Association of genes in bacterial population genomics

Piyachat Udomwong

PhD

University of York

Biology

September 2015

Abstract

This study aims to gain insight into the distribution of genes within a local bacterial population. It is based on genomic and phenotypic data for seventy-two *Rhizobium leguminosarum* strains isolated from *Trifolium repens* and *Vicia sativa* plants growing in a 1 m^2 area of roadside vegetation located between Wentworth College and Walmgate Stray on the University of York campus, UK. These were previously classified into five genospecies based on a phylogeny of 305 universal core genes. Because of the incongruence in the phylogeny of genes from different replicons shown in the previous study, the congruence of each rhizobium gene was first analysed by clanistics. Clanistics were applied to each gene tree of the population and can not only identify conserved genes in symbiovar and genospecies but also genes shared between two symbiovars and within genospecies. Genes of the rhizobium population were investigated along with their occurrence patterns in the population using Pearson's correlation. A broader view of occurrence of genes in the population was illustrated in the gene co-occurrence network, which reflected the organisation of genes with favoured and disfavoured co-occurrence in the population. The computation demonstrated clusters of genes involved in the nodulation process including both annotated and unannotated genes. Due to the diversity of ability to utilise carbon substrates in the rhizobium population, class association rule was chosen as the method to relate the ability to utilise carbon substrates with the distribution of genes in the population. Results demonstrated that there not only exists a relationship between the ability to utilise a substrate and the distribution of genes in the population, but also cooperation of genes involved in the substrate utilisation. The methods discovered genes involved in the utilisation of γ-hydroxybutyric acid, which were consistent with evidence from experiments and the literature. Hence, it can be concluded that gene transfer and loss can cause variation in the gene content of a population, resulting in recognisable sets of genes present in a particular symbiovar or genospecies, or associated with phenotypes such as substrate utilisation.

Table of Contents

Chapter 4 Analysis of phenotype-genotype data of the local population of *Rhizobium leguminosarum*

List of Figures

Chapter 1 Introduction

Chapter 2 Clanistics as a tool to explore gene distributions in a local population of *Rhizobium leguminosarum*

Chapter 3 Co-occurrence of genes in the local population of *Rhizobium leguminosarum*

Chapter 4 Analysis of phenotype-genotype data of the local population of *Rhizobium leguminosarum*

Appendix III

List of Tables

Chapter 2 Clanistics as a tool to explore gene distributions in a local population of *Rhizobium leguminosarum*

Chapter 3 Co-occurrence of genes in the local population of *Rhizobium leguminosarum*

Chapter 4 Analysis of phenotype-genotype data of the local population of *Rhizobium leguminosarum*

Appendix I

Appendix II

Appendix III

Acknowledgments

First and foremost, I would like to thank my supervisor Professor Peter Young for his guidance and support throughout my doctoral studies. His patience, enthusiasm, motivation and immense knowledge inspired me to carry on research as a bioinformatician. I would like to acknowledge the help and advice offered by the members of my thesis advisory panel, Dr Peter Ashton and Dr James Cussens and the thought provoking questions raised by them during each meeting.

I wish to thank Shakeel, Nayoung, Nitin, and Kailin, who have contributed to my personal and professional development at the University of York. I am also grateful for my Thai friend in UK and Thailand. A special thanks to Ganesh; his knowledge of biology and warm companionship fulfilled the needs of friend and family during my study.

Finally, my family and relatives, I would like to thank my parents Mr. Mongkol and Mrs. Onwisa Udomwong for their unconditional love, support and encouragement. I also would like to thank my grandma, Mrs. Chaaon Srisup and my sister, Patcharin Udomwong, for their consistent support.

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Part of the work described in Chapter 4 has been published in the following research article:

Kumar, N., Lad, G., Giuntini, E., Kaye, M. E., Udomwong, P., Shamsani, N. J., Young, J. P. W. & Bailly, X. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biol,* 5.

Chapter 1 Introduction

1.1 Introduction

Since the sequencing of the first bacterial genome *(Haemophilus influenzae)* (Fleischmann *et al.*, 1995), the attempt to gain an insight into bacterial genomes has never stopped. The first comparative study of bacterial genomes, involving two strains of *Helicobacter pylori* (Alm *et al.*, 1999), was published five years later. It revealed two types of genes, viz. genes present in both the strains and genes present in only one of the two strains. Even if the findings in the study did not officially constitute a novel branch of comparative genomics, it triggered an intense study of comparative genomics to answer the question whether an individual sequenced genome would be enough to describe a bacterial species.

Tettelin *et al.* (2005) introduced the concept of "pan-genome" by comparing eight strains of *Streptococcus agalactiae.* The pan-genome was described as a gene pool containing all genes that can potentially be present in that species. As a result of advances in techniques used in the biological sciences, sequencing the genome of an organism is no longer a time-consuming task but a technology that provides more accurate outcomes at less expense resulting in useful high-throughput data. As a result, many bacterial genomes have been sequenced and studied based on the concept of pan-genome. Comparative studies have revealed many aspects and facts of the bacterial genome including distinct features of genes in the pan-genome (Young *et al.*, 2006), pathogenesis (Hogg *et al.*, 2007, Donati *et al.*, 2010, Tomida *et al.*, 2013, Méric *et al.*, 2014), ecological niche adaptation (Cadillo-Quiroz *et al.*, 2012, Ellegaard *et al.*, 2013, Ooi *et al.*, 2013, Zhang *et al.*, 2014), host specificity (Black *et al.*, 2012) and redefinition of bacterial species by using genes shared by all strains (Lan *et al.*, 1996).

This study aims to explore and investigate the distribution of genes within a bacterial population. This chapter provides an introduction to replicon architecture in bacteria, genetic exchange in bacteria, and *Rhizobium leguminosarum*, which used in this study as a model species.

1.2 Replicon architecture in bacteria

The domain Bacteria is one of three primary domains of life, the other two being Archaea and Eukarya (Woese *et al.*, 1990). Bacteria are minute, single-celled microorganisms. Cellular architecture of bacteria is considerably simpler than eukaryotes. Like Archaea, bacteria have no membrane-bound organelles. Chloroplasts and mitochondria, membrane-bound organelles in eukaryotes, may have originated from bacteria as a result of endosymbiosis and hence carry their own genetic material (Gray, 1999). The hereditary information in bacteria is carried on three classes of replicon, viz. chromosome, chromid and plasmid (Harrison *et al.* 2010).

1.2.1 Chromosome

The bacterial chromosome is usually circular and the largest replicon (Ochman, 2002), and carries a set of core genes responsible for encoding fundamental processes, the basic aspects of biology of a species and its major phenotypic traits. The chromosome generally has a higher percentage of guanine and cytosine (GC) than plasmids. Previous studies (Bentley *et al.*, 2004, Nishida, 2012) have reported that chromosomal size and base composition have a linear positive correlation. Larger genomes tend to have a high GC content (i.e. CG-rich), while smaller genomes tend to have low GC content (alternately, AT-rich).

A traditional view has been that the bacterial genome is contained in a single replicon, the chromosome (unipartite genomes) as in *Streptococcus pneumoniae* (Tettelin et al., 2001) and *Porphyromonas gingivalis* (Nelson et al., 2003); but some bacteria always carry replicons (chromids and plasmids) in addition to the chromosome (multipartite genomes).

1.2.2 Chromids

Chromids were first described as “secondary chromosomes” or “megaplasmids” when discovered in the *Rhodobacter sphaeroides* 2.4.1 (Suwanto *et al.*, 1989). Harrison *et al.* (2010) performed a systematic analysis on these secondary chromosomes. These replicons reflect some features of a chromosome and others of a plasmid. These replicons are normally larger in size than plasmids but smaller than the chromosome. Harrison *et al.* (2010) called them 'chromids' and established three criteria to clarify and distinguish chromids from accompanying plasmids and primary chromosome: (i) chromids have plasmid-type maintenance and replication systems, (ii) chromids have a nucleotide composition close to that of the primary chromosome, and (iii) chromids carry several essential (core) genes that are found on the chromosome in other species.

1.2.3 Plasmids

Plasmids are described as the smallest self-replicating elements (Ochman, 2002, Frost *et al.*, 2005). A majority of genes on plasmids are accessory genes that are responsible for additional properties such as virulence (Duangsonk *et al.*, 2006), ecological determinants (Prosser *et al.*, 2007, Norman *et al.*, 2009) and host specificity (Paulsson, 2002). Genes on plasmids exhibit diversity that can be observed in strains of the same species. Horizontal gene transfer is the key mechanism for transmission of plasmid genes. The transmission can be in the form of transient genes or the entire replicon may be transferred between bacteria. Genes and their dynamic transmission in plasmids lead to genetic diversity within the species and drive bacterial evolution. The plasmids carry some conserved genes (viz. 'backbone genes') that are present in all plasmids, such as the replication system *repABC* in alpha-proteobacterial plasmids (Cevallos *et al.*, 2008).

1.3 Genetic exchange in bacteria

Asexual reproduction processes in bacteria, including bipolar division, binary fission, budding, and multiple fission (Angert, 2005) do not provide an opportunity for genetic exchange. Genetic material in the bacteria can be exchanged via homologous recombination and horizontal gene transfer (HGT). The modified genetic material is passed from donor (providing a genetic element) to recipient (receiving the genetic element) in a single direction.

1.3.1 Homologous recombination

Homologous recombination is described as the migration of homologous genes or short pieces of genes from a donor cell to a recipient cell, occurring in closely related microbes. Abundant genes in a species show a greater impact of homologous recombination than rare (Ochman *et al.*, 2000). Hence, this type of genetic exchange affects core genes more than accessory genes (Didelot *et al.*, 2012). Rather than introducing uniqueness to the genome, homologous recombination retains the fitness of the bacterial species (Donati *et al.*, 2010). This suggests that this type of genetic exchange does not play an important role in creating novel ecological and physiological adaptation but helps in the maintenance of adaptation. In addition, the degree of homologous recombination is higher within a subgroup than between different subgroups of the species (Ellegaard *et al.*, 2013, Méric *et al.*, 2014).

1.3.2 Horizontal gene transfer

Horizontal Gene Transfer (HGT) (also known as Lateral Gene Transfer (LGT)) is defined as the acquisition of non-homologous genes or short pieces of genes from distantly related or unrelated taxa (Goldenfeld *et al.*, 2007, Lawrence *et al.*, 2009, Boto, 2010). This genetic transfer involves acquiring non-homologous DNA, so its impact is generally observed to be more frequent on accessory genes (shared by some strains in the species) than the core genes (shared by all strains). This type of genetic exchange helps bacteria to obtain benefits from acquisition of novel genes conferring abilities such as environmental adaptation (Espinosa-Urgel *et al.*, 1998) and pathogenesis (Furuya *et al.*, 2006). Furthermore, the genetic exchange also promotes bacterial speciation (Ochman *et al.*, 2000). Hence, the acquisition of genes by HGT possibly increases the uniqueness of bacteria.

Genetic transfer can be described as the mechanism by which DNA of a donor cell passes to a recipient cell. For homologous recombination and HGT, genes can be transferred by three distinct mechanisms viz. transformation, transduction and conjugation (Figure 1.1).

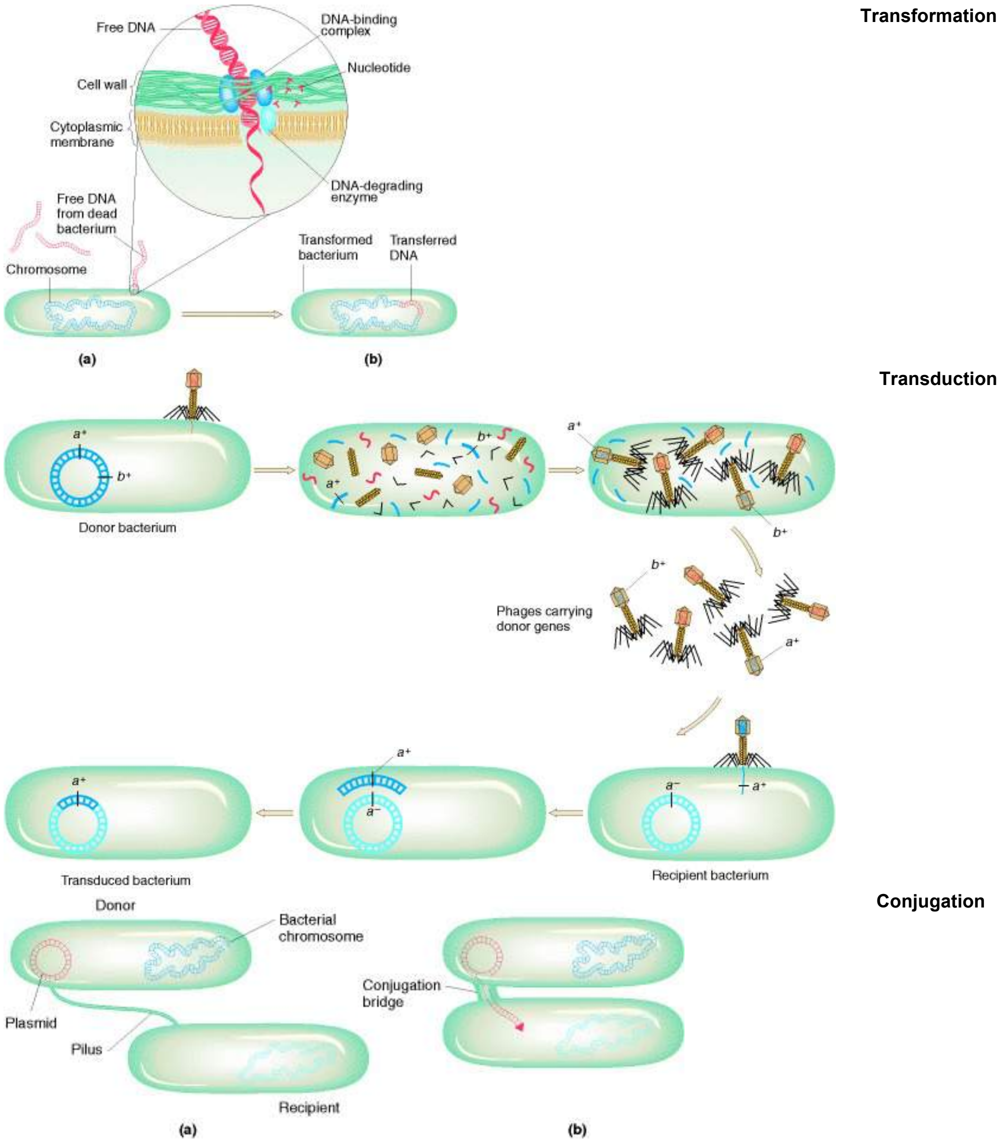


Figure 1.1 Mechanism of genetic exchange in bacteria (Griffiths *et al.*, 2000).

Transformation is an event in which a recipient cell takes up exogenous DNA directly from the environment. Griffith (1928) demonstrated transformation in *Streptococcus pneumoniae,* which can adapt itself under the host immune system and does so by acquiring DNA from dead cells.

Transduction is a mechanism driven by viruses called bacteriophages, which are able to infect bacterial cells and use the bacterial cells as host for reproduction. During multiplication, the bacteriophage genome may integrate into the host genome. When it leaves the old host and infects the new host, it may accidentally transfers gene from the donor to the recipient cell.

Conjugation is a genetic transfer mechanism based on cell-to-cell contact. During conjugation, the donor cell produces a tube-like structure, called a pilus which recognises and binds to the surface of the recipient cell forming a bridge. After the donor and recipient cells are connected by the pilus, genetic transfer is initiated. Plasmids carrying transfer (*tra*) genes and *oriT* (origin of transfer) sequences are able to organise conjugation. However, plasmids that are not carrying *tra* genes but having *oriT* sequences are known as mobilizable plasmids and are able to use the conjugation machinery provided by self-transmissible conjugative plasmids.

1.4 The bacterial pan-genome

Tettelin *et al.* (2005) introduced the concept of "pan-genome" by comparing eight strains of *Streptococcus agalactiae.* The pan-genome ("pan" – "παν" means "whole" in Greek) is a gene pool containing all genes that can potentially be present in that species (Medini *et al.*, 2005).

The Core Genome Hypothesis (CGH) was proposed to describe distinct features of genes in the bacterial genome (Lan *et al.*, 1996). The bacterial genome can be differentiated into a core genome and an accessory genome (Medini *et al.*, 2005, Riley *et al.*, 2009).

The core genome is the genome shared by all member of a species and contains mostly chromosomal genes that have a stable organisation. The typical core genes also have a high GC content and are passed from the parent cell to their daughter cells through vertical gene transfer (VGT). These core genes carry out essential functions relating to survival (Rogel *et al.*, 2011) such as assembly of the transcription machinery, ribosome synthesis, chaperones, and cell division. They have orthologs in related species and their phylogenies harmonize with those inferred from rRNA sequences. These core genes are therefore used to define general bacterial or species-specific characteristics.

The accessory genome contains accessory genes, which are mostly genes on plasmids and other genes that might be present on genomic islands on the chromosome. The accessory genome has a low GC content and confers functions in supplementary biological pathways encoding products that serve ecological adaptation, antibiotic resistance, etc. The accessory genes vary greatly from strain to strain in a bacterial population and are transferred by HGT (Lawrence, 1999, Ragan, 2001, Riley *et al.*, 2009).

The CGH (Lan *et al.*, 1996) points out that during the evolution of bacteria, core genes will, on average, display a neutral rate of evolution, while accessory genes will experience a variety of selective pressures. Thus, the average rate of evolution for accessory genes could be about anything and the variance around this rate could be extreme. Hence the distribution of genes may have underlying patterns, which may reveal the characteristics of genetic organization by their occurrence.

As the pan-genome size grows with an increase in the number of sequenced strains, the increment in accessory genes after sequencing of new strains characterises the pan-genome as an "open" pan-genome. An open pan-genome is usually found in species that colonize multiple niches and exchange genes in multiple ways such as *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). A pan-genome is characterised as a "closed" pan-genome when new strains are sequenced but the pan-genome size does not change. A closed pan-genome is more conserved, niche-isolated and has limited genetic exchange, like *Bacillus anthracis* (Medini *et al.*, 2005), *Campylobacter coli* and *C. jejuni* (Lefébure *et al.*, 2010).

1.5 Rhizobium used in this study as model bacterium

Rhizobial data were used in order to test the performance of research methods. Rhizobia are bacteria that play a very important role in agriculture by inducing nitrogen-fixing nodules on roots of leguminous plants including vetch, pea, bean, clover and alfalfa. Many rhizobia have been described, including a group of closely related species in the genus *Rhizobium* consisting of *R. etli, R. fabae, R. pisi, R. phaseoli* and *R. leguminosarum*. The type species of the genus *Rhizobium* is *R. leguminosarum*. Symbiovars in rhizobia have been used to distinguish symbiotically distinct subgroups within a single rhizobial species (Long, 1989, Rogel *et al.*, 2011). In *Rhizobium leguminosarum*, three symbiovars have been defined using host specificity, viz.: *viciae* (nodulating pea, broad bean and vetch), *trifolii* (nodulating clover), and *phaseoli* (nodulating kidney bean).

The complete sequence of *R. leguminosarum* biovar *viciae* strain 3841 (*Rlv*3841) was published in 2006 (Young *et al.*, 2006). The 7.75 Mb genome comprises one circular chromosome and six circular plasmids viz. pRL7, pRL8, pRL9, pRL10, pRL11, and pRL12 (Figure 1.2). Bacterial genomes can be distinguished into two distinct components: a conserved core and a variable accessory genome. Although core genes are considered functionally essential in the genome and carried by the chromosome, Harrison *et al.* (2010) pointed out that plasmids pRL11 and pRL12 also contains core genes and called them "chromids" because their properties were intermediate between those of chromosomes and plasmids.

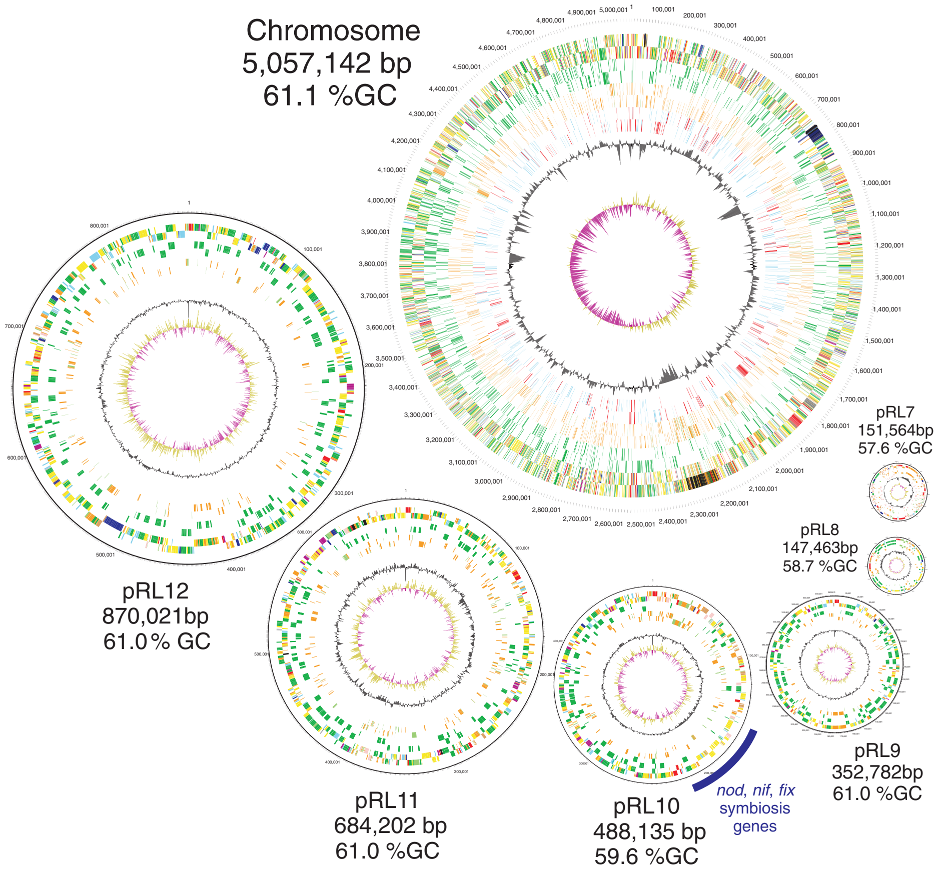


Figure 1.2 Genomic structure of *R. leguminosarum* biovar *viciae* strain 3841 comprising one circular chromosome and six plasmids (taken from Young *et al.* (2006)).

Seventy-two *R. leguminosarum* strains were isolated from an area of one metre squared behind Wentworth College, University of York. These 72 local strains consisted of 36 bv. *viciae* strains isolated from *Vicia sativa* (vetch) and 36 bv. *trifolii* strains isolated from *Trifolium repens* (clover). The variation amongst the isolates was investigated by mapping the 72 Wentworth strains of *R. leguminosarum* along with isolates from different geographical locations to *Rlv*3841 (similarity sequence search based on GSmapper). A phylogeny based on 305 universal core genes showed five clear clades within *R. leguminosarum* (Figure 1.3), which were called genospecies A-E (Kumar *et al.*, 2015).

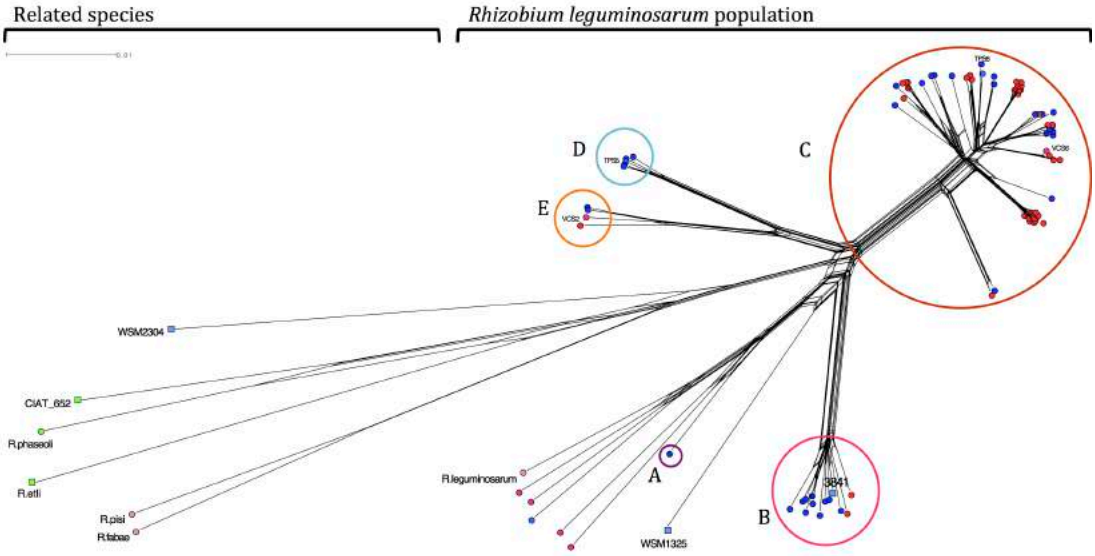


Figure 1.3 The 305 core-gene phylogeny constructed from rhizobium strains obtained from different locations shows differences between the genospecies (gs). With Wentworth strains (bv. *trifolii* as ● and bv. *viciae* as ●), Swedish strains as ●, and Scottish strains as ●: gs A has only one strain, gs B includes *Rlv*3841 (reference strain ■) and 12 more strains, gs C includes 52 strains, gs D and E contain 4 and 3 strains, respectively (taken from Kumar *et al.* (2015)).

1.6 Aims of the study

The overall aim of the study is to explore and investigate the distribution and contribution of core and accessory genes in the bacterial population by computational approaches. The study is carried out bearing in mind the following questions.

1. How do patterns of gene transfer relate to genomic location, gene function, and population substructure in a bacterial population?
2. Can genes with related functions be identified from patterns of gene co-occurrence in a bacterial population?
3. Can gene functions be discovered by examining the relationship between genotype and phenotype in a bacterial population?

To answer these questions, each chapter has objectives as follows:

Chapter 2 Clanistics as a tool to explore gene distributions in a local population of *Rhizobium leguminosarum*

- Investigate discordant and concordant genes in subspecies and symbiovars of the rhizobium population.
- Investigate degree of gene transfer between subspecies and between symbiovars.
- Explore relationship of discordant and concordant genes and their function.

Chapter 3 Co-occurrence of genes in a local population of *Rhizobium leguminosarum*

- Investigate and explore co-occurrence of genes in the bacterial population by correlational computation.
- Compare and find the most suitable computational method for extracting co-occurrence of genes.
- Apply network analysis to disclose latent relationships of genes in the bacterial population.

Chapter 4 Analysis of phenotype-genotype data of the local population of *Rhizobium leguminosarum*

- Assess association between sequence and functional features to identify contribution of presence of genes responsible for a phenotype.
- Compare and find the most suitable computational method for identifying the genes responsible for a phenotype.

Chapter 2 Clanistics as a tool to explore gene distributions in a local population of *Rhizobium leguminosarum*

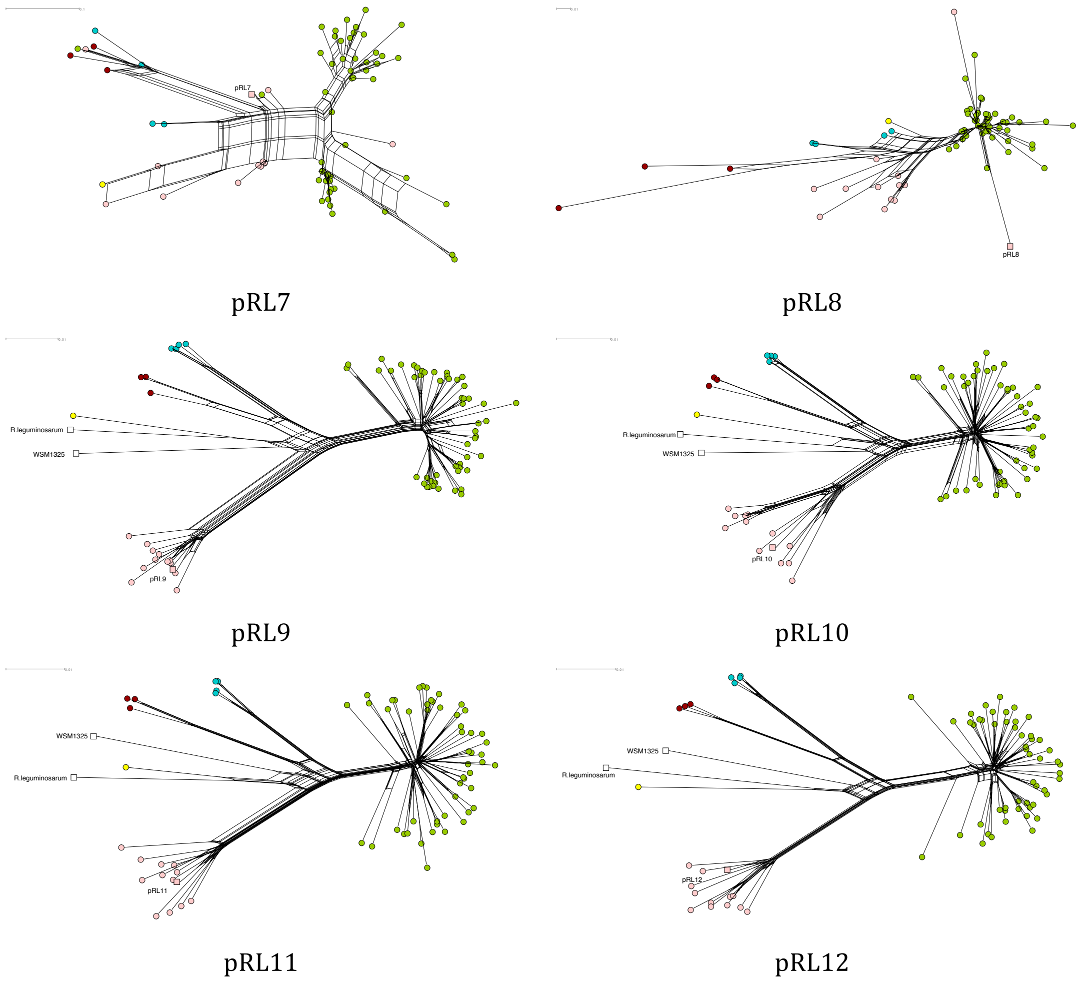
2.1 Abstract

Evolutionary signals in the local rhizobium population were observed via an incongruence in the phylogenetic trees constructed from distinct replicons in the genome. The patterns of gene acquisition in a local population are explored by classifying the arrangement of the symbiovar or the genospecies categories on the phylogeny of each gene in turn, an approach known as "clanistics". An unrooted-tree-based analysis was applied to 6,509 unrooted gene trees in order to detect candidate genes with remarkable events of gene transfer. Results from the analysis will reveal the plurality of evolutionary signals in the local population and dominant features of genes affected by genetic transfer events.

2.2 Introduction

Prokaryotic or bacterial evolution is not only influenced by duplication, losses, and vertical gene transfer events but also homologous recombination and horizontal gene transfer (HGT). Shared evolution of genes in the rhizobium population by recombination and/or HGT between strains necessarily leads to incongruence among single gene phylogenies (Miller *et al.*, 2007, Tian *et al.*, 2010, Rogel *et al.*, 2011).

Phylogenies of 6 non-chromosomal replicons (Figure 2.1) demonstrate incongruence between them and the five genospecies defined by the chromosomal phylogeny. While pRL9, pRL10, and the two chromids (pRL11 and pRL12) do not demonstrate much incongruence, noticeable phylogenetic incongruence is observed on the phylogenies of pRL7 and pRL8, which are conjugative plasmids and contain mainly accessory genes (Young *et al.*, 2006). This phylogenetic incongruence may result from gene transfer in the rhizobium population that has occurred between genospecies. The shared evolution of genes is considerably driven by recombination and/or relatively frequent HGT between strains.



pRL7 pRL8

pRL9 pRL10

pRL11 pRL12

Figure 2.1 Phylogenies of six plasmids are constructed from Rhizobium strains from different locations. 72 rhizobium strains are coloured by their genospecies (●: genospecies A, ●: genospecies B, ●: genospecies C, ●: genospecies D, and ●: genospecies E) (taken from Kumar *et al.* (2015)).

However, manifest gene transfer events are not observed on the chromosomal, pRL9, pRL10, pRL11 and pRL12 phylogenies (Figure 2.1) because these phylogenies are obtained from many genes on each replicon. Many studies (Mutch *et al.*, 2004, Tian *et al.*, 2010, Bailly *et al.*, 2011) attempt to find evidence of gene transfer in the population using a comparative genome approach.

Phylogenetic analysis of nodulation genes on pRL10 (Figure 2.2) has revealed that the nod gene phylogeny conflicts with the core gene phylogeny. Strains that are genetically closely linked in the core gene phylogeny have varied nod genes that are genetically distant on the nod gene phylogeny. These nod genes leave a trace of gene transfer events (Kumar *et al.*, 2015). Another study on the 72 Wentworth strains conducted by Kailin Hui demonstrated that there are five genes carried on pRL8 (which is the smallest replicon) that are present only in *bv. viciae* and are absent from *bv. trifolii*, i.e. symbiovar *viciae* specific (*bvs* genes). It is plausible that these five *bvs* genes serve a specific adaptive function in symbiovar *viciae*. Consequently, the exploration of incongruence in phylogenetic gene trees in the rhizobium genome may disclose gene transfer signals in the rhizobium population.

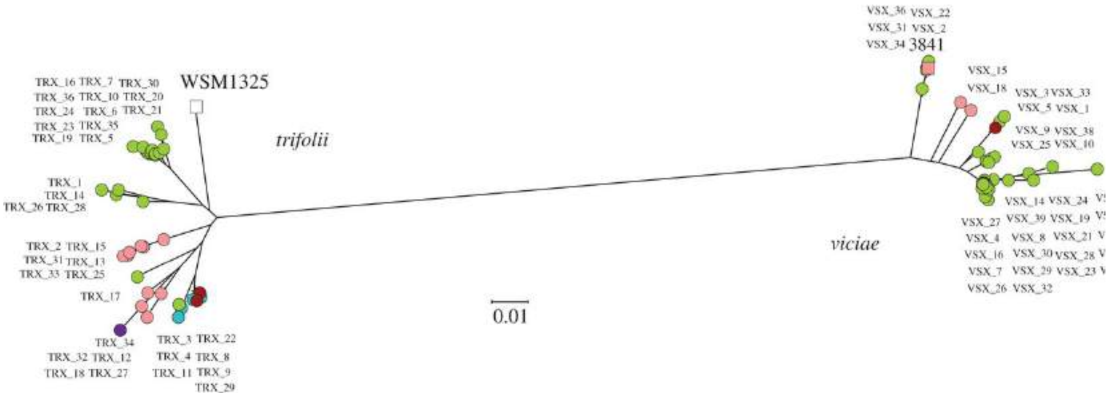


Figure 2.2 Phylogeny of 11 nodulation genes (*nodABCDEFIJLMN*) of *Rlv* 3841 and *Rlt* WSM1325 for two symbiovars *viciae* and *trifolii*. Each strain is coloured by their genospecies (●: genospecies A, ●: genospecies B, ●: genospecies C, ●: genospecies D, and ●: genospecies E) (taken from Kumar *et al.* (2015)).

Phylogenetic trees are a key tool for studying evolution. However, rooting a phylogeny of bacterial genes is a difficult task because bacterial genes are regularly disturbed during evolution by processes such as loss, duplication and horizontal gene transfer. For this reason, study of the evolution of bacteria is conducted using unrooted phylogenetic trees. The terminology of relationship in an unrooted phylogenetic trees is slightly different from that of a rooted phylogenetic tree (Wilkinson *et al.*, 2007) (Figure 2.3). The properties of an unrooted phylogenetic tree were developed by Lapointe *et al.* (2010) as a tool for phylogenomic study to detect a genetic sharing signal.

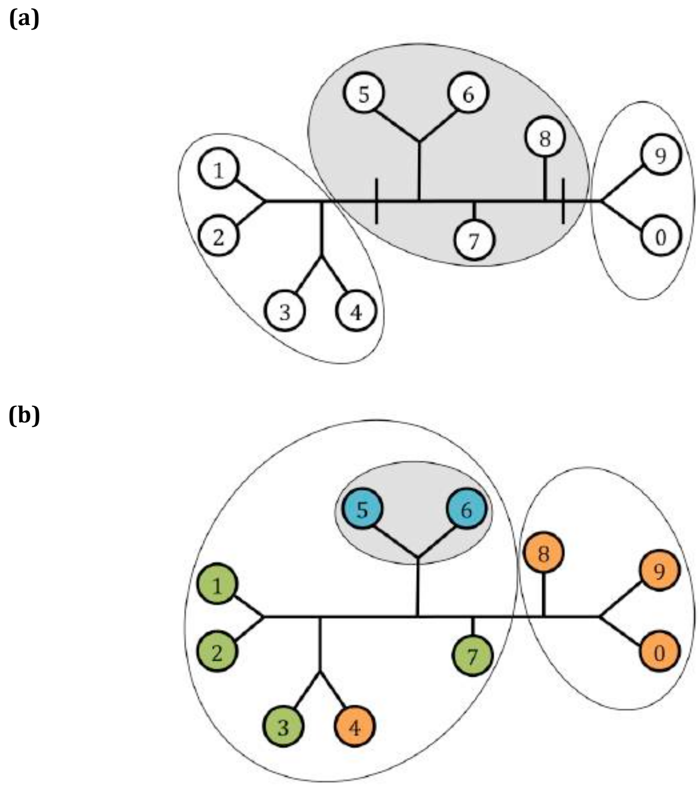


Figure 2.3 Illustration of clanistics concept for an unrooted tree (a) A **clan** is identified by every bipartition of an unrooted tree. In this figure {1,2,3,4} and {0,9} are clans. A **slice** is identified by every tripartition of an unrooted tree. In this figure {5,6,7,8} is slice. **(b)** Categories are defined with three different colours, for example, if ● is defined as native to a group then the other colours are intruders if they occur within that group. A **complete group** is defined as a clan or slice including all members from a single category. Clan {1,2,3,4,5,6,7} is complete because it includes all natives. A **homogeneous group** is defined a clan or slice with members from the same category. Clan {8,9,0} is homogeneous because it contains elements from a single category. A **perfect group** is defined a clan or slice that contains all members from a single category and no intruders. Clan {5,6} is a perfect clan (redrawn from Lapointe *et al.* (2010)).

Clanistic analysis (Lapointe *et al.*, 2010) employs features of an unrooted tree such as clans and introduces novel features like slices (Figure 2.3) to harvest a signal of genetic sharing by partitioning unrooted phylogenetic trees in a forest (a collection of unrooted trees). A principal idea of clanistic analysis is to provide coherent partition(s) on unrooted trees in order to capture evolutionary signals among operational taxonomical units (OTUs) which are labelled either “native” or “intruder”, in which native is a member of the category of interest, while intruder is not. The selection of the interesting category is dependent on what evolutionary questions need to be answered. For example, in the case of pathogenicity, pathogenic strains are defined as natives and non-pathogenic are defined as intruders (Beauregard-Racine *et al.*, 2011). Lifestyle adaption of bacteria is also used for category definition, such as anaerobes defined as native and their complement defined as intruder (Schliep *et al.*, 2011). After partitioning, genes with evidence for recombination are specified when the unrooted gene tree either gets more than one split or contains a heterogeneous group (a group that includes native and intruder) in the same partition. Alternately, genes without recombinant evidence are specified when the unrooted gene tree gets no more than one split or contains a homogeneous group (a group that includes single category).

In this chapter, we focus on genetic sharing in the Wentworth rhizobium population. At the beginning, genes of the local rhizobium population that have orthologs in *Rlv*3841 are investigated to find evidence of genetic transfer by using either the two symbiovars or the five genospecies for category definition. Next, genes with or without evidence for recombination are further investigated for their location in the genome and function.

2.3 Chapter aims

1. Identifying rhizobium genes that have or have not been transferred between species-related categories, based on clanistic analysis.
2. Highlighting genes with remarkable clanistic patterns specific to biological entities as species specificity and/or host specificity.
3. Illustrating the distribution of gene sharing patterns that vary among host plants, replicons, and subspecies.
4. Illustrating the relationship between gene transfer patterns and their genomic function.

2.4 Methods

2.4.1 Sequence data

The gene sequences of *R. leguminosarum* used for this analysis are those that map to the *Rlv*3841 genome. Analysis was carried out on 7,148 genes found in 81 strains of which 72 are Wentworth strains (36 bv. *viciae* and 36 bv. *trifolii* strains), seven Swedish strains (from Dr. Kerstin Huss-Danell), and two Scottish strains (from Dr. Euan James). Five of seven Swedish strains were isolated from *Vicia cracca*, so they belong to symbiovar viciae (VCS*n*). Two of them were isolated from *Trifolium pratense*, so they belong to symbiovar *trifolii* (TPS*n*). One of the Scottish strains came from *Vicia cracca* (VCS6) and another came from *Trifolium pratense* (TPS6), hence these strains belong to symbiovar *viciae* and *trifolii*, respectively. Lists of rhizobium strains and other rhizobium used in this work are shown in Table 2.1 and Table 2.2.

2.4.2 Multiple sequence alignment

In order to achieve phylogenetic analysis, an alignment of each gene was conducted using MUSCLE (MUltiple Sequence Comparison by Log-Expectation) (Edgar, 2004) at the nucleotide level. Elements of the algorithm include fast distance estimation using contiguous subsequence of length k (also known as *k*mer) counting, progressive alignment using a new profile function called the log-expectation score, and refinement using tree-dependent restricted partitioning. To exclude sequences that were too incomplete for reliable phylogeny, the mean sequence length (excluding sequence of *Rlv*3841), was computed for each gene file. Gene sequences with length shorter than the mean length were removed from the analysis. Also, since unrooted trees required at least four strains to construct a tree, then the number of retained strains for a given gene file was considered. The gene files having less than four strains were excluded from the analysis. This step was implemented by bio3d package (Grant *et al.*, 2006).

Table 2.1 List of strains of the two symbiovars (*trifolii*: trx for Wentworth strains, otherwise TPS and *viciae*: vsx for Wentworth strains, otherwise VCS) that are classified into 5 genospecies based on the 305 core genes

Genospecies	trx (*bv. trifolii*)	vsx (*bv.viciae*)	(TRX/VSX) Total
A	trx34 TPS1	VCS1 VCS3 VCS4 VCS5	(2/4) 6
B	trx2 trx12 trx13 trx15 trx18 trx25 trx27 trx31 trx32 trx33	vsx15 vsx18	(10/2) 12
C	TPS6 trx1 trx3 trx5 trx6 trx7 trx10 trx14 trx16 trx17 trx19 trx20 trx21 trx23 trx24 trx26 trx28 trx30 trx35 trx36	VCS6 vsx1 vsx2 vsx3 vsx4 vsx5 vsx6 vsx7 vsx8 vsx9 vsx10 vsx11 vsx14 vsx16 vsx17 vsx19 vsx21 vsx22 vsx23 vsx24 vsx25 vsx26 vsx27 vsx28 vsx29 vsx30 vsx31 vsx32 vsx34 vsx35 vsx36 vsx37 vsx38 vsx39	(20/34) 54
D	TPS5 trx4 trx8 trx11 trx29		(5/0) 5
E	trx9 trx22	VCS2 vsx33	(2/2) 4

Table 2.2 Other *R. leguminosarum* strains sequenced in this study

Genospecies	Strains	Host	Location
A	KHDVB 646.3 (VCS1)	*Vicia cracca*	Lappland, Sorsele, Ammarnäs meadow
	KHDVB 717.3 (TPS1)	*Trifolium pratense*	
	OYAVB 169.1 (VCS3)	*Vicia cracca*	Västerbotten, Umeå, Ängersjö edge of field
	OYAVB 296.5 (VCS4)	*Vicia cracca*	
	OYAVB 349.6 (VCS5)	*Vicia cracca*	Västerbotten, Umeå, Ålidhem roadside
C	S 273.16(VCS6)	*Vicia cracca*	Tayport, Shanwell Farm farm track
	S 272.1 (TPS6)	*Trifolium pratense*	
D	OYAVB 371.3 (TPS5)	*Trifolium pratense*	Västerbotten, Umeå, Ålidhem roadside
E	KHDVB 902.1 (VCS2)	*Vicia cracca*	Lappland, Sorsele, Kraddsele meadow

2.4.3 Phylogenetic analysis

Neighbour-joining (Saitou *et al.*, 1987) is a distance-based tree method. Tree construction is carried out by finding neighbours (which are two strains whose distance leads to the largest reduction in tree length if they are connected). The distance matrix used in this study was obtained from LogDet transformation (Lockhart *et al.*, 1994), which allows for unequal rates of substitution (e.g. comparison between GC rich and AT rich taxa) at different sites. The reliability of inferred phylogenetic trees was derived from 100 bootstrap replicates (Efron *et al.*, 1983). These were implemented in the R language using the Ape package (Paradis *et al.*, 2004), which provides functions for manipulating phylogenetic trees.

2.4.4 Clanistic Analysis

2.4.4.1 Coherent partitions on unrooted phylogenies

Initially, each gene on the unrooted tree is labelled as either native or intruder. Natives are defined as genes from strains that belong to the category of interest, or native category, while intruders are defined genes from strains that belong to an intruder category (alternatively defined as the complement of natives).

Optimal partitions completely separate natives from intruders on the unrooted tree. The number of splits, described by parsimony score (*p-score*) and homogeneity score within partitions of the unrooted tree, was used to derive the optimal partitions with significant signals (Schliep *et al.* 2011).

First, *p-score* was used for describing the number of splits. The *p-score* is the number of cuts required to separate natives from intruders. A single split or a perfect clan gives a *p-score* of 1. A pair of splits or a perfect slice gives a *p-score* of 2. Higher parsimony scores (*p-score* > 2) indicate that there is combination of the natives and intruders in the tree that would require additional partitioning to split them.

Second, the homogeneity of the partitions in the tree (alternately, quantification of the distribution of intruders within partitions) is measured by the equitability index (E), which is derived from the Shannon diversity index (H) (Shannon, 1948).

$$E = H/H' \text{ where } H' = \log n$$

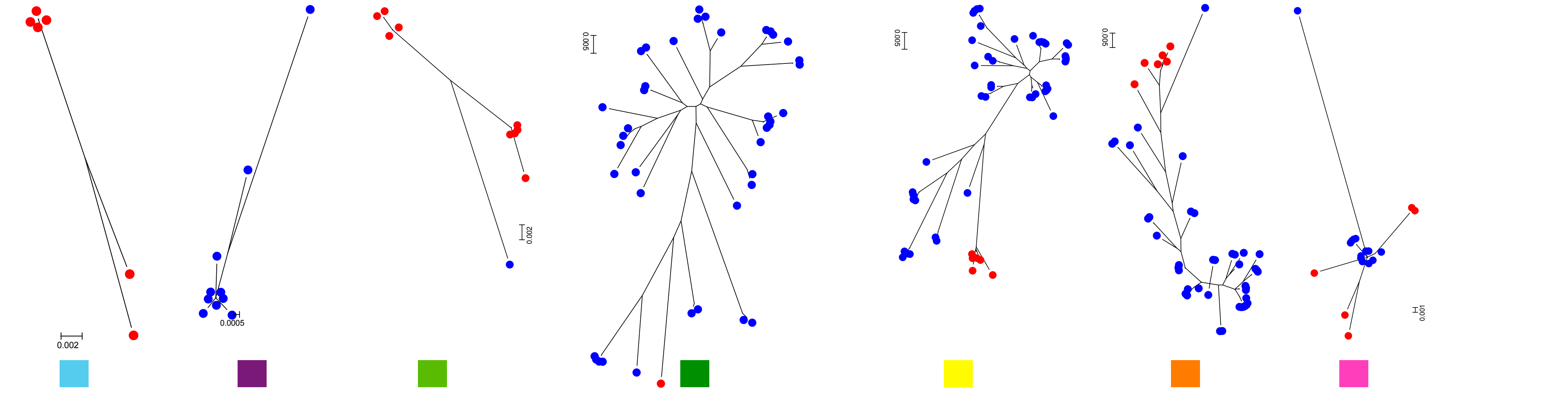
where n is a number of natives of a given category. $-\sum_{i=1}^{k}(p_i \log p_i)$ was denoted the Shannon diversity index (H). The k largest homogeneous clans contain the n natives with relative sizes p_i. A null value of E demonstrated that all intruders of a given category are completely separated into a perfect partition within the smallest complete partition of natives. The larger values of E values, the more dispersion of the intruders. The maximum equitability ($E = 1$) is reached when the n natives split into n partitions (Lapointe *et al.*, 2010).

The optimal partitions with evolutionary significance were derived herein by the R function called *getDiversity* in the Phangorn package (Schliep, 2011).

2.4.4.2 Evolutionary signal detection

Clanistic analysis was performed and obtained coherent partitions (later called clanistic patterns) on each rhizobial gene. The clanistic patterns based on genospecies (Figure 2.4) were herein classified into two classes, viz. "concordant" pattern describing a gene without evidence for HGT, and "discordant" pattern describing a gene with evidence for HGT.

The concordant pattern included trees with natives only (null E and *p-score* of 0) (Figure 2.4 (a)), trees with intruders only (null E and *p-score* of 0) (Figure 2.4 (b)), and trees with a perfect clan (null E and *p-score* of 1) (Figure 2.4 (e)).



(a.) pRL110057| hypothetical protein

(b.) pRL110176| hypothetical protein

(c.) pRL110521| hypothetical protein

(d.) pRL110067| putative glycosyltransferase

(e.) pRL110018| major facilitator subfamily transporter protein

(f.) pRL110077 | ABC transporter permease

(g.) pRL110131 | hypothetical protein

Figure 2.4 Examples of different clanistic patterns found in analysing genes on pRL11, which define *genospecies B* in 🔴 as natives and *complement of genospecies B* in 🔵 as intruders (a.) tree containing only natives (🟦), (b.) tree containing only intruders (🟪), (c.) trees with natives + single intruder (🟩), (d.) trees with intruders + single native (🟩), (e.) tree with perfect clan (🟨), (f.) tree with perfect slices (🟧), and (g.) mélange tree without any kind of perfect group (🟪). All of these patterns were defined using p-score of the tree, clan and slice, respectively. Patterns (c.), (d.), (f.), and (g.) can be explained in terms of phylogeny by invoking at least one HGT.

The natives-only pattern represented candidate genes that might be specific genes of the native category. Genes with intruders-only pattern are absent from the native category. The clan patterns represent genes for which the assigned category can be separated from its complement.

Other clanistic patterns were specified as discordant patterns. Trees with intruders + single native (*E*> 0 and *p-score* of 1) (Figure 2.4 (d)), trees with natives + single intruder (*E*> 0 and *p-score* of 1) (Figure 2.4 (c)), trees with slice (*E*> 0 and *p-score* of 2) (Figure 2.4 (f)), and trees with a mélange of natives and intruders (*E*> 0 and *p-score*>2) (Figure 2.4 (g)) patterns notably show at least one species having evidence for HGT (Schliep *et al.*, 2011).

A similar analysis was also performed based on symbiovar categories rather than genospecies. In this case, the concordant pattern with "no evidence for HGT" should be interpreted as identifying genes that have the same distribution as the symbiosis genes, and hence might be involved in the symbiosis. The discordant pattern with "evidence for HGT" should be interpreted as genes shared in the two symbiovars.

2.4.5 Host specificity and genospecies categories

The analysis categorised the strains either by symbiovar (bv. *trifolii* and bv. *viciae*) or genospecies (A to E), and considered each category in turn as native. For example, if the genospecies B was considered as native then all rhizobium strains belonging to genospecies B were considered as native and strains belonging to others genospecies were intruders (in this case, the intruders were implicitly genospecies A, genospecies C, genospecies D and genospecies E).

Another use of these defined categories was to discover sets of genes that may be transferred together among the five genospecies by defining a discordance index. The discordance index is the number of trees (out of five) using genospecies as native category, which have clanistic patterns representing an evidence of HGT. For example, pRL80044 demonstrated mélange using genospecies A as native, intruders only for using genospecies B, clan for using genospecies E and slice for using genospecies C and D as native. Hence, discordance index of pRL80044 is 3. Fisher's exact test (Routledge, 2005) examined the distribution of clanistic patterns by different native categories and the association between discordance index and each replicon. Fisher's exact test was computed using fisher.test() in R.

2.4.6 Functional Analysis

The relationship between gene transfer and gene function was examined by an enrichment analysis. To determine functions that involved frequently or infrequently transferred genes with statistical significance, each tree was first assigned to a function based on 23 categories of the proteins they encode defined in the COG database (Tatusov *et al.*, 2001) and then classified by distinct native categories and discordance index. Fisher's exact test examined the distribution of clanistic patterns in each functional category by different native categories and the association between discordance index and functional categories.

2.4.7 Computational Resources

The whole procedure was run on an Apple MacBook Pro Intel Core i5 2.3 GHz CPU with 4GB 667 MHz RAM running Mac OS X 10.8.4. R version 2.15.2 was used in this chapter.

2.5 Results

The results on clanistic pattern of each gene by different replicon for each of the two symbiovars and five genospecies are in Figures 2.5-2.12.

2.5.1 Clanistic Patterns Distribution based on Native Categories of two symbiovars

First, the host-specific categories of bv. *viciae* and bv. *trifolii* were assigned as natives (Figure 2.12) and their patterns were compared for genes on different replicons. For example, using bv. *trifolii* as native, the frequency of genes with a perfect clan on pRL7 was compared with the frequency for genes located elsewhere, using Fisher's exact test (Table 2.3).

Table 2.3 Frequency of genes with clan pattern on pRL7 and other replicons when bv. *trifolii* is the native category

		Replicon	
		pRL7	Not pRL7
Clanistic pattern	Clan	9	78
	Not clan	147	6,914

(*p-value* = 0.00011, *two-sided Fisher's exact test*)

Using bv. *trifolii* as native, genes with a perfect clan were overrepresented on pRL7.

Features of the distributions of the clanistic patterns on the two symbiovar categories (Figures 2.5-2.11) can be explained as follows.

The two symbiovars are complementary categories. For example, a gene that demonstrates a pattern of natives only, given bv. *viciae* as native category, will demonstrate a pattern of intruders only when bv. *trifolii* is the native category. Genes with a single-category pattern might be presumed to be host-specific genes of either bv. *viciae* or bv. *trifolii*. However, *Rlv*3841, used as a reference strain for mapping sequence data in this study, is bv. *viciae*. In the case of genes with natives-only pattern for bv. *trifolii* as native, these genes are not really absent in all bv. *viciae* because they are present in *Rlv*3841.

Chromosomal genes are core genes and reflect the core phylogeny. Consequently mélange patterns for symbiovar as native are overrepresented in chromosomal genes (🟥 in Figure 2.12).

In contrast to clanistic patterns found in chromosome, clanistic patterns of plasmid genes may be either concordant or discordant. Genes having concordant patterns of clanistics are present in other replicons. Salient features of these genes are as follows.

Because the reference strain, *Rlv*3841, is bv. *viciae*, genes with a natives-only pattern for bv. *viciae* as native were highlighted to be bv. *viciae* specific genes (🟦 in Figure 2.12). The *bvs* genes (*biovar viciae* specific genes) are five genes on pRL8 (pRL80073 - pRL80077) whose clanistic patterns are illustrated in Figure 2.13. The *nod*, *nif*, and *fix* genes on pRL10 are also identified by clanistic analysis to be bv. *viciae* specific genes. A list of detected bv. *viciae* specific genes on pRL8 and pRL10 is given in Table 2.4. Genes probably transferred between the two symbiovars were found on pRL7, pRL9, pRL10, pRL11, and pRL12.

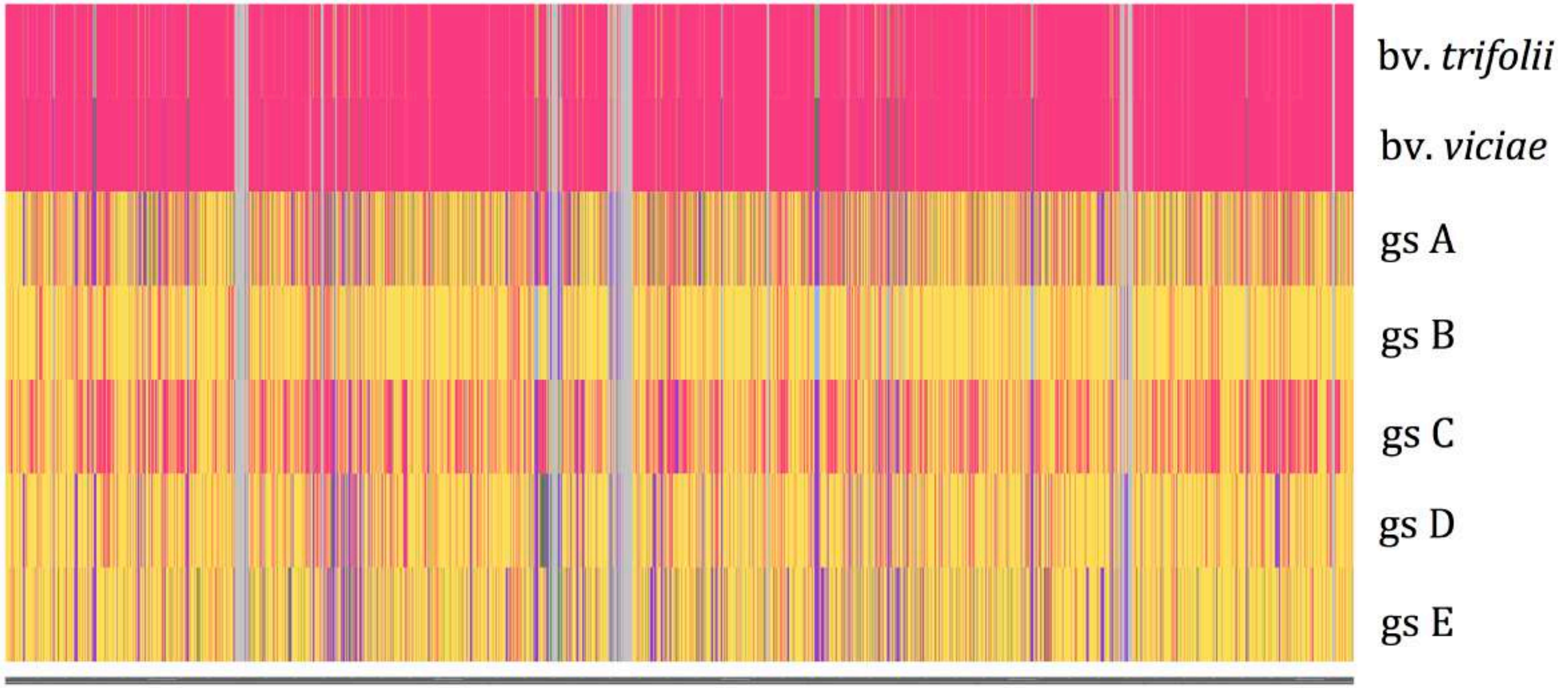


Figure 2.5 Clanistic patterns of chromosomal genes for different native categories. Rows represent defined native categories, and columns represent chromosomal genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ◼ genes excluded from analysis).

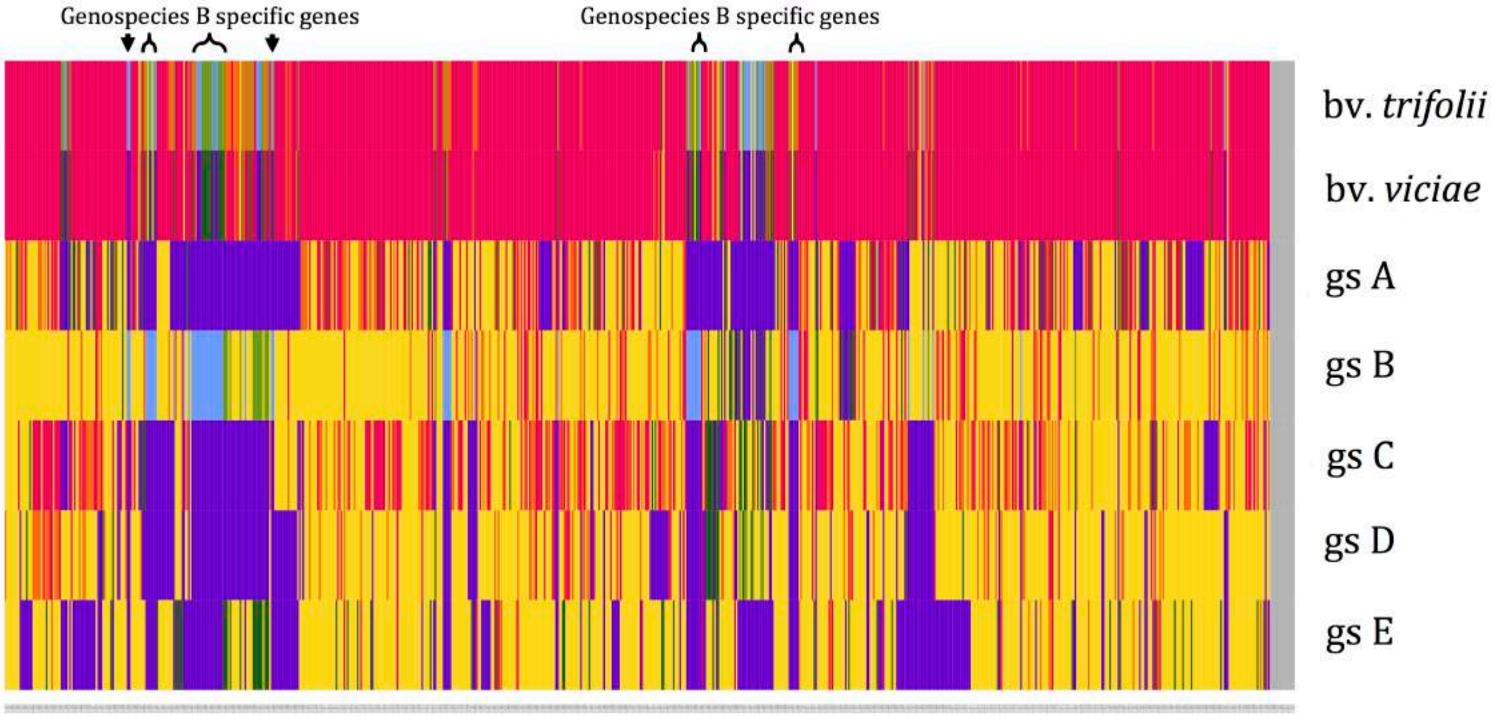


Figure 2.6 Clanistic patterns of pRL12 genes for different native categories. Rows represent defined native categories, and columns represent pRL12 genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ◼ genes excluded from analysis).

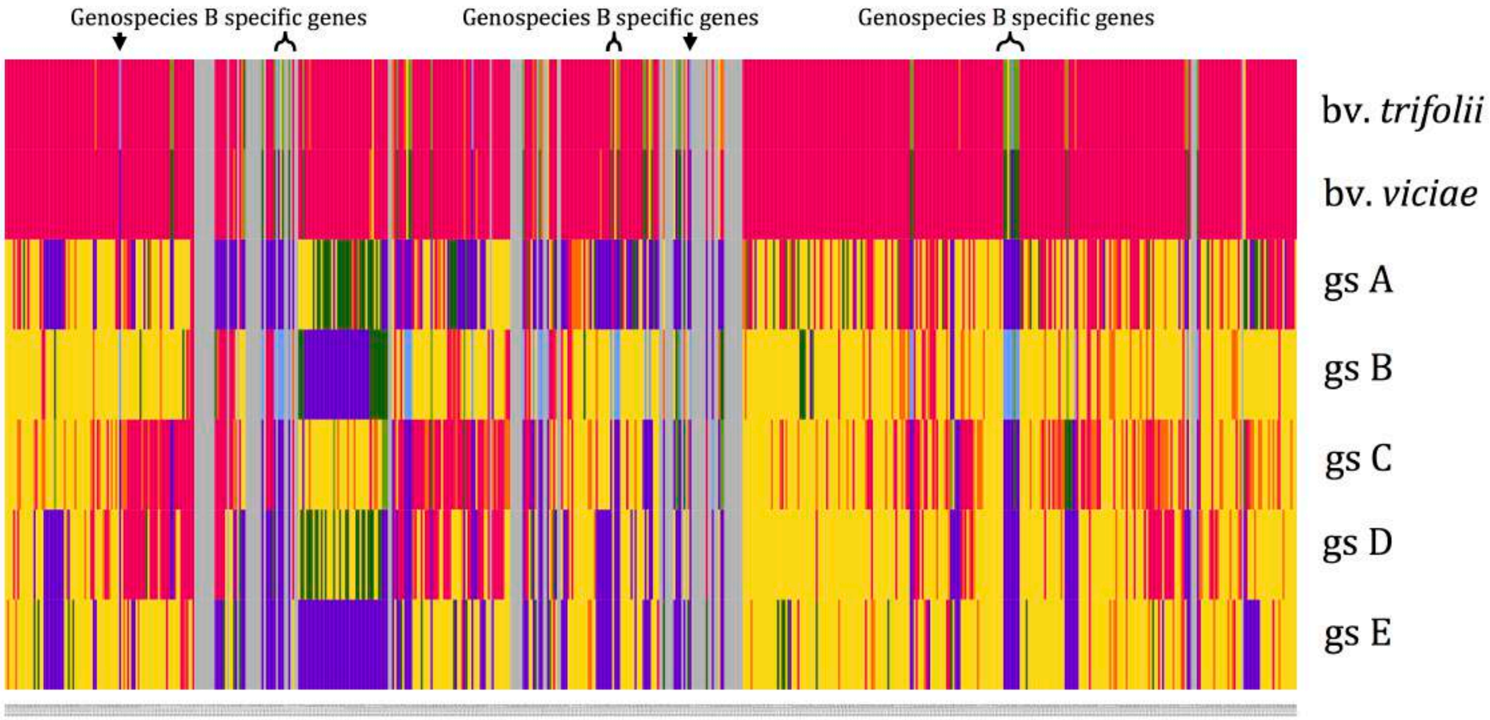


Figure 2.7 Clanistic patterns of pRL11 genes for different native categories. Rows represent defined native categories, and columns represent pRL11 genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ■ genes excluded from analysis).

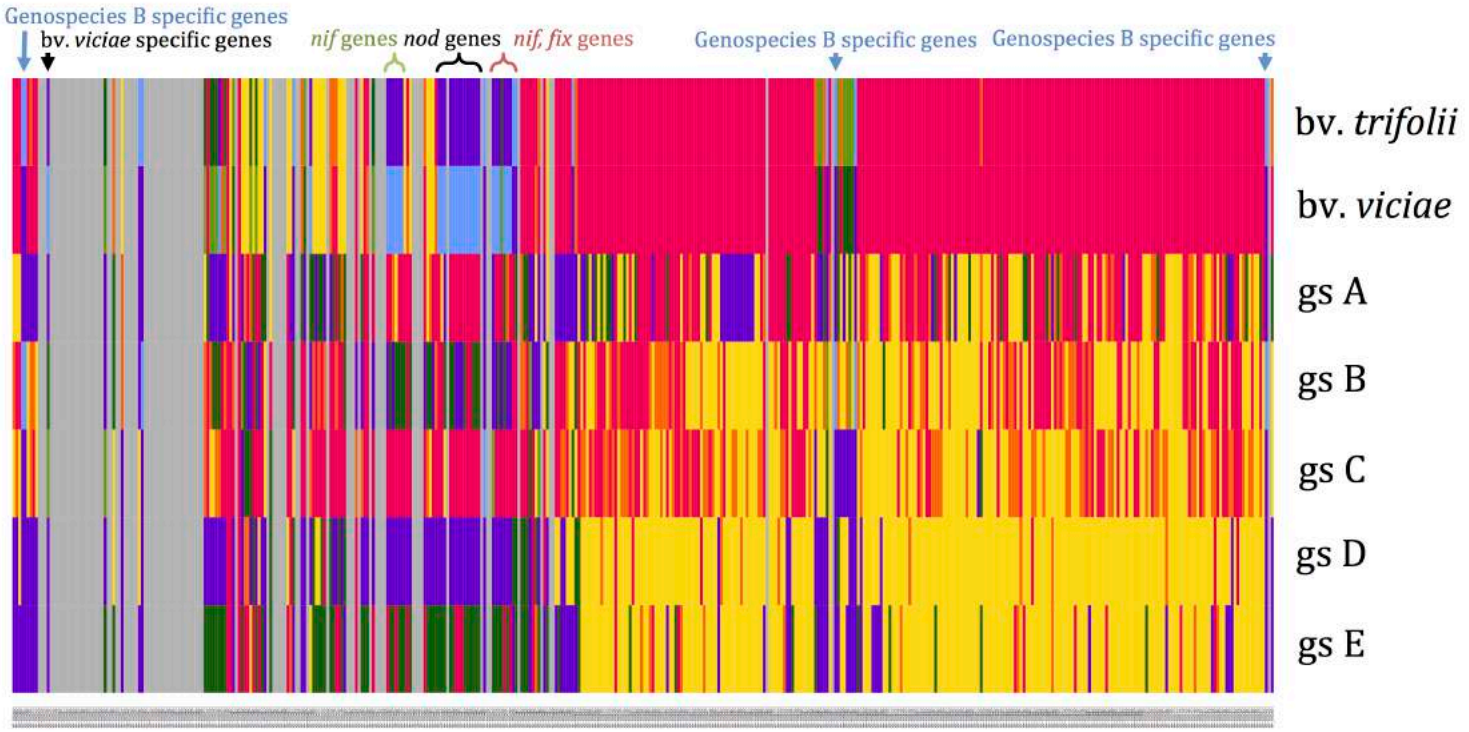


Figure 2.8 Clanistic patterns of pRL10 genes for different native categories. Rows represent defined native categories, and columns represent pRL10 genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ◼ genes excluded from analysis).

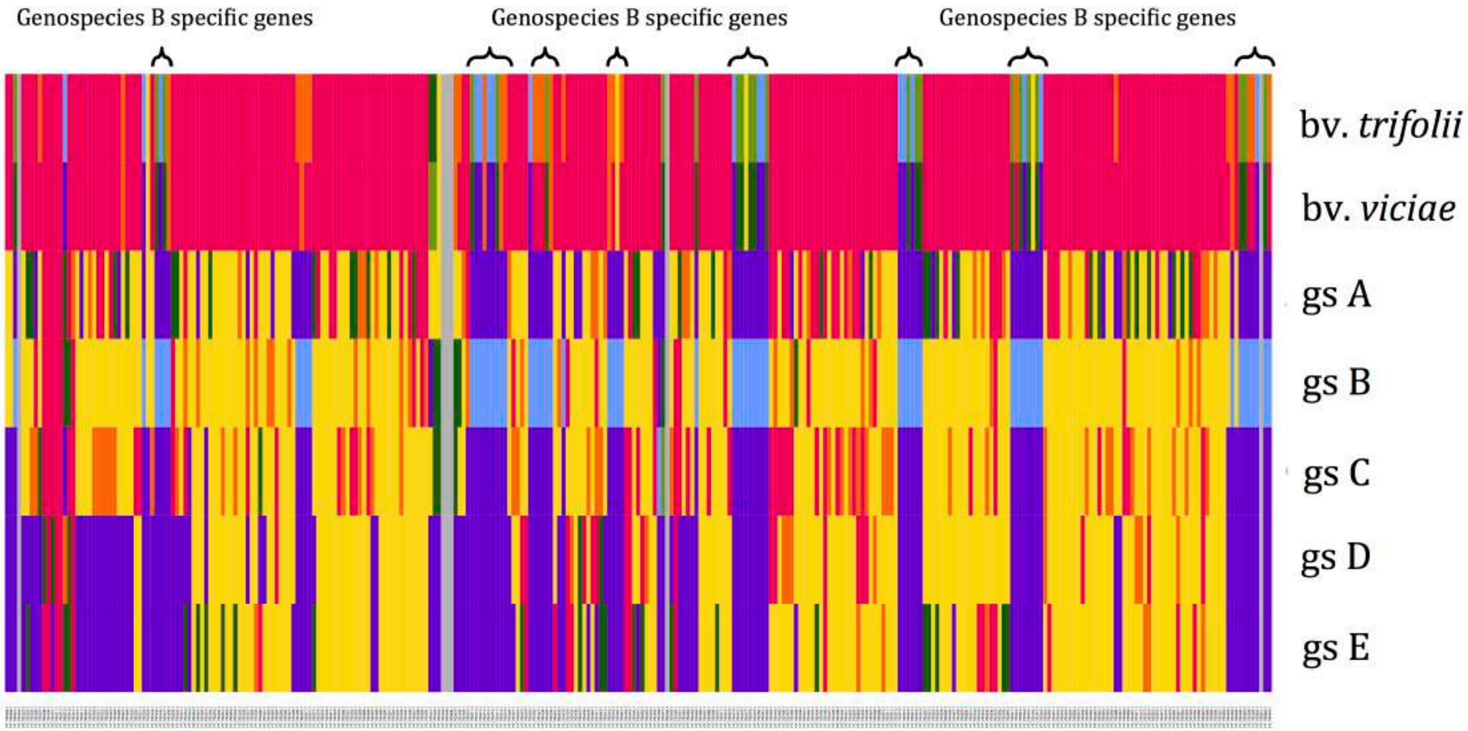


Figure 2.9 Clanistic patterns of pRL9 genes for different native categories. Rows represent defined native categories, and columns represent pRL9 genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis).

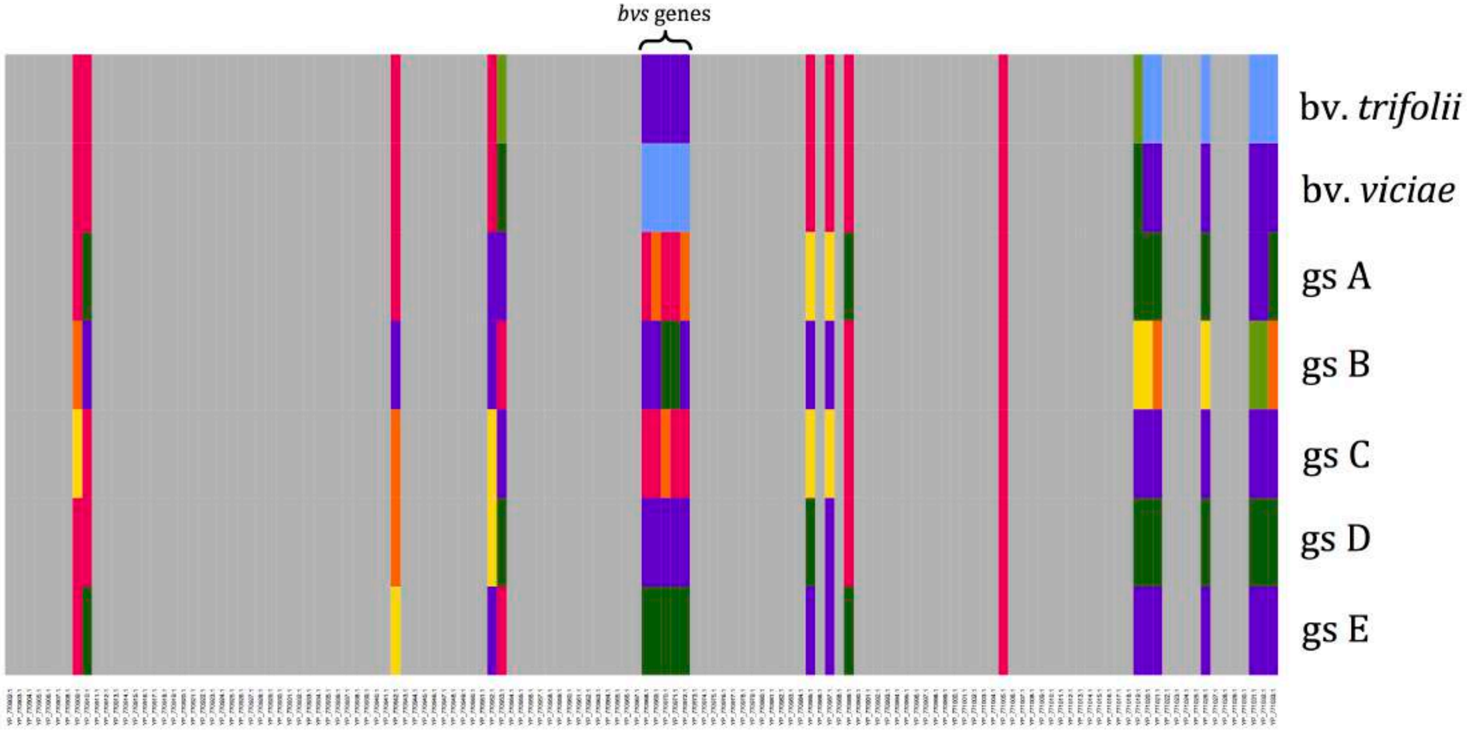


Figure 2.10 Clanistic patterns of pRL8 genes for different native categories. Rows represent defined native categories, and columns represent pRL8 genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis).

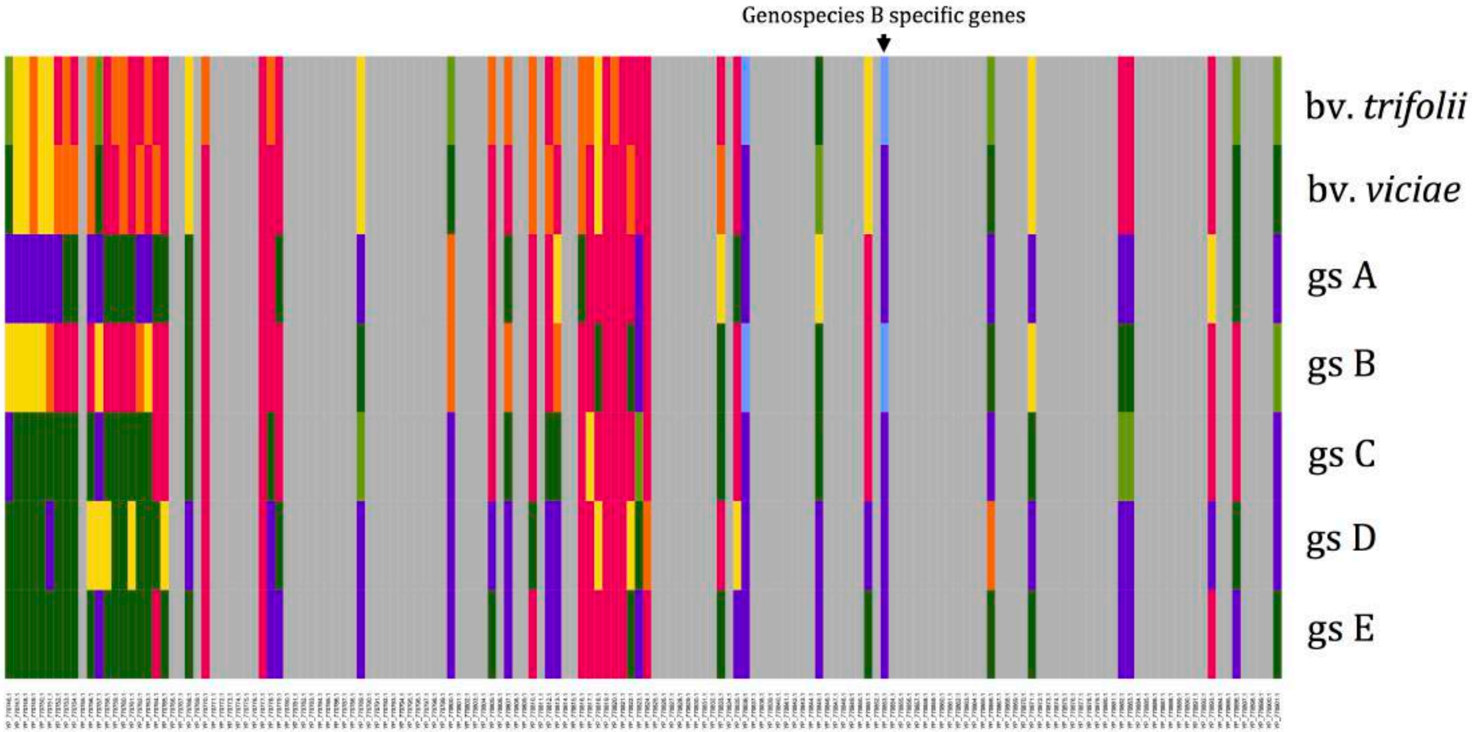


Figure 2.11 Clanistic patterns of pRL7 genes for different native categories. Rows represent defined native categories, and columns represent pRL7 genes arranged according their locus tag (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruders, 🟩 trees with intruders+ single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis).

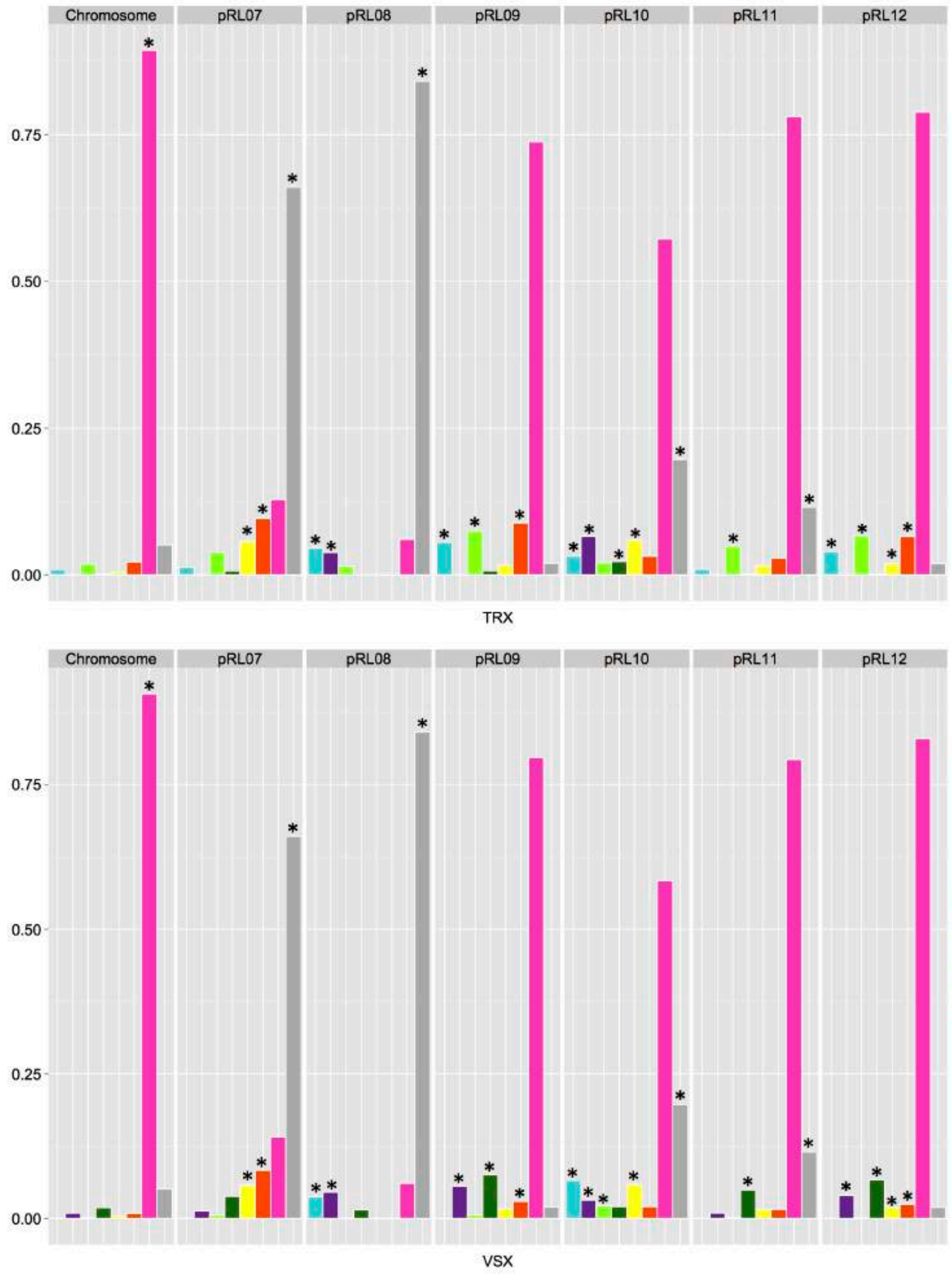
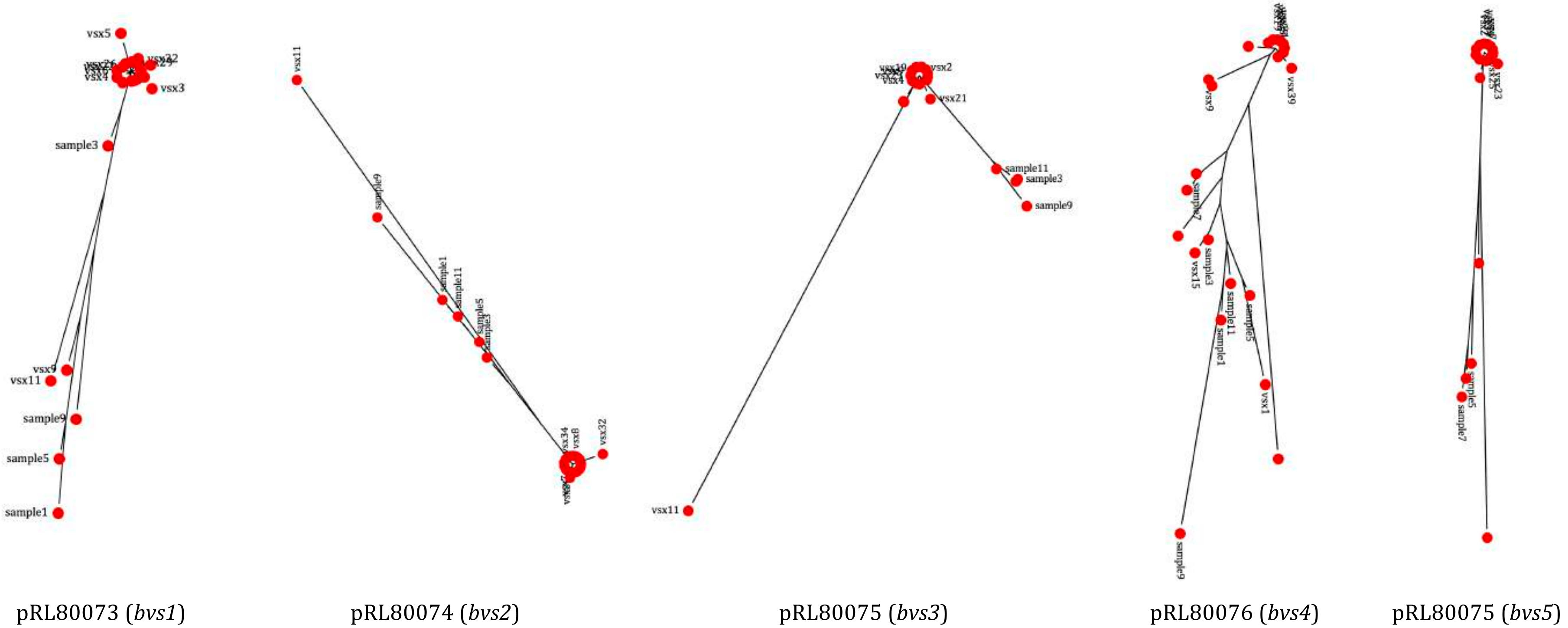


Figure 2.12 Distribution of the rhizobium gene trees when strains are classified into the two symbiovars. Each plot indicates, for a given category, the fraction of gene trees associated with the assigned category (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ■ genes excluded from analysis). Fisher's exact tests identified (*) a category which is overrepresented relative to its abundance on other replicons.

Table 2.4 List of genes on pRL8 and pRL10 with natives-only clanistic patterns with *bv. viciae* (vsx) as native category. Locus tags of biovar-specific genes are in bold.

Locus tag	Gene symbol	Protein accession	Annotated function	COG function
pRL80073	*bvs1*	YP_770968.1	Selenocysteine lyase/Cysteine desulfurase	Amino acid transport and metabolism
pRL80074	*bvs2*	YP_770969.1	DNA-binding transcriptional regulator, LysR family	Transcription
pRL80075	*bvs3*	YP_770970.1	Enamine deaminase RidA, house cleaning of reactive enamine intermediates, YjgF/YER057c/UK114 family	Translation, ribosomal structure and biogenesis, and Defense mechanisms
pRL80076	*bvs4*	YP_770971.1	Predicted amidohydrolase	General function prediction only
pRL80077	*bvs5*	YP_770972.1	Periplasmic DMSO/TMAO reductase YedYZ, molybdopterin-dependent catalytic subunit	Energy production and conversion
pRL100158	*nifN*	YP_770436.1	Nitrogenase molybdenum-iron protein, alpha and beta chains	Energy production and conversion
pRL100159	*nifE*	YP_770437.1	Nitrogenase molybdenum-iron protein, alpha and beta chains	Energy production and conversion
pRL100160	*nifK*	YP_770438.1	Nitrogenase molybdenum-iron protein, alpha and beta chains	Energy production and conversion
pRL100161	*nifD*	YP_770439.1	Nitrogenase molybdenum-iron protein, alpha and beta chains	Energy production and conversion
pRL100162	*nifH*	YP_770440.1	Nitrogenase subunit NifH, an ATPase	General function prediction only
pRL100175	*nodO*	YP_770454.1	-	-
pRL100177	-	YP_770455.1	-	-
pRL100179	*nodN*	YP_770457.1	Acyl dehydratase	Lipid transport and metabolism
pRL100180	*nodM*	YP_770458.1	Glucosamine 6-phosphate synthetase, contains amidotransferase and phosphosugar isomerase domains	Cell wall/membrane/envelope biogenesis
pRL100181	*nodL*	YP_770459.1	Acetyltransferase (isoleucine patch superfamily)	General function prediction only
pRL100182	*nodE*	YP_770460.1	3-oxoacyl-(acyl-carrier-protein) synthase	Lipid transport and metabolism , and Q

Locus tag	Gene symbol	Protein accession	Annotated function	COG function
pRL100183	*nodF*	YP_770461.1	Acyl carrier protein	Lipid transport and metabolism, and Q
pRL100184	*nodD*	YP_770462.1	DNA-binding transcriptional regulator, LysR family	Transcription
pRL100185	*nodA*	YP_770463.1	-	-
pRL100186	*nodB*	YP_770464.1	Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family	Carbohydrate transport and metabolism, and Cell wall/membrane/envelope biogenesis
pRL100187	*nodC*	YP_770465.1	Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine synthase	Cell wall/membrane/envelope biogenesis
pRL100188	*nodI*	YP_770466.1	ABC-type multidrug transport system, ATPase component	Defense mechanisms
pRL100189	*nodJ*	YP_770467.1	ABC-type polysaccharide/polyol phosphate export permease	Carbohydrate transport and metabolism, and Cell wall/membrane/envelope biogenesis
pRL100195	*nifB*	YP_770473.1	Radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family	General function prediction only
pRL100196	*nifA*	YP_770474.1	Transcriptional regulator containing GAF, AAA-type ATPase, and DNA-binding Fis domains	Transcription, and Signal transduction mechanisms
pRL100198	*fixC*	YP_770476.1	Dehydrogenase (flavoprotein)	Energy production and conversion
pRL100199	*fixB*	YP_770477.1	Electron transfer flavoprotein, alpha subunit	Energy production and conversion
pRL100200	*fixA*	YP_770478.1	Electron transfer flavoprotein, alpha and beta subunits	Energy production and conversion



pRL80073 (*bvs1*) pRL80074 (*bvs2*) pRL80075 (*bvs3*) pRL80076 (*bvs4*) pRL80075 (*bvs5*)

Figure 2.13 Clanistic patterns of the five *bvs* genes. All five genes exhibited unrooted trees with natives-only pattern for bv. *viciae* as native category.

2.5.2 Clanistic pattern distribution based on native categories of five genospecies

In general, the fraction of concordant clanistic patterns is higher for genospecies than for symbiovar. Features of the distributions of the clanistic patterns when each genospecies is, in turn, taken as native (Figure 2.14-2.16) can be described as follows.

Chromosomal genes of all five genospecies show overrepresented proportions of trees with a perfect clan. A majority of chromosomal genes are not transferred among the five genospecies and are well conserved in each genospecies. Some of chromosomal genes with evidence for HGT represent clusters of genes transferred between genospecies and reveal that chromosomal genes can behave like accessory genes transferred among the five genospecies.

Genes on pRL8 represents a significant fraction of patterns without evidence for HGT as they have an intruders-only pattern for genospecies B (🟪 in Figure 2.14), implying that these genes are absent in genospecies B. However, *Rlv*3841, used as a reference strain for mapping sequence data in this study, is in genospecies B and does, of course, have the genes. Then genes with intruders pattern only in this case are not really absent in genospecies B.

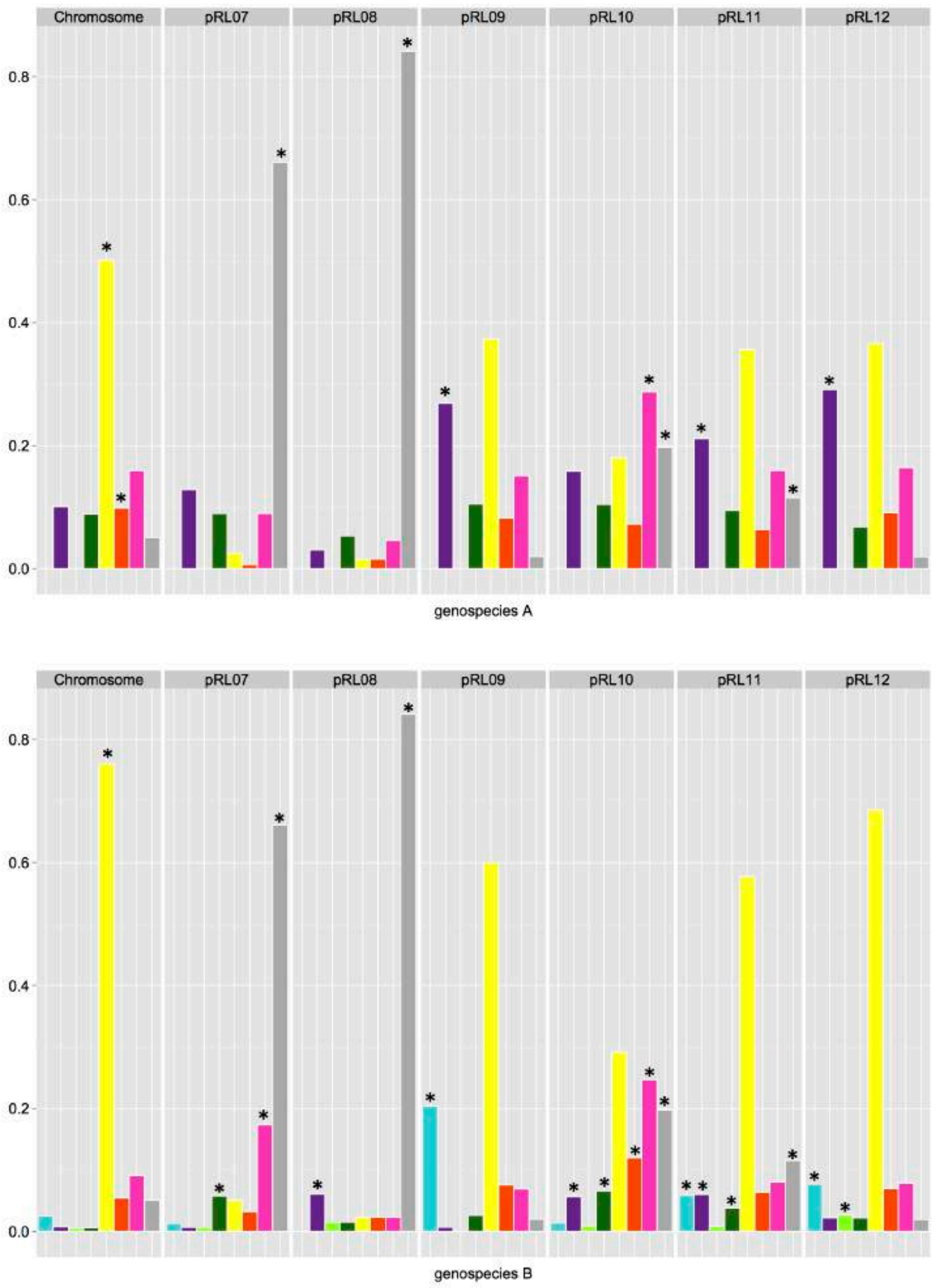


Figure 2.14 Distribution of the rhizobium gene trees in genospecies A and B based on different tree patterns. Each plot indicates, for a given category, the fraction of gene trees associated with the assigned category (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis). Fisher's exact tests identified (*) a category which is overrepresented relative to its abundance on other replicons.

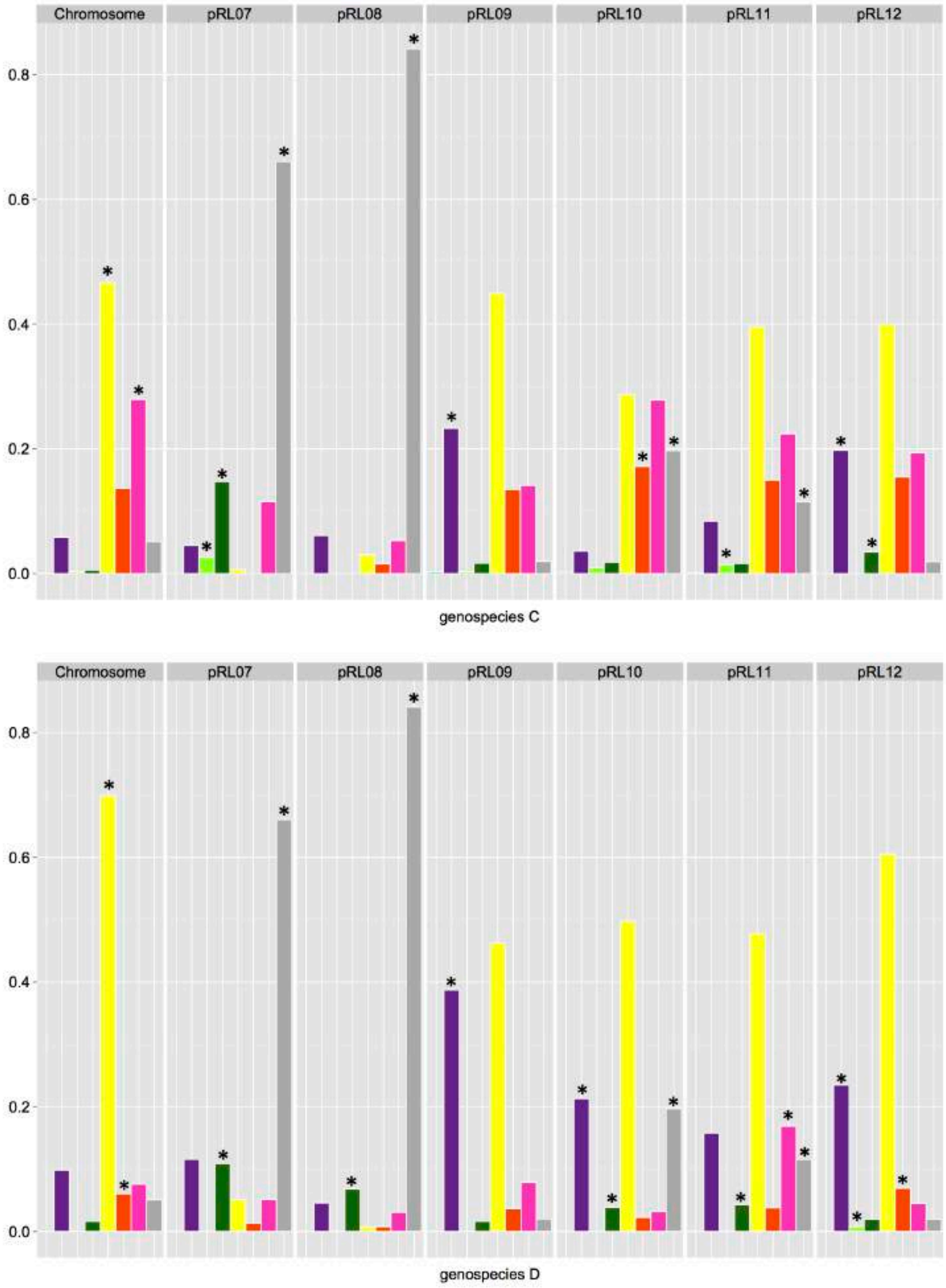


Figure 2.15 Distribution of the rhizobium gene trees in genospecies C and D based on different tree patterns. Each plot indicates, for a given category, the fraction of gene trees associated with the assigned category (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ◼ genes excluded from analysis). Fisher's exact tests identified (*) a category which is overrepresented relative to its abundance on other replicons.

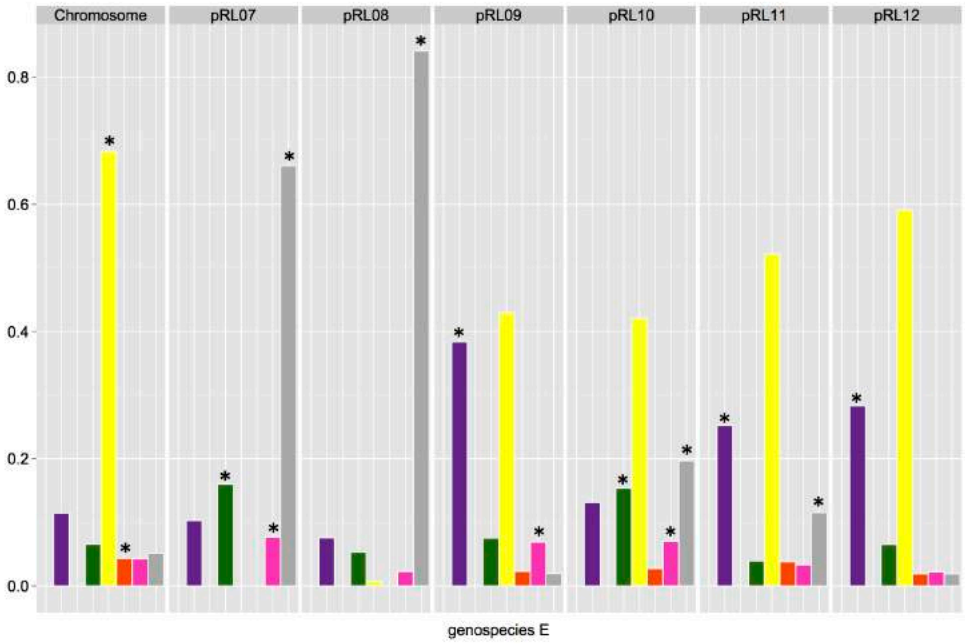


Figure 2.16 Distribution of the rhizobium gene trees in genospecies E based on different tree patterns. Each plot indicates, for a given category, the fraction of gene trees associated with the assigned category (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ■ genes excluded from analysis). Fisher's exact tests identified (*) a category which is overrepresented relative to its abundance on other replicons.

pRL9 genes strongly exhibit patterns without evidence for HGT in all five genospecies. There is a cluster of genes in only genospecies B (🟦 in Figure 2.14). Table 2.5 and 2.6 represent lists of genes carried by strains in genospecies B.

Table 2.5 List of genes with natives-only clanistic patterns on pRL9, pRL11, and pRL12 with genospecies B as native category. Locus tags and other information of 26 genes held by all 12 members of genospecies B.

Locus tag	Gene symbol	Protein accession	Annotated function
pRL90043	-	YP_765336.1	Multidrug resistance efflux pump
pRL90045	-	YP_765338.1	ABC-type multidrug transport system, permease component
pRL90121	-	YP_765413.1	Predicted ATPase
pRL90122	-	YP_765414.1	DNA-binding transcriptional regulator, LacI/PurR family
pRL90124	-	YP_765416.1	ABC-type sugar transport system, permease component

Locus tag	Gene symbol	Protein accession	Annotated function
pRL90125	-	YP_765417.1	ABC-type glycerol-3-phosphate transport system, permease component
pRL90126	-	YP_765418.1	ABC-type glycerol-3-phosphate transport system, periplasmic component
pRL90188	-	YP_765475.1	Cupin domain protein related to quercetin dioxygenase
pRL90189	-	YP_765476.1	Predicted dehydrogenase
pRL90190	-	YP_765477.1	ABC-type glycerol-3-phosphate transport system, permease component
pRL90192	-	YP_765479.1	ABC-type glycerol-3-phosphate transport system, periplasmic component
pRL90231	-	YP_765518.1	ABC-type transport system, periplasmic component
pRL90232	-	YP_765519.1	Arylsulfatase A or related enzyme
pRL90255	-	YP_765541.1	Glycine cleavage system T protein (aminomethyltransferase)
pRL90256	-	YP_765542.1	5,10-methylenetetrahydrofolate reductase
pRL90315	-	YP_765597.1	Predicted ATPase
pRL90317	-	YP_765599.1	Predicted enzyme related to lactoylglutathione lyase
pRL90318	*ohr*	YP_765600.1	Organic hydroperoxide reductase OsmC/OhrA
pRL110057	-	YP_771090.1	-
pRL110132	-	YP_771166.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL110198	-	YP_771232.1	-
pRL110199	-	YP_771233.1	-
pRL110301	-	YP_771334.1	-
pRL110302	-	YP_771335.1	-
pRL120089	-	YP_764606.1	-
pRL120500	-	YP_765005.1	TRAP-type mannitol/chloroaromatic compound transport system, periplasmic component

Table 2.6 List of genes with natives-only clanistic patterns on pRL7, pRL9, pRL10, pRL11, and pRL12 with genospecies B as native category. Locus tags and other information of 48 genes held by 6-11 members of genospecies B.

Locus tag	Gene symbol	Protein accession	Annotated function
pRL70123	-	YP_770853.1	Plasmid stabilization system protein ParE
pRL90041	*groEL*	YP_765335.1	Chaperonin GroEL (HSP60 family)
pRL90119	-	YP_765411.1	DNA-binding transcriptional regulator, LysR family
pRL90120	-	YP_765412.1	Uncharacterized conserved protein YurZ, alkylhydroperoxidase/carboxymuconola ctone decarboxylase family

Locus tag	Gene symbol	Protein accession	Annotated function
pRL90157	-	YP_765446.1	-
pRL90257	-	YP_765543.1	DNA-binding transcriptional regulator, GntR family
pRL90259	-	YP_765545.1	Branched-chain amino acid ABC-type transport system, permease component
pRL90314	-	YP_765596.1	Tryptophan-rich sensory protein (mitochondrial benzodiazepine receptor homolog)
pRL100005	-	YP_770307.1	Uncharacterized protein, contains PIN domain
pRL100006	-	YP_770308.1	Uncharacterized protein
pRL100316	-	YP_770592.1	-
pRL100468	-	YP_770743.1	-
pRL110134	-	YP_771168.1	NADPH:quinone reductase or related Zn-dependent oxidoreductase
pRL110135	-	YP_771169.1	Phenylpyruvate tautomerase PptA, 4-oxalocrotonate tautomerase family
pRL110139	-	YP_771173.1	Predicted dehydrogenase
pRL110338	-	YP_771370.1	-
pRL110494	-	YP_771528.1	-
pRL110497	-	YP_771531.1	-
pRL110585	-	YP_771619.1	-
pRL120075	*stbC*	YP_764592.1	Plasmid stability protein
pRL120076	*stbB*	YP_764593.1	Predicted nucleic acid-binding protein, contains PIN domain
pRL120086	-	YP_764603.1	Phage shock protein A
pRL120092	-	YP_764609.1	Glutathionylspermidine synthase
pRL120103	-	YP_764620.1	-
pRL120118	-	YP_764633.1	Predicted oxidoreductase
pRL120119	-	YP_764634.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL120120	-	YP_764635.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL120121	-	YP_764636.1	Dihydroorotase or related cyclic amidohydrolase
pRL120123	-	YP_764638.1	Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family
pRL120124	-	YP_764639.1	Nucleoside-diphosphate-sugar epimerase
pRL120125	-	YP_764640.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL120126	-	YP_764641.1	Dihydroorotase or related cyclic amidohydrolase
pRL120127	-	YP_764642.1	-
pRL120128	-	YP_764643.1	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component

Locus tag	Gene symbol	Protein accession	Annotated function
pRL120129	-	YP_764644.1	ABC-type dipeptide/oligopeptide/nickel transport system, permease component
pRL120130	-	YP_764645.1	ABC-type dipeptide/oligopeptide/nickel transport system, permease component
pRL120132	-	YP_764647.1	Transcriptional regulator GlxA family, contains an amidase domain and an AraC-type DNA-binding HTH domain
pRL120133	*stbB*	YP_764648.1	Predicted nucleic acid-binding protein, contains PIN domain
pRL120134	*stbC*	YP_764649.1	Plasmid stability protein
pRL120168	-	YP_764680.1	DNA-binding transcriptional regulator, LysR family
pRL120428	-	YP_764935.1	DNA-binding transcriptional regulator, MurR/RpiR family, contains HTH and SIS domains
pRL120429	-	YP_764936.1	Asp/Glu/hydantoin racemase
pRL120430	-	YP_764937.1	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
pRL120433	-	YP_764940.1	ABC-type dipeptide/oligopeptide/nickel transport system, permease component
pRL120434	-	YP_764941.1	ABC-type transport system, periplasmic component
pRL120497	-	YP_765002.1	-
pRL120498	-	YP_765003.1	TRAP-type mannitol/chloroaromatic compound transport system, large permease component
pRL120499	-	YP_765004.1	TRAP-type mannitol/chloroaromatic compound transport system, small permease component

Contemplating genospecies B, pRL11 and pRL12 genes exhibit the clanistic pattern without evidence for HGT as natives-only pattern (🟦 in Figure 2.14). Those genes were carried by strains genospecies B. These genes may be considered as genospecies B specific genes (Table 2.4 and 2.5).

Discordance index was employed to clarify the degrees of gene transfer within the five genospecies by computing the discordance index from results on clanistic pattern number of recombinant species to number of excluded genes from analysis of each replicon (Figure 2.17).

For example, a null hypothesis, tested by Fisher's exact test, was that the distribution of genes on pRL9 with discordance index of 0 was evenly distributed within five genospecies. Under the null hypothesis, a 2×2 contingency table was drawn (Table 2.7).

Table 2.7 Frequency of genes with discordance index of 0 on pRL9 and other replicons under the defined hypothesis.

		Replicon	
		pRL9	Not pRL9
Discordance index	0	126	1,883
	> 0	179	4,960

(*p-value* ≤0.0001, *two-sided Fisher's exact test*)

Genes with discordance index of 0 were overrepresented on pRL9.

Genes with discordance index of 0 were found on the chromosome, pRL9, and pRL12 (🟥 Figure 2.17) because chromosomal genes have a large number of genes with clans, which can distinguish the native genospecies from the other genospecies, while pRL9 and pRL12 carried genes specific to genospecies B.

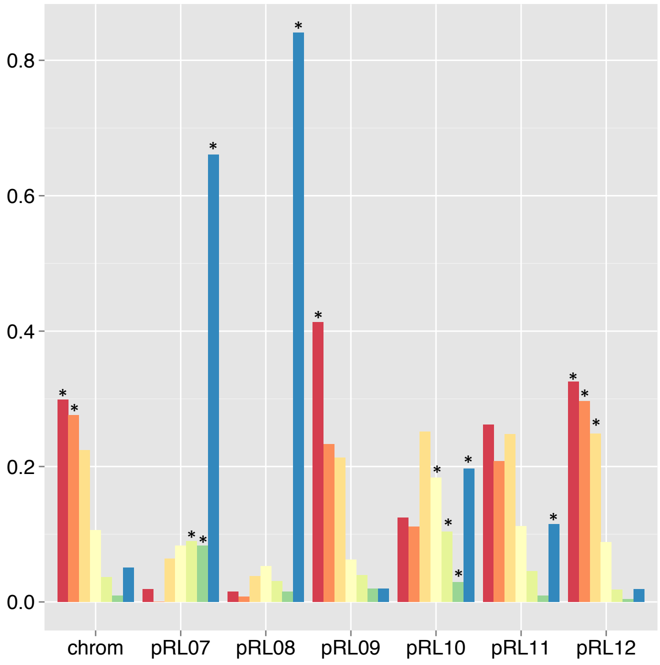


Figure 2.17 Distribution of genes in each replicon, classified by discordance index (discordance index of 0 🟥, 1 🟧, 2 🟨, 3 🟨, 4 🟩, 5 🟩, and 🟦 genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

Genes with discordance index of at least 1 were present in significant numbers on every replicon apart from pRL8, pRL9, and pRL11. These genes may be transferred between one genospecies, which was considered as native, to at least one of the other four genospecies. These implied that HGT within five genospecies could be found on every type of replicon in the genome.

Genes excluded from the analysis represented significant fractions on pRL7, pRL8, pRL10 and pRL11 (🟦 in Figure 2.17). This situation can be described by genome architecture that pRL8, pRL7 and pRL10 are plasmids, which genes are present in some strain and a number of strains carrying these genes not qualifying the assumption of analysing method, similarly to a fraction of excluded genes of pRL11, which is a chromid.

2.5.3 Overrepresented functional categories in clanistic patterns using the two symbiovars native categories

Plots of the two symbiovars viz. bv. *trifolii* (Figure 2.18) and bv. *viciae* (Figure 2.19) defined as native category and functional categories reveal that genes with evidence for HGT were found in almost all 22 functional categories.

Clanistic patterns of genes associated with their function were hypothesised and investigated by different replicons. For example, a null hypothesis that genes with native-only pattern in function category Q (Secondary metabolites biosynthesis, transport and catabolism) evenly distributed using genospecies B as native was established (Table 2.8).

Table 2.8 Frequency of genes with natives-only pattern in category Q and other categories under the defined hypothesis.

		Clanistic pattern	
		Native only	Not native only
Function	Q	14	172
	Not Q	138	7,470

(*p-value* = 0.02302, *two-sided Fisher's exact test*)

It was concluded that the genes with native-only pattern were overrepresented in category Q.

Symbiovar-specific genes were overrepresented in inorganic transport and metabolism. Of 343 genes involving in inorganic transport and metabolism, 5 genes, *nifNEKDH* (pRL100158-pRL100162) on pRL10, have intruders-only pattern for bv. *trifolii* as native (🟪 in Figure 2.18) and natives-only pattern for bv. *viciae* as native (🟦 in Figure 2.19).

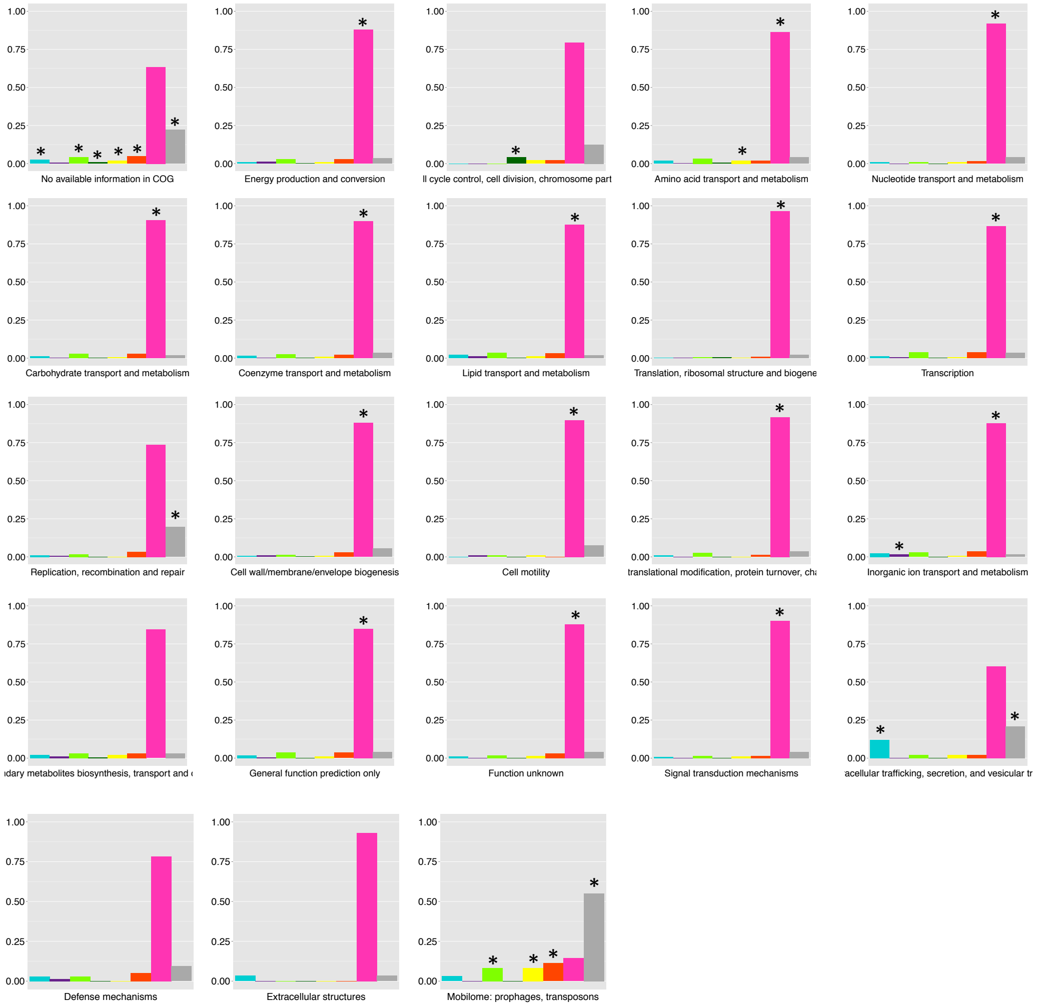


Figure 2.18 Distribution of gene trees in 22 functional categories based on different tree patterns, given bv. *trifolii* as a native category. Each graph describes the percentage of gene trees associated with a given COG function (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ■ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

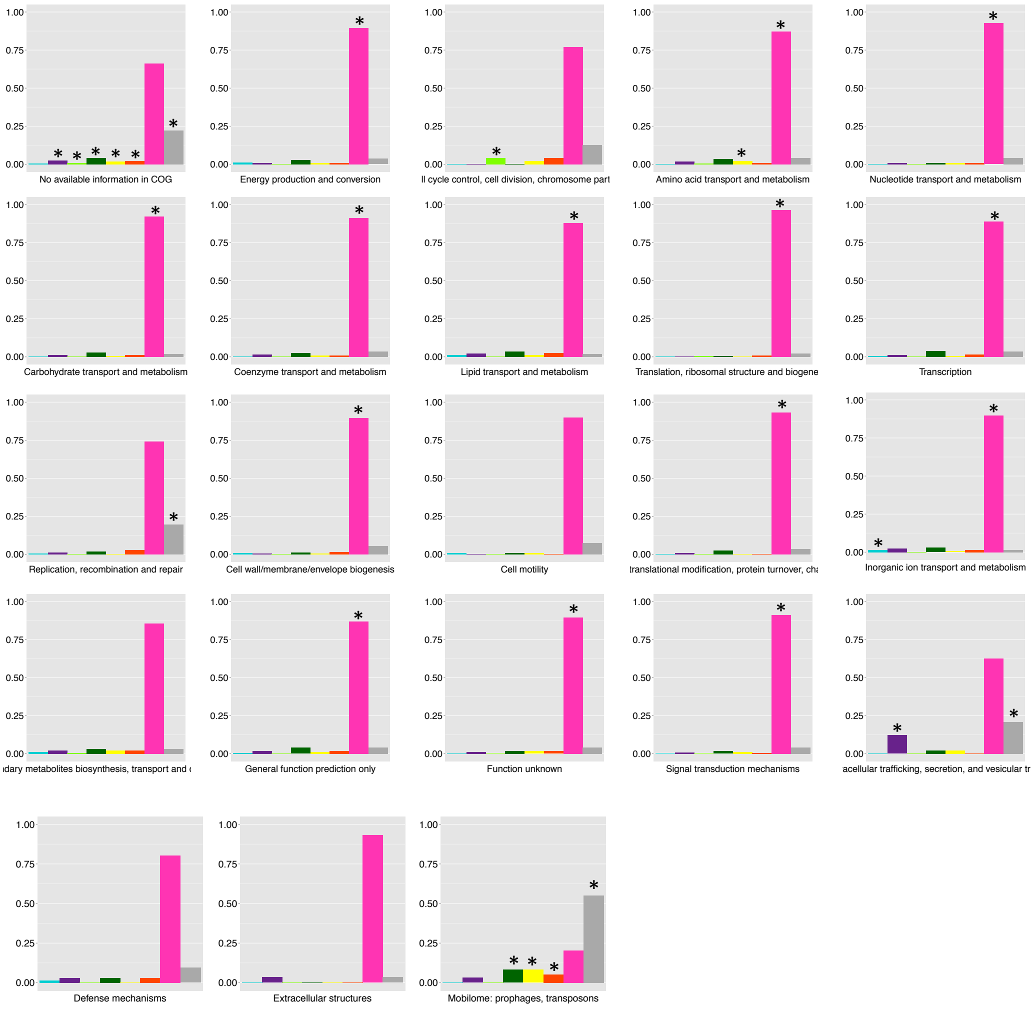


Figure 2.19 Distribution of gene trees in 22 functional categories based on different tree patterns, given symbiovar *viciae* as a native category. Each graph describes the percentage of gene trees associated with a given COG function. (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with native + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ■ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

2.5.4 Overrepresented functional categories in clanistic patterns using native categories as five genospecies

Plots of functional categories when each of the five genospecies was defined as native show the proportion of genes with and without evidence of HGT associated with a given COG (Figure 2.20–2.24).

The five genospecies (Figure 2.20-2.24) showed gene trees having evidence of HGT were overrepresented in amino acid transport and metabolism, carbohydrate transport and metabolism, inorganic ion transport and metabolism, translation, ribosomal structure and biogenesis, replication, recombination and repair, intracellular trafficking, secretion, and vesicular transport, mobilome: prophages, transposons, posttranslational modification, protein turnover, chaperones, secondary metabolites biosynthesis, transport and catabolism, signal transduction mechanisms, energy production and conversion, cell motility, extracellular structures, general function prediction only, unknown function, and genes with no information in COG.

Genes with natives-only pattern (🟦 in Figure 2.21) were considered to be candidates of genospecies B specific genes. The genes were predominantly involved in secondary metabolites biosynthesis, transport and catabolism, general function prediction only, mobilome: prophages, transposons, and as genes with no available information in COG.

For given different native categories, genes excluded from the analysis (■ in Figure 2.18-2.24) were overrepresented in replication, recombination and repair, intracellular trafficking, secretion, and vesicular transport, mobilome: prophages, transposons, and as genes with no available information in COG database.

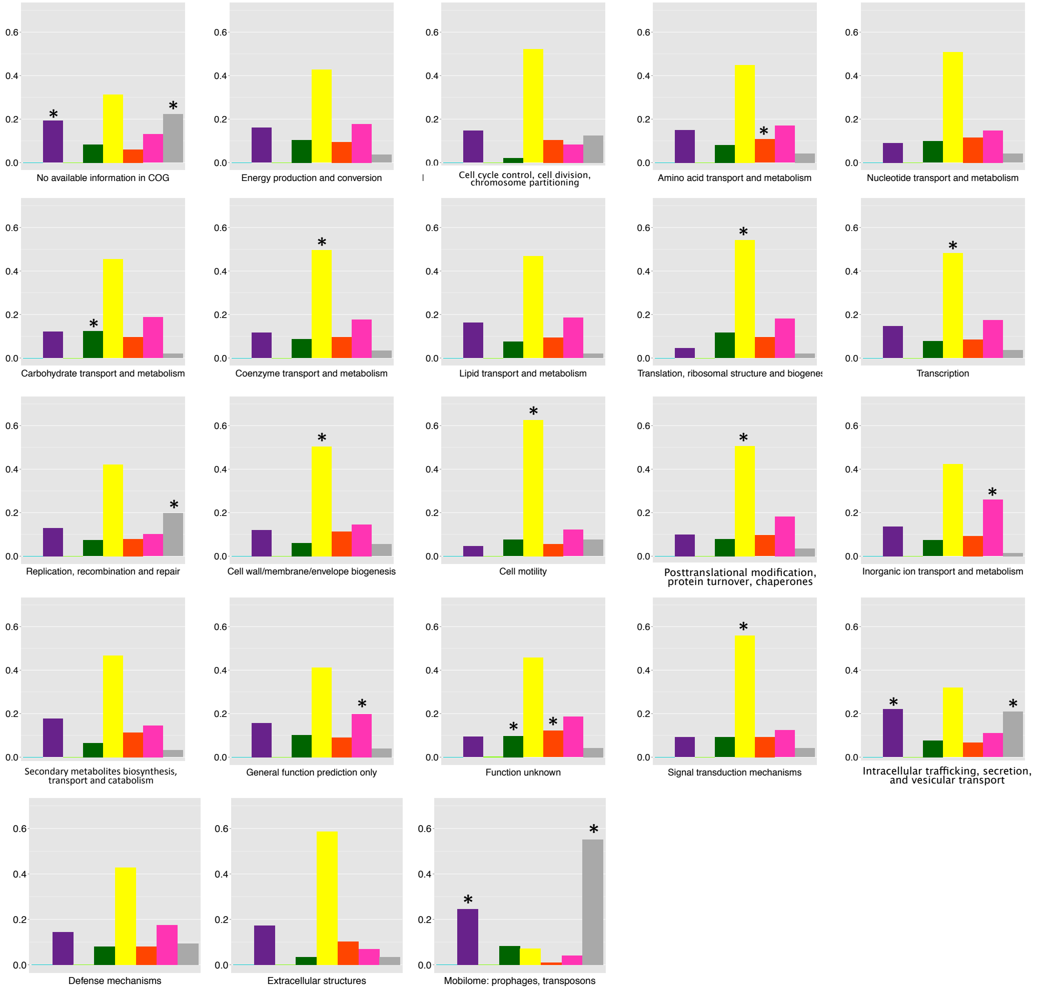


Figure 2.20 Distribution of gene trees in 22 functional categories based on different tree patterns, given genospecies A as a native category. Each graph describes the percentage of gene trees associated with a given COG function (🟦 Trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

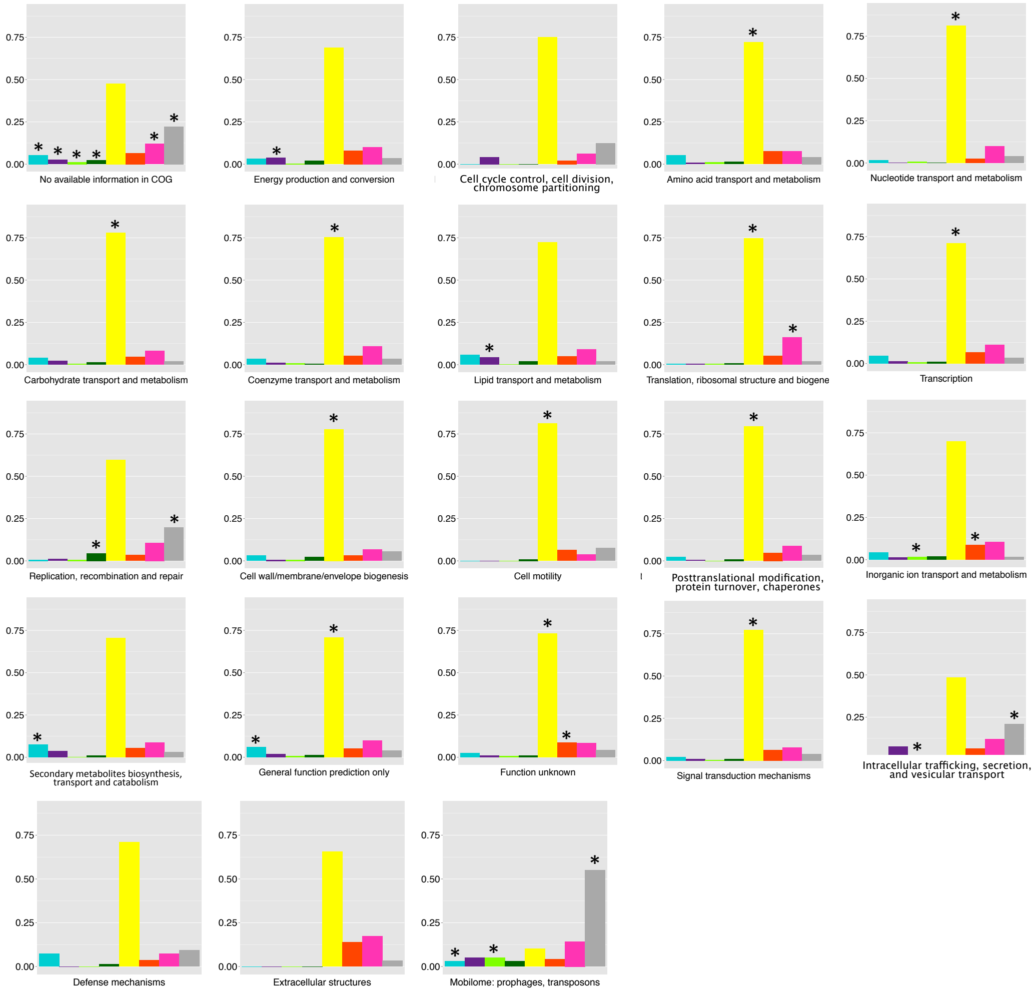


Figure 2.21 Distribution of gene trees in 22 functional categories based on different tree patterns, given genospecies B as a native category. Each graph describes the percentage of gene trees associated with a given COG function (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

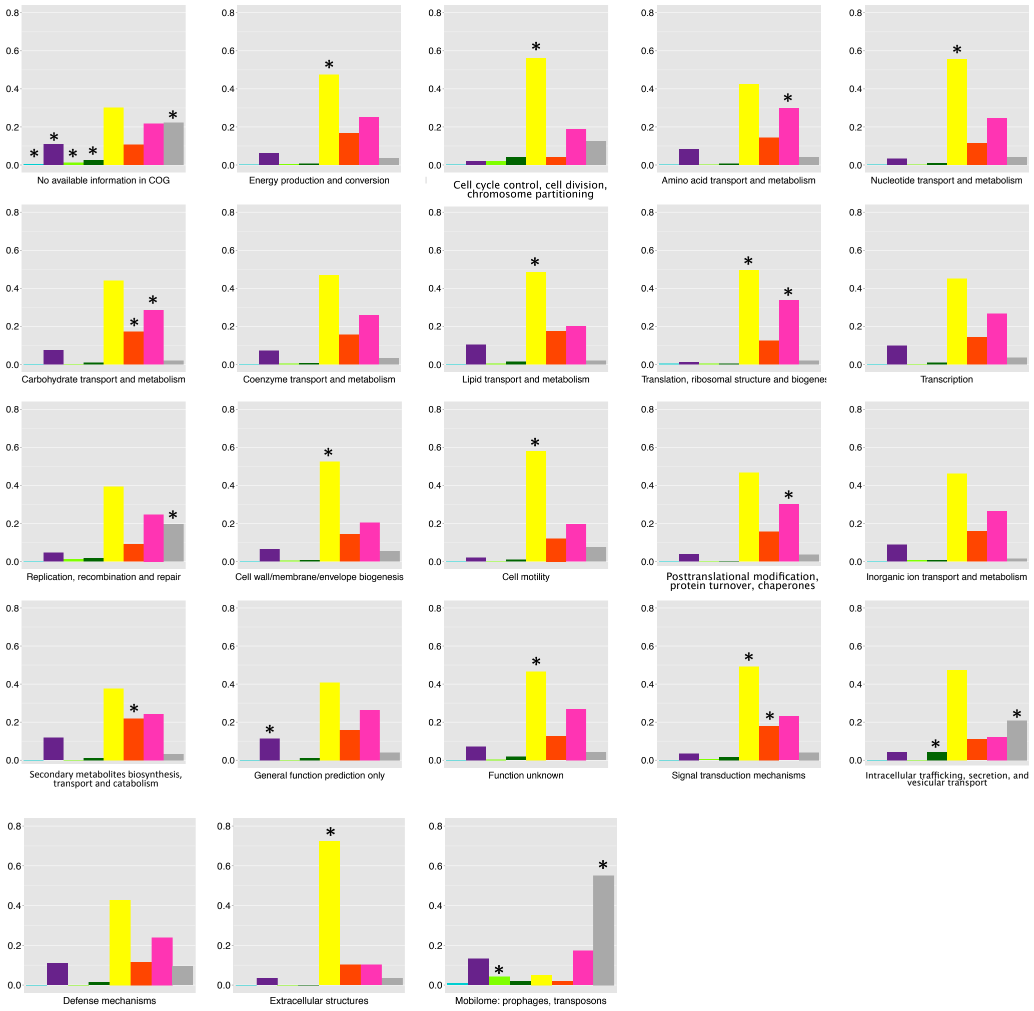


Figure 2.22 Distribution of gene trees in 22 functional categories based on different tree patterns, given genospecies C as a native category. Each graph describes the percentage of gene trees associated with a given COG function (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬜ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

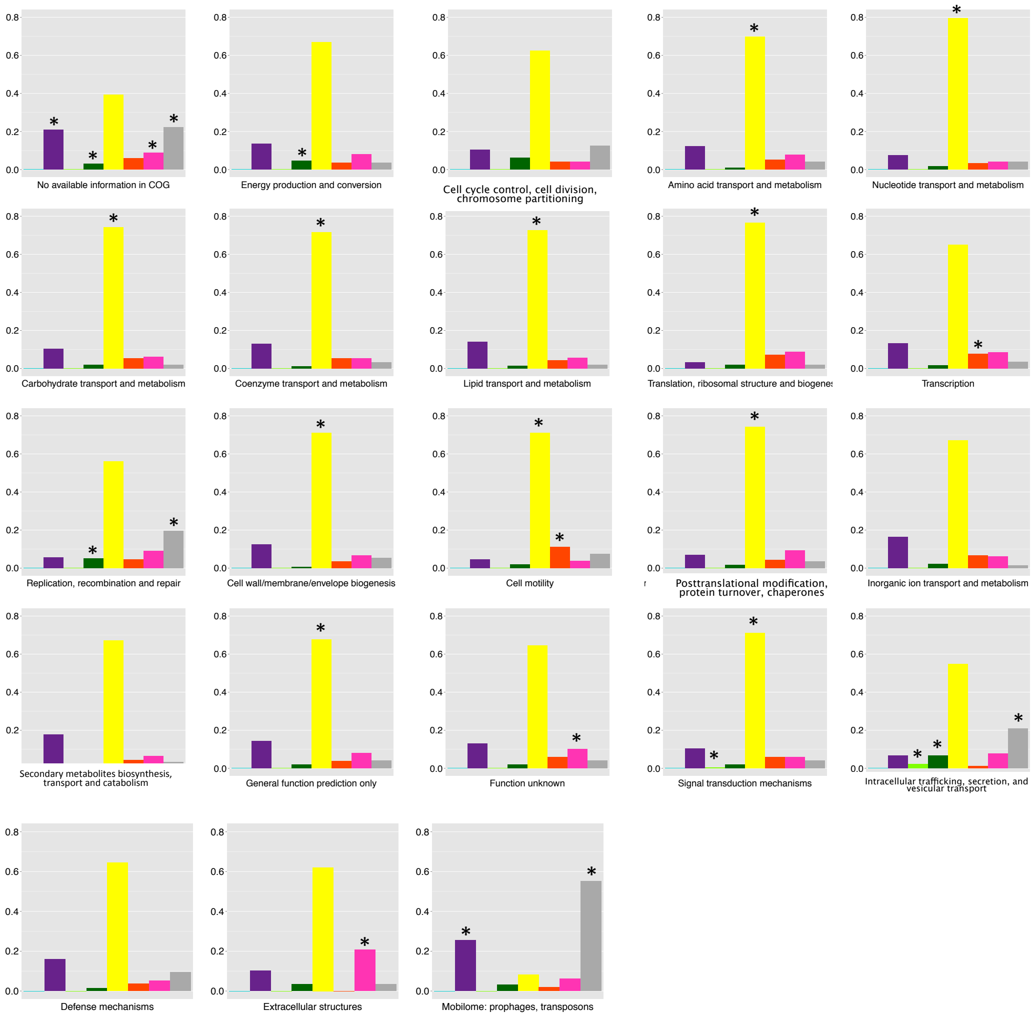


Figure 2.23 Distribution of gene trees in 22 functional categories based on different tree patterns, given genospecies D as a native category. Each graph describes the percentage of gene trees associated with a given COG function (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬜ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

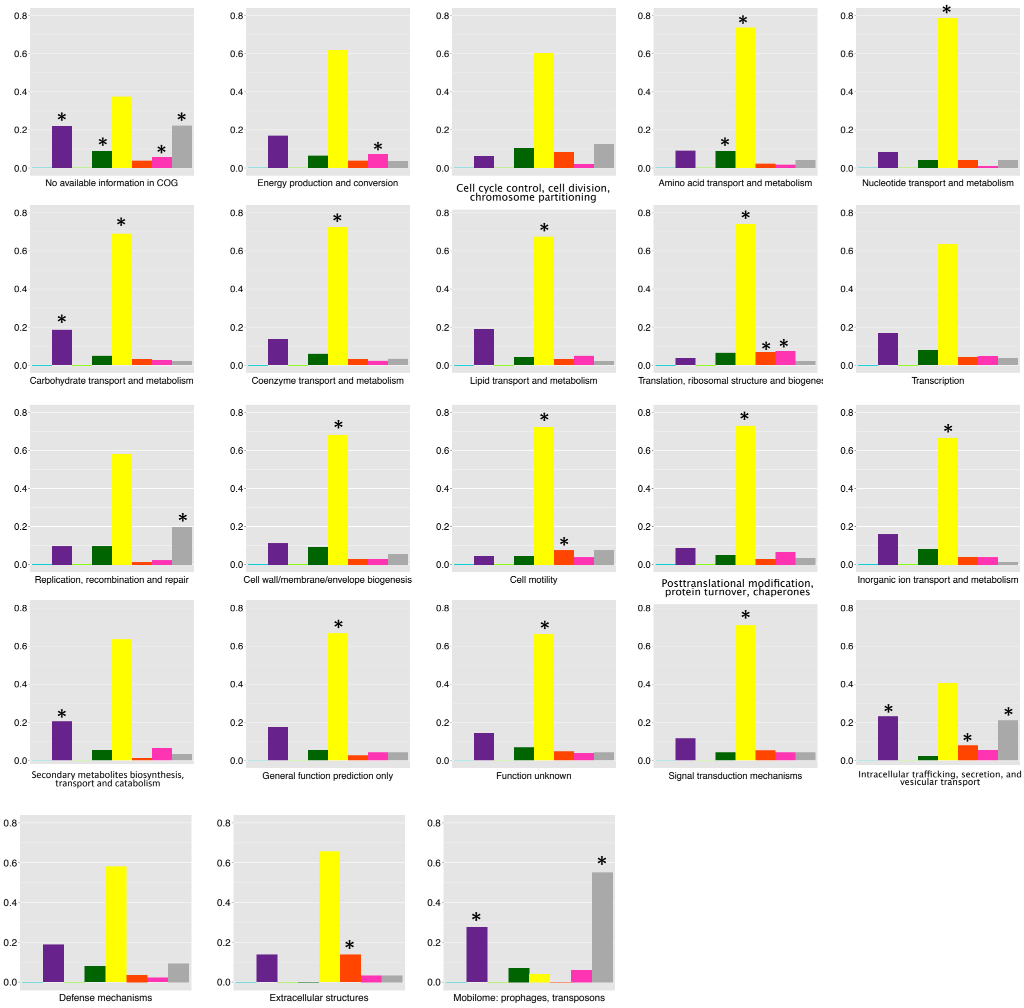


Figure 2.24 Distribution of gene trees in 22 functional categories based on different tree patterns, given genospecies E as a native category. Each graph describes the percentage of gene trees associated with a given COG function (🟦 trees with natives only, 🟪 trees with intruders only, 🟩 trees with natives + single intruder, 🟩 trees with intruders + single native, 🟨 trees with perfect clan, 🟧 trees with perfect slice, 🟥 trees with mélange of natives and intruders, and ⬛ genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

To delineate association between degree of gene transfer and functional categories, discordance index from results on clanistic patterns number of recombinant species to number of excluded genes from analysis were plotted against 22 functional categories (Figure 2.25).

Degrees of transferred gene within five genospecies associated with their function were also hypothesised and investigated. It was exemplified that the distribution of genes with discordance index of 0 and function F (Nucleotide transport and metabolism) were evenly distributed. 2×2 contingency table was drawn (Table 2.9)

Table 2.9 Frequency of genes with discordance index of 0 in category F and other categories under the defined hypothesis.

		Discordance index	
		0	>0
Function	F	49	73
	Not F	2,212	5,640

(p-value = 0.0031, two-sided Fisher's exact test)

It was concluded that the genes with discordance index of 0 were overrepresented in F category.

Genes with discordance index of zero represented strong association with nucleotide transport and metabolism, lipid transport and metabolism and cell wall/membrane/envelope biogenesis.

Genes with discordance index of 1 and greater represented strong association with amino acid transport and metabolism, carbohydrate transport and metabolism, inorganic ion transport and metabolism, energy production and conversion, transcription, secondary metabolites biosynthesis, transport and catabolism, signal transduction mechanisms, translation, ribosomal structure and biogenesis, mobilome: prophages, transposons and genes with no available information in COG database.

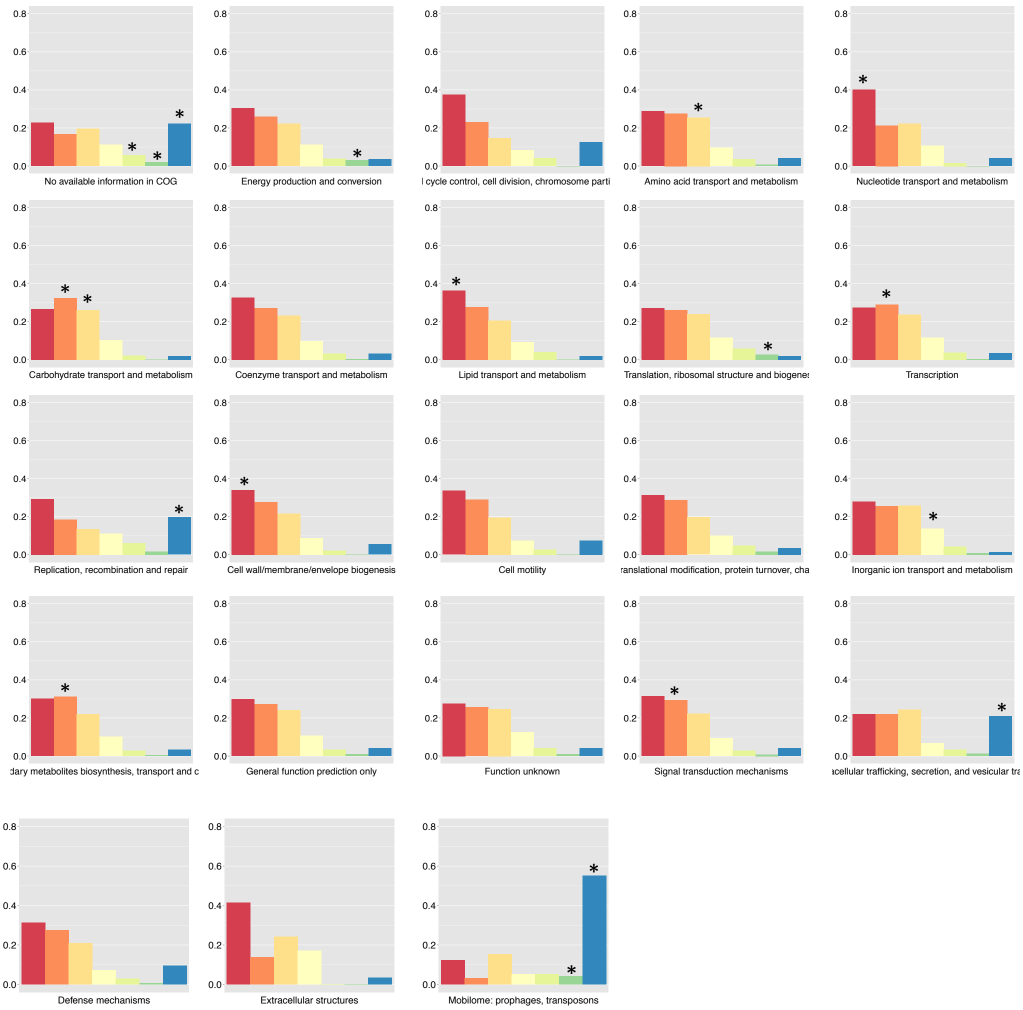


Figure 2.25 Distribution of gene trees in 22 functional categories based on different discordance index. Each graph describes the percentage of gene trees associated with a given COG function (discordance index of 🟥 0, 🟧 1, 🟨 2, 🟨 3, 🟩 4, 🟩 5, and 🟦 genes excluded from analysis). Fisher's exact tests identified (*) fraction of genes with overrepresented in analysed data.

2.6 Discussion

2.6.1 Interspersed and conserved genes between two symbiovars

In this study, the two symbiovars, *trifolii* and *viciae*, have an interspersed distribution on the phylogenies of the majority of genes in the genome, not just those on the chromosome (🟥 gene trees with mélange in Figure 2.12 for given bv. *viciae* as native). For chromosomal genes, a mélange was expected because of the symbiovar definition. Symbiovars are strains within a bacterial species that share chromosomal genes but are differentiated by symbiosis genes (Rogel *et al.*, 2011).

However, for bv. *viciae* as native, it was found that the two symbiovars also shared genes that showed symbiovar-related patterns on plasmids pRL7 (🟧 gene trees with perfect slice), pRL9 (🟩 trees with intruders + single native and 🟧 trees with perfect slice), pRL10 (🟩 trees with natives + single intruder and 🟩 trees with intruders + single native), pRL11 (🟩 trees with intruders + single native), and pRL12 (🟩 trees with intruders + single native and 🟧 trees with perfect slice Figure 2.12). These genes suggest that evolution related to these two symbiovars may also be seen on other replicons.

When a symbiovar is defined as the native category, the natives-only pattern represents differences between the two symbiovars. Some genes with the natives-only pattern were the well-known symbiosis-related genes on pRL10 (Young *et al.*, 2006). pRL100177, which was newly identified as a bv. *viciae* specific gene, was carried by VCS2, VSX15, VSX22 VSX31, and VSX36. The recently-described *bvs* genes on pRL8 (Kumar *et al.*, 2015) were also highlighted, though their annotated functions are not related to symbiosis (Table 2.4).

2.6.2 Effects of genes transferred within five genospecies

This study has revealed that the genomes of the five genospecies have a complex mix of genes with or without evidence of HGT.

The five genospecies were defined by a 305-gene core phylogeny (Figure 1.3 in Chapter 1). The 305 genes are the universal genes described by Harrison *et al.* (2010). Clanistics identified 201 genes with evidence for HGT between genospecies that were in the set of universal genes (discordance index of all 305 genes in Appendix Table I.I). Similar instability of the core genome was also found in other bacterial populations (Didelot *et al.*, 2010, Beauregard-Racine *et al.*, 2011, Didelot *et al.*, 2011, Cadillo-Quiroz *et al.*, 2012).

Clanistics demonstrated the roles that each replicon has taken in evolution. The chromosome not only carried genes with no evidence for HGT, of which a majority formed clans (Figure 2.14-2.16), supporting the concept of species maintained by barriers to gene transfer, but also carried mobile genes. Genes on pRL9 and pRL12 with concordance index of zero and with a natives-only pattern (Figure 2.14) were identified as genospecies B specific genes (Table 2.8 and 2.9), while genes with evidence of HGT were found on pRL7, pRL10, and pRL12 (Figure 2.17). Genes on chromids and plasmids are generally recognised as adaptive and mobilisable genes in bacterial population genomics studies (Heuer *et al.*, 2012, Galardini *et al.*, 2013, Sentchilo *et al.*, 2013).

The genes without evidence of HGT were overrepresented in the categories of nucleotide transport and metabolism, and lipid transport and metabolism, cell wall/ membrane/ envelope biogenesis (Figure 2.25), which include many housekeeping functions that one would expect to be part of the core genome. Genes relevant to mobilome: prophages, transposons were overrepresented with evidence for HGT within five genospecies because transposons and plasmids are mobilome elements (Siefert, 2009) and naturally transferred between species (Nakamura *et al.*, 2004, Beiko *et al.*, 2005, Tamminen *et al.*, 2012). Genes with evidence for HGT are also overrepresented in the functional categories of energy production and conversion, amino acid transport and metabolism, carbohydrate transport and metabolism, inorganic ion transport and metabolism, secondary metabolites biosynthesis, transport and catabolism, translation, ribosomal structure and biogenesis, transcription, signal transduction mechanisms, and as genes with no available information in COG (Figure 2.25). Genes related to operational categories including amino acid biosynthesis, biosynthesis of cofactors, cell envelope proteins, intermediary metabolism, fatty acid and phospholipid biosynthesis, nucleotide biosynthesis, and regulatory genes tended to be transferred. Overrepresentation of horizontally transferred operational genes has been reported in previous studies (Jain *et al.*, 1999, Nakamura *et al.*, 2004, Zhaxybayeva *et al.*, 2006, Kanhere *et al.*, 2009). The complexity hypothesis (Jain *et al.*, 1999) stated that the operational genes are usually members of small assemblies of a few genes products which makes them more portable. Another possible explanation is that these genes encode enzymes required for niche adaptation (Wisniewski-Dyé *et al.*, 2012, Dziewit *et al.*, 2014, Guo *et al.*, 2015).

Genes with evidence for HGT were also preferentially found in translation, ribosomal structure and biogenesis, transcription, and signal transduction mechanisms categories, which contradicts the complexity hypothesis of Jain *et al.* (1999), but is in agreement with Kanhere *et al.* (2009), Wisniewski-Dyé *et al.* (2012), Dziewit *et al.* (2014), and Epstein *et al.* (2014). The genes belonging to the transcription category often produced proteins associated with transcription that was required for the stable maintenance of bacterial plasmids and regulation of accessory genes.

It is noticeable that some of the genes with evidence for HGT were in a "poorly characterized" category, including genes with no available information in COG, general function unknown and function unknown. This finding corresponded with studies of comparative genomics in other bacteria (Wisniewski-Dyé *et al.*, 2012, Dziewit *et al.*, 2014, Epstein *et al.*, 2014) and reflects the fact that the functions of accessory genes are generally less well understood than those of core genes. It should be noted that Choi *et al.* (2007) reported a conflicting result, in that they did not find any association between functional categories and gene transferred.

2.6.3 Limitations of the analysis

Sequences with insufficient information, for example, length of sequence lower than the threshold, were excluded from the analysis. This limitation may result in the analysis suffering because many genes on pRL7, pRL8, and pRL10 were excluded because they did not qualify on the criteria of sequence length and number of sequences. For example, orthologues of pRL80012 were carried by eight strains (VSX2, VSX3, VSX5, VSX22, VSX34, VSX36, VSX31, and *Rlv*3841) but their length ranged from 175 to 1,125 nucleotides. After removing sequences whose length was less than mean of length of all eight sequences, three sequences were remained, which were not sufficient for further analysis. As another example, pRL80002 had orthologues in just two strains, TRX18 and *Rlv*3841, reflected the fact that pRL7 and pRL8 carried accessory genes, which are often rare in the population.

The criterion of sequence length also affected identification of candidate native-specific genes. For example, for bv. *viciae* specific genes, clanistics identified six candidate genes, but five of them were not only carried by bv. *viciae* but also by bv. *trifolii* strains, whose sequences were removed because they were too short. However, this tool is friendly with a large amount of data, provides reasonable time on computation, and computational results with statistics.

The reference strain has an effect in this study because *Rlv*3841 is bv. *viciae* and genospecies B. When a gene with intruders-only pattern is found for genospecies B or bv. *viciae* as native, this means that *Rlv*3841 is an exception compared to strains of the same genospecies or symbiovar in our population. The gene is not actually absent in genospecies B or bv. *viciae* in general, as it is in *Rlv*3841, but it is absent from these groups in our population. This means genes absent from genospecies B cannot be detected. Conversely, when a gene with natives-only pattern is found for given genospecies A, and C-E or bv. *trifolii* as native, the gene is not actually present only in genospecies A, and C-E or bv. *trifolii*. There are genes in the population which are not present in *Rlv*3841. These genes may be specific to genospecies A, and C-E or to bv. *trifolii*, as shown in Figure 4 of Kumar *et al.* (2015). Consequently, using a different reference strain, which is from genospecies A, and C-E or bv. *trifolii*, might help to disclose further gene transfer events in the population.

Some genes have well-known functions, but these are not recorded in the COG database; for example *nodO* (pRL100175, genes with no information in COG), *nodA* (pRL100185, genes with no information in COG), *nifB* (pRL100195, General function prediction only). In general, the COG database has more accurate information on core genes that on accessory genes, and this may bias the apparent relationship between functional characteristics and gene transfer.

In this chapter, clanistics can detect dispersed and conserved genes between two symbiovars by looking at concordance and discordance on clanistic patterns. Genes specific to bv. *viciae* were confirmed in line with previous studies. Within five genospecies, rhizobium genes were identified that have or have not been transferred. Many more genes were well conserved within five genospecies than between two symbiovars. Use of the discordance index demonstrated that chromosomal, pRL9 and pRL12 genes were less affected by HGT, some of these had been used for core phylogeny construction. The clanistics results could be used to identify a set of core genes that were reliable markers for identifying the genospecies of new isolates. Clanistics also detected genes specific to genospecies B that were located on pRL7, pRL9, pRL10, pRL11 and pRL12, so it appears that the differences between genospecies are not confined to chromosomal genes but also extend to plasmid-encoded genes (or, at least, genes that are plasmid-encoded in 3841).

Chapter 3 Co-occurrence of genes in the local population of *Rhizobium leguminosarum*

3.1 Abstract

In the previous chapter, gene transfers were observed in the rhizobium population which reflected the genetic diversity within the five genospecies. Within the genospecies, the genes tend to be either lost or conserved in the genome, thus increasing the distinctiveness of a genospecies or its ability to adapt to a specific environment during evolution. Patterns of gene presence/absence (i.e. gene distribution in the population) are investigated in this chapter. To obtain the patterns within the occurrence data, gene presence/absence profiles (alternatively called phylogenetic profiles) were first evaluated as relationships of gene pairs by different measures. Co-occurrence and anti co-occurrence of genes quantified by these different measures were later compared in order to find an optimal measure. Before co-occurrence and anti co-occurrence relationships were converted to a co-occurrence gene network, an optimal threshold of correlation values was determined to prevent network construction from spurious correlations. The co-occurrence gene network was constructed based on optimal parameters to visualise a broader view of genomic relationships in the rhizobium population. The final network included 2,663 genes and 33,318 interactions of which a majority were co-occurrence. Co-occurrence subnetworks contained neighbouring genes, genes participating in the same biological process (e.g. symbiosis genes) or genes present in the same subpopulation (e.g. genes specific to genospecies). Anti co-occurrence relationships between subnetworks were detected for chromosomal genes on genomic islands and for plasmid or chromid genes with replaceable functions.

3.2 Introduction

The diversity of ecological properties, such as ability to utilise carbon substrates and symbiotic host range, or symbiovar (Rogel *et al.*, 2011), is a consequence of interplay between numerous genes. These genetic dependencies require the co-occurrence of those genes in the population (Huynen *et al.*, 2000, Cui, 2010, Kim *et al.*, 2011). Comparative studies of genomes in bacterial populations revealed that there are many mechanisms, including recombination and horizontal gene transfer, that relate to adaptation of organisms to their specific environmental conditions during evolution (Tian *et al.*, 2010, Bailly *et al.*, 2011, Kumar *et al.*, 2015). Homologous recombination or horizontal gene transfer can result in variation of gene distribution in the bacterial population (Yerrapragada *et al.*, 2009, Mallet *et al.*, 2010, Smokvina *et al.*, 2013, Sugawara *et al.*, 2013, Méric *et al.*, 2014).

Table 3.1 Phylogenetic profiles. Rows represent strains, and columns represent gene presence/absence in a strain, which are encoded with binary values (0: absent gene and 1: present gene) (redrawn from Cokus *et al.* (2007)).

	Gene 1	***Gene 2***	***Gene 3***	***Gene 4***
Strain A	1	0	1	1
Strain B	0	0	0	1
Strain C	0	0	1	0
Strain D	1	0	1	0
Strain E	1	1	0	0

Patterns of genetic elements were firstly investigated as gene homologs (Overbeek *et al.*, 1999) and protein homologs (Pellegrini *et al.*, 1999, Cokus *et al.*, 2007). The presence/absence profile of genes, called the phylogenetic profile, is in a binary format (0: absence of gene and 1: presence of gene) (Table 3.1). Analyses of gene occurrence have been done previously by using data from the COG database (Wu *et al.*, 2003, Kim *et al.*, 2011, Cohen *et al.*, 2012). Gene occurrence in *Mycoplasma genitalium* was compared with evidence from gene fusions and genomic neighbourhood (Huynen *et al.*, 2000). These studies classify the relationship of genes into co-occurrence (genes that tend to occur in the same genome) and anti co-occurrence (genes that tend not to be found in the same genome). Their findings indicate that genes with co-occurrence relationships participate or play a role in the same pathway or under a specific condition, while anti-co-occurrence genes may encode the same function but perform in a different way.

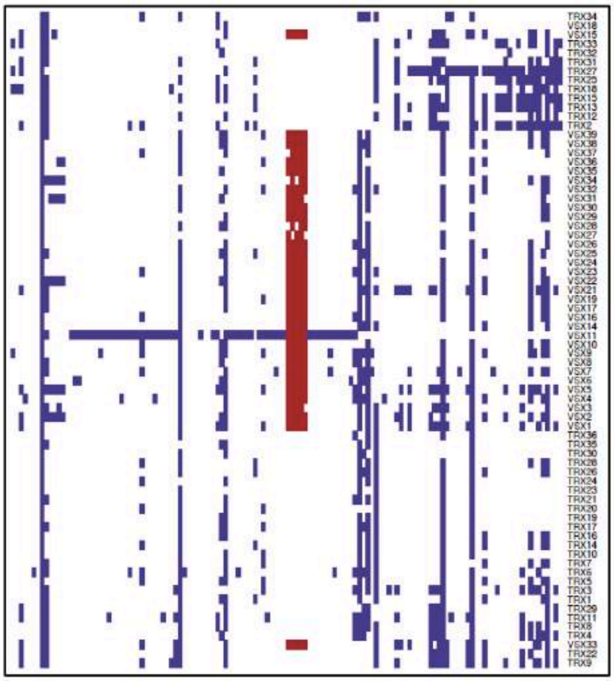


Figure 3.1 Presence/absence matrix of pRL8 obtained for 72 R. leguminosarum strains using *Rlv*3841 small plasmid genes. The presence of genes is shown in blue, absent genes are in white. Rows represent 72 strains, and columns represent plasmid genes that are longer than 100 bp. Strains are arranged according to their respective genospecies (A-E). The *bvs* (symbiovar *viciae* specific) genes in pRL8 are in brown (taken from Kumar *et al.* (2015)).

In the comparative study of 72 rhizobium strains by Kumar *et al.* (2015), the diversity of gene occurrence in the population was represented in a gene presence/absence matrix format with reference to the replicons of *Rlv*3841 (e. g. Figure 3.1). The presence of many genes was found to be restricted to some strains in the population, but dispersed among the five genospecies. Focusing on the five *bvs* genes on pRL8 (Table 2.4 in Chapter 2), the results of the previous chapter demonstrated that the unrooted trees of these five genes represented natives-only patterns for bv. *viciae* as the native category (Figure 2.13 in Chapter 2). The occurrence patterns and clanistic results of these five genes supported the co-occurrence relationships of these *bvs* genes.

This chapter aims to quantify co-occurrence and anti co-occurrence relationships of genes in the rhizobium population, which may not only give us a better insight into gene organisation of the rhizobium genome but also allow us to infer function of unknown genes. To achieve this, co-occurrence relationships measured by different coefficients will be compared to find the optimal measure. In order to display the co-occurrence of genes that may be on distinct replicons, significant co-occurrence relationships of genes will be converted to a gene co-occurrence network. The significant co-occurrence relationships are greater than or equal to an optimal threshold, which is a minimum correlation value used for the gene co-occurrence network construction. Finding the optimal threshold is a challenge because defining a low threshold may include spurious relationships in the network, whereas defining a high threshold may exclude significant relationships from the network. The threshold of the gene co-occurrence network will be selected by using methods based on graph theory.

3.3 Chapter aims

1. Evaluating co-occurrence and anti co-occurrence relationships of gene pairs in a rhizobium population.
2. Constructing the gene co-occurrence network from optimised parameters and analysing the gene co-occurrence network based on graph theory.
3. Extracting biological features of co-occurrence and anti co-occurrence relationships reflected from the gene co-occurrence network.

3.4 Materials and methods

3.4.1 Gene presence/absence data

The gene data included in the study were obtained from 85 rhizobium strains as detailed in Chapter 2. The data used herein were profiles containing gene distribution in terms of gene presence/absence from the rhizobial population for each gene in the genome of the reference strain *Rlv*3841. They are illustrated in Figure 3 of Kumar *et al.* (2015). According to Nitin Kumar (personal communication), the phylogenetic profile of each gene was collected in binary format (1: present and 0: absent) from reference-based assembly against *Rlv*3841 (Newbler 2.5 software with 90% sequence identity and 40-bp minimum overlap as parameters).

3.4.2 Quantification of co-occurrence genes

To gain insight into the interaction of co-occurring genes, standard measures of profile similarity were considered.

Jaccard similarity index (*J*) (Jaccard, 1912) is used for quantifying association of either similarity or difference between two genes present in the population. The Jaccard similarity index of two genes is a ratio between the number of strains containing the two genes and the number of strains containing either of them, where

$$J_{gene_A,gene_B} = \frac{\text{Number of strains where both } gene_A \text{ and } gene_B \text{ are present}}{\text{Number of strains where } gene_A \text{ and/or } gene_B \text{ are present}}$$

The Jaccard similarity index ranges between 0 and 1. The index equals 1 if the occurrence of the two genes is identical in the population, while the index equals 0 if there is no strain carrying both genes. The Jaccard index is considered to be an effective measure when there are strains carrying both the two genes. A weakness of the Jaccard is that when the two genes have presence patterns that are complements of each other, the Jaccard index of the two genes is "0", although actually the two genes have a negative relationship.

Pearson correlation coefficient ($r_{\text{gene}_A,\text{gene}_B}$) for binary variables, used by Kim *et al.* (2011), is defined as follows :

$$r_{gene_A,gene_B} = \frac{C_{gene_A,gene_B}N - E_{gene_A}E_{gene_B}}{\sqrt{E_{gene_A}E_{gene_B}\left(N - E_{gene_A}\right)\left(N - E_{gene_B}\right)}}$$

$r_{gene_A,gene_B}$ denotes Pearson's correlation in occurrence between *gene*$_A$ and *gene*$_B$. $C_{\text{gene}_A,\text{gene}_B}$ is the number of strains carrying both *gene*$_A$ and *gene*$_B$. E_{gene_A} is the number of strains containing *gene*$_A$. *N* denotes the total number of strains in the dataset. The value of $r_{gene_A,gene_B}$ is standardized to lie between -1 and 1. 0 means that the strains having *gene*$_A$ are completely independent of the presence of *gene*$_B$. There is no relationship between the presence of *gene$_A$* and *gene$_B$*. 1 means *gene$_A$* and *gene$_B$* have a maximal co-occurrence relationship because the presence of *gene$_A$* and *gene$_B$* is identical, *i. e.*, every strain that carries *gene$_A$* has to carry *gene$_B$*. −1 means means *gene$_A$* and *gene$_B$* have an anti co-occurrence relationship: every strain that carries *gene$_A$*, does not carry *gene$_B$* and vice-versa.

Mutual information ($M_{\text{gene}_A,\text{gene}_B}$) (Huynen *et al.*, 2000, Steinhauser *et al.*, 2007, Kensche *et al.*, 2008) quantifies dependence between *gene$_A$* and *gene$_B$*. $M_{\text{gene}_A,\text{gene}_B}$ is derived from the entropy of *gene$_A$* (H_{gene_A}) and *gene$_B$* (H_{gene_B}) . The mutual information is described mathematically by the log-odds ratio of the expected co-occurrence of pairs of genes, based on their individual frequencies, to the observed frequency of occurrence.

$$H_{\text{gene}_A} = -\sum\nolimits_{\text{gene}_A} P_{\text{gene}_A}\, logP_{\text{gene}_A}$$

$$H_{\text{gene}_B} = -\sum\nolimits_{\text{gene}_B} P_{\text{gene}_B}\, logP_{\text{gene}_B}$$

$$H_{\text{gene}_A,\text{gene}_B} = -\sum\nolimits_{\text{gene}_A,\text{gene}_B} P_{\text{gene}_A,\text{gene}_B}\, logP_{\text{gene}_A,\text{gene}_B}$$

$$\begin{aligned} M_{\text{gene}_A,\text{gene}_B} &= H_{\text{gene}_A} + H_{\text{gene}_B} - H_{\text{gene}_A,\text{gene}_B} \\ &= \sum\nolimits_{\text{gene}_A,\text{gene}_B} P_{\text{gene}_A,\text{gene}_B} / P_{\text{gene}_A}\, P_{\text{gene}_B} \end{aligned}$$

P_{gene_A} represents the probability of all possible events relating to *gene$_A$* (herein events mean the presence or absence of *gene$_A$*). $P_{\text{gene}_A,\text{gene}_B}$ represents the probability of all possible events relating to *gene$_A$* and *gene$_B$* (herein events mean the combination of presence or absence of *gene$_A$* and *gene$_B$*). $M_{\text{gene}_A,\text{gene}_B}$ is 0 if and only if the measurements on the *gene$_A$* and *gene$_B$* are statistically independent. The mutual information of *gene$_A$* and *gene$_B$* gets higher if the occurrence of *gene$_A$* and *gene$_B$* is more similar. The maximum mutual information is given when (1) both genes are present in about 50% of the genomes (the individual entropies of the genes are maximum), and (2) the genes are completely present together (the combined entropy is minimum). The combined entropy is theoretically minimal when the genes are never found together. That event does not occur when studying the genes from one genome.

The probability (*p-value*) was computed in order to verify the significance of correlation strength. If this *p-value* is lower than 0.05, the correlation coefficient is considered to be statistically significant (Steinhauser *et al.*, 2007).

3.4.3 Threshold estimation

The co-occurrence genes network consisted of genes and their interactions that are represented as nodes and edges, respectively. In order to derive a biologically meaningful network, threshold selection is a crucial step. By setting too low a threshold, either a false negative or a false positive relationship can be retained in the network. Likewise if this threshold is set too high, important relationships can be lost from the network.

Graph-based topology (including network density, connected components, and clustering coefficient) and spectral graph theory were introduced here in order to optimise threshold of the network.

Density of network (Pavlopoulos *et al.*, 2011) is the proportion of all possible edges that are actually found.

$$density = \frac{2|E|}{|V|(|V|-1)}$$

where $|V|$ is the number of nodes and $|E|$ is the number of edges. Density of network tends to decrease when the correlation threshold is increased. An increment in threshold removes edges having a correlation value lower than the threshold from the network, and also removes isolated nodes. At high correlation, the density increases again as the number of nodes tends to be stable while the number of edges gradually decreases. Network threshold is expected to be found at a minimal value of network density (Aoki *et al.*, 2007, Ozaki *et al.*, 2010).

Connected component in the network can be described as a subnetwork in which any two nodes are connected to each other. A number of connected components are studied in order to find the network threshold. As the threshold is increased, edges in the network are removed, the network shows a tendency to be sparse. The number of connected components then increases. At the correlation value showing a sharp transition in the number of connected components, the optimal threshold is observed (Fukushima *et al.*, 2011).

Clustering Coefficient (Barabasi *et al.*, 2004) measures the tendency of a node to form a cluster. Considering node$_i$, the local clustering coefficient of node$_i$ (C_i) is defined by a ratio between the number of links between neighbour nodes of a node$_i$ (e_i) and the number of neighbour nodes (k_i).

$$C_i = \frac{2|e_i|}{k_i(k_i - 1)}$$

The average Clustering Coefficient of the whole network $C_{average}$ is given by

$$C_{average} = \frac{1}{N}\sum_{i=1}^{N} \frac{2|e_i|}{k_i(k_i-1)}$$

where $N = |V|$ represents the number of nodes. This clustering coefficient value ranges between 0-1. The network is likely to be clustered, if this average local clustering coefficient is closer to 1. When the threshold is increased, edges are gradually removed from the network resulting in a decrease in the average clustering coefficient. When the average clustering coefficient is lowered the network becomes highly ‘cliquish’, referring to disconnected subnetworks, and average clustering coefficient starts increasing again. A transition of average clustering coefficient is observed at the potential threshold (Gupta *et al.,* 2006, Elo *et al.*, 2007).

Spectral graph theory method (Perkins *et al.*, 2009) uses eigenvalues and eigenvectors of the largest component in the network to quantify a number of spectral clusters. Different numbers of spectral clusters are found by increasing the cutoff value. A potential threshold is identified when a peak in the number of spectral clusters is seen. The spectrum of the largest connected component is considered because it contains a majority of nodes in the network. Analysis of spectral clusters of the network is conducted on eigenvalues and eigenvectors of Laplacian matrix. The Laplacian matrix is derived from an adjacency matrix (A) and a degree matrix (D) (Ding *et al.,* 2001). The smallest eigenvalue will be zero and the remaining eigenvalues will not be zero; the smallest non-zero eigenvalue is named the algebraic connectivity. At the lowest algebraic connectivity, the network will have nearly-connected components (Ding *et al.*, 2001), the potential threshold will then be identified. The number of spectral clusters is maximised by detecting the cluster on the eigenvector associated with the smallest non-zero eigenvalue. Cluster detection is conducted based on searching the gap between eigenvector values by using a sliding window technique. With a sliding window five elements wide, a new cluster is detected when the difference between the lowest and the highest value in the window is greater than $m + \frac{s}{2}$, where m is the median of all difference in positions windowsize apart and s is the standard deviation of this set of values.

3.4.4 Community analysis

The property of community in the networks is defined by densely connected nodes within them but sparse connections to the other nodes. Pons *et al.* (2005) developed a random-walk algorithm named *Walktrap. Walktrap* combines a distance optimization for measuring node similarity and a modularity evaluation for investigating community. Intuitively, random walks on the network tend to be "trapped" within highly connected parts of local communities. Hence, this algorithm is called *Walktrap* and is implemented in *igraph* (Csardi *et al.*, 2006).

The unweighted graph G has an associated adjacency matrix A; $A_{ij} = 1$ if there is an edge between i and j, and $A_{ij} = 0$ otherwise. In the journey between i and j, if the visited nodes are chosen randomly and uniformly, it is referred to as a Markov chain. At each time point, the random walk process starts at node i and in the walk of length t, a random step is taken to an adjacent node j. The transition probability from i to j is defined as $P_{ij} = \frac{A_{ij}}{d(i)}$, where $d(i)$ is the degree of i, $d(i) = \sum_j A_{ij}$. These transition probabilities define a transition matrix P. The transition matrix P satisfies two general properties of the random walk process. P_{ij}^{t} stands for the probability of going from i to j through a random walk of length t.

The basic idea of community detection is to partition densely connected nodes from sparsely connected nodes. A partition ($\mathcal{P} = \{C_1, C_2, \dots, C_k\}$) of the network is defined as an optimal community providing maximum modularity.

The modularity is computed by the fraction of edges inside the community compared to the fraction of edges bound to community in the partition.

In a random walk length t starting at i towards infinity, the probability of being on node j tends to be the degree of node j:

$$\forall_i \lim_{t \to +\infty} P_{ij}^t = \frac{d(j)}{\sum_k d(k)}$$

where k is an index of all nodes n in graph G.

The probabilities of walking from i to j and from j to i through a random walk of a length t have a ratio that only depends on the degrees $d(i)$ and $d(j)$:

$$\forall i, \forall j, d(i)P_{ij}^t = d(j)P_{ji}^t$$

The transition probabilities P_{ji}^t are not only used to measure structural similarity between vertices but also similarity between communities.

A partition $\mathcal{P}_1$ of the graph into n communities is reduced to a single node. The distances between all adjacent nodes are computed. Iterative methods are used to find the optimal communities in the network. An overview of the algorithm at step k is provided.

- Choose two communities C_1 and C_2 in $\mathcal{P}_k$ on a criterion based on the distance between the communities (of which details are provided later).
- Merge C_1 and C_2 into a new community C_3 and create the new partition: $\mathcal{P}_{k+1}$= $(\mathcal{P}_k \setminus \{C_1, C_2\}) \cup \{C_3\}$.
- Update the distances between communities.

This similarity can be used in hierarchical clustering. The distance between the two vertices i and j, (r_{ij}) is computed by:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}^t - P_{jk}^t\right)^2}{d(k)}}$$

The probability of going from community C to node j at time t is:

$$P_{Cj}^{t} = \frac{1}{|C|}\sum_{i\in C} P_{ij}^{t}$$

Likewise, the distance between two communities C_1 and C_2 $(r_{C_1C_2})$ is:

$$r_{C_1C_2} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{C_1k}^{t} - P_{C_2k}^{t}\right)^2}{d(k)}}$$

where $P_{C_1k}^{t}$ measures the probability of traversing from a node in C_j to node k (j=1,2). At each step k in the merge algorithm, two communities (C_1, C_2) are determined to merge by minimizing the mean σ_k of the squared distances between each vertex and its community.

$$\sigma_k = \frac{1}{n}\sum_{C\in R_k}\sum_{i\in C} r_{iC}^2$$

After the merge of C_1 and C_2, the squared distances $(\Delta\sigma)$ between the two communities is updated and calculated by:

$$\Delta\sigma(C_1, C_2) = \frac{1}{n}\frac{|C_1||C_2|}{|C_1| + |C_2|} r_{C_1C_2}^2$$

Maximizing modularity (Q) is performed to find the optimal communities in the network. The modularity is derived by

$$Q(\mathcal{P}) = \sum_{C\in\mathcal{P}} e_C - a_C^2$$

where e_c represents the fraction of edges inside the community. a_c represents the fraction of edges bound to community C. Further background and details of the *Walktrap* implementation can be found in the original work (Pons *et al.*, 2005).

3.5 Results

3.5.1 Results on different measures of association

All standard association measures were compared in order to find the optimal measure for use in this work. Because this work focuses on association of gene presence/absence, we considered the effect of gene abundance, *i. e.* the number of strains carrying the gene in the population, on each measure. In the distribution of abundance of genes on pRL10 (Figure 3.2), coloured bars were identified as genes with high or low abundance in the population that will be considered later by different association measures.

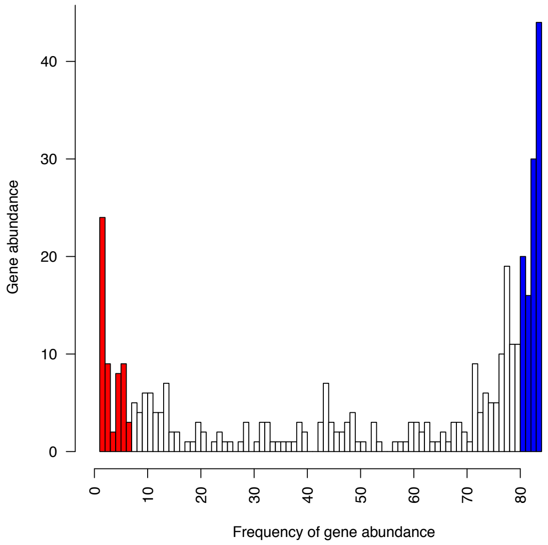


Figure 3.2 Histogram of gene abundance on pRL10. Plots are coloured by gene abundance (🟦 genes with high are found in at least 79 strains, 🟥 genes with low abundance are found in 0-6 strains, and ☐ genes with low abundance are found in 7-78 strains).

Jaccard was the basic measure to compare similarity between gene profiles. Jaccard worked well when gene presence was the focus. Genes (🟥 in Figure 3.2) with low abundance in the population can represent high values of Jaccard because pairs of them were possibly present in the same few strains and usually absent from the rest of the strains (🔴 in Figure 3.3).

When genes had high abundance (🟦 in Figure 3.2), pairs of such genes (🔵 in Figure 3.3) had a high value of Jaccard that was not surprising because those genes were found in almost all strains in the population. However, these gene pairs were not interesting in this work because they were found generally in the population and may not represent specific functions in a subgroup of the population. A drawback of this measure was found in the case of two genes that tend to be present in complementary strains. These two genes may be considered as either independent or anti co-occurring genes. If these two genes are anti co-occurring genes, Jaccard cannot enumerate their association. This measure then was ultimately discarded for further analysis.

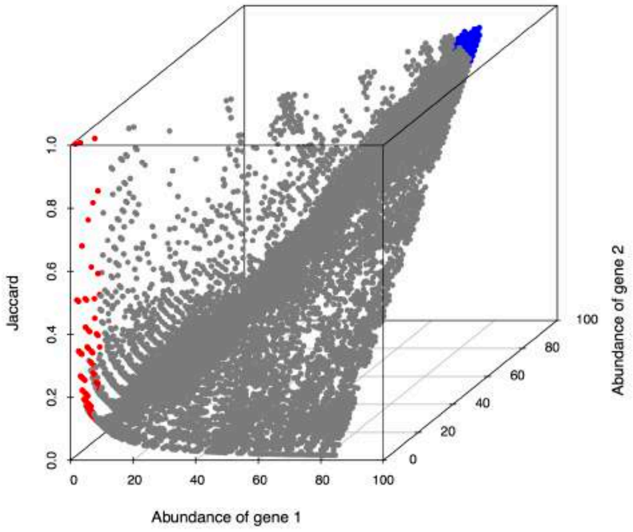


Figure 3.3 Plot of frequency of pRL10 gene pairs and their Jaccard values. Nodes are coloured by abundance of genes in a pair (● gene pairs with low abundance, found in 0-6 strains, ● gene pairs with high abundance, found in at least 79 strains; and ● gene pairs found in 7-78 strains).

Next, Pearson's correlation was considered. Pearson's correlation can quantify not only the strength of association of a gene pair by amount of correlation value but also the type of their interaction as shown by the sign of the correlation value. Co-occurring and anti co-occurring genes can be described with positive and negative signs of the correlation, respectively. The strength of the interaction can be expressed as the amount of Pearson's correlation. An independent gene pair will have a Pearson's correlation value of zero that indicates that the two genes were present independently of each other.

Pairs of genes with low abundance in the population (■ in Figure 3.2) presented a wide range of Pearson's correlation values (● in Figure 3.4). High positive correlation of pairs of genes with low abundance was observed when they were present in the same few strains and absent from the rest of the strains. Pairs of genes with high abundance in the population (■ in Figure 3.2) also represented a wide range of Pearson's correlation values (● in Figure 3.4), differing from Jaccard since Pearson's correlation not only counted the gene presence but also gene absence in the population. Gene pairs with high abundance had high values of Pearson's correlation if those genes were absent in the same few strains. Gene pairs with high abundance had low values of Pearson's correlation if those genes were missing from different strains. These associations are shown in ● in Figure 3.4, near the zero of Pearson's correlation.

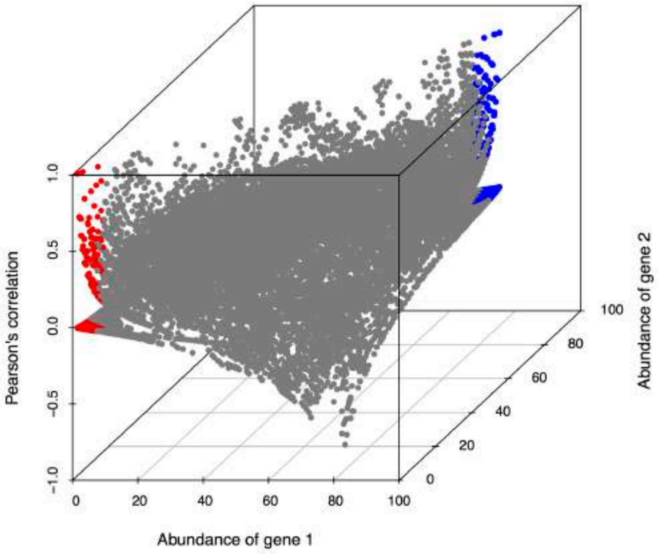


Figure 3.4 Plot of frequency of pRL10 gene pairs and their Pearson's correlation values. Nodes are coloured by abundance of genes in a pair (● gene pairs with low abundance, found in 0-6 strains; ● gene pairs with high abundance, found in 79 strains; and ● gene pairs found in 7-78 strains).

Mutual information was the last association measure considered here. Because mutual information measures the information that two genes give about each other, high values of mutual information were represented in the middle of the plot shown in Figure 3.5, belonging to two genes present in about half of the population. It was also found that genes with low and high abundance in the population had low mutual information values. Considering two pairs of gene with complementary frequencies, their mutual information values are equal, so this measure generates a symmetrical plot (Figure 3.5) (Knobbe *et al.,* 1996). Mutual information was robust to gene pairs with high or low abundance. Mutual information values of these genes were not as high as correlation values of these genes from Jaccard and Pearson's correlation.

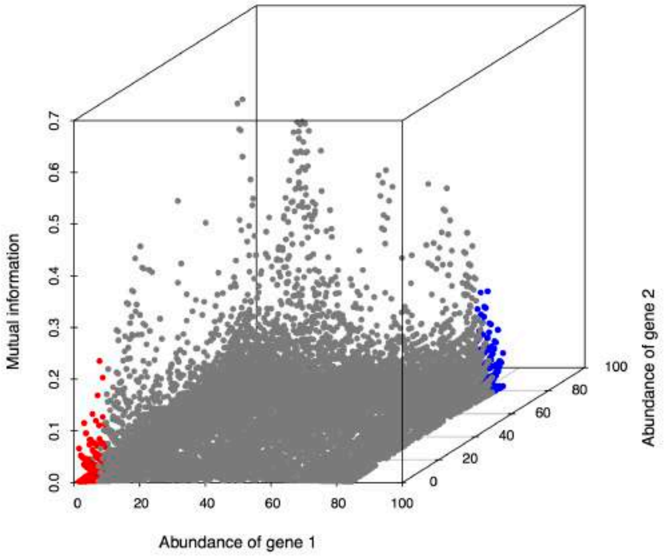


Figure 3.5 Plot of frequency of pRL10 gene pairs and their mutual information values. Nodes are coloured by abundance of genes in a pair (● gene pairs with low abundance, found in 0-6 strains; ● gene pairs with high abundance, found in 79-85 strains; and ● gene pairs found in 7-78 strains).

Even though mutual information was able to quantify the amount of association between two genes without gene abundance bias and find non-linear relationships, it cannot specify the direction of the relationship. It was also found that the computed Pearson's correlation and mutual information values for the data used in this study were similar (Figure 3.6). This situation was also found in a study on co-expressed genes (Steuer *et al.,* 2002).

In order to achieve the objectives of this chapter, Pearson's correlation was therefore chosen to be the association measure for this work.

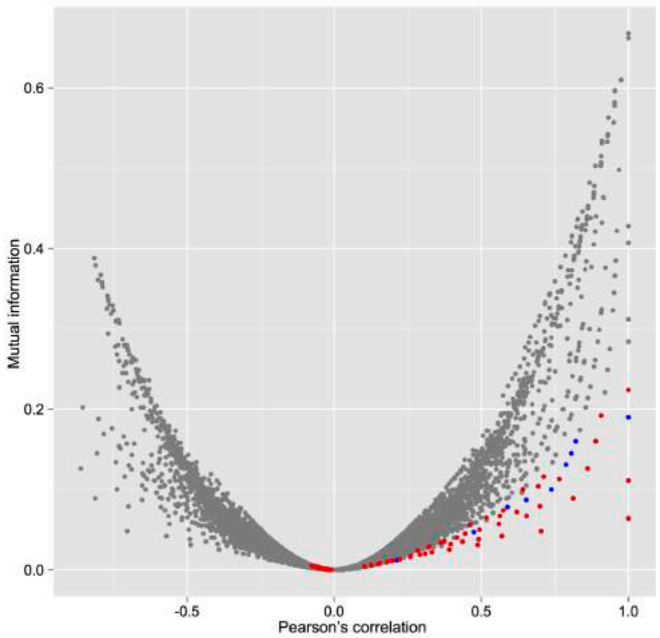


Figure 3.6 A comparison between the Pearson's correlation and mutual information of pRL10 genes. Nodes are coloured by abundance of genes in a pair (● gene pairs with low abundance, found in 0-6 strains; ● gene pairs with high abundance, found in 79-85 strains; and ● gene pairs found in 7-78 strains).

3.5.2 Results on threshold estimation

An optimal threshold of association value is important for the study of a co-occurring gene network because the threshold affects the conclusions acquired from the resulting co-occurring gene network. An absolute value of Pearson's correlation was evaluated for threshold selection.

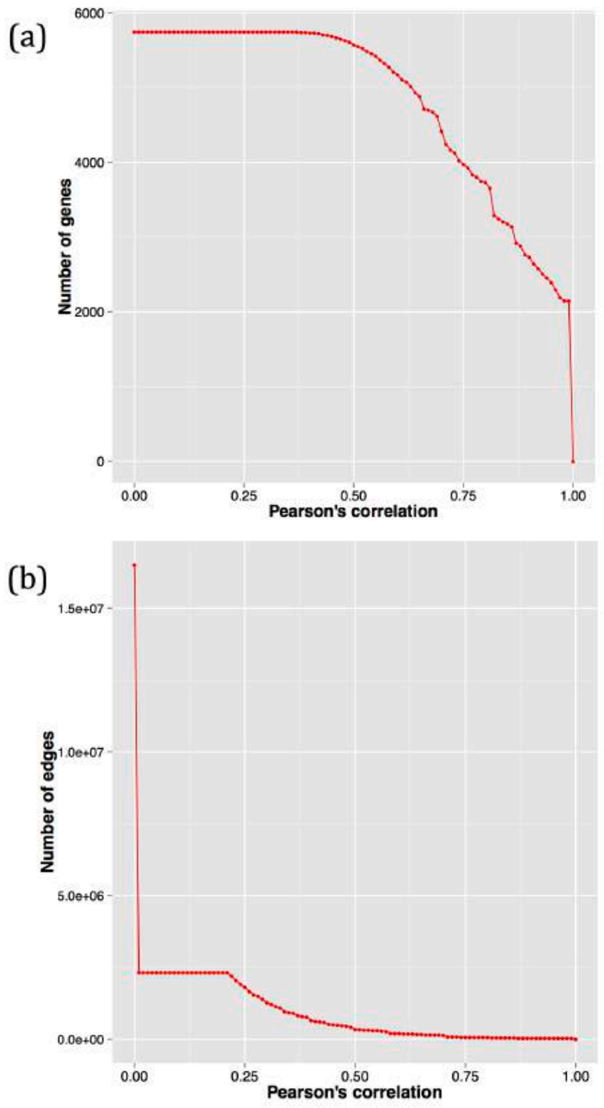


Figure 3.7 Overview of co-occurrence gene networks derived from 72 strains of the rhizobium population. (a) The number of genes in the network at various thresholds of absolute values of Pearson's correlation for co-occurrence genes network (b) The number of edges in the network at various thresholds for co-occurrence genes network.

An average clustering coefficient was computed at different thresholds (Figure 3.8). At the threshold value of 0, every node was linked. Hence, the average clustering coefficient values was 1. After zero, it declined because many edges in the network were removed. Then it increased again as genes that were isolated were removed. At threshold values approaching 1, the clustering coefficient was close to 1 again because almost all genes and edges had been removed. At the threshold of 1, the value cannot be enumerated because there was no genes or edges in the network. This is a property found in large networks (Elo *et al.*, 2007). When the threshold value was increased, the network tended to be sparse and more cliquish, which can be described as every node in the connected components having connections. The optimal threshold was identified at a sharp increase (Gupta *et al.*, 2006), which was observed in our study at 0.70.

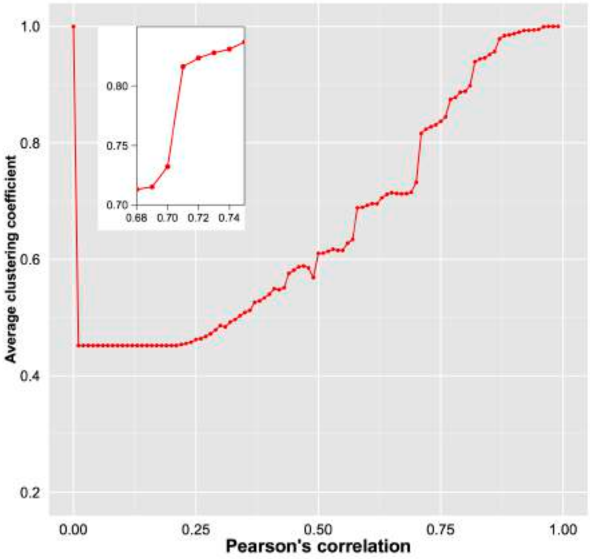


Figure 3.8 Average clustering coefficient was enumerated at various thresholds of absolute values of Pearson's correlation for co-occurrence genes network. An inset figure represents a sharp increase at the optimal threshold value.

The number of connected components was counted at different thresholds (Figure 3.9). The number of connected components increased gradually from zero after that. As the threshold increased, edges were removed from the network and the network became sparser, resulting in the number of connected components increasing. The threshold was determined at the steepest slope of the number of connected components in the network (Fukushima *et al.*, 2011). In this study, there were two sharp increases at Pearson's correlation 0.70 and 0.80.

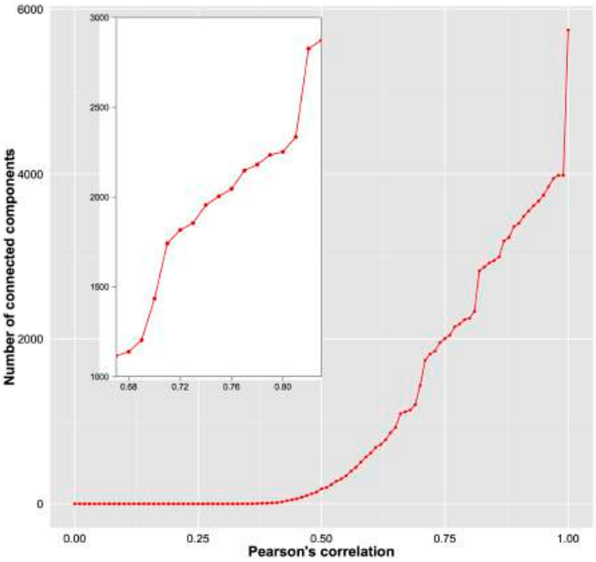


Figure 3.9 Number of connected components were counted at various thresholds of absolute values of Pearson’s correlation for co-occurrence genes network. An inset figure represents two sharp increases at two threshold values, Pearson’s correlation of 0.70 and 0.80.

The density of network was calculated at various thresholds (Figure 3.10); at zero, all genes in the network were connected. The density value, which is the ratio between the actual number of edges and the number of possible edges in the network, was equal to 1. Above zero, the density of network decreased suddenly because edges with correlation values lower than the threshold were removed. The network density value was reduced to the lowest network density due to removal of isolated genes and edges having Pearson’s correlation lower than the threshold of the network. After that, the network density increased. At high threshold values, the number of nodes and the number of edges decreased, resulting in the number of possible edges in the network being not much different from that found in the beginning. This led to an increment in network density. The threshold was seen at the lowest value (Aoki *et al.*, 2007, Ozaki *et al.*, 2010). The threshold was selected at 0.80 by this principle. It was also found that correlation values of 0.71-0.86 had values of network density lower than 0.009.

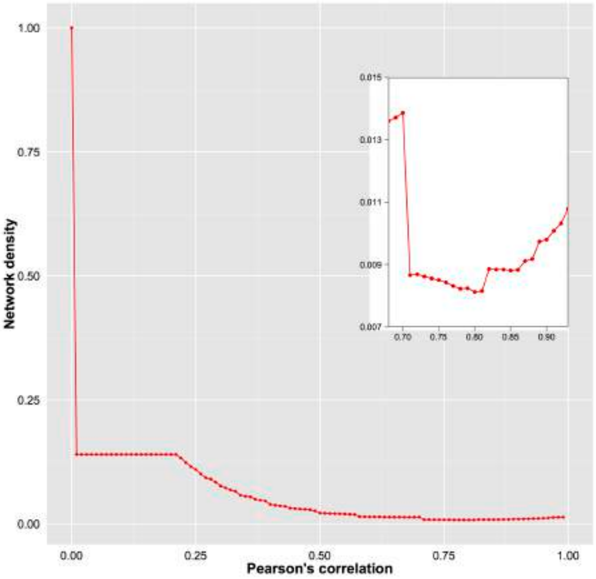


Figure 3.10 Network density was measured at various thresholds of absolute value of Pearson's correlation for the co-occurrence genes network. An inset figure represents thresholds with density values lower than 0.009.

Threshold selection using spectral graph theory was based on decomposing the network into eigenvalues and eigenvectors, conducted on the largest connected components in the network (Perkins *et al.*, 2009) because these contained the majority of genes in the network (Figure 3.11). When doing the spectrum analysis on the largest component, the smallest eigenvalue was zero and the rest were nonzero. The smallest nonzero eigenvalue is called "algebraic connectivity" (λ_1). Lower algebraic connectivity represented the network becoming sparser or in nearly-disconnected components (Ding *et al.*, 2001). The algebraic connectivity identified an optimal threshold at its lowest value. The algebraic connectivity value was computed at different Pearson's correlation thresholds (Figure 3.12). The co-occurring gene network reached the lowest algebraic connectivity at 0.80 Pearson's correlation. Some of the thresholds computed represented relatively low values of algebraic connectivity (less than 0.05 as shown in 🟥 in Figure 3.12). This method provided many possible thresholds.

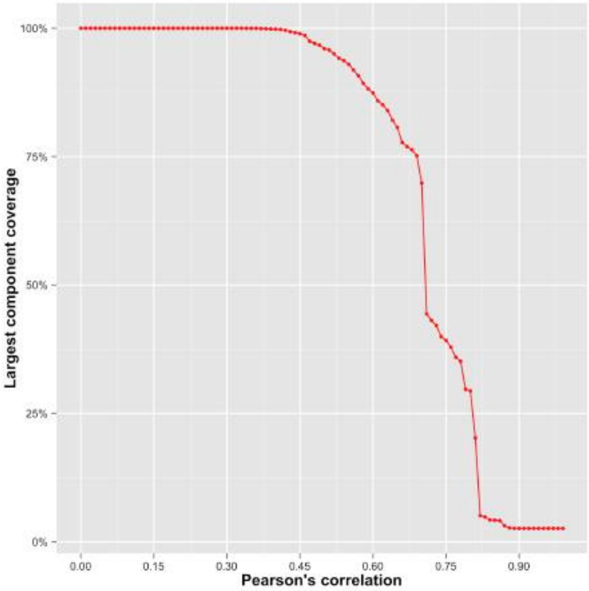


Figure 3.11 Largest component coverage was computed by the percentage of genes included in the largest connected components at various thresholds of absolute value of Pearson's correlation.

Spectral clustering was performed next to detect the threshold. An eigenvector associated with λ_1 (defined as v_1) was selected and sorted in ascending order as a step function (Figure 3.13). A number of clusters were detected by a sliding window approach 5 elements wide. The number of clusters was quantified at different thresholds (Figure 3.14). Forty-four clusters, which is the highest number of clusters, were detected at a correlation value of 0.74. As a result, 0.74 was chosen as the optimal threshold using spectral clustering.

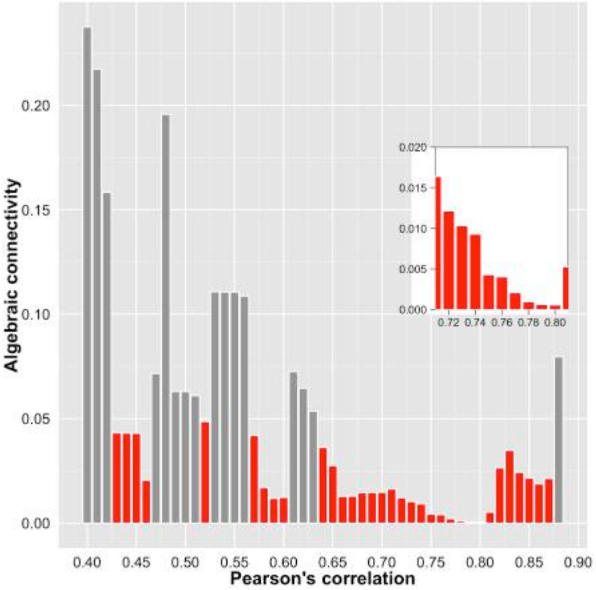


Figure 3.12 Algebraic connectivity computed at various thresholds of absolute value of Pearson's correlation for co-occurrence genes network. Pearson's correlation values in 🟥 also exhibited relatively low values of algebraic connectivity (less than 0.05).

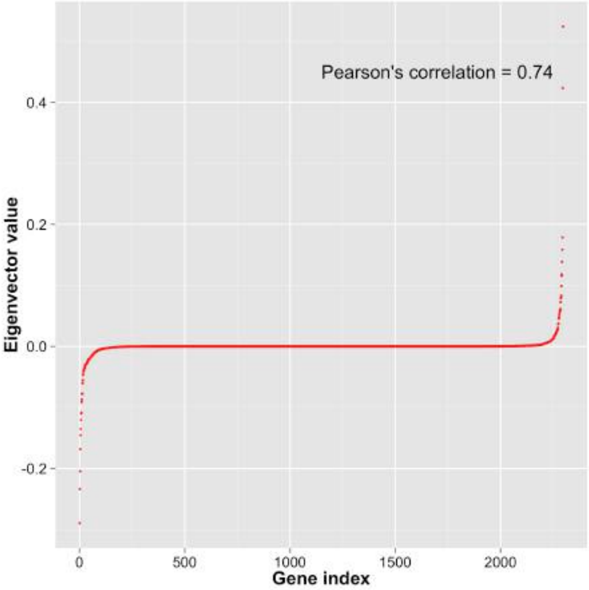


Figure 3.13 Sorted eigenvector associated with λ_1 in the co-occurrence network at threshold of 0.74. The associated eigenvector was plotted for each gene.

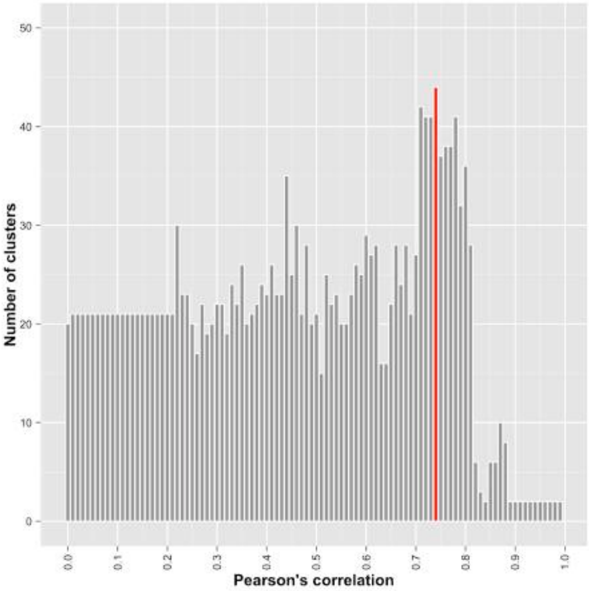


Figure 3.14 Number of clusters computed based on spectral graph theory for co-occurrence genes network at various thresholds of absolute value of Pearson's correlation. The highest number of clusters was observed at 0.74 (🟥), which was therefore determined to be the threshold.

When comparing the threshold determined by different methods (Table 3.2), it was observed that the optimal threshold ranged from 0.70 to 0.80. Pearson's correlation of 0.70 was selected by Average Clustering coefficient, Number of connected components, and Algebraic connectivity. Ultimately the optimal threshold was chosen at Pearson's correlation = 0.70.

Table 3.2 Thresholds were determined by different methods

Selection methods	Selected threshold
Average Clustering coefficient	0.70
Number of connected components	0.70, 0.80
Network density	0.71-0.86
Algebraic connectivity	0.64-0.87
Spectral clustering	0.74

3.5.3 Results on gene abundance in the population

Pearson's correlation can be strong between two genes rarely or generally found in the population (Figure 3.15), but these values are based on very few strains so may be unreliable. In order to explore effects of gene abundance on the co-occurrence network, the co-occurrence gene network was then created without removing genes carried by fewer than 6 or more than 79 strains.

Genes with low and high abundance played a major role in the structure of the network (Figure 3.16). This may lead to network relationships that are not biologically significant. In the case of strong correlation among genes with high abundance, many of these genes were probably core genes that were present in all strains but were sometimes not detected in strains with low sequence coverage. These genes with high abundance were therefore removed from the network. Strong correlations among genes with low abundance could possibly reflect rare clusters of genes operating under the same conditions or confined to a group of closely-related strains. To avoid problems caused by missing genes during the sequencing process or false negative sequencing results, all genes with low abundance were also removed.

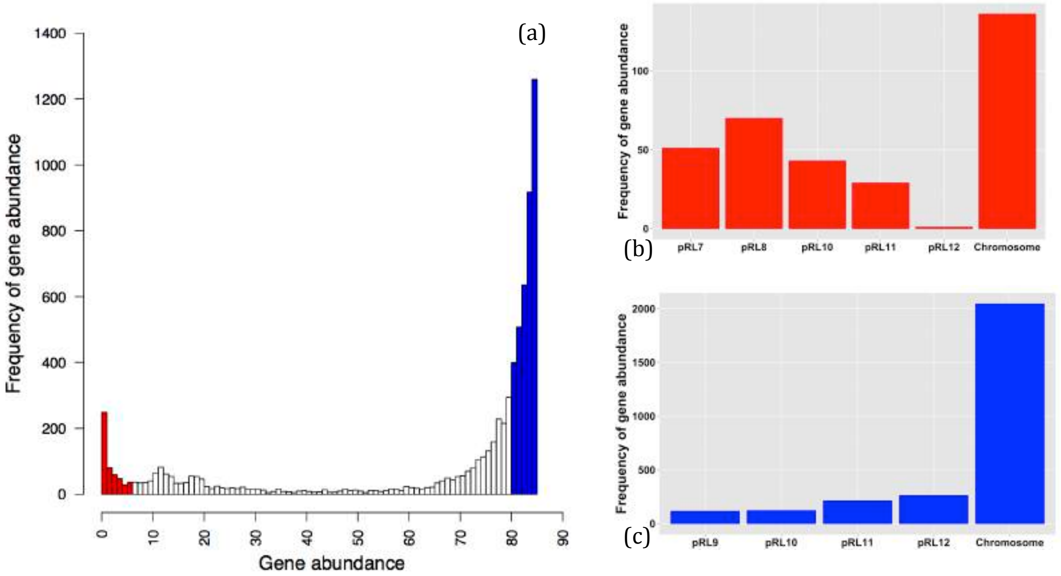


Figure 3.15 Gene abundance in the population. (a) Strains numbers carrying genes in red and blue are fewer than 6 and more than 79, respectively. Two bar plots on the right represent frequency of gene abundance by each replicon; (b) genes carried by fewer than 6 strains and (c) genes carried by more than 79.

Genes with low abundance were generally found on every replicon apart from pRL9 (Figure 3.15(b)). Genes with high abundance can be found on every replicon apart from pRL7 and pRL8 (Figure 3.15(c)). This is possibly because pRL7 and pRL8 are relatively small conjugative plasmids containing genes carried by just some strains in the population (Young *et al.*, 2006). It is not surprising then that a majority of them were excluded from the analysis.

3.5.4 Results on network construction

The resulting network was constructed using Pearson's correlation threshold of 0.70 with a p-value ≤ 0.05 (Figure 3.16 (a)) and excluding genes carried by less than 6 or greater than 79 strains (Figure 3.16 (b)). The co-occurrence gene network contains 2,663 genes, 33,318 interactions (consisting of 1,150 anti co-occurrence interactions and 32,168 co-occurrence interactions), 0.7152 average clustering coefficient, and 6.4636 average path length with diameter of 20. The largest connected component in the network consists of 1,211 genes. There are connected components containing 44, 22, 21, 14, 9, and 5 nodes (one of each). 8-node, 7-node, 6-node, and 4-node connected components are found 2, 3, 4, and 6 times respectively, in the resulting network. There are 14 triplets, 22 pairs, and 1,122 singletons.

Co-occurrence relationships are mostly observed between neighbouring genes on the same replicon. A spatial arrangement of these neighbouring genes may reflect their translocation in the population controlled by the same elements or their participation in the same biological process. However, missing genes in an ordering in some clusters were seen because they did not qualify under the abundance criterion. Most genes connected to each other by co-occurrence interactions (positive correlation values) are carried by similar numbers of strains in the population and presumably have similar profiles (Figure 3.17(a)). In contrast, the genes connected with anti co-occurrence interactions (negative correlation values) were inversely related in abundance (Figure 3.17(a)).

To find co-occurrence and anti co-occurrence relationships on distinct replicons, genes in the network were coloured by seven colours, each colour denoting a different replicon (Figure 3.17(b)). Co-occurrence relationships are observed between chromosome-chromosome, chromosome-plasmid (especially large plasmids like pRL9 and pRL10), chromosome-chromid, chromid-plasmid, plasmid-plasmid, and chromid-chromid genes. Co-occurrence relationships between genes on the chromosome and small plasmids, like pRL7 and pRL8, were not observed. The genes on pRL7 and pRL8 are accessory genes and rarely found in the population, in contrast to chromosomal genes most of which are carried by most isolates in the population.

Anti co-occurrence relationships were found within and between replicons consisting of chromosome-chromosome, chromid-chromid, plasmid-plasmid, and plasmid-chromid genes. It is noticeable that there is no anti co-occurrence relationship observed between chromosome-chromid and chromosome-plasmid genes.

The anti co-occurrence relationship is basically considered as the relationship of genes with replaceable functions. It appears that the functions of plasmid/chromid and accessory chromosomal genes are not interchangeable. Most of the co-occurrence relationships were observed between neighbouring genes, in contrast to the anti co-occurrence that can be observed genes between distant genes.

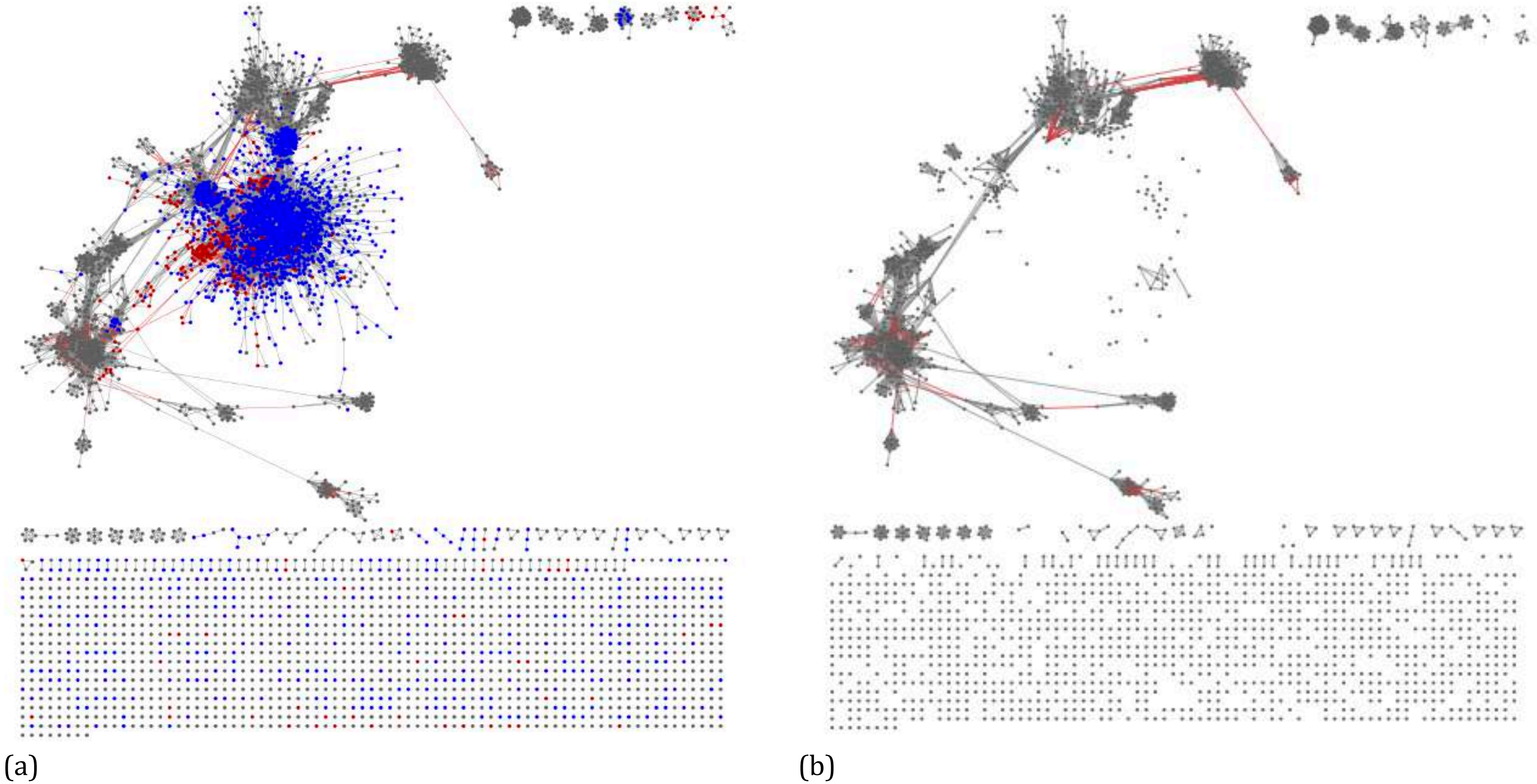


Figure 3.16 Co-occurrence gene networks derived from 72 strains of the rhizobium population. (a) The network consists of 5,748 genes (nodes) and 135,385 interactions (edges). Red (●) and blue (●) nodes represent genes with low (genes carried by fewer than 6 strains) and high (genes carried by more than 79 strains) abundance in the population, respectively. Grey (—) and red (—) edges specify co-occurrence and anti co-occurrence interactions, respectively (b) The resulting network after removing low and high abundance genes consists of 2,663 genes and 33,318 interactions.

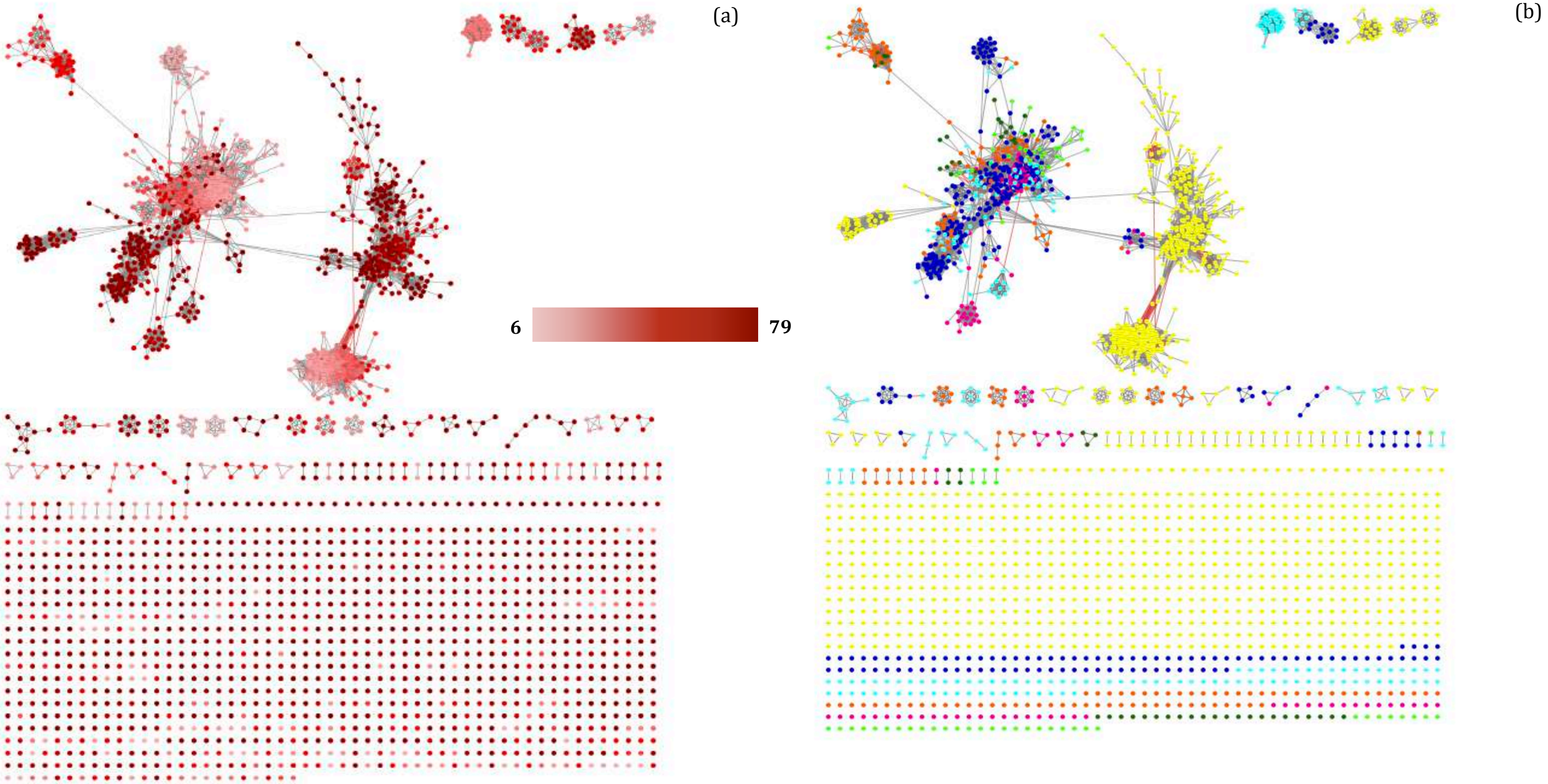


Figure 3.17 The resulting co-occurrence genes network. The co-occurrence genes network consists of 2,663 genes and 33,318 interactions. Grey (—) and red (—) edges specify co-occurrence and anti co-occurrence interactions, respectively (a) Nodes are coloured by their abundance in the population. (b) Nodes as genes are coloured by 7 different replicons (● chromosome, ● pRL7, ● pRL8, ● pRL9, ● pRL10, ● pRL11, and ● pRL12).

3.5.5 Results on co-occurrence relationships

The largest subnetwork contains co-occurrence and anti co-occurrence relationships (Figure 3.17) whereas the rest of the network contains predominantly co-occurrence relationships. Co-occurrence relationships are hence described as two different original subnetworks consisting of 1) the subnetworks excluding the largest subnetwork and 2) the largest subnetwork.

3.5.5.1 Results on co-occurrence of genes in the network excluding the largest subnetwork

1,202 subnetworks are observed in the network excluding the largest network. The second largest subnetwork has 44 genes. Most of subnetworks are singletons. The distribution of subnetwork size is shown in Figure 3.18. It was observed that a majority of them included genes on a single replicon (Figure 3.19).

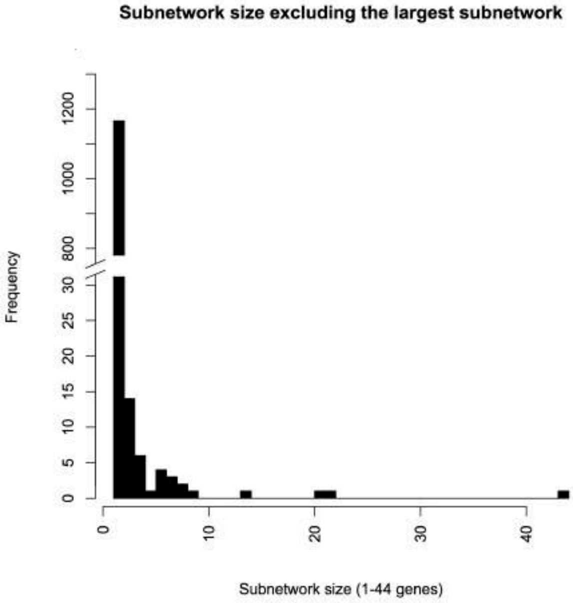


Figure 3.18 Distribution of subnetwork sizes in the network excluding the largest subnetwork. The largest subnetwork has 44 genes. 8-node, 7-node, 6-node, and 4-node connected components were found 2, 3, 4, and 6 times respectively, in the resulting network. There were 14 triplets, 22 pairs, and 1,122 singletons.

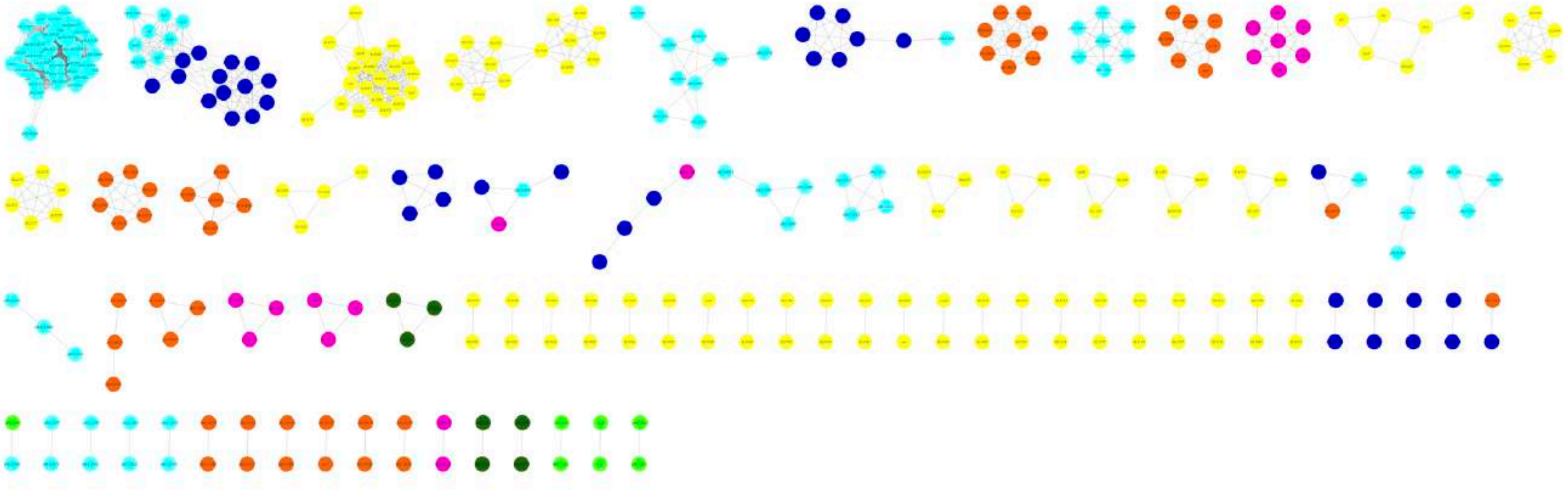


Figure 3.19 Subnetworks in the network excluding the largest subnetwork and isolated nodes Grey (—) and red (—) edges specify co-occurrence and anti co-occurrence interactions, respectively. Nodes represent genes and are coloured by 7 different replicons (● chromosome, ● pRL7, ● pRL8, ● pRL9, ● pRL10, ● pRL11, and ● pRL12).

3.5.5.2 Results on co-occurrence relationships in the largest subnetwork

In contrast to the subnetworks observed in the previous section, the largest subnetwork, containing 1,211 nodes, includes both co-occurrence and anti co-occurrence relationships. Focusing on co-occurrence relationships in the largest subnetwork, anti co-occurrence relationships were first removed from the subnetwork. Then the remaining subnetworks contain only co-occurrence relationships to explore. The remaining co-occurrence subnetworks (Figure 3.20) include one subnetwork each of 480 genes, 374 genes, 272 genes, 28 genes, 19 genes, 12 genes, and 11 genes, four subnetworks of pair of genes, and seven of singletons.

Since three large subnetworks having more than 100 genes are observed, a community detection technique is applied to those three large subnetworks to decipher co-occurrence relationships in the large network. Results on community detection of each large subnetwork are as described below.

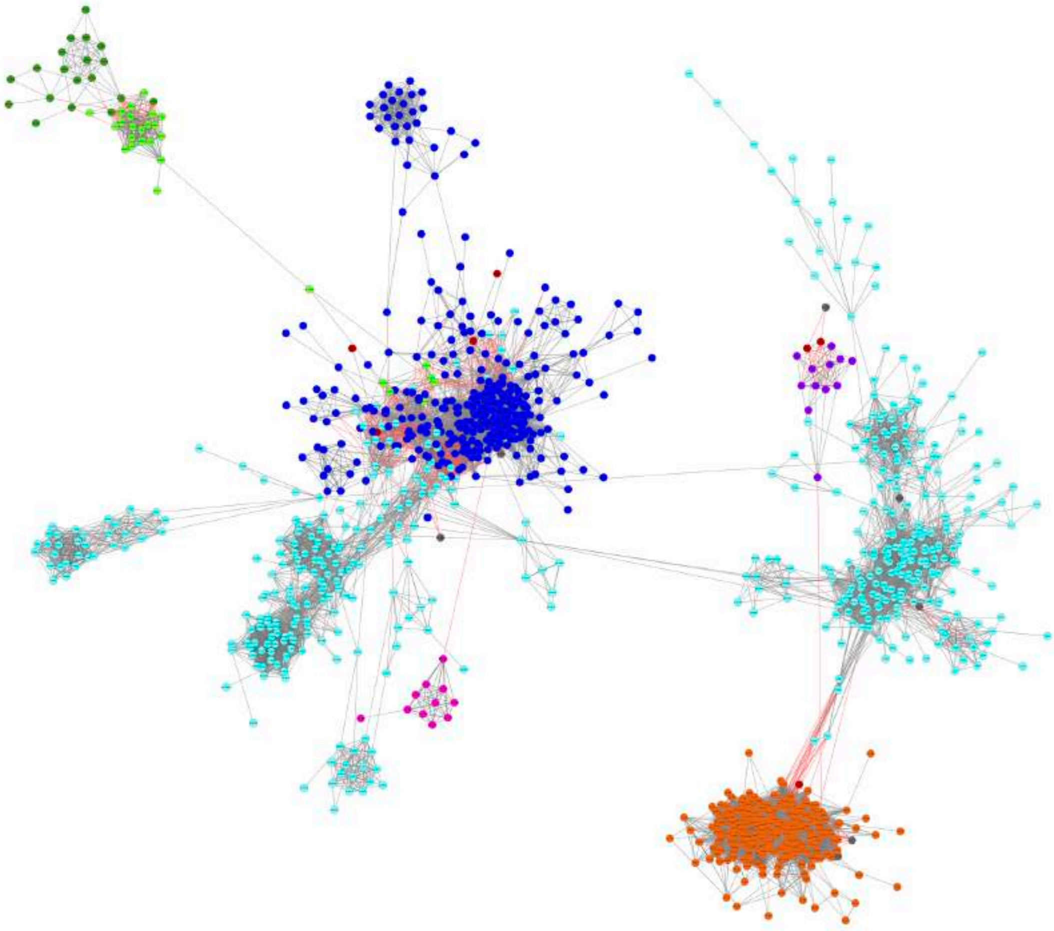


Figure 3.20 The largest subnetwork. Nodes are coloured by clusters of genes connected with co-occurrence interactions. Grey (—) and red (—) edges specify co-occurrence and anti co-occurrence interactions, respectively (● 480-gene subnetwork, ● 374-gene subnetwork, ● 272-gene subnetwork, ● 28-gene subnetwork, ● 19-gene subnetwork, ● 12-gene subnetwork, ● 11-gene subnetwork, four ● 2-gene subnetworks and seven ● singletons).

The 480-gene subnetwork (● in Figure 3.20) contains 27 pRL9, 30 pRL10, 36 pRL11, 114 pRL12 and 273 chromosome genes. Community detection reveals that there are 36 communities (Figure 3.21). The largest community has 77 genes, all of which are chromosomal genes. The second largest community, whose size is not much different from the largest community, has 60 genes from plasmids and chromids. Investigating the type of genes in every single community shows that relationships within each community are strongly conserved on types of replicon. It is observed that chromosomal gene communities and chromid-plasmid gene communities are distinct (Table 3.3).

These imply that the distributions of chromosomal accessory genes and of plasmid-chromid genes are independent.

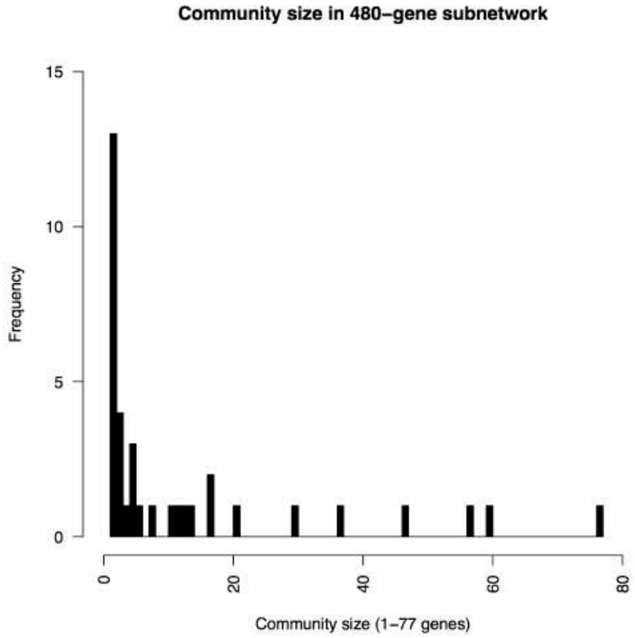


Figure 3.21 Distribution of community sizes in the 480-gene subnetwork. Community sizes of this cluster in the range of 1-77 genes are shown. More than half of the communities contain ≤ 5 genes. The remaining communities contain 6-77 genes. The largest community has 77 genes.

Table 3.3 Location of genes found in each community of the 480-gene subnetwork

Number of genes	Number of communities	Location of genes
77	1	Chromosome (1)
57	1	Chromosome (1)
46	1	pRL12 (1)
37	1	Chromosome (1)
24	1	pRL12 (1)
23	1	pRL12 (1)
21	1	Chromosome (1)
17	2	Chromosome (1) pRL9 (1)
14	1	Chromosome (1)
13	1	Chromosome (1)
12	1	pRL10 (1)

Number of genes	Number of communities	Location of genes
11	2	pRL11 (2)
7	1	pRL11 (1)
6	2	pRL10 (1)
		pRL12 (1)
5	4	Chromosome (2)
		pRL10 (1)
		pRL12 (1)
4	3	Chromosome (1)
		pRL9 (1)
		pRL11 (1)
3	4	Chromosome (2)
		pRL12 (2)
2	11	Chromosome (7)
		pRL9 (1)
		pRL10 (2)
		pRL12 (1)
1	15	Chromosome (3)
		pRL9 (4)
		pRL10 (3)
		pRL11 (3)
		pRL12 (2)

The subnetwork of 374 genes (● in Figure 3.20) has plasmid (pRL7, pRL8, pRL9, and pRL10) and chromid (pRL11 and pRL12) genes. Forty communities are observed in the subnetwork of 374 genes. Distribution of community size of the 374-gene subnetwork is shown in Figure 3.21. The largest community has 170 genes. Among dense co-occurrence relationships in the 170-gene community (● in Figure 3.23), some pRL9 and pRL12 genes from this community have been reported as genomic islands carried by genospecies B (appendix table I.I and I.II). Another community reflecting species specificity is community of 26 pRL12 genes (● in Figure 3.23). 9 out of 26 genes are absent in genospecies C.

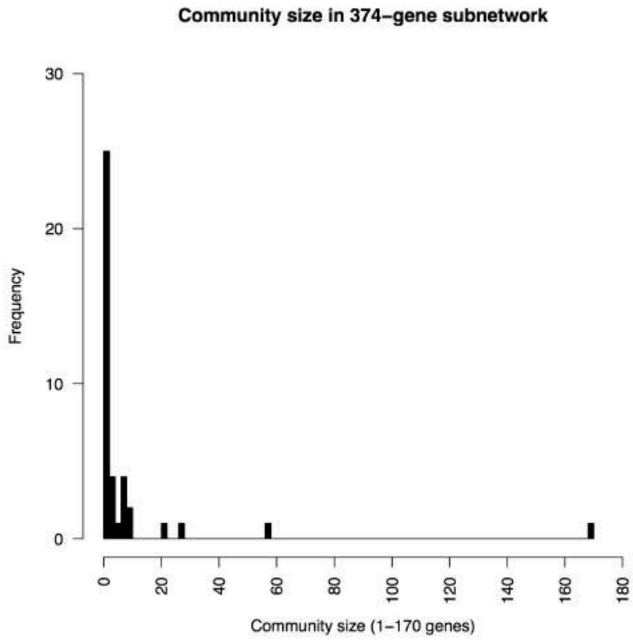


Figure 3.22 Distribution of community sizes in the 374-gene subnetwork. Community sizes of this subnetwork range from 1 to 170 genes as shown in a histogram. Most communities (21) are singletons. Apart from the largest community, which has 170 genes, the rest of the communities contain 2-57 genes.

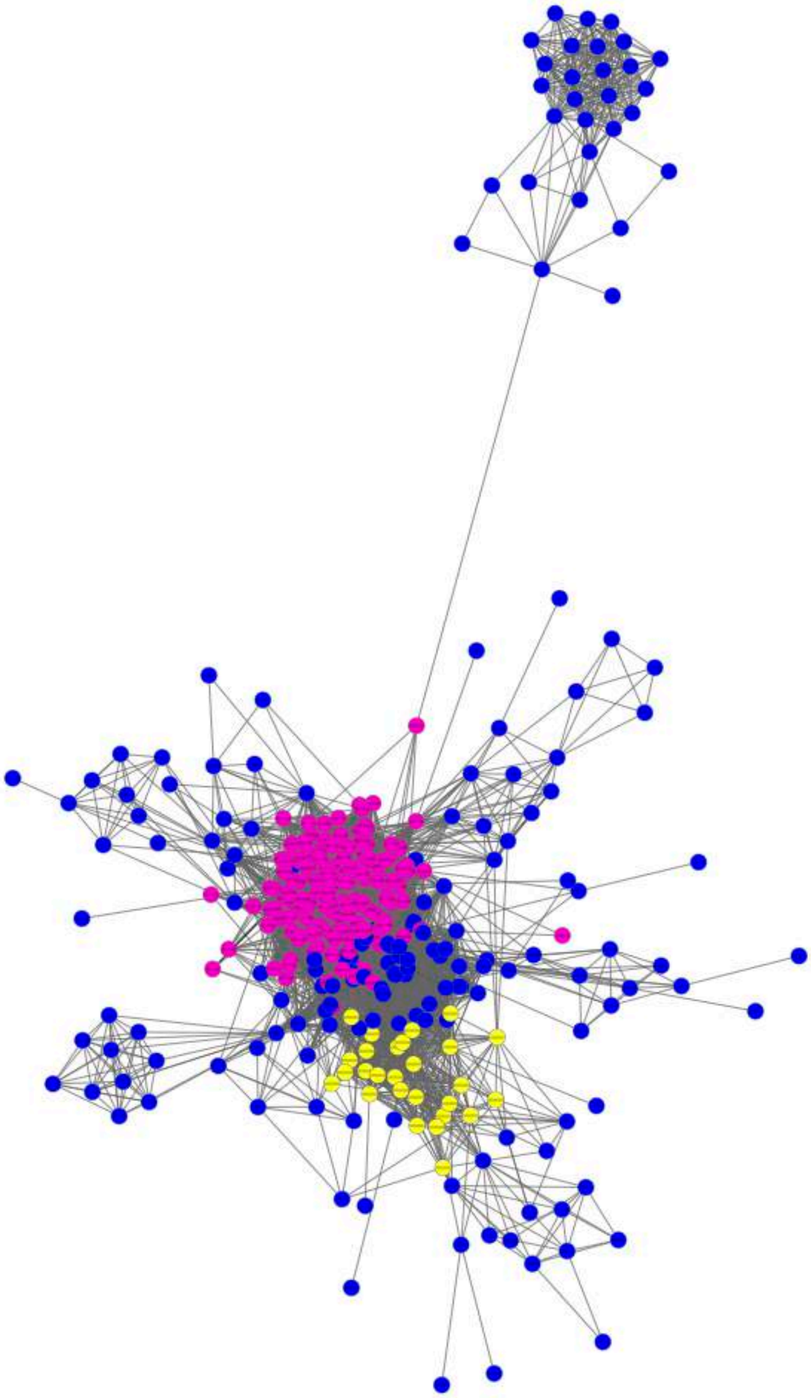


Figure 3.23 The 374-gene subnetwork. A community of 170 genes is highlighted in magenta (●), which contains genospecies B specific genes. A community of 26 genes is highlighted in yellow (●), which contains genes absent in genospecies C.

The 272-chromosomal-gene subnetwork comprised 9 communities. It is noticeable that the chromosomal genes in this subnetwork are sporadically distributed across strains, which is different from the general pattern of chromosomal genes that are ubiquitously distributed. However, the detected communities in this subnetwork do not represent specificity to a particular genospecies or any symbiovar. The distribution of community size of the 272-gene subnetwork is shown in Figure 3.24.

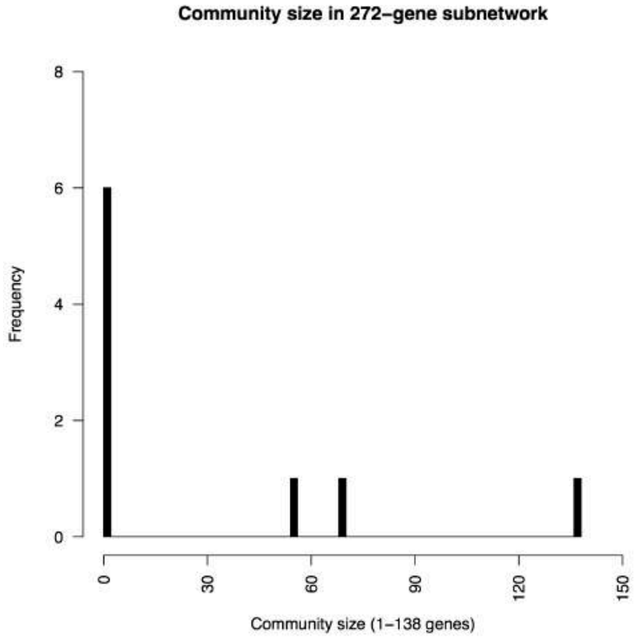


Figure 3.24 Distribution of community sizes in the 272-gene subnetwork. Community sizes of this subnetwork range from 1 to 138 genes as shown in a histogram. This subnetwork has 3 singletons, 3 pairs of genes, one each of 56-gene, 69-gene, and 138-gene communities. The largest community has 138 genes.

The remaining subnetworks, having fewer than 30 genes, are explored individually.

The 28-gene subnetwork (Figure 3.25) is a subnetwork of pRL8 and pRL10 genes. All five pRL8 genes included in this subnetwork have been reported as symbiovar *viciae* specific genes (Table 3.2). These five genes are present only in symbiovar *viciae* strains. The rest of the genes in this subnetwork are pRL10 genes. Some of them are symbiosis genes (Table 3.4), which were relevant to nodulation of this symbiovar. These genes then are used to distinguish symbiovar *viciae* and *trifolii*. It could be inferred that genes in the 28-gene subnetwork are abundant in a particular symbiovar, namely *viciae*. It was also found that the subnetwork also contained a cluster of pRL100162A, pRL100163, pRL100164 (*rhiI*), pRL100169 (*rhiA*), pRL100170 (*rhiB*), pRL100171 (*rhiC*), and pRL100172 (*rhiR*), which are indirectly involved in nodulation (Rodelas *et al.*, 1999, Wisniewski-Dye *et al.*, 2002). Consequently, genes in the subnetwork were functionally related to the nodulation.

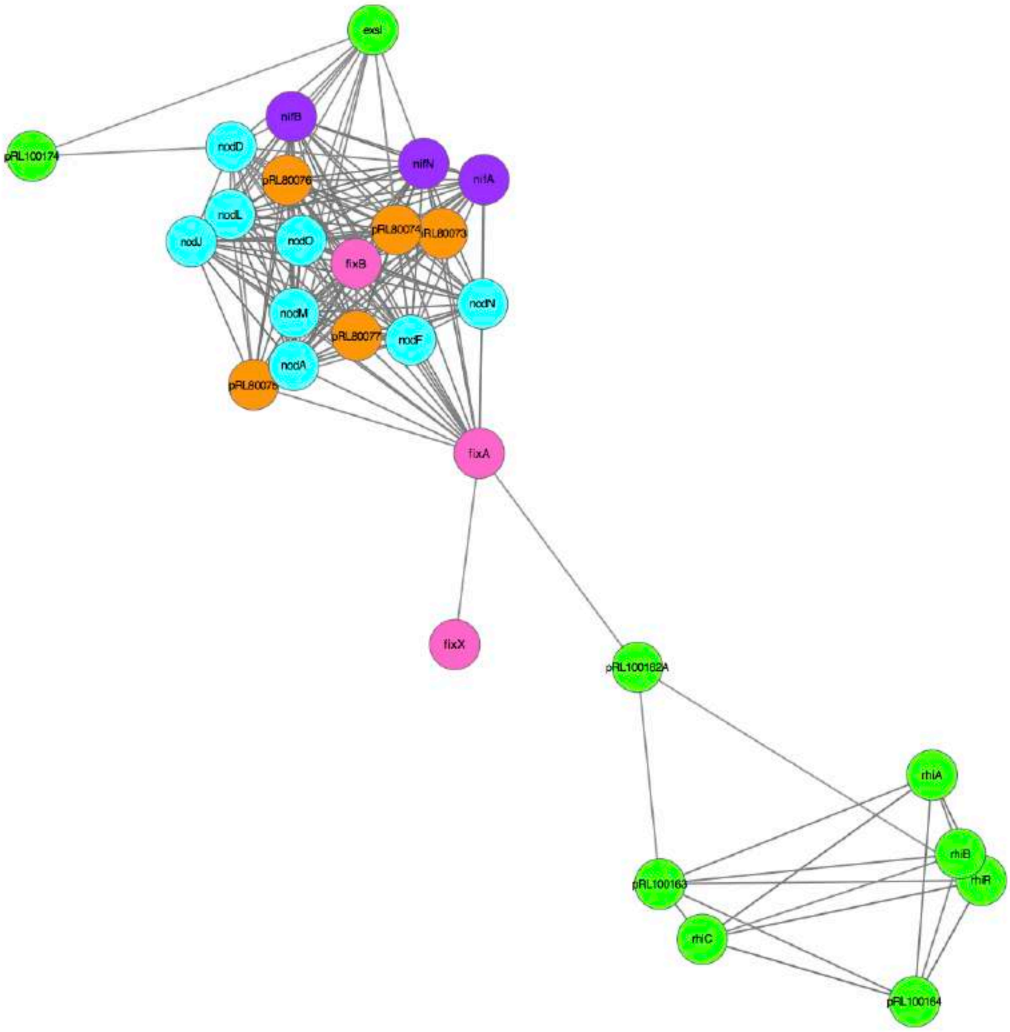


Figure 3.25 The 28-gene subnetwork This subnetwork represents co-occurrence relationships between pRL8 and pRL10 genes, all of which are abundant in bv. *viciae*. These genes are symbiosis genes on pRL10 (including *nod* (●), *nif* (●), and *fix* (●) genes) and Bvs genes (●) on pRL8.

Table 3.4 List of genes in the 28-gene subnetwork. Locus tag of symbiosis genes is in bold.

Locus tag	Gene symbol	Protein accession	Annotated function
pRL80073	*bvs1*	YP_770968.1	cysteine desulfurase
pRL80074	*bvs2*	YP_770969.1	LysR family transcriptional regulator
pRL80075	*bvs3*	YP_770970.1	endoribonuclease L-PSP
pRL80076	*bvs4*	YP_770971.1	aliphatic nitrilase

Locus tag	Gene symbol	Protein accession	Annotated function
pRL80077	*bvs5*	YP_770972.1	molybdenum-binding oxidoreductase
pRL100151	*exsI*	YP_770430.1	transccriptional regulator
pRL100158	*nifN*	YP_770436.1	nitrogenase molybdenum-cofactor biosynthesis protein NifN
pRL100162A	-	YP_770441.1	asparagine synthase
pRL100163	-	YP_770442.1	asparagine synthase
pRL100164	*rhiI*	YP_770443.1	autoinducer synthesis protein
pRL100169	*rhiA*	YP_770448.1	hypothetical protein
pRL100170	*rhiB*	YP_770449.1	rhizosphere induced protein RhiB
pRL100171	*rhiC*	YP_770450.1	hypothetical protein
pRL100172	*rhiR*	YP_770451.1	transcriptional regulator
pRL100174	-	YP_770453.1	hypothetical protein
pRL100175	*nodO*	YP_770454.1	nodulation protein
pRL100179	*nodN*	YP_770457.1	nodulation protein
pRL100180	*nodM*	YP_770458.1	glucosamine--fructose-6-phosphate aminotransferase
pRL100181	*nodL*	YP_770459.1	nodulation protein
pRL100183	*nodF*	YP_770461.1	nodulation protein F
pRL100184	*nodD*	YP_770462.1	nodulation protein D
pRL100185	*nodA*	YP_770463.1	acyltransferase NodA
pRL100189	*nodJ*	YP_770467.1	nodulation protein
pRL100195	*nifB*	YP_770473.1	FeMo cofactor biosynthesis protein
pRL100196	*nifA*	YP_770474.1	nifA transcriptional regulator
pRL100197	*fixX*	YP_770475.1	ferredoxin-like protein
pRL100199	*fixB*	YP_770477.1	FixB electron transfer protein
pRL100200	*fixA*	YP_770478.1	FixA electron transfer protein

The 19-gene subnetwork contains 2 genes on pRL7 and 17 genes on pRL10 (Figure 3.26). An exploration in this subnetwork reveals that three pRL10 genes carried in this subnetwork, which are called *attKLM* (pRL100134, pRL100135, and pRL100136), have been reported as involved in hydroxybutyrate metabolism (Chai *et al.*, 2007, Prell *et al.*, 2009, White *et al.*, 2009).

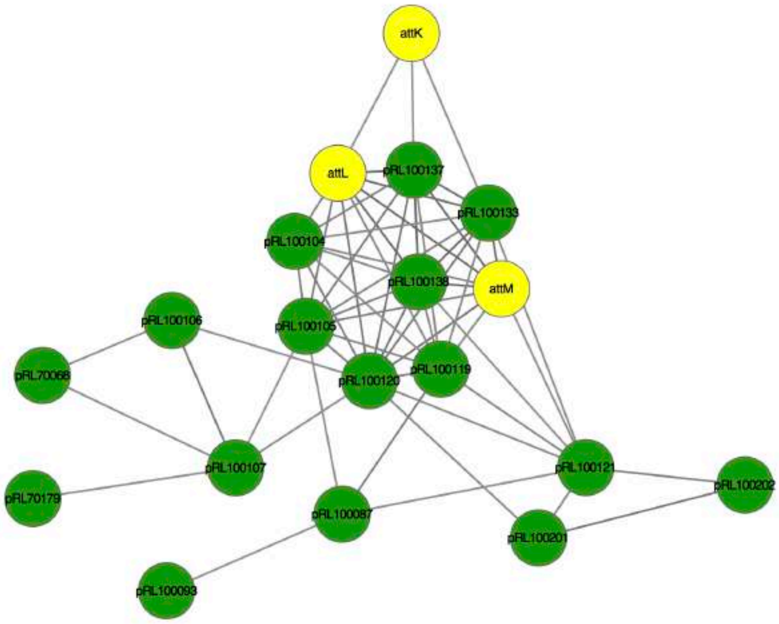
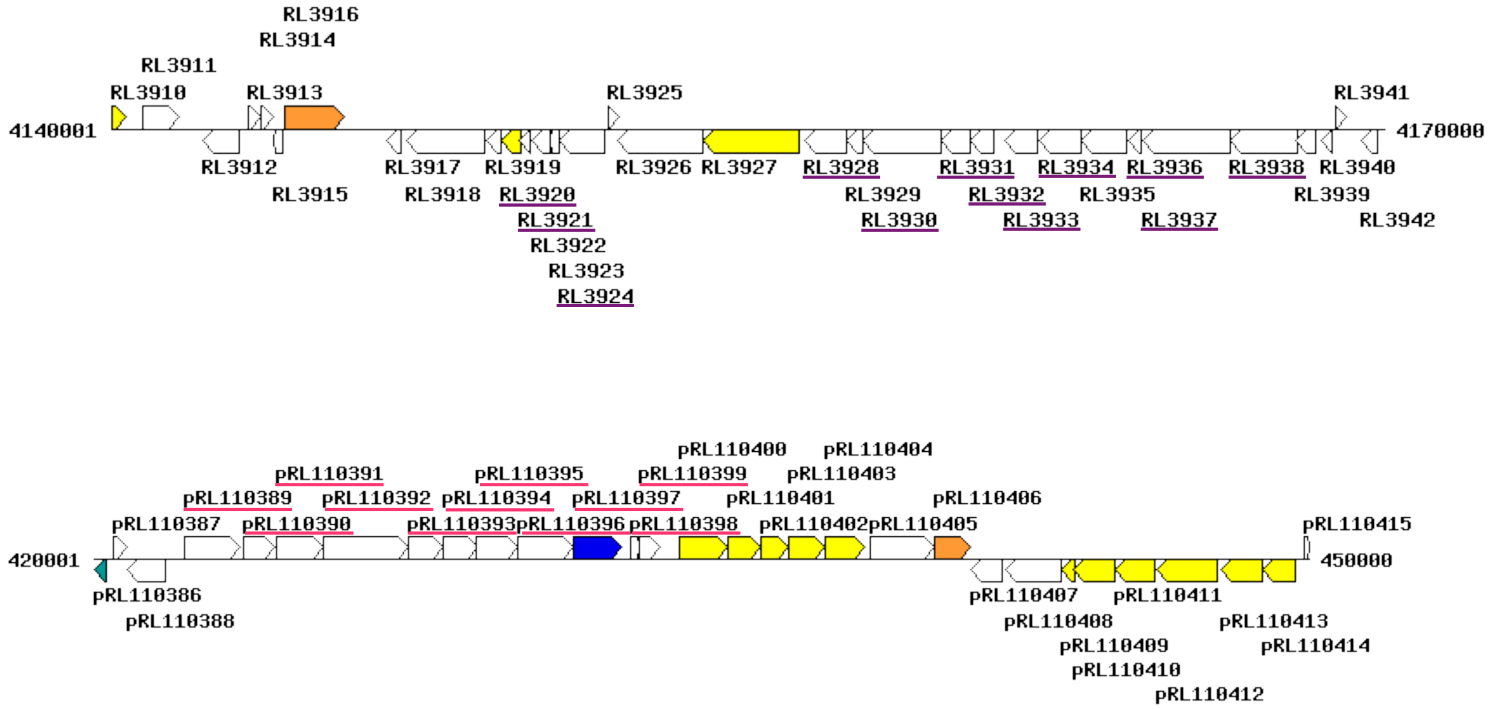


Figure 3.26 The 19-gene subnetwork. This subnetwork contains three pRL10 genes (●) relating to γ-hydroxybutyrate metabolism.

The 12-gene (●) subnetwork (Figure 3.20) is a chromosomal gene subnetwork and includes RL3920, RL3921, RL3924, RL3928, RL3930-RL3934, and RL3936-RL3938. The 11-gene (●) (Figure 3.20) subnetwork is a pRL11 gene subnetwork and includes pRL110389-pRL110399. It is observed that those two subnetworks are subnetworks of neighbouring genes (Figure 3.27).



(a.)

(b.)

Figure 3.27 Genome map of the two subnetworks from KEGG database (Tanabe *et al.*, 2002) (a.) the 12-gene (underlined 🟣) subnetwork and (b.) the 11-gene (underlined 🔴) subnetwork. The numbers at the end represent genomic location in base pairs (bp). Genes are shown as arrows representing the direction of transcription and coloured by KEGG pathway categories (🟩 lipid metabolism, 🟦 carbohydrate metabolism, 🟧 amino metabolism, 🟨 environmental information processing and ☐ unclassified).

3.5.6 Results on anti co-occurrence relationships

Anti co-occurrence relationships are identified by negative values of correlation between pairs of genes. The anti co-occurrence relationships are mainly seen in the largest subnetwork. An exploration on the largest subnetwork reveals that there are nine types of subnetworks classified by genes contained in each subnetwork, which each subnetwork connected to others by anti co-occurrence relationships (Figure 3.20). These nine types of subnetwork are one each of 480-gene subnetwork, 374-gene subnetwork, 272-gene subnetwork, 28-gene subnetwork, 19-gene subnetwork, 11-gene subnetwork, 12-gene subnetwork, 4 pairs of genes, and 7 singletons. Prominent features of each subnetwork are described as follows:

The largest genes subnetwork contains 480 genes (● in Figure 3.20). Genes in this subnetwork are carried by most of the population and negatively correlated to two subnetworks of genes including 1) the 374-gene (● in Figure 3.20) subnetwork, which is a subnetwork of plasmid (pRL7, pRL8, pRL9, and pRL10) and chromid (pRL11 and pRL12) genes, and 2) the 272-gene (● in Figure 3.20) subnetwork, which is a chromosomal gene subnetwork. Pairs of genes showing anti co-occurrence relationship between the 480-gene and the 374-gene subnetwork were investigated their distribution in the population (Figure 3.28). It was found that the anti co-occurrence relationships between the 480-gene and the 374-gene subnetwork arose because isolates that carried the genes in the 480-gene subnetwork did not carry genes in the 374-gene subnetwork.

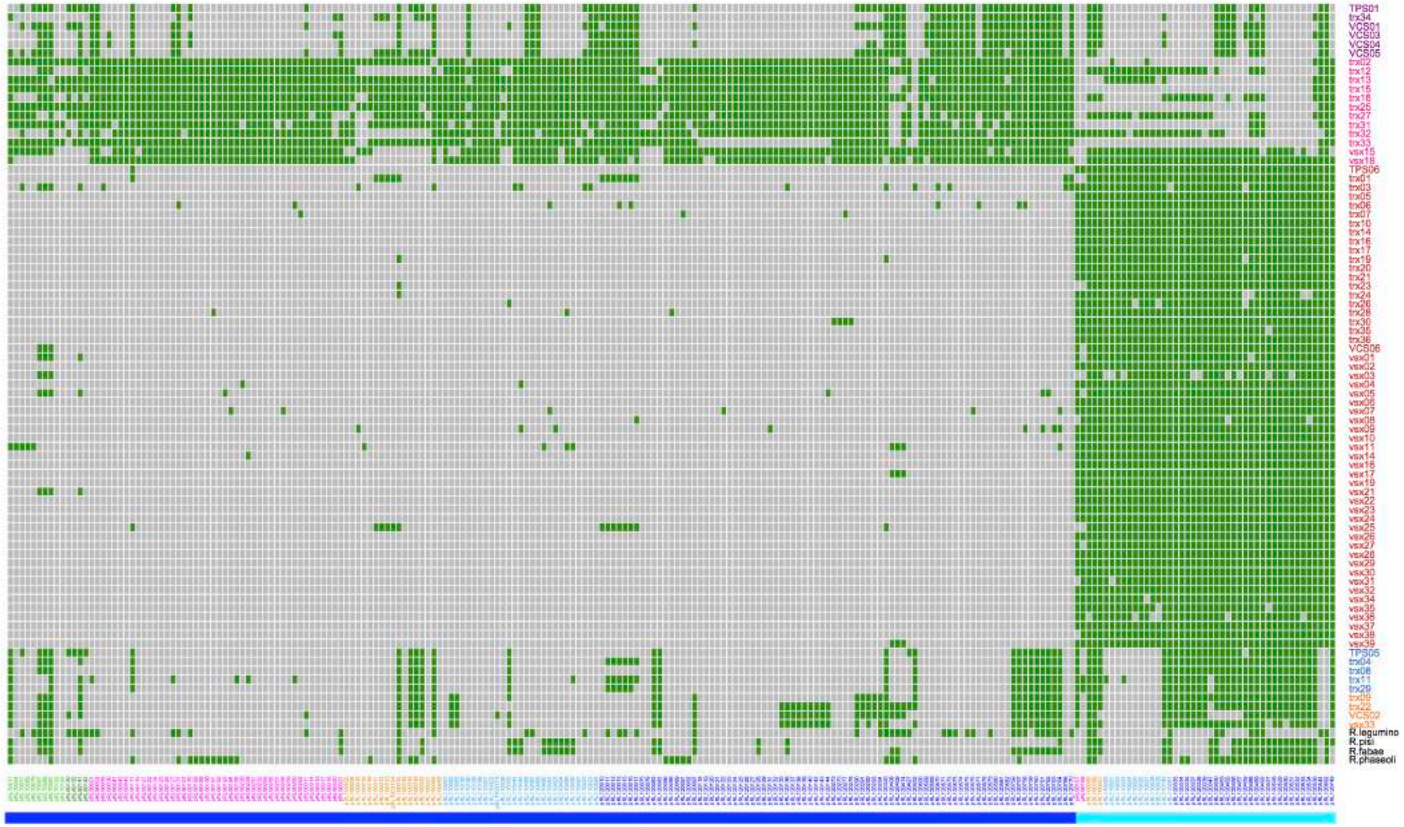


Figure 3.28 Phylogenetic profiles of genes in the 480-gene and the 374-gene subnetworks having anti co-occurrence relationships. The presence of genes is shown in 🟩, the absence of genes is shown in ■. Columns represent genes that have anti co-occurrence relationships and are arranged according to their subnetwork (■ 480-gene and ■ 374-gene subnetwork). Rows represent 85 strains and are arranged according to their respective genospecies (●: A, ●: B, ●: C, ●: D, ●: E) and type strains (●).

Likewise, there are anti co-occurrence relationships between the 480-gene (● in Figure 3.20) and the 272-gene (● in Figure 3.20) subnetworks, both of which are chromosomal gene subnetworks. The 272-gene subnetwork may be carried on genomic islands on the chromosome. It can be observed that those genomic islands were partially absent from strains carrying genes in the 480-gene subnetwork (Figure 3.29).

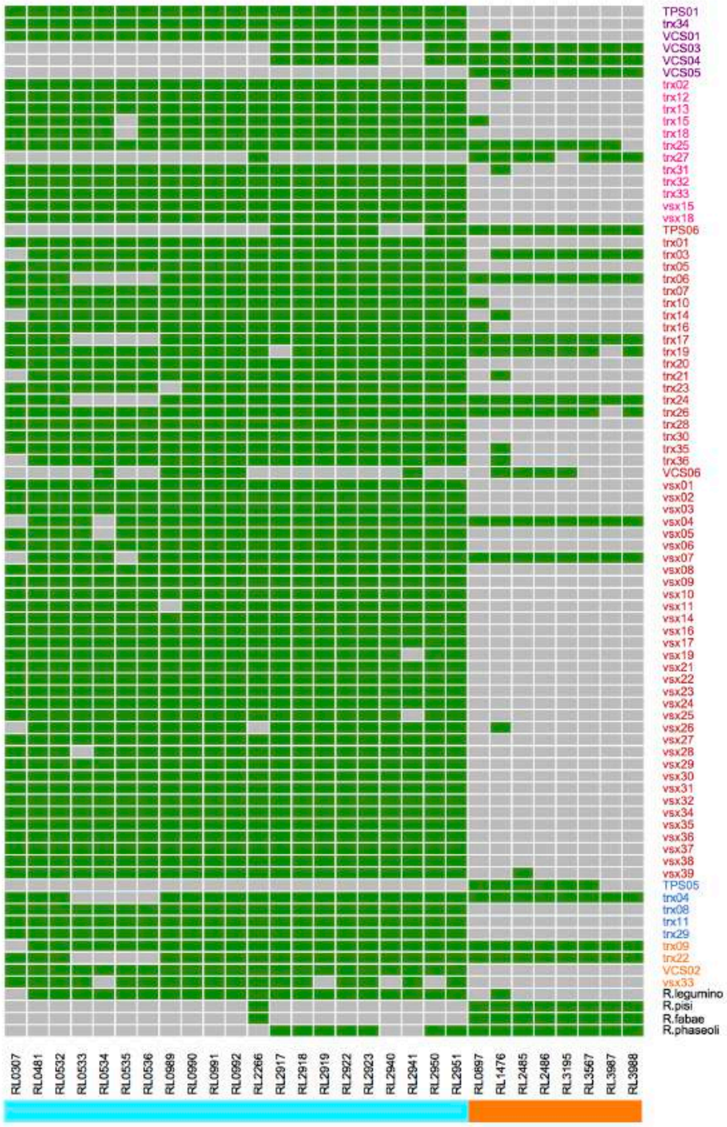


Figure 3.29 Phylogenetic profiles of genes in the 480-gene and the 272-gene subnetworks having anti co-occurrence relationships. The presence of genes is shown in 🟩, the absence of genes is shown in ■. Columns represent genes that have anti co-occurrence relationships and are arranged according to their subnetwork (■ 480-gene and ■ 272-gene subnetwork). Rows represent 85 strains and are arranged according to their respective genospecies (●: A, ●: B, ●: C, ●: D, ●: E) and type strains (●).

The 374-gene (● in Figure 3.20) subnetwork associates negatively with six genes of pRL10 including (pRL100163, pRL100164 (*rhiI*), pRL100169 (*rhiA*), pRL100170 (*rhiB*), pRL100171 (*rhiC*), and pRL100172 (*rhiR*)) in the 28-gene (● in Figure 3.20) subnetwork. Genes in the 374-gene subnetwork are largely absent in genospecies C, in contrast to the six genes in the 28-gene subnetwork (Figure 3.30).

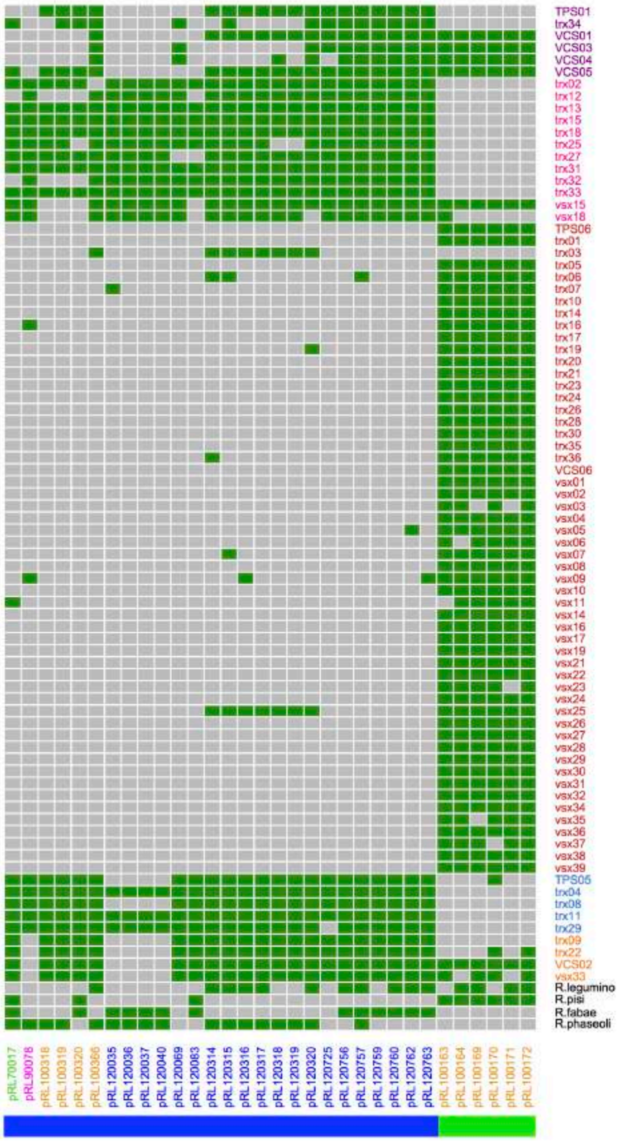


Figure 3.30 Phylogenetic profiles of genes in the 374-gene and the 28-gene subnetworks having anti co-occurrence relationships. The presence of genes is shown in ■, the absence of genes is shown in ■. Columns represent genes that have anti co-occurrence relationships and are arranged according to their subnetwork (■ 374-gene and ■ 28-gene subnetwork). Rows represent 85 strains and are arranged according to their respective genospecies (●: A, ●: B, ●: C, ●: D, ●: E) and type strains (●).

The other two subnetworks connected with anti co-occurrence relationships are the 28-gene (● in Figure 3.20) and the 19-gene (● in Figure 3.20) subnetworks. Six genes on pRL10 included in the 19-gene subnetwork have anti co-occurrence relationships to 18 genes on pRL8 and pRL10 in the 28-gene subnetwork. Most of the 18 genes in the 28-gene subnetwork are symbiosis genes on pRL10 and all five bvs (biovar *viciae* specific) genes on pRL8 (Figure 3.31).

3.6 Discussion

3.6.1 Observation on neighbouring genes and gene within the same replicon in the same sub co-occurrence network

The systematic study revealed that a majority of genes in the same co-occurrence subnetwork were neighbouring genes, for example in the 12-gene and 11-gene subnetworks (Figure 3.27). The plausible explanation was they may functionally related genes and tend to be conserved their presence in the genome (Tamames *et al.*, 1997, Tamames, 2001, Huyen *et al.*, 2000). As a result, they may be transferred within the population together through horizontal gene transfer (Achtman *et al.,* 2008).

After conducting community detection to reveal subtle relationships within the 480-gene subnetwork, the presence of chromosomal and non-chromosomal communities (Table 3.3) implied that these replicons took responsibilities for different functions in the population.

3.6.2 Detection on genes with specific characteristic

A subnetwork of bv. *viciae* specific genes including genes abundant in bv. *viciae* (Figure 3.25 and Table 3.4) was discovered. Five *bvs* genes on pRL8 (Kumar *et al.,* 2015) were found only in bv. *viciae*. Symbiosis genes on pRL10, *nod*, *nif*, and *fix* genes (Young *et al.,* 2006) directly involving in nodulation and nitrogen fixation, were identified. The analysis also identified *rhi* genes which are indirectly involved in nodulation (Rodelas *et al.,* 1999). In addition, pRL100162 and pRL100163, encoded asparagine synthase, were in this subnetwork Asparagine, which is an amino acid produced by nodules, may regulate nodulation (Oti-Boateng *et al.*, 1993, Lodwig *et al.*, 2003). These genes exhibited strong correlation within the population with significant statistics.

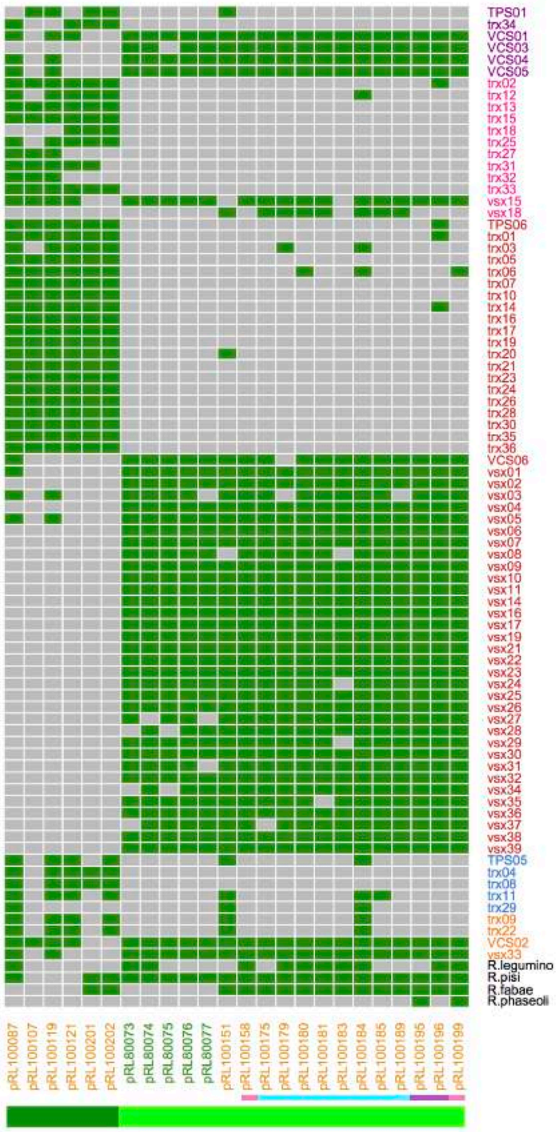


Figure 3.31 Phylogenetic profiles of genes in the 28-gene and the 19-gene subnetworks having anti co-occurrence relationships. The presence of genes is shown in 🟩, the absence of genes is shown in ⬜. Columns represent genes that have anti co-occurrence relationships and are arranged according to their subnetwork (🟩28-gene and 🟩 19-gene subnetwork). Genes in the 28 genes subnetwork are highlighted by their annotated function including *nod* (🟦), *nif* (🟪), and *fix* (🟥) genes on pRL10 and *bvs* genes pRL180073-pRL80077 on pRL8. Rows represent 85 strains and are arranged according to their respective genospecies (●: A, ●: B, ●: C, ●: D, ●: E) and type strains (●).

Community detection was applied to unravel complicated relationships in the subnetworks containing more than 100 genes. In the community of 170 genes in the 374-gene cluster, 79 out of them (Figure 3.23) are highly correlated with genospecies B. 26 genes (appendix table I.I) are observed in all genospecies B isolates and have been reported as a genospecies B specific-island (Kumar, 2013). Another community in this cluster, including 27 genes, also represents nine pRL12 genes that are absent in genospecies C (appendix table I.III). It is interesting to note that there are no chromosomal genes detected as candidate genospecies-specific or symbiovar-specific genes.

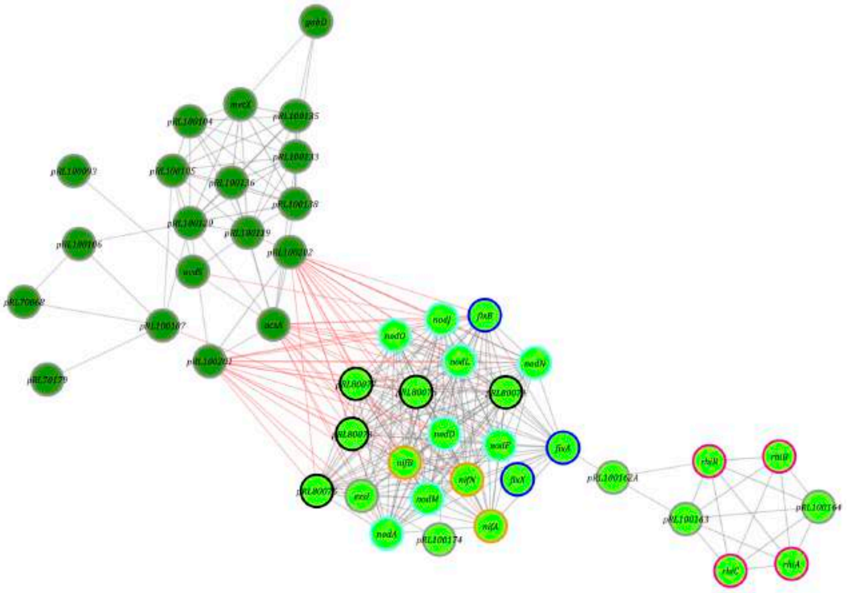


Figure 3.32 The 28-gene (●) and the 19-gene (●) subnetworks having anti co-occurrence relationships. Genes in the 28 genes subnetwork are highlighted by their annotated function including *nod* (○), *nif* (○), *fix* (○), and *rhi* (○) genes on pRL10 and Bvs genes (○) on pRL8. Grey (—) and red (—) edges specify co-occurrence and anti co-occurrence interactions.

3.6.3 Observations on anti co-occurrence relationships that represent mobile genetic elements, and the negative relationship between bvs genes, symbiosis genes and γ-hydroxybutyric acid genes

One possible explanation for the anti co-occurrence relationship between the 480-gene and the 272-gene subnetwork (● and ● in Figure 3.20, respectively) could be the activity of mobile genetic elements which are moveable and present in some strains. 20 chromosomal genes in the 480-gene and 43 chromosomal genes in the 272-gene subnetworks are identified as GIs (Table 3.5 and 3.6, respectively) by using IslandViewer 3 (Dhillon *et al.,* 2015).

Table 3.5 List of 20 chromosomal genes found in the 480-gene subnetwork and annotated as genomic islands by using IslandViewer 3

Locus tag	Gene symbol	Protein accession	Annotated function
RL0461	-	YP_766070.1	hypothetical protein
RL0462	-	YP_766071.1	transmembrane protein
RL0465	-	YP_766074.1	hypothetical protein
RL0466	-	YP_766075.1	hypothetical protein
RL1894	-	YP_767497.1	hypothetical protein
RL2261	-	YP_767855.1	hypothetical protein
RL2262	*cya3*	YP_767856.1	adenylate cyclase
RL2263	-	YP_767857.1	transmembrane protein
RL2264	-	YP_767858.1	arylsulfatase
RL2265	-	YP_767859.1	hypothetical protein
RL2266	-	YP_767860.1	hypothetical protein
RL2267	-	YP_767861.1	arylsulfatase
RL2272	-	YP_767866.1	hypothetical protein
RL3833	-	YP_769412.1	short-chain dehydrogenase/reductase
RL3834	-	YP_769413.1	ErfK/YbiS/YhnG oxidoreductase
RL3835	-	YP_769414.1	hypothetical protein
RL3836	-	YP_769415.1	transmembrane protein
RL3837	-	YP_769416.1	hypothetical protein
RL3838	-	YP_769417.1	transmembrane dehydrogenase/oxidoreductase
RL4664	-	YP_770227.1	transmembrane protein

Table 3.6 List of 43 chromosomal genes found in the 272-gene subnetwork and annotated as genomic islands by using IslandViewer 3

Locus tag	Gene symbol	Protein accession	Annotated function
RL0458	-	YP_766067.1	adenylate cyclase
RL0459	-	YP_766068.1	hypothetical protein
RL0460	-	YP_766069.1	hypothetical protein
RL0793	-	YP_766402.1	transposase-related protein
RL1135	-	YP_766745.1	transmembrane copper resistance protein
RL1136	-	YP_766746.1	hypothetical protein
RL1137	-	YP_766747.1	hypothetical protein

Locus tag	Gene symbol	Protein accession	Annotated function
RL1138	-	YP_766748.1	ECF sigma factor
RL1139	-	YP_766749.1	transmembrane protein
RL1846	-	YP_767450.1	two component response regulator transcriptional regulatory protein
RL1847	-	YP_767451.1	transcriptional regulator
RL1848	-	YP_767452.1	epoxide hydrolase
RL1849	-	YP_767453.1	hypothetical protein
RL1850	-	YP_767454.1	isochorismatase
RL1851	-	YP_767455.1	hypothetical protein
RL1852	-	YP_767456.1	enoyl-CoA hydratase
RL1853	-	YP_767457.1	endoribonuclease L-PSP family protein
RL1854	-	YP_767458.1	oxidoreductase
RL1858	-	YP_767461.1	protease
RL1859	-	YP_767462.1	hypothetical protein
RL1887	-	YP_767490.1	transmembrane protein
RL1888	-	YP_767491.1	hypothetical protein
RL1890	-	YP_767493.1	transmembrane protein
RL1891	-	YP_767494.1	transmembrane protein
RL1892	-	YP_767495.1	cation transporting P-type ATPase
RL1893	-	YP_767496.1	transmembrane protein
RL1895	-	YP_767498.1	hypothetical protein
RL2258	-	YP_767852.1	hypothetical protein
RL2279	-	YP_767873.1	hypothetical protein
RL2333	-	YP_767924.1	hypothetical protein
RL2334	-	YP_767925.1	AraC family transcriptional regulator
RL2843	-	YP_768428.1	transmembrane component of ABC transporter
RL2844	-	YP_768429.1	solute-binding component of ABC transporter
RL2845	*aroE*	YP_768430.1	shikimate dehydrogenase
RL2846	-	YP_768431.1	glyoxalase/dioxygenase
RL2847	-	YP_768432.1	shikimate dehydrogenase
RL2848	-	YP_768433.1	LysR family transcriptional regulator
RL2849	-	YP_768434.1	2-pyrone-4,6-dicarboxylic acid hydrolase
RL2850	-	YP_768435.1	hypothetical protein
RL3100	-	YP_768679.1	hypothetical protein
RL3101	-	YP_768680.1	transmembrane protein
RL3102	-	YP_768681.1	PadR family transcriptional regulator
RL3855	-	YP_769434.1	cytochrome c protein

Another possible explanation for an anti co-occurrence relationship between subnetworks might be their replaceable functions. For example, genes from the 480-gene and the 374-gene subnetworks having anti co-occurrence relationships, and some of them have similar annotated functions (Table 3.7).

Table 3.7 List of genes with anti co-occurrence relationship and their annotated function in the 374-gene and 480-gene subnetworks

Annotated function	Locus tag of gene in the 374-gene subnetwork	Locus tag of gene in the 480-gene subnetwork
ABC-type branched-chain amino acid transport system, periplasmic component	pRL90258	pRL120493
ABC-type branched-chain amino acid transport system, permease component	pRL90260	pRL120489
ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component	pRL120128	pRL120338
ABC-type dipeptide/oligopeptide/nickel transport system, permease component	pRL90228 pRL120129	pRL120337
ABC-type oligopeptide transport system, ATPase component	pRL90229	pRL120339
ABC-type transport system, periplasmic component	pRL90231 pRL120131	pRL120333
DNA-binding transcriptional regulator, AcrR family	pRL110082	pRL100457
DNA-binding transcriptional regulator, GntR family	pRL90193 pRL90257	pRL120527
DNA-binding transcriptional regulator, LysR family	pRL90119 pRL120275 pRL120294	pRL110026 pRL120457 pRL120347 pRL120544
NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	pRL120120 pRL120119 pRL120125 pRL110132 pRL120143 pRL120144 pRL120569 pRL120292	pRL120488
Pimeloyl-ACP methyl ester carboxylesterase	pRL120770 pRL120290	pRL120450

However, the 19-gene subnetwork (● in figure 3.20) contains genes relating to γ-hydroxybutyric acid utilisation such as *attKLM* (pRL100134, pRL100135, and pRL100136) (Chai *et al.*, 2007, Prell *et al.*, 2009, White *et al.*, 2009), and presents an anti co-occurrrence relationship to the 28-gene subnetwork (● in figure 3.20), which contains genes abundant in bv. *viciae*, such as symbiosis genes on pRL10 (including *nifN* (pRL100158), *nodO* (pRL100175), *nodNML* (pRL100179-pRL100181), *nodF* (pRL100183), *nodA* (pRL100185), *nodJ* (pRL100189), *nifB* (pRL100195), and *fixB* (pRL100199) (Figure 3.32). It is implied that most isolates that can utilise γ-hydroxybutyrate belong to symbiovar *trifolii*. Of 34 strains in the population utilising γ-hydroxybutyrate, 25 strains are bv. *trifolii* and 9 strains are bv. *viciae*. However, *attKLM* are absent from trx32, vsx18, vsx16, vsx26, vsx27, and vsx37. Consequently, there is no evidence for any functional relationship between *att* and symbiosis genes.

This chapter aimed to explore and investigate co-occurrence and anti co-occurrence relationship of genes in the rhizobium population by employing correlational computation and network analysis. The different correlational computation methods were first evaluated to find the most suitable for the data. In the step of network construction, correlation values were converted to network using the optimal threshold. In order to retain biological significance of the constructed network, many criteria were applied to find a compromised optimal threshold. The network then was generated with the optimal threshold. The gene co-occurrence network reflected global relationship of genes in the population, which occurrence of gene did not occur by chance. Genes with positive correlation, co-occurrence genes, were shown in the same subnetwork and included functionally related genes like the symbiosis and *rhi* genes on pRL10. Moreover, the community detection allowed us to gain insight into genes present under the same conditions despite complicated relationship in the subnetwork, for example, a community of genes specific to genospecies B in the 374-gene subnetwork. Genes with negative correlation, anti co-occurrence genes, were also shown in the constructed network. Anti co-occurrence relationship reflected mobile genetic elements as GIs on chromosome which were found in the 272 and 480-gene subnetworks. The relationship between anti co-occurring genes was possibly that they may be involved in adaptation to the same conditions but they may achieve this by alternative means. For example, the anti co-occurrence of the 374 and 480-gene subnetworks included genes with potentially replaceable functions. On the other hand, the anti co-occurrence relationship between the 19 and 28-

gene subnetwork cannot be ascribed to functional equivalence. Genes in the 19-gene subnetwork, annotated as genes involving γ-hydroxybutyric acid utilisation, were negatively correlated to symbiosis genes in the 28-gene subnetwork. This will be investigated in the next chapter.

Chapter 4 Analysis of phenotype-genotype data of the local population of *Rhizobium leguminosarum*

4.1 Abstract

The co-occurrence gene network in the previous chapter identified clusters of genes with either favoured or disfavoured occurrence. The co-occurrence gene network reflected the diversity of presence/absence of genes in the population. Ability of the population to utilise different carbon substrates was also examined in the laboratory and there is diversity in the ability of carbon substrate utilisation within the population. The substrate utilisation is usually a result of cooperation between a number of genes. Hence the gene presence/absence of the population might be reflected in the ability to utilise carbon substrates. This chapter aims to identify genes relevant to the utilisation of carbon substrates, by using computational methods to select important genes. The phylogenetic profiles of 72 Wentworth strains were used as genotypic data. A phenotypic data or metabolism profile of each rhizobium strain was obtained from the GN2 Biolog Microplate. The gene selection was computed by different computational methods. The most suitable method and its results were included in further analysis. The selected genes or candidate genes were investigated for their properties including function, distribution in the population and presence in the co-occurrence gene network.

4.2 Introduction

4.2.1 Diversity of phenotypic traits in the microbial population

Comparative genomic study in bacteria has demonstrated variation within a bacterial species not only in the gene content but also in metabolism, which is a result of the concerted action of multiple encoding genes. Study of growth-related phenotypes of *Pseudomonas aeruginosa* and its mutants demonstrated the different phenotypic patterns amongst them (Pommerenke *et al.*, 2010). Multiple *Lactococcus lactis* strains exhibited diversity in substrate utilisation (Bayjanov *et al.*, 2013). These studies suggest that the diversity in metabolic ability is probably a consequence of the distribution of genes. Studies of genotype-phenotype relationships involve voluminous data and computational analyses (Kell, 2004, Price *et al.*, 2004).

4.2.2 Computational methods for studying interdependence between phenotype and genotype data

The identification of genes closely relevant to the phenotype, called feature selection in machine learning (Guyon *et al.*, 2003, Saeys *et al.*, 2007, Blaby-Haas *et al.*, 2011), aims to quantify interdependence between phenotype and genotype data. Genotypic data used in a study of phenotype-genotype association can be single nucleotide polymorphisms (Gamazon *et al.*, 2012), orthologous groups (Goh *et al.*, 2006, Slonim *et al.*, 2006, Tamura *et al.*, 2008, Bayjanov *et al.*, 2013, Li *et al.*, 2014) or gene expression profiles from microarrays (Dudoit *et al.*, 2002, Schadt *et al.*, 2005). Phenotypic data in the study can be, for example, disease (Dudoit *et al.*, 2002, Schadt *et al.*, 2005, Gamazon *et al.*, 2012, Li *et al.*, 2014), metabolomics (Pommerenke *et al.*, 2010, Bayjanov *et al.*, 2013), or lifestyle traits (Goh *et al.*, 2006, Slonim *et al.*, 2006, Pommerenke *et al.*, 2010).

4.2.2.1 Pairwise association

The relationships between genotype and phenotype were measured directly by pairwise association metrics such as mutual information (Slonim *et al.*, 2006, Wu *et al.*, 2009) and Pearson's correlation (Goh *et al.*, 2006, Li *et al.*, 2014). An advantage of mutual information is that the method is not influenced by the relationships found very rarely or abundantly in the population, but the method cannot identify the direction of the relationship (gene present or absent when the substrate was utilised). Pearson's correlation has the advantage that it specifies the direction of the relationship. Weaknesses of Pearson's correlation are that the measure cannot be used for capturing non-linear relationship in the data and is unduly influenced by relationships found rarely or abundantly in the population, resulting in identification of spurious associations.

4.2.2.2 False discovery rate

Genotype-phenotype data that is represented by a small number of observations (n) and a large number of features (p) can result in acquiring "significant" correlation between gene and phenotype that occurs by chance, sometimes called the "small n, large p" problem. One approach to control the significant correlation between gene and phenotype that occurs by chance, or type I error, is to control false discovery rate (FDR)(Reiner *et al.*, 2003, Storey *et al.*, 2003). False discovery rate (FDR)(*q-value*)(Schweder *et al.*, 1982, Benjamini *et al.*, 1995) is the expected proportion of rejected null hypotheses that were incorrectly rejected.

4.2.2.3 Random forest

Ensemble methods (Dietterich, 2000) such as random forest were introduced to evaluate correlation on phenotype-genotype data (Kursa *et al.*, 2010b, Bayjanov *et al.*, 2012) due to the voluminous data. The random forest method, first introduced by Breiman (2001), is conducted based on **B**ootstrap **agg**regat**ing** (**Bagging**) (Breiman, 1996). Each tree in a forest is grown using a bootstrap sample of learning data without pruning and predictions are made by the majority vote, with all trees (i.e. the forest) having same weight. The random forest slightly differs from Bagging in that its algorithm is keen on partitioning variables present in a random sample instead of all the variables (Breiman, 2001). Random forest construction can be divided into two main steps, namely bootstrap sampling and aggregation. Assume data on n strains and p genes. In the step of bootstrap sampling, learning data ($\mathcal{L}$) containing m strains are generated from n-strain data where $m < n$. $\mathcal{L}_i$ is drawn randomly but with replacement to construct a tree classifier C_i(where $i = 1,2, \dots, k$), as a result $\mathcal{L}_i$ may or may not contained replicated samples. To obtain classification trees (C_i where $i = 1,2, \dots, k$) in the forest, growth of each classification tree is preceded by selecting variables that are present in each learning dataset. Some of the learning data is excluded from the tree construction, which are denoted as **o**ut-**o**f-**b**ag (OOB) data similar to a cross-validation method, are use to test performance of the classification tree. Aggregation then takes place to find prediction results by majority vote among the resulting decision trees, and these predictions are used to estimate an error rate given from the forest.

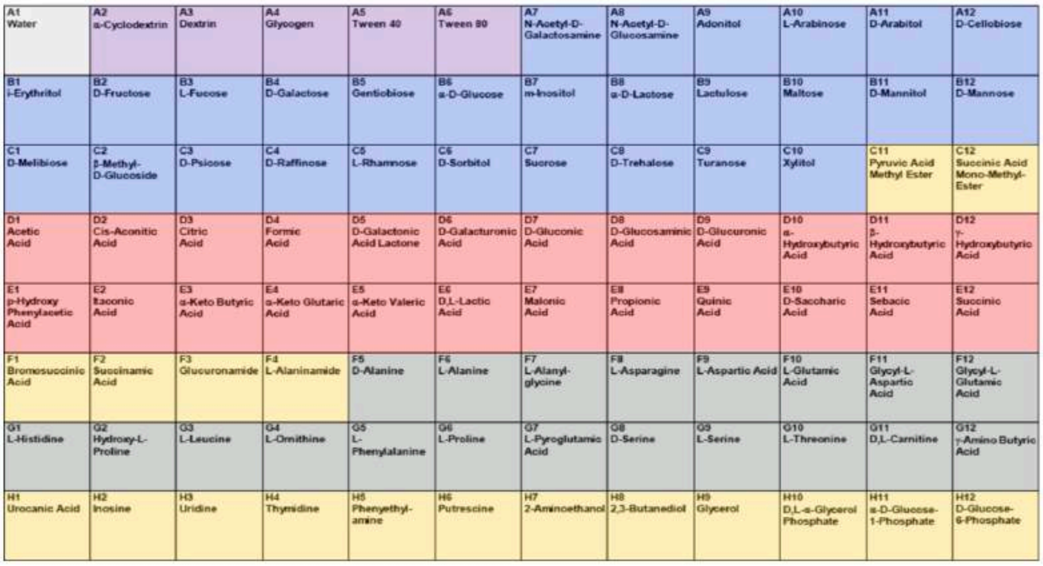
Procedures of feature selection in random forest are based on computation of feature importance. The feature or gene importance computation is basically to replace the contribution of the selected gene with random noise. If the random noise decreases the *accuracy* of the prediction of the ability to use the substrate, the selected gene is determined to be relevant to the substrate utilisation. On the contrary, if the random noise exhibits less effect on the *accuracy* of the prediction, the selected gene is determined to be irrelevant to the substrate utilisation.

4.2.2.4 Class association rules

Pairwise association computation and random forests were used for finding one-to-one relationships. One-to-one relationships were found to be less powerful than many-to-one relationships. To analyse multiple genes involved in the same phenotype (a many-to-one relationship), class association rules were introduced (Carmona-Saez *et al.*, 2006, Tamura *et al.*, 2008). Association rules (AR henceforth) (Agrawal *et al.*, 1993, Agrawal *et al.*, 1994) have been used as a tool for identifying relationships between items in a large database. The idea of association rules is to predict the occurrence of an item based on the occurrences of other items in the database. Alternately, the association rules are generated following if–then syntax i.e. (*Set of items*)$_1$ ⇒ (*Set of items*)$_2$, where (*Set of items*)$_1$ and (*Set of items*)$_2$ are disjoint. A class association rule (CAR henceforth), a subset of association rules, defined (*Set of items*)$_1$ as gene presence and (*Set of items*)$_2$ as ability of substrate utilisation (utilising/not utilising substrate), which is called a class (i.e. class association rule).

The association rules algorithm can be broken down into two main steps; 1) Rules and their frequencies are generated and 2) All generated rules in the previous step are evaluated for their strength. NETCAR (Tamura *et al.*, 2008), a CAR mining algorithm, is conducted to find sets of genes relevant to the substrate utilisation. First, rules and their frequency are generated, since genes and substrate utilisation profiles are different from market basket data, which AR was initially applied to analyse. The genes and substrate utilisation profiles contain a larger number of genes (variables) than strains (observations), resulting in a large space of rules and irrelevant rule generation (Liu, 2007). Mutual information is proposed to narrow down the size of generated rules by constructing co-occurrence gene networks. The co-occurrence network contains nodes as genes and edges computed from mutual information. If the mutual information between their phylogenetic profiles is greater than a threshold, rules generated from genes are kept. Not only does this reduce computational time but rules generated from genes having close phylogenetic profiles may provide more information on biological function.

In a previous study by Kumar *et al.* (2015), the utilisation of carbon substrates in the 72 Wentworth strains was studied by using the Biolog GN2 Microplate. Each Biolog plate has 96 wells with 95 different carbon substrates and a blank well with water (Figure 4.1). These 95 substrates can be classified into 5 groups based on substrate classes that are 1) neutral, 2) sugar / sugar derivative, 3) carboxylic / dicarboxylic acid, 4) amino acid / amino acid derivative, and 5) miscellaneous intermediates of metabolism. This system employs a universal reporter of metabolism involving a redox dye and shows results based on utilisation of the substrate (as not utilised/poorly utilised/strongly utilised) in each well. The diversity in ability to utilise substrate in the 72 Wentworth strains is shown in Figure 4.2.



A1 Water	A2 α-Cyclodextrin	A3 Dextrin	A4 Glycogen	A5 Tween 40	A6 Tween 80	A7 N-Acetyl-D-Galactosamine	A8 N-Acetyl-D-Glucosamine	A9 Adonitol	A10 L-Arabinose	A11 D-Arabitol	A12 D-Cellobiose
B1 i-Erythritol	B2 D-Fructose	B3 L-Fucose	B4 D-Galactose	B5 Gentiobiose	B6 α-D-Glucose	B7 m-Inositol	B8 α-D-Lactose	B9 Lactulose	B10 Maltose	B11 D-Mannitol	B12 D-Mannose
C1 D-Melibiose	C2 β-Methyl-D-Glucoside	C3 D-Psicose	C4 D-Raffinose	C5 L-Rhamnose	C6 D-Sorbitol	C7 Sucrose	C8 D-Trehalose	C9 Turanose	C10 Xylitol	C11 Pyruvic Acid Methyl Ester	C12 Succinic Acid Mono-Methyl-Ester
D1 Acetic Acid	D2 Cis-Aconitic Acid	D3 Citric Acid	D4 Formic Acid	D5 D-Galactonic Acid Lactone	D6 D-Galacturonic Acid	D7 D-Gluconic Acid	D8 D-Glucosaminic Acid	D9 D-Glucuronic Acid	D10 α-Hydroxybutyric Acid	D11 β-Hydroxybutyric Acid	D12 γ-Hydroxybutyric Acid
E1 p-Hydroxy Phenylacetic Acid	E2 Itaconic Acid	E3 α-Keto Butyric Acid	E4 α-Keto Glutaric Acid	E5 α-Keto Valeric Acid	E6 D,L-Lactic Acid	E7 Malonic Acid	E8 Propionic Acid	E9 Quinic Acid	E10 D-Saccharic Acid	E11 Sebacic Acid	E12 Succinic Acid
F1 Bromosuccinic Acid	F2 Succinamic Acid	F3 Glucuronamide	F4 L-Alaninamide	F5 D-Alanine	F6 L-Alanine	F7 L-Alanyl-glycine	F8 L-Asparagine	F9 L-Aspartic Acid	F10 L-Glutamic Acid	F11 Glycyl-L-Aspartic Acid	F12 Glycyl-L-Glutamic Acid
G1 L-Histidine	G2 Hydroxy-L-Proline	G3 L-Leucine	G4 L-Ornithine	G5 L-Phenylalanine	G6 L-Proline	G7 L-Pyroglutamic Acid	G8 D-Serine	G9 L-Serine	G10 L-Threonine	G11 D,L-Carnitine	G12 γ-Amino Butyric Acid
H1 Urocanic Acid	H2 Inosine	H3 Uridine	H4 Thymidine	H5 Phenyethyl-amine	H6 Putrescine	H7 2-Aminoethanol	H8 2,3-Butanediol	H9 Glycerol	H10 D,L-α-Glycerol Phosphate	H11 α-D-Glucose-1-Phosphate	H12 D-Glucose-6-Phosphate

Figure 4.1 The Biolog GN2 substrate panel. The substrates are coloured coded as followed: 🟪 polymers, 🟦 sugars/sugar derivatives, 🟥 carboxylic/dicarboxylic acid, ◼ amino acid/amino acid derivative and 🟨 miscellaneous intermediates of metabolism (Bochner, 1989).

This chapter aims to find genes relevant to ability of utilisation of specific substrates in the rhizobium population. Initially, candidate genes relevant to the substrate utilisation were selected by using different measures including 1) Pairwise associations and FDR, 2) random forest and 3) class association rules. Later, candidate genes were investigated for their function, related pathways and occurrence pattern in the genome.

4.3 Chapter aims

Analysing and exploring associations between genotype and phenotype of the rhizobium population from phylogenetic gene profiles of the local rhizobial population and their phenotypic data.

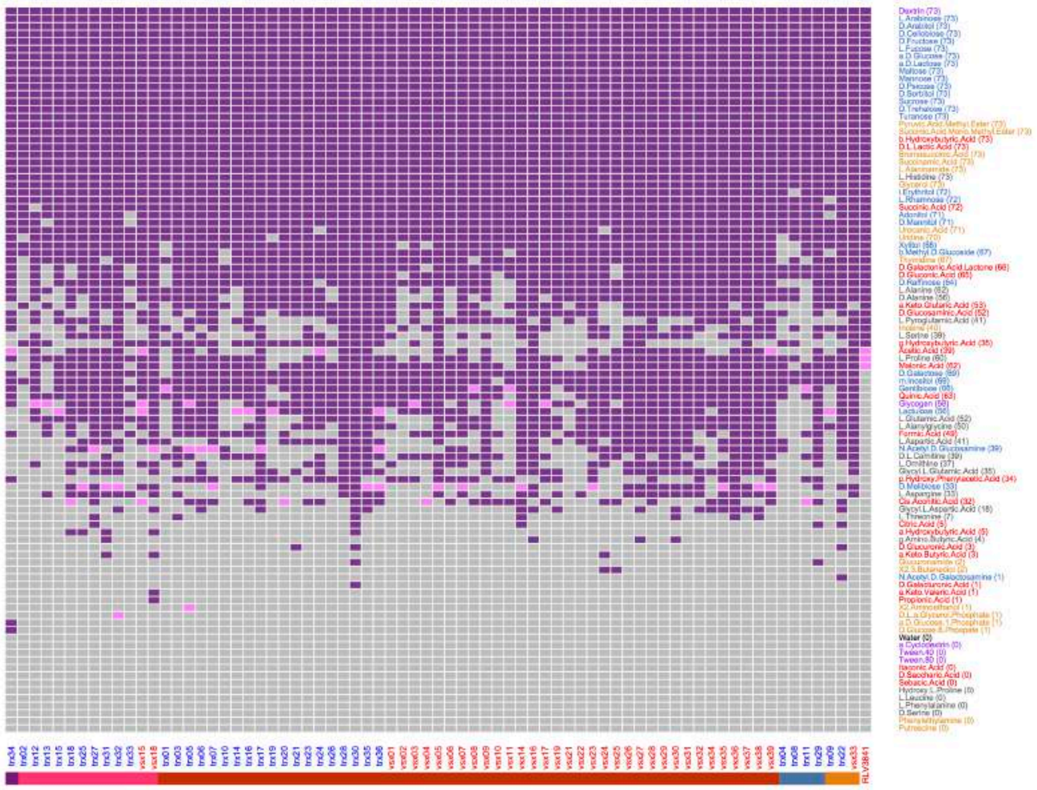


Figure 4.2 Profile of substrate utilisation of 72 Wentworth strains and the reference strain, *Rlv*3841. Rows represent substrates with number of strains utilising the substrate in the brackets and colour-labelled by carbon substrate (🟪 polymers, 🟦 sugars, 🟥 acids, ■ amino acids and 🟧 miscellaneous intermediates). Columns represent strains in this study arranged according to five genospecies (■: A, ■: B, 🟥: C, 🟦: D, and 🟧: E) and colour-labelled by their symbiovar (🔵 bv. *trifolii* and 🔴 bv. *viciae*). Ability of substrate utilization is defined by ■ strain able to utilise the substrate, ■ strain able to partially utilise the substrate and ■ strain unable to utilise the substrate.

4.4 Materials and methods

4.4.1 Genotype and phenotype data

The genotype data of 72 Wentworth strains were used in the study. The phenotypic test data of 72 Wentworth strains were obtained using Biolog GN2 MicroPlate Gram-negative identification test panel system (Kumar *et al.*, 2015). The level of substrate utilisation was designated at 3 levels that were '1' designating the ability of the strain to utilise the substrate, '0.5' designating the ability of the strain to partially utilise the substrate and '0' designating the inability of the strain to utilise the substrate.

4.4.2 Pre-processing data

Substrates utilised both by *Rlv*3841 and by some strains in the population were focused on, because presence or absence of genes in the genotypic data was inferred from the *Rlv*3841 genome. Substrates that were not utilised by any strains, utilised by all 72 Wentworth strains, not utilised by *Rlv*3841 or partially used by *Rlv*3841 were excluded from further analysis. Hence, of the 95 carbon substrates, only 21 substrates that were utilised by *Rlv*3841 and some, but not all, of the Wentworth isolates were included for further analysis (Figure 4.2). The full list of the 95 substrates in the Biolog plate can be seen in Appendix Table III.I.

4.4.3 Computational methods on substrate utilisation and gene occurrence

4.4.3.1 Pearson correlation coefficient

Pearson correlation coefficient ($r_{\mathrm{gene}_A \mathrm{substrate}_B}$) is defined as below (Goh *et al.*, 2006).

$$r_{\mathrm{gene}_A,\mathrm{substrate}_B} = \frac{\sum_{j=1}^{N}\left(X_{gene_A,j} - \overline{X_{gene_A}}\right)\left(Y_{j,\mathrm{substrate}_B} - \overline{Y_{\mathrm{substrate}_B}}\right)}{\sqrt{\sum_{j=1}^{N}\left(X_{gene_A,j} - \overline{X_{gene_A}}\right)^2}\sqrt{\sum_{j=1}^{N}\left(Y_{j,\mathrm{substrate}_B} - \overline{Y_{\mathrm{substrate}_B}}\right)^2}}$$

$X_{gene_A,j}$ is the presence or absence of *gene_A* within *strain_j*, and $Y_{j,\mathrm{substrate}_B}$ represents the utilisation of *substrate$_B$* of *strain$_j$*. $\overline{X_{gene_A}}$ is defined as the mean of the distribution of *gene$_A$* for the population. $\overline{Y_{\mathrm{substrate}_B}}$ is the mean of utilisation of *substrate$_B$* for the population. The value of $r_{\mathrm{gene}_A,\mathrm{substrate}_B}$ ranged between -1 to 1. A value of 0 means that the presence of *gene$_A$* is completely independent of the utilisation of *substrate$_B$*. A value of 1 means that all the strains carrying *gene$_A$*, and none of the others, are able to utilise *substrate$_B$*. A correlation of −1 means that every strain that carries *gene$_A$* is unable to utilise *substrate$_B$* and vice-versa. The computation of Pearson's correlation was implemented with WGCNA (Langfelder *et al.*, 2008).

4.4.3.2 Mutual information

Mutual information ($M_{\text{gene}_A,\text{substrate}_B}$) (Slonim *et al.*, 2006, Wu *et al.*, 2009) The empirical mutual information can be estimated as follows.

$$
\begin{aligned}
& M_{\text{gene}_A,\text{substrate}_B} \\
& \quad = P_{\text{gene}_A{}^+,\text{substrate}_B{}^+} log \frac{P_{\text{gene}_A{}^+,\text{substrate}_B{}^+}}{P_{\text{gene}_A{}^+} \times P_{\text{substrate}_B{}^+}} \\
& \quad + P_{\text{gene}_A{}^+,\text{substrate}_B{}^-} log \frac{P_{\text{gene}_A{}^+,\text{substrate}_B{}^-}}{P_{\text{gene}_A{}^+} \times P_{\text{substrate}_B{}^-}} \\
& \quad + P_{\text{gene}_A{}^-,\text{substrate}_B{}^+} log \frac{P_{\text{gene}_A{}^-,\text{substrate}_B{}^+}}{P_{\text{gene}_A{}^-} \times P_{\text{substrate}_B{}^+}} \\
& \quad + P_{\text{gene}_A{}^-,\text{substrate}_B{}^-} log \frac{P_{\text{gene}_A{}^-,\text{substrate}_B{}^-}}{P_{\text{gene}_A{}^-} \times P_{\text{substrate}_B{}^-}}
\end{aligned}
$$

P_{gene_A} represents the probability of the presence $(P_{\text{gene}_A{}^+})$ or absence $(P_{\text{gene}_A{}^-})$ of *gene_A*. $P_{\text{substrate}_B}$ is the probability of the utilisation of *substrate$_B$*, which can be either the substrate was utilised $\left(P_{\text{substrate}_B{}^+}\right)$ or not $(P_{\text{substrate}_B{}^-})$. $P_{\text{gene}_A,\text{substrate}_B}$ represents the probability of the distribution of *gene$_A$* and the utilisation of *substrate$_B$*, which can be the probability of the presence of *gene$_A$* in strains utilising *substrate$_B$* $\left(P_{\text{gene}_A{}^+,\text{substrate}_B{}^+}\right)$, the presence of *gene$_A$* in strains not utilising *substrate$_B$* $(P_{\text{gene}_A{}^+,\text{substrate}_B{}^-})$, the absence of *gene$_A$* in strains utilising *substrate$_B$* $\left(P_{\text{gene}_A{}^-,\text{substrate}_B{}^+}\right)$, and the absence of *gene$_A$* in strains not utilising *substrate$_B$* $(P_{\text{gene}_A{}^-,\text{substrate}_B{}^-})$. $M_{\text{gene}_A,\text{substrate}_B}$ is 0 if and only if the measurements on the distribution of *gene$_A$* and the ability to utilise *substrate$_B$* are independent. The mutual information of *gene$_A$* and *substrate$_B$* is greater if the distribution of *gene$_A$* and the ability to utilise *substrate$_B$* are relevant, it could be said that disappearance of *gene$_A$* was found in stains able to utilise *substrate$_B$* or appearance of *gene$_A$* was found in stains able to utilise *substrate$_B$*. The mutual information computation was implemented with infotheo (Meyer, 2009), an R package.

4.4.3.3 False Discovery Rate (FDR) control

To control the number of genes with false discoveries, the approach of Benjamini *et al.* (1995) was used. Given *m* tested null hypotheses, for each hypothesis H_i (i=1,...,m), *p-value* is calculated along with the corresponding p_i (i=1,...,m). *R* denotes the number of null hypotheses rejected by a procedure. *V* represents the number of true null hypotheses with incorrect rejection. The *FDR* (*q-value*) is defined as *E(V/R)*.

First, the *p*-values are ordered so that $p_{(1)} \leq ... \leq p_{(m)}$. Second, each value $p_{(i)}$ is compared with $q\frac{i}{m}$, where *q* is the desired FDR level. Finally, with *k* = max(*i* : $p_{(i)} \leq q\frac{i}{m}$) all hypotheses belonging to $p_{(1)}$,...,$p_{(k)}$ are rejected. By using Benjamini-Hochberg (BH) rule, the following is simple correction of *p*-values:

$$p_i^{BH} = p_i \frac{m}{order(p_i)}, i = 1, \dots, m.$$

where order (p_i) equals one for the smallest and *m* for the largest *p*-value, respectively. The fdrtool package in R (Strimmer, 2008) was used to compute *FDR* in this study.

4.4.3.4 Random forest

Selection of genes related to the substrate utilisation was performed using the Boruta (Kursa *et al.*, 2010b) and randomForest (Liaw *et al.*, 2002) packages for R. Boruta is a wrapper algorithm. The wrapper approach quantifies the subset of variables, which provides the maximum prediction *accuracy*, by using the training data and the classifier as part of the evaluation (Kohavi *et al.*, 1997). The variable importance of the Boruta algorithm is enumerated from the *Z*-scores of the original RF variable importance score against the randomly shuffled original observations for each variable, which latter are called shadow variables, to determine which variables are truly important (Kursa *et al.*, 2010b). The following are the steps in the Boruta algorithm.

- Extend the information system by adding at least 5 shadow attributes
- Shuffle the added shadow attributes to remove their correlations with the response.
- Run a random forest classifier on the extended information system and gather the Z scores computed.

- Quantify the maximum Z score among shadow attributes (MZSA)
- Mark variables with Z score lower than MZSA as 'irrelevant' and remove them from the analysis.
- Mark variables with Z score higher MZSA as 'relevant'.
- Remove all shadow attributes.
- Repeat the procedure until no further variables are marked irrelevant or until the maximum number of user-defined iterations has been reached.

Since random forests chose genes based on random selection, iteration on gene selection was operated three times in order to stabilise the results of gene selection (Van Landeghem *et al.*, 2010, Bayjanov *et al.*, 2013). Genes with three times "Confirmed" result were considered as genes relevant in the substrate utilisation.

4.4.3.5 Class association rules

The strength of rules in the NETCAR algorithm is evaluated by using *Confidence* and *Support* (Agrawal *et al.*, 1994). *Confidence* is the conditional probability of the ability of substrate utilisation (utilising/not utilising substrate) given by the set of genes. *Support* is the fraction of strains in which the rule is valid in the data. For example, $gene_A$ *and* $gene_B \Rightarrow$ *utilising* $substrate_C$ with 100% *Confidence*, it means that in all strains carrying $gene_A$ *and* $gene_B$ it is observed that $substrate_C$ is utilised. Strength of the converse relation is evaluated by mutual information to ensure the generated rule has biological relationships following the syntax, *set of present genes* ⇔ *ability of substrate utilisation*. The basic algorithm is described as follows:

- Select *Parent* genes whose profile shows a strong relationship with the substrate utilisation phenotype.
- Construct a connectivity graph of gene presence/absence (a co-occurrence network). An edge is present if the mutual information between two genes is greater than the threshold.
- Select *Child* genes that are within $s-1$ (s is denoted as width of rule or number of genes present in a rule) steps from a *Parent* on the gene presence/absence graph. For $s \geq 5$, unconnected subgraphs

start being generated, resulting in production of redundant rules and long computational times. The maximum width of rule is therefore 4 genes.

- Generate rules or candidate sets of genes containing at least one *Parent gene* and Child genes from a connected subgraph on the gene presence/absence connectivity graph. At this step, $gene_A$ *and* $gene_B$ ⇒ *utilising substrate*$_C$ is generated.
- Evaluate mutual information of the converse relationship of *Set of present genes ⇔ ability of substrate utilisation*. If the converse relationship exhibits mutual information that is larger than the defined threshold, the rule is kept.

The class association rule uses *accuracy, F-score, precision*, and *recall* to select significant rules.

$$Accuracy = \frac{\begin{array}{c}Number\ of\ strains\ carrying\ genes\ and\ not\ utilising\ substrate\\ +\ Number\ of\ strains\ not\ carrying\ genes\ and\ utilising\ substrate\end{array}}{Number\ of\ all\ strains}$$

$$F\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

$$Precision = \frac{Number\ of\ strains\ carrying\ genes\ and\ not\ utilising\ substrate}{\begin{array}{c}Number\ of\ strains\ carrying\ genes\ and\ not\ utilising\ substrate\ +\\ Number\ of\ strains\ carrying\ genes\ and\ utilising\ substrate\end{array}}$$

$$Recall = \frac{Number\ of\ strains\ carrying\ genes\ and\ not\ utilising\ substrate}{\begin{array}{c}Number\ of\ strains\ carrying\ genes\ and\ not\ utilising\ substrate\ +\\ Number\ of\ strains\ not\ carrying\ genes\ and\ not\ utilising\ substrate\end{array}}$$

In this study, the width of rules was varied between 2 and 4 genes. Genes were considered to relate to the substrate utilisation when the genes were present in significant rules.

4.4.4 Computational Resources

The correlation and random forest were run on an Apple MacBook Pro Intel Core i5 2.3 GHz CPU with 16GB 1333 MHz RAM running Mac OS X 10.10.4. R version 3.1.3 were used in this chapter. NETCAR, a java application, was computed the University of York Biology Linux grid, comprising of 27 quad core machines with 16GB RAM, controlled by Sun Microsystems Grid engine 6.1u2. java version 1.7.0_79 was used in this chapter.

4.5 Results

4.5.1 Results on computational methods

4.5.1.1 Results on pairwise association methods

The threshold of correlation value and mutual information was chosen as the area above the 80th percentile of the entire correlation and mutual information values (Li *et al.*, 2014). By using the criterion of genes with correlation value lying above about 81.2 % of the entire correlation values of the analysed data (Figure 4.3), genes with absolute correlation value equal to or greater than 0.14 were considered to be potential genes. Of 125,592 correlation values in total, 23,712 correlation values were considered to involve potential genes by this criterion.

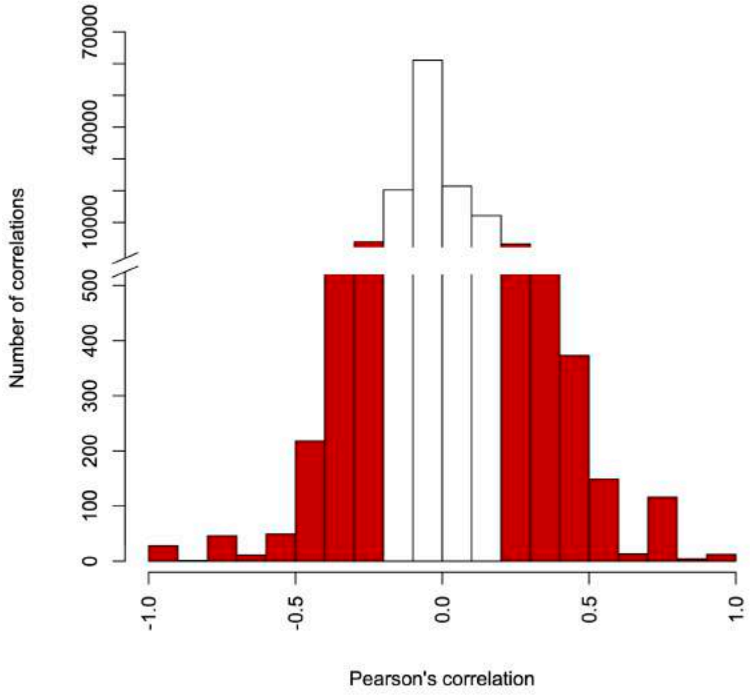


Figure 4.3 The distribution of correlation between substrate utilisation and genes. Genes with correlation value greater than 0.14 or less than -0.14 in 🟥 were considered to be potential genes involved in substrate utilisation.

When the number of strains utilising a substrate was equal to the number of strains with genes and no genes were found in strains not utilising substrate, a

Pearson's correlation of 1 was obtained (● in circle Figure 4.4). In this study, genes with a Pearson's correlation of 1 were from substrates that were used by 71 of 72 strains in the population, viz. i-erythritol, L-rhamnose, and succinic acid. The genes with a Pearson's correlation of 1 were present in the strains utilising the substrates and absent from the single strain that did not use the corresponding substrate. Pearson's correlation of -1 (● in circle Figure 4.4) was seen in substrates that were utilised by 71 of 72 strains in the population, viz. adonitol and L-rhamnose. Genes representing a Pearson's correlation of -1 were absent from the strains utilising adonitol and L-rhamnose.

Though Pearson's correlation of -1 and 1 identified genes completely dissociated or associated with the substrate utilisation ability, associations that depend on a single isolate can readily arise by chance, which is a drawback of Pearson's correlation.

There were some genes in the analysed data in which the gene was present in all (or nearly all) strains that utilised the substrate, but their correlation values were not equal to 1 or -1 (● and ● outside the circles Figure 4.4) because the correlation value was reduced by the presence of strains that had the gene but did not utilise the substrate.

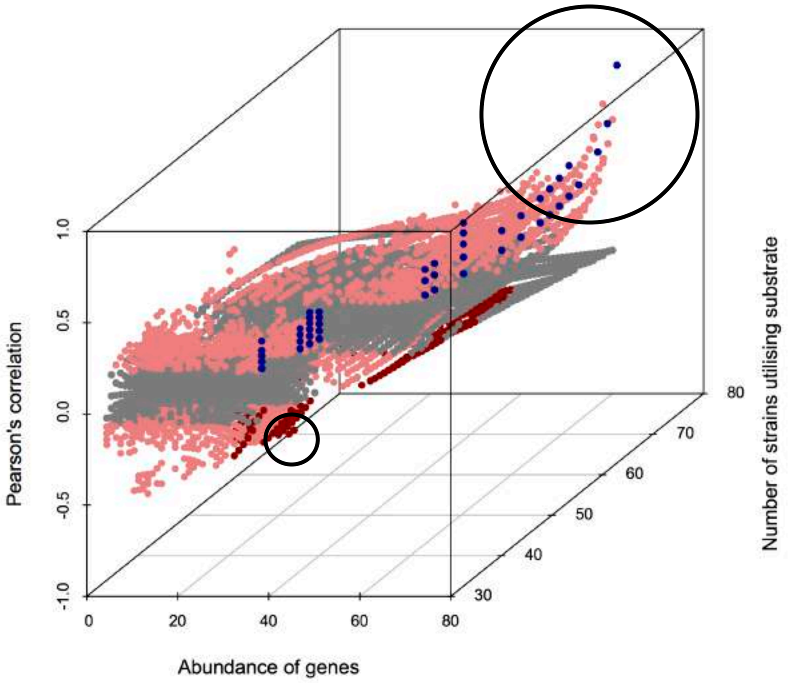


Figure 4.4 Plot of number of strains utilising substrate, abundance of genes in strains utilising the substrate and their Pearson's correlation values. Nodes are coloured by abundance of genes, the number of strains utilising substrate and their Pearson's correlation values (● gene with correlation value equal or less than -0.14 and having equal numbers of strains utilising substrate and an abundance of genes in strains not utilising the substrate, ● gene with correlation value equal or greater than 0.14 and having equal number of strains utilising substrate and an abundance of genes in strains that utilise the substrate, ● gene with absolute of correlation value equal to or greater than 0.14 and having unequal numbers of genes and strains utilising substrate, and ● gene with absolute of correlation value less than 0.14).

Mutual information values were calculated between the presence of genes and the ability to utilise substrate (Figure 4.5), and genes with values of mutual information equal or greater than 0.011, lying above more than 80.3% of mutual information values (24,804 mutual information values of 125,592 mutual information values in total), were considered to be potential genes associated with the substrate utilisation by this criterion.

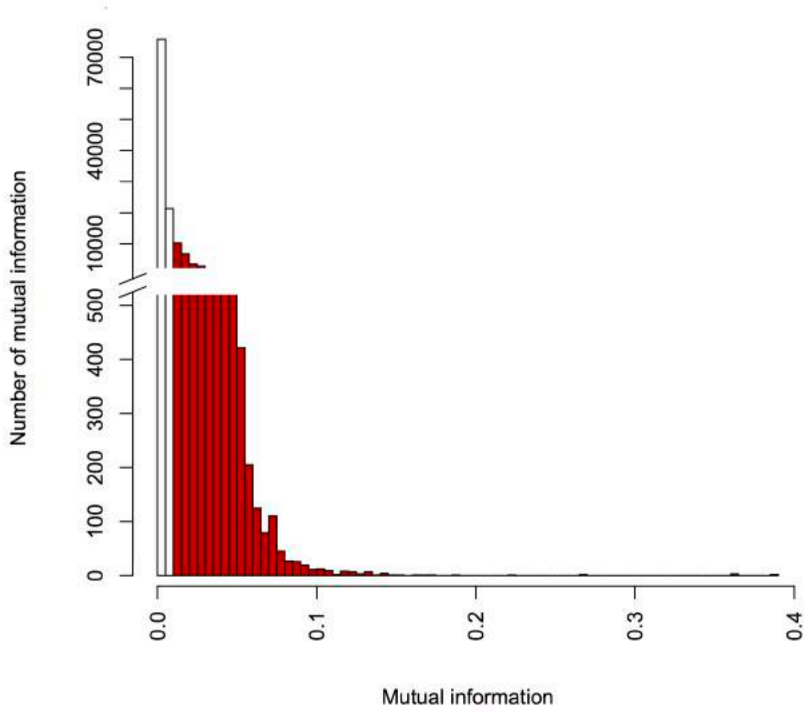


Figure 4.5 The distribution of mutual information between substrate utilisation and genes. Genes with mutual information value equal or greater than 0.011 in 🟥 were considered to be potential genes for utilising substrate.

The mutual information was robust with regards to genotype-phenotype relationships that were from substrates used by most strains in the population and abundance of genes carried by the strains (Figure 4.6). The mutual information values between substrates utilised by 71 strains and genes present in the same 71 strains (giving Pearson's correlation of 1) were 🔵 encircled in Figure 4.6. Likewise, the mutual information value between substrate utilised by 71 strains without the gene and one strain not utilising substrate but carrying the gene (giving Pearson's correlation of -1) were 🔴 encircled in Figure 4.6. Mutual information values of these genes are not as high as their Pearson's correlation values. Consequently, cases of 1 and -1 of Pearson's correlation were handled better by mutual information.

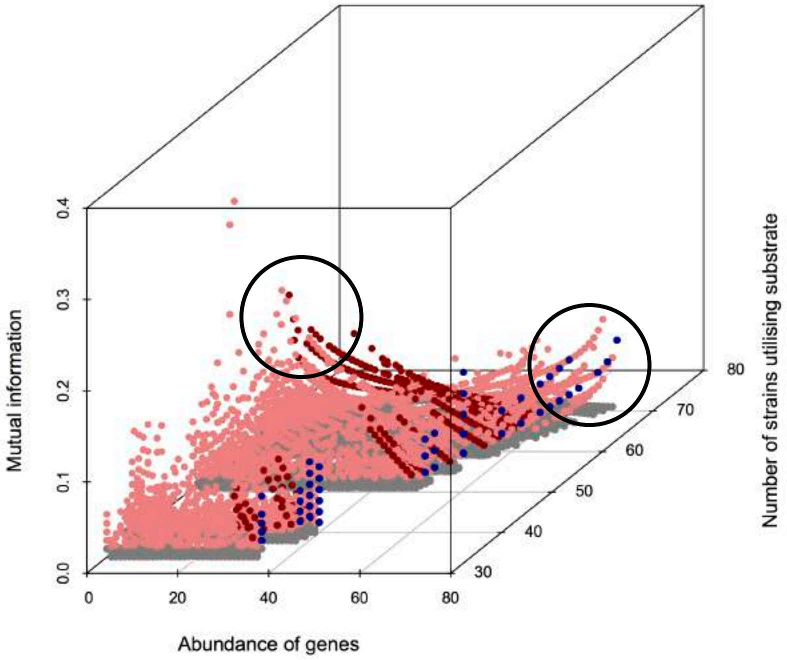


Figure 4.6 Plot of frequency of number of strains utilising substrate, abundance of genes in strains utilising the substrate and their mutual information values. Nodes are coloured by abundance of genes, the number of strains utilising substrate and their mutual information values (● gene with mutual information equal to or greater than 0.011 and having equal number of strains utilising substrate and an abundance of genes in strains not utilising the substrate, ● gene with mutual information equal to or greater than 0.011 and having equal number of strains utilising the substrate and an abundance of genes in strains that utilise the substrate, ● gene with mutual information value equal to or greater than 0.011 and having unequal numbers of genes and strains utilising substrate and ● gene with mutual information value less than 0.011).

It was found that Pearson's correlation and mutual information provided results in the same way (Figure 4.7 (a)). Using the criterion of selecting Pearson's correlation and mutual information above the 80th percentile, there were 27,699 significant associations, including associations between the substrates utilised by most of strains in the population and genes with high abundance, which might be random associations.

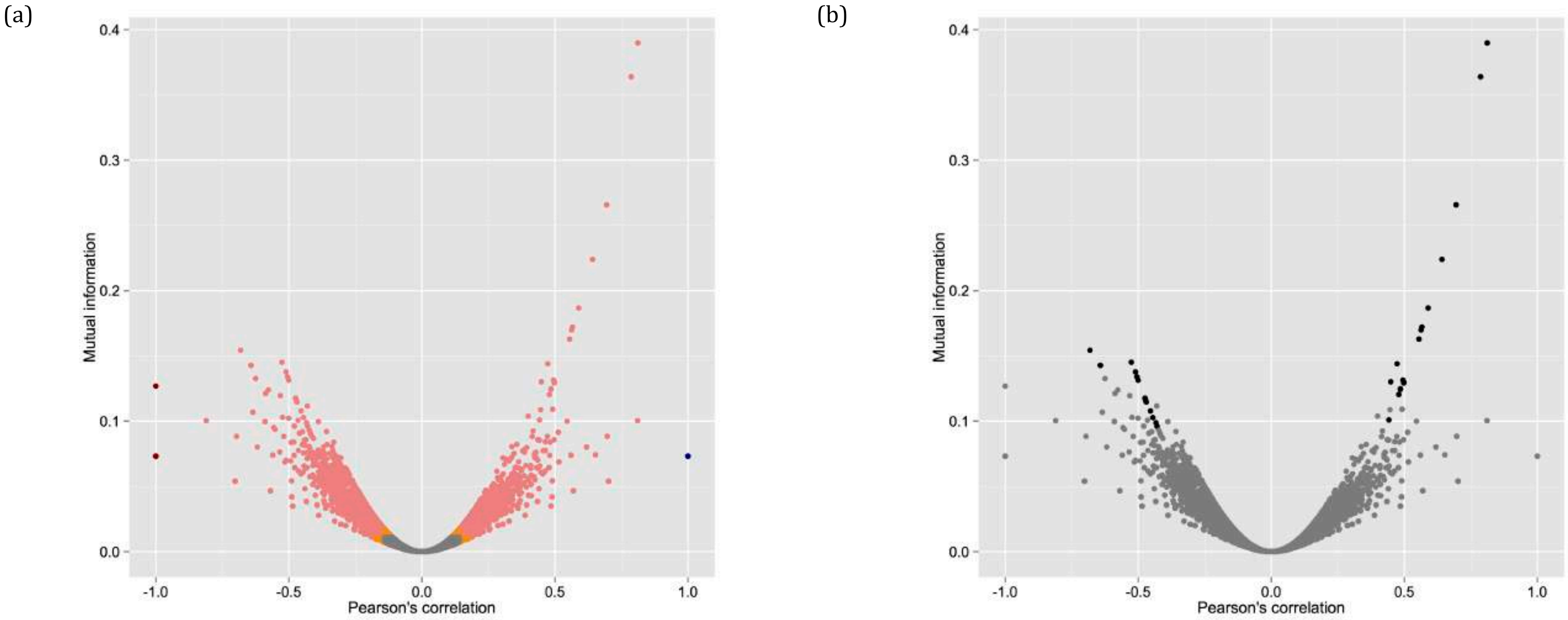


Figure 4.7 Plots of Pearson's correlation and mutual information value. (a) Nodes are coloured by association value from different measures and significance (● gene with significant mutual information and correlation value and having equal number of strains utilising substrate and abundance of genes of strains that did not utilise the substrate, ● gene with significant mutual information and correlation value and having equal number of strains utilising the substrate and abundance of genes of strains that utilise the substrate, ● gene with significant mutual information and significant correlation, ● gene with significant mutual information or significant correlation value, and ● gene with insignificant mutual information and correlation values). (b) Nodes are coloured by association value from different measure and False Discovery Rate q-value (● gene with significant mutual information, correlation, and *q-value* ≤ 0.05, ● gene with insignificant mutual information or correlation and *q-value* > 0.05)

In order to alleviate effects of associations between the substrates utilised by most strains in the population and genes with high abundance considered as potential genes, FDR was applied to the data (Figure 4.7 (b)). After FDR was applied to the analysed data, of 27,699 associations, 44 associations between genes and substrates exhibited significant relationships. The 44 associations were all involved in the utilisation of GHB (γ-hydroxybutyric acid) and D-galactonic acid lactone.

4.5.1.2 Results on random forests

Random forest identified genes involving 11 substrates, consisting of 29 genes involved in the utilisation of GHB (Figure 4.8), 8 genes involved in the utilisation of D-galactonic acid lactone (Figure 4.9), 4 genes involved in the utilisation of D-mannitol, 2 genes involved in D-alanine, D-gluconic acid, and uridine utilisation, 1 gene involved in α-keto glutaric acid, D-glucosaminic acid, D-raffinose, succinic acid, and thymidine utilisation (Appendix Figure III.I-III.IX).

The utilisation of GHB and D-galactonic acid lactone were demonstrated because the results of the two substrates exhibited strong relationships as found by other methods. Figures 4.8 and 4.9 display genes that were identified as relevant and tentatively relevant to the substrate utilisation. Genes identified as not important to the substrate utilisation were not displayed because of their abundance.

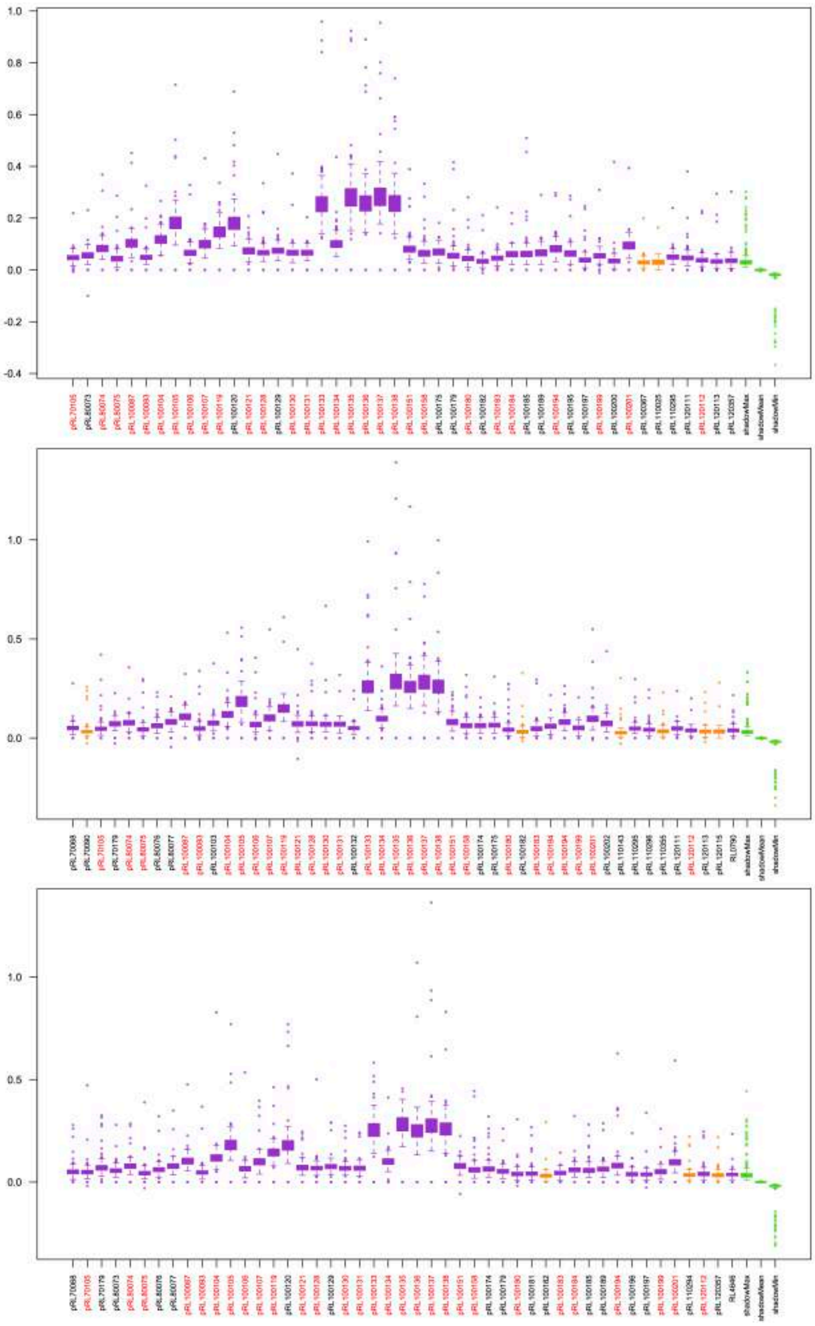


Figure 4.8 Z-score profiles of genes implicated in the utilisation of GHB by the Random Forest method. Three independent iterations of the gene selection algorithm are illustrated. Vertical axis represents Z-score. Columns represent genes coloured by the confirmation of their involvement in the three iterations (🟥 genes confirmed as relevant genes by 3 iterations and ■ genes confirmed relevant genes by 1-2 iterations). Each plot was coloured by Z-score by comparing to the maximum Z-score among shadow attributes (MZSA) (■ identified as genes involved in the substrate utilisation, 🟧 identified as tentative genes involved in the substrate utilisation, and 🟩 Z-score of shadow attributes). In box-and-whisker plot, the horizontal center of each box represents median, boxes represent 25th to 75th percentiles, and whiskers represent 10th and 90th percentiles. Dots located out of the box represent outliers.

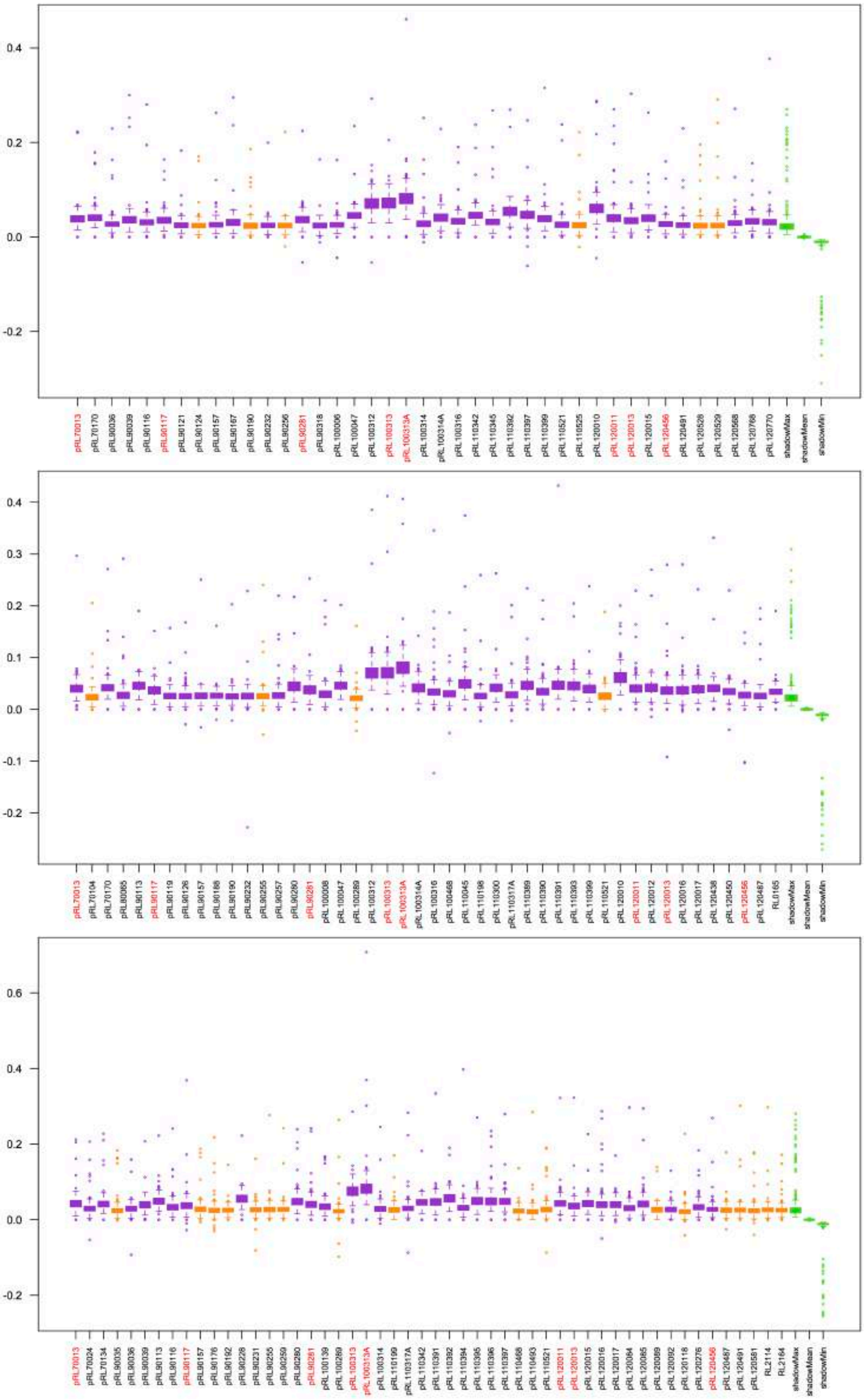


Figure 4.9 Z-score profiles of genes implicated in the utilisation of D-galactonic acid lactone by the Random Forest method. See legend to Figure 4.8 for details.

4.5.1.3 Results on class association rules

Of 21 substrates, two substrates, viz. GHB and D-galactonic acid lactone, presented groups of genes whose presence was strongly associated with substrate utilisation. The suggested width of rules was three but in this study the width of rules was varied from 1-4 genes in order to find the optimal width of rules for further analysis. The rules constructed by NETCAR were ordered by their *Accuracy* value. To reduce redundancy, rules with mutual information ≥ 0.25 were selected for further analysis. Table 4.1 represents parameters of different width rules with mutual information ≥ 0.25 and minimum accuracy.

In the case of GHB, all rules were generated with positive correlation, which mean that genes in the rules were present in strains utilising GHB. *Confidence* values increased when width of rules was changed from 1 to 2, and then stabilised. This meant that more than one gene participated in a pathway of substrate utilisation. In contrast, *support* values decreased when width of rules was changed from 1 to 2, and then stabilised because the number of strains utilising GHB and genes carried in the qualified rules compared against all strains in the analysed data were much different in size. However, the stable confidence and support values demonstrated consistency in the rules. *Accuracy* and *F1-score* values indicated performance of the constructed rules, when width of rules was changed from 1 to 2, and then stabilised. The decrease was acceptable. There was no significant change of the *support* and *accuracy* of the rules when the width of rules was increased from 2 to 4, while *confidence* increased when the width of rules was increased from 1 to 2 and stabilised after that. Hence the optimal width of rule was chosen by the number of genes involved in the GHB utilisation; the optimal width of rules was 4.

Unlike the correlation of genes found to be involved in GHB utilisation, genes correlated with D-galactonic acid lactone were absent from strains utilising D-galactonic acid lactone, or alternately negatively correlated genes. There was no qualified rule when width of rule was varied from 1-3 genes. It was observed that rules were generated with mutual information less than 0.25. However, qualified rules were observed when the width of rules was 4 genes.

The strength of the rules was not higher because the number of strains not utilising D-galactonic acid lactone was small compared to the ones utilising D-galactonic acid lactone. This resulted in low *strength* of the rules relating to the number of strains carrying the genes. Considering *accuracy* and *F-score* value in the previously mentioned equations, the *accuracy* was high because the number of strains that can utilise D-galactonic acid lactone (66 of 72 strains, and most of them did not carry the genes) was much larger than of those that cannot utilise D-galactonic acid lactone. The *F-score* was low compared to that for the utilisation of GHB. The *F-score* focused on the number of strains carrying the genes and not utilising D-galactonic acid lactone, which was a small number (6 of 72 strains).

Table 4.1 Number of rules and genes with mutual information ≥ 0.25. The *accuracy*, *confidence*, and *support* value is shown for the rule with the minimum F-score in each case. The optimal width of rules used in the study is in bold.

Substrate	**Width of rules**	**Number of rules**	**Number of genes**	***Accuracy***	***Confidence* (%)**	***Support* (%)**	**Min *F-score (%)***
GHB	1	9	9	0.792	85.2	31.9	75.9
	2	102	20	0.764	94.7	25	67.9
	3	757	24	0.764	94.7	25	67.9
	4	**6,919**	**57**	**0.764**	**94.7**	**25**	**67.9**
D-galactonic acid lactone	1	-	-	-	-	-	-
	2	**-**	**-**	**-**	**-**	**-**	**-**
	3	-	-	-	-	-	-
	4	**272**	**28**	**0.958**	**25**	**2.8**	**5.5**

4.5.2 Genes associated with GHB utilisation in the population

GHB utilisation is brought about by the activity of two dehydrogenases. The first dehydrogenase converts GHB to succinic semialdehyde (SSA). Another dehydrogenase converts SSA to SA (Figure 4.10), which is hence completely metabolised via the electron transport chain and is an intermediate of the tricarboxylic acid (TCA) cycle. TCA is beneficial to rhizobium for colonising plant host and developing effective nitrogen fixation (Dunn, 1998).

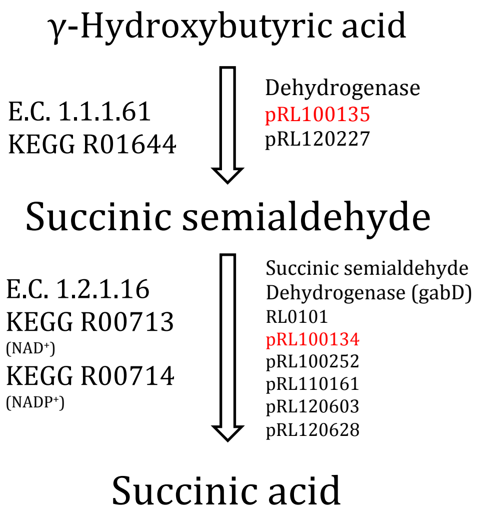


Figure 4.10 Pathway for utilisation of GHB from KEGG database (Tanabe *et al.*, 2002). Arrows define the direction of metabolism of each intermediate (EC enzyme numbers (E.C. X.X.X.XX), KEGG reaction number (KEGG RXXXXX) and locus tag of genes involved in each metabolic step). Locus tags in red were genes also found by using computational methods in this study.

In *Agrobacterium tumefaciens* C58, which is in a genus related to *Rhizobium*, GHB was involved in the utilisation of GBL as an intermediate (Carlier *et al.*, 2004). The sequence of reactions starts with a lactonase encoded by the gene *attM* (Zhang *et al.*, 2002) converting a ring-structured GBL to the open ring GHB. Subsequently, GHB is oxidised by *attL* to SSA, which is finally oxidised by *attK* (*gabD* (Prell *et al.*, 2009)) to SA. These genes were regulated by *attKLM* operon (Chevrot *et al.*, 2006, Wang *et al.*, 2006, White *et al.*, 2009), whose expression was repressed by *attJ* (Zhang *et al.*, 2002). The existence of GBL, GHB, and SSA in the cell was reported as *attKLM* operon inducers and *attJ* repressor (Carlier *et al.*, 2004, Chai *et al.*, 2007). Furthermore, *attM* can also catabolise *N*-acyl-homoserine lactones, which is known as a quorum-sensing signal in *Rhizobium* (Wisniewski-Dye *et al.*, 2002), to an open ring, *N*-acylhomoserine (Carlier *et al.*, 2004), resulting in inactivation of quorum sensing (Chai *et al.*, 2007).

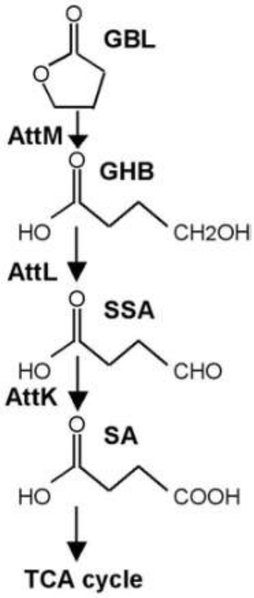


Figure 4.11 Pathway for γ-butyrolactone utilisation (γ-butyrolactone: GBL, GHB: GHB, succinic semialdehyde: SSA, and succinic acid: SA) (taken from Carlier *et al.* (2004)).

Genes identified as candidate genes relating to GHB utilisation were found on pRL7, pRL8, pRL10, pRL11 and pRL12 (gene list in Appendix Table III.II). A network of candidate genes involved in the GHB utilisation was constructed from rules generated by class association rules (Figure 4.12). The size of node represents frequency of nodes in all the generated rules. The six adjacent genes of large size were more frequently found in significant generated rules for GHB utilisation than the small ones. The four adjacent genes, viz. pRL100133 (*attJ* with 60% amino acid identity), pRL100134 (*gabD*) (*attK* with 78% amino acid identity), pRL100135 (*attL* with 83% amino acid identity), and pRL100136 (*attM* with 90% amino acid identity), are homologs of *attJKLM* genes of *Agrobacterium tumefaciens* C58. The adjacent genes pRL100137 (*metX*, homoserine O-acetyltransferase) and pRL100138 (MerR-family transcriptional regulator) are also identified as genes related to GHB utilisation (Kumar *et al.*, 2015). The role of pRL100135 (*attL*) and pRL100136 (*attM*) in GHB utilisation in *Rlv*3841 has been investigated and confirmed by mutational knock-out (Lad, 2013). These genes were discovered by all three computational methods. Two other genes identified by all three computational methods and confirmed as relevant genes in the utilisation are pRL100104 and pRL100105, which possibly encode subunits of polyhydroxybutyrate synthase. Another gene, pRL100103, identified by class association rule and pairwise correlation, is an alcohol dehydrogenase and homolog of *attL* with 51% amino acid identity.

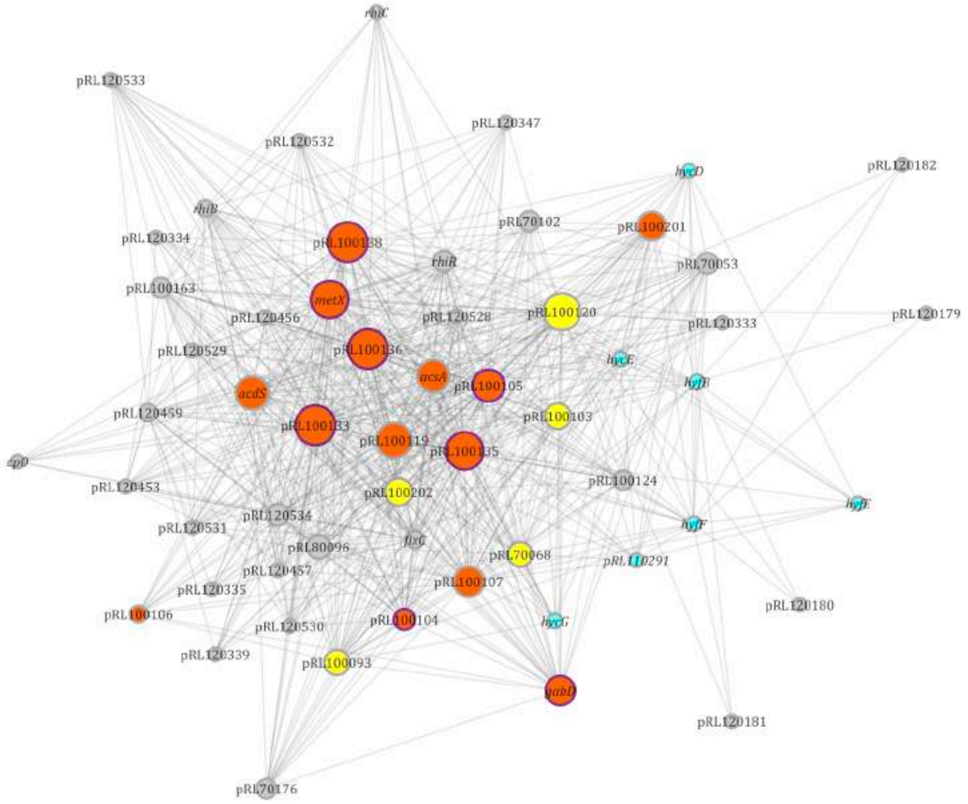


Figure 4.12 Genes involving in the utilisation of GHB. Nodes represent genes. Edges connect genes present in the same rules. The size of node represents frequency of nodes in the extracted rules. The orange nodes (●) represent genes found by pairwise association, random forest, and class association rules. The yellow nodes (●) represent genes found by two of these methods. The nodes with purple border represent annotated γ-hydroxybutyric-acid-related genes. The cyan nodes (●) represent genes that are found in strains that can utilise the substrate but do not have annotated GHB-related genes.

The distribution of genes associated with GHB in the population is shown in Figure 4.13, 34 strains in the population were able to utilise GHB. Of 34 strains, 28 strains carried genes pRL100103-pRL100105 and pRL100133-pRL100138, which were annotated with functions related to the utilisation of GHB.

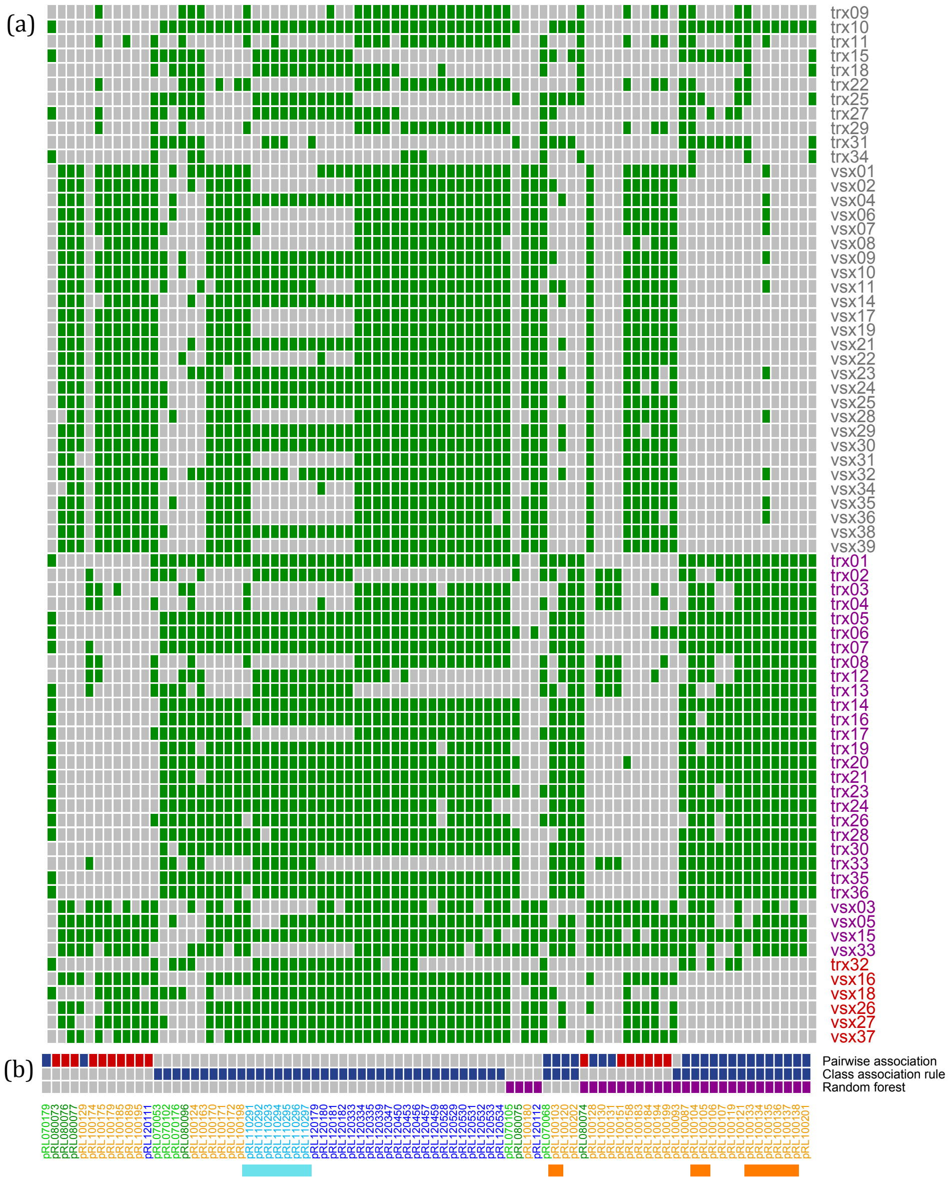


Figure 4.13 The distribution of genes associated with GHB utilisation in the population and profile of computational methods (a.) The distribution of genes associated with GHB utilisation. Rows represent rhizobium strains used in this study (● strains that cannot utilise substrate, ● strains that can utilise the substrate and have annotated γ-hydroxybutyric-acid-related genes and ● strains that can utilise the substrate but do not have annotated γ-hydroxybutyric-acid-related genes). Columns are genes associated with substrate utilisation (■ gene absent and ■ present). (b.) Profile of computational method identifying associated genes. Rows represent the computational methods and are coloured by their identified relation (■ gene not associated with substrate utilisation, ■ gene associated with substrate utilisation, ■ gene presence associated with substrate utilisation, and ■ gene absence associated with substrate utilisation). Column labels are coloured by their location (● chromosome, ● pRL7, ● pRL8, ● pRL9, ● pRL10, ● pRL11, and ● pRL12).

The genes confirmed as genes related to GHB utilisation were absent from trx32, vsx18, vsx16, vsx26, vsx27, and vsx37. However, the class association rules identified genes on pRL11 (● Figure 4.12 and ■ in Figure 4.13) that might be involved in the GHB utilisation and are carried by trx32, vsx18, vsx16, vsx26, vsx27, and vsx37. These genes were interesting because they were abundant in trx32, vsx18, vsx16, vsx26, vsx27, and vsx37 and absent in some of the other strains utilising GHB.

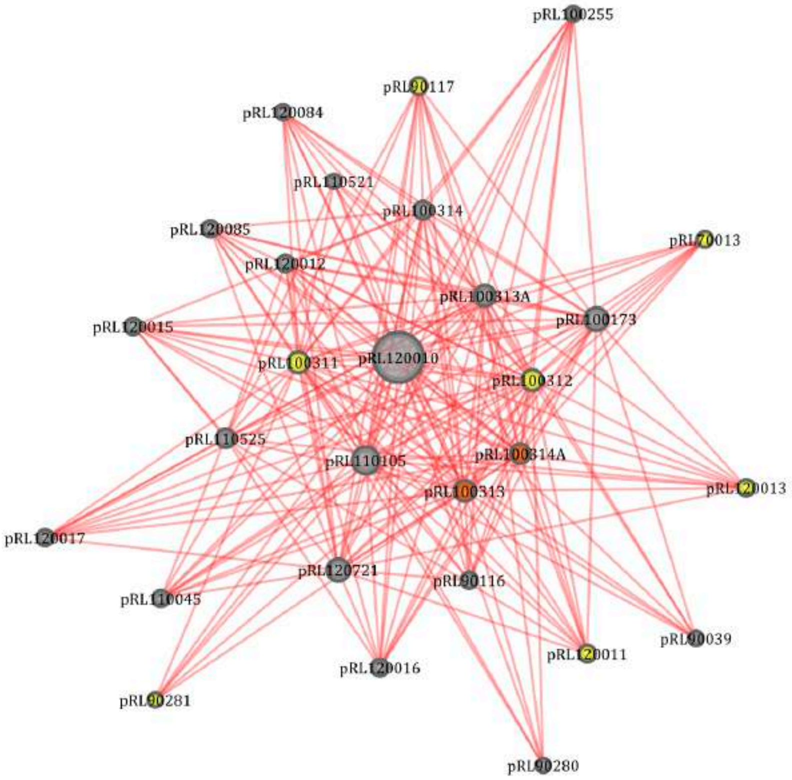


Figure 4.14 Genes involved in the utilisation of D-galactonic acid lactone. Nodes represent genes. Edges represent genes present in the same rule. The size of the node represents frequency of nodes in the extracted rules. The orange nodes (●) represent genes found by pairwise association, random forest, and class association rules. The yellow nodes (●) represent genes found by two of these methods.

4.5.3 Genes associated with D-galactonic acid lactone utilisation in the population

In contrast to the rules that generated the utilisation profile of GHB, results of class association rules for D-galactonic acid Lactone consisted of genes absent from strains able to use the substrate (Figure 4.14). Genes whose occurrence was involved in D-galactonic acid Lactone utilisation in the population were scattered on pRL7, pRL9, pRL10, pRL11, and pRL12 (genes list in Appendix Table III.III) (Figure 4.15). Amongst the significant generated rules, pRL120010 was frequently found because pRL120010 was absent in a majority of strains utilising the substrate (Figure 4.15).

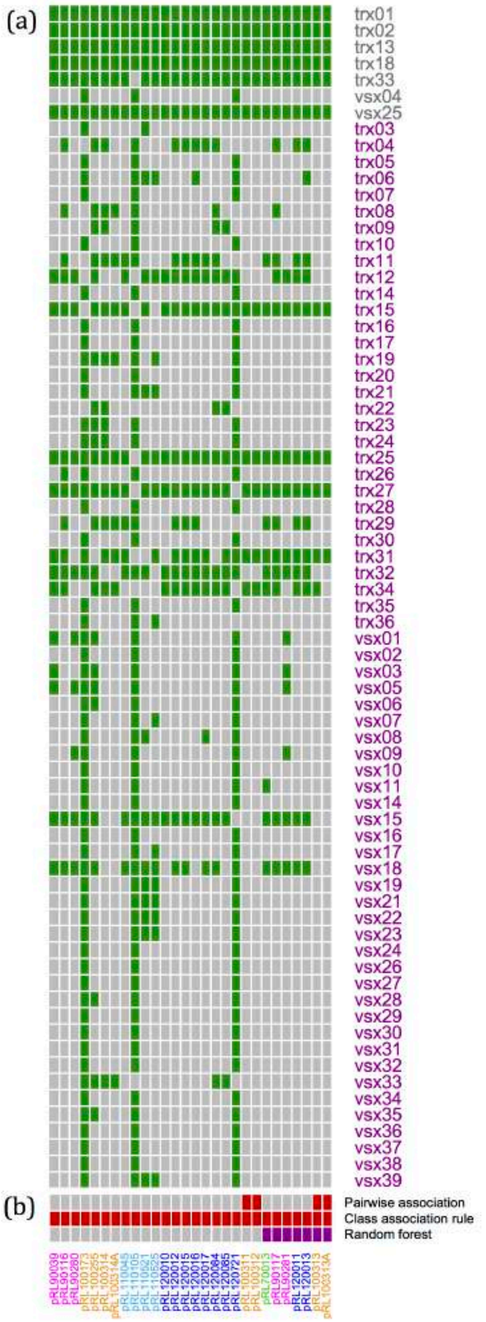


Figure 4.15 The distribution of genes associated with D-galactonic acid lactone utilisation in the population and profile of computational methods (a.) The distribution of genes associated with D-galactonic Acid Lactone utilisation. Rows represent rhizobium strains used in this study (● strains that cannot utilise the substrate and ● strains that can utilise the substrate). Columns are genes associated with substrate utilisation (■ gene absent gene and ■ present). (b.) Profile of computational method identifying associated genes. Rows represent the computational methods and are coloured by their identified relation (■ gene not associated with substrate utilisation, ■ gene associated with substrate utilisation, and ■ gene absence associated with substrate utilisation). Column labels are coloured by their location (● pRL7, ● pRL9, ● pRL10, ● pRL11, and ● pRL12).

4.5.4 Presence and absence of genes related to substrate utilisation in the population

Occurrence patterns of genes whose presence was involved in the substrate utilisation was investigated in order to find relationships between occurrence patterns of genes and their participation in the utilisation of the same substrate. In the co-occurrence network obtained in Chapter 3 with threshold of Pearson's correlation ≥ |0.70| (Figure 3.17 in Chapter 3), genes involved in the utilisation of the same substrate were in the same cluster (Figure 4.16). However, in the case of genes associated with GHB, there was more than one cluster of genes observed. Cluster 1 (Figure 4.16 (b)) contained genes that were annotated as genes involved in GHB metabolism. It was found that genes in Cluster 1 were carried by the almost same strains and most of them can utilise GHB. Cluster 3 (Figure 4.16 (b)) contained genes that have not been annotated as genes involved in GHB utilisation (Table 4.2). A difference between these two disjointed clusters was the distribution of genes in the population. Genes in the two disjointed clusters might be able to use GHB by an alternative pathway.

Table 4.2 List of candidate genes involved in GHB utilisation present in trx32, vsx18, vsx16, vsx26, vsx27, and vsx37.

Locus tag	Gene Symbol	Protein accession	Annotated function
pRL110292	*hycG*	YP_771325.1	putative formate hydrogenlyase subunit 7
pRL110293	*hycE*	YP_771326.1	putative formate hydrogenlyase subunit 5
pRL110294	*hyfF*	YP_771327.1	hydrogenase 4 subunit F
pRL110295	*hyfE*	YP_771328.1	putative hydrogenase-4 component E
pRL110296	*hycD*	YP_771329.1	putative hydrogenase protein
pRL110297	*hyfB*	YP_771330.1	hydrogenase 4 subunit B

Cluster 2 (Figure 4.16 (b)) exhibited an anti co-occurrence relationship between genes required in the utilisation of GHB and genes absent from strains utilising D-galactonic acid lactone. One plausible explanation is related to the strains utilisation pattern of the two substrates i.e. most strains able to utilise GHB also utilised D-galactonic acid lactone. Genes were negatively correlated with the utilisation of D-galactonic acid lactone and found in strains

that did not utilise the substrate. Cluster 2 implies that the genes involved in GHB utilisation are also positively associated with D-galactonic acid lactone utilisation.

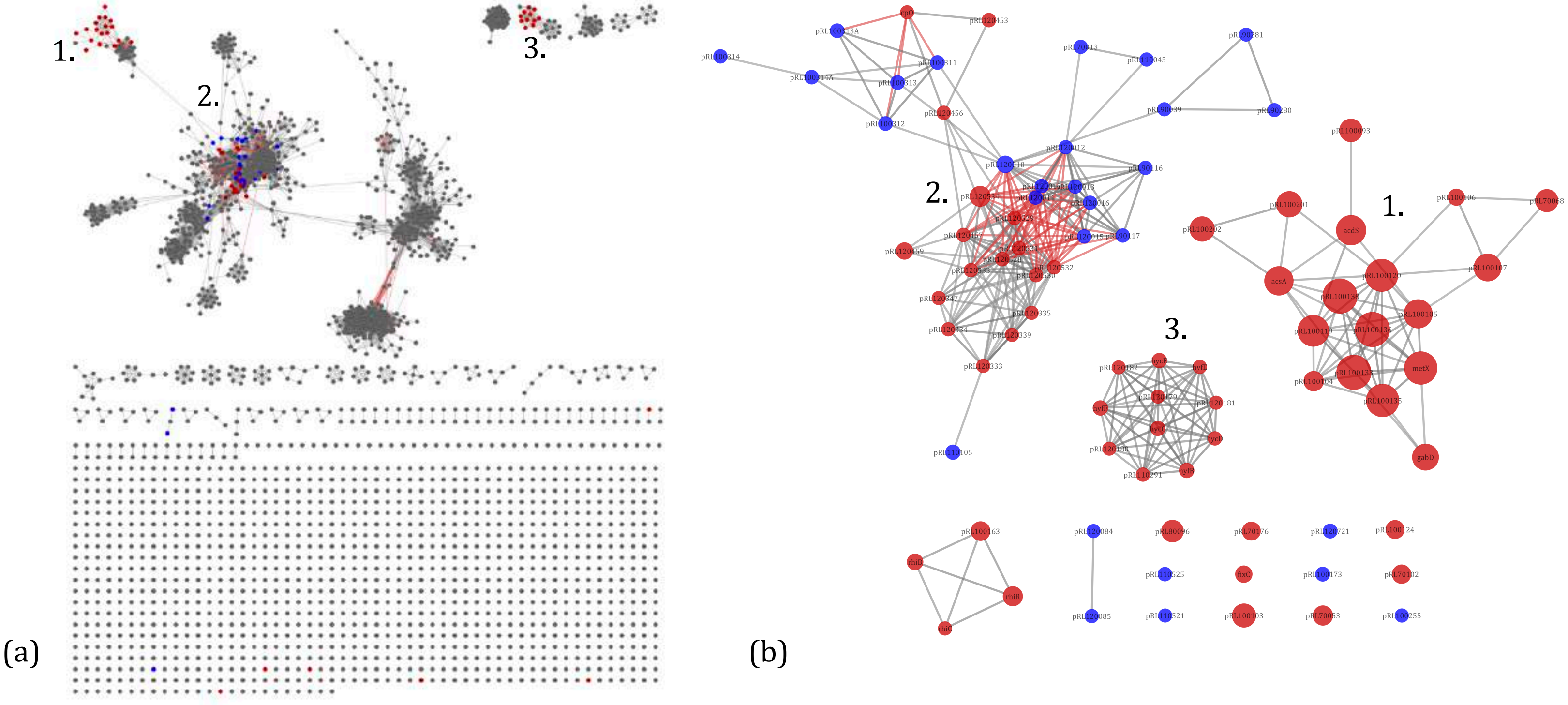


Figure 4.16 Genes associated with substrate utilisation in the population according to the gene co-occurrence network (as seen in Chapter 3). Nodes represent genes coloured by substrates (● and ● genes involved in GHB utilisation, D-galactonic acid lactone utilisation, respectively and ● genes not involved in utilisation of these substrates). Node size represents frequency of gene found in rules. Edges specify co-occurrence (—) and anti co-occurrence (—) interactions. (a) the entire gene co-occurrence network (b) genes involved in the substrate utilisation are enlarged.

4.6 Discussion

4.6.1 Gene selection by different methods

In this study, pairwise association (including Pearson's correlation and mutual information), random forest and class association rule were applied to analyse the data in order to select genes relevant to the utilisation of substrate.

The pairwise association methods were considered to be simply methods for extracting relationships between genes relating to substrate utilisation. The methods selected genes from the frequency of genes carried by strains utilising and not utilising substrate (Figure 4.4 and 4.6)(Dash *et al.*, 1997). These resulted in generating false positives when random high correlation values were identified, even though FDR was applied to remove results from these spurious correlations. The FDR was found to provide restricted results with the analysed data. For example, in the case of utilisation of GHB, candidate genes identified to be involved in the substrate utilisation were genes that were present in the majority of strains utilising the substrate (Figure 4.13). The methods missed genes in some strains which were able to use the substrate but did not carry genes with significant pairwise association.

Another interesting point observed when using pairwise association to select genes was that a majority of strains utilising GHB were bv. *trifolii*. The method identified bv. *viciae* specific genes having negative relationship to the substrate utilisation with significant statistics ($p \leq 0.001$, *two-sided Fisher's exact test*). However, not only bv. *trifolii* but bv. *viciae* too can utilise the substrate (Figure 4.13). This means that there exists a significant correlation between symbiovar and GHB utilisation, but this could be a chance association because utilisation is present, and absent, in strains of both symbiovars.

In this study, we used the implementation of random forest called the Boruta algorithm (Kursa *et al.*, 2010a). Due to random selection of genes to investigate potential of each gene, the random forest algorithm required iterative computation to obtain consistent computational results. Herein, the suggested three-time iteration was applied (Bayjanov *et al.*, 2013). However, the analysed data was considered to have an imbalance of class data (Chawla *et al.*, 2004), in which the number of strains utilising substrate and strains not utilising were much different. This is considered to be a general problem in feature selection (herein gene selection) using machine learning because the method focused on features that were significant to the majority of observations (referring to substrate utilising or non-utilising strains in this study) and ignored the minority of observations. To clarify this situation in the case of strains utilising GHB, the methods also could not identify genes in some strains, which were able to use the substrate but did not carry genes confirmed to be important to substrate utilisation (Figure 4.13). Considering identified genes relevant to substrates utilised by most of the population (ranging 52-71 strains), the method usually selected genes whose presence or absence was found in the majority class (which was either utilising strains or non-utilising strains) (Appendix Figure III.I-III.IX). The imbalance of class problem was overcome by balancing statistical measure performance of binary classification (Menon *et al.*, 2013). The implementation of random forest, Boruta, used in this study did not provide any statistical measure to guard against the imbalance of class problem.

Class association rule was another computational method used in this work. NETCAR (Tamura *et al.*, 2008), which is an implement of class association rule, was applied to the data. Potential chosen genes from NETCAR were evaluated by mutual information and *F-score*, as well as traditional measures such as *confidence* and *support* (Agrawal *et al.*, 1994). The mutual information can deal with redundancy of generated rules, while the *F-score* introduces measures for reducing cost of imbalance in class data (Menon *et al.*, 2013). It is noticeable that rules involving multiple genes were more powerful. This is exemplified by the pRL11 genes identified as candidate genes involved in the utilisation of GHB (Figure 4.13). The pRL11 genes were not detected by the other methods because of the distribution of the pRL11 genes in the population. Hence the class association rule was considered to be the most informative method in this study.

Pairwise association methods including mutual information and Pearson's correlation were beneficial to discover the relationship between an individual gene and substrate utilisation (one-to-one relationship). Results of gene selection based on pairwise association depend on the distribution of genes and not on observations (herein, rhizobium strains), like the random forest. The random forest cannot specify the direction of the relationship between genes and phenotype. However, the previous studies (Bayjanov *et al.*, 2013) suggested that the direction of the relationship can be inferred by comparing the distribution of genes in the analysed data. For example, in the case of GHB utilisation, genes annotated as relating to GHB utilisation on pRL10 (pRL100133-138) were all discovered by all computational methods (Figure 4.9), the pRL10 genes were carried by a majority of strains utilising GHB. Some strains utilising GHB did not carry the pRL10 genes. The class association rule method demonstrated genes on pRL11 and pRL12 as potential genes involving in GHB utilisation apart from the pRL10 genes and present in strains not carrying the pRL10 genes. This might be a result of more than one pathway with the capability for substrate utilisation. The class association rule can deal with many-to-one relationships and relationship direction. Hence, the class association rule was selected as the method of choice for gene selection in this study.

4.6.2 Genes relevant to the substrate utilisation

Substrate utilisation ability did not relate closely to the symbiovars and genospecies of the population (Figure 4.2). Diversity in the substrate utilisation ability was controlled by genes present in the rhizobium strains (Kumar *et al.*, 2015). This behaviour is also evident in *Pseudomonas aeruginosa* (Pommerenke *et al.*, 2010), *Lactococcus lactis* (Bayjanov *et al.*, 2013) and *Myxococcus xanthus* (Yan et al., 2014).

In this study, genes not present in every strain in the population were focused on because these genes represented the gene diversity of the population. Such genes are known as accessory genes (Young *et al.*, 2006). Hence all candidate genes were accessory genes, of which a majority were located on plasmids. The results of the study supported the hypothesis that the diversity in the distribution of accessory genes was relevant to the substrate utilisation.

Alternatively, accessory genes were carried by some strains for specific purposes. In the co-occurrence genes network, genes that participate in the same substrate utilisation have short distance in the network and were transferred together in the population. When genes utilised the substrate by different pathways, the genes might possibly present themselves as unconnected subnetworks in the co-occurrence genes subnetwork. A plausible explanation might be that the strain acquiring either of them was able to utilise the substrate.

In summary, this chapter made use of computational methods for selecting genes involved in the metabolic ability of the rhizobium population. To achieve this, the computational results from different methods were compared to find the most suitable method for the data. The most suitable tool, class association rule, was selected by referring to prior knowledge. The class association rule could identify nine genes on pRL10 that were annotated as GHB related genes. The method also discovered candidate genes carried in strains utilising GHB but not carrying the annotated genes. The results of this study demonstrated that the ability of substrate utilisation did not relate to symbiovar or genospecies but the ability of substrate utilisation depended on genes carried by the strains. Furthermore, most of the genes required for the substrate utilisation were transferred together in the population. However, an exception was found in the co-occurrence network, as disjointed subnetworks were found when genes possibly utilised the substrate by alternate pathways.

Chapter 5 General discussion

The distribution of genes across seventy-two strains of a local population of *Rhizobium leguminosarum* was investigated in a comprehensive way using computational approaches. Chapters 2 and 3 demonstrate uses of computational methods to identify clusters of genes that are functionally related by examining their distribution across strains in a population without prior knowledge of experimental data. In Chapter 2, gene transfer within the population affecting the distribution of genes in the bacterial population was investigated through incongruence of the gene tree of each gene. In Chapter 3, occurrence of genes in the population was studied in order to find patterns of gene co-occurrence. Chapter 4 demonstrates an integrated analysis using experimental data and the distribution of genes. Chapter 4 studied relationship, in the population, of occurrence of genes and ability to utilise substrate. The study illustrates some of the insights that can be gained by integrating comparatively simple data across multiple individuals, rather than by studying single individuals in detail. This chapter provides a synopsis that assesses the contribution and limitations of each chapter. Finally, some directions for future analysis and conclusions are presented, using knowledge obtained through this work.

5.1 Synopsis

In Chapter 2, gene distributions in a local population of *Rhizobium leguminosarum were* explored. Compositional methods, including atypical nucleotide content (Lawrence *et al.*, 1997, Karlin, 2001, van Passel *et al.*, 2005, Putonti *et al.*, 2006) and codon usage (Lawrence *et al.*, 1998) was applied to detect transferred genes. A comparison of sequence similarity between analysed genes and genes in a public database (Lawrence *et al.*, 1998, Lefébure *et al.*, 2010) has been used for assessing HGT using tools for searching similarity of sequences (i.e. BLAST (Altschul *et al.*, 1990)). Horizontally transferred genes were identified when the analysed genes and hit genes from the database had high similarity score. Incongruence on phylogenetic trees is another approach for gene transfer detection. A tree of analysed genes (e.g. set of orthologs or genes in the same family) is constructed and compared to a reference tree (i.e. core gene phylogeny). Genes transferred horizontally were observed when the analysed gene tree was not consonant with the reference tree. For example, phylogenies were constructed from core genes on distinct replicons compared to the core phylogeny for example *Sinorhizobium* (Bailly *et al.*, 2011) and *Rhizobium leguminosarum* (Tian *et al.*, 2010, Kumar *et al.*, 2015).

In this study, clanistics was used as a tool to explore gene distribution in a local population of *Rhizobium leguminosarum*. Clanistics (Lapointe *et al.*, 2010) was chosen as a tool because the method was flexible for detecting shared genes between two symbiovars and transferred genes within five genospecies. Using each of the two symbiovars as natives, many genes, particularly chromosomal genes, showed discordant patterns, reflecting the common background of core genes shared by the two symbiovars (Young *et al.*, 1987, Young *et al.*, 2006, Mauchline *et al.*, 2014). On the other hand, genes with concordant patterns were found in a particular symbiovar, of which a majority were known symbiosis genes (*nod*, *nif*, and *fix* genes) and used for defining the symbiovar (Long, 2001, Miller *et al.*, 2007, Rogel *et al.*, 2011). The genes with concordant patterns were overrepresented on plasmids. The overrepresentation of symbiosis genes with concordant clanistic patterns reflected the fact that the symbiosis genes on plasmids served to differentiate the two symbiovars. Not only were annotated symbiosis genes discovered using clanistics, but also a novel gene (pRL100177) on pRL10 that is specific to bv. *viciae* but has no annotated function.

Clanistics was also used for detecting transferred and non-transferred genes within the five genospecies. The chromosome not only carried genes with no evidence for HGT but also carried mobile genes. The observation of core genes shared either by the whole population or just a single genospecies in the bacterial population has been reported not only in our study but also in other bacteria such as *Streptococcus* (Lefébure *et al.*, 2007), *Sinorhizobium* (Bailly *et al.*, 2011, Sugawara *et al.*, 2013), *Campylobacter* (Lefébure *et al.*, 2010, Méric *et al.*, 2014), *Agrobacterium tumefaciens* (Lassalle *et al.*, 2011) and *Lactobacillus* *paracasei* (Smokvina et al., 2013). Genospecies-specific genes were found on plasmids and chromids, which are known to carry adaptive and mobilisable genes in bacterial population genomics studies (Heuer *et al.*, 2012, Galardini *et al.*, 2013, Sentchilo *et al.*, 2013). Some of these specific genes on pRL9 and pRL12 have also been annotated as genospecies B specific genes by a different approach (Kumar *et al.*, 2015). The novel genospecies B specific genes emphasised that not only chromosomal genes but also plasmid-encoded genes (or, at least, genes that are plasmid-encoded in *Rlv*3841) may be responsible for the differences between genospecies.

Investigation into the relationship between functional assignment and gene transfer in the population demonstrated that genes relevant to the mobilome (prophages, transposons) were overrepresented among those with evidence for HGT within the five genospecies because transposons and plasmids are mobilome elements (Siefert, 2009) and generally transferred within the population (Nakamura *et al.*, 2004, Beiko *et al.*, 2005, Tamminen *et al.*, 2012).

Genes relevant to operational categories tended to be transferred, as noted in previous studies (Jain *et al.*, 1999, Nakamura *et al.*, 2004, Zhaxybayeva *et al.*, 2006, Kanhere *et al.*, 2009). Genes with evidence for HGT were also preferentially found in translation, ribosomal structure and biogenesis, transcription, and signal transduction mechanisms categories, which contradicts the complexity hypothesis of Jain *et al.* (1999), but is in agreement with Kanhere *et al.* (2009), Wisniewski-Dyé *et al.* (2012), Dziewit *et al.* (2014), and Epstein *et al.* (2014). Genes in the “poorly characterized” category frequently had evidence for HGT, which has also been seen in previous studies of comparative genomics in other bacteria (Wisniewski-Dyé *et al.*, 2012, Dziewit *et al.*, 2014, Epstein *et al.*, 2014) and reflects the fact that the functions of accessory genes are generally less well understood than those of core genes. However, in contrast to many other studies, Choi *et al.* (2007) reported that there was no association between functional categories and gene transferred.

This chapter confirmed that there was instability of the bacterial genomes within the local rhizobium population, which were affected by homologous recombination or HGT. Similar instability of the core genome was also found in other bacterial populations (Didelot *et al.*, 2010, Beauregard-Racine *et al.*, 2011, Didelot *et al.*, 2011, Cadillo-Quiroz *et al.*, 2012).

Studies of gene content in bacterial populations using multiple genomes can identify contributions of gene transfer to the population, for example genospecies maintenance (Bailly *et al.*, 2011, Lassalle *et al.*, 2011, Smokvina *et al.*, 2013, Sugawara *et al.*, 2013, Kumar *et al.*, 2015), host specificity (Rogel *et al.*, 2011, Sugawara *et al.*, 2013), virulence (Hogg *et al.*, 2007, Lefébure *et al.*, 2007, Lefébure *et al.*, 2010, Beauregard-Racine *et al.*, 2011, Méric *et al.*, 2014) and environmental adaptation (Shapiro *et al.*, 2012, Smokvina *et al.*, 2013).

Chapter 3 investigated occurrence patterns of genes in the rhizobium population. This chapter exploited correlational computation and network analysis on gene present/absent data of multiple genomes. A comparable study of *Mycoplasma genitalium* (Huynen *et al.*, 2000) revealed genes related to the same function using mutual information for relationship quantification. Rather than using the phylogenetic profile of genes of one species, Kim *et al.* (2011) employed occurrence of genes in multiple bacterial species, and their results also showed gene occurrence patterns across species. These two studies reported results on co-occurrence and anti co-occurrence relationships compatible with our study.

Pearson's correlation was chosen to find gene occurrence relationships classified into co-occurrence genes with positive correlation and anti co-occurrence genes. Network analysis was applied to a massive set of correlations by converting numerical values, filtered by an optimal threshold, to a view of the gene co-occurrence network. The gene co-occurrence network reflected global relationships of genes in the population.

Genes with positive correlation, i.e. co-occurrence genes, were frequently found to be neighbouring genes (Tamames, 2001) which might be transferred via horizontal gene transfer (Achtman *et al.*, 2008). Co-occurrence genes were placed in the same subnetwork and subnetworks could be used to identify functionally related sets of genes like, for example, the symbiosis genes (Young *et al.*, 2006) and *rhi* genes (Rodelas *et al.*, 1999, Wisniewski-Dye *et al.*, 2002) on pRL10, which are relevant to nodulation and nitrogen fixation. Furthermore, co-occurrence relationships of *attKLM* (pRL100134, pRL100135, and pRL100136) (Chai *et al.*, 2007, Prell *et al.*, 2009, White *et al.*, 2009) represent genes necessary for γ-hydroxybutyric acid utilisation. Another set of co-occurrence genes, five *bvs* genes on pRL8, were found only in bv. *viciae* (Kumar *et al.*, 2015). pRL100162A and pRL100163, encoding asparagine synthase, were also in this subnetwork. Asparagine, which is an amino acid produced by nodules, may regulate nodulation (Oti-Boateng *et al.*, 1993, Lodwig *et al.*, 2003). Moreover, community detection not only allowed us to recognise genes present under the same conditions amongst complicated relationships in the subnetwork, for example, the community of genes specific to genospecies B in the 374-gene subnetwork, which have been reported as a genospecies B specific-island (Kumar, 2013), but also revealed the subtle relationships within the 480-gene subnetwork, the presence of chromosomal and non-chromosomal communities implied that these replicons took responsibilities for different functions in the population.

Genes with negative correlation, anti co-occurrence genes, also appeared in the constructed network. These anti co-occurrence genes might be on alternative mobile genetic elements such as GIs which could be identified by IslandViewer 3 (Dhillon *et al.*, 2015). The other possible explanation for the anti co-occurrence relationship between subnetworks might be their replaceable functions.

Chapter 4 aimed to investigate phenotypic differences of the population in relation to the distribution of genes in the population. The chapter employed computational methods for selecting genes involved in the metabolic ability of the rhizobium population. Class association rules were identified as the most suitable tool for the data and can identify genes involved in γ-hydroxybutyric acid utilisation. Some of the candidate genes on pRL10 were annotated as γ-hydroxybutyric acid related genes (Chai *et al.*, 2007, Prell *et al.*, 2009, White *et al.*, 2009). Interestingly, the genes reported as γ-hydroxybutyric acid related genes were not present in all strains utilising the substrate. The class association rule can identify other candidate genes carried in strains utilising γ-hydroxybutyric acid besides the annotated genes. The study emphasised that the diversity of presence of accessory genes was relevant to phenotypic differences in the population such as substrate utilisation abilities. The study demonstrated that the ability to use substrates did not relate to symbiovar or genospecies, but depended on genes carried in the strains. This is compatible with studies in *Pseudomonas aeruginosa* (Pommerenke *et al.*, 2010), *Lactococcus lactis* (Bayjanov *et al.*, 2013), *Myxococcus xanthus* (Yan et al., 2014), and *Lactobacillus rhamnosus* (Ceapa et al., 2015). A majority of the genes that were identified as relevant to substrate utilisation were on plasmids and shared by some, but not all, strains in the population.

5.2 Directions for future study

In general, this study exploited computational approaches to the raw data obtained from laboratory studies, such as sequencing data in Chapter 2, phylogenetic profiles of genes in Chapter 3, and substrate utilisation profiles in Chapter 4, all of which data were obtained from multiple strains of bacteria. Comparative genomic study here illustrates some of the insights that can be gained by integrating comparatively simple data across multiple individuals, rather than by studying single individuals in detail. Clanistics as a tool to explore gene distribution (Chapter 2) can be beneficial for studying the distribution of genes in the population by using a user-defined category that varies with the research question, such as pathogenicity (Beauregard-Racine *et al.*, 2011, Xu *et al.*, 2014) or life style of organism (Schliep *et al.*, 2011). This could be extended to other user-defined categories in order to answer evolutionary questions. Correlational computation and network analysis of gene occurrence would be useful for viewing gene organisation at the genome level. Some subnetworks were validated with the experimental results or available literature, but many other subnetworks were identified for which there is not yet supporting evidence from either the laboratory or literature. To extend this study, laboratory-based experiments could be conducted to extend the interpretation of the network. Rather than using phylogenetic profiles of genes to study variation of gene content, as in the bacteria presented in this study, or functional annotation (Huynen *et al.*, 2000, Kim *et al.*, 2011), the principle of the method can apply to other data in a comparable format with different research questions. The presence/absence of bacteria has been used to explore the relationship of bacteria and their habitat (Barberan *et al.*, 2012, Faust *et al.*, 2012). Phylogenetic profiles of mutant and non mutant genes were informative for drug discovery (Cui, 2010). In Chapter 4, the analysis on integrated data (phylogenetic profiles of gene and substrate utilisation profile) allows us to comprehend genes involved in substrate utilisation. Although some candidate genes identified by the class association rule method were supported by literature and their annotated function, there remained genes without supporting evidence. To validate these candidate genes, laboratory studies could be done. Apart from correlating the distribution of gene and substrate utilisation in the population, this approach can apply to the other analysis of phenotype-genotype relationships with analogous data. Published examples include genes across multiple microbial species and profiles of intracellular pathogenicity phenotype obtained from the NCBI database (Slonim *et al.*, 2006) or COG database (Goh *et al.*, 2006, Tamura *et al.*, 2008), and genes associated with disease (Li *et al.*, 2014), for which phenotype-genotype data were generated from PubMed, containing information of symptom and disease, disease gene association databases (PharmCKB, OMIM, and CTD), and protein databases (genotype data) (MINT, DIP, HPRD, and IntAct).

As we are witnessing an incredible increment of massive genomic data, the systematic computation used in this study will hopefully be increasingly useful for comparative genomic studies.

Appendix I

Table I.I The 305 core genes held by pRL8, pRL9, pRL10, pRL11, pRL12 and chromosome (Harrison *et al.,* 2010). Locus tag, gene symbol, protein accession, replicon, and HGT index are mentioned in Chapter 2.

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
pRL080044	*acsA*	YP_770942.1	pRL8	3
pRL090212	-	YP_765499.1	pRL9	2
pRL100453	-	YP_770729.1	pRL10	3
pRL110033	-	YP_771066.1	pRL11	0
pRL110442	*thiE*	YP_771476.1	pRL11	2
pRL120279	*prC*	YP_764789.1	pRL12	2
pRL120359	*panC*	YP_764869.1	pRL12	3
pRL120360	*panB*	YP_764870.1	pRL12	1
pRL120416	*dadX*	YP_764923.1	pRL12	2
pRL120642	*groEL*	YP_765148.1	pRL12	0
pRL120643	*groS*	YP_765149.1	pRL12	0
RL0003	*aroE*	YP_765607.1	Chromosome	0
RL0004	*coaE*	YP_765608.1	Chromosome	0
RL0012	*gyrB*	YP_765616.1	Chromosome	0
RL0021	*trpB*	YP_765625.1	Chromosome	3
RL0022	*trpA*	YP_765626.1	Chromosome	1
RL0024	*folC*	YP_765628.1	Chromosome	3
RL0025	-	YP_765629.1	Chromosome	5
RL0029	-	YP_765633.1	Chromosome	0
RL0042	*hisF*	YP_765646.1	Chromosome	0
RL0043	*hisA*	YP_765647.1	Chromosome	0
RL0046	*hisH*	YP_765650.1	Chromosome	0
RL0048	*hisB*	YP_765652.1	Chromosome	2
RL0106	*rpsA*	YP_765710.1	Chromosome	1
RL0108	*aroA*	YP_765712.1	Chromosome	2
RL0120	*pnp*	YP_765724.1	Chromosome	1
RL0123	*truB*	YP_765727.1	Chromosome	2
RL0125	*infB*	YP_765729.1	Chromosome	2
RL0127	*nusA*	YP_765731.1	Chromosome	1
RL0131A	*recR*	YP_765736.1	Chromosome	2
RL0134	*dnaX*	YP_765739.1	Chromosome	0
RL0139	-	YP_765744.1	Chromosome	0
RL0151	*dnaJ*	YP_765756.1	Chromosome	0
RL0152	*dnaK*	YP_765757.1	Chromosome	1
RL0160	*polA*	YP_765765.1	Chromosome	1

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL0161	-	YP_765766.1	Chromosome	0
RL0181	-	YP_765786.1	Chromosome	1
RL0254	*lepA*	YP_765860.1	Chromosome	2
RL0268	*rplT*	YP_765874.1	Chromosome	2
RL0269	*pheS*	YP_765875.1	Chromosome	1
RL0270	*pheT*	YP_765876.1	Chromosome	0
RL0282	*xseA*	YP_765888.1	Chromosome	1
RL0315	*guaA*	YP_765921.1	Chromosome	1
RL0326	-	YP_765932.1	Chromosome	2
RL0328	-	YP_765934.1	Chromosome	2
RL0334	*dnaN*	YP_765940.1	Chromosome	0
RL0335	-	YP_765941.1	Chromosome	1
RL0357	*coaBC*	YP_765964.1	Chromosome	3
RL0371	*ubiE*	YP_765978.1	Chromosome	1
RL0375	*dnaA*	YP_765982.1	Chromosome	1
RL0377	*hemN*	YP_765984.1	Chromosome	0
RL0378	-	YP_765985.1	Chromosome	1
RL0382	-	YP_765989.1	Chromosome	0
RL0388	*trmB*	YP_765995.1	Chromosome	0
RL0389	*metK*	YP_765996.1	Chromosome	1
RL0393	-	YP_766000.1	Chromosome	0
RL0394	*phoH*	YP_766001.1	Chromosome	0
RL0395	*miaB*	YP_766002.1	Chromosome	0
RL0404	*mviN*	YP_766011.1	Chromosome	0
RL0406	*mutS*	YP_766013.1	Chromosome	0
RL0421	-	YP_766028.1	Chromosome	0
RL0433	*fmt*	YP_766040.1	Chromosome	0
RL0445	*argB*	YP_766052.1	Chromosome	0
RL0504	*pgi*	YP_766113.1	Chromosome	1
RL0550	*argF*	YP_766160.1	Chromosome	3
RL0572	-	YP_766181.1	Chromosome	0
RL0611	*murA*	YP_766221.1	Chromosome	3
RL0613	*hisD*	YP_766223.1	Chromosome	2
RL0616	*infA*	YP_766226.1	Chromosome	2
RL0680	-	YP_766290.1	Chromosome	0
RL0743	-	YP_766353.1	Chromosome	0
RL0847	*guaB*	YP_766458.1	Chromosome	0
RL0877	*hisS*	YP_766489.1	Chromosome	0
RL0883	*groEL*	YP_766495.1	Chromosome	1

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL0884	*groES*	YP_766496.1	Chromosome	4
RL0886	*ribF*	YP_766498.1	Chromosome	1
RL0889	*ileS*	YP_766501.1	Chromosome	2
RL0891	-	YP_766503.1	Chromosome	4
RL0892	-	YP_766504.1	Chromosome	2
RL0910	*mutL*	YP_766522.1	Chromosome	0
RL0920	-	YP_766532.1	Chromosome	0
RL0930	*rnhB*	YP_766542.1	Chromosome	0
RL0937	*ispB*	YP_766549.1	Chromosome	1
RL0945	*aroA*	YP_766557.1	Chromosome	1
RL0947	*purD*	YP_766559.1	Chromosome	0
RL0956	*ubiA*	YP_766568.1	Chromosome	0
RL0960	-	YP_766572.1	Chromosome	0
RL0969	*rumA*	YP_766581.1	Chromosome	3
RL0973	*dxs*	YP_766585.1	Chromosome	2
RL1007	*aroC*	YP_766618.1	Chromosome	2
RL1014	*pdxH*	YP_766625.1	Chromosome	1
RL1030	*ispH*	YP_766641.1	Chromosome	0
RL1078	*mutY*	YP_766689.1	Chromosome	4
RL1262	-	YP_766867.1	Chromosome	1
RL1370	*msrB*	YP_766976.1	Chromosome	0
RL1412	*groEL*	YP_767017.1	Chromosome	0
RL1503	*smpB*	YP_767107.1	Chromosome	3
RL1510	*sipS*	YP_767114.1	Chromosome	1
RL1543	*cysS*	YP_767147.1	Chromosome	2
RL1546	*purF*	YP_767150.1	Chromosome	2
RL1548	*radA*	YP_767152.1	Chromosome	0
RL1550	-	YP_767154.1	Chromosome	0
RL1551	*dnaC*	YP_767155.1	Chromosome	4
RL1552	*rplI*	YP_767156.1	Chromosome	2
RL1554	*rpsR*	YP_767158.1	Chromosome	2
RL1558	*fabG*	YP_767162.1	Chromosome	1
RL1564	*ksgA*	YP_767168.1	Chromosome	0
RL1580	*ndk*	YP_767184.1	Chromosome	3
RL1595	*purN*	YP_767199.1	Chromosome	2
RL1596	*purM*	YP_767200.1	Chromosome	2
RL1605	*aspS*	YP_767209.1	Chromosome	1
RL1616	*hemB*	YP_767220.1	Chromosome	1
RL1620	*glyA*	YP_767224.1	Chromosome	2

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL1621	*ribD*	YP_767225.1	Chromosome	2
RL1632	*ribH*	YP_767236.1	Chromosome	1
RL1668	*argC*	YP_767272.1	Chromosome	1
RL1672	*rpsI*	YP_767276.1	Chromosome	2
RL1673	*rplM*	YP_767277.1	Chromosome	4
RL1688	*clpP*	YP_767292.1	Chromosome	1
RL1723	*dnaE*	YP_767327.1	Chromosome	2
RL1735	*topA*	YP_767339.1	Chromosome	0
RL1736	*smf*	YP_767340.1	Chromosome	0
RL1737	-	YP_767341.1	Chromosome	0
RL1739	*pyrB*	YP_767343.1	Chromosome	0
RL1760	*nusG*	YP_767364.1	Chromosome	0
RL1761	*rplK*	YP_767365.1	Chromosome	5
RL1762	*rplA*	YP_767366.1	Chromosome	2
RL1764	*rplJ*	YP_767368.1	Chromosome	2
RL1765	*rplL*	YP_767369.1	Chromosome	2
RL1767	*rpoC*	YP_767371.1	Chromosome	2
RL1770	*rpsG*	YP_767374.1	Chromosome	1
RL1771	*fus*	YP_767375.1	Chromosome	0
RL1774	*rplC*	YP_767378.1	Chromosome	3
RL1775	*rplD*	YP_767379.1	Chromosome	1
RL1776	*rplW*	YP_767380.1	Chromosome	3
RL1777	*rplB*	YP_767381.1	Chromosome	1
RL1778	*rpsS*	YP_767382.1	Chromosome	2
RL1779	*rplV*	YP_767383.1	Chromosome	3
RL1780	*rpsC*	YP_767384.1	Chromosome	1
RL1781	*rplP*	YP_767385.1	Chromosome	5
RL1783	*rpsQ*	YP_767387.1	Chromosome	1
RL1784	*rplN*	YP_767388.1	Chromosome	3
RL1785	*rplX*	YP_767389.1	Chromosome	3
RL1786	*rplE*	YP_767390.1	Chromosome	2
RL1788	*rpsH*	YP_767392.1	Chromosome	3
RL1789	*rplF*	YP_767393.1	Chromosome	3
RL1790	*rplR*	YP_767394.1	Chromosome	3
RL1791	*rpsE*	YP_767395.1	Chromosome	2
RL1793	*rplO*	YP_767397.1	Chromosome	1
RL1794	*secY*	YP_767398.1	Chromosome	1
RL1795	*adk*	YP_767399.1	Chromosome	1
RL1797	*rpsK*	YP_767401.1	Chromosome	1

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL1798	*rpoA*	YP_767402.1	Chromosome	0
RL1799	*rplQ*	YP_767403.1	Chromosome	2
RL1803	*ilvD*	YP_767407.1	Chromosome	1
RL2035	*valS*	YP_767633.1	Chromosome	1
RL2041	*argS*	YP_767639.1	Chromosome	0
RL2043	*nagZ*	YP_767641.1	Chromosome	1
RL2048	*tatC*	YP_767646.1	Chromosome	0
RL2049	*serS*	YP_767647.1	Chromosome	1
RL2050	*surE*	YP_767648.1	Chromosome	1
RL2055	*secD*	YP_767653.1	Chromosome	0
RL2069	*map*	YP_767667.1	Chromosome	2
RL2099	*recJ*	YP_767697.1	Chromosome	0
RL2221	*rpsB*	YP_767815.1	Chromosome	0
RL2222	*tsf*	YP_767816.1	Chromosome	0
RL2223	*pyrH*	YP_767817.1	Chromosome	0
RL2224	*frr*	YP_767818.1	Chromosome	0
RL2225	*uppS*	YP_767819.1	Chromosome	2
RL2227	*ecfE*	YP_767821.1	Chromosome	1
RL2238	*kdsA*	YP_767832.1	Chromosome	0
RL2239	*eno*	YP_767833.1	Chromosome	1
RL2249	-	YP_767843.1	Chromosome	3
RL2254	*ispDF*	YP_767848.1	Chromosome	1
RL2255	*dus*	YP_767849.1	Chromosome	3
RL2288	*cysG2*	YP_767882.1	Chromosome	3
RL2381	*glmU*	YP_767971.1	Chromosome	2
RL2382	*glmS*	YP_767972.1	Chromosome	2
RL2384	*recG*	YP_767974.1	Chromosome	2
RL2386	*mfd*	YP_767976.1	Chromosome	2
RL2392	*glnA*	YP_767982.1	Chromosome	1
RL2393	*glnB*	YP_767983.1	Chromosome	0
RL2398	*uvrA*	YP_767988.1	Chromosome	3
RL2399	*ssb*	YP_767989.1	Chromosome	3
RL2401	*gyrA*	YP_767991.1	Chromosome	1
RL2403	*coaD*	YP_767993.1	Chromosome	2
RL2406	*queA*	YP_767996.1	Chromosome	1
RL2407	*tgt*	YP_767997.1	Chromosome	1
RL2442	*ilvI*	YP_768032.1	Chromosome	1
RL2472	-	YP_768057.1	Chromosome	1
RL2473	*metG*	YP_768058.1	Chromosome	1

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL2476	*tmk*	YP_768061.1	Chromosome	2
RL2493	*trpD*	YP_768077.1	Chromosome	4
RL2494	*trpC*	YP_768078.1	Chromosome	3
RL2511	*pyrG*	YP_768095.1	Chromosome	0
RL2528	*thrS*	YP_768112.1	Chromosome	0
RL2532	*hisI*	YP_768116.1	Chromosome	0
RL2555	*lipB*	YP_768139.1	Chromosome	0
RL2588	*tyrS*	YP_768172.1	Chromosome	0
RL2598	*rpe*	YP_768182.1	Chromosome	0
RL2612	*purL*	YP_768196.1	Chromosome	0
RL2624	*rpsD*	YP_768208.1	Chromosome	2
RL2627	*murI*	YP_768211.1	Chromosome	2
RL2636	*alaS*	YP_768220.1	Chromosome	0
RL2637	*recA*	YP_768221.1	Chromosome	0
RL2648	-	YP_768232.1	Chromosome	1
RL2650	*folC*	YP_768234.1	Chromosome	0
RL2691	-	YP_768276.1	Chromosome	1
RL2798	*leuS*	YP_768383.1	Chromosome	2
RL2801	*ddl*	YP_768386.1	Chromosome	1
RL2824	*cobA*	YP_768409.1	Chromosome	3
RL2957	*uvrB*	YP_768542.1	Chromosome	2
RL2987	*argG*	YP_768571.1	Chromosome	2
RL2990	*ubiA*	YP_768573.1	Chromosome	2
RL3013	*tyrS*	YP_768596.1	Chromosome	1
RL3071	*ftsZ*	YP_768653.1	Chromosome	3
RL3170	-	YP_768750.1	Chromosome	2
RL3205	*ilvC*	YP_768785.1	Chromosome	3
RL3244	*ilvH*	YP_768824.1	Chromosome	1
RL3245	*ilvI*	YP_768825.1	Chromosome	0
RL3249	*miaA*	YP_768830.1	Chromosome	1
RL3276	*pcrA*	YP_768857.1	Chromosome	0
RL3293	*ligA*	YP_768872.1	Chromosome	0
RL3295	*recN*	YP_768874.1	Chromosome	1
RL3298	*ftsZ*	YP_768877.1	Chromosome	1
RL3301	*ddl*	YP_768880.1	Chromosome	2
RL3306	*murC*	YP_768885.1	Chromosome	0
RL3307	*murG*	YP_768886.1	Chromosome	1
RL3309	*murD*	YP_768888.1	Chromosome	1
RL3310	*mraY*	YP_768889.1	Chromosome	0

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL3311	*murF*	YP_768890.1	Chromosome	0
RL3312	*murE*	YP_768891.1	Chromosome	0
RL3313	-	YP_768892.1	Chromosome	0
RL3315	*mraW*	YP_768894.1	Chromosome	0
RL3402	*rpoD*	YP_768982.1	Chromosome	1
RL3408	*dnaG*	YP_768988.1	Chromosome	2
RL3411	*carA*	YP_768991.1	Chromosome	3
RL3419	*carB*	YP_768999.1	Chromosome	3
RL3460	*proC*	YP_769040.1	Chromosome	1
RL3465	-	YP_769045.1	Chromosome	0
RL3468	*prs*	YP_769048.1	Chromosome	0
RL3471	-	YP_769051.1	Chromosome	0
RL3474	*pth*	YP_769054.1	Chromosome	1
RL3479	*ychF*	YP_769059.1	Chromosome	0
RL3521	*trpE*	YP_769101.1	Chromosome	0
RL3553	*engA*	YP_769133.1	Chromosome	1
RL3765	*rLuD*	YP_769344.1	Chromosome	2
RL3768	*purA*	YP_769347.1	Chromosome	2
RL3957	*mnmA*	YP_769535.1	Chromosome	0
RL3965	*ftsH*	YP_769543.1	Chromosome	1
RL3983	-	YP_769560.1	Chromosome	1
RL3986	*ruvC*	YP_769563.1	Chromosome	2
RL3989	*ruvA*	YP_769566.1	Chromosome	1
RL3990	*ruvB*	YP_769567.1	Chromosome	0
RL4006	*cbbT*	YP_769583.1	Chromosome	0
RL4007	*gap*	YP_769584.1	Chromosome	1
RL4017	*rpmE*	YP_769594.1	Chromosome	0
RL4044	*purE*	YP_769621.1	Chromosome	0
RL4060	*pykA*	YP_769637.1	Chromosome	1
RL4085	*gltA*	YP_769660.1	Chromosome	1
RL4184	*gltX*	YP_769759.1	Chromosome	1
RL4203	*talB*	YP_769778.1	Chromosome	0
RL4207	-	YP_769782.1	Chromosome	1
RL4265	*msrB*	YP_769840.1	Chromosome	0
RL4279	*clpB*	YP_769854.1	Chromosome	0
RL4281	*hemK*	YP_769856.1	Chromosome	1
RL4282	*prfA*	YP_769857.1	Chromosome	0
RL4298	*secA*	YP_769872.1	Chromosome	1
RL4323	*argH*	YP_769896.1	Chromosome	2

Locus tag	Gene symbol	Protein accession	Replicon	HGT index
RL4325	*lysA*	YP_769898.1	Chromosome	1
RL4352	*aroB*	YP_769923.1	Chromosome	1
RL4353	*aroK*	YP_769924.1	Chromosome	0
RL4412	*priA*	YP_769982.1	Chromosome	3
RL4436	*sucD*	YP_770006.1	Chromosome	2
RL4438	*sucC*	YP_770008.1	Chromosome	2
RL4493	*gpsA*	YP_770060.1	Chromosome	4
RL4494	*gcp*	YP_770061.1	Chromosome	2
RL4495	*hemC*	YP_770062.1	Chromosome	2
RL4506	*typA*	YP_770073.1	Chromosome	2
RL4507	*dcp*	YP_770074.1	Chromosome	0
RL4515	*argG*	YP_770080.1	Chromosome	3
RL4522	-	YP_770087.1	Chromosome	2
RL4550	*rimM*	YP_770115.1	Chromosome	1
RL4551	*trmD*	YP_770116.1	Chromosome	0
RL4552	*rplS*	YP_770117.1	Chromosome	2
RL4555	-	YP_770120.1	Chromosome	0
RL4563	-	YP_770128.1	Chromosome	1
RL4565	*glnB*	YP_770130.1	Chromosome	1
RL4630	*ispG*	YP_770194.1	Chromosome	0
RL4677	*rpmA*	YP_770239.1	Chromosome	4
RL4681	*obgE*	YP_770243.1	Chromosome	2
RL4682	*proB*	YP_770244.1	Chromosome	1
RL4683	*proA*	YP_770245.1	Chromosome	0
RL4689	-	YP_770251.1	Chromosome	1
RL4692	*ctpA*	YP_770254.1	Chromosome	1
RL4705	*leuD*	YP_770267.1	Chromosome	1
RL4707	*leuB*	YP_770269.1	Chromosome	1
RL4722	*purH*	YP_770284.1	Chromosome	2
RL4727	*acs*	YP_770289.1	Chromosome	4
RL4731	-	YP_770293.1	Chromosome	0
RL4732	*leuS*	YP_770294.1	Chromosome	0
RL4735	*parB*	YP_770297.1	Chromosome	1
RL4736	*parA*	YP_770298.1	Chromosome	0
RL4738	*gidA*	YP_770300.1	Chromosome	0
RL4739	*trmE*	YP_770301.1	Chromosome	1

Appendix II

Table II.I The genospecies B specific islands in pRL9 and pRL12. Locus tags and other informations of 29 genes of pRL9 and pRL12 held by all 12 members of genospecies B.

Locus tag	Gene name	Protein accession	Annotated function
pRL90041	groEL	YP_765335.1	Chaperonin GroEL (HSP60 family)
pRL90043	-	YP_765336.1	Multidrug resistance efflux pump
pRL90045	-	YP_765338.1	ABC-type multidrug transport system, permease component
pRL90119	-	YP_765411.1	DNA-binding transcriptional regulator, LysR family
pRL90120	-	YP_765412.1	Uncharacterized conserved protein YurZ, alkylhydroperoxidase/carboxymuconolactone decarboxylase family
pRL90121	-	YP_765413.1	Predicted ATPase
pRL90122	-	YP_765414.1	DNA-binding transcriptional regulator, LacI/PurR family
pRL90124	-	YP_765416.1	ABC-type sugar transport system, permease component
pRL90125	-	YP_765417.1	ABC-type glycerol-3-phosphate transport system, permease component
pRL90126	-	YP_765418.1	ABC-type glycerol-3-phosphate transport system, periplasmic component
pRL90255	-	YP_765541.1	Glycine cleavage system T protein (aminomethyltransferase)
pRL90256	-	YP_765542.1	5,10-methylenetetrahydrofolate reductase
pRL90257	-	YP_765543.1	DNA-binding transcriptional regulator, GntR family
pRL90259	-	YP_765545.1	Branched-chain amino acid ABC-type transport system, permease component
pRL120118	-	YP_764633.1	Predicted oxidoreductase
pRL120119	-	YP_764634.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL120120	-	YP_764635.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL120121	-	YP_764636.1	Dihydroorotase or related cyclic amidohydrolase
pRL120123	-	YP_764638.1	Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family
pRL120124	-	YP_764639.1	Nucleoside-diphosphate-sugar epimerase
pRL120125	-	YP_764640.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL120126	-	YP_764641.1	Dihydroorotase or related cyclic amidohydrolase
pRL120127	-	YP_764642.1	-
pRL120128	-	YP_764643.1	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component

Locus tag	Gene name	Protein accession	Annotated function
pRL120129	-	YP_764644.1	ABC-type dipeptide/oligopeptide/nickel transport system, permease component
pRL120130	-	YP_764645.1	ABC-type dipeptide/oligopeptide/nickel transport system, permease component
pRL120132	-	YP_764647.1	Transcriptional regulator GlxA family, contains an amidase domain and an AraC-type DNA-binding HTH domain
pRL120133	stbB	YP_764648.1	Predicted nucleic acid-binding protein, contains PIN domain
pRL120134	stbC	YP_764649.1	Plasmid stability protein

Table II.II Candidate specific islands of genospecies B in pRL7, pRL9, pRL10, pRL11 and pRL12. Locus tags and other informations of 50 genes of pRL7, pRL9, pRL10, pRL11 and pRL12 held by genospecies B.

Locus tag	Gene name	Protein accession	Annotated function
pRL70123	-	YP_770853.1	Plasmid stabilization system protein ParE
pRL70124	-	YP_770854.1	-
pRL90157	-	YP_765446.1	-
pRL90188	-	YP_765475.1	Cupin domain protein related to quercetin dioxygenase
pRL90189	-	YP_765476.1	Predicted dehydrogenase
pRL90190	-	YP_765477.1	ABC-type glycerol-3-phosphate transport system, permease component
pRL90192	-	YP_765479.1	ABC-type glycerol-3-phosphate transport system, periplasmic component
pRL90231	-	YP_765518.1	ABC-type transport system, periplasmic component
pRL90232	-	YP_765519.1	Arylsulfatase A or related enzyme
pRL90314	-	YP_765596.1	Tryptophan-rich sensory protein (mitochondrial benzodiazepine receptor homolog)
pRL90315	-	YP_765597.1	Predicted ATPase
pRL90317	-	YP_765599.1	Predicted enzyme related to lactoylglutathione lyase
pRL90318	ohr	YP_765600.1	Organic hydroperoxide reductase OsmC/OhrA
pRL100005	-	YP_770307.1	Uncharacterized protein, contains PIN domain
pRL100006	-	YP_770308.1	Uncharacterized protein
pRL100139	-	YP_770421.1	Site-specific DNA recombinase related to the DNA invertase Pin
pRL100316	-	YP_770592.1	-
pRL100468	-	YP_770743.1	-
pRL110057	-	YP_771090.1	-
pRL110132	-	YP_771166.1	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family
pRL110133	-	YP_771167.1	Predicted ATPase
pRL110134	-	YP_771168.1	NADPH:quinone reductase or related Zn-dependent oxidoreductase
pRL110135	-	YP_771169.1	Phenylpyruvate tautomerase PptA, 4-oxalocrotonate tautomerase family
pRL110137	-	YP_771171.1	Glyoxylase or a related metal-dependent hydrolase, beta-lactamase superfamily II
pRL110139	-	YP_771173.1	Predicted dehydrogenase
pRL110189	-	YP_771223.1	-
pRL110198	-	YP_771232.1	-
pRL110199	-	YP_771233.1	-

Locus tag	Gene name	Protein accession	Annotated function
pRL110301	-	YP_771334.1	-
pRL110302	-	YP_771335.1	-
pRL110338	-	YP_771370.1	-
pRL110494	-	YP_771528.1	-
pRL110497	-	YP_771531.1	-
pRL110585	-	YP_771619.1	-
pRL110607	-	YP_771641.1	Transposase
pRL120075	stbC	YP_764592.1	Plasmid stability protein
pRL120076	stbB	YP_764593.1	Predicted nucleic acid-binding protein, contains PIN domain
pRL120086	-	YP_764603.1	Phage shock protein A
pRL120089	-	YP_764609.1	-
pRL120092	-	YP_764620.1	Glutathionylspermidine synthase
pRL120103	-	YP_764592.1	-
pRL120168	-	YP_764680.1	DNA-binding transcriptional regulator, LysR family
pRL120428	-	YP_764935.1	DNA-binding transcriptional regulator, MurR/RpiR family, contains HTH and SIS domains
pRL120429	-	YP_764936.1	Asp/Glu/hydantoin racemase
pRL120430	-	YP_764937.1	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
pRL120433	-	YP_764940.1	ABC-type dipeptide/oligopeptide/nickel transport system, permease component
pRL120434	-	YP_764941.1	ABC-type transport system, periplasmic component
pRL120498	-	YP_765003.1	TRAP-type mannitol/chloroaromatic compound transport system, large permease component
pRL120499	-	YP_765004.1	TRAP-type mannitol/chloroaromatic compound transport system, small permease component
pRL120500	-	YP_765005.1	TRAP-type mannitol/chloroaromatic compound transport system, periplasmic component

Table II.III pRL12 genes absent in genospecies C. Locus tags and other informations of 9 genes of pRL12 absent in members of genospecies C.

Locus tag	Protein accession	Annotated function
pRL120756	YP_765259.1	ABC-type sugar transport system, periplasmic component, contains N-terminal xre family HTH domain
pRL120757	YP_765260.1	ABC-type sugar transport system, ATPase component
pRL120758	YP_765261.1	Ribose/xylose/arabinose/galactoside ABC-type transport system, permease component
pRL120759	YP_765262.1	DNA-binding transcriptional regulator LsrR, DeoR family
pRL120760	YP_765263.1	Predicted oxidoreductase (related to aryl-alcohol dehydrogenase)
pRL120761	YP_765264.1	Glycerol-3-phosphate dehydrogenase
pRL120762	YP_765265.1	Fructose-bisphosphate aldolase class Ia, DhnA family
pRL120763	YP_765266.1	Sugar (pentulose or hexulose) kinase
pRL120764	YP_765267.1	Choline dehydrogenase or related flavoprotein

Appendix III

Table III.I Carbon substrates in Biolog GN2 microplate and classified by the utilisation of *Rlv*3841. Substrates used in this study is in bold (🟪 polymers, 🟦 sugars/sugar derivatives, 🟥 carboxylic/dicarboxylic acid, ◼ amino acid/amino acid derivative and 🟩 miscellaneous intermediates of metabolism).

	Carbon substrates	
Not Used	**Used**	**Used Partially**
N-Acetyl-D-Galactosamine	**Adonitol**	Malonic Acid
N-Acetyl-D-Glucosamine	L-Arabinose	Acetic Acid
D-Galactose	D-Arabitol	
Gentiobiose	D-cellobiose	L-Proline
m-Inositol	**i-Erythritol**	
Lactulose	D-Fructose	
D-Melibiose	L-Fucose	
	α-D-Glucose	
Cis-Aconitic Acid	α-D-Lactose	
Citric Acid	Maltose	
Formic Acid	**D-Mannitol**	
D-Galacturonic Acid	D-Mannose	
D-Glucuronic Acid	**β-Methyl-D-Glucoside**	
α-Hydroxy Butyric Acid	D-Psicose	
p-Hydroxy Phenylacetic Acid	**D-Raffinose**	
Itaconic Acid	**L-Rhamnose**	
α-Keto Butyric Acid	D-Sorbitol	
α-Keto Valeric Acid	Sucrose	
Propionic Acid	D-Trehalose	
Quinic Acid	Turanose	
D-Saccharic Acid	**Xylitol**	
Sebacic Acid		
	D-Galactonic Acid Lactone	
L-Alanyl-glycine	**D-Gluconic Acid**	
L-Asparagine	**D-Glucosaminic Acid**	
L-Aspartic Acid	β-Hydroxy Butyric Acid	
L-Glutamic Acid	**γ-Hydroxy Butyric Acid**	
Glycyl-L-AsparticAcid	**α-Keto Glutaric Acid**	
Glycyl-L-Glutamic Acid	D,L-Lactic Acid	
Hydroxy-L-Proline	**Succinic Acid**	
L-Leucine		
L-Ornithine	**D-Alanine**	
L-Phenylalanine	**L-Alanine**	
D-Serine	L-Histidine	
L-Threonine	**L-Pyroglutamic Acid**	
D,L-Carnitine	**L-Serine**	
γ-Amino Butyric Acid		
	Urocanic Acid	
2,3-Butanediol	**Inosine**	
Phenyethylamine	**Thymidine**	
Putrescine	**Uridine**	
2-Aminoethanol	L-Alaninamide	
Glucuronamide	Methyl Pyruvate	
Glucose-1-Phosphate	Mono-Methyl-Succinate	
D,L-α-glycerol Phosphate	Glycerol	
Glucose-6-Phosphate	Bromo Succinic Acid	
	Succinamic Acid	
Glycogen		
α-Cyclodextrin	Dextrin	
Tween 40		
Tween 80		

Table III.II List of genes whose presence is significantly related to γ-Hydroxybutyric acid utilisation by using class association rule. Bold Locus tags were annotated as γ-Hydroxybutyric-acid-related genes.

Locus tag	Gene symbol	Protein accession	Annotated function
pRL70053	-	YP_770796.1	transmembrane protein
pRL70068	-	YP_770805.1	transposase-like protein
pRL70102	-	YP_770836.1	hypothetical protein
pRL70176	-	YP_770893.1	transposase-related protein
pRL80096	-	YP_770989.1	IS30 family transposase
pRL100087	*acdS*	YP_770380.1	1-aminocyclopropane-1-carboxylate deaminase
pRL100093	-	YP_770383.1	hypothetical protein
pRL100103	-	YP_770388.1	alcohol dehydrogenase
pRL100104	-	YP_770389.1	hypothetical protein
pRL100105	-	YP_770390.1	polyhydroxyalkanoate synthase subunit C
pRL100106	-	YP_770391.1	hypothetical protein
pRL100107	-	YP_770392.1	hypothetical protein
pRL100119	-	YP_770400.1	propionate CoA-transferase
pRL100120	-	YP_770401.1	hypothetical protein
pRL100121	*acsA*	YP_770402.1	acetyl-coenzyme A synthetase
pRL100124	-	YP_770405.1	transposase family protein
pRL100133	-	YP_770415.1	IclR family transcriptional regulatory protein
pRL100134	*gabD*	YP_770416.1	succinate-semialdehyde dehydrogenase
pRL100135	-	YP_770417.1	1,3-propanediol dehydrogenase
pRL100136	-	YP_770418.1	beta lactamase/homoserine lactonase
pRL100137	*metX*	YP_770419.1	homoserine O-acetyltransferase
pRL100138	-	YP_770420.1	MerR family transcriptional regulator
pRL100163	-	YP_770442.1	hypothetical protein
pRL100170	*rhiB*	YP_770449.1	rhizosphere induced protein RhiB
pRL100171	*rhiC*	YP_770450.1	hypothetical protein
pRL100172	*rhiR*	YP_770451.1	transcriptional regulator
pRL100198	*fixC*	YP_770476.1	nitrogen fixation FixC protein
pRL100201	-	YP_770479.1	hypothetical protein
pRL100202	-	YP_770480.1	hypothetical protein
pRL110291	-	YP_771324.1	hypothetical protein
pRL110292	*hycG*	YP_771325.1	putative formate hydrogenlyase subunit 7
pRL110293	*hycE*	YP_771326.1	putative formate hydrogenlyase subunit 5
pRL110294	*hyfF*	YP_771327.1	hydrogenase 4 subunit F
pRL110295	*hyfE*	YP_771328.1	putative hydrogenase-4 component E
pRL110296	*hycD*	YP_771329.1	putative hydrogenase protein
pRL110297	*hyfB*	YP_771330.1	hydrogenase 4 subunit B

Locus tag	Gene symbol	Protein accession	Annotated function
pRL120179	-	YP_764691.1	indolepyruvate ferredoxin oxidoreductase
pRL120180	-	YP_764692.1	oxidoreductase
pRL120181	-	YP_764693.1	GntR family transcriptional regulator
pRL120182	-	YP_764694.1	alcohol dehydrogenase
pRL120333	-	YP_764843.1	ABC transporter substrate-binding protein
pRL120334	-	YP_764844.1	hydrolase
pRL120335	-	YP_764845.1	acylase
pRL120339	-	YP_764849.1	ABC transporter ATP-binding protein
pRL120347	-	YP_764857.1	LysR family transcriptional regulator
pRL120450	*cpO*	YP_764956.1	chloroperoxidase
pRL120453	-	YP_764959.1	transcriptional regulator
pRL120456	-	YP_764962.1	dioxygenase
pRL120457	-	YP_764963.1	LysR family transcriptional regulator
pRL120459	-	YP_764965.1	hypothetical protein
pRL120528	-	YP_765033.1	dihydrodipicolinate synthase
pRL120529	-	YP_765034.1	aldehyde dehydrogenase
pRL120530	-	YP_765035.1	dehydrogenase/oxidoreductase
pRL120531	-	YP_765036.1	ABC transporter substrate-binding protein
pRL120532	-	YP_765037.1	ABC transporter ATP-binding protein
pRL120533	-	YP_765038.1	ABC transporter permease
pRL120534	-	YP_765039.1	ABC transporter permease

Table III.III List of genes whose presence is significantly related to D-Galactonic acid lactone utilisation by using class association rule.

Locus tag	Protein accession	Annotated function
pRL070013	YP_770758.1	hypothetical protein
pRL090039	YP_765334.1	hypothetical protein
pRL090116	YP_765408.1	putative ribitol 2-dehydrogenase
pRL090117	YP_765409.1	putative D-ribulokinase/ribitol kinase
pRL090280	YP_765566.1	hypothetical protein
pRL090281	YP_765567.1	hypothetical protein
pRL100173	YP_770452.1	outer-membrane immunogenic protein
pRL100255	YP_770534.1	hypothetical protein
pRL100311	YP_770586.1	transmembrane polysaccharide synthesis protein
pRL100312	YP_770587.1	hydrolase
pRL100313	YP_770588.1	hypothetical protein
pRL100313A	YP_770589.1	hypothetical protein
pRL100314	YP_770590.1	hypothetical protein
pRL100314A	YP_770591.1	hypothetical protein
pRL110045	YP_771078.1	hypothetical protein
pRL110105	YP_771137.1	LysR family transcriptional regulator
pRL110521	YP_771555.1	hypothetical protein
pRL110525	YP_771559.1	hypothetical protein
pRL120010	YP_764527.1	hypothetical protein
pRL120011	YP_764528.1	substrate-binding periplasmic protein precursor
pRL120012	YP_764529.1	ABC transporter permease
pRL120013	YP_764530.1	ABC transporter permease
pRL120015	YP_764532.1	glycerophosphoryl diester phosphodiesterase
pRL120016	YP_764533.1	DeoR family transcriptional regulator
pRL120017	YP_764534.1	hydrolase
pRL120084	YP_764601.1	hypothetical protein
pRL120085	YP_764602.1	hypothetical protein
pRL120721	YP_765224.1	hypothetical protein

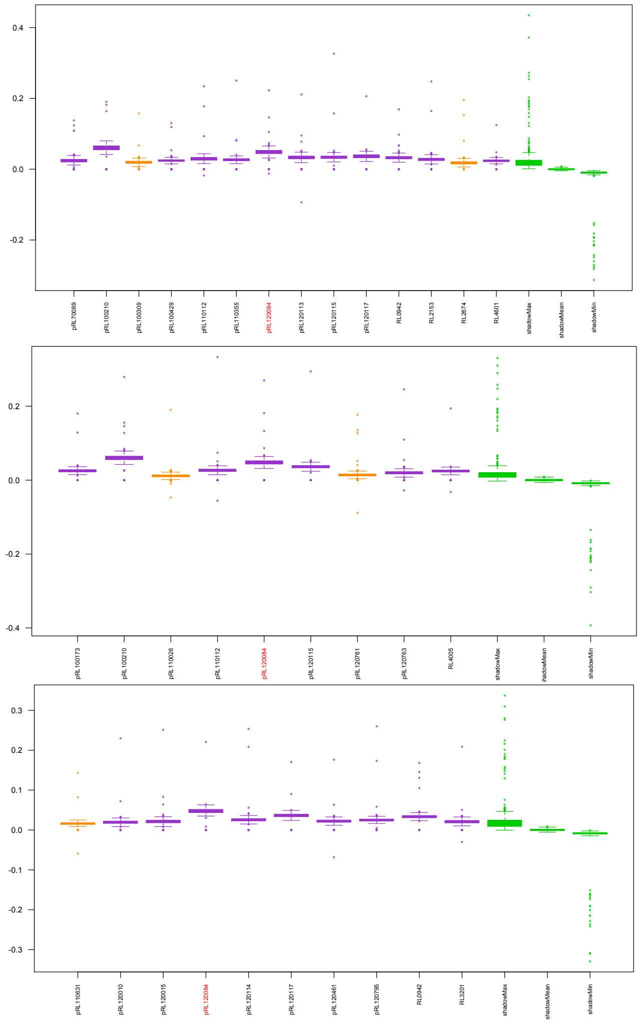


Figure III.I Z-score profiles of genes involving in the utilisation of α-keto glutaric acid. Three independent iterations of the gene selection algorithm are illustrated. Vertical axis represents Z-score. Columns represent genes coloured by the confirmation of their involvement in the three iterations (🟥 genes confirmed as relevant genes by 3 iterations and ⬛ genes confirmed relevant genes by 1-2 iterations). Each plot was coloured by Z-score by comparing to the maximum Z-score among shadow attributes (MZSA) (🟪 identified as genes involved in the substrate utilisation, 🟧 identified as tentative genes involved in the substrate utilisation, and 🟩 Z-score of shadow attributes). In box-and-whisker plot, the horizontal center of each box represents median, boxes represent 25th to 75th percentiles, and whiskers represent 10th and 90th percentiles. Dots are located out of the box represent outliers.

Table III.IV Contingency table of candidate genes of the utilisation of α-keto glutaric acid identified by random forest.

	pRL120084	
	Present	**Absent**
Utilising	9	43
Not utilising	11	9

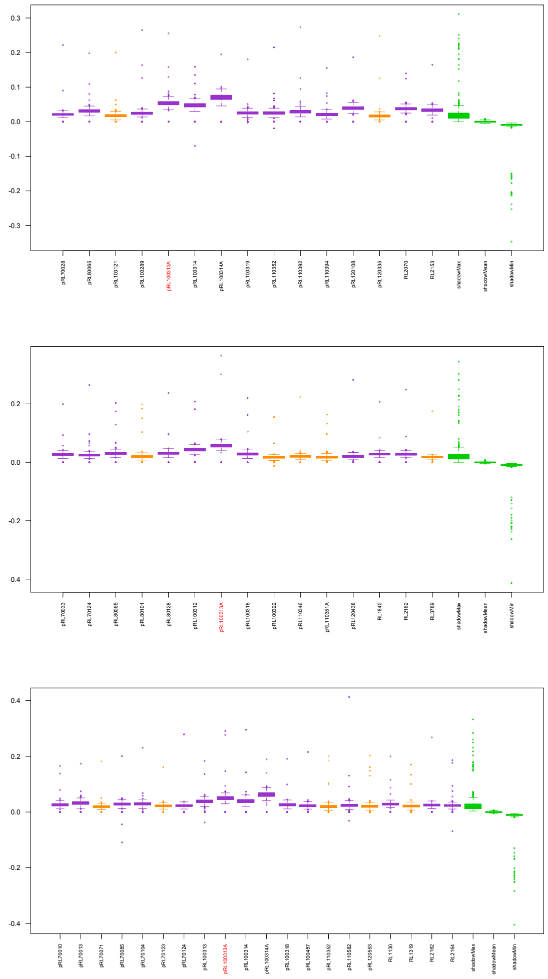


Figure III.II Z-score profiles of genes involving in the utilisation of D-glucosaminic acid. See legend to Figure III.I for details.

Table III.V Contingency table of candidate genes of the utilisation of D-glucosaminic acid identified by random forest.

	pRL120084	
	Present	**Absent**
Utilising	3	48
Not utilising	8	13

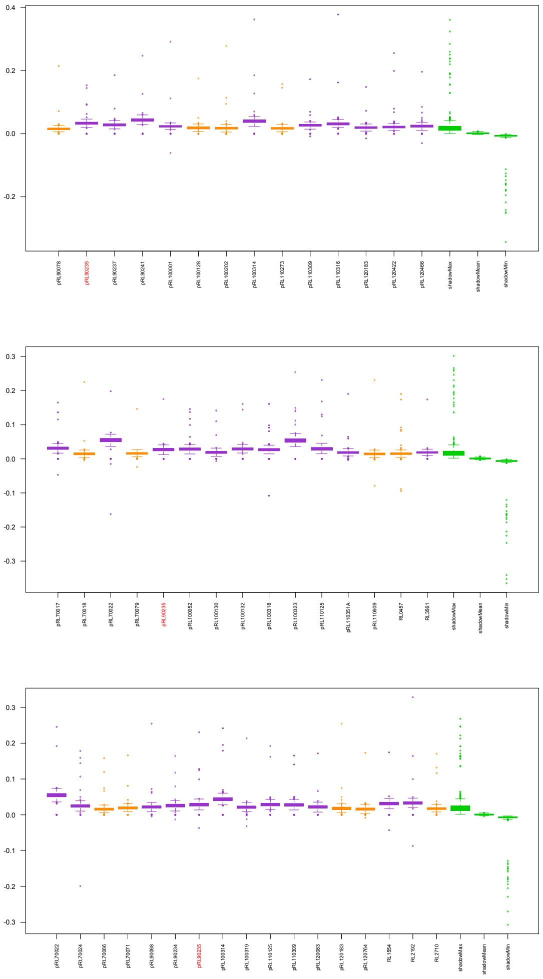


Figure III.III Z-score profiles of genes involving in the utilisation of D-raffinose. See legend to Figure III.I for details.

Table III.VI Contingency table of candidate genes of the utilisation of D-raffinose identified by random forest.

	pRL90235	
	Present	**Absent**
Utilising	24	39
Not utilising	8	1

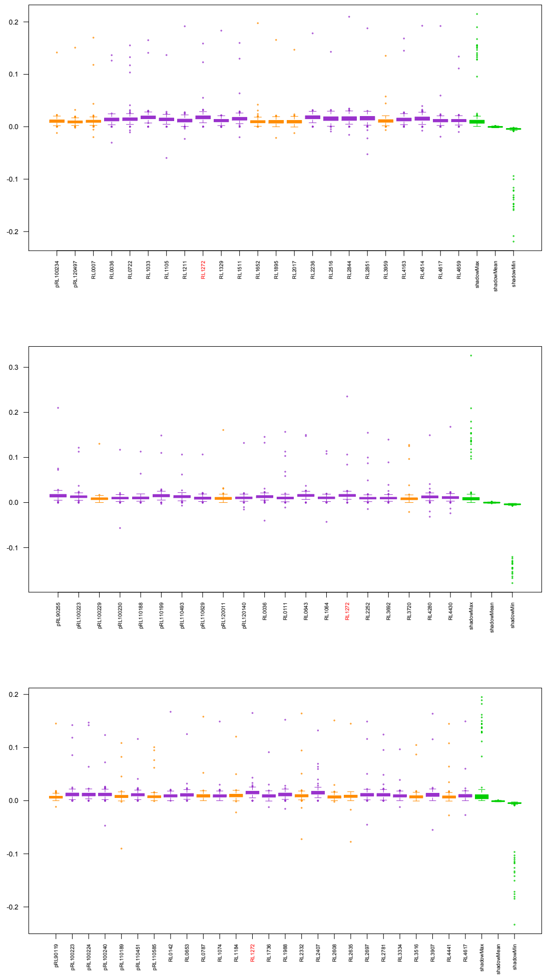


Figure III.IV Z-score profiles of genes involving in the utilisation of succinic acid. See legend to Figure III.I for details.

Table III.VII Contingency table of candidate genes of the utilisation of Succinic Acid identified by random forest.

	RL1272	
	Present	**Absent**
Utilising	71	0
Not utilising	0	1

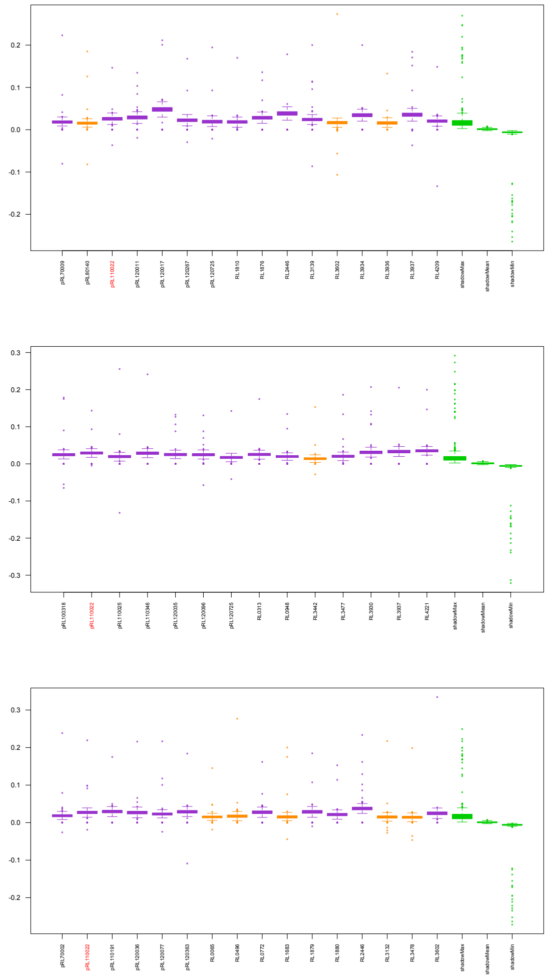


Figure III.V Z-score profiles of genes involving in the utilisation of Thymidine. See legend to Figure III.I for details.

Table III.VIII Contingency table of candidate genes of the utilisation of Thymidine identified by random forest.

	pRL110022	
	Present	**Absent**
Utilising	56	10
Not utilising	2	4

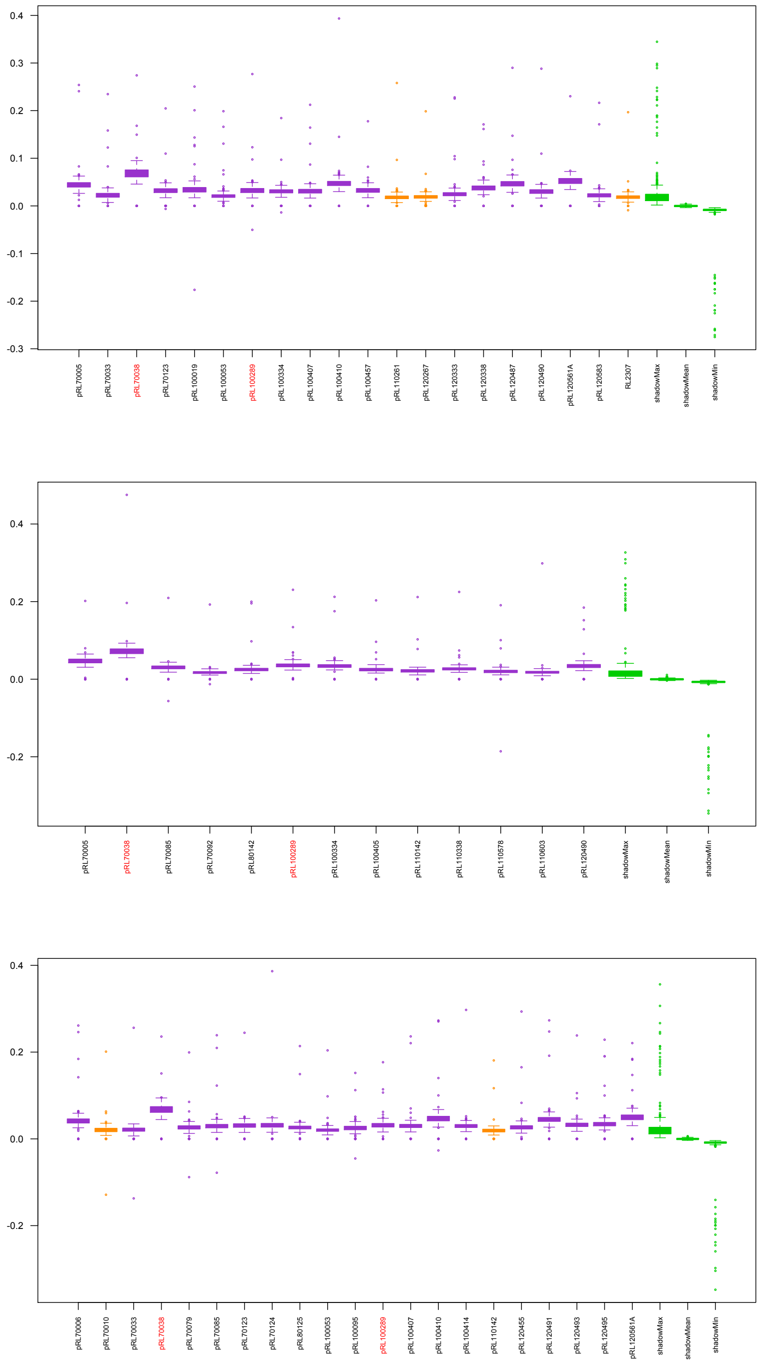


Figure III.VI Z-score profiles of genes involving in the utilisation of D-alanine. See legend to Figure III.I for details.

Table III.IX Contingency table of candidate genes of the utilisation of D-alanine identified by random forest.

	pRL70038	
	Present	**Absent**
Utilising	1	54
Not utilising	7	10

	pRL100313	
	Present	**Absent**
Utilising	54	1
Not utilising	12	5

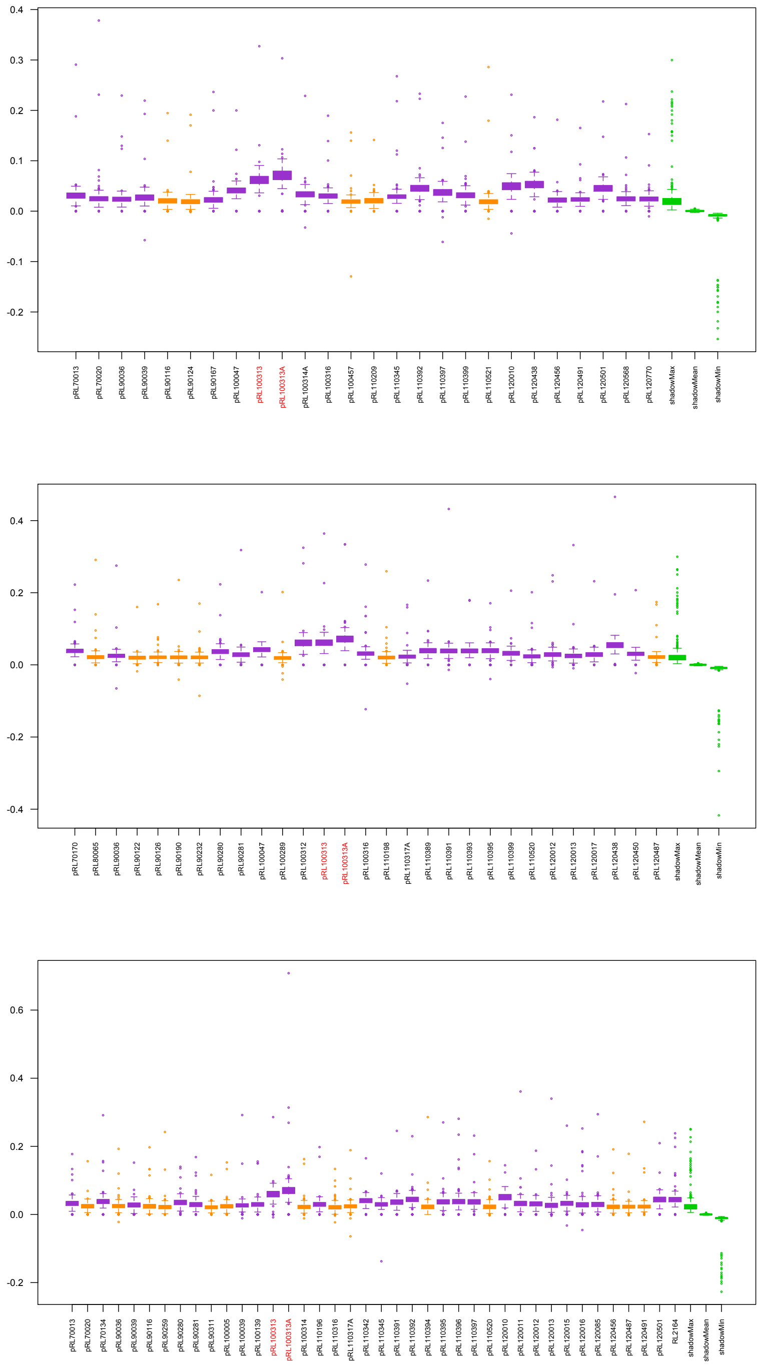


Figure III.VII Z-score profiles of genes involving in the utilisation of D-gluconic acid. See legend to Figure III.I for details.

Table III.X Contingency table of candidate genes of the utilisation of D-gluconic acid identified by random forest.

	pRL100313	
	Present	**Absent**
Utilising	5	59
Not utilising	6	2

	pRL100313A	
	Present	**Absent**
Utilising	4	60
Not utilising	6	2

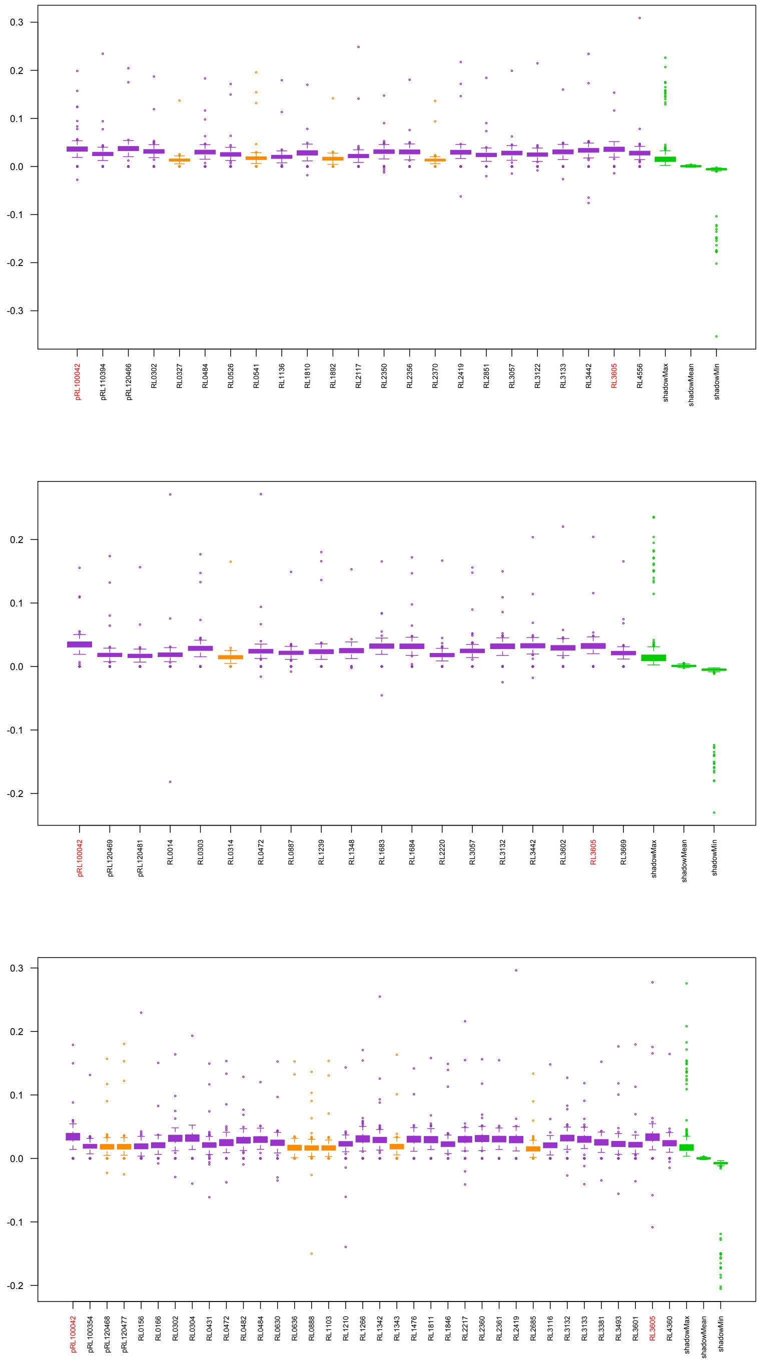


Figure III.VIII Z-score profiles of genes involving in the utilisation of uridine. See legend to Figure III.I for details.

Table III.XI Contingency table of candidate genes of the utilisation of uridine identified by random forest.

	RL3605	
	Present	**Absent**
Utilising	8	61
Not utilising	1	2

	pRL100042	
	Present	**Absent**
Utilising	7	62
Not utilising	3	0

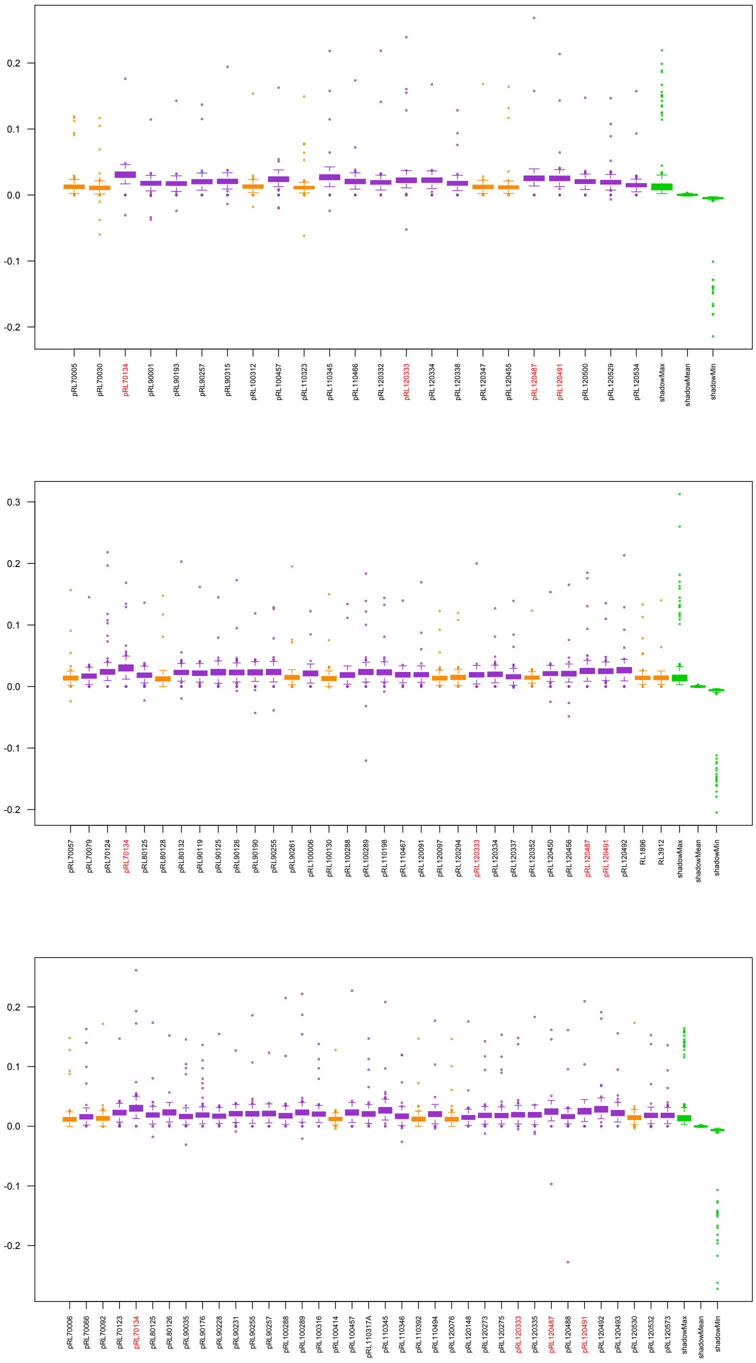


Figure III.IX Z-score profiles of genes involving in the utilisation of D-mannitol. See legend to Figure III.I for details.

Table III.XII Contingency table of candidate genes of the utilisation of D-Mannitol identified by random forest.

	pRL70134	
	Present	**Absent**
Utilising	1	69
Not utilising	2	0

	pRL120333	
	Present	**Absent**
Utilising	65	5
Not utilising	0	2

	pRL120487	
	Present	**Absent**
Utilising	67	3
Not utilising	0	2

	pRL120491	
	Present	**Absent**
Utilising	67	3
Not utilising	0	2

References

Achtman, M. & Wagner, M. 2008. Microbial diversity and the genetic nature of microbial species. *Nat Rev Micro,* 6**,** 431-440.

Agrawal, R., Imieliński, T. & Swami, A. Mining association rules between sets of items in large databases. ACM SIGMOD Record, 1993. ACM, 207-216.

Agrawal, R. & Srikant, R. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, VLDB, 1994. 487-499.

Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., Dejonge, B. L., Carmel, G., Tummino, P. J., Caruso, A., Uria-Nickelsen, M., Mills, D. M., Ives, C., Gibson, R., Merberg, D., Mills, S. D., Jiang, Q., Taylor, D. E., Vovis, G. F. & Trust, T. J. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature,* 397**,** 176-180.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

Angert, E. R. 2005. Alternatives to binary fission in bacteria. *Nat Rev Micro,* 3**,** 214-224.

Aoki, K., Ogata, Y. & Shibata, D. 2007. Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology. *Plant and Cell Physiology,* 48**,** 381-390.

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**,** 1722-1734.

Barabasi, A. L. & Oltvai, Z. N. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet,* 5**,** 101-113.

Barberan, A., Bates, S. T., Casamayor, E. O. & Fierer, N. 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. *Isme j,* 6**,** 343-351.

Bayjanov, J. R., Molenaar, D., Tzeneva, V., Siezen, R. J. & Van Hijum, S. A. 2012. PhenoLink-a web-tool for linking phenotype to~ omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. *BMC genomics,* 13, 170.

Bayjanov, J. R., Starrenburg, M. J., Van Der Sijde, M. R., Siezen, R. J. & Van Hijum, S. A. 2013. Genotype-phenotype matching analysis of 38 *Lactococcus lactis* strains using random forest methods. *BMC microbiology,* 13, 68.

Beauregard-Racine, J., Bicep, C., Schliep, K., Lopez, P., Lapointe, F. J. & Bapteste, E. 2011. Of woods and webs: possible alternatives to the tree of life for studying genomic fluidity in *E. coli*. *Biol Direct,* 6, 39; discussion 39.

Beiko, R. G., Harlow, T. J. & Ragan, M. A. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America,* 102, 14332-14337.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bentley, S. D. & Parkhill, J. 2004. Comparative genomic structure of prokaryotes. *Annu Rev Genet,* 38, 771-792.

Blaby-Haas, C. E. & De Crécy-Lagard, V. 2011. Mining high-throughput experimental data to link gene and function. *Trends in Biotechnology,* 29, 174-182.

Black, M., Moolhuijzen, P., Chapman, B., Barrero, R., Howieson, J., Hungria, M. & Bellgard, M. 2012. The genetics of symbiotic nitrogen fixation: comparative genomics of 14 rhizobia strains by resolution of protein clusters. *Genes,* 3, 138-166.

Bochner, B. R. 1989. Sleuthing out bacterial identities. *Nature,* 339, 157-158.

Boto, L. 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society of London B: Biological Sciences,* 277, 819-827.

Breiman, L. 1996. Bagging Predictors. *Machine Learning,* 24, 123-140.

Breiman, L. 2001. Random Forests. *Machine Learning,* 45, 5-32.

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.

Carlier, A., Chevrot, R., Dessaux, Y. & Faure, D. 2004. The assimilation of gamma-butyrolactone in Agrobacterium tumefaciens C58 interferes with the accumulation of the N-acyl-homoserine lactone signal. *Mol Plant Microbe Interact,* 17**,** 951-957.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. & Pascual-Montano, A. 2006. Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics,* 7**,** 54.

Ceapa, C., Lambert, J., Van Limpt, K., Wels, M., Smokvina, T., Knol, J. & Kleerebezem, M. 2015. Correlation of *Lactobacillus rhamnosus* Genotypes and Carbohydrate Utilization Signatures Determined by Phenotype Profiling. *Applied and Environmental Microbiology,* 81**,** 5458-5470.

Cevallos, M. A., Cervantes-Rivera, R. & Gutiérrez-Ríos, R. M. 2008. The repABC plasmid family. *Plasmid,* 60**,** 19-37.

Chai, Y., Tsai, C. S., Cho, H. & Winans, S. C. 2007. Reconstitution of the biochemical activities of the AttJ repressor and the AttK, AttL, and AttM catabolic enzymes of *Agrobacterium tumefaciens. J Bacteriol,* 189**,** 3674-3679.

Chawla, N. V., Japkowicz, N. & Kotcz, A. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter,* 6**,** 1-6.

Chevrot, R., Rosen, R., Haudecoeur, E., Cirou, A., Shelp, B. J., Ron, E. & Faure, D. 2006. GABA controls the level of quorum-sensing signal in Agrobacterium tumefaciens. *Proc Natl Acad Sci U S A,* 103**,** 7460-7464.

Choi, I. G. & Kim, S. H. 2007. Global extent of horizontal gene transfer. *Proc Natl Acad Sci USA,* 104.

Cohen, O., Ashkenazy, H., Burstein, D. & Pupko, T. 2012. Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics,* 28**,** i389-i394.

Cokus, S., Mizutani, S. & Pellegrini, M. 2007. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics,* 8**,** 1-12.

Csardi, G. & Nepusz, T. 2006. The igraph Software Package for Complex Network Research. *InterJournal,* Complex Systems.

Cui, Q. 2010. A Network of Cancer Genes with Co-Occurring and Anti-Co-Occurring Mutations. *PLoS ONE,* 5, e13180.

Dash, M. & Liu, H. 1997. Feature selection for classification. *Intelligent Data Analysis,* 1, 131-156.

Dhillon, B. K., Laird, M. R., Shay, J. A., Winsor, G. L., Lo, R., Nizam, F., Pereira, S. K., Waglechner, N., Mcarthur, A. G., Langille, M. G. I. & Brinkman, F. S. L. 2015. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Research,* 43, W104-W108.

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. & Maiden, M. C. J. 2010. Impact of recombination on bacterial evolution. *Trends in Microbiology,* 18, 315-322.

Didelot, X., Meric, G., Falush, D. & Darling, A. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics,* 13, 256.

Dietterich, T. 2000. Ensemble Methods in Machine Learning. *Multiple Classifier Systems.* Springer Berlin Heidelberg.

Ding, C. H. Q., He, X. & Zha, H. 2001. A spectral method to separate disconnected and nearly-disconnected web graph components. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.* San Francisco, California: ACM.

Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N., Angiuoli, S., Oggioni, M., Dunning Hotopp, J., Hu, F., Riley, D., Covacci, A., Mitchell, T., Bentley, S., Kilian, M., Ehrlich, G., Rappuoli, R., Moxon, E. R. & Masignani, V. 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology,* 11, R107.

Duangsonk, K., Gal, D., Mayo, M., Hart, C. A., Currie, B. J. & Winstanley, C. 2006. Use of a variable amplicon typing scheme reveals considerable variation in the accessory genomes of isolates of *Burkholderia pseudomallei*. *J Clin Microbiol,* 44, 1323-1334.

Dudoit, S., Fridlyand, J. & Speed, T. P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association,* 97, 77-87.

Dunn, M. F. 1998. Tricarboxylic acid cycle and anaplerotic enzymes in rhizobia. *FEMS Microbiology Reviews,* 22, 105-123.

Dziewit, L. & Bartosik, D. 2014. Plasmids of psychrophilic and psychrotolerant bacteria and their role in adaptation to cold environments. *Frontiers in Microbiology,* 5.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32, 1792-1797.

Efron, B. & Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician,* 37, 36-48.

Ellegaard, K. M., Klasson, L., Näslund, K., Bourtzis, K. & Andersson, S. G. E. 2013. Comparative Genomics of *Wolbachia* and the Bacterial Species Concept. *PLoS Genet,* 9, e1003381.

Elo, L. L., Järvenpää, H., Orešič, M., Lahesmaa, R. & Aittokallio, T. 2007. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. *Bioinformatics,* 23, 2096-2103.

Epstein, B., Sadowsky, M. J. & Tiffin, P. 2014. Selection on Horizontally Transferred and Duplicated Genes in Sinorhizobium (Ensifer), the Root-Nodule Symbionts of Medicago. *Genome Biology and Evolution,* 6, 1199-1209.

Espinosa-Urgel, M. & Kolter, R. 1998. Escherichia coli genes expressed preferentially in an aquatic environment. *Mol Microbiol,* 28, 325-332.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J. & Huttenhower, C. 2012. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput Biol,* 8, e1002606.

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.

Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. 2005. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology,* 3**,** 722-732.

Fukushima, A., Kusano, M., Redestig, H., Arita, M. & Saito, K. 2011. Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Syst Biol,* 5**,** 1.

Furuya, E. Y. & Lowy, F. D. 2006. Antimicrobial-resistant bacteria in the community setting. *Nat Rev Micro,* 4**,** 36-45.

Galardini, M., Pini, F., Bazzicalupo, M., Biondi, E. G. & Mengoni, A. 2013. Replicon-Dependent Bacterial Genome Evolution: The Case of Sinorhizobium meliloti. *Genome Biology and Evolution,* 5**,** 542-558.

Gamazon, E. R., Huang, R. S., Dolan, M. E., Cox, N. J. & Im, H. K. 2012. Integrative genomics: quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. *Frontiers in genetics,* 3.

Goh, C. S., Gianoulis, T. A., Liu, Y., Li, J., Paccanaro, A., Lussier, Y. A. & Gerstein, M. 2006. Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics,* 7**,** 257.

Goldenfeld, N. & Woese, C. 2007. Biology's next revolution. *Nature,* 445**,** 369-369.

Grant, B. J., Rodrigues, A. P. C., Elsawy, K. M., Mccammon, J. A. & Caves, L. S. D. 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics,* 22**,** 2695-2696.

Gray, M. W. 1999. Evolution of organellar genomes. *Curr Opin Genet Dev,* 9**,** 678-687.

Griffith, F. 1928. The Significance of Pneumococcal Types. *J Hyg (Lond),* 27**,** 113-159.

Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. 2000. Population genetics.

Guo, J., Wang, Q., Wang, X., Wang, F., Yao, J. & Zhu, H. 2015. Horizontal gene transfer in an acid mine drainage microbial community. *BMC Genomics,* 16**,** 1-11.

Gupta, A., Maranas, C. D. & Albert, R. 2006. Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites. *Bioinformatics,* 22**,** 209-214.

Guyon, I. & Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research,* 3**,** 1157-1182.

Harrison, P. W., Lower, R. P. J., Kim, N. K. D. & Young, J. P. W. 2010. Introducing the bacterial chromid : not a chromosome, not a plasmid. *Trends in microbiology,* 18**,** 141-148.

Heuer, H. & Smalla, K. 2012. Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiology Reviews,* 36**,** 1083-1104.

Hogg, J., Hu, F., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C. & Ehrlich, G. 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 Biology,* 8**,** R103.

Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res,* 10**,** 1204-1210.

Jaccard, P. 1912. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist,* 11**,** 37-50.

Jain, R., Rivera, M. C. & Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences,* 96**,** 3801-3806.

Kanhere, A. & Vingron, M. 2009. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evolutionary Biology,* 9**,** 1-13.

Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology,* 9**,** 335-343.

Kell, D. B. 2004. Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol,* 7**,** 296-307.

Kensche, P. R., Van Noort, V., Dutilh, B. E. & Huynen, M. A. 2008. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J R Soc Interface,* 5, 151-170.

Kim, P.-J. & Price, N. D. 2011. Genetic Co-Occurrence Network across Sequenced Microbes. *PLoS Comput Biol,* 7, e1002340.

Knobbe, A. J. & Adriaans, P. W. Analysing Binary Associations. KDD, 1996. 311.

Kohavi, R. & John, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence,* 97, 273-324.

Kumar, N. 2013. *Are bacterial species really ecotypes?* PhD, University of York.

Kumar, N., Lad, G., Giuntini, E., Kaye, M. E., Udomwong, P., Shamsani, N. J., Young, J. P. W. & Bailly, X. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biol,* 5.

Kursa, M. B., Jankowski, A. & Rudnicki, W. R. 2010a. Boruta A System for Feature Selection. *Fundamenta Informaticae,* 101, 271-285.

Kursa, M. B. & Rudnicki, W. R. 2010b. Feature selection with the Boruta package. Journal.

Lad, G. 2013. *Phenomic and genomic diversity of a bacterial species in a local population.* Doctor of Philosophy, University of York.

Lan, R. & Reeves, P. R. 1996. Gene transfer is a major factor in bacterial evolution. *Molecular Biology and Evolution,* 13, 47-55.

Langfelder, P. & Horvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9, 559.

Lapointe, F. J., Lopez, P., Boucher, Y., Koenig, J. & Bapteste, E. 2010. Clanistics: a multi-level perspective for harvesting unrooted gene trees. *Trends Microbiol,* 18, 341-347.

Lassalle, F., Campillo, T., Vial, L., Baude, J., Costechareyre, D., Chapulliot, D., Shams, M., Abrouk, D., Lavire, C., Oger-Desfeux, C., Hommais, F., Guéguen, L., Daubin, V., Muller, D. & Nesme, X. 2011. Genomic Species Are Ecological Species as Revealed by Comparative Genomics in *Agrobacterium tumefaciens*. *Genome Biology and Evolution,* 3, 762-781.

Lawrence, J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr Opin Genet Dev,* 9, 642-648.

Lawrence, J. & Retchless, A. 2009. The Interplay of Homologous Recombination and Horizontal Gene Transfer in Bacterial Speciation. *In:* GOGARTEN, M., et al. (eds.) *Horizontal Gene Transfer.* Humana Press.

Lawrence, J. G. & Ochman, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol,* 44**,** 383-397.

Lawrence, J. G. & Ochman, H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A,* 95**,** 9413-9417.

Lefébure, T., Pavinski Bitar, P. D., Suzuki, H. & Stanhope, M. J. 2010. Evolutionary Dynamics of Complete *Campylobacter* Pan-Genomes and the Bacterial Species Concept. *Genome Biology and Evolution,* 2**,** 646-655.

Lefébure, T. & Stanhope, M. J. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol,* 8**,** R71.

Li, X., Zhou, X., Peng, Y., Liu, B., Zhang, R., Hu, J., Yu, J., Jia, C. & Sun, C. 2014. Network based integrated analysis of phenotype-genotype data for prioritization of candidate symptom genes. *BioMed research international,* 2014.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news,* 2**,** 18-22.

Liu, B. 2007. *Web data mining: exploring hyperlinks, contents, and usage data*, Springer Science & Business Media.

Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular biology and evolution,* 11**,** 605-612.

Lodwig, E. M., Hosie, A. H. F., Bourdes, A., Findlay, K., Allaway, D., Karunakaran, R., Downie, J. A. & Poole, P. S. 2003. Amino-acid cycling drives nitrogen fixation in the legume-Rhizobium symbiosis. *Nature,* 422**,** 722-726.

Long, S. R. 1989. Rhizobium genetics. *Annu Rev Genet,* 23**,** 483-506.

Long, S. R. 2001. Genes and Signals in the Rhizobium-Legume Symbiosis. *Plant Physiology,* 125**,** 69-72.

Mallet, L., Becq, J. & Deschavanne, P. 2010. Whole genome evaluation of horizontal transfers in the pathogenic fungus Aspergillus fumigatus. *BMC Genomics,* 11**,** 171.

Mauchline, T. H., Hayat, R., Roberts, R., Powers, S. J. & Hirsch, P. R. 2014. Assessment of core and accessory genetic variation in Rhizobium leguminosarum symbiovar trifolii strains from diverse locations and host plants using PCR-based methods. *Lett Appl Microbiol,* 59**,** 238-246.

Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. 2005. The microbial pan-genome. *Curr Opin Genet Dev,* 15**,** 589-594.

Menon, A., Narasimhan, H., Agarwal, S. & Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. Proceedings of The 30th International Conference on Machine Learning, 2013. 603-611.

Méric, G., Yahara, K., Mageiros, L., Pascoe, B., Maiden, M. C. J., Jolley, K. A. & Sheppard, S. K. 2014. A Reference Pan-Genome Approach to Comparative Bacterial Genomics: Identification of Novel Epidemiological Markers in Pathogenic *Campylobacter*. *PLoS ONE,* 9**,** e92798.

Meyer, P. E. 2009. infotheo: Information-theoretic measures. *R package. Version,* 1.

Miller, S. H., Elliot, R. M., Sullivan, J. T. & Ronson, C. W. 2007. Host-specific regulation of symbiotic nitrogen fixation in *Rhizobium leguminosarum biovar trifolii*. *Microbiology,* 153**,** 3184-3195.

Mutch, L. A. & Young, J. P. 2004. Diversity and specificity of *Rhizobium leguminosarum biovar viciae* on wild and cultivated legumes. *Mol Ecol,* 13**,** 2435-2444.

Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet,* 36**,** 760-766.

Nishida, H. 2012. Evolution of genome base composition and genome size in bacteria. *Frontiers in Microbiology,* 3**,** 420.

Norman, A., Hansen, L. H. & Sorensen, S. J. 2009. Conjugative plasmids: vessels of the communal gene pool. *Philos Trans R Soc Lond B Biol Sci,* 364**,** 2275-2289.

Ochman, H. 2002. Bacterial Evolution: Chromosome Arithmetic and Geometry. *Current Biology,* 12**,** R427-R428.

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

Ooi, W. F., Ong, C., Nandi, T., Kreisberg, J. F., Chua, H. H., Sun, G., Chen, Y., Mueller, C., Conejero, L. & Eshaghi, M. 2013. The condition-dependent transcriptional landscape of *Burkholderia pseudomallei*. *PLoS genetics,* 9**,** e1003795.

Oti-Boateng, C. & Silbury, J. H. 1993. The Effects of Exogenous Amino Acid on Acetylene Reduction Activity of *Vicia faba* L. cv. Fiord. *Annals of Botany,* 71**,** 71-74.

Overbeek, R., Fonstein, M., D'souza, M., Pusch, G. D. & Maltsev, N. 1999. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences,* 96**,** 2896-2901.

Ozaki, S., Ogata, Y., Suda, K., Kurabayashi, A., Suzuki, T., Yamamoto, N., Iijima, Y., Tsugane, T., Fujii, T. & Konishi, C. 2010. Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module. *DNA research,* 17**,** 105-116.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics,* 20**,** 289-290.

Paulsson, J. 2002. Multileveled selection on plasmid replication. *Genetics,* 161**,** 1373-1384.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R. & Bagos, P. G. 2011. Using graph theory to analyze biological networks. *BioData Min,* 4**,** 10.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A,* 96**,** 4285-4288.

Perkins, A. & Langston, M. 2009. Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinformatics,* 10**,** S4.

Pommerenke, C., Musken, M., Becker, T., Dotsch, A., Klawonn, F. & Haussler, S. 2010. Global genotype-phenotype correlations in *Pseudomonas aeruginosa*. *PLoS Pathog,* 6**,** e1001074.

Pons, P. & Latapy, M. 2005. Computing Communities in Large Networks Using Random Walks. *In:* YOLUM, P., et al. (eds.) *Computer and Information Sciences - ISCIS 2005.* Springer Berlin Heidelberg.

Prell, J., Bourdes, A., Karunakaran, R., Lopez-Gomez, M. & Poole, P. 2009. Pathway of γ-Aminobutyrate metabolism in *Rhizobium leguminosarum* 3841 and Its Role in Symbiosis. *Journal of bacteriology,* 191**,** 2177-2186.

Price, N. D., Reed, J. L. & Palsson, B. O. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol,* 2**,** 886-897.

Prosser, J. I., Bohannan, B. J. M., 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. W. 2007. The role of ecological theory in microbial ecology. *Nat Rev Micro,* 5**,** 384-392.

Putonti, C., Luo, Y., Katili, C., Chumakov, S., Fox, G. E., Graur, D. & Fofanov, Y. 2006. A Computational Tool for the Genomic Identification of Regions of Unusual Compositional Properties and Its Utilization in the Detection of Horizontally Transferred Sequences. *Molecular Biology and Evolution,* 23**,** 1863-1868.

Ragan, M. A. 2001. Detection of lateral gene transfer among microbial genomes. *Curr Opin Genet Dev,* 11**,** 620-626.

Reiner, A., Yekutieli, D. & Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics,* 19**,** 368-375.

Riley, M. A. & Lizotte-Waniewski, M. 2009. Population genomics and the bacterial species concept. *Methods Mol Biol,* 532**,** 367-377.

Rodelas, B., Lithgow, J. K., Wisniewski-Dye, F., Hardman, A., Wilkinson, A., Economou, A., Williams, P. & Downie, J. A. 1999. Analysis of Quorum-Sensing-Dependent Control of Rhizosphere-Expressed (rhi) Genes in *Rhizobium leguminosarum bv. viciae*. *Journal of Bacteriology,* 181**,** 3816-3823.

Rogel, M. A., Ormeno-Orrillo, E. & Martinez Romero, E. 2011. Symbiovars in rhizobia reflect bacterial adaptation to legumes. *Syst Appl Microbiol,* 34**,** 96-104.

Routledge, R. 2005. Fisher's Exact Test. *Encyclopedia of Biostatistics.* John Wiley & Sons, Ltd.

Saeys, Y., Inza, I. & Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics,* 23**,** 2507-2517.

Saitou, N. & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol,* 4**,** 406-425.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A. & Lusis, A. J. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet,* 37**,** 710-717.

Schliep, K., Lopez, P., Lapointe, F. J. & Bapteste, E. 2011. Harvesting evolutionary signals in a forest of prokaryotic gene trees. *Mol Biol Evol,* 28**,** 1393-1405.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics,* 27**,** 592-593.

Schweder, T. & Spjøtvoll, E. 1982. Plots of p-values to evaluate many tests simultaneously. *Biometrika,* 69**,** 493-502.

Sentchilo, V., Mayer, A. P., Guy, L., Miyazaki, R., Green Tringe, S., Barry, K., Malfatti, S., Goessmann, A., Robinson-Rechavi, M. & Van Der Meer, J. R. 2013. Community-wide plasmid gene mobilization and selection. *ISME J,* 7**,** 1173-1186.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal,* 27**,** 379-423.

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F. & Alm, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *science,* 336**,** 48-51.

Siefert, J. 2009. Defining the Mobilome. *In:* GOGARTEN, M., et al. (eds.) *Horizontal Gene Transfer.* Humana Press.

Slonim, N., Elemento, O. & Tavazoie, S. 2006. Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Biol,* 2**,** 2006.0005.

Smokvina, T., Wels, M., Polka, J., Chervaux, C., Brisse, S., Boekhorst, J., Van Hylckama Vlieg, J. E. & Siezen, R. J. 2013. *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. *PLoS One,* 8**,** e68731.

Steinhauser, D., Krall, L., Müssig, C., Büssis, D. & Usadel, B. 2007. Correlation Networks.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics,* 18 Suppl 2**,** S231-240.

Storey, J. & Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA,* 100**,** 9440 - 9445.

Strimmer, K. 2008. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics,* 24**,** 1461-1462.

Sugawara, M., Epstein, B., Badgley, B., Unno, T., Xu, L., Reese, J., Gyaneshwar, P., Denny, R., Mudge, J., Bharti, A., Farmer, A., May, G., Woodward, J., Medigue, C., Vallenet, D., Lajus, A., Rouy, Z., Martinez-Vaz, B., Tiffin, P., Young, N. & Sadowsky, M. 2013. Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biology,* 14**,** R17.

Suwanto, A. & Kaplan, S. 1989. Physical and genetic mapping of the Rhodobacter sphaeroides 2.4. 1 genome: presence of two unique circular chromosomes. *Journal of Bacteriology,* 171**,** 5850-5859.

Tamames, J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol,* 2**,** 1-0020.0011.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. 1997. Conserved Clusters of Functionally Related Genes in Two Bacterial Genomes. *Journal of Molecular Evolution,* 44**,** 66-73.

Tamminen, M., Virta, M., Fani, R. & Fondi, M. 2012. Large-scale analysis of plasmid relationships through gene-sharing networks. *Mol Biol Evol,* 29**,** 1225-1240.

Tamura, M. & D'haeseleer, P. 2008. Microbial genotype-phenotype mapping by class association rule mining. *Bioinformatics,* 24**,** 1523-1529.

Tanabe, M. & Kanehisa, M. 2002. Using the KEGG Database Resource. *Current Protocols in Bioinformatics.* John Wiley & Sons, Inc.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. & Koonin, E. V. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research,* 29**,** 22-28.

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. B., 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". *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 13950-13955.

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**,** 563-576.

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, e00003-00013.

Van Landeghem, S., Abeel, T., Saeys, Y. & Van De Peer, Y. 2010. Discriminative and informative features for biomolecular text mining with ensemble feature selection. *Bioinformatics,* 26, i554-560.

Van Passel, M., Bart, A., Thygesen, H., Luyf, A., Van Kampen, A. & Van Der Ende, A. 2005. An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics,* 6, 163.

Wang, C., Zhang, H.-B., Wang, L.-H. & Zhang, L.-H. 2006. Succinic semialdehyde couples stress response to quorum-sensing signal decay in Agrobacterium tumefaciens. *Molecular Microbiology,* 62, 45-56.

White, J. P., Prell, J., Ramachandran, V. K. & Poole, P. S. 2009. Characterization of a γ-aminobutyric acid transport system of *Rhizobium leguminosarum bv. viciae* 3841. *J Bacteriol,* 191, 1547-1555.

Wilkinson, M., Mcinerney, J. O., Hirt, R. P., Foster, P. G. & Embley, T. M. 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol Evol,* 22, 114-115.

Wisniewski-Dye, F. & Downie, J. A. 2002. Quorum-sensing in Rhizobium. *Antonie Van Leeuwenhoek,* 81, 397-407.

Wisniewski-Dyé, F., Lozano, L., Acosta-Cruz, E., Borland, S., Drogue, B., Prigent-Combaret, C., Rouy, Z., Barbe, V., Herrera, A. M. & González, V. 2012. Genome sequence of Azospirillum brasilense CBG497 and comparative analyses of Azospirillum core and accessory genomes provide insight into niche adaptation. *Genes,* 3, 576-602.

Woese, C. R., Kandler, O. & Wheelis, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences,* 87, 4576-4579.

Wu, J., Kasif, S. & Delisi, C. 2003. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics,* 19, 1524-1530.

Wu, X., Jin, L. & Xiong, M. 2009. Mutual Information for Testing Gene-Environment Interaction. *PLoS ONE,* 4, e4578.

Xu, F., Jerlström-Hultqvist, J., Einarsson, E., Ástvaldsson, Á., Svärd, S. G. & Andersson, J. O. 2014. The Genome of *Spironucleus salmonicida*

Highlights a Fish Pathogen Adapted to Fluctuating Environments. *PLoS Genet,* 10**,** e1004053.

Yan, J., Bradley, M. D., Friedman, J. & Welch, R. D. 2014. Phenotypic profiling of ABC transporter coding genes in *Myxococcus xanthus*. *Frontiers in Microbiology,* 5**,** 352.

Yerrapragada, S., Siefert, J. L. & Fox, G. E. 2009. Horizontal gene transfer in cyanobacterial signature genes. *Methods Mol Biol,* 532**,** 339-366.

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., Demetriou, L. & Apte, R. G. 1987. Rhizobium Population Genetics: Enzyme Polymorphism in *Rhizobium leguminosarum* from Plants and Soil in a Pea Crop. *Applied and Environmental Microbiology,* 53**,** 397-402.

Zhang, H.-B., Wang, L.-H. & Zhang, L.-H. 2002. Genetic control of quorum-sensing signal turnover in Agrobacterium tumefaciens. *Proceedings of the National Academy of Sciences,* 99**,** 4638-4643.

Zhang, Y. & Sievert, S. M. 2014. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in Epsilonproteobacteria. *Frontiers in Microbiology,* 5.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F. & Papke, R. T. 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Research,* 16**,** 1099-1108.