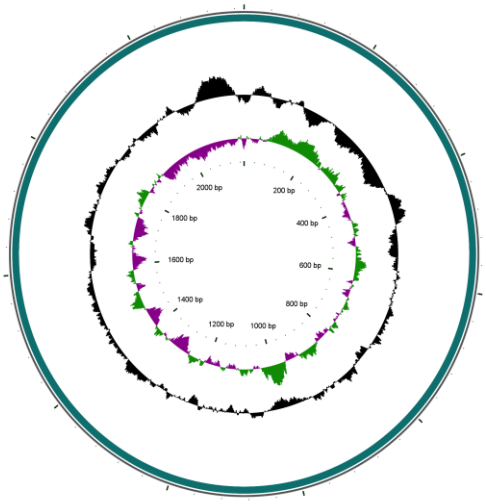
Figure 5.4 | The circular maps of Scaffolds 7 (A) and 8 (B) of TRX_6. From outside to centre: Alignments are shown in different ring colours: chromosome (blue-green), pRL12 (cyan), pRL11 (gold), pRL10 (violet), pRL9 (red), pRL8 (blue) and pRL7 (olive green) of 3841 using BLASTn, GC content (black), GC skew (+: green, -: dark purple). The conserved region is shown in the black rectangle in A.



Scaffold 9



Scaffold 10



Scaffold 11

Figure 5.5 | The circular maps of Scaffolds 9, 10 and 11 of TRX_6. From outside to centre: Alignments are shown in different ring colours: chromosome (blue-green), pRL12 (cyan), pRL11 (gold), pRL10 (violet), pRL9 (red), pRL8 (blue) and pRL7 (olive green) of 3841 using BLASTn, GC content (black), GC skew (+: green, -: dark purple).

Table 5.3 | Description of genes of pRL10, pRL11 and pRL12 that are allocated in the homologous region of scaffolds 3, 4, 5 and 7.

Locus tag	Strand	Name	Function
pRL100214	+	-	LysR family transcriptional regulator
pRL100215	+	-	putative DNA conjugation-related protein
pRL100216	-	-	hypothetical protein
pRL100217	-	-	putative conjugal transfer protein
pRL100218	-	-	putative mobilization protein
pRL100219	+	*traA*	conjugal transfer protein TraA
pRL110270	+	-	putative ArdC antirestriction protein
pRL110271	-	-	putative WGR-family protein
pRL110272	+	-	putative HTH-type transcriptional regulator
pRL110274	-	-	putative transmembrane traG transfer-related protein
pRL110275	-	-	putative mobilization protein MobS
pRL110276	+	*traA*	putative conjugal transfer protein TraA
pRL120562	+	-	LysR family transcriptional regulator
pRL120563	+	*ardC*	putative ArdC antirestriction protein
pRL120564	-	-	hypothetical protein
pRL120565	-	-	putative transmembrane traG homologue/component of type IV secretion system
pRL120566	-	*mobC*	putative mobilization protein C
pRL120567	+	*traA*	putative conjugal transfer protein TraA

Based on these BLAST results, the scaffolds were arranged into TRX_6 replicons (Table 5.4). The genes of the TRX_6 chromosome were scattered on four scaffolds (1, 2, 6 and 8). Smaller and insignificant scaffolds (9-11) were ignored. Two chromids (scaffolds 3 and 4) and two plasmids (scaffolds 5 and 7) were other replicons of this genome.

Table 5.4 | Eleven scaffolds of TRX_6 are arranged on the basis of their similarity with 3841 replicons.

TRX_6	3841
Scaffold 1,2,6,8	Chromosome
Scaffold 3	pRL12
Scaffold 4	pRL11
Scaffold 5	pRL9 & pRL10
Scaffold 7	Absent

To determine the major rearrangements in the scaffolds, we performed whole genome alignment of scaffolds (1-8) against 3841 replicons (Figure 5.6). Each dot (red for the forward strand and blue denotes the reverse strand) in the figure represents a MUM (Maximal Unique Match). The replicons of 3841 are on the X-axis, while TRX_6 scaffolds are located on the Y-axis. The results were mirrored by the BLAST results such as chromosomal genes are scattered on the four scaffolds 1, 2, 6 and 8. Some genomic rearrangements in the chromid and plasmids were detected such as insertion in pRL9, insertion in scaffold 5 and clear example of inversion in pRL12/scaffold3.

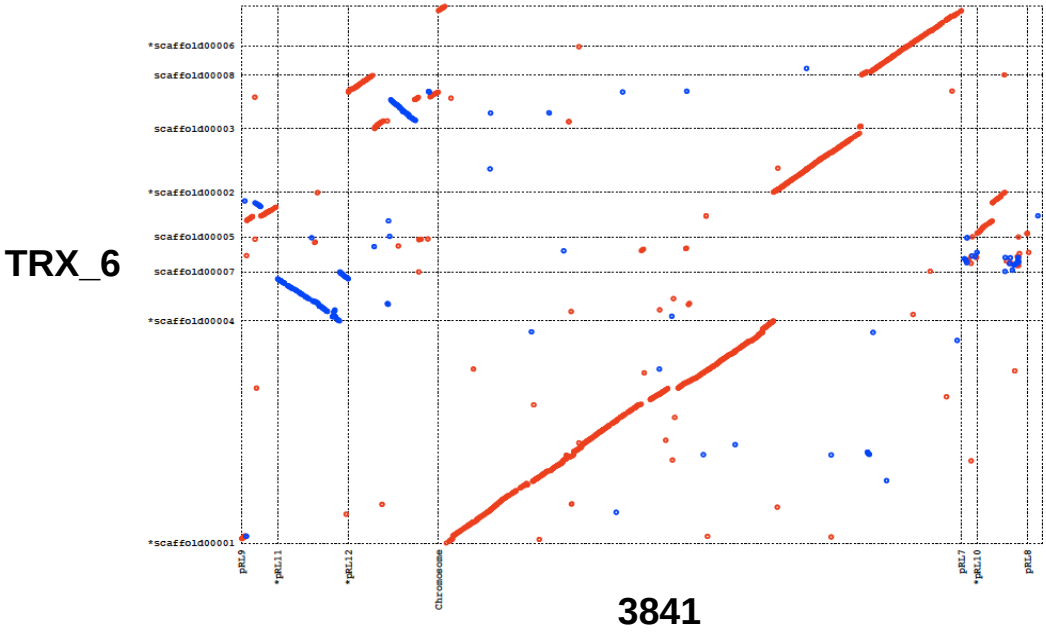


Figure 5.6 | The Nucmer plots of TRX_6 scaffolds and 3841 replicons highlighting synteny relationship between them. Each dot represents a MUM (Maximal Unique Match). Red dots: MUMs in forward direction. Blue dots: MUMs in reverse direction.

5.4.4 Correct assembly of TRX_6 chromosomal related scaffolds using CONTIGuator

To assemble the chromosomal related scaffolds of TRX_6 in the correct order and orientation, we extracted and mapped the contigs of scaffolds 1, 2, 6 and 8-11 against the *Rlv* 3841 chromosome. The CONTIGuator considered only contigs of the scaffolds 1, 2, 6 and 8 to construct a whole sequence of the Chromosome. The CONTIGuator results were observed using MUMmer. In Figure 5.7, the TRX_6 chromosome and scaffolds 3, 4, 5 and 7 were aligned against 3841 replicons. Each dot (red for a forward strand and blue for a reverse strand) in the figure represents a MUM (Maximal Unique Match). The replicons of 3841 are on the X-axis, while the chromosome and scaffolds of TRX_6 were located on the Y-axis. Extensive synteny was observed between the TRX_6 chromosome and 3841, which is interrupted by two large clear as well as small hidden genomic islands.

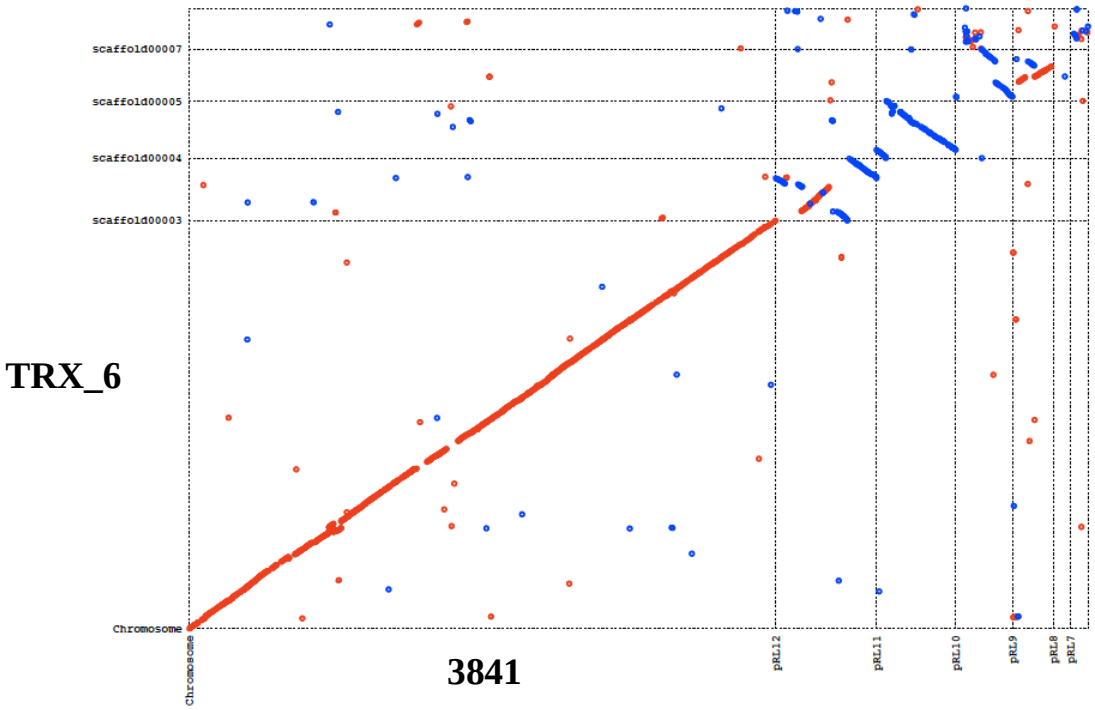


Figure 5.7 | The Nucmer plots of TRX_6 and 3841 replicons highlighting synteny relationship between them and correct ordering of chromosomal related scaffolds. Each dot represents a MUM (Maximal Unique Match). Red dots: MUMs in forward direction. Blue dots: MUMs in reverse direction.

5.4.5 Automatic annotation of TRX_6 genome

The functional annotation from the RAST server identified 7551 CDS. The genes were functionally categorized under 421 subsystems (Figure 5.8). In Figure 5.8, the histogram shows the coverage of subsystems (40%) determined by RAST server in this genome. The pie chart (Figure 5.8) obtained from RAST server reflects the distribution of predicted genes in different subsystems. There are many carbohydrate subsystems including central carbohydrate and monosaccharide metabolism. Another major subsystem feature is amino acids and derivatives, including lysine, threonine, methionine, and cysteine.

The replication systems (*repABC*) were located in each of the chromid and plasmid scaffolds. The replication system of scaffold 3 (contig00360) showed high similarity with the pRL12 system. Scaffold 4 (contig00380) carried a replication system that shared high similarity with the pRL11 system. The replication system of scaffold 5 (contig00423) shared high similarity with the pRL10 system. The replication system of scaffold 7 (contig00509) was absent in 3841 and displayed high similarity with pR132503 (*Rlt* WSM1325) system.

The expected nodulation genes (*nod N*, *M*, *L*, *E*, *F*, *D*, *A*, *B*, *C*, *I*, *J*) were present in this genome. The coding sequences of eleven nodulation genes (*nod A*, *B*, *C*, *D*, *E*, *F*, *I*, *J*, *L*, *M*, *N*) were clustered on the contig 00484 of scaffold 7 (Figure 5.9). In addition to these nod genes, *nodX* (Davis et al., 1988) was also present on the same contig (Figure 5.9) of scaffold 7 with other nod genes.

As found earlier in this chapter, the RAST server also suggested the presence of coding sequences of conjugation genes and mobilization genes on each chromid and plasmid related scaffold. Scaffold 3 harbors a sequence of putative conjugal transfer protein *traA* in contig00374 and putative mobilization protein C in contig00328. Scaffold 4 includes a sequence of putative conjugal transfer protein *traA* located in contig00400 and putative mobilization protein *mobS*, which is present on contig00402 adjacent to contig00400. Scaffold 5 includes a sequence of conjugal transfer protein and putative mobilization protein C. Both of them are located in contig00451. Scaffold 7 includes a

sequence of putative mobilization protein C in contig00517, while a conjugal transfer protein is located in contig00518 adjacent to contig00517.

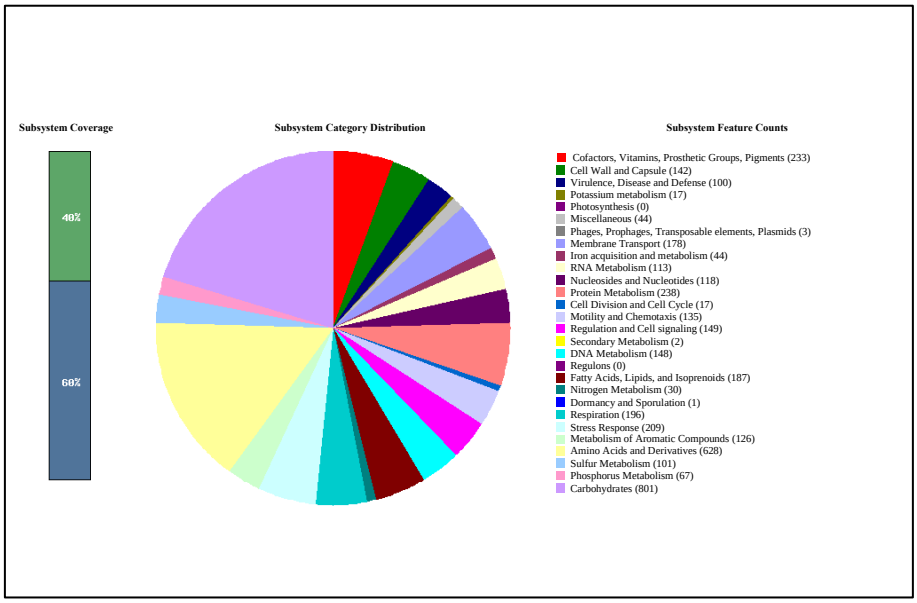


Figure 5.8 | The functional subsystems present in TRX_6 determined by the RAST server.

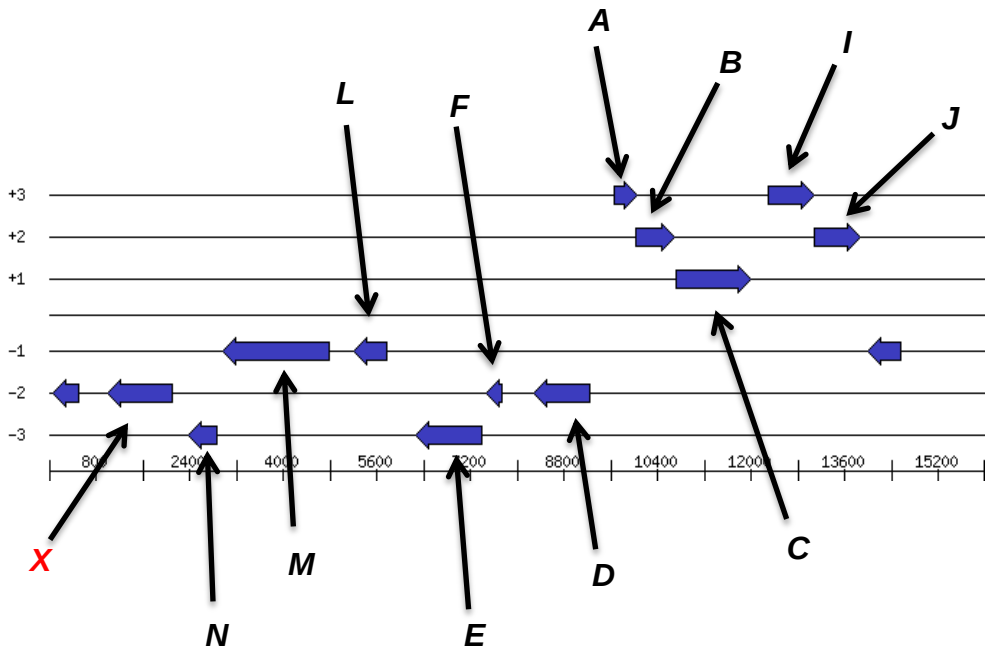


Figure 5.9 | The twelve *nod* genes of TRX_6 (symbiovar trifolii) are located in scaffold 7. The scale (x-axis) describes the length of the related contig 00484. The Y-axis represents the codon phase information (GFF format). The symbiovar *trifolii* specific *nod* gene, X, is shown in red.

5.5 Discussion

This chapter shows a comparative genomic analysis of two genospecies (B and C) by selecting one strain from each of the species as their representative (TRX_6: genospecies C and 3841: genospecies B) strain.

The results (CGView and MUMmer) suggested that TRX_6 has one chromosome, two chromids (scaffolds 3 and 4) and two plasmids (scaffolds 5 and 7). A syntenic relationship between both chromosomes and chromids (pRL12: scaffold3, pRL11: scaffold4) indicates a common origin. These observations are confirmed by the similarity of the replication system (*repABC*) in chromid pairs (Figure 4.2).

Interestingly, scaffold 5 is the integrated plasmid that contained homologous genes of pRL10 and pRL9, but replication system of pRL10 only indicating a common ancestor of pRL10 and scaffold 5. Although homologous genes of pRL9 were present in TRX_6 and other members of genospecies C (Figure 4.5B), homologous replication system of pRL9 (Kim, 2012) was absent in members of genospecies C (Figure 4.2) including TRX_6. We conclude that all genospecies C members may have this type of cointegrate plasmid. On the other hand, all members of genospecies B possessed the homologous replication system of pRL9, which suggested the presence of separate pRL10-like and pRL9-like plasmids, like 3841, rather than a cointegrate like TRX_6.

Another plasmid (scaffold 7) of TRX_6 shared its ancestry with pR132503 (Reeve et al., 2010a) of *Rlt* WSM1325 and was absent in *Rlv* 3841. The homologous genes of replication system of pR132503 were conserved in members of genospecies C except TRX_3, VSX_1, VSX_3 and VSX_5 (Figure 4.2), while only a few members of genospecies B (TRX_12, TRX_18, TRX_31, VSX_15, and VSX_18) had this replication system (Kim, 2012). Based on these results, we conclude that pR132503-type plasmid has ubiquitously distributed in genospecies C, but rarely in members of genospecies B.

Nodulation (*nod*) genes are essential for fixing nitrogen in leguminous plants. These genes are generally located on the plasmids known as symbiosis plasmids or within

genomic islands. Studies (Crossman et al., 2008; Gonzalez et al., 2010; Lozano et al., 2010) suggest that symbiosis plasmids have a common origin. In TRX_6, these genes were located on scaffold 7 that shared its ancestry with the non-symbiosis plasmid (pR132503) of *Rlt* WSM1325 (Reeve et al., 2010a) not with the symbiosis plasmid (pR132501) of *Rlt* WSM1325 (Reeve et al., 2010a). These results allow us to conclude that *nod* genes are transferred between plasmids, so that symbiosis plasmids are not necessarily homologous outside the symbiosis gene region.

Additionally, a symbiovar *trifolii* (host) specific gene known as *nodX* (Davis et al., 1988) has been detected in TRX_6, which is absent in *Rlv* 3841. However, there are a few exceptional *viciae* strains such as *Rlv* TOM that harbor *nodX* genes and can nodulate Afghanistan peas (Davis et al., 1988). The sequence of *nodX* genes is found in all the draft genomes of the *trifolii*, but absent in *viciae* strains except VSX_1. Based on these results, we conclude that *nodX* is a highly conserved gene in the symbiovar *trifolii*, but uncommon in symbiovar *viciae* strains.

Conjugative plasmids are the key agents for the bacterial diversity and harbor specialized accessory genes. Conjugative plasmids are also observed in *Rhizobium* species such as *R. etli* CFN 42 that has two conjugative plasmids: p42a and p42d (Crossman et al., 2008; Tun-Garrido et al., 2003). The replication system (*repABC*) of two conjugative plasmids (pRL7 and pRL8) of 3841 (Young et al., 2006) is absent in the TRX_6 genome. Two chromids (scaffold 3 and 4) and a plasmid (scaffold 5) in TRX_6 displayed the presence of mobilisation genes related to those of non-self transmissible replicons (pRL12, pRL11, and pRL10) of 3841 (Ding et al., 2013; Young et al., 2006), whereas the presence of almost all the conjugative genes (*traA* encoding Dtr system *oriT* relaxase, *trbB*, *trbI*, *trbH* and *traG*) suggests the self-transmissible characteristic of scaffold 7.

A further detailed comparative study could be useful to predict genomic islands that are present or absent in these two genomes. In order to explore genetic diversity, strains of different genospecies can be compared against the reference genome of TRX_6.

In conclusion, we report a significant similarity between genospecies C (TRX_6) and genospecies B (3841), which are two major genospecies of *R. leguminosarum*. The draft genome of TRX_6 includes one chromosome, two chromids and two plasmids. The two plasmids of TRX_6 are the main source of genetic differences. One of them is a unique self-transmissible plasmid that shares evolutionary history with the *Rlt* WSM1325 plasmid and present in all members of genospecies C except TRX_3, VSX_1, VSX_3 and VSX_5, but absent in 3841 and other members of genospecies B. However, another cointegrate plasmid is present in all members of genospecies C, but absent in members of genospecies B.

Chapter 6. General Discussion

The purpose of this study was to explore the nature of bacterial species and the role of recombination barriers, based on a set of population genomic data. The ecotype model is one of the most famous species models in which bacterial species are further classified into ecologically distinct populations (ecotypes) that are governed by the cohesive force of periodic selection, while the role of recombination in the divergence of these species is insignificant. Sequence clusters obtained by massive sequencing and new methodologies strongly correspond with bacterial ecotypes. However, this thesis has shown that genospecies based on core genes can be different from ecological clusters that are specialized for separate host plants. In contrast to periodic selection, recombination can have a significant role in the divergence of different genospecies.

6.1 Synopsis

Chapter 2 described the presence of genospecies that are different from symbiotic ecotypes. It was found that a local population of *R. leguminosarum* falls into five discrete clusters that correspond to five cryptic genospecies (A-E) based on core genes and ANI analysis (a robust alternative to DDH) that did not seem to be congruent with bacterial species based on ecological properties, in which ecotypes would correspond to distinguishable genetic clusters (Cohan and Perry, 2007; Connor et al., 2010; Didelot et al., 2011; Ward et al., 2008). This finding is much the same for the species within *Bacillus cereus* sensu lato (Zwick et al., 2012) where plasmid-encoded pathogenic properties are used to classify three closely related species (*B. cereus, B. anthracis, B. thuringiensis*) of this group (Guinebretiere et al., 2008; Helgason et al., 2000). However, clusters of mixed strains of these species based on MLST (Priest et al., 2004) and whole genome sequencing (Zwick et al., 2012) indicated the presence of common core genomic backgrounds as well as evidence of HGT between these species that allows species members to have different pathogenic properties. For example, strains of *B. cereus* (Klee et al., 2010; Oh et al., 2011) that are closely related to *B. anthracis* have ability to cause anthrax because these *B. cereus* strains contained plasmids similar to the anthrax related plasmids present in *B. anthracis* (Okinaka et al., 1999). Similarly, the presence of insecticidal crystal toxin gene is the only genetic difference that differentiate *B. thuringiensis* from *B. cereus*. Another related study involves the *Shigella* bacterium (human pathogen) evolved from commnesal *Escherichia coli* (*E. coli*). Initially, *Shigella* was classified as a separate and closely related genus to *Escherichia coli* (Ewing, 1949), which is no longer valid because genomic evidences such as common chromosomal genes, acquistion of virulence plasmids and loss of virulence genes indicates that these are human-adapted pathovars of *Escherichia coli* that can invade the intestinal epithelium and cause dysentry (Holt et al., 2012; Pupo et al., 2000; Yang et al., 2005).

It was observed that each genospecies was comprised of mixed symbiovars; for example, the reference genome was clustered with symbiovars in one of the genospecies. This genospecies has been observed in other regions of the world such as

Sweden, Scotland and China, indicating that this genospecies is not confined to this isolated site (Wentworth, York).

Population genomics of the *R. leguminosarum* species complex (*R. leguminosarum, R. etli, R. pisi, R. fabae and R. phaseoli*) explored the relatedness of five genospecies with other *R. leguminosarum* strains and existence of different members of one of the genospecies that are isolated from different locations. Moreover, this study made significant contribution to the rhizobium taxonomy by endorsing a study (Lopez-Guerrero et al., 2012) that suggests reclassifying CIAT652 as a strain of *R. phaseoli* and confirmed the classification of two novel species (*R. pisi* and *R. fabae*) into one single species as discussed by Alvarez-Martinez et al. (2009) based on a relatively small amount of data.

Chapter 3 shed light on the role of recombination in the maintenance of five genospecies. We observed that most of the core genes were affected by recombination supporting the importance of chromosomal recombination in the evolution of *R. leguminosarum*; this is interesting because recombination in the chromid and plasmid was more frequent than the chromosome in a population of *S. medicae* (Bailly et al., 2011). The results from ClonalFrame analysis supported the presence of five genospecies in our dataset and indicated higher rates of recombination between members of the same genospecies rather than between them. These data endorsed the theory of recombination (Fraser et al., 2009; Fraser et al., 2007; Majewski et al., 2000; Roberts and Cohan, 1993; Zawadzki et al., 1995) in which decline of recombination rate is associated with increase in sequence divergence. Moreover, the Structure analysis inferred five ancestral populations in our dataset that strongly correlated with these biological genospecies. ClonalFrame and Structure analysis have been widely used for detecting the population structure and bacterial species in different bacterial populations (Cadillo-Quiroz et al., 2012; Didelot et al., 2012b; Doroghazi and Buckley 2010; Joseph et al., 2012; Shapiro et al., 2012). The recently diverged populations observed in ocean bacteria (Shapiro et al., 2012) and a thermoacidophilic archaeon (Cadillo-Quiroz et al., 2012) are compatible with the ecotype model, but our study involves a population of symbiovars, which is a much later stage in bacterial speciation and displays a lack of association between genetic clusters and ecological adaptation.

Chapter 4 investigated the genetic diversity and phylogenetic structure of five genospecies in the accessory genome of *R. leguminosarum*. The presence/absence matrices based on the genes of reference replicons indicated that core genes were not only located on the chromome and chromids, but also on large plasmids. Although genetic information of small plasmids was almost absent, symbiovar specialization of five specific (Bvs) genes in one of the smallest plasmids reflected genetic plasticity in symbiovar *viciae* isolates. In the context of phylogeny, reference based replicon phylogenies displayed the same five genospecies structure as observed in core genes phylogeny implies lack of transfer between genospecies in different replicons of *R. leguminosarum*. It is interesting to note that one of the reference based plasmid phylogenies showed clusters of mixed strains of different genospecies with signals of symbiotic properties. The *nod* genes are among the host specific symbiosis genes that are generally maintained by homologous recombination or HGT. Findings show that these genes were maintained by HGT, which allow them to cross the genospecies barriers. It was found that a pool of accessory genes was available in a local population, which is absent in reference genome. On the other hand, genetic variation in the shared population of *S. medicae* (Bailly et al., 2011) was very low. Remarkably, clusters of genes with unknown functions were able to differentiate five genospecies. Potentially, these genes could explain the phenotypic differences between these genospecies. This finding parallels the findings of Lassalle et al. (2011), where genes specific to a genospecies of *Agrobacterium tumefaciens* were identified, functionally annotated and experimentally confirmed.

Chapter 5 described the genetic differences between two major genospecies (B and C) of *R. leguminosarum* based on the comparative analysis of two representative strains. Genospecies C is represented by the draft genome of TRX_6, which is classified into one chromosome, two chromids and two plasmids, whereas the fully sequenced genome of 3841 represents genospecies B. A significant similarity was observed in the regions of chromosomes and chromids between TRX_6 and 3841 strains. The majority of genetic differences were observed in the regions of plasmids. One of the TRX_6 plasmids was present in other members of genospecies C, but was absent in 3841 and in most of the other members of genospecies B. Another plasmid of TRX_6 was a cointegrate plasmid comprised of homologous genes of two 3841 plasmids (pRL9 and

pRLl0), and shared its ancestry with pRL10. This type of plasmid was present in all members of genospecies C. On the other hand, all genospecies B members displayed the characteristics of two distinct plasmids like 3841. Based on these results, species members with different host specificity can share similar genomic strategies for adaptations that are important in the specific conditions of our sampling site. This reinforces an observation made many years ago, that both symbiovars had a common genetic background (Young, 1985).

6.2 Direction for future research

The results of the population genomics study have revealed true bacterial species in a closely related bacterial population. The construction of a genospecies based on core genes and ANI analysis as described in chapter 2 can be fruitful for the correct identification of bacterial species in different genera and was recently recommended by Chan et al. (2012). Inclusion of samples from different geographical locations can be useful to explore the global wide distribution of observed genospecies and identification of the rest of the genospecies in *R. leguminosarum*. Exploration of major donor and recipient genospecies using different computationally intensive tools such as ClonalOrigin could extend the work of chapter 3. A putative extension of the work of chapter 4 would be useful in determining the incongruent genes located on chromids and plasmids that highlight the evidence of HGT within and between genospecies. Moreover, phylogeny of one of the smallest plasmids (pRL7) reflected the signals of ecological species indicating the possibility of host specific genes on this plasmid. However, these signals were weak in the individual pRL7 gene trees and could be identified with high sequence information in another study. Further research could be completed with high sequencing information involving the identification of more population specific genes that are present in our dataset, but absent in the reference genome. The number of these genes will increase with additional genomic data and this could be useful to identify the distinct phenotypic traits of a local population. It would be interesting to explore the genetic properties such as *nod* genes in the other plasmids present in this population, and this would reflect the genomic plasticity in each member of the population.

Finally, in terms of bacterial fitness, the observed population could be analyzed to determine the population fitness, species competitiveness and beneficial role of homologous recombination. It has been inferred (Muller, 1932) that homologous recombination reduces the competition between beneficial mutations (clonal interference) by incorporating beneficial mutations together that will eventually speed up adaptation. Moreover, recombination facilitates natural selection by breaking the negative disequilibrium generated by epistasis (interaction between mutations) (Vos, 2009; Barton and Otto, 2005). Future work will be required to characterize the effects of homologous recombination in five genospecies of *R. leguminosarum*.

6.3 Conclusions

The main objective of this thesis was to explore the nature of bacterial species by performing a population genomics of symbiovars of a bacterium isolated from the same location. In summary, this work has demonstrated that species based on core gene sequences are completely different from major ecological clusters that have been diverged from each other a long time ago. Instead, it is the core genes that delineate the bacterial species and the species members are grouped together by the cohesive force of recombination. The cosmopolitan nature is another characteristic of bacterial species observed in this study. This study also demonstrated that core genes are not restricted to the chromosome or chromids, but can be located on plasmids. The structure of core genes phylogeny was reflected in the phylogenies of replicons indicates the occurrence of low inter- genospecies recombination in many replicons. Furthermore, the host specific symbiotic genes (*nod* genes) are one of the accessory genes that are maintained by inter- genospecies HGT and have an insignificant role in defining bacterial species.

Although we are in the sequencing era, the theoretical definition of prokaryotic species is still debatable (Cohan, 2002; Cohan and Perry, 2007; Fraser et al., 2007; Fraser et al., 2009; Shapiro et al., 2012). A universal species concept for bacteria is essential in the fields of medicine, veterinary medicine, agriculture, and industry and requires correct identification of species, for example, identification of pathogenic bacteria and their associated diseases. Since gene transfer is more likely within than between genospecies, it is our belief that the most appropriate model produced to date for microbial species is

the model (Fraser et al., 2009) that utilizes the combination of genomic clusters and ecological diversity. Hopefully, this study will make a significant contribution to the construction of a theoretical definition of bacterial species.

Appendix I

Table I.I | The 305 core genes held by all chromid-possessing bacteria (Harrison et al., 2010). Locus tags and other information are based on *Rlv* 3841. The 100 genes with best coverage have bold Locus tags and are used in Chapter 3.

Locus tag	Gene symbol	Location	Position
RL0003	aroE	Chromosome	1921..2778
RL0004	coaE	Chromosome	2778..3389
RL0012	gyrB	Chromosome	8285..10720
RL0021	trpB	Chromosome	17640..18860
RL0022	trpA	Chromosome	18864..19703
RL0024	folC	Chromosome	20778..22130
RL0025	-	Chromosome	22203..22523
RL0029	-	Chromosome	30083..31600
RL0042	hisF	Chromosome	44389..45177
RL0043	hisA	Chromosome	45178..45924
RL0046	hisH	Chromosome	46921..47571
RL0048	hisB	Chromosome	48082..48690
RL0106	rpsA	Chromosome	135131..136834
RL0108	aroA	Chromosome	137731..139089
RL0120	pnp	Chromosome	147985..150123

RL0123	truB	Chromosome	151863..152795
RL0125	infB	Chromosome	153364..156117
RL0127	nusA	Chromosome	156908..158509
RL0131A	recR	Chromosome	162809..163414
RL0134	dnaX	Chromosome	164470..166347
RL0139	-	Chromosome	169654..170508
RL0151	dnaJ	Chromosome	180070..181197
RL0152	dnaK	Chromosome	181283..183199
RL0160	polA	Chromosome	192994..196044
RL0161	-	Chromosome	196319..198664
RL0181	-	Chromosome	215180..216967
RL0254	lepA	Chromosome	290983..292815
RL0268	rplT	Chromosome	302898..303302
RL0269	pheS	Chromosome	303453..304535
RL0270	pheT	Chromosome	304552..306975
RL0282	xseA	Chromosome	317593..319173
RL0315	guaA	Chromosome	348345..349907
RL0326	-	Chromosome	360897..362054

RL0328	-	Chromosome	362992..363789
RL0334	dnaN	Chromosome	367988..369106
RL0335	-	Chromosome	369366..370283
RL0357	coaBC	Chromosome	390460..391665
RL0371	ubiE	Chromosome	406332..407108
RL0375	dnaA	Chromosome	410392..411840
RL0377	hemN	Chromosome	413129..414331
RL0378	-	Chromosome	414341..414985
RL0382	-	Chromosome	417455..418087
RL0388	trmB	Chromosome	421197..421898
RL0389	metK	Chromosome	421907..423145
RL0393	-	Chromosome	426720..427235
RL0394	phoH	Chromosome	427235..428284
RL0395	miaB	Chromosome	428303..429724
RL0404	mviN	Chromosome	436353..437933
RL0406	mutS	Chromosome	441335..444061
RL0421	-	Chromosome	457817..458617
RL0433	fmt	Chromosome	470957..471916

RL0445	argB	Chromosome	483090..483977
RL0504	pgi	Chromosome	541877..543559
RL0550	argF	Chromosome	594964..595878
RL0572	-	Chromosome	616769..617857
RL0611	murA	Chromosome	662438..663730
RL0613	hisD	Chromosome	664326..665624
RL0616	infA	Chromosome	666708..666926
RL0680	-	Chromosome	729689..732250
RL0743	-	Chromosome	791485..792384
RL0847	guaB	Chromosome	915252..916736
RL0877	hisS	Chromosome	944612..946183
RL0883	groEL	Chromosome	950963..952606
RL0884	groES	Chromosome	952679..952975
RL0886	ribF	Chromosome	954249..955232
RL0889	ileS	Chromosome	956316..959264
RL0891	-	Chromosome	960076..960531
RL0892	-	Chromosome	960604..962709
RL0910	mutL	Chromosome	979136..980938

RL0920	-	Chromosome	990927..992087
RL0930	rnhB	Chromosome	1000331..1001020
RL0937	ispB	Chromosome	1007149..1008165
RL0945	aroA	Chromosome	1013898..1015160
RL0947	purD	Chromosome	1017197..1018474
RL0956	ubiA	Chromosome	1029958..1030917
RL0960	-	Chromosome	1036681..1037655
RL0969	rumA	Chromosome	1046331..1047581
RL0973	dxs	Chromosome	1051866..1053782
RL1007	aroC	Chromosome	1088479..1089576
RL1014	pdxH	Chromosome	1097268..1097888
RL1030	ispH	Chromosome	1111897..1112898
RL1078	mutY	Chromosome	1158581..1159684
RL1262	-	Chromosome	1328143..1328901
RL1370	msrB	Chromosome	1434172..1434585
RL1412	groEL	Chromosome	1470011..1471645
RL1503	smpB	Chromosome	1573100..1573555
RL1510	sipS	Chromosome	1579124..1579867

RL1543	cysS	Chromosome	1609853..1611244
RL1546	purF	Chromosome	1613195..1614685
RL1548	radA	Chromosome	1615550..1616953
RL1550	-	Chromosome	1617758..1618927
RL1551	dnaC	Chromosome	1619016..1620551
RL1552	rplI	Chromosome	1620888..1621466
RL1554	rpsR	Chromosome	1622632..1622880
RL1558	fabG	Chromosome	1625930..1626667
RL1564	ksgA	Chromosome	1631562..1632389
RL1580	ndk	Chromosome	1650820..1651242
RL1595	purN	Chromosome	1664648..1665319
RL1596	purM	Chromosome	1665316..1666389
RL1605	aspS	Chromosome	1678144..1679934
RL1616	hemB	Chromosome	1689408..1690421
RL1620	glyA	Chromosome	1693675..1694973
RL1621	ribD	Chromosome	1695463..1696809
RL1632	ribH	Chromosome	1706382..1706837
RL1668	argC	Chromosome	1747476..1748408

RL1672	rpsI	Chromosome	1751149..1751616
RL1673	rplM	Chromosome	1751618..1752082
RL1688	clpP	Chromosome	1768136..1768765
RL1723	dnaE	Chromosome	1807797..1811294
RL1735	topA	Chromosome	1822923..1825577
RL1736	smf	Chromosome	1825826..1826968
RL1737	-	Chromosome	1826968..1827588
RL1739	pyrB	Chromosome	1828951..1829907
RL1760	nusG	Chromosome	1859388..1859918
RL1761	rplK	Chromosome	1860096..1860527
RL1762	rplA	Chromosome	1860532..1861233
RL1764	rplJ	Chromosome	1861563..1862081
RL1765	rplL	Chromosome	1862140..1862517
RL1767	rpoC	Chromosome	1867111..1871319
RL1770	rpsG	Chromosome	1872701..1873171
RL1771	fus	Chromosome	1873201..1875300
RL1774	rplC	Chromosome	1877047..1877688
RL1775	rplD	Chromosome	1877702..1878322

RL1776	rplW	Chromosome	1878319..1878612
RL1777	rplB	Chromosome	1878624..1879460
RL1778	rpsS	Chromosome	1879476..1879754
RL1779	rplV	Chromosome	1879757..1880146
RL1780	rpsC	Chromosome	1880146..1880877
RL1781	rplP	Chromosome	1880914..1881327
RL1783	rpsQ	Chromosome	1881552..1881791
RL1784	rplN	Chromosome	1882050..1882418
RL1785	rplX	Chromosome	1882430..1882738
RL1786	rplE	Chromosome	1882731..1883288
RL1788	rpsH	Chromosome	1883639..1884037
RL1789	rplF	Chromosome	1884080..1884613
RL1790	rplR	Chromosome	1884626..1884988
RL1791	rpsE	Chromosome	1885120..1885689
RL1793	rplO	Chromosome	1885930..1886406
RL1794	secY	Chromosome	1886641..1887981
RL1795	adk	Chromosome	1887978..1888628
RL1797	rpsK	Chromosome	1889353..1889742

RL1798	rpoA	Chromosome	1889840..1890850
RL1799	rplQ	Chromosome	1890974..1891396
RL1803	ilvD	Chromosome	1893842..1895680
RL2035	valS	Chromosome	2142990..2145833
RL2041	argS	Chromosome	2151831..2153588
RL2043	nagZ	Chromosome	2157013..2158026
RL2048	tatC	Chromosome	2160649..2161476
RL2049	serS	Chromosome	2161677..2162960
RL2050	surE	Chromosome	2162965..2163738
RL2055	secD	Chromosome	2167972..2170512
RL2069	map	Chromosome	2186377..2187213
RL2099	recJ	Chromosome	2210649..2212409
RL2221	rpsB	Chromosome	2340782..2341549
RL2222	tsf	Chromosome	2341791..2342717
RL2223	pyrH	Chromosome	2342814..2343536
RL2224	frr	Chromosome	2343701..2344261
RL2225	uppS	Chromosome	2344299..2345042
RL2227	ecfE	Chromosome	2345901..2347034

RL2238	kdsA	Chromosome	2359852..2360697
RL2239	eno	Chromosome	2360835..2362109
RL2249	-	Chromosome	2369909..2370880
RL2254	ispDF	Chromosome	2373263..2374480
RL2255	dus	Chromosome	2374626..2375642
RL2288	cysG2	Chromosome	2410570..2412012
RL2381	glmU	Chromosome	2503587..2504948
RL2382	glmS	Chromosome	2505058..2506884
RL2384	recG	Chromosome	2507724..2509829
RL2386	mfd	Chromosome	2510224..2513724
RL2392	glnA	Chromosome	2518499..2519908
RL2393	glnB	Chromosome	2519987..2520325
RL2398	uvrA	Chromosome	2524791..2527712
RL2399	ssb	Chromosome	2527975..2528484
RL2401	gyrA	Chromosome	2529493..2532330
RL2403	coaD	Chromosome	2533660..2534154
RL2406	queA	Chromosome	2535384..2536469
RL2407	tgt	Chromosome	2536466..2537596

RL2442	ilvI	Chromosome	2573933..2575639
RL2472	-	Chromosome	2607249..2608031
RL2473	metG	Chromosome	2608036..2609586
RL2476	tmk	Chromosome	2611683..2612363
RL2493	trpD	Chromosome	2628498..2629514
RL2494	trpC	Chromosome	2629524..2630336
RL2511	pyrG	Chromosome	2648019..2649647
RL2528	thrS	Chromosome	2666110..2668095
RL2532	hisI	Chromosome	2670560..2671012
RL2555	lipB	Chromosome	2691150..2691836
RL2588	tyrS	Chromosome	2723170..2724426
RL2598	rpe	Chromosome	2735612..2736289
RL2612	purL	Chromosome	2749632..2751866
RL2624	rpsD	Chromosome	2762956..2763573
RL2627	murI	Chromosome	2766794..2767597
RL2636	alaS	Chromosome	2775464..2778118
RL2637	recA	Chromosome	2778272..2779366
RL2648	-	Chromosome	2791538..2792056

RL2650	folC	Chromosome	2792411..2793280
RL2691	-	Chromosome	2846841..2848568
RL2798	leuS	Chromosome	2959236..2961752
RL2801	ddl	Chromosome	2963872..2964948
RL2824	cobA	Chromosome	2983476..2984315
RL2957	uvrB	Chromosome	3122747..3125776
RL2987	argG	Chromosome	3152923..3154134
RL2990	ubiA	Chromosome	3155285..3156199
RL3013	tyrS	Chromosome	3175005..3176270
RL3071	ftsZ	Chromosome	3233697..3234716
RL3170	-	Chromosome	3326081..3326611
RL3205	ilvC	Chromosome	3356982..3358001
RL3244	ilvH	Chromosome	3391354..3391926
RL3245	ilvI	Chromosome	3391949..3393739
RL3249	miaA	Chromosome	3399304..3400197
RL3276	pcrA	Chromosome	3433255..3435744
RL3293	ligA	Chromosome	3449396..3451552
RL3295	recN	Chromosome	3452287..3453960

RL3298	ftsZ	Chromosome	3456313..3458031
RL3301	ddl	Chromosome	3460336..3461262
RL3306	murC	Chromosome	3465824..3467239
RL3307	murG	Chromosome	3467236..3468360
RL3309	murD	Chromosome	3469523..3470935
RL3310	mraY	Chromosome	3470943..3472043
RL3311	murF	Chromosome	3472066..3473499
RL3312	murE	Chromosome	3473496..3474965
RL3313	-	Chromosome	3475023..3476783
RL3315	mraW	Chromosome	3477309..3478334
RL3402	rpoD	Chromosome	3568480..3570537
RL3408	dnaG	Chromosome	3576509..3578518
RL3411	carA	Chromosome	3580808..3582013
RL3419	carB	Chromosome	3589727..3593215
RL3460	proC	Chromosome	3632148..3632975
RL3465	-	Chromosome	3636408..3637202
RL3468	prs	Chromosome	3639378..3640310
RL3471	-	Chromosome	3644565..3645173

RL3474	pth	Chromosome	3648639..3649370
RL3479	ychF	Chromosome	3651075..3652190
RL3521	trpE	Chromosome	3693219..3695408
RL3553	engA	Chromosome	3733251..3734672
RL3765	rLuD	Chromosome	3965270..3966295
RL3768	purA	Chromosome	3970117..3971415
RL3957	mnmA	Chromosome	4180678..4181904
RL3965	ftsH	Chromosome	4191430..4193361
RL3983	-	Chromosome	4209142..4209888
RL3986	ruvC	Chromosome	4213671..4214180
RL3989	ruvA	Chromosome	4214909..4215523
RL3990	ruvB	Chromosome	4215607..4216647
RL4006	cbbT	Chromosome	4232574..4234547
RL4007	gap	Chromosome	4234629..4235639
RL4017	rpmE	Chromosome	4245616..4245837
RL4044	purE	Chromosome	4273781..4274275
RL4060	pykA	Chromosome	4288486..4289925
RL4085	gltA	Chromosome	4316437..4321158

RL4184	gltX	Chromosome	4436061..4437527
RL4203	talB	Chromosome	4454319..4455290
RL4207	-	Chromosome	4458217..4459023
RL4265	msrB	Chromosome	4523218..4523703
RL4279	clpB	Chromosome	4549225..4551825
RL4281	hemK	Chromosome	4553188..4554048
RL4282	prfA	Chromosome	4554045..4555124
RL4298	secA	Chromosome	4567724..4570441
RL4323	argH	Chromosome	4592706..4594070
RL4325	lysA	Chromosome	4594453..4595721
RL4352	aroB	Chromosome	4621118..4622248
RL4353	aroK	Chromosome	4622245..4622802
RL4412	priA	Chromosome	4690057..4692273
RL4436	sucD	Chromosome	4718517..4719419
RL4438	sucC	Chromosome	4719933..4721126
RL4493	gpsA	Chromosome	4775166..4776149
RL4494	gcp	Chromosome	4776146..4777243
RL4495	hemC	Chromosome	4777316..4778245

RL4506	typA	Chromosome	4786672..4788492
RL4507	dcp	Chromosome	4788651..4790738
RL4515	argG	Chromosome	4797133..4798356
RL4522	-	Chromosome	4803714..4804940
RL4550	rimM	Chromosome	4830168..4830749
RL4551	trmD	Chromosome	4830746..4831480
RL4552	rplS	Chromosome	4831767..4832306
RL4555	-	Chromosome	4834364..4835773
RL4563	-	Chromosome	4841412..4844102
RL4565	glnB	Chromosome	4845808..4846146
RL4630	ispG	Chromosome	4921382..4922632
RL4677	rpmA	Chromosome	4981224..4981493
RL4681	obgE	Chromosome	4983849..4984943
RL4682	proB	Chromosome	4984897..4986099
RL4683	proA	Chromosome	4986092..4987375
RL4689	-	Chromosome	4991923..4992372
RL4692	ctpA	Chromosome	4994535..4995857
RL4705	leuD	Chromosome	5007551..5008159

RL4707	leuB	Chromosome	5009319..5010431
RL4722	purH	Chromosome	5030574..5032190
RL4727	acs	Chromosome	5037147..5039102
RL4731	-	Chromosome	5043305..5043964
RL4732	leuS	Chromosome	5044165..5046795
RL4735	parB	Chromosome	5048439..5049320
RL4736	parA	Chromosome	5049353..5050147
RL4738	gidA	Chromosome	5050809..5052686
RL4739	trmE	Chromosome	5052759..5054078
pRL120279	prC	pRL12	294782..296896
pRL120359	panC	pRL12	388343..389254
pRL120360	panB	pRL12	389251..390072
pRL120416	dadX	pRL12	450216..451349
pRL120642	groEL	pRL12	696038..697666
pRL120643	groS	pRL12	697820..698134
pRL110033	-	pRL11	40088..42031
pRL110442	thiE	pRL11	475772..476407
pRL100453	-	pRL10	467953..469989

pRL90212	-	pRL9	231444..231953
pRL80044	acsA	pRL8	48652..50592

Appendix II

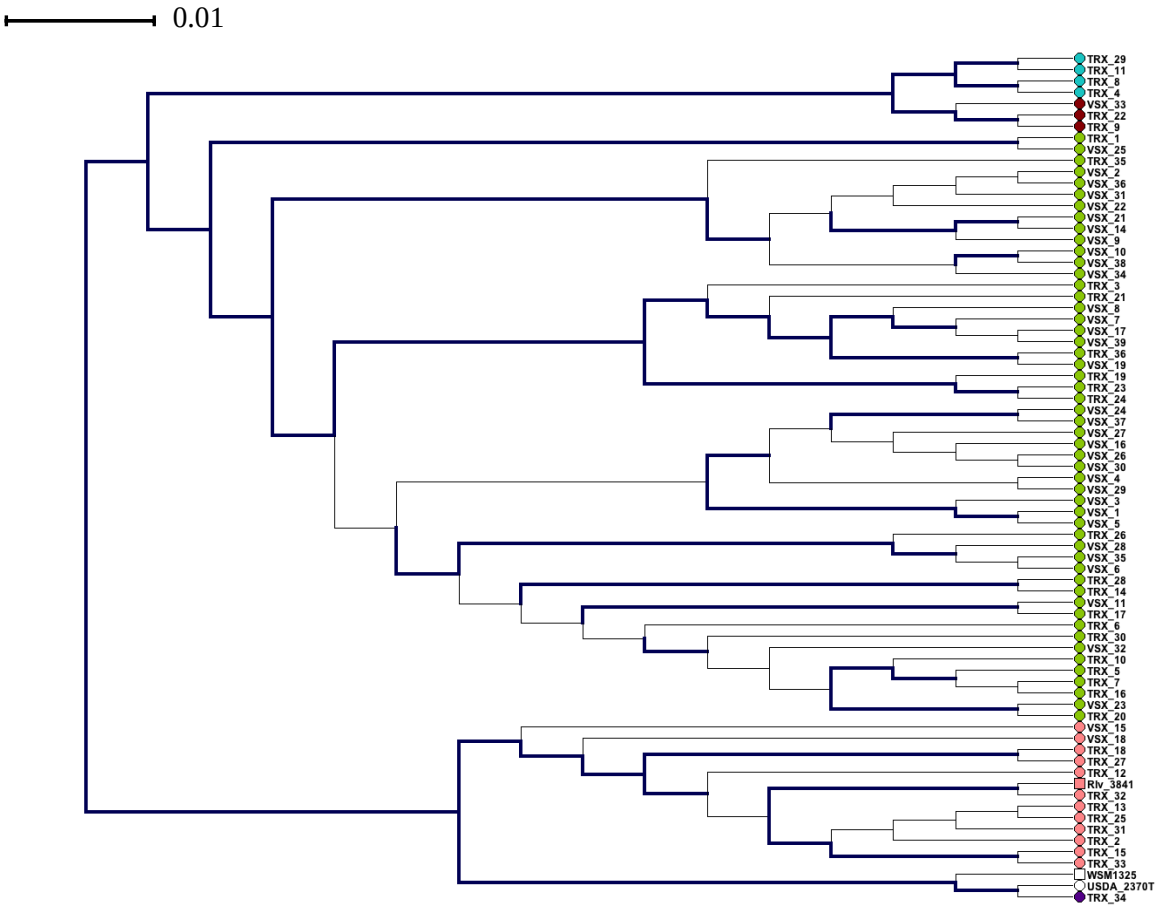


Figure II.I | Maximum Likelihood tree based on 100-gene alignment showing the position of 75 *R. leguminosarum* strains. Strain nodes are colored on the basis of their genospecies (A: purple, B: salmon, C: green, D: cyan and E: dark red). WSM1325 (white square) represents *Rlt* WSM1325. USDA_2370^{T} is shown by white circle. *Rlv* 3841 is shown by red square. Branches with bootstrap values > 70% are colored.

Appendix III

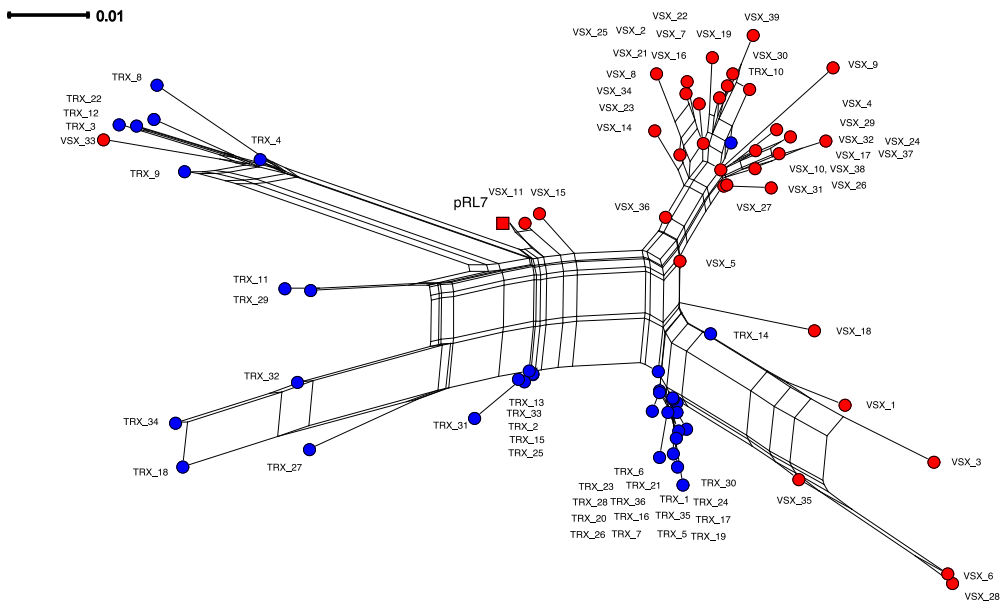


Figure III.I | Phylogenetic network obtained from a small plasmid (pRL7) of *Rlv* 3841 for 72 *R. leguminosarum* strains. Strain nodes are circled according to their symbiovars (red: *viciae* and blue *trifolii*). *Rlv* 3841 is shown by the salmon square.

Table III.I | The genospecies B specific-island in Chromid (pRL12). Locus tags and other informations of 46 genes of pRL12 held by all 12 members of genospecies B.

Locus tag	Strand	Function
pRL120118	+	putative aldo-keto reductase/oxidoreductase
pRL120119	-	putative short-chain dehydrogenase
pRL120120	-	putative short-chain dehydrogenase
pRL120121	-	putative dihydroorotase
pRL120122	-	hypothetical protein
pRL120123	-	putative polysaccharide deacetylase
pRL120124	-	putative NAD-dependent epimerase/dehydratase
pRL120125	-	putative short-chain dehydrogenase
pRL120126	-	putative D-hydantoinase
pRL120127	-	MFS family transporter
pRL120128	-	putative ATP-binding component of ABC transporter
pRL120129	-	putative permease component of ABC transporter
pRL120130	-	putative permease component of ABC transporter
pRL120131	-	putative substrate binding component of ABC transporter
pRL120132	-	AraC family transcriptional regulator
pRL120133	-	putative plasmid stability protein

pRL120134	-	putative plasmid stability protein
pRL120135	-	putative cyclase
pRL120136	-	hypothetical protein
pRL120137	-	hypothetical protein
pRL120138	-	cobW family cobalimin synthesis protein
pRL120139	-	hypothetical protein
pRL120140	-	putative imidase
pRL120141	-	putative ATP-binding component of ABC transporter
pRL120142	-	putative substrate-binding component of ABC transporter
pRL120143	-	putative short-chain dehydrogenase/reductase
pRL120144	-	putative 3-oxoacyl-[acyl-carrier-protein] reductase (3-ketoacyl-acyl carrier protein reductase)
pRL120145	-	putative permease component of ABC transporter
pRL120146	-	putative permease component of ABC transporter
pRL120147	-	putative ATP-binding component of ABC transporter
pRL120148	+	hypothetical protein
pRL120149	+	GntR family transcriptional regulator
pRL120150	-	putative urea amidolyase related protein
pRL120151	-	putative urea amidohydrolyase homologue

pRL120152	-	acetyl-CoA carboxylase biotin carboxylase subunit
pRL120154	-	hypothetical protein
pRL120155	+	LysR family transcriptional regulator
pRL120157	-	hypothetical protein
pRL120158	-	hypothetical protein
pRL120159	-	allantoate amidohydrolase
pRL120160	-	DeoR family transcriptional regulator
pRL120161	-	putative substrate-binding component of ABC transporter
pRL120162	-	putative permease component of ABC transporter
pRL120163	-	putative permease component of ABC transporter
pRL120164	-	putative component of ABC transporter

Table III.II | The first genospecies B specific-island in large plasmid (pRL9). Locus tags and other informations of 9 genes held by all 12 members of genospecies B.

Locus tag	Strand	Function
pRL90119	-	LysR family transcriptional regulator
pRL90120	+	putative 4-carboxymuconolactone decarboxylase
pRL90121	-	hypothetical protein
pRL90122	-	putative LacI/HTH-type transcriptional regulator
pRL90123	+	putative lactose transport ATP-binding protein
pRL90124	+	putative transmembrane binding-protein-dependent transporter
pRL90125	+	putative permease transporter component
pRL90126	+	putative solute binding-protein component of transporter
pRL90127	+	putative glycosyl hydrolase

Table III.III | The second genospecies B specific-island in pRL9. Locus tags and other informations of 7 genes held by all 12 members of genospecies B.

Locus tag	Strand	Function
pRL90035	-	putative ATP_binding protein of ABC transporter,pseudogene
pRL90036	-	putative attachment-related protein
pRL90039	+	hypothetical protein
pRL90041	-	chaperonin GroEL
pRL90043	+	putative transmembrane transport protein
pRL90044	+	putative transmembrane ABC transporter
pRL90045	+	putative ABC transporter permease component

Table III.IV | The third genospecies B specific-island in pRL9. Locus tags and other informations of 8 genes held by all 12 members of genospecies B.

Locus tag	Strand	Function
pRL90088	+	hypothetical protein
pRL90089	+	hypothetical protein
pRL90090	+	putative ATP-binding component of ABC transporter
pRL90091	+	putative permease component of ABC transporter
pRL90092	-	hypothetical protein
pRL90093	+	hypothetical protein
pRL90094	+	hypothetical protein
pRL90095	-	putative plasmid stability protein

Table III.V | The fourth genospecies B specific-island in pRL9. Locus tags and other informations of 8 genes held by all 12 members of genospecies B.

Locus tag	Strand	Function
pRL90255	-	putative glycine cleavage protein/aminomethyltransferase
pRL90256	-	putative 5,10-methylenetetrahydrofolate reductase
pRL90257	+	GntR family transcriptional regulator
pRL90258	+	putative substrate-binding ABC transporter protein
pRL90259	+	putative permease component of ABC transporter
pRL90260	+	putative permease component of ABC transporter
pRL90261	+	putative ATP-binding ABC transporter
pRL90262	+	putative ATP-binding component of ABC transporter

References

ACHTMAN, M. & WAGNER, M. 2008. Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol,* 6**,** 431-40.

ACOSTA, J. L., EGUIARTE, L. E., SANTAMARIA, R. I., BUSTOS, P., VINUESA, P., MARTINEZ-ROMERO, E., DAVILA, G. & GONZALEZ, V. 2011. Genomic lineages of *Rhizobium etli* revealed by the extent of nucleotide polymorphisms and low recombination. *BMC Evol Biol,* 11**,** 305.

ALIKHAN, N. F., PETTY, N. K., BEN ZAKOUR, N. L. & BEATSON, S. A. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics,* 12**,** 402.

ALVAREZ-MARTINEZ, E. R., VALVERDE, A., RAMIREZ-BAHENA, M. H., GARCIA-FRAILE, P., TEJEDOR, C., MATEOS, P. F., SANTILLANA, N., ZUNIGA, D., PEIX, A. & VELAZQUEZ, E. 2009. The analysis of core and symbiotic genes of rhizobia nodulating *Vicia* from different continents reveals their common phylogenetic origin and suggests the distribution of *Rhizobium leguminosarum* strains together with *Vicia* seeds. *Arch Microbiol,* 191**,** 659-68.

ATWOOD, K. C., SCHNEIDER, L. K. & RYAN, F. J. 1951. Periodic selection in *Escherichia coli*. *Proc Natl Acad Sci U S A,* 37**,** 146-55.

AUCH, A. F., VON JAN, M., KLENK, H. P. & GOKER, M. 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci,* 2**,** 117-34.

AZIZ, R. K., BARTELS, D., BEST, A. A., DEJONGH, M., DISZ, T., EDWARDS, R. A., FORMSMA, K., GERDES, S., GLASS, E. M., KUBAL, M., MEYER, F., OLSEN, G. J., OLSON, R., OSTERMAN, A. L., OVERBEEK, R. A., MCNEIL, L. K., PAARMANN, D., PACZIAN, T., PARRELLO, B., PUSCH, G. D., REICH, C., STEVENS, R., VASSIEVA, O., VONSTEIN, V., WILKE, A. & ZAGNITKO, O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics,* 9**,** 75.

BACKERT, S., KWOK, T. & KONIG, W. 2005. Conjugative plasmid DNA transfer in *Helicobacter pylori* mediated by chromosomally encoded relaxase and TraG-like proteins. *Microbiology,* 151**,** 3493-503.

BAILLY, X., GIUNTINI, E., SEXTON, M. C., LOWER, R. P., HARRISON, P. W., KUMAR, N. & YOUNG, J. P. 2011. Population genomics of *Sinorhizobium medicae* based on low-coverage sequencing of sympatric isolates. *ISME J,* 5**,** 1722-34.

BAILLY, X., OLIVIERI, I., BRUNEL, B., CLEYET-MAREL, J. C. & BENA, G. 2007. Horizontal gene transfer and homologous recombination drive the evolution of the nitrogen-fixing symbionts of *Medicago* species. *J Bacteriol,* 189**,** 5223-36.

BARTON, N. H. & OTTO, S. P. 2005. Evolution of Recombination Due to Random Drift. *Genetics,* 169**,** 2353-2370.

BEIKO, R. G., HARLOW, T. J. & RAGAN, M. A. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A,* 102**,** 14332-7.

BENNETT, G. M. & MORAN, N. A. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biol Evol,* 5**,** 1675-88.

BENNETT, J. S., JOLLEY, K. A., SPARLING, P. F., SAUNDERS, N. J., HART, C. A., FEAVERS, I. M. & MAIDEN, M. C. 2007. Species status of *Neisseria gonorrhoeae*: evolutionary and epidemiological inferences from multilocus sequence typing. *BMC Biol,* 5**,** 35.

BENTLEY, S. D. & PARKHILL, J. 2004. Comparative genomic structure of prokaryotes. *Annu Rev Genet,* 38**,** 771-92.

BISHOP, C. J., AANENSEN, D. M., JORDAN, G. E., KILIAN, M., HANAGE, W. P. & SPRATT, B. G. 2009. Assigning strains to bacterial species via the internet. *BMC Biol,* 7**,** 3.

BLOEMBERG, G. V., KAMST, E., HARTEVELD, M., VAN DER DRIFT, K. M., HAVERKAMP, J., THOMAS-OATES, J. E., LUGTENBERG, B. J. & SPAINK, H. P. 1995. A central domain of *Rhizobium* NodE protein mediates host specificity by determining the hydrophobicity of fatty acyl moieties of nodulation factors. *Mol Microbiol,* 16**,** 1123-36.

BOTO, L. 2010. Horizontal gene transfer in evolution: facts and challenges. *Proc Biol Sci,* 277**,** 819-27.

BOUCHER, Y., DOUADY, C. J., SHARMA, A. K., KAMEKURA, M. & DOOLITTLE, W. F. 2004. Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J Bacteriol,* 186**,** 3980-90.

CADILLO-QUIROZ, H., DIDELOT, X., HELD, N. L., HERRERA, A., DARLING, A., RENO, M. L., KRAUSE, D. J. & WHITAKER, R. J. 2012. Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol,* 10**,** e1001265.

CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., RAJANDREAM, M. A., BARRELL, B. G. & PARKHILL, J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics,* 21**,** 3422-3.

CEVALLOS, M. A., CERVANTES-RIVERA, R. & GUTIERREZ-RIOS, R. M. 2008. The repABC plasmid family. *Plasmid,* 60**,** 19-37.

CHAN, J. Z., HALACHEV, M. R., LOMAN, N. J., CONSTANTINIDOU, C. & PALLEN, M. J. 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol,* 12**,** 302.

CHANG, Y. J., LAND, M., HAUSER, L., CHERTKOV, O., DEL RIO, T. G., NOLAN, M., COPELAND, A., TICE, H., CHENG, J. F., LUCAS, S., HAN, C., GOODWIN, L., PITLUCK, S., IVANOVA, N., OVCHINIKOVA, G., PATI, A., CHEN, A., PALANIAPPAN, K., MAVROMATIS, K., LIOLIOS, K., BRETTIN, T., FIEBIG, A.,

ROHDE, M., ABT, B., GOKER, M., DETTER, J. C., WOYKE, T., BRISTOW, J., EISEN, J. A., MARKOWITZ, V., HUGENHOLTZ, P., KYRPIDES, N. C., KLENK, H. P. & LAPIDUS, A. 2011a. Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium *Ktedonobacter racemifer* type strain (SOSP1-21). *Stand Genomic Sci,* 5**,** 97-111.

CHANG, Y. L., WANG, E. T., SUI, X. H., ZHANG, X. X. & CHEN, W. X. 2011b. Molecular diversity and phylogeny of rhizobia associated with *Lablab purpureus* (Linn.) grown in Southern China. *Syst Appl Microbiol,* 34**,** 276-84.

CHEN, I. & DUBNAU, D. 2004. DNA uptake during bacterial transformation. *Nat Rev Microbiol,* 2**,** 241-9.

CICCARELLI, F. D., DOERKS, T., VON MERING, C., CREEVEY, C. J., SNEL, B. & BORK, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science,* 311**,** 1283-7.

CLARRIDGE, J. E., 3RD 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev,* 17**,** 840-62, table of contents.

COHAN, F. 2005. Periodic Selection and Ecological Diversity in Bacteria. *In:* NURMINSKY, D. (ed.) *Selective Sweep.* Springer US.

COHAN, F. M. 2001. Bacterial species and speciation. *Syst Biol,* 50**,** 513-24.

COHAN, F. M. 2002. What are bacterial species? *Annu Rev Microbiol,* 56**,** 457-87.

COHAN, F. M. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci,* 361**,** 1985-96.

COHAN, F. M. & PERRY, E. B. 2007. A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol,* 17**,** R373-86.

COLWELL, R. R. 1970. Polyphasic taxonomy of the genus *vibrio*: numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. *J Bacteriol,* 104**,** 410-33.

CONNOR, N., SIKORSKI, J., ROONEY, A. P., KOPAC, S., KOEPPEL, A. F., BURGER, A., COLE, S. G., PERRY, E. B., KRIZANC, D., FIELD, N. C., SLATON, M. & COHAN, F. M. 2010. Ecology of speciation in the genus *Bacillus*. *Appl Environ Microbiol,* 76**,** 1349-58.

CROSSMAN, L. C., CASTILLO-RAMIREZ, S., MCANNULA, C., LOZANO, L., VERNIKOS, G. S., ACOSTA, J. L., GHAZOUI, Z. F., HERNANDEZ-GONZALEZ, I., MEAKIN, G., WALKER, A. W., HYNES, M. F., YOUNG, J. P., DOWNIE, J. A., ROMERO, D., JOHNSTON, A. W., DAVILA, G., PARKHILL, J. & GONZALEZ, V. 2008. A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria. *PLoS One,* 3**,** e2567.

DAGAN, T. 2011. Phylogenomic networks. *Trends Microbiol,* 19**,** 483-91.

DARLING, A. C., MAU, B., BLATTNER, F. R. & PERNA, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res,* 14**,** 1394-403.

DAVIS, E. O., EVANS, I. J. & JOHNSTON, A. W. 1988. Identification of *nodX*, a gene that allows *Rhizobium leguminosarum* biovar *viciae* strain TOM to nodulate Afghanistan peas. *Mol Gen Genet,* 212**,** 531-5.

DEBEVEC, A. H. & WHITFIELD, J. B. 2013. Introduction to Phylogenetic Networks.—David A. Morrison. *Systematic Biology,* 62**,** 177-178.

DELCHER, A. L., PHILLIPPY, A., CARLTON, J. & SALZBERG, S. L. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res,* 30**,** 2478-83.

DIDELOT, X., BOWDEN, R., STREET, T., GOLUBCHIK, T., SPENCER, C., MCVEAN, G., SANGAL, V., ANJUM, M. F., ACHTMAN, M., FALUSH, D. & DONNELLY, P. 2011. Recombination and population structure in *Salmonella enterica*. *PLoS Genet,* 7**,** e1002191.

DIDELOT, X., BOWDEN, R., WILSON, D. J., PETO, T. E. & CROOK, D. W. 2012a. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet,* 13**,** 601-12.

DIDELOT, X. & FALUSH, D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics,* 175**,** 1251-66.

DIDELOT, X., LAWSON, D., DARLING, A. & FALUSH, D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics,* 186**,** 1435-49.

DIDELOT, X., MERIC, G., FALUSH, D. & DARLING, A. E. 2012b. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics,* 13**,** 256.

DING, H., YIP, C. B. & HYNES, M. F. 2013. Genetic characterization of a novel rhizobial plasmid conjugation system in *Rhizobium leguminosarum* bv. *viciae* strain VF39SM. *J Bacteriol,* 195**,** 328-39.

DOBRINDT, U., HOCHHUT, B., HENTSCHEL, U. & HACKER, J. 2004. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol,* 2**,** 414-24.

DONATI, C., HILLER, N. L., TETTELIN, H., MUZZI, A., CROUCHER, N. J., ANGIUOLI, S. V., OGGIONI, M., DUNNING HOTOPP, J. C., HU, F. Z., RILEY, D. R., COVACCI, A., MITCHELL, T. J., BENTLEY, S. D., KILIAN, M., EHRLICH, G. D., RAPPUOLI, R., MOXON, E. R. & MASIGNANI, V. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol,* 11**,** R107.

DOROGHAZI, J. R. & BUCKLEY, D. H. 2010. Widespread homologous recombination within and between *Streptomyces* species. *ISME J,* 4**,** 1136-43.

DYKHUIZEN, D. E. & GREEN, L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol,* 173**,** 7257-68.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EDWARDS, D. J. & HOLT, K. E. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp,* 3**,** 2.

EISEN, J. A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev,* 10**,** 606-11.

ELLEGAARD, K. M., KLASSON, L., NASLUND, K., BOURTZIS, K. & ANDERSSON, S. G. 2013. Comparative genomics of *Wolbachia* and the bacterial species concept. *PLoS Genet,* 9**,** e1003381.

EVANNO, G., REGNAUT, S. & GOUDET, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol,* 14**,** 2611-20.

EWING, W. H. 1949. *Shigella* Nomenclature. *J Bacteriol,* 57**,** 633-8.

FALUSH, D., WIRTH, T., LINZ, B., PRITCHARD, J. K., STEPHENS, M., KIDD, M., BLASER, M. J., GRAHAM, D. Y., VACHER, S., PEREZ-PEREZ, G. I., YAMAOKA, Y., MEGRAUD, F., OTTO, K., REICHARD, U., KATZOWITSCH, E., WANG, X., ACHTMAN, M. & SUERBAUM, S. 2003. Traces of human migrations in *Helicobacter pylori* populations. *Science,* 299**,** 1582-5.

FINLAY, B. J. 2002. Global dispersal of free-living microbial eukaryote species. *Science,* 296**,** 1061-3.

FIRMIN, J. L., WILSON, K. E., CARLSON, R. W., DAVIES, A. E. & DOWNIE, J. A. 1993. Resistance to nodulation of cv. Afghanistan peas is overcome by *nodX*, which mediates an O-acetylation of the *Rhizobium leguminosarum* lipo-oligosaccharide nodulation factor. *Mol Microbiol,* 10**,** 351-60.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A., MERRICK, J. M. & ET AL. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* 269**,** 496-512.

FRASER, C., ALM, E. J., POLZ, M. F., SPRATT, B. G. & HANAGE, W. P. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. *Science,* 323**,** 741-6.

FRASER, C., HANAGE, W. P. & SPRATT, B. G. 2007. Recombination and the nature of bacterial speciation. *Science,* 315**,** 476-80.

FROST, L. S., LEPLAE, R., SUMMERS, A. O. & TOUSSAINT, A. 2005. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol,* 3**,** 722-32.

FUKUSHIMA, M., KAKINUMA, K. & KAWAGUCHI, R. 2002. Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. *J Clin Microbiol,* 40**,** 2779-85.

FUNK, D. J., HELBLING, L., WERNEGREEN, J. J. & MORAN, N. A. 2000. Intraspecific phylogenetic congruence among multiple symbiont genomes. *Proc Biol Sci,* 267**,** 2517-21.

FURUYA, E. Y. & LOWY, F. D. 2006. Antimicrobial-resistant bacteria in the community setting. *Nat Rev Microbiol,* 4**,** 36-45.

GALARDINI, M., BIONDI, E. G., BAZZICALUPO, M. & MENGONI, A. 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med,* 6**,** 11.

GELMAN, A. & RUBIN, D. B. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science,* 7**,** 457-472.

GEVERS, D., COHAN, F. M., LAWRENCE, J. G., SPRATT, B. G., COENYE, T., FEIL, E. J., STACKEBRANDT, E., VAN DE PEER, Y., VANDAMME, P., THOMPSON, F. L. & SWINGS, J. 2005. Opinion: Re-evaluating prokaryotic species. *Nat Rev Microbiol,* 3**,** 733-9.

GIRAUD, E., MOULIN, L., VALLENET, D., BARBE, V., CYTRYN, E., AVARRE, J. C., JAUBERT, M., SIMON, D., CARTIEAUX, F., PRIN, Y., BENA, G., HANNIBAL, L., FARDOUX, J., KOJADINOVIC, M., VUILLET, L., LAJUS, A., CRUVEILLER, S., ROUY, Z., MANGENOT, S., SEGURENS, B., DOSSAT, C., FRANCK, W. L., CHANG, W. S., SAUNDERS, E., BRUCE, D., RICHARDSON, P., NORMAND, P., DREYFUS, B., PIGNOL, D., STACEY, G., EMERICH, D., VERMEGLIO, A., MEDIGUE, C. & SADOWSKY, M. 2007. Legumes symbioses: absence of *Nod* genes in photosynthetic *bradyrhizobia*. *Science,* 316**,** 1307-12.

GOECKS, J., NEKRUTENKO, A., TAYLOR, J. & GALAXY, T. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol,* 11**,** R86.

GOGARTEN, J. P., DOOLITTLE, W. F. & LAWRENCE, J. G. 2002. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol,* 19**,** 2226-38.

GOLDENFELD, N. & WOESE, C. 2007. Biology's next revolution. *Nature,* 445**,** 369.

GOMEZ-VALERO, L., RUSNIOK, C., JARRAUD, S., VACHERIE, B., ROUY, Z., BARBE, V., MEDIGUE, C., ETIENNE, J. & BUCHRIESER, C. 2011. Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics,* 12**,** 536.

GONZALEZ, V., ACOSTA, J. L., SANTAMARIA, R. I., BUSTOS, P., FERNANDEZ, J. L., HERNANDEZ GONZALEZ, I. L., DIAZ, R., FLORES, M., PALACIOS, R., MORA, J. & DAVILA, G. 2010. Conserved symbiotic plasmid DNA sequences in the multireplicon pangenomic structure of *Rhizobium etli*. *Appl Environ Microbiol,* 76**,** 1604-14.

GONZALEZ, V., SANTAMARIA, R. I., BUSTOS, P., HERNANDEZ-GONZALEZ, I., MEDRANO-SOTO, A., MORENO-HAGELSIEB, G., JANGA, S. C., RAMIREZ, M. A., JIMENEZ-JACINTO, V., COLLADO-VIDES, J. & DAVILA, G. 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci U S A,* 103**,** 3834-9.

GORIS, J., KONSTANTINIDIS, K. T., KLAPPENBACH, J. A., COENYE, T., VANDAMME, P. & TIEDJE, J. M. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol,* 57**,** 81-91.

GRIMONT, P. A. 1988. Use of DNA reassociation in bacterial classification. *Can J Microbiol,* 34**,** 541-6.

GUINDON, S., DUFAYARD, J. F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol,* 59**,** 307-21.

GUINEBRETIERE, M. H., THOMPSON, F. L., SOROKIN, A., NORMAND, P., DAWYNDT, P., EHLING-SCHULZ, M., SVENSSON, B., SANCHIS, V., NGUYEN-THE, C., HEYNDRICKX, M. & DE VOS, P. 2008. Ecological diversification in the *Bacillus cereus* Group. *Environ Microbiol,* 10**,** 851-65.

HALEY, B. J., GRIM, C. J., HASAN, N. A., CHOI, S. Y., CHUN, J., BRETTIN, T. S., BRUCE, D. C., CHALLACOMBE, J. F., DETTER, J. C., HAN, C. S., HUQ, A. & COLWELL, R. R. 2010. Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. *BMC Microbiol,* 10**,** 154.

HAMMERL, J. A., KLEIN, I., LANKA, E., APPEL, B. & HERTWIG, S. 2008. Genetic and functional properties of the self-transmissible *Yersinia enterocolitica* plasmid pYE854, which mobilizes the virulence plasmid pYV. *J Bacteriol,* 190**,** 991-1010.

HANAGE, W. P., FRASER, C. & SPRATT, B. G. 2005. Fuzzy species among recombinogenic bacteria. *BMC Biol,* 3**,** 6.

HANAGE, W. P., FRASER, C. & SPRATT, B. G. 2006. Sequences, sequence clusters and bacterial species. *Philos Trans R Soc Lond B Biol Sci,* 361**,** 1917-27.

HARRISON, P. W., LOWER, R. P., KIM, N. K. & YOUNG, J. P. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol,* 18**,** 141-8.

HELGASON, E., OKSTAD, O. A., CAUGANT, D. A., JOHANSEN, H. A., FOUET, A., MOCK, M., HEGNA, I. & KOLSTO, A. B. 2000. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*--one species on the basis of genetic evidence. *Appl Environ Microbiol,* 66**,** 2627-30.

HOGG, J. S., HU, F. Z., JANTO, B., BOISSY, R., HAYES, J., KEEFE, R., POST, J. C. & EHRLICH, G. D. 2007. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol,* 8**,** R103.

HOLT, K. E., BAKER, S., WEILL, F. X., HOLMES, E. C., KITCHEN, A., YU, J., SANGAL, V., BROWN, D. J., COIA, J. E., KIM, D. W., CHOI, S. Y., KIM, S. H., DA SILVEIRA, W. D., PICKARD, D. J., FARRAR, J. J., PARKHILL, J., DOUGAN, G. & THOMSON, N. R. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet,* 44**,** 1056-9.

HUSON, D. H. & BRYANT, D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol,* 23**,** 254-67.

HUSON, D. H. & SCORNAVACCA, C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol,* 61**,** 1061-7.

JORDAN, D. 1984. Family III. Rhizobiaceae. *Bergey's manual of systematic bacteriology,* 1**,** 234-242.

JORDAN, D. C. 1982. NOTES: Transfer of *Rhizobium japonicum* Buchanan 1980 to *Bradyrhizobium* gen. nov., a Genus of Slow-Growing, Root Nodule Bacteria from Leguminous Plants. *International Journal of Systematic Bacteriology,* 32**,** 136-139.

JOSEPH, S. J., DIDELOT, X., ROTHSCHILD, J., DE VRIES, H. J., MORRE, S. A., READ, T. D. & DEAN, D. 2012. Population genomics of *Chlamydia trachomatis*: insights on drift, selection, recombination, and population structure. *Mol Biol Evol,* 29**,** 3933-46.

KATOH, K. & STANDLEY, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol,* 30**,** 772-80.

KEYMER, D. P. & BOEHM, A. B. 2011. Recombination shapes the structure of an environmental *Vibrio cholerae* population. *Appl Environ Microbiol,* 77**,** 537-44.

KIM, N. 2012. Investigation into plasmid replication and partitioning systems in proteobacteria. PhD University of York.

KLEE, S. R., BRZUSZKIEWICZ, E. B., NATTERMANN, H., BRUGGEMANN, H., DUPKE, S., WOLLHERR, A., FRANZ, T., PAULI, G., APPEL, B., LIEBL, W., COUACY-HYMANN, E., BOESCH, C., MEYER, F. D., LEENDERTZ, F. H., ELLERBROK, H., GOTTSCHALK, G., GRUNOW, R. & LIESEGANG, H. 2010. The genome of a *Bacillus* isolate causing anthrax in chimpanzees combines chromosomal properties of *B. cereus* with *B. anthracis* virulence plasmids. *PLoS One,* 5**,** e10986.

KOCH, A. L. 1974. The pertinence of the periodic selection phenomenon to prokaryote evolution. *Genetics,* 77**,** 127-42.

KONSTANTINIDIS, K. T. & TIEDJE, J. M. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A,* 102**,** 2567-72.

KOSKI, L. B., MORTON, R. A. & GOLDING, G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol,* 18**,** 404-12.

KRAWIEC, S. & RILEY, M. 1990. Organization of the bacterial chromosome. *Microbiol Rev,* 54**,** 502-39.

KRZYWINSKI, M., SCHEIN, J., BIROL, I., CONNORS, J., GASCOYNE, R., HORSMAN, D., JONES, S. J. & MARRA, M. A. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res,* 19**,** 1639-45.

LAN, R. & REEVES, P. R. 1996. Gene transfer is a major factor in bacterial evolution. *Mol Biol Evol,* 13**,** 47-55.

LARANJO, M., ALEXANDRE, A., RIVAS, R., VELAZQUEZ, E., YOUNG, J. P. & OLIVEIRA, S. 2008. Chickpea rhizobia symbiosis genes are highly conserved across multiple *Mesorhizobium* species. *FEMS Microbiol Ecol,* 66**,** 391-400.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-8.

LASSALLE, F., CAMPILLO, T., VIAL, L., BAUDE, J., COSTECHAREYRE, D., CHAPULLIOT, D., SHAMS, M., ABROUK, D., LAVIRE, C., OGER-DESFEUX, C., HOMMAIS, F., GUEGUEN, L., DAUBIN, V., MULLER, D. & NESME, X. 2011. Genomic species are ecological species as revealed by comparative genomics in *Agrobacterium tumefaciens*. *Genome Biol Evol,* 3**,** 762-81.

LAWRENCE, J. G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol,* 2**,** 519-23.

LAWRENCE, J. G. & OCHMAN, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol,* 44**,** 383-97.

LAWRENCE, J. G. & OCHMAN, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A,* 95**,** 9413-7.

LEFEBURE, T., BITAR, P. D., SUZUKI, H. & STANHOPE, M. J. 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. *Genome Biol Evol,* 2**,** 646-55.

LEVIN, B. R. 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics,* 99**,** 1-23.

LINZ, B., BALLOUX, F., MOODLEY, Y., MANICA, A., LIU, H., ROUMAGNAC, P., FALUSH, D., STAMER, C., PRUGNOLLE, F., VAN DER MERWE, S. W., YAMAOKA, Y., GRAHAM, D. Y., PEREZ-TRALLERO, E., WADSTROM, T., SUERBAUM, S. & ACHTMAN, M. 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature,* 445**,** 915-8.

LOPEZ-GUERRERO, M. G., ORMENO-ORRILLO, E., VELAZQUEZ, E., ROGEL, M. A., ACOSTA, J. L., GONZALEZ, V., MARTINEZ, J. & MARTINEZ-ROMERO, E. 2012. *Rhizobium etli* taxonomy revised with novel genomic data and analyses. *Syst Appl Microbiol,* 35**,** 353-8.

LOPEZ-MADRIGAL, S., LATORRE, A., PORCAR, M., MOYA, A. & GIL, R. 2011. Complete genome sequence of "*Candidatus Tremblaya princeps*" strain PCVAL, an intriguing translational machine below the living-cell status. *J Bacteriol,* 193**,** 5587-8.

LOZANO, L., HERNANDEZ-GONZALEZ, I., BUSTOS, P., SANTAMARIA, R. I., SOUZA, V., YOUNG, J. P., DAVILA, G. & GONZALEZ, V. 2010. Evolutionary dynamics of insertion sequences in relation to the evolutionary histories of the chromosome and symbiotic plasmid genes of *Rhizobium etli* populations. *Appl Environ Microbiol,* 76**,** 6504-13.

MADDISON, W. P. 1997. Gene Trees in Species Trees. *Systematic Biology,* 46**,** 523-536.

MAIDEN, M. C., BYGRAVES, J. A., FEIL, E., MORELLI, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M. & SPRATT, B. G. 1998. Multilocus sequence typing: a portable

approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A,* 95**,** 3140-5.

MAJEWSKI, J., ZAWADZKI, P., PICKERILL, P., COHAN, F. M. & DOWSON, C. G. 2000. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol,* 182**,** 1016-23.

MANN, R. A., SMITS, T. H., BUHLMANN, A., BLOM, J., GOESMANN, A., FREY, J. E., PLUMMER, K. M., BEER, S. V., LUCK, J., DUFFY, B. & RODONI, B. 2013. Comparative genomics of 12 strains of *Erwinia amylovora* identifies a pan-genome with a large conserved core. *PLoS One,* 8**,** e55644.

MASSON-BOIVIN, C., GIRAUD, E., PERRET, X. & BATUT, J. 2009. Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? *Trends Microbiol,* 17**,** 458-66.

MATIC, I., RAYSSIGUIER, C. & RADMAN, M. 1995. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell,* 80**,** 507-15.

MAZUR, A., STASIAK, G., WIELBO, J., KUBIK-KOMAR, A., MAREK-KOZACZUK, M. & SKORUPSKA, A. 2011. Intragenomic diversity of *Rhizobium leguminosarum* bv. *trifolii* clover nodule isolates. *BMC Microbiol,* 11**,** 123.

MCCUTCHEON, J. P., MCDONALD, B. R. & MORAN, N. A. 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet,* 5**,** e1000565.

MEDINI, D., DONATI, C., TETTELIN, H., MASIGNANI, V. & RAPPUOLI, R. 2005. The microbial pan-genome. *Curr Opin Genet Dev,* 15**,** 589-94.

MENDE, D. R., SUNAGAWA, S., ZELLER, G. & BORK, P. 2013. Accurate and universal delineation of prokaryotic species. *Nat Methods,* 10**,** 881-4.

MENNA, P. & HUNGRIA, M. 2011. Phylogeny of nodulation and nitrogen-fixation genes in *Bradyrhizobium*: supporting evidence for the theory of monophyletic origin, and spread and maintenance by both horizontal and vertical transfer. *Int J Syst Evol Microbiol,* 61**,** 3052-67.

MILNE, I., LINDNER, D., BAYER, M., HUSMEIER, D., MCGUIRE, G., MARSHALL, D. F. & WRIGHT, F. 2009. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics,* 25**,** 126-7.

MORAN, N. A. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell,* 108**,** 583-6.

MOULIN, L., BENA, G., BOIVIN-MASSON, C. & STEPKOWSKI, T. 2004. Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the *Bradyrhizobium* genus. *Mol Phylogenet Evol,* 30**,** 720-32.

MULLER, H. J. 1932. Some Genetic Aspects of Sex. *The American Naturalist,* 66**,** 118-138.

NARRA, H. P. & OCHMAN, H. 2006. Of what use is sex to bacteria? *Curr Biol,* 16**,** R705-10.

NELSON, K. E., FLEISCHMANN, R. D., DEBOY, R. T., PAULSEN, I. T., FOUTS, D. E., EISEN, J. A., DAUGHERTY, S. C., DODSON, R. J., DURKIN, A. S., GWINN, M., HAFT, D. H., KOLONAY, J. F., NELSON, W. C., MASON, T., TALLON, L., GRAY, J., GRANGER, D., TETTELIN, H., DONG, H., GALVIN, J. L., DUNCAN, M. J., DEWHIRST, F. E. & FRASER, C. M. 2003. Complete genome sequence of the oral pathogenic Bacterium *porphyromonas gingivalis* strain W83. *J Bacteriol,* 185**,** 5591-601.

NISHIDA, H. 2012a. Comparative analyses of base compositions, DNA sizes, and dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. *Int J Evol Biol,* 2012**,** 342482.

NISHIDA, H. 2012b. Evolution of genome base composition and genome size in bacteria. *Front Microbiol,* 3**,** 420.

NORMAND, P., ORSO, S., COURNOYER, B., JEANNIN, P., CHAPELON, C., DAWSON, J., EVTUSHENKO, L. & MISRA, A. K. 1996. Molecular phylogeny of the genus *Frankia* and related genera and emendation of the family *Frankiaceae*. *Int J Syst Bacteriol,* 46**,** 1-9.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol,* 302**,** 205-17.

OCHMAN, H., LAWRENCE, J. G. & GROISMAN, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature,* 405**,** 299-304.

OH, S. Y., BUDZIK, J. M., GARUFI, G. & SCHNEEWIND, O. 2011. Two capsular polysaccharides enable *Bacillus cereus* G9241 to cause anthrax-like disease. *Mol Microbiol,* 80**,** 455-70.

OKINAKA, R., CLOUD, K., HAMPTON, O., HOFFMASTER, A., HILL, K., KEIM, P., KOEHLER, T., LAMKE, G., KUMANO, S., MANTER, D., MARTINEZ, Y., RICKE, D., SVENSSON, R. & JACKSON, P. 1999. Sequence, assembly and analysis of pX01 and pX02. *J Appl Microbiol,* 87**,** 261-2.

OLSEN, G. J. & WOESE, C. R. 1993. Ribosomal RNA: a key to phylogeny. *FASEB J,* 7**,** 113-23.

PACHULEC, E. & VAN DER DOES, C. 2010. Conjugative plasmids of *Neisseria gonorrhoeae*. *PLoS One,* 5**,** e9962.

PALLEN, M. J. & WREN, B. W. 2007. Bacterial pathogenomics. *Nature,* 449**,** 835-42.

PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One,* 5**,** e9490.

PRIEST, F. G., BARKER, M., BAILLIE, L. W., HOLMES, E. C. & MAIDEN, M. C. 2004. Population structure and evolution of the *Bacillus cereus* group. *J Bacteriol,* 186**,** 7959-70.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-59.

PROSSER, J. I., BOHANNAN, B. J., CURTIS, T. P., ELLIS, R. J., FIRESTONE, M. K., FRECKLETON, R. P., GREEN, J. L., GREEN, L. E., KILLHAM, K., LENNON, J. J., OSBORN, A. M., SOLAN, M., VAN DER GAST, C. J. & YOUNG, J. P. 2007. The role of ecological theory in microbial ecology. *Nat Rev Microbiol,* 5**,** 384-92.

PUPO, G. M., LAN, R. & REEVES, P. R. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A,* 97**,** 10567-72.

RAMIREZ-BAHENA, M. H., GARCIA-FRAILE, P., PEIX, A., VALVERDE, A., RIVAS, R., IGUAL, J. M., MATEOS, P. F., MARTINEZ-MOLINA, E. & VELAZQUEZ, E. 2008. Revision of the taxonomic status of the species *Rhizobium leguminosarum* (Frank 1879) Frank 1889AL, *Rhizobium phaseoli* Dangeard 1926AL and *Rhizobium trifolii* Dangeard 1926AL. *R. trifolii* is a later synonym of *R. leguminosarum*. Reclassification of the strain *R. leguminosarum* DSM 30132 (=NCIMB 11478) as *Rhizobium pisi* sp. nov. *Int J Syst Evol Microbiol,* 58**,** 2484-90.

REDFIELD, R. J. 2001. Do bacteria have sex? *Nat Rev Genet,* 2**,** 634-9.

REEVE, W., O'HARA, G., CHAIN, P., ARDLEY, J., BRAU, L., NANDESENA, K., TIWARI, R., COPELAND, A., NOLAN, M., HAN, C., BRETTIN, T., LAND, M., OVCHINIKOVA, G., IVANOVA, N., MAVROMATIS, K., MARKOWITZ, V., KYRPIDES, N., MELINO, V., DENTON, M., YATES, R. & HOWIESON, J. 2010a. Complete genome sequence of *Rhizobium leguminosarum* bv. *trifolii* strain WSM1325, an effective microsymbiont of annual Mediterranean clovers. *Stand Genomic Sci,* 2**,** 347-56.

REEVE, W., O'HARA, G., CHAIN, P., ARDLEY, J., BRAU, L., NANDESENA, K., TIWARI, R., MALFATTI, S., KISS, H., LAPIDUS, A., COPELAND, A., NOLAN, M., LAND, M., IVANOVA, N., MAVROMATIS, K., MARKOWITZ, V., KYRPIDES, N., MELINO, V., DENTON, M., YATES, R. & HOWIESON, J. 2010b. Complete genome sequence of *Rhizobium leguminosarum* bv *trifolii* strain WSM2304, an effective microsymbiont of the South American clover *Trifolium* polymorphum. *Stand Genomic Sci,* 2**,** 66-76.

RELIC, B., PERRET, X., ESTRADA-GARCIA, M. T., KOPCINSKA, J., GOLINOWSKI, W., KRISHNAN, H. B., PUEPPKE, S. G. & BROUGHTON, W. J. 1994. Nod factors of *Rhizobium* are a key to the legume door. *Mol Microbiol,* 13**,** 171-8.

RICHTER, M. & ROSSELLO-MORA, R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A,* 106**,** 19126-31.

ROBERTS, M. S. & COHAN, F. M. 1993. The effect of DNA sequence divergence on sexual isolation in *Bacillus*. *Genetics,* 134**,** 401-8.

ROBLEDO, M., VELAZQUEZ, E., RAMIREZ-BAHENA, M. H., GARCIA-FRAILE, P., PEREZ-ALONSO, A., RIVAS, R., MARTINEZ-MOLINA, E. & MATEOS, P. F. 2011. The celC gene, a new phylogenetic marker useful for taxonomic studies in *Rhizobium*. *Syst Appl Microbiol,* 34**,** 393-9.

ROCHA, E. P. & DANCHIN, A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet,* 18**,** 291-4.

ROCHE, P., MAILLET, F., PLAZANET, C., DEBELLE, F., FERRO, M., TRUCHET, G., PROME, J. C. & DENARIE, J. 1996. The common nodABC genes of *Rhizobium meliloti* are host-range determinants. *Proc Natl Acad Sci U S A,* 93**,** 15305-10.

ROGEL, M. A., ORMENO-ORRILLO, E. & MARTINEZ ROMERO, E. 2011. Symbiovars in rhizobia reflect bacterial adaptation to legumes. *Syst Appl Microbiol,* 34**,** 96-104.

SANTILLANA, N., RAMIREZ-BAHENA, M. H., GARCIA-FRAILE, P., VELAZQUEZ, E. & ZUNIGA, D. 2008. Phylogenetic diversity based on rrs, atpD, recA genes and 16S-23S intergenic sequence analyses of rhizobial strains isolated from *Vicia faba* and *Pisum sativum* in Peru. *Arch Microbiol,* 189**,** 239-47.

SCHLIEP, K., LOPEZ, P., LAPOINTE, F. J. & BAPTESTE, E. 2011. Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Mol Biol Evol,* 28**,** 1393-405.

SCHLIEP, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics,* 27**,** 592-3.

SCORTICHINI, M., MARCELLETTI, S., FERRANTE, P. & FIRRAO, G. 2013. A Genomic Redefinition of *Pseudomonas avellanae* species. *PLoS One,* 8**,** e75794.

SHAPIRO, B. J., FRIEDMAN, J., CORDERO, O. X., PREHEIM, S. P., TIMBERLAKE, S. C., SZABO, G., POLZ, M. F. & ALM, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science,* 336**,** 48-51.

SHEPPARD, S. K., DIDELOT, X., JOLLEY, K. A., DARLING, A. E., PASCOE, B., MERIC, G., KELLY, D. J., CODY, A., COLLES, F. M., STRACHAN, N. J., OGDEN, I. D., FORBES, K., FRENCH, N. P., CARTER, P., MILLER, W. G., MCCARTHY, N. D., OWEN, R., LITRUP, E., EGHOLM, M., AFFOURTIT, J. P., BENTLEY, S. D., PARKHILL, J., MAIDEN, M. C. & FALUSH, D. 2013. Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Mol Ecol,* 22**,** 1051-64.

SHEPPARD, S. K., MCCARTHY, N. D., FALUSH, D. & MAIDEN, M. C. 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science,* 320**,** 237-9.

SHI, T. & FALKOWSKI, P. G. 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. *Proc Natl Acad Sci U S A,* 105**,** 2510-5.

SHIMODAIRA, H. & HASEGAWA, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17**,** 1246-7.

SMITH, N. H., KREMER, K., INWALD, J., DALE, J., DRISCOLL, J. R., GORDON, S. V., VAN SOOLINGEN, D., HEWINSON, R. G. & SMITH, J. M. 2006. Ecotypes of the *Mycobacterium tuberculosis* complex. *J Theor Biol,* 239**,** 220-5.

SNEATH, P. H. A. & SOKAL, R. R. 1973. Numerical Taxonomy: The Principles and Practice of Numerical Classification, W H Freeman Limited.

SOREK, R., ZHU, Y., CREEVEY, C. J., FRANCINO, M. P., BORK, P. & RUBIN, E. M. 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science,* 318**,** 1449-52.

SPAINK, H. P. 1994. The molecular basis of the host specificity of the *Rhizobium* bacteria. *Antonie Van Leeuwenhoek,* 65**,** 81-98.

SPAINK, H. P., SHEELEY, D. M., VAN BRUSSEL, A. A., GLUSHKA, J., YORK, W. S., TAK, T., GEIGER, O., KENNEDY, E. P., REINHOLD, V. N. & LUGTENBERG, B. J. 1991. A novel highly unsaturated fatty acid moiety of lipo-oligosaccharide signals determines host specificity of *Rhizobium*. *Nature,* 354**,** 125-30.

SPAINK, H. P., WIJFJES, A. H. & LUGTENBERG, B. J. 1995. *Rhizobium* NodI and NodJ proteins play a role in the efficiency of secretion of lipochitin oligosaccharides. *J Bacteriol,* 177**,** 6276-81.

SPRATT, B. G. 1999. Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. *Curr Opin Microbiol,* 2**,** 312-6.

STACKEBRANDT, E., FREDERIKSEN, W., GARRITY, G. M., GRIMONT, P. A., KAMPFER, P., MAIDEN, M. C., NESME, X., ROSSELLO-MORA, R., SWINGS, J., TRUPER, H. G., VAUTERIN, L., WARD, A. C. & WHITMAN, W. B. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol,* 52**,** 1043-7.

STACKEBRANDT, E. & GOEBEL, B. M. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology,* 44**,** 846-849.

STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G., KORF, I., LAPP, H., LEHVASLAIHO, H., MATSALLA, C., MUNGALL, C. J., OSBORNE, B. I., POCOCK, M. R., SCHATTNER, P., SENGER, M., STEIN, L. D., STUPKA, E., WILKINSON, M. D. & BIRNEY, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12**,** 1611-8.

STOTHARD, P. & WISHART, D. S. 2005. Circular genome visualization and exploration using CGView. *Bioinformatics,* 21**,** 537-9.

SUGAWARA, M., EPSTEIN, B., BADGLEY, B. D., UNNO, T., XU, L., REESE, J., GYANESHWAR, P., DENNY, R., MUDGE, J., BHARTI, A. K., FARMER, A. D., MAY, G. D., WOODWARD, J. E., MEDIGUE, C., VALLENET, D., LAJUS, A., ROUY, Z., MARTINEZ-VAZ, B., TIFFIN, P., YOUNG, N. D. & SADOWSKY, M. J. 2013. Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol,* 14**,** R17.

SUSKO, E., LEIGH, J., DOOLITTLE, W. F. & BAPTESTE, E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Mol Biol Evol,* 23**,** 1019-30.

SUWANTO, A. & KAPLAN, S. 1989a. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: genome size, fragment identification, and gene localization. *J Bacteriol,* 171**,** 5840-9.

SUWANTO, A. & KAPLAN, S. 1989b. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. *J Bacteriol,* 171**,** 5850-9.

TENAILLON, O., SKURNIK, D., PICARD, B. & DENAMUR, E. 2010. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol,* 8**,** 207-17.

TETTELIN, H., MASIGNANI, V., CIESLEWICZ, M. J., DONATI, C., MEDINI, D., WARD, N. L., ANGIUOLI, S. V., CRABTREE, J., JONES, A. L., DURKIN, A. S., DEBOY, R. T., DAVIDSEN, T. M., MORA, M., SCARSELLI, M., MARGARIT Y ROS, I., PETERSON, J. D., HAUSER, C. R., SUNDARAM, J. P., NELSON, W. C., MADUPU, R., BRINKAC, L. M., DODSON, R. J., ROSOVITZ, M. J., SULLIVAN, S. A., DAUGHERTY, S. C., HAFT, D. H., SELENGUT, J., GWINN, M. L., ZHOU, L., ZAFAR, N., KHOURI, H., RADUNE, D., DIMITROV, G., WATKINS, K., O'CONNOR, K. J., SMITH, S., UTTERBACK, T. R., WHITE, O., RUBENS, C. E., GRANDI, G., MADOFF, L. C., KASPER, D. L., TELFORD, J. L., WESSELS, M. R., RAPPUOLI, R. & FRASER, C. M. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A,* 102**,** 13950-5.

TETTELIN, H., NELSON, K. E., PAULSEN, I. T., EISEN, J. A., READ, T. D., PETERSON, S., HEIDELBERG, J., DEBOY, R. T., HAFT, D. H., DODSON, R. J., DURKIN, A. S., GWINN, M., KOLONAY, J. F., NELSON, W. C., PETERSON, J. D., UMAYAM, L. A., WHITE, O., SALZBERG, S. L., LEWIS, M. R., RADUNE, D., HOLTZAPPLE, E., KHOURI, H., WOLF, A. M., UTTERBACK, T. R., HANSEN, C. L., MCDONALD, L. A., FELDBLYUM, T. V., ANGIUOLI, S., DICKINSON, T., HICKEY, E. K., HOLT, I. E., LOFTUS, B. J., YANG, F., SMITH, H. O., VENTER, J. C., DOUGHERTY, B. A., MORRISON, D. A., HOLLINGSHEAD, S. K. & FRASER, C. M. 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science,* 293**,** 498-506.

THOMPSON, C. C., THOMPSON, F. L. & VICENTE, A. C. 2008. Identification of *vibrio cholerae* and *vibrio mimicus* by multilocus sequence analysis (MLSA). *Int J Syst Evol Microbiol,* 58**,** 617-21.

TIAN, C. F., WANG, E. T., WU, L. J., HAN, T. X., CHEN, W. F., GU, C. T., GU, J. G. & CHEN, W. X. 2008. *Rhizobium fabae* sp. nov., a bacterium that nodulates *Vicia faba*. *Int J Syst Evol Microbiol,* 58**,** 2871-5.

TIAN, C. F., YOUNG, J. P., WANG, E. T., TAMIMI, S. M. & CHEN, W. X. 2010. Population mixing of *Rhizobium leguminosarum* bv. *viciae* nodulating *Vicia faba*: the role of recombination and lateral gene transfer. *FEMS Microbiol Ecol,* 73**,** 563-76.

TOMIDA, S., NGUYEN, L., CHIU, B. H., LIU, J., SODERGREN, E., WEINSTOCK, G. M. & LI, H. 2013. Pan-genome and comparative genome analyses of *propionibacterium acnes* reveal its genomic diversity in the healthy and diseased human skin microbiome. *MBio,* 4**,** e00003-13.

TUN-GARRIDO, C., BUSTOS, P., GONZALEZ, V. & BROM, S. 2003. Conjugative transfer of p42a from *rhizobium etli* CFN42, which is required for mobilization of the symbiotic plasmid, is regulated by quorum sensing. *J Bacteriol,* 185**,** 1681-92.

VAN BRUSSEL, A. A., RECOURT, K., PEES, E., SPAINK, H. P., TAK, T., WIJFFELMAN, C. A., KIJNE, J. W. & LUGTENBERG, B. J. 1990. A biovar-specific signal of *Rhizobium leguminosarum* bv. *viciae* induces increased nodulation gene-inducing activity in root exudate of *Vicia sativa* subsp. nigra. *J Bacteriol,* 172**,** 5394-401.

VAN DER AUWERA, G. A., ANDRUP, L. & MAHILLON, J. 2005. Conjugative plasmid pAW63 brings new insights into the genesis of the *Bacillus anthracis* virulence plasmid pXO2 and of the *Bacillus thuringiensis* plasmid pBT9727. *BMC Genomics,* 6**,** 103.

VANDAMME, P., POT, B., GILLIS, M., DE VOS, P., KERSTERS, K. & SWINGS, J. 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev,* 60**,** 407-38.

VERNIKOS, G. S., THOMSON, N. R. & PARKHILL, J. 2007. Genetic flux over time in the *Salmonella* lineage. *Genome Biol,* 8**,** R100.

VINUESA, P., ROJAS-JIMENEZ, K., CONTRERAS-MOREIRA, B., MAHNA, S. K., PRASAD, B. N., MOE, H., SELVARAJU, S. B., THIERFELDER, H. & WERNER, D. 2008. Multilocus sequence analysis for assessment of the biogeography and evolutionary genetics of four *Bradyrhizobium* species that nodulate soybeans on the asiatic continent. *Appl Environ Microbiol,* 74**,** 6987-96.

VOS, M. 2009. Why do bacteria engage in homologous recombination? *Trends Microbiol,* 17**,** 226-32.

VOS, M. & DIDELOT, X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J,* 3**,** 199-208.

WAGNER, S. C. 2012. Biological Nitrogen Fixation. Nature Education Knowledge 3(10):15

WANG, T., CHEN, Z., CHENG, Q., ZHOU, M., TIAN, X., XIE, P., ZHONG, L., SHEN, M. & QIN, Z. 2012. Characterization of replication and conjugation of plasmid pWTY27 from a widely distributed *Streptomyces* species. *BMC Microbiol,* 12**,** 253.

WARD, D. M., COHAN, F. M., BHAYA, D., HEIDELBERG, J. F., KUHL, M. & GROSSMAN, A. 2008. Genomics, environmental genomics and the issue of microbial species. *Heredity (Edinb),* 100**,** 207-19.

WAYNE, L. G., BRENNER, D. J., COLWELL, R. R., GRIMONT, P. A. D., KANDLER, O., KRICHEVSKY, M. I., MOORE, L. H., MOORE, W. E. C., MURRAY, R. G. E., STACKEBRANDT, E., STARR, M. P. & TRUPER, H. G. 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic Bacteriology,* 37**,** 463-464.

WHITAKER, R. J. 2006. Allopatric origins of microbial species. *Philos Trans R Soc Lond B Biol Sci,* 361**,** 1975-84.

WHITAKER, R. J., GROGAN, D. W. & TAYLOR, J. W. 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science,* 301**,** 976-8.

WISNIEWSKI-DYE, F., BORZIAK, K., KHALSA-MOYERS, G., ALEXANDRE, G., SUKHARNIKOV, L. O., WUICHET, K., HURST, G. B., MCDONALD, W. H., ROBERTSON, J. S., BARBE, V., CALTEAU, A., ROUY, Z., MANGENOT, S., PRIGENT-COMBARET, C., NORMAND, P., BOYER, M., SIGUIER, P., DESSAUX, Y., ELMERICH, C., CONDEMINE, G., KRISHNEN, G., KENNEDY, I., PATERSON, A. H., GONZALEZ, V., MAVINGUI, P. & ZHULIN, I. B. 2011. *Azospirillum* genomes reveal transition of bacteria from aquatic to terrestrial environments. *PLoS Genet,* 7**,** e1002430.

WU, H., ZHANG, Z., HU, S. & YU, J. 2012. On the molecular mechanism of GC content variation among eubacterial genomes. *Biol Direct,* 7**,** 2.

YANG, F., YANG, J., ZHANG, X., CHEN, L., JIANG, Y., YAN, Y., TANG, X., WANG, J., XIONG, Z., DONG, J., XUE, Y., ZHU, Y., XU, X., SUN, L., CHEN, S., NIE, H., PENG, J., XU, J., WANG, Y., YUAN, Z., WEN, Y., YAO, Z., SHEN, Y., QIANG, B., HOU, Y., YU, J. & JIN, Q. 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res,* 33**,** 6445-58.

YOUNG, J. P., CROSSMAN, L. C., JOHNSTON, A. W., THOMSON, N. R., GHAZOUI, Z. F., HULL, K. H., WEXLER, M., CURSON, A. R., TODD, J. D., POOLE, P. S., MAUCHLINE, T. H., EAST, A. K., QUAIL, M. A., CHURCHER, C., ARROWSMITH, C., CHEREVACH, I., CHILLINGWORTH, T., CLARKE, K., CRONIN, A., DAVIS, P., FRASER, A., HANCE, Z., HAUSER, H., JAGELS, K., MOULE, S., MUNGALL, K., NORBERTCZAK, H., RABBINOWITSCH, E., SANDERS, M., SIMMONDS, M., WHITEHEAD, S. & PARKHILL, J. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol,* 7**,** R34.

YOUNG, J. P. W. 1985. Rhizobium Population Genetics: Enzyme Polymorphism in Isolates from Peas, Clover, Beans and Lucerne Grown at the Same Site. *Journal of General Microbiology,* 131**,** 2399-2408.

YOUNG, J. P. W. 1996. Phylogeny and taxonomy of rhizobia. *Plant and Soil,* 186**,** 45-52.

ZAWADZKI, P., ROBERTS, M. S. & COHAN, F. M. 1995. The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. *Genetics,* 140**,** 917-32.

ZHANG, Y. M., TIAN, C. F., SUI, X. H., CHEN, W. F. & CHEN, W. X. 2012. Robust markers reflecting phylogeny and taxonomy of rhizobia. *PLoS One,* 7**,** e44936.

ZWICK, M. E., JOSEPH, S. J., DIDELOT, X., CHEN, P. E., BISHOP-LILLY, K. A., STEWART, A. C., WILLNER, K., NOLAN, N., LENTZ, S., THOMASON, M. K., SOZHAMANNAN, S., MATECZUN, A. J., DU, L. & READ, T. D. 2012. Genomic characterization of the *Bacillus cereus* sensu lato species: backdrop to the evolution of *Bacillus anthracis*. *Genome Res,* 22**,** 1512-24.