Understanding the intraspecies genetic and phenotypic diversity of the clover symbiont *Rhizobium leguminosarum*

Bryden May Fields

Doctor of Philosophy

University of York
Department of Biology

July 2020

Abstract

Rhizobia are agriculturally important bacteria capable of forming symbiosis with legumes and fixing atmospheric nitrogen which sustainably improves plant productivity and soil fertility. The *Rhizobium leguminosarum* species complex is highly genetically diverse and contains five genetically distinct genospecies. Significant phenotypic diversity is also displayed within *Rhizobium leguminosarum*; however, no phenotypes are genospecies-exclusive. The importance of the broad genetic diversity of *Rhizobium leguminosarum* and its influence on phenotypic diversity and rhizosphere-associated interactions are unclear. In this thesis, *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) intraspecies diversity was investigated by assessing the genetic and phenotypic variation of white clover nodule *Rlt* from agricultural field managements across Europe.

This thesis identified that the significant genetic diversity of *Rlt* can manifest in substantial transcriptional and phenotypic variation across strains, and this diversity can influence plant-mediated symbiont selectivity and competitive strain interactions. A novel multiplexed high-throughput amplicon sequencing approach, MAUI-seq, was developed to improve detection of chimeras and other erroneous sequences for confident determination of intraspecies diversity from environmental samples. Using this method, significant *Rlt* nodule population diversity was identified between clover genotypes due to the combined effects of plant-host filtering and geospatial variation in allele frequencies of individual genes. Investigation of multiple *Rlt* strain transcriptomes demonstrated that genospecies displayed differences in core genome expression which was associated with phenotypic growth traits and putative differences in bacterial metabolism. Genomic and transcriptomic variation was utilised to identify transcriptional units conserved across strains. Pairwise growth competition experiments between *Rlt* strains further showed that significant competitive variation is evident and potentially associated with genospecies differences. This research demonstrates that utilising multiple strains can aid identification of species-specific traits by considering the representative variation within a species. The work presented here has laid the groundwork for future investigation into the implications of intraspecies diversity for symbiotic effectiveness in the rhizobia-legume symbiosis.

Table of Contents

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Accompanying Material

Additional File 1

PDF for **Chapter 2**: MAUI-seq Laboratory Protocol and additional method recommendations.

Additional File 2

Excel file containing additional tables for **Chapter 2**: Detailed output sequences for all three HTAS denoising methods (MAUI-seq, DADA2 and UNOISE3) for each gene (*rpoB*, *recA*, *nodA* and *nodD*).

Additional File 3

Text document for **Chapter 2**: DADA2 tutorial code.

Additional File 4

Text document for **Chapter 2**: UNOISE3 tutorial code.

Additional File 5

Text document for **Chapter 2**: R code for Principal Components Analysis, observed vs expected ratios analysis, and allele heatmap generation.

Additional File 6

Excel file containing additional tables for **Chapters 4, 5 and 6**.

Chapter 4:

Table S1. RNA-seq data quality statistics from samtools, rseqc and HISAT2.

Table S2. Between-genospecies and within-genospecies core genome expression principal component pathway enrichment with metacyc and KEGG.

Table S3. WGCNA co-expressed gene modules. Modules are numbered in order of decreasing size. Number of genes in each module are provided along with assigned WGCNA module colours, identified eigengene value differences between genospecies, significant correlations to Tryptone Yeast broth growth phenotypes, associated KEGG pathways, metacyc pathway enrichment, and identification of symbiosis genes in modules.

Chapter 5:

Table S4. 94 conserved transcriptional units across genospecies. Gene names, gene type, and putative gene function are listed.

Table S5. The number of transcriptional units maintained within each *Rlt* WGCNA core gene module is displayed for each genospecies. WGCNA modules were generated in Chapter 4.

Table S6. List of transcriptional units identified for each genospecies detailing the orthologous gene groups each unit contains.

Chapter 6:

Table S7. Percentage identity of *Rhizobium leguminosarum* bacteriocins and quorum sensing molecules found in 24 *Rlt* strains to reference sequences.

Acknowledgements

Firstly, I would like to thank my supervisor Ville-Petri Friman for believing there was a PhD in me and for his support, guidance, patience and consistent positivity throughout the entire project; it has been a true privilege to work with him for the past four years. Thank you to my co-supervisor Thorunn Helgason for her support and positive strategies for combatting project challenges. I would also like to express my gratitude to my unofficial supervisor J. Peter W. Young for his guidance and expert advice on all thing's rhizobia and bacterial genomics. Thank you to Stig U. Andersen, head of the NCHAIN research consortium at Aarhus University, from which my project was funded, for his useful advice on transcriptome analyses and time dedicated to all of the skype meetings. I must also thank Jon Pitchford and Luke Mackinder, my fantastic Thesis Advisory Panel, for always asking interesting questions and providing me with advice and ideas for how to improve upon my work. I would not have been able to achieve as much as I have without the diverse range of support from all of these advisors, and I am very grateful to have been able to learn and grow as a researcher from their support.

Thank you to Emma Smith for carrying out the phenotype work which gave more clarity to the genomics of Chapters 4 and 6. I'd also like to thank Ellie Harrison, from the University of Sheffield, for inviting me to collaborate with her on some cool rhizobia interaction ideas that formed Chapter 6. Thank you to David Sherlock for sharing his HTAS knowledge and tips. I would also like to give a huge thank you to staff in the University of York Technology Facility, particularly Sally James, John Davey and Lesley Gilbert for all the sequencing work and support, and for answering my many long-winded questions. Thank you to the PhD TF Facilities Awards Panel for awarding me funding which enabled me to carry out additional transcriptomics work for this project.

I'd like to*concomitantly* thank everyone in the Friman Lab group, past and present, who have created such a fun, helpful and supportive lab atmosphere, which has been a joy to be a part of. I'd like to thank the Helgason lab group for offering their HTAS advice and letting me take over their lab spaces from time to time. In today's world research is rarely done alone, and I am very lucky to have also undertaken this project as part of the large collaborative effort led by the international, interdisciplinary NCHAIN research consortium. This project and its findings would not have been achievable without the research contributions, insights and discussions from consortium PhD students, post-doctoral researchers, principal investigators and industrial partners involved. It has specifically been a pleasure to collaborate with Sara Moeskjær and M. Izabel A. Cavassim within the consortium; they made me feel very welcome during my times in Denmark, and their field work and genomic analysis also provided a strong foundation for the work undertaken in this project. Thank you also to Niels Roulund from DLF Trifolium who was always very accommodating and helpful when I came to sample their clover trials. I am also very grateful for the numerous opportunities given to me as a result of being a part of the consortium including spending time in Denmark carrying out fieldwork and attending consortium meetings, and also for the chance to present my project at multiple international conferences.

Finally, I would like to thank my parents, sisters and grandparents for their unconditional love, encouragement and support throughout not just my PhD but the entirety of my very long education which has led me to this point. To my parents, thank you for always giving me a home, and for reminding me to not fixate so much on the milestones, to enjoy life along the way, and to find the positives in the difficult times. To my sisters, thank you particularly for the relatable memes and the film night commentaries which boosted my morale to complete this project. My family have been my rock, and this thesis would not have happened without their endless patience to listen to me ramble on about germs for hours on end.

Declaration by Author

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.

Exception to declaration:

Chapter 2.
This chapter has been submitted for publication in April 2020 to Molecular Ecology Resources and is reproduced here as the submitted version.
This paper was undertaken with joint first authorship between Bryden Fields and Sara Moeskjær. J. Peter W. Young developed and conceptualised the MAUI-seq method, wrote the scripts, and implemented them. Bryden Fields performed all laboratory work including PCRs, Nextera labelling, and preparation of the strains for sequencing as well as the PEAR assembly of the reads, and the UNOISE analysis. Bryden Fields and Sara Moeskjær did all field sampling and collected the nodules, performed all analysis in parallel, and wrote the manuscript. Sara Moeskjær prepared the figures and carried out the DADA2 analysis. The text has been written and revised jointly between Bryden Fields, Sara Moeskjær, Ville-Petri Friman, Stig U. Andersen and J. Peter W. Young. This paper has also been improved as a result of comments from five anonymous reviewers during the submission and revision process.

Chapter 4.
Tryptone Yeast broth growth experiments: The laboratory work was undertaken by Emma Smith for inclusion in her Master's thesis.

Chapter 6.
Ecoplate Biolog growth experiments: The laboratory work was undertaken by Emma Smith for inclusion in her Master's thesis.

Bryden May Fields

THE UNIVERSITY *of York*

DEPARTMENT OF BIOLOGY
P O Box No 373 York YO10 5YW U K

Statement about co-authorship of work for inclusion in PhD thesis.

The chapter based on the publication entitled "MAUI-seq: Multiplexed, high-throughput amplicon diversity profiling using unique molecular identifiers" is co-authored by Bryden Fields and Sara Moeskjær and will be included in both their PhD theses at the University of York, UK, and Aarhus University, Denmark as a components of their PhD work.

The following author's contributions summarise the respective contributions of both authors to this work.

Author contributions

J. Peter W. Young developed and conceptualised the MAUI-seq method, wrote the scripts, and implemented them.

Bryden Fields performed all laboratory work including PCRs, Nextera labelling, and preparation of the strains for sequencing as well as the PEAR assembly of the reads.

Bryden Fields and Sara Moeskjær did all field sampling and collected the nodules, performed all analysis in parallel, and wrote the manuscript.

Sara Moeskjær prepared the figures.

Signed on 31st May 2019

Bryden Fields

Sara Moeskjær

Gavin H. Thomas
Director of Graduate Studies,
Department of Biology,
University of York.

Chapter 1. Introduction

1.1. Overview

Securing access to an increased global food supply using sustainable solutions that also take advantage of nutrient poor soils is a major scientific issue for the future of agricultural practices (Godfray *et al.*, 2010; Tilman *et al.*, 2011). Despite the prevalence of dinitrogen (N_2) gas in the atmosphere, nitrogen in a biologically accessible form is commonly limiting in agricultural systems impeding plant growth. In order to circumvent this restriction, specific plants called legumes form a beneficial symbiotic relationship with soil bacteria called rhizobia which are capable of converting dinitrogen gas into a plant-accessible form, such as ammonia (NH_3), in a natural process called biological nitrogen fixation (BNF) (Oldroyd *et al.*, 2011; Terpolilli *et al.*, 2014). Within this mutualism, the diazotrophic rhizobia fix nitrogen for the legume using the microbial nitrogenase enzyme in exchange for carbon sources such as photosynthates. Not only is the symbiosis beneficial to the symbiotic partners, but the process also re-fertilises soil nitrogen reserves which is then accessible to non-leguminous plants (Bohlool *et al.*, 1992). This non-obligatory symbiosis is of agricultural importance as it increases the amount of available nitrogen to the legume which enables increased plant growth and subsequent yield. Therefore, BNF through rhizobia inoculation into agricultural systems provides an alternative sustainable, environmentally friendly, and economically attractive method for improving soil fertility over exogenous nitrogen fertilizer (Bohlool *et al.*, 1992; Wang *et al.*, 2012). On organic farms and forage pastures rhizobial BNF is vital to productivity.

In the UK alone, 70% of total land area is agriculturally managed, and a significant proportion of this is pastureland (Andrews et al., 2007). Forage pastures are commonly used for animal husbandry such as sheep and cow feed for dairy and meat production. In cultivated pastures across temperate agricultural systems, BNF is predominantly driven by symbiosis between *Trifolium repens* (white clover) and the highly strictly specific microsymbiont *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) (Dénarié, Debelle and Rosenberg, 1992; Annicchiarico et al., 2015).

Although the nitrogen fixing symbiosis between white clover and *Rlt* is strictly species specific, there is also a large amount of variation in the symbiotic compatibility between different strains and clover genotypes. This can result in varied nitrogen fixing efficiencies depending on the symbiotic partners and environmental context (Terpolilli *et al.*, 2014). Consequently, there is a need to identify rhizobium inoculants that are compatible with both the legume host and its soil environment microbiota. Intraspecies genetic and phenotypic diversity of rhizobia is large, and it is unclear what causes the maintenance of this diversity in soils, but possible explanations include through strain-strain interactions, strain-plant interactions and heterogeneity of soil environments. Similarly, how this genetic diversity translates into phenotypic differences is unclear. It is thus important to understand to what extent rhizobium intraspecies diversity is linked with its fitness and symbiotic specificity and how this diversity is maintained. Improving our understanding of the relevance of intraspecies diversity and using this knowledge to applying precision farming techniques will aid progression to securing sustainable global food security.

This study aimed to investigate the level of observed intraspecies diversity within white clover nodule rhizobia collected from agricultural fields across Europe, and to explore the potential mechanisms behind the maintenance of this rhizobial intraspecies diversity and its implications for the efficacy of the legume-rhizobia symbiosis. This work contributed to a broad collaborative effort with Aarhus University and industrial partners DLF Trifolium, SEGES and Legume Technology as part of the NCHAIN research consortium funded by Innovation Denmark. Additionally, strain interaction work in this study also forms part of a collaboration with the University of Sheffield.

In this introduction, background information for the research undertaken within this thesis is provided, and the aims and objectives of the project are defined. The background context regarding the rhizobium-legume symbiosis will be discussed, firstly introducing the importance of the symbiosis for agriculture, the mechanisms and specificity of rhizobia-legume interactions, and the challenges for exploiting the symbiosis commercially. Additionally, there is a focus on the intraspecific genetic, transcriptomic and phenotypic variation observed within *Rhizobium leguminosarum*, and how these three topics are investigated in the thesis. This introduction will also discuss the particular rhizobia strains of interest to this study. To conclude the chapter, the aims and objectives are summarised with an outline of the thesis.

1.2. The rhizobia-legume mutualism

1.2.1. Importance of the legume rhizobia symbiosis in agriculture

BNF is the principal source of naturally fixed nitrogen and occurs through the conversion of atmospheric dinitrogen into ammonia compounds using the microbial nitrogenase enzyme. Only certain species of bacteria and archaea, called diazotrophs, are capable of performing BNF, and exist in free-living and symbiotic lifestyles. Specifically, symbiotic bacterial diazotrophs called rhizobia can perform BNF when engaging in symbiosis with legume plant species. The exploitation of BNF provides an ecologically sustainable and environmentally friendly method to reduce the amount of external nitrogen input into agricultural systems and improve management of its resources compared to Haber-Bosch generated fertilisers (Bohlool *et al*., 1992). The symbiotic mutualism between crop/forage legumes and rhizobia are the most important nitrogen fixing partners in agriculture (Herridge, Peoples and Boddey, 2008), and drives the largest natural source of nitrogen into agricultural systems (Galloway *et al*., 2004). It has been estimated that symbiotic partnerships fix up to 80% of BNF in agriculture (O'Hara, 1998). However, since the mass production and application of inorganic artificial nitrogen fertiliser from the Haber-Bosch process, the number of farming managements utilising legumes has drastically reduced (Galloway *et al.*, 2004; Erisman *et al.*, 2008). Although high applications of fertilisers generated by the Haber-Bosch process have aided attainment of increased crop yields, intensive farming practices have also caused a substantial negative impact on the environment through various processes: reduction in soil condition causing eutrophication of aquatic environments and risks to human health; nitrate leaching; increasing greenhouse gas nitrogen oxide emissions that contribute to air pollution; and loss of biodiversity (Tilman *et al*., 2002, 2011; Jensen and Hauggaard-Nielsen, 2003; Erisman *et al*., 2008; Cameron, Di and Moir, 2013). Additionally, a large amount of fossil fuels are required for Haber-Bosch nitrogen fertiliser production in comparison to nitrogen generated by BNF, which is essentially 'free' due to being generated from the exchange with legume photosynthates (Jensen and Hauggaard-Nielsen, 2003). Subsequently, in order to meet the demand of intensive agriculture, and to reduce the unsustainable ramifications of increased chemical fertiliser use, efforts are being made to optimise the production of symbiotic BNF which is a more environmentally sustainable source of nitrogen (Jensen and Hauggaard-Nielsen, 2003; Reeve *et al.*, 2015).

Nitrogen fixation in soils with low rhizobium abundance or activity can be improved by inoculating legumes with rhizobia strains that are highly efficient for BNF (Sessitsch *et al.*, 2002). Plant growth promoting rhizobacteria, such as rhizobia, provide not only beneficial effects to their associated legumes but also to non-leguminous neighbouring plants and its surrounding soil environment. The most evident benefit for legumes is the biofertilisation from forming symbiosis with rhizobia, whereby rhizobia facilitate the uptake of accessible nitrogen by the legume (Vessey, 2003). Use of the rhizobia-legume symbiosis to re-fertilise soils in crop rotations also helps to maintain soil nitrogen reserves and provides bio-actively accessible nitrogen to non-leguminous plants which in turn increases their crop yields (Bohlool *et al.*, 1992; Graham and Vance, 2000; Sessitsch *et al.*, 2002; Lüscher *et al.*, 2014). The use of legumes in agriculture itself is additionally an important protein source for both humans and livestock, and can improve soil structure, aid crop disease and pest control, and promote biological diversity (Jensen and Hauggaard-Nielsen, 2003; Lindström *et al.*, 2010). Taken together, the use of the legume-rhizobia symbiosis in agriculture provides a variety of ecosystem benefits over artificial nitrogen fertiliser, which overall help to reduce fossil energy use and greenhouse gas emissions, increase crop yields, and ultimately provide food to humans and animals in a way that maintains environment integrity (Jensen *et al.*, 2012; Lüscher *et al.*, 2014; Phelan *et al.*, 2015). BNF alone may not match the yields produced by intensive agriculture with artificial fertiliser, however BNF in combination with chemical fertiliser and animal manure could also provide a promising alternative for future more sustainable agricultural systems (Tilman *et al*., 2002, 2011; Jensen and Hauggaard-Nielsen, 2003).

A globally important cultivated forage crop is the perennial legume white clover species that has been introduced across temperate regions from the UK. It is usually planted in combination with perennial ryegrass for pasture systems to improve livestock nutrition (Graham and Vance, 2003; Phelan *et al.*, 2015). In this way, nitrogen input into pastures is driven by the symbiosis between clover and rhizobia which causes nitrogen to become available also to the grass via legume exudates, legume senescence, and manure from livestock digesting the clover-grass sown mixture (Ledgard, 2001; Andrews *et al.*, 2007). Forage legume pastures are vital for maintaining animal husbandry practices and have been used for centuries to provide protein and energy for increased cow dairy and meat production which are ranked 1st and 3rd top food commodities across the globe, respectively (Graham and Vance, 2003; Lindström *et al.*, 2010; Lüscher *et al.*, 2014). Countries such as New Zealand and Australia rely on the BNF from white clover to maintain their pasture productivity, and additionally approximately 40% of agricultural land in Europe is grassland used as ruminant animal feed (Andrews *et al.*, 2007). It has also been suggested that 20 million hectares of land is used for forage legume monocultures across the world, not including their land area in mixtures with grasses which is likely to be far greater than monocultures (Graham and Vance, 2003; Phelan *et al.*, 2015). Furthermore, land used for grazing has been suggested to total the most widespread form of agriculture accounting for 25% of land use (Asner *et al.*, 2004). Clover species are additionally used as cover crops in order to re-fertilise the soil for other non-leguminous species (Fageria, Baligar and Bailey, 2005). Therefore, the benefits and significance to optimising the efficiency of nitrogen input for these clover systems are evident.

Global nitrogen fixation input estimates for symbiosis in pasture and fodder legumes has been suggested to be around 12–25 Tg, with a total estimate of 50–70 Tg N fixed when agricultural legume crops are also included (Herridge, Peoples and Boddey, 2008; Lindström *et al.*, 2010). Between 35%-60% less fossil fuel energy is used by legume crops and pastures than artificially fertilised grasslands and cereal crops, which is predominantly due to the reduced requirements for artificial fertilisers by legumes (Jensen *et al.*, 2012). It has also been suggested that clover could have between £125–160 ha^{-1} annum^{-1} advantage over N fertiliser for low maintenance perennial white clover – ryegrass pastures in the UK (Andrews *et al.*, 2007). This highlights that the clover-rhizobia symbiosis as a means of agricultural N management is not only considered environmentally sustainable but also economically feasible.

1.2.2. Symbiosis mechanism between *Rhizobium leguminosarum* sv. *trifolii* and white clover

Rhizobium leguminosarum symbiovar *trifolii* (*Rlt*) strains are facultative microsymbionts that can form a mutualistic symbiotic relationship with white clover (*Trifolium repens*) roots. Rhizobia have two different life stages; a free-living physiology in the soil, and a symbiotic bacteroid physiology within plant cells. For symbiosis to occur, rhizobia are required to contain symbiosis genes usually located on a symbiosis plasmid. The rhizobial symbiosis genes mediate the interactions with the legume species and can be categorised into three main components: nodulation (*nod*), *nif* and *fix* genes. In brief, *nod* genes expressed by free-living rhizosphere soil rhizobia elicit the formation of a symbiotic root organ called a nodule (Göttfert, 1993), whereas *nif* and *fix* genes are expressed at later stage in symbiotic establishment and are involved in nitrogen fixation within the legume root nodule (Fischer, 1994).

In order to initiate symbiotic establishment, the clover root releases clover-specific flavonoids that are detected by the NodD protein in free-living, soil rhizosphere-dwelling *Rlt* strains. NodD is a LysR family transcriptional regulator of *nod* genes, and is activated by specific clover flavonoids, which enables NodD to transcriptionally activate *nod* gene expression (Hong, Burn and Johnston, 1987a). Expression of *nod* genes induces production and transport of rhizobia Nod factors, which are lipochitooligosaccharides (LCOs) that are secreted from the rhizobia and initiate symbiotic establishment. The LCO backbone is synthesised by *nodABC* gene products, and additional *nod* gene products (NodFEHGPQ), and *nol* and *noe* gene products, modify the chitin backbone by adding species-specific substituents which determines strain host range (Haeze and Holsters, 2002; Lupwayi, Clayton and Rice, 2006; Wang, Liu and Zhu, 2018). This is important, as it is vital that Nod factors are recognised by Nod factor receptors-like kinases on the legume root in order to instigate nodule formation (Herman P. Spaink *et al.*, 1987; Oldroyd *et al.*, 2011; Downie, 2014). Expression of *nod* genes can also induce chemotaxis of the rhizobia to the clover root (Munoz Aguilar *et al.*, 1988). Recognition of structurally specific Nod factors by the legume instigates root-hair curling and formation of an infection thread from which the root colonised rhizobia grow through to reach the developing root nodule

(Downie, 2014). *nod* genes are crucial for nodulation and symbiosis as mutations result in a Nod- phenotype (Downie *et al.*, 1985; Jacobs, Egelhoff and Long, 1985).

After being internalised into a plant-derived symbiosome membrane, *Rlt* cells proliferate and terminally differentiate into non-motile, nitrogen-fixing bacteroids (Oldroyd *et al.*, 2011; Wang, Liu and Zhu, 2018). This proliferation aids creation of the new root nodule structure. In white clover, the nodules are indeterminate and so maintain an apical meristem (Łotocka, Kopcińska and Skalniak, 2012). Terminal differentiation of rhizobia into bacteroids includes downregulation of cell growth and reproduction related gene expression, endoreduplication of genomes, cell elongation and increased membrane permeability (Mergaert *et al.*, 2006; Kereszt, Mergaert and Kondorosi, 2011; Oldroyd *et al.*, 2011; Haag *et al.*, 2013). This dramatic alteration in cell physiology and gene expression enables bacteroids to be better adapted to the intracellular plant environment (Mergaert *et al.*, 2006; Haag *et al.*, 2013; Wang, Liu and Zhu, 2018). Additionally, expression of *nif* and *fix* genes for nitrogen fixation are upregulated, which control the conversion of dinitrogen (N_2) into ammonia (NH_3) catalysed by a nitrogenase enzyme (Oldroyd *et al.*, 2011). Nitrogen fixing nodules are pink, as a result of the production of leghaemoglobin by the plant cells, which enables a low enough oxygen concentration for the oxygen-sensitive nitrogenase to work, while also ensuring rhizobium can undergo aerobic respiration (Ott *et al.*, 2005). Consequently, a symbiosis is then established within the nodule whereby the plant supplies photosynthates and a safe niche to the rhizobia, and in turn, the rhizobia supplies vital accessible nitrogen to the plant through Biological Nitrogen Fixation. In indeterminate nodules, different zones of bacterial differentiation are evident whereby some rhizobia will not have yet terminally differentiated into bacteroids. It is these non-differentiated free-living rhizobia within the nodule that will go on to recolonise the soil at the end of symbiosis when the nodule senesces, as the terminally differentiated bacteroids are unable to reproduce (Sprent, Sutherland and De Faria, 1987; Mergaert *et al.*, 2006).

1.2.3. Partner choice for symbiotic establishment

While description of symbiosis mechanisms predominantly highlights the single pairwise interaction between legume host and *Rhizobium* strain, in reality there can be many rhizobial strains interacting and forming symbiosis with the legume at any one time, with varying degrees of nitrogen fixing effectiveness (Mytton, 1975; Denison, 2000). For a successful symbiosis to occur, both the rhizobium and the legume must be compatible throughout symbiotic establishment. Incompatibility between symbiotic partners is often observed by nodule organogenesis failure on a specific legume host or where there is an absence of nitrogen fixation in nodules. Effectiveness of symbiosis is predominantly determined by the nitrogen fixing ability of the rhizobial strain within the legume nodule (Laranjo, Alexandre and Oliveira, 2014).

In the majority of cases, it is well known that only particular rhizobia will form a symbiosis with certain legumes (symbiovars), thereby providing interaction specificity. Some rhizobial strains are more promiscuous than others; for example, *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) nodulates only *Trifolium* spp. (clover), although other rhizobia species can form symbiosis with multiple legume species and have a wider host-range within their symbiovar (Perret, Staehelin and Broughton, 2000; Hirsch, Lum and Downie, 2001). This rhizobia-legume interaction specificity occurs at both species level and also individual genotype level (Perret, Staehelin and Broughton, 2000; Wang *et al.*, 2012; Wang, Liu and Zhu, 2018). Notable examples include ICC105's interaction with white and Caucasian clovers, and WSM1689's ability to only fix N_2 efficiently in *Trifolium uniflorum* (Miller *et al.*, 2007; Terpolilli *et al.*, 2014). Similarly, some legume genotypes are suggested to differ in their range of compatibility with rhizobia strains, as wild clovers have been shown to be compatible with a potentially greater number of rhizobial strains than their domesticated crop species equivalents (Mutch and Young, 2004; Wang *et al.*, 2012).

Multiple genetic and molecular pathways have been identified that regulate the symbiotic specificity and compatibility between symbiotic partners (Wang, Liu and Zhu, 2018). Legume selection for a *Rhizobium* strain is largely influenced by the genetic compatibility with rhizobial symbiosis genes. In particular, specificity can be mediated by: 1) legume flavonoid interaction with the rhizobial NodD transcriptional regulator; and 2) legume recognition of rhizobium Nod factors (Haeze and Holsters, 2002; Wang, Liu and Zhu, 2018). Compatibility for NodD is crucial for initiation of symbiosis as it is activated by legume flavonoids and functions as a regulator of *nod* gene activation (Redmond *et al.*, 1986; Maj *et al.*, 2010; Hassan and Mathesius, 2012). NodD proteins from different rhizobia strains have evolved to be activated by different legume flavonoids (Perret, Staehelin and Broughton, 2000). Only certain legume flavonoids will induce specific NodD proteins and instigate a bacterial response, thereby ensuring legume host specificity of rhizobial partner choice. In addition to activation of NodD, legume flavonoids can act as chemo-attractants to the root and the strength of their attraction can differ for specific rhizobial strains (Hassan and Mathesius, 2012). Furthermore, once *nod* genes are activated, rhizobia can produce multiple types of Nod factors (Wang, Liu and Zhu, 2018). Plants can also produce multiple types of flavonoids, for example, multiple flavonoids have been detected from white clover secretions which can activate and inhibit symbiosis (Redmond *et al.*, 1986; Djordjevic *et al.*, 1987; Carlsen *et al.*, 2012). The wide-variety of plant-specific flavonoid and rhizobial-specific Nod factor combinations available, in addition to further downstream interactions, are suggested to determine the specificity and compatibility of symbioses at both the inter- and intra-species levels (Perret, Staehelin and Broughton, 2000; Wang, Liu and Zhu, 2018).

Molecular interactions between the rhizobia and legume are complex, and other signalling mechanisms in addition to Nod factors are important for determining symbiotic specificity and host range. Symbiotic specificity is also known to be mediated by legume detection of rhizobial exopolysaccharides, lipopolysaccharides, secretion systems, their secreted effectors, and microbe-associated molecular patterns (MAMPs), as part of the plant immune response for pathogen detection (Kannenberg, Rathbun and Brewin, 1992; Fauvart and Michiels, 2008; Deakin and Broughton, 2009; Downie, 2010; Okazaki *et al.*, 2013, 2016; Kawaharada *et al.*, 2015; Zipfel and Oldroyd, 2017; Wang, Liu and Zhu, 2018). For example, the production, composition and structure of exopolysaccharides can vary between rhizobial species and has been shown to determine nodulation capability in *Rlt* (Fraysse, Couderc and Poinsot, 2003; Skorupska *et al.*, 2006; Janczarek and Rachwał, 2013; Ghosh and Maiti, 2016; Rachwal *et al.*, 2016). Despite the multiple levels of specificity required for symbiotic establishment, this does not necessarily lead to optimal, efficient, successful symbioses. Some rhizobia strains reach the nodule and are subsequently unable to fix nitrogen efficiently, for example, due to early nodule sensencence induced by plant immune response (S. Yang *et al.*, 2017; Wang, Liu and Zhu, 2018). On the other hand, some poor nitrogen-fixing strains are able to remain in nodules and benefit from the host, giving way to the symbiotic manipulation of defective "cheater" strains (Simms and Taylor, 2002).

Taken together, potential rhizobia and legume interactions are largely diverse at both inter- and intraspecies levels. This diversity of interactions is accounted for during symbiotic establishment by the implementation of multiple mechanisms to select for the most compatible symbiotic partners, thereby introducing symbiotic specificity between and within rhizobia and legume species.

1.2.4. Challenges for commercially exploiting the rhizobia-legume symbiosis

Optimisation of the symbiosis has involved choosing the most compatible rhizobia strains and legume genotypes for agricultural application. However, in many instances the legume genotype and the environmental field factors are predetermined, leaving only the selection of rhizobia strains as the flexible parameter for improving symbiotic efficiency in the field (Rys and Bonish, 1981; Lupwayi, Clayton and Rice, 2006). For a rhizobia strain to be considered a suitable inoculant it must be able to colonise the soil and endure it's abiotic environmental conditions, compete with native rhizobia and other microbes, form nodules successfully, be an effective nitrogen fixer, and have no adverse effects on non-target hosts (Brockwell and Bottomley, 1995; Howieson, Malden and Yates, 2000; Checcucci *et al.*, 2017; Zaidi, Khan and Musarrat, 2017). This is not to mention the strain's ability to survive farmers practices and manufacturing standards including method of sowing, pesticide usage, and having a long shelf life (Catroux, Hartmann and Revellin, 2001). However, with the long list of inoculant prerequisites it is not surprising that the majority of inoculants produced globally are suggested to be of suboptimal quality (Catroux, Hartmann and Revellin, 2001).

Rhizobium leguminosarum symbiovar *trifolii* strains RCR221/TA1, VAR1, CC275e and U204 are commonly used as commercial inoculants of white clover (Brockwell, McIlroy and Hebb, 1998; Batista *et al.*, 2015; Delestre *et al.*, 2015; Roberts *et al.*, 2017; Tartaglia *et al.*, 2019). The process of choosing inoculant strains has been largely based on testing growth ability under laboratory conditions and subsequently trialling strains for resilience in the field. Previously, strains have been chosen due to their efficient nitrogen-fixing capabilities, and strain have also been selected based on their interactions within chosen legume cultivars (Howieson, Malden and Yates, 2000; Denton *et al.*, 2003; Checcucci *et al.*, 2017).

Rhizobial intraspecies diversity has been known to cause various challenges for symbiotic establishment of commercial rhizobia inoculants in the field. It is well documented that inoculants are commonly unable to compete with the diverse native soil rhizobia in the field (Denton *et al.*, 2003; Batista *et al.*, 2015; Checcucci *et al.*, 2017; Irisarri *et al.*, 2019; Tartaglia *et al.*, 2019). This is likely because *Rlt* strains vary in ability to form nodules and fix nitrogen (Rys and Bonish, 1981; Vanlauwe *et al.*, 2019), and while commercial strains are suggested to generally be better nitrogen fixers, they are often found to be poor competitors for nodule occupancy compared to the adapted native soil rhizobia. Additionally, long-term growing pastures and self-regenerating crop rotations have shown the highest levels of competition compared to annually sown forages and crops where inoculants more often have the advantage of legume seeds being densely inoculated before sowing which provides a greater chance of symbiotic establishment (Sessitsch *et al.*, 2002).

To improve the efficiency of commercial rhizobia inoculants, strains must be able to outcompete indigenous strains, but also remain optimally compatible with the host legume for nitrogen fixation (Sessitsch *et al.*, 2002; Lupwayi, Clayton and Rice, 2006; Checcucci *et al.*, 2017). In order to achieve this, improving our understanding of the level of intraspecies diversity present within a rhizobia species and the potential mechanisms maintaining this diversity (e.g. through interactions with legume hosts, and interactions with other rhizobial strains) could be utilised to aid future strain selection (Checcucci *et al.*, 2017). Furthermore, by exploring the wide diversity of native soil rhizobia, this provides additional opportunities to isolate more strains matching desirable inoculant traits for development of improved inoculants (Lindström *et al.*, 2010; Batista *et al.*, 2015; Santos, Nogueira and Hungria, 2019).

1.3. Intraspecies diversity of *Rhizobium leguminosarum*

Multiple genetic and phenotypic traits have been used to identify and categorise strains of rhizobia into taxonomic groups (Pongsilip, 2012; Shamseldin, Abdelkhalek and Sadowsky, 2017). Rhizobial speciation of a broad variety of species and genotypes across genera has been enabled by rhizobia adaptation to diverse soil and plant niches which provide substantial opportunities for gene transfer (Provorov, Andronov and Onishchuk, 2017). However, despite the categorisation of strains into species there still exist a large amount of genetic diversity within rhizobia species (Fagerli and Svenning, 2005). In particular, *Rhizobium leguminosarum* contains a large group of strains and indigenous soil populations commonly display vast genetic diversity which is also reflected as variation in strains' gene expression and phenotypic properties.

1.3.1. Genetic classification of *Rhizobium leguminosarum*

Symbiotic nitrogen-fixers from legumes were classed into 6 *Rhizobium* species until the 1980s (*R. leguminosarum*, *R. trifolii*, *R. phaseoli*, *R. meliloti*, *R. japonicum* and *R. lupini*), and taxonomy of rhizobia has since largely developed and is still rapidly evolving through current research (Somasegaran and Hoben, 1985; Young and Haukka, 1996). By 2017, the number of defined rhizobia species reached 238 species across 18 genera and two clades (Shamseldin, Abdelkhalek and Sadowsky, 2017). *Rhizobium* and *Bradyrhizobium* remain the two largest genera of rhizobia, with *Rhizobium* itself containing 69 species able to infect distinct legume hosts (Shamseldin, Abdelkhalek and Sadowsky, 2017). The consistent fluctuations in taxonomic reassignment of rhizobial species is largely attributed to the development of molecular techniques used for species classification, from multilocus sequence typing (MLST) (Ribeiro *et al.*, 2009), sequence analysis of core genome markers (González *et al.*, 2019), average nucleotide identity between shared genomic regions (Kumar *et al.*, 2015; Rashid *et al.*, 2015; González *et al.*, 2019), and whole genome sequencing alignments (Ormeño-Orrillo *et al.*, 2015). Therefore, the realised diversity of rhizobia is continually increasing as research methods develop and as the agronomic importance of rhizobial diversity becomes more apparent for agricultural practices (Masson-Boivin *et al.*, 2009; Shamseldin, Abdelkhalek and Sadowsky, 2017).

Today, *Rhizobium leguminosarum* is classified as a bacterial species belonging to the *Rhizobiaceae* family and *Alphaproteobacteria* class (Stackebrandt, Murray and Truper, 1988). *Rhizobium leguminosarum* can be further subdivided into a species complex of genetically distinct sub-species called genospecies (Kumar *et al.*, 2015). The genospecies concept is used to define a group of bacterial genotypes that: 1) are genetically related through a common ancestor; 2) can undergo horizontal gene transfer and recombination within their group; 3) have diverged and subsequently restricted gene flow although not necessarily completely; and 4) but that are phenotypically heterogeneous and show no exclusive phenotype (Ravin, 1960, 1963). Five sympatric genospecies (gsA, gsB, gsC, gsD, and gsE) of *Rhizobium leguminosarum* have previously been identified from multiple soils globally, and seem to account for the majority of *Rhizobium leguminosarum* diversity in Northern Europe (Kumar *et al.*, 2015; Cavassim *et al.*, 2020). Additionally, two additional genospecies groups gsF-1 and gsF-2 were identified from European soil samples, which include *R. laguerreae* strains. However, their frequencies are generally much lower compared to other genospecies (Boivin *et al.*, 2020). These genospecies consistently display a clear average nucleotide identity (ANI) above 95% (equating to the 70% DNA-DNA hybridisation measure for species distinction) based on core genes and also from a standard whole-genome measure of ANI further confirming the definitive species distinction of *Rhizobium leguminosarum* genospecies (Kumar *et al.*, 2015; Rashid *et al.*, 2015; Cavassim *et al.*, 2020). Genospecies distinction is not evident using 16S gene markers as the 16S sequence can be too conserved for genospecies distinction. However, multiple housekeeping genes, or core gene markers, can be used to determine intraspecies diversity (Ramírez-Babena *et al.*, 2008; Ramírez-Bahena *et al.*, 2009; Shamseldin, Abdelkhalek and Sadowsky, 2017). Further sub-structuring in the *Rhizobium* species complex indicate a continuous evolutionary divergence of genomes leading to species divergence (Pérez Carrascal *et al.*, 2016; González *et al.*, 2019).

The species naming convention for rhizobia is determined based on genome organisation, whereby the species name is determined by the core genes, and the symbiovar is indicative of accessory genes important to the symbiotic function of the bacteria (Young, 2016). In this species-naming methodology, the nodulation genes on the symbiosis plasmid are useful for determining a strain’s symbiovar (sv.). In 1984, *Rhizobium trifolii*, *Rhizobium leguminosarum* and *Rhizobium phaseoli* species were reclassified into symbiovars of the *Rhizobium leguminosarum* species to become *R. leguminosarum* sv. *trifolli* (*Rlt*), sv. *viciae* (*Rlv*), and sv. *phaseoli*, respectively (Ramírez-Bahena *et al.*, 2009; Rogel, Ormeño-Orrillo and Martinez Romero, 2011).

The taxonomy of *Rhizobium* species is predominantly currently based on symbiotic compatibility with the host-legume and does not consider the expansive free-living phenotypic diversity, although modern taxonomic approaches (such as the *R. leguminosarum* genospecies classification) identify species using core genome phylogenies and pairwise strain ANI (Richter and Rosselló-Móra, 2009; Kumar *et al.*, 2015; Pérez Carrascal *et al.*, 2016). For example, *Rhizobium leguminosarum* genospecies are not necessarily symbiovar specific and are not defined by legume host range. As a result, a *Rhizobium leguminosarum* genospecies can contain strains that can be isolated from either pea or clover, at least for gsB, gsC and gsE (Kumar *et al.*, 2015; Boivin *et al.*, 2020). Furthermore, while strains of *viciae* and *trifolii* symbiovars do not share symbiosis genes, they are likely to share other accessory genes that are not host-specific, as demonstrated by their genospecies classifications (Kumar *et al.*, 2015).

1.3.2. Genetic variation of Rhizobium leguminosarum

1.3.2.1. Diversity within a small geographical range

The large and versatile genomes of rhizobia (<10.5Mbp) indicate the complexity of their life cycle transitioning from free-living to bacteroid physiologies and the heterogeneity of their ecological niche (MacLean, Finan and Sadowsky, 2007; Pini *et al.*, 2011; Sánchez-Cañizares *et al.*, 2018). For example, *Rlt* strain WSM1689 has a genome size of 6,903,379bp containing 6,709 protein-encoding genes, similar to *Rlt* strains WSM2304, WSM1325, CC275e, with the common inoculant strain TA1 containing a larger genome with around 8,493 protein-coding genes (Reeve, O'Hara, Chain, Ardley, Brau, Nandesena, Tiwari, Copeland, *et al.*, 2010; Reeve, O'Hara, Chain, Ardley, Brau, Nandesena, Tiwari, Malfatti, *et al.*, 2010; Reeve *et al.*, 2013; Terpolilli *et al.*, 2014; Delestre *et al.*, 2015). While the number of rhizobial genomes continues to increase in open source databases, few genomes are fully annotated to enable in depth genomic comparisons (Sánchez-Cañizares *et al.*, 2018). However, with the consistent development of next generation sequencing methodologies, the number of complete fully annotated rhizobial genomes is also increasing, and from this large scale analysis of species diversity is feasible (Cavassim *et al.*, 2020).

The level of intraspecies diversity is suggested to be correlated with the heterogeneity of a species' environment (Brockhurst *et al.*, 2019). The soil environment provides a multitude of opportunities for rhizobial species diversification through genetic transfer between rhizobia strains, as strains are likely to share overlapping niches and be in close proximity to other strains within diverse soil microbial communities (Kloesges *et al.*, 2011; Brockhurst *et al.*, 2019). Additionally, the large size of rhizobial genomes are suggested to be advantageous for competition under varied soil conditions because larger genomes enable strains to harbour multiple metabolic capabilities for accessing soil nutrients, providing a selective advantage in changing environments (Young *et al.*, 2006; Wielbo *et al.*, 2010).

Intraspecies diversity of *Rhizobium leguminosarum* is extensive, and several studies have shown that up to five distinct genospecies can occur sympatrically within a small plot of soil (Kumar *et al.*, 2015; Boivin *et al.*, 2020; Cavassim *et al.*, 2020). While the genetic composition of *Rhizobium leguminosarum* communities has been shown to differ at small geographic scales (Stefan *et al.*, 2018), distinct genotypes and genospecies have also been identified across the world from geographic regions with differing environmental conditions (Ramírez-Bahena *et al.*, 2009; Mauchline *et al.*, 2014; Kumar *et al.*, 2015; Cavassim *et al.*, 2020). Furthermore, these co-occurring genospecies can display different preferences for legume hosts, which is determined by their symbiosis genes (Mauchline *et al.*, 2014; Kumar *et al.*, 2015). It has been shown that a greater amount of intraspecies diversity is identified in the soil compared to from root nodules (Duodu et al., 2006). This suggests that the studies using legume-host trapping to determine rhizobium diversity have so far likely underestimated the level of intraspecies rhizobial diversity, even though some studies have still found high genetic diversity of *R. leguminosarum* from nodule populations (Stefan *et al.*, 2018).

1.3.2.2. Diversity at the global level: Pangenomes

Within a bacterial species, strains can differ genetically in the presence and absence of a substantial number of genes. The pangenome refers to all of the genes identified within a species. This can be further partitioned into the core genome, containing genes present in all strains, and the accessory genome, consisting of genes not present in all strains (Young *et al.*, 2006; Brockhurst *et al.*, 2019). Pangenomes occur due to the dynamic nature of prokaryotic genomes being able to gain genes from other bacterial species through horizontal gene transfer, and also readily lose genes if they confer a large fitness disadvantage (Brockhurst *et al.*, 2019). Species with large and long-term populations that are able to migrate to new niches are most likely to develop pangenomes as a consequence of adaptive evolution. This fits the lifestyle of rhizobia well as they spend the majority of their lives in heterogeneous soil environments and later navigate plant tissues for nodulation (McInerney, McNally and O'Connell, 2017).

The core and accessory genomes were defined to aid understanding intraspecies genetic variation, its origins, and the genomic structure of a species (McInerney, McNally and O'Connell, 2017). Core genes are generally essential genes with functions related to vital cell maintenance and are usually chromosomally encoded (although they can be found on plasmids) with a characteristically high G+C content (around 60% in *Rhizobium*) (Young *et al.*, 2006). Core genes tend to reflect the same phylogenies as 16S sequences for species, whereas accessory genes can move more freely between strains and tend to show diverging phylogenies that deviate from the accepted species tree (Young *et al.*, 2006). That being said, core genes can still display substantial intraspecies diversity at the sequence level (Wielbo *et al.*, 2010).

On the other hand, accessory genes are often attributed to more specialised adaptive functions, are located on auxiliary plasmids and chromosomal islands, and their G+C content can range from core-like to a lower G+C composition (Young *et al.*, 2006; Cavassim *et al.*, 2020). As a result, the accessory genome is suggested to influence strain adaptation to specific niches. While characteristic functional traits have been assigned to core and accessory genes, it is important to note that the majority of these are putative as most genes still have unknown functions. The accessory genome can be further subdivided; each genospecies contain a specific set of genospecies-exclusive genes which could confer genospecies-specific traits (Kumar *et al.*, 2015). Furthermore, the similarity of gene contents between *Rhizobium leguminosarum* strains has previously been found to cluster by genospecies, with additional underlying substructure based on whether strains were isolated from a similar geographic origin (Cavassim *et al.*, 2020). Additionally, there are often genes which have only ever been identified in one strain (ORFans), which are suggested either be recently acquired to the species, or be present unevenly across a species (Young *et al.*, 2006). Taken together, the accessory genome has maintained distinctive characteristics despite a long history of coexisting alongside core genes, and could reflect differences in gene transferability, mutation rate, or genomic location on either chromosomes or plasmids (Young *et al.*, 2006; Jiao *et al.*, 2018). As a result, the accessory genome is often evolving more rapidly than the core genome (Crossman *et al.*, 2008)

The size of the pangenome is known to vary widely amongst species, and the pangenome itself can be considered open or closed (Brockhurst *et al.*, 2019). Pangenome size has been suggested to be influenced by both effective population size and bacterial ability to migrate to new niches and environments (Kimes *et al.*, 2014; McInerney, McNally and O'Connell, 2017). The potential for expansion of the pangenome size is also dependent upon the diversity and size of the gene pool a strain is exposed to, which is likely to be larger and more diverse in spatially and temporally variable environments such as the soil (Brockhurst *et al.*, 2019). Open pangenomes characteristically contain a greater number of genes, of which a smaller proportion are considered core and the majority are accessory and gained through horizontal gene transfer (Brockhurst *et al.*, 2019). Closed pangenomes have a smaller number of genes and the majority are considered core and have a lower frequency of gene acquisition through horizontal gene transfer (Brockhurst *et al.*, 2019). As a result, species core genome sizes are known to range from totalling 3% to 84% of a species pangenome (McInerney, McNally and O'Connell, 2017). *Rhizobium leguminosarum* has an open pangenome, with a consistent core genome containing a large number of genes also shared by other rhizobia species, and an expansive accessory genome (Crossman *et al.*, 2008; González *et al.*, 2019). It was additionally shown from analysis of the *Rhizobium leguminosarum* genospecies complex that the total number of unique accessory genes increases indefinitely for the species when more genomes are continually added (Cavassim *et al.*, 2020). As an example of the pangenome diversity of *Rlt* (not considering ORFan genes), 196 strains were identified to have a stable core genome of 4,204 orthologous gene groups (19% of the pangenome), and 17,911 accessory orthologous gene groups (Cavassim *et al.*, 2020). Therefore, *Rhizobium leguminosarum* has a large open pangenome, which is likely to confer a diverse range of adaptive traits to enable survival and competition in the soil environment.

1.3.2.3. Plasmids, recombination and introgression

Rhizobium leguminosarum genomes are characteristically multipartite, whereby the genome is split across a chromosome and one or more plasmids (Young *et al.*, 2006;

Terpolilli *et al.*, 2014; diCenzo and Finan, 2017; Provorov, Andronov and Onishchuk, 2017; Zahran, 2017; Sánchez-Cañizares *et al.*, 2018). Commonly, the general term of a "replicon" is used to denote a DNA molecule within the genome architecture (diCenzo and Finan, 2017). Around 10% of bacterial genomes are divided into two or more replicons (Harrison *et al.*, 2010; diCenzo and Finan, 2017). Within this category, *Rhizobium leguminosarum* can contain multiple plasmids which can vary in number and size between strains, for example, common inoculant *Rlt* strain TA1 has 5 replicons, and *Rlv* strain WSM3841 has 65% genomic material in the chromosome with the remainder organised into six plasmids (Young *et al.*, 2006; Krol *et al.*, 2008). Additionally, analysis of 196 *Rlt* strains identified single strains containing up to 8 replicons (Cavassim *et al.*, 2020). Overall, the multipartite genome organisation has been suggested to be advantageous for enabling easier management of large genomes containing a diverse range of rhizobial functions while also keeping replication systems small to permit shorter generation times under heterogeneous and fluctuating environments (Zahran, 2017).

Some genomic features are known to differ between bacterial species, such as differences in G+C content, codon usage and dinucleotide (A-T, G-C pairs) relative abundance (diCenzo and Finan, 2017). The accessory genome has been shown to differ in many of these characteristics to the core genome, and consequently, it is assumed that many accessory genome components have been introduced through transfer of mobile plasmids and phages from other bacterial species, with alleles further incorporated onto the chromosome by recombination (Harrison and Brockhurst, 2012; Pérez Carrascal *et al.*, 2016; Brockhurst *et al.*, 2019). Therefore, it is common that replicons within a single multipartite genome also display distinction in these genomic characteristics. In order to categorise these differences, replicons have been classified into three general types: the chromosome, chromid and plasmid (diCenzo and Finan, 2017).

The chromosome predominantly contains the core genome functions, is the largest replicon in the genome, and is the most genetically stable (Harrison *et al.*, 2010; diCenzo and Finan, 2017; Zahran, 2017; Sánchez-Cañizares *et al.*, 2018). On the other hand, the extra-chromosomal plasmids usually encode the accessory genes, and in *Rhizobium leguminosarum* usually have a lower G+C content which indicates towards their foreign origin (Harrison *et al.*, 2010; Zahran, 2017; Sánchez-Cañizares *et al.*, 2018). The chromid is a replicon that is in between a chromosome and a plasmid. This is because the chromid can contain some core genes and its G+C content and codon usage is similar to the host chromosome, however they have plasmid replication systems (i.e. *repABC*) and the majority of genes are considered part of the accessory genome (Harrison *et al.*, 2010; Zahran, 2017).

The diversity of plasmids, including their number and size, varies between strains of *Rhizobium leguminosarum*. The *Rhizobium leguminosarum* pangenome has been shown to be largely diverse at both the level of the chromosome and plasmids (González *et al.*, 2019). However, in most instances, strains of *Rhizobium leguminosarum* display highly syntenic and colinear chromosomes, but have extensively non-uniform, highly varied plasmid profiles with a large amount of within-replicon 'mosaic structured' genetic diversity (Crossman *et al.*, 2008; Krol *et al.*, 2008; Wielbo *et al.*, 2010; Mazur *et al.*, 2011; Sánchez-Cañizares *et al.*, 2018). This can involve strains containing similar chromosomes but completely different symbiosis plasmids (Fagerli and Svenning, 2005; Kumar *et al.*, 2015; Sánchez-Cañizares *et al.*, 2018). Additionally, the high synteny of the chromosomes is suggested to reflect genome conservation at the species and genus levels. However, differences in chromosomal types have also previously been shown to be strongly associated with geographic origin (Fagerli and Svenning, 2005; diCenzo and Finan, 2017; Stefan *et al.*, 2018). The genetic conservation of chromids is almost as stable as chromosomes but has only been shown to sustain genetic similarity at the species level and not genus level (Harrison *et al.*, 2010). Across *Rhizobium leguminosarum* strains, plasmids are the most genetically variable, and numbers have ranged from 2-8 plasmids with sizes of approximately 200 kb to 1 Mb (Krol *et al.*, 2008; Wielbo *et al.*, 2010; Provorov, Andronov and Onishchuk, 2017; Cavassim *et al.*, 2020). However, analysis of 196 Rlt strains demonstrated while up to 20 distinct *repABC* sequence group plasmid families could be identified, eight of those plasmid types accounted for the majority of plasmids identified (Cavassim *et al.*, 2020). This is also supported by another study which found that although extra-chromosomal replicons showed significant diversity, only a few replicon families were identified overall (González *et al.*, 2019). Therefore, the accessory genome which is predominantly encoded in the plasmids largely confers the genomic traits and phenotypic capabilities that are strain-specific and is also largely influenced by the significant genetic diversity within plasmid families.

Studies have also suggested that chromosomal recombination is rare within *Rhizobium leguminosarum* symbiovars and recombination of core genes predominantly occurs within species boundaries (Harrison, Jones and Young, 1989; Kumar *et al.*, 2015). However, plasmid diversity is not explained by *Rhizobium leguminosarum* genospecies complex, as plasmids are not exclusive to a single genospecies, although some plasmids are more overrepresented in some genospecies (Cavassim *et al.*, 2020). Additionally, symbiosis genes have been found to be encoded by different plasmid families within the *Rhizobium leguminosarum* genospecies complex, and these different symbiosis plasmids can co-exist in the same geographic location (Cavassim *et al.*, 2020). This diversity of symbiosis plasmid types has also been found in other studies (Black *et al.*, 2012). Some of these plasmid families have additionally been found to contain conjugal transfer proteins, suggesting that there could be different methods and rates of both plasmid and symbiosis plasmid transfer within *Rhizobium leguminosarum* (Cavassim *et al.*, 2020). Consequently, phylogenetic analysis indicates symbiosis plasmids have crossed genospecies boundaries (Cavassim *et al.*, 2020). The inter-strain transfer of the symbiosis plasmids and genes has also been identified in other previous investigations (Kumar *et al.*, 2015; González *et al.*, 2019). However, diversification of plasmid profiles due to conjugal plasmid transfer is likely somewhat restricted by geographic location as bacterial cells must be within close proximity for transfer to occur, although soil pH has also been suggested to contribute to plasmid profile composition (Ramírez-Bahena *et al.*, 2009; Stefan *et al.*, 2018). Overall, symbiosis plasmids are suggested to display the highest recombination rates within rhizobia populations and between species, followed by accessory plasmids, and core chromosomes display the lowest recombination rates (Carrascal et al., 2019).

Overall, horizontal gene transfer of plasmids and genetic elements facilitates acquisition of adaptive genes from closely related and distantly related strains and thereby increasing bacterial inter- and intraspecies diversity (Wiedenbeck and Cohan, 2011). In this way, horizontal gene transfer can dissociate phenotypic traits from their inferred species and enable overlapping of environmental niches between distantly related strains, as demonstrated by the transfer of symbiosis plasmids between species (Kumar *et al*., 2015; Pérez Carrascal *et al*., 2016; González et al., 2019; Cavassim *et al*., 2020). Analysis of *Rlt* strains identified that genes can travel across genetically distinct genospecies boundaries, however nearly all genes otherwise showed no evidence of introgression and those that did were predominantly plasmid-localised (Cavassim *et al.*, 2020). Consequently, even though genospecies have evolved to display defined boundaries of recombination within their species boundaries, rarer introgression events can still introduce adaptive genes from more distantly related species and provide unique phenotypes at the individual strain level (Kumar *et al.*, 2015; Pérez Carrascal *et al.*, 2016).

1.3.2.4. Operons

In prokaryotic genomes, genes within replicons are commonly thought to be organised into operons. An operon is a group of genes arranged consecutively along the genome and co-directionally transcribed by a common promoter and terminator. The first defined classical operon is the well-studied *E. coli lac* operon, which consists of genes required for the transportation and metabolism of lactose sugars (Jacob and Monod, 1961). Therefore, it is thought that genes that share functions in related cellular pathways are arranged into non-random operon units for efficient co-expression via co-transcription into a single stand of polycistronic mRNA (Jacob and Monod, 1961; Wolf *et al.*, 2001; De Hoon *et al.*, 2004; Koonin, 2009; Osbourn and Field, 2009). However, some operons are suggested to contain genes from different functional pathways but are grouped into operons because they are required under the same environmental conditions (Osbourn and Field, 2009). Operon structures have also been shown to be dynamic and altered by environmental influences (Okuda *et al.*, 2007; Osbourn and Field, 2009; Fortino *et al.*, 2014). Some operons can be subdivided into multiple transcriptional units with their own internal promoter and terminators that are regulated differently depending on the external stimuli (Okuda *et al.*, 2007). Therefore, the environmental context of the species in question is likely to substantially influence operon organisation and diversity in prokaryotic genomes.

The evolutionary persistence of operon structures across bacteria is debated by different theories (Lawrence, 1999; Rocha, 2008). The Selfish Operon Model is currently the most accepted theory, which rationalises that operon gene organisation is maintained in prokaryotes because genes required for a selectable phenotype can be transferred by both horizontal co-transfer and vertical transmission and are consequently maintained due to their close proximity (Lawrence, 1999; Koonin, 2009; Osbourn and Field, 2009). Gene clustering by operon organisation is therefore considered advantageous to the constituent genes, instead of solely due to the importance of functional coregulation for the organism itself (Lawrence, 1999).

The coverage of operon organisation within genomes significantly differs between bacterial species (Wolf *et al.*, 2001; Koonin, 2009). Little is known regarding the difference in operon organisation between rhizobia species and considering the dynamic nature of their large multipartite genomes this could be vast. This is especially considering initial comparisons of bacterial genome operon organisation revealed low conservation of gene order beyond the extent of operons, which was further confirmed by variability in gene order across the *Rhizobium leguminosarum* species complex (Koonin, 2009; Cavassim et al., 2020). Despite the limited investigation of operon organisation diversity in rhizobia, many individual common rhizobial operons have been extensively studied previously, including: nodulation gene operons which are vital for symbiotic establishment (*nodABCIJ* and *nodEF*) (Herman P. Spaink *et al.*, 1987; Hong, Burn and Johnston, 1987a); operons involved in metabolism and nutrient acquisition (Yeoman *et al.*, 1997; Poole *et al.*, 1999); and operons involved in the rhizosphere and quorum-sensing such as the *tra-trb* operon system used for conjugational transfer of the symbiosis plasmid (Wisniewski-Dyé and Downie, 2002; Danino *et al.*, 2003). Taken together, operon structure and function can be largely diverse between strains and should be considered with both environmental context and bacterial species in mind.

1.3.3. Transcriptomic variation of *Rhizobium leguminosarum*

Gene expression is measured in abundance of transcribed mRNA, and in most instances is highly correlated to protein levels. Therefore, expression levels are considered a good proxy for identifying different phenotypic responses from an organism exposed to different environmental conditions, and conversely, understanding expression differences between different organisms grown in the same environment. Analysis of gene expression can highlight ecologically crucial phenotypes that have previously been difficult to define and measure because differences are not translated into morphologically distinct phenotypes (Pavey *et al.*, 2010). Consequently, transcriptome profiling has become an insightful tool for investigating phenotypic differences among organisms, and developments in microarray and RNA-Seq technologies have advanced transcriptomic analysis capabilities enabling affordable wider-scope gene expression investigations (MacLean, Finan and Sadowsky, 2007; Wang, Gerstein and Snyder, 2009; Yoder-Himes *et al.*, 2009; Filiatrault, 2011; Peng *et al.*, 2014; Jiménez-Guerrero *et al.*, 2017; diCenzo *et al.*, 2019)

Both inter- and intra-species transcriptomic variation investigations have predominantly been driven by the most genetically well-characterised bacterial species, including *Escherichia coli*, *Staphylococcus aureus*, *Pseudomonas aeruginosa* and *Salmonella enterica* (Carrasco, Tan and Duman, 2011; Zarrineh *et al.*, 2014; Hosseinkhan *et al.*, 2015; Vital *et al.*, 2015; Hosseinkhan, Mousavian and Masoudi-Nejad, 2018; Hornischer *et al.*, 2019). To identify transcriptomic differences between bacterial strains at an inter- and intra-species level, previous prokaryotic analyses have predominantly used one or two isolates, and a maximum of 4, as the representative strains of a species' transcriptome profile (Scaria *et al.*, 2013; Kimes *et al.*, 2014; González-Torres *et al.*, 2015; Vital *et al.*, 2015; Connolly *et al.*, 2019). Despite only using a few strains, variation in gene expression is evident at the intraspecies level for core and accessory genes (Scaria et al., 2013; González-Torres et al., 2015; Vital et al., 2015; Connolly et al., 2019). Consequently, it is anticipated that the large genomic diversity of rhizobia species would to some extent reflect variation across individual strain phenotypes.

Rhizobia have been extensively utilised for investigating transcription variation in bacteria, for several reasons. The substantial number of full sequenced genomes available in sequencing repositories and the elaborateness of their multipartite genomes (Young *et al.*, 2006; diCenzo and Finan, 2017) facilitates interesting investigations of transcriptional regulation. The additional complexity of their lifestyle involving transition between a motile free-living soil form and a non-motile symbiotic bacteroid form is a physiological transformation associated with significant alterations to gene expression, and as a result has received a great amount of focus in rhizobial transcriptomics (Yoder-Himes *et al.*, 2009; Vercruysse *et al.*, 2011; Lopez-Leal *et al.*, 2014; diCenzo *et al.*, 2019). Studies have principally focused on rhizobial transcriptomic responses to altered nutrient resources, symbiosis development across different hosts (Karunakaran *et al.*, 2009; Ramachandran *et al.*, 2011; Krysciak *et al.*, 2014; Peng *et al.*, 2014; Roux *et al.*, 2014; Perez-Montano *et al.*, 2016; Green *et* *al.*, 2019), free-living versus bacteroid physiologies (Yoder-Himes *et al.*, 2009; Vercruysse *et al.*, 2011; Lopez-Leal *et al.*, 2014) and responses to known stress conditions (Vercruysse *et al.*, 2011; Liu *et al.*, 2014; Lopez-Leal *et al.*, 2014). These investigations have improved our understanding of important functional and regulatory interactions of genes involved in quorum-sensing, symbiotic establishment and metabolism in rhizobia. However, there has been limited investigation into the extent of transcriptional variation of rhizobia at either the inter- or intra-species level, and only a few studies have undertaken direct transcriptome comparisons between rhizobia strains differing in core and accessory genome organisation (Heath, Burke and Stinchcombe, 2012; Galardini *et al.*, 2015; Rachwal, Matczynska and Janczarek, 2015; Jiao *et al.*, 2018; Green *et al.*, 2019). Genetic diversity is clearly an important factor in transcriptomic and phenotypic variation in symbiosis, as it has been shown that *Sinorhizobium meliloti* gene expression in symbiotic establishment varies between strains and is also dependent on the interaction with specific plant genotypes (Heath, Burke and Stinchcombe, 2012).

One recent study investigated how rhizobial transcriptional profiles were affected by the differing organisation of core and accessory genes in two *Sinorhizobium fredii* strains (Jiao *et al.*, 2018). Analysis of accessory genes was enabled by sub-setting the genes in the genomes based on their frequency in ten published *Sinorhizobium* genomes (Jiao *et al.*, 2018). The analysis identified that while expression of core genes were similar in both strains across growth phases and symbiotic conditions, intraspecies accessory genes displayed larger variations in expression and could contribute to rhizobia diversification (Jiao *et al.*, 2018). Additionally, the study highlighted the relevance of genomic architecture for gene incorporation into replicon regulatory networks. A large amount of between-replicon co-regulation of genes was found to occur, and importantly the symbiosis plasmid was found to display more between-replicon gene co-expression than within-replicon gene co-expression (Jiao *et al.*, 2018). As it is known that the genomic architecture of *Rhizobium leguminosarum* replicons is diverse, this could have profound influences on the intraspecies transcriptional variation between strains as a result in differences in gene connectivity. Furthermore, it has already been shown from analysis of 51 *Sinorhizobium meliloti* strain pangenome that significant variation in gene connectivity can occur at the intraspecies level largely as a result of accessory genome content variability (Galardini *et al.*, 2015).

Variation in intraspecies-level gene expression has also been suggested to promote bacterial speciation (Pavey *et al.*, 2010). This is suggested to be achieved by expression variation enabling populations to colonise new ecological niches where regulation of expression could then become vital for population persistence in the new niche, which could lead to species diversification and potential reproductive isolation (Pavey *et al.*, 2010; Ng *et al.*, 2019). The extent to which this has contributed to the diversification of *Rhizobium leguminosarum* genetic and phenotypic variation is unknown and consequently requires further investigation.

1.3.4. Phenotypic variation of *Rhizobium leguminosarum*

Despite the high level of observed genotypic diversity and advancements in genetic analyses, a significant proportion of genes are still annotated with putative, unknown or hypothetical functions (Sánchez-Cañizares *et al.*, 2018). Consequently, in order to fully understand and confirm the functional capacities of strains there is a need for direct phenotypic investigations.

Reflecting their genetic variation, rhizobia are phenotypically diverse. *Rhizobium leguminosarum* is characterised as a gram negative, motile, non-spore forming, rod-shaped bacterium that grows into white colonies after 3-5 days growth. The bacteria grow optimally at 28°C but have a temperature range between 10 – 35°C. Additionally, *Rhizobium leguminosarum* grows optimally at pH 7.75, but is collectively as a species able to grow within a pH range of 5-8.5 (Ramírez-Bahena *et al.*, 2009; Reeve, O'Hara, Chain, Ardley, Brau, Nandesena, Tiwari, Copeland, *et al.*, 2010; Mazur *et al.*, 2013; Delestre *et al.*, 2015; Howieson and Dilworth, 2016).

Rhizobia strains can dwell in the soil for years in between opportunities for symbiotic interaction, and are suggested to display a versatile, metabolically active state in the soil between these periods (Young *et al.*, 2006). Similarly, it is anticipated that rhizobia are able to metabolise a diverse variety of compounds considering they are found in various complex environments from the soil to inside plant cells where many of the compounds are unknown (Mazur *et al.*, 2013; Ormeño-Orrillo and

Martínez-Romero, 2013). *Rhizobium leguminosarum* strains grow at different rates, and this could be a result of the different metabolic capabilities of strains (Wielbo *et al.*, 2010; Mazur *et al.*, 2013). A high proportion of field isolates have been described as metabolically versatile (i.e. they have not specialised to utilize only one particular type of substrate), and there is a wide variation in the number of substrates that individual strains can metabolise (Wielbo *et al.*, 2010; Mazur *et al.*, 2013). For example, in one study, the lowest number of substrates utilised by a strain was 113 (Mazur *et al.*, 2013). Strains of *Rhizobium leguminosarum* were shown to vary predominantly in their ability to metabolise sugar compounds, and secondly acids and amino acids (Wielbo *et al.*, 2010; Mazur *et al.*, 2013). Polysaccharides, sugar acids and D-amino were shown to be metabolised the least by strains (Mazur *et al.*, 2013). It has been suggested that metabolic versatility is associated with replicon diversity, with vital metabolic functions located on the chromosome and metabolic capabilities providing resilience in fluctuating heterogeneous environments encoded by plasmids (Mazur *et al.*, 2013; Ormeño-Orrillo and Martínez-Romero, 2013). Furthermore, strains with no large replicons were shown to use a significantly lower number of monosaccharides and oligosaccharides but a higher utilisation of sugar acids, modified carboxylic acids, and nitrogen compared to strains with large plasmids (Mazur *et al.*, 2013). Similar to the varied gene and plasmid distribution across genospecies, no metabolic capability or substrate utilisation profile has been found exclusive to a single genospecies or other defined genotype group (Wielbo *et al.*, 2010; Kumar *et al.*, 2015). The ability to be metabolically versatile is likely to be an advantageous long-term approach for soil survival, colonisation and adaptation as strains can metabolise more diverse plant-secreted compounds and nutrients available in the soil rhizosphere soil (Wielbo *et al.*, 2010; Mazur *et al.*, 2013). Moreover, it has even been proposed that *Rhizobium* metabolic capacity and diversity is underestimated due to the large number of unknown gene functions which may have functional associations to metabolism of soil and plant compounds (Ormeño-Orrillo and Martínez-Romero, 2013).

Rhizobium leguminosarum strains also vary in their symbiotic capacity. While *Rhizobium leguminosarum* strains differ by which legume hosts they infect (sv. *trifolii*, *viciae* and *phaseoli*), symbiovars display differences in their response to legume flavonoids, competitive ability for nodulation and effectiveness for nitrogen fixation (Leung, Wanjage and Bottomley, 1994; Ramírez-Bahena et al., 2009; Maj et al., 2010;

Wielbo et al., 2011; Bourion et al., 2018). Furthermore, the legume host species that *Rhizobium leguminosarum* strains form symbiosis with are not explained by genospecies classification, as strains of the same genospecies have been shown to form specific symbioses with either clover (sv. *trifolii* strains) or pea and faba bean (sv. *viciae* strains) (Kumar *et al.*, 2015; Boivin *et al.*, 2020). Increased metabolic versatility was shown to not be advantageous for competitive nodulation with clover, however strains at low frequency in the soil population with a specialised metabolism were suggested to be more symbiotically effective (Wielbo *et al.*, 2010). In addition, the types of secondary metabolites produced by *Rhizobium leguminosarum* are also diverse at the intraspecies level. For example, the number and combinations of quorum-sensing pathways, such as *cinI/cinR, rhiI/rhiR, traI/traR* and *raiI/raiR*, contribute to the diversity of strain interactions with neighbouring soil microbes and legumes (Wisniewski-Dyé and Downie, 2002; Sanchez-Contreras *et al.*, 2007).

Although phenotypic diversity is high within *Rhizobium leguminosarum*, some previous studies have found limited association to genetic diversity of sequence types and taxonomic classifications (Kumar *et al.*, 2015; Stefan *et al.*, 2018). Currently, there are no phenotypes that are exclusive to a single genospecies, although it is not yet known the extent to which genospecies boundaries are associated with differences in symbiotic capabilities (Kumar *et al.*, 2015; Boivin *et al.*, 2020). Therefore, understanding the connection between genotype and phenotype is paramount to understanding both intraspecies diversity and symbiotic potential within *Rhizobium leguminosarum,* and is most likely dependent on the environmental context and relevant organism interactions.

1.4. Significance of intraspecies diversity in the rhizosphere

The rhizosphere is the section of soil closest to the plant root where interactions between soil microorganisms can influence plant growth, plant health, resilience to environmental stresses, competition for resources and nutrient cycling (Philippot *et al.*, 2013; Jones *et al.*, 2019). In addition to direct influences from plant root secretions, the interspecies and intraspecies interactions between microorganisms and the plant root are dynamic. Therefore, understanding the importance and specificity of these interactions is crucial for refining and improving crop production (Philippot *et al.*, 2013; Liu *et al.*, 2019). The sequencing of genomes from soil and rhizosphere environments has demonstrated the extensiveness of rhizobial species pangenomes and the subsequent metabolic diversity of rhizobia populations (Poole, Ramachandran and Terpolilli, 2018). This rhizobial intraspecies diversity can lead to variation in the potential interspecies and intraspecies interactions between strains and also affect rhizobial interactions with the host legume.

Rhizobial intraspecies diversity in the soil rhizosphere can be affected by a number of factors including interspecies and intraspecies plant variation (Kiers and Denison, 2008; Miranda-Sánchez, Rivera and Vinuesa, 2016; Kroll, Agler and Kemen, 2017; Vuong, Thrall and Barrett, 2017; Clúa *et al.*, 2018), abiotic soil factors (Rice, Penney and Nyborg, 1977; Harrison, Jones and Young, 1989; Xiong *et al.*, 2017; Igiehon and Babalola, 2018; Liu *et al.*, 2019), and different agricultural management practices such as organic or conventional farming (Kiers, West and Denison, 2002; Lupwayi, Clayton and Rice, 2006; Shu *et al.*, 2012; Weese *et al.*, 2015). In addition, microbial factors such as the interspecific competition with other soil bacteria, and intraspecific competition with other rhizobia strains for nutrient resources and nodulation, can affect intraspecies diversity (Pugashetti, Angle and Wagner, 1982; Villacieros *et al.*, 2003; Denison and Kiers, 2004; Kiers and Denison, 2008; Blanco, Sicardi and Frioni, 2010; Hibbing *et al.*, 2010; Wielbo *et al.*, 2011; Barrett *et al.*, 2015; Teng *et al.*, 2015; Lu *et al.*, 2017). As an example, rhizobia can indirectly interact through competition for nutrient resources. These indirect competitive interactions, where some strains more effectively metabolise a resource which limits its availability for other strains is one method that could suppress growth of niche-sharing strains and subsequently could reduce symbiont diversity within a community (Ramachandran *et al.*, 2011; Becker *et al.*, 2012). Siderophores that sequester iron can also be used by rhizobia as a resource competition mechanism to inhibit growth of competitor strains (Joshi *et al.*, 2008; diCenzo *et al.*, 2014; Kramer, Özkaya and Kümmerli, 2019). In the rhizosphere, there is a high likelihood of overlapping resource utilisation between strains which provides many opportunities for interference and cheating interactions to occur within communities (Jousset *et al.*, 2011; Barrett *et al.*, 2015). Additionally, quorum sensing is a powerful direct signalling mechanism between strains of bacteria, capable of modulating growth through regulation of gene expression by intra- and interspecies communication in a cell-density dependent manner (Miller and Bassler, 2001; Wisniewski-Dyé and Downie, 2002; Gonzalez and Marketon, 2003; Checcucci *et al.*, 2017). *N*-acyl homoserine lactones (AHLs) are the most commonly identified quorum sensing molecules, and *Rhizobium* species produce the largest diversity of AHLs among soil bacteria (Cha *et al.*, 1998; Wisniewski-Dyé and Downie, 2002). Quorum sensing is an important aspect of symbiotic establishment, as this signalling mechanism between cells can modulate the rhizosphere community interactions and increases bacterial densities around the root surface in preparation for nodule development (Schwinghamer and Brockwell, 1978; Miller and Bassler, 2001; Wisniewski-Dyé and Downie, 2002; Gonzalez and Marketon, 2003; Downie, 2010). This signalling capability varies between strains, but strains with this ability can use it to their competitive advantage to co-ordinate bacterial competitor interactions with plant hosts (Gonzalez and Marketon, 2003).

Intraspecies diversity of rhizobia is important in the rhizosphere because symbiont community diversity and the associated competitive interactions could potentially influence symbiont function and effectiveness of legume-rhizobia symbiosis by determining which strains form symbiosis with the legume (Hibbing *et al.*, 2010; Bolnick *et al.*, 2011; Barrett *et al.*, 2015; Pahua *et al.*, 2018; Liu *et al.*, 2019). Increased intraspecies diversity can be beneficial by providing legumes with more opportunity to form symbiosis with many strains, however this increased diversity might also lead to a prevalence of cheating behaviours and antagonistic intraspecies interactions within the population which are detrimental to plant growth (Becker *et al.*, 2012; Barrett *et al.*, 2015). For example, the success of rhizobial inoculant can be limited by competition with indigenous soil rhizobia that can outcompete inoculants for nodule occupancy but are less effective nitrogen fixers themselves (Berg *et al.*, 1988; Triplett and Sadowsky, 1992; Blanco, Sicardi and Frioni, 2010).

Taken together, to develop methods for increasing plant productivity by optimising the legume-rhizobia symbiosis it is crucial to gain an understanding of the importance of intraspecies rhizobial diversity and competitive interactions which may explain why specific strains form symbiosis with the legume over others (Barrett *et al.*, 2015; Pahua *et al.*, 2018; Liu *et al.*, 2019).

1.5. *Rhizobium* strains of interest in this study

This project concentrated on *Rhizobium leguminosarum* symbiovar *trifolii* strains that form symbiosis with white clover. Strains used within this study were selected from a 196 *Rlt* genome-sequenced isolate collection generated by the NCHAIN consortium (Cavassim *et al.*, 2019, 2020). Strains were isolated from white clover root nodules collected from DLF Trifolium conventional breeding trial sites in the UK (32 isolates), Denmark (43 isolates), France (40 isolates), and 50 organic fields across Denmark (81 isolates) (Figure 1.1). Clover roots were sampled from 40 plots within each conventional trial site, and in total 170 plots were sampled overall. The 196 strains all have respective Illumina sequenced whole-genome assemblies and additionally 8 strains were also re-sequenced with PacBio technology (Pacific Biosciences of California, USA) (Cavassim *et al.*, 2020).

The 196 strains are categorised into five genetically distinct *Rhizobium leguminosarum* genospecies (gsA, gsB, gsC, gsD and gsE) (Kumar *et al.*, 2015). These genospecies were determined previously by constructing a phylogeny from *rpoB* gene sequences using known genospecies strain representatives in addition to the 196 strains. Genospecies classification was then determined for all 196 strains based on their relative positions within the phylogeny compared to the known representatives (Cavassim *et al.*, 2020). Pairwise average nucleotide identity (ANI) based on 6,529 genes present in at least 100 strains demonstrated that strains clustered by genospecies, with some additional substructure determined by geographic origin (Cavassim *et al.*, 2020).

Therefore, these *Rlt* strains were used in this study because it enabled analysis of strains that were both genetically distinct and geographically different, with the exception of gsA and gsB strains, which were exclusively isolated from Danish organic sites and a UK conventional site, respectively (Table 1.1). Additionally, all strains were similar in that they could form symbiosis with white clover genotypes.

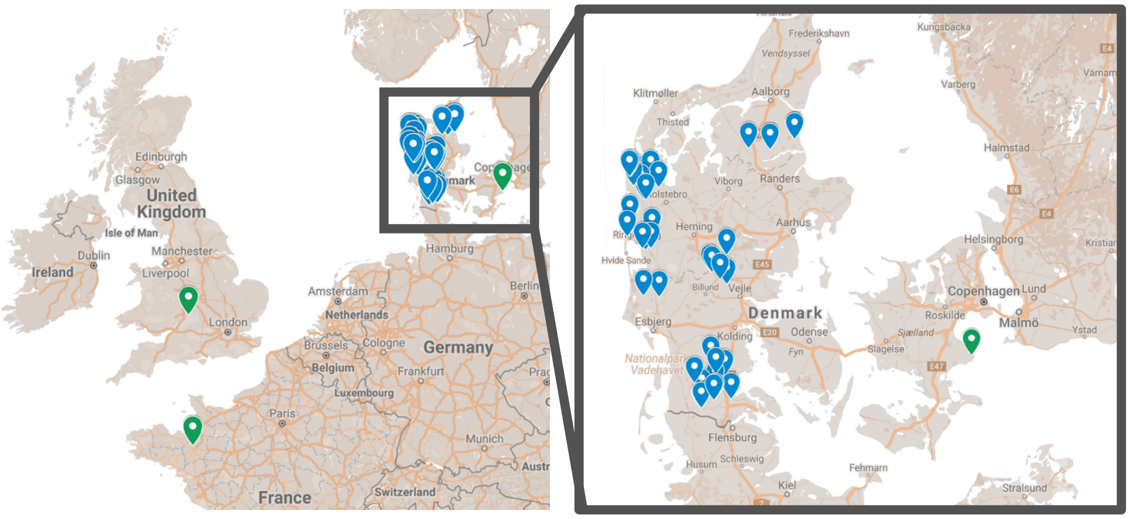


Figure 1.1 Field sites sampled across the UK, France and Denmark. Green pins represent the locations of DLF Trifolium conventional breeding trial sites, and blue pins represent the organic field sites.

Table 1.1 The dataset of 196 *Rhizobium leguminosarum* symbiovar *trifolii* strains used in this study. Strains are grouped by geographic origin and genospecies classification.

Geographic Origin	Genospecies					**Total**
	A	B	C	D	E	
UK conventional farms	-	32	-	-	-	**32**
France conventional farms	-	-	40	-	-	**40**
Denmark conventional farms	-	-	30	4	9	**43**
Denmark organic farms	32	-	46	1	2	**81**
Total	**32**	**32**	**116**	**5**	**11**	**196**

1.6. Project Background

1.6.1. Project Aims and Objectives

This project is part of a large research consortium effort called NCHAIN, led by Aarhus University, Denmark, and includes both academic and industrial collaborators. NCHAIN aims to improve white clover and grass mixture production for animal feed on organic farms through development of a quantitative model of nitrogen transfer from rhizobia to clover to grass, based on genetic data from these three organisms. One aim of the NCHAIN consortium is to identify optimal *Rlt* – white clover genotype partnerships that can be used to increase crop yields on agricultural land. Multiple studies have proposed that selection of optimal rhizobia should consider legume cultivar, environment, and soil microbiota (Bolnick *et al.*, 2011; Busby *et al.*, 2017; Pahua *et al.*, 2018; diCenzo *et al.*, 2019), thereby suggesting that intraspecific variations in rhizobial interactions are important factors determining inoculant success. This thesis project focuses on understanding the extent and importance of intraspecies diversity in rhizobial genetics and phenotypic interactions associated with the *Rlt* – white clover symbiosis.

Questions that currently remain unanswered include:

- what mechanisms maintain *Rhizobium* intraspecies genetic diversity;
- how intraspecies genetic diversity translates transcriptionally and phenotypically;
- whether *Rhizobium leguminosarum* genetic diversity determines intraspecific interactions between strains.

Therefore, the overall purpose of this PhD project was to determine the extent of intraspecies diversity of *Rlt* at the genetic and phenotypic levels, particularly with regard to identifying functional differences between *Rhizobium leguminosarum* genospecies. More specifically the objectives were to:

1) Determine if the diversity of *Rlt* populations can be explained by the selective differences of white clover genotypes;
2) understand if *Rlt* genetic diversity manifests itself in the gene expression profiles and growth phenotypes of strains between and within genospecies;
3) identify whether intraspecific *Rlt* interactions can be determined by genetic differences between genospecies and environmental origins of strains.

1.6.2. Thesis chapter outline

This thesis includes the following chapters, presented in the form of research papers:

Chapter 2: MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction

In this chapter, a multiplexed High Throughput Amplicon Sequencing method was developed and validated for characterising intraspecies diversity of DNA samples. The method, named MAUI-seq, uses unique molecular identifiers to improve sequencing error correction by eliminating chimeric and other erroneous reads. MAUI-seq was validated with white clover nodule DNA samples and by comparing its error correction results to alternative known methods, DADA2 and UNOISE3.
This chapter is currently under peer review in Molecular Ecology Resources.

Chapter 3: *Rhizobium* nodule diversity is determined by both clover host genotype and local growth conditions

This chapter utilised the MAUI-seq amplicon sequencing pipeline to compare how symbiotic selection by five different white clover genotypes affected *Rlt* nodule community diversity under field conditions. *Rlt* diversity was determined based on the allelic diversity of two chromosomal housekeeping genes, *rpoB* and *recA*, and two auxiliary plasmid-bound symbiosis genes, *nodA* and *nodD*.

Chapter 4: *Rhizobium leguminosarum* symbiovar *trifolii* sub-species display distinct intraspecies transcriptomic variation

In this chapter, transcriptional differences between and within *Rlt* genospecies were investigated, to gain an insight into how genetic distance is associated with gene expression patterns. In total, 79 *Rlt* strains were grown under the same *in vitro* conditions and transcriptome profiles were evaluated for fundamental core gene expression differences. Transcriptional differences between genospecies were further associated to phenotypic and putative metabolic traits in order to determine functional differences between genospecies.

Chapter 5: Identifying conserved operonic transcriptional units in *Rhizobium leguminosarum* symbiovar *trifolii* genospecies

This chapter used genome and transcriptome data from 26 *Rlt* strains to identify transcriptional units (putative operons) conserved at the genospecies- and *Rhizobium leguminosarum* species-level. Multiple parameters were used to define conserved transcriptional units for each genospecies, including; gene ortholog group classification, adjacent gene pair identification; mean intergenic distance calculations; and detection of gene co-expression using correlation coefficients and expression deviance scores. Species-conserved transcriptional units were identified by cross comparing genospecies-conserved transcriptional units. The pipeline was further validated by determining if known *Rlt* operons were identified using the method. This study generated a database of putative operons for *Rhizobium leguminosarum*.

Chapter 6: Competitive rhizobial intraspecies interactions are genospecies specific

In this chapter, variation of pairwise intraspecies competitive interactions between 24 *Rlt* strains was investigated. This was undertaken to determine whether interactions were predictable based on genetic background and environmental origin. Pairwise competitive interactions were determined *in vitro* in two ways: 1) indirectly, mediated by interactions with cell-free supernatants; and 2) directly, by observing growth inhibition when strains were grown on spot agar plate assays. To identify potential underlying competition mechanisms, comparative genomics was used to determine differences in strain metabolic capacities and presence of genes associated with quorum-sensing, bacteriocins, secondary metabolites and prophages.

Chapter 7: General Discussion

An overview of the project is discussed in the context of answering the three main project research questions. The potential for future research directions based on the presented work are suggested.

The methods used for each chapter are outlined within the respective chapters. References are provided at the end of the thesis. The supplementary information for each chapter is shown at the end of the thesis in separate chapter Appendices (A-E), and in specified Additional Files as Accompanying Material.

Chapter 2. MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction

Authors:

Bryden Fields[1*] and Sara Moeskjær[2*], Ville-Petri Friman[1], Stig U. Andersen[2], and J. Peter W. Young[1]

*: These authors contributed equally to the work.

Author affiliations:

[1]Department of Biology, University of York, York, United Kingdom

[2]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

Acknowledgements

We thank David Sherlock for his experimental expertise in developing this method, the University of York Technology Facility for sequencing, Simon Kelly for DADA2 expertise, Asger Bachmann, Terry Mun, M. Izabel A. Cavassim, and Marni Tausen for preliminary data analysis and script development, and DLF for access to their clover field trials. This work was funded by grant no. 4105-00007A from Innovation Fund Denmark (S.U.A.). Initial development of the method was funded by the EU FP7-KBBE project LEGATO (J.P.W.Y).

Author contributions

Conceptualization: JPWY; Methodology: JPWY, SUA; Software: BF, SM, JPWY; Validation: BF, SM, JPWY, SUA; Formal analysis: BF, SM, JPWY; Investigation: BF, SM; Resources: JPWY, SUA, VPF; Data curation: BF, SM, JPWY; Writing - original draft: BF, SM; Writing - review and editing: BF, SM, JPWY, SUA, VPF; Visualisation: BF, SM; Supervision: JPWY, SUA, VPF; Project administration: JPWY, SUA; Funding acquisition: JPWY, SUA.

2.1. Abstract

Background: Sequencing and PCR errors are a major challenge when characterising genetic diversity using high-throughput amplicon sequencing (HTAS).

Results: We have developed a multiplexed HTAS method, MAUI-seq, which uses unique molecular identifiers (UMIs) to improve error correction by exploiting variation among sequences associated with a single UMI. We show that two main advantages of this approach are efficient elimination of chimeric and other erroneous reads, outperforming DADA2 and UNOISE3, and the ability to confidently recognise genuine alleles that are present at low abundance or resemble chimeras.

Conclusions: The method provides sensitive and flexible profiling of diversity and is readily adaptable to most HTAS applications, including microbial 16S rRNA profiling and metabarcoding of environmental DNA.

Keywords:

Metabarcoding, High-throughput amplicon sequencing, Error correction, Chimeric amplicons, Amplicon sequence variant

2.2. Introduction

The evaluation of DNA diversity in environmental samples has become a pivotal approach in microbial ecology (Birtel *et al.*, 2015) and is increasingly also used to assess the distribution of larger organisms (Deiner *et al.*, 2017). If a core gene can be amplified from environmental DNA with universal primers, the relative abundance of species in the community can be estimated from the proportions of species-specific variants among the amplicons. High throughput amplicon sequencing (HTAS), often termed metabarcoding, is a cost-effective way to detect multiple species simultaneously within a range of environmental samples (Poisot, Péquin and Gravel, 2013; Elbrecht and Leese, 2015; Gohl *et al.*, 2016; Tessler *et al.*, 2017; Fonseca, 2018; Krehenwinkel *et al.*, 2018). While shotgun sequencing of the whole community (metagenomics) can provide a richer description of the functions in a community, HTAS remains a more efficient tool for comparing the species diversity of a large number of community samples. Despite the extensive use of HTAS for interspecies ecological diversity studies, few investigations have utilised HTAS for intraspecies analysis (Kinoti *et al.*, 2017; Poirier *et al.*, 2018). As 16S rRNA amplicons are too highly conserved to estimate microbial within-species diversity, other target gene candidates need to be considered in order to sufficiently discern intraspecies sequence variation.

Many studies have evaluated the extent of PCR-based amplification errors and bias for HTAS diversity studies (Elbrecht and Leese, 2015; Kebschull and Zador, 2015; Gohl *et al.*, 2016; Krehenwinkel *et al.*, 2018). Numerous known PCR biases reduce the accuracy of diversity and abundance estimations, with the major concern being the inability to confidently distinguish PCR error from natural sequence variation in environmental samples, which is an especially limiting factor for intraspecific studies.

Polymerase errors, production of chimeric sequences by template switching, and the stochasticity of PCR amplification can be major causes of PCR errors (Edgar *et al.*, 2011; Kebschull and Zador, 2015; Edgar, 2016a). Polymerase errors introduce new sequences into the template population during amplification. These sequence errors include not only substitutions but also insertions and deletions. The use of proofreading polymerases, optimised DNA template concentration, and reduced PCR cycle number have been suggested to reduce these errors (Kebschull and Zador, 2015; Oliver *et al.*, 2015; Gohl *et al.*, 2016).

In order to account for the introduction of sequence variants in PCR amplification, several sequence-classification approaches have been established to manage diversity estimates. The most common method is the use of operational taxonomic units (OTUs) in microbial diversity studies which analyse target gene sequences and cluster based on an arbitrary fixed similarity threshold (QIIME (Bokulich *et al.*, 2018); UPARSE (Huse *et al.*, 2010; Edgar, 2013; Lindahl *et al.*, 2013; Poisot, Péquin and Gravel, 2013; Callahan *et al.*, 2016; Fierer, Brewer and Choudoir, 2017). Within species boundaries this technique could dramatically reduce the resolution of naturally occurring sequence variation.

Most recent methods rely on the formation of sequence groups called amplicon sequence variants (ASVs) (DADA2, (Callahan *et al.*, 2016); UNOISE3, (Edgar, 2016b; Fierer, Brewer and Choudoir, 2017). This approach allows sequence resolution down to one nucleotide, which is advantageous for determining intraspecies allelic variation, but noise from PCR errors is also more evident. Variation induced by PCR errors often cannot be differentiated from rare natural allelic variation without the use of sequence denoising methods (Kebschull and Zador, 2015). DADA2 relies on a quality-aware parametric error model, which is developed on a per sequencing run basis. This increases the run time compared to UNOISE3, which uses a one-pass technique (Nearing *et al.*, 2018).

An approach that can reduce sequencing noise is to assign a unique molecular identifier (UMI) to every initial DNA template within an HTAS sample, which also enables evaluation of PCR amplification bias (Lundberg *et al.*, 2013). Additionally, the UMI provides a potential route to address polymerase errors in metabarcoding studies. The UMI is provided by a set of random bases in the gene-specific forward inner primer, which introduces a unique DNA sequence into every initial DNA template upstream of the amplicon region during the first round of amplification. Once all original DNA template strands are assigned a unique UMI, an outer forward primer and the gene-specific reverse primer can be used for further amplification. Consequently, all subsequent DNA amplified from the original template will have the same UMI, so the number of reads amplified from the initial template can be calculated. Grouping sequences by shared UMI allows identification of a consensus, which is assumed to be the correct sequence (Kou *et al.*, 2016). To our knowledge, UMIs have previously only been used for single-amplicon interspecies investigations (Jabara *et al.*, 2011; Kinde *et al.*, 2011; Faith *et al.*, 2013; Hoshino and Inagaki, 2017).

Here, we present a method for metabarcoding using amplicons with unique molecular identifiers to improve error correction – MAUI-seq. The innovative approach is that we use variation among sequences associated with a single UMI to identify erroneous sequences, and we show that this improves error correction compared to non-UMI based analysis using the state-of-the-art software packages DADA2 and UNOISE3.

2.3. Materials and methods

2.3.1. Aim, design, and setting

MAUI-seq is a HTAS method designed to assess genetic diversity within or across species, using global UMI-based errors rates to detect potential PCR artefacts such as chimeras and single-base substitutions. To evaluate MAUI-seq, we compared its performance with the widely-used ASV clustering methods, DADA2 and UNOISE3 on DNA mixtures of two *Rlt* strains to assess accuracy on a set of known sequences, and two sets of environmental samples of white clover root nodules to assess the performance on a complex set of sequences.

2.3.2. Preparation of DNA mixtures

Two *Rlt* strains (SM3 and SM170C) were chosen based on their *recA*, *rpoB*, *nodA,* and *nodD* sequence divergence, with a minimum of 3 base pair differences in the amplicon region required for each gene. Strains were grown on Tryptone Yeast agar (28°C, 48hrs). Culture was resuspended in 750ul of the DNeasy Powerlyzer PowerSoil DNA isolation kit (QIAGEN, USA) and DNA was extracted following the manufacturer's instructions. DNA sample concentrations were calculated using QuBit (Thermofisher Scientific Inc., USA). DNA samples of the two strains were diluted to the same concentration and mixed in various ratios (Appendix Table A.1).

2.3.3. Preparation of environmental samples

For Field-Samples-1 data, white clover (*Trifolium repens*) root nodules were collected from two locations: Store Heddinge, Denmark (6 plots) and Aarhus University Science Park, Aarhus, Denmark (2 plots) (Appendix Figure A.2). The clover varieties sampled were Klondike (Store Heddinge) and wild white clover, (Aarhus). 100 large pink nodules were collected from 4 points on each plot, making a total of 32 samples. Nodules were stored at -20°C until DNA extraction. Nodule samples were thawed at room temperature and crushed using a sterile homogeniser stick. Crushed nodules were mixed with 750μl Bead Solution from the DNeasy PowerLyzer PowerSoil DNA isolation kit (QIAGEN, USA) and DNA was extracted following the manufacturer's instructions. DNA sample concentrations were measured using a Nanodrop 3300 instrument (Thermofisher Scientific Inc., USA).

For Field-Samples-2 data, root nodules were additionally sampled from 13 white clover conventionally-managed field trial plots at Store Heddinge, Denmark (Sample 1A-13A, Additional File 2). All plots were sown under the same conditions in 2017. Three to ten clover plants were sampled from one point in each plot and the 100 largest nodules collected. Nodules were stored at -20°C, and DNA was extracted for each sample using the Qiagen DNeasy PowerLyzer PowerSoil DNA isolation kit, as above. Samples were processed independently with Platinum (non-proofreading) and Phusion (proofreading) polymerases to evaluate the method dependency on polymerase choice, as described in the following sections.

2.3.4. PCR and purification

Primer sequences were designed for two *Rlt* housekeeping genes, recombinase A (*recA*) and RNA polymerase B (*rpoB*), and for two *Rlt* specific symbiosis genes, *nodA* and *nodD* (Additional File 1: Table S1).

The three primers are a target-gene forward inner primer, a universal forward outer primer, and a target-gene reverse primer. The concentration of the inner forward primer was 100-fold lower than the universal forward outer primer and the reverse primer (Figure 2.1) in order to reduce the competitiveness of this primer compared to the outer primer. The inner primer is essential for the first round of amplification, but its participation is undesirable in later rounds as it would assign a new unique UMI to an existing amplicon. The PCR reaction mixture and thermocycler programme are provided (Additional File 1: Tables S2 and S3).

PCRs were undertaken individually for each primer set using Platinum Taq DNA polymerase (Thermofisher Scientific Inc., USA) (Additional File 1: Table S2) and subsequently pooled and purified using AMPure XP Beads following the manufacturer's instructions (Additional File 1: Table S5) (Beckman Coulter, USA). Successful PCR amplification was confirmed by running a 0.5X TBE 2% agarose gel at 90V for 2 hours.

For the DNA mixture samples, PCRs were run in triplicate. DNA from single strains was also processed as a control to determine the level of cross contamination between samples. Some samples were also amplified using Phusion High-Fidelity polymerase (Thermofisher Scientific Inc., USA), to evaluate whether use of a proof-reading polymerase improved the quality of the results using the PCR program described in Additional File 1: Table S2 and Table S4.

2.3.5. Nextera indexing for multiplexing and MiSeq sequencing

Samples were indexed for multiplexed sequencing libraries with Nextera XT DNA Library Preparation Kit v2 set A (Illumina, USA) using the Phusion High-Fidelity DNA polymerase (Thermofisher Scientific Inc., USA). PCR reaction mixture and programme are detailed in Additional File 1: Tables S6 and S7 Indices were added in unique combinations as specified in the manufacturer's instructions (Illumina, USA).

The PCR product was purified on a 0.5X TBE 1.5% agarose gel and extracted with the QIAQuick gel extraction kit (QIAGEN, USA) (expected band length: ~454bp). PCR amplicon concentrations were quantified using GelAnalyzer2010a and normalised to 10nM (Lazar and Lazar, 2012). A pooled sample was quantified and checked for quality by Bioanalyzer (Agilent, USA) before sequencing using Illumina MiSeq (2x300bp paired end reads) by the University of York Technology Facility. A detailed protocol is available in Additional File 1.

2.3.6. Read processing and data analysis

The PEAR assembler was used to merge paired ends (Zhang *et al.*, 2014). Python scripts were used to separate the merged reads by gene (MAUIsortgenes.py) and to calculate allele frequencies both with and without the use of UMIs (MAUIcount.py). The scripts are available in the GitHub repository https://github.com/jpwyoung/MAUI. Sequences were clustered by UMI, and the number of unique UMIs was counted for each distinct sequence, provided that sequence had at least two more reads with that UMI than any other sequence. In cases where two or more sequences were associated with the same UMI, the second most abundant sequence was noted, and sequences that occurred more than 0.7 times as often as second sequences than as the main sequence associated with a UMI were filtered out of the results as putative PCR-induced chimeras or other errors. Sequences with primers removed (ignoring UMIs) were also clustered using DADA2 (version 1.8) (Callahan *et al.*, 2016) and UNOISE3 (USEARCH version 11.0.667) (Edgar, 2016b) with default settings. An overall read frequency filter of 0.1% was applied to DADA2 and UNOISE3 outputs to match MAUI-seq accepted sequences filtering.

Raw Illumina reads are available in the SRA repositories with accession numbers SRP221010 (Synthetic mix and Field-Samples-1) and SRP238323 (Field-Samples-2). Detailed output sequences for all three methods are available in Additional File 2. Scripts used for DADA2, UNOISE3, and figure generation are available in Additional File 3, 4, and 5, respectively. Output abundance data were then processed for statistical analysis and figure generation using various R packages (Additional File 3, 4, and 5; (Wickham, 2009; Team, 2015)). Principal components were calculated with the R 'prcomp' package using singular value decomposition to explain the *Rhizobium* diversity and abundance within each sub-plot sample. Differences in allele frequencies between samples were quantified using Bray-Curtis beta-diversity estimation using the R package 'vegdist.' PERMANOVA tests were performed using the R package 'adonis'. Empirical Bayes estimator of F_{ST} was calculated using the R package 'FinePop' as previously described (Kitada, Nakamichi and Kishino, 2017).

2.4. Results

2.4.1. Laboratory protocol: UMI labelling and amplicon multiplexing

We developed a procedure (MAUI-seq) to amplify multiple target genes from environmental samples, while assigning a random UMI to each initial copy of a template. We opted for a straightforward protocol using a "one-pot" initiation and amplification system. Forward primers consist of two modules; an inner primer bearing the UMI and designed to amplify the target gene, and a universal outer primer that binds only to a linker on the inner primer (Figure 2.1a). We used a 12-base UMI that allowed over 4 million distinct sequences, which is adequate to ensure that duplicate use is negligible for samples with a few thousand sequenced UMIs. For studies with greater sequencing depth, a longer UMI can easily be designed. As a test case, we used MAUI-seq to investigate the genetic diversity of the nitrogen-fixing bacterium *Rhizobium leguminosarum* symbiovar *trifolii (Rlt)* by characterising amplicons from the chromosomal core genes *rpoB* and *recA* and the plasmid-borne nodulation genes *nodA* and *nodD*. Each gene was amplified separately in a single reaction, using a target-specific inner forward primer (at low concentration) to assign the UMI and a universal outer primer (at high concentration) to amplify the resulting molecules (Figure 2.1a). The resulting amplicons were pooled and tagged by Nextera to identify the sample, then further pooled for high-throughput paired-end sequencing (Figure 2.1b). The full MAUI-seq step-by-step laboratory protocol can be found in Additional File 1.

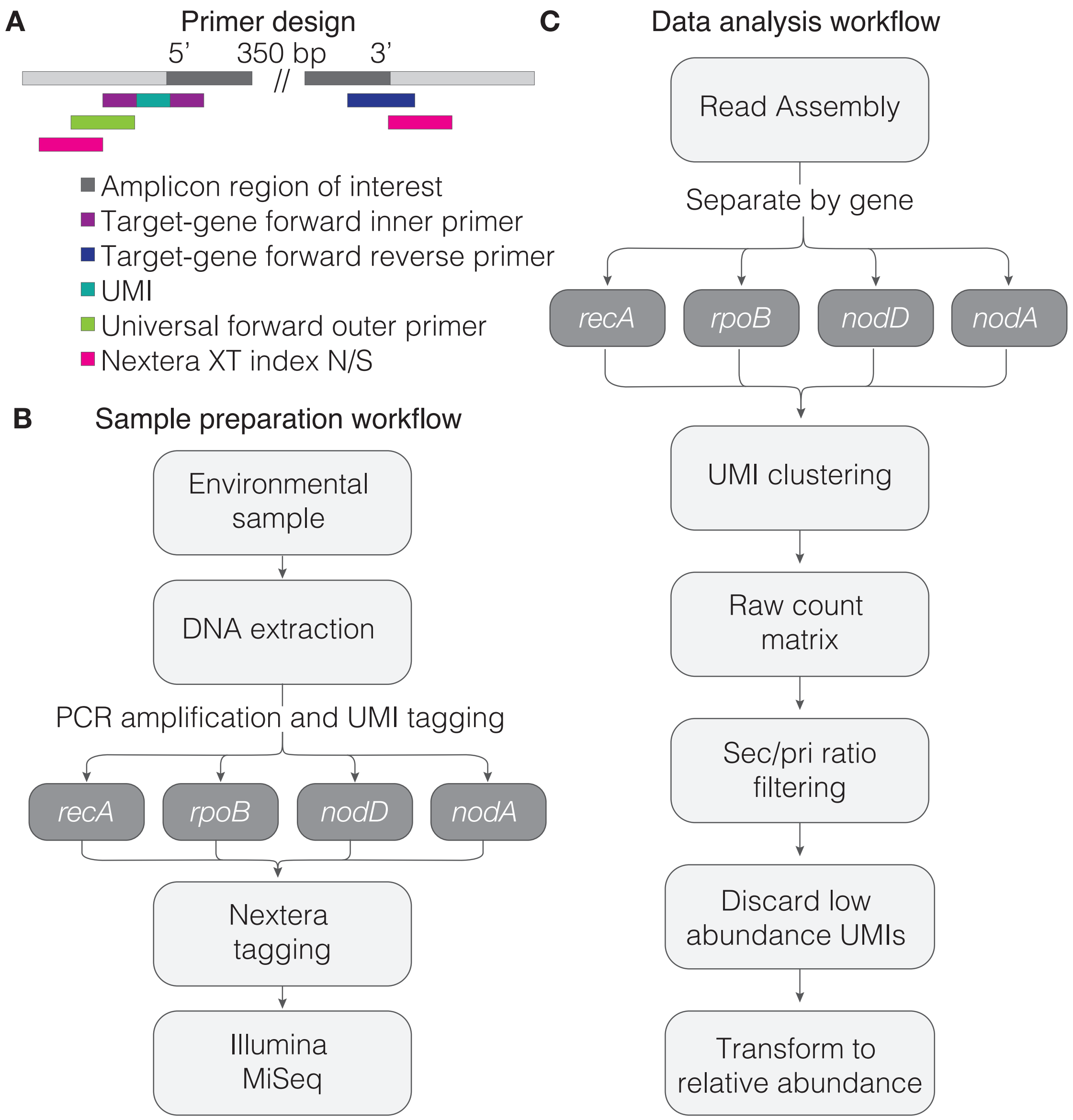


Figure 2.1 Primer design and method workflow. **a)** Primer design using the sense strand of the target DNA template as an example. The amplicon region of interest should be no longer than 500bp. The target-gene forward inner primer, universal forward outer primer and the target-gene reverse primer are all used in the initial PCR. The Nextera XT indices provide sample barcodes in a separate PCR step. The unique molecular identifier (UMI) region is shown in turquoise on the target-gene forward inner primer. **b)** Sample preparation workflow. **c)** MAUI-seq data analysis workflow.

2.4.2. Analysis protocol: filtering using UMI-based error rates

The resulting paired-end reads were merged and then separated by gene prior to downstream analysis, where UMIs are critical in two ways. Firstly, sequences are clustered by UMI, and the number of unique UMIs is counted for each distinct sequence, selecting the most abundant sequence associated with each UMI (Figure 2.1c). UMIs are discarded as ambiguous if the most abundant sequence does not have at least two reads more than the next in abundance. The most abundant sequence will usually be the correct one (Figure 2.2a Case 1) but, because most UMIs are represented by just a small number of reads, it can sometimes happen that an erroneous sequence is sampled more often than the true sequence, so the primary sequence of the UMI becomes this erroneous sequence (Figure 2.2a Case 2). Secondly, we reasoned that it may be possible to eliminate these errors by using the UMIs to provide information on global error rates across all samples. We implemented this in MAUI-seq by noting both the most abundant (primary) and the second most abundant (secondary) sequence if two or more sequences were associated with the same UMI. MAUI-seq then distinguishes between true and erroneous sequences based on the ratio of primary and secondary occurrences of each sequence, eliminating sequences that show a high ratio (default is 0.7) of secondary to primary occurrences (Figure 2.1c and Figure 2.2b). The 0.7 threshold was chosen empirically, based on the ratios observed for known true and erroneous sequences, but it is a compromise because the incidence of secondary sequences varies across genes and studies. An examination of the results may suggest choosing different thresholds in other studies. Finally, globally rare sequences are discarded (default threshold is 0.1% averaged across samples - a lower threshold could be used if samples were sequenced to a greater depth). Python scripts for separating the genes and for the UMI analysis are available at https://github.com/jpwyoung/MAUI.

To summarise, in the case where two or more sequences were associated with the same UMI, the more prevalent primary sequence was accepted. The secondary sequences associated with UMIs are considered errors and therefore they are only used to determine whether primary sequences are likely to be erroneous sequences (Figure 2.2a Case 2). In this method, the number of unique UMIs associated with each (primary) sequence are totalled to determine the abundance of a sequence across samples (note that number of UMIs are counted, not individual reads as with other ASV methods). For each unique amplicon sequence, we count the number of times it is found as the primary sequence for a unique UMI, and we also count the number of times it is a secondary sequence for a UMI. If the number of times a sequence is found as a secondary sequence compared to a primary sequence is a ratio of 0.7 or higher, then it is considered erroneous and the sequence is rejected from all samples. This is because the confidence of selecting the sequence as genuine is low, because it is highly associated among secondary sequences (which are assumed to be errors).

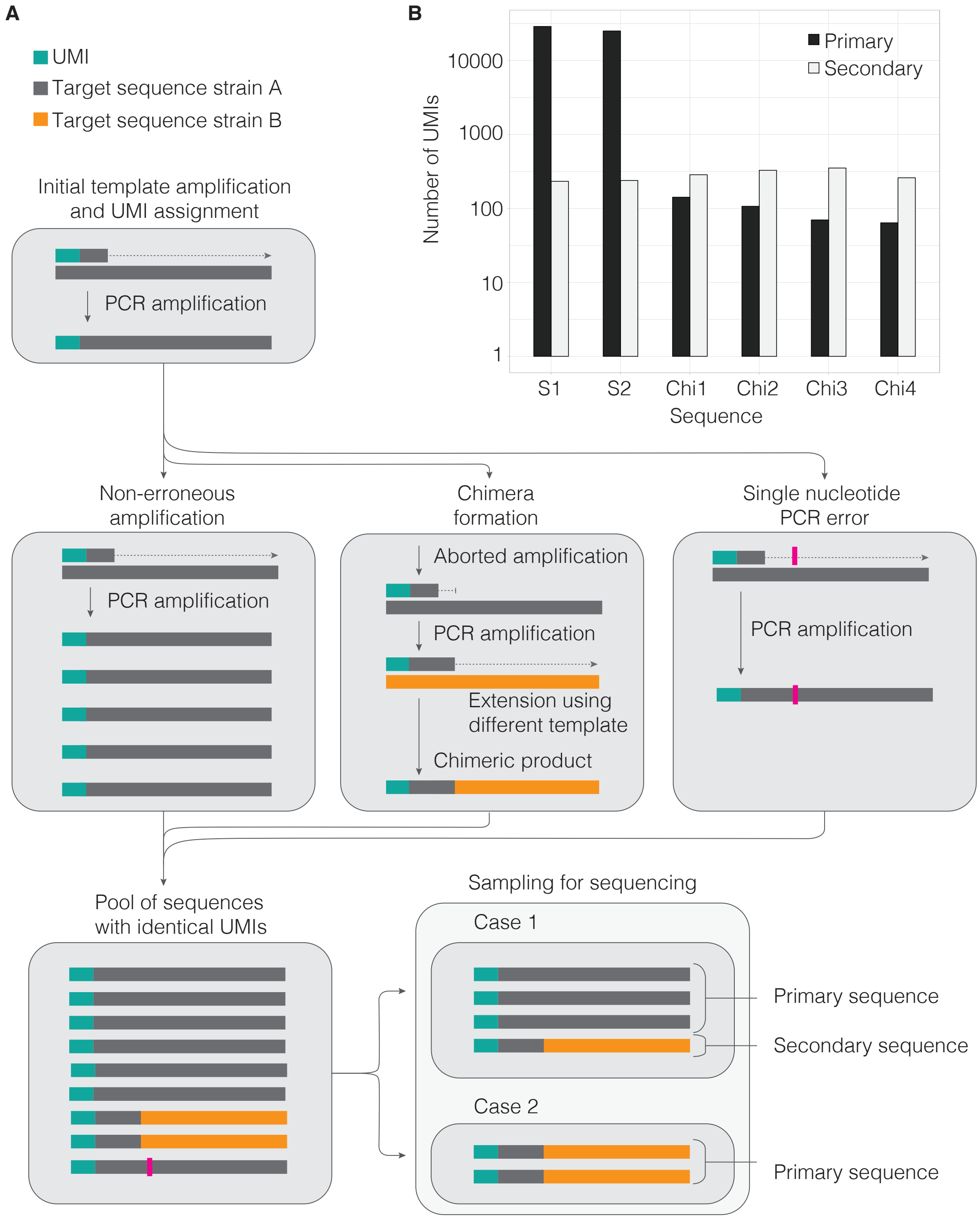


Figure 2.2 Erroneous read formation and filtering. **a)** Schematic showing the formation of different sequences with identical UMIs, and bias introduced when sampling for sequencing. **b)** Example data showing the occurrence of real and chimeric *rpoB* sequences as primary and secondary sequence (log scale). S1 and S2: Real sequences derived from two different rhizobium strains (SM170C and SM3). Chi1-4: Chimeric sequences.

2.4.3. Validation using purified DNA mixed in known proportions

We first evaluated the accuracy of MAUI-seq by profiling DNA mixtures with known strain DNA ratios. DNA was extracted from two *Rlt* strains differing by a minimum of 3 bp in each of their *recA*, *rpoB*, *nodA,* and *nodD* amplicon sequences, and the extracted DNA was mixed in different ratios (Appendix Table A.1). After amplification and sequencing, assembled reads were assigned to their target gene and analysed using MAUI-seq and two programs frequently used for de-noising of amplicon sequencing data, DADA2 and UNOISE3 (Callahan *et al.*, 2016; Edgar, 2016b). Since rare sequences have a high error rate, we discarded (for each of the three methods) sequences that fell below a threshold frequency of 0.1% of accepted sequences. The observed and expected strain ratios were highly correlated for all four genes across the three analysis methods, and we found that the performances of the proofreading (Phusion) and non-proofreading (Platinum) polymerases were gene-dependent, which could be due to differences in amplification efficiency for the four templates (Table 2.1 and Appendix Figure A.2-Appendix Figure A.5). On average, MAUI-seq detected between 98.5% and 100% true sequences exactly matching those of the two strains in the mixture, while DADA2 ranged from 89.7% to 100%, and UNOISE3 from 79.8% to 100% (Table 2.1). The better performance of MAUI-seq was due to more effective elimination of chimeras, which were especially abundant when the PCR reaction was carried out using the Platinum non-proofreading polymerase (Table 2.1 and Appendix Figure A.2-Appendix Figure A.5). For the proofreading polymerase, DADA2 detected 100% true sequences for all four genes, whereas MAUI-seq detected 99.03% for *nodA*, failing to eliminate three rare sequences that did not have sufficient secondary counts. This suggests that DADA2 can perform equally well or even slightly better than MAUI-seq, when a proofreading polymerase is used to amplify DNA from a simple, two-component mix. The prevalence of secondary sequences varied with gene and polymerase: the secondary/primary ratio for accepted sequences was 0.0322 for *rpoB* using Phusion, but just 0.0002 for *nodD* using Platinum. When the ratio was very low, there were insufficient secondary counts for MAUI-seq to eliminate erroneous sequences effectively.

Table 2.1 Total number of detected sequences in the synthetic mix samples using MAUI-seq, DADA2 and UNOISE3. The percentage of true sequences is averaged over 23 samples for Platinum (non-proofreading) and 14 samples for Phusion (proofreading). **n seq** is the total number of sequences occurring across all samples. **%true** is calculated by dividing the number of counts for the true sequences by the total number of counts accepted by the method. **%true-overall** is based on summed counts for all four genes. **Cor.exp/obs** is the Pearson correlation for the observed proportion of SM170C reads versus the expected proportion. **Chim.freq** is the proportion of chimeras compared to total reads at 0.5 expected proportion of sequences. **Exp.seq** is the expected number of detected sequences.† SM170C has a second copy of *nodD* (Cavassim *et al.*, 2019).

		Platinum			**Phusion**			
		MAUI-seq	**DADA2**	**UNOISE3**	**MAUI-seq**	**DADA2**	**UNOISE3**	**exp. seq***
rpoB	**n seq***	2	3	4	2	2	2	2
	%true*	100	96.96	93.80	100	100	100	-
	Cor.exp/obs*	0.956	0.977	0.981	0.996	0.999	0.9998	-
	chim.freq*	0	0.07	0.13	0	0	0	-
recA	**n seq**	2	2	2	2	2	2	2
	%true	100	100	100	100	100	100	-
	Cor.exp/obs	0.984	0.991	0.989	0.948	0.952	0.947	-
	chim.freq	0	0	0	0	0	0	-
nodA	**n seq**	6	5	4	5	2	4	2
	%true	99.04	89.70	89.93	99.03	100	90.43	-
	Cor.exp/obs	0.985	0.998	0.999	0.989	0.999	0.999	-
	chim.freq	0.10	0.25	0.22	0.04	0	0.16	-
nodD	**n seq**	7	6	21	3	3	14	3†
	%true	98.49	93.93	90.10	100	100	79.83	-
	Cor.exp/obs	0.998	0.998	0.995	0.990	0.998	0.995	-
	chim.freq	0.05	0.05	0.13	0	0	0.11	-
all	**%true-overall***	99.76	93.73	91.93	99.74	100	91.71	-

2.4.4. Validation using environmental samples

To test the method on more complex samples, we compared *Rlt* populations in root nodules from two locations in Denmark: a clover trial station in Store Heddinge on Zealand and a lawn at Aarhus University in Jutland (the Field-Samples-1 dataset; Appendix Figure A.1). One hundred nodules were pooled for each sample and each plot was sampled in four replicates. Platinum Taq polymerase enzyme was used for amplification. Each clover root nodule is usually colonised by a single *Rhizobium* strain, so a maximum of 100 unique sequences per gene is expected per sample.

For Field-Samples-1, the total number of distinct sequences for MAUI-seq and DADA2 were in the same range as the number of distinct alleles observed in a population of 196 natural European *Rlt* isolates (Cavassim *et al.*, 2019) (Table 2.2). In contrast, UNOISE3 produced a substantially higher number of distinct sequences, suggesting that its default filtering might be too lenient for our data (Table 2.2). The sequences accepted as true by MAUI-seq were nearly all also included in the DADA2 and UNOISE3 outputs (Figure 2.3). On the other hand, DADA2 and UNOISE3 both accepted a number of sequences that were filtered out by MAUI-seq, and many of these were eliminated by MAUI-seq because a high ratio of secondary to primary occurrences strongly suggested that they represent errors and not real sequences (Figure 2.3 and Additional File 2: Field-Samples-1 tables). To provide independent evidence as to whether sequences were likely to be genuine, we checked whether they matched (or differed by a single nucleotide from) known sequences in either a reference database of 196 natural European *Rlt* isolates (Cavassim *et al.*, 2019), or the NCBI whole-genome shotgun database (Figure 2.3). The great majority of sequences rejected by MAUI-seq did not have exact matches to these known sequences. A few sequences that exactly matched known alleles were included by DADA2 and UNOISE, but not by MAUI-seq. These sequences were not reported by MAUI-seq because their UMI counts were below the abundance threshold, not because the secondary/primary occurrence filter identified them as erroneous (Figure 2.3). The count threshold could be lowered to include rarer sequences, if the study required it.

The allele frequency distributions were different at Aarhus and Store Heddinge (Figure 2.3), and the two sites were clearly separated by the first principal component in a Principal Component analysis (PCA) for MAUI-seq, DADA2 and

UNOISE3 sequences. (Figure 2.4 and Appendix Figure A.6-Appendix Figure A.8). The amplicon sequencing has sufficient resolution to characterize geospatial variation in allele frequencies. For example, MAUI-seq, DADA2 and UNOISE3 can all clearly identify several highly abundant sequences from one location that are either absent or present in very low frequency in samples from the other location (Figure 2.3). To quantify the genetic differentiation between the Aarhus and Store Heddinge sites, we calculated fixation indices (F_{ST}). Considering all four target genes combined, the MAUI-seq output resulted in the highest F_{ST} value followed by DADA2 and UNOISE3 (Table 2.2, Figure 2.4 and Appendix Figure A.9-Appendix Figure A.11). For all individual genes, MAUI-seq also produced the highest F_{ST} estimates, and the differences were especially pronounced for *nodA*, which also showed the highest overall level of differentiation (Table 2.2 and Appendix Figure A.9-Appendix Figure A.11). The lower genetic differentiation estimated based on DADA2 and UNOISE3 results, compared to those of MAUI-seq, reflects the inclusion of an increased number of erroneous sequences, which are less differentiated between the two sampled sites than the real sequences (Figure 2.3).

Since it was clear from the DNA mixture experiment that the choice of DNA polymerase could significantly affect error rates, we sampled root nodules from 13 additional clover field plots (the Field-Samples-2 dataset) and amplified each sample (a pool of one hundred root nodules) using Platinum and Phusion polymerases in parallel. For samples amplified using Platinum, MAUI-seq detected fewer sequences than DADA2 and UNOISE3 for the two core genes, but the same number of reference sequences were detected (Table 2.3). DADA2 included two chimeric sequences that were filtered out by MAUI-seq due to a high ratio of secondary to primary occurrences (Additional File 2: Field-Samples-2-platinum-rpoB). UNOISE3 detected twice as many sequences as DADA2 and MAUI-seq for the accessory genes, but most of the additional sequences had no associated UMIs and were classified as “other” (Table 2.3, Additional File 2: Field-Samples-2-platinum-nodA and Field-Samples-2-platinum-nodD). For samples amplified using Phusion, MAUI-seq and DADA2 detected a similar number of sequences (Table 2.3). All nine UNOISE3 *rpoB* sequences that were not accepted by either MAUI-seq or DADA2 (Additional File 2: Field-Samples-2-phusion-rpoB) are putative chimeric sequences with two parental sequences of higher abundance. For *nodA*, MAUI-seq includes three sequences that have a single nucleotide difference from a reference sequence, but all have a good ratio of secondary to primary reads, so we hypothesise that these are true sequences. Some reference or exact blast hit sequences were included by DADA2 but not by MAUI-seq because their abundance was estimated by DADA2 to be above the 0.001 threshold, but MAUI-seq estimated that they were rarer.

Both MAUI-seq and DADA2 identify and remove sequences that appear to be errors (base substitutions or chimeras), but they use completely different evidence. As a result, they do not always make the same decision, as illustrated for a small set of representative data in Table 2.4 (the *rpoB* sequences amplified by Phusion). While DADA2 examines the sequences and rejects those that are likely to be generated from more abundant sequences in the sample, MAUI-seq does not use the actual sequence but bases decisions on how frequently a sequence occurs as a secondary sequence with the same UMI as another (primary) sequence. Sequences ranked 5 and 6 (Table 2.4) are both potential chimeras of the more abundant sequences 1-4. Both DADA2 and MAUI-seq reject sequence 6 and accept sequence 5. Sequence 6 has a secondary/primary ratio of 103/118, which is above the default threshold of 0.7, so MAUI-seq rejects it as a likely error. On the other hand, the ratio for sequence 5 is 71/229. This is well below the threshold, but it is higher than other sequences with a similar primary count, e.g. sequence 9 (15/270). A possible explanation is that some of the reads for sequence 5 are generated as chimeras but others are genuine, since it is entirely plausible that new alleles are generated by recombination between existing alleles. To some extent, MAUI-seq compensates for this because it allocates sequence 5 a relatively low count and hence lower ranking (8) than it has in the raw reads or the DADA2 analysis. There are two further sequences, 10 and 29, that are rejected by DADA2 as potential chimeras but accepted by MAUI-seq (Additional File 2: Field-Samples-2-phusion-rpoB); in both cases they have secondary sequence counts well below the threshold, so MAUI-seq accepts them as genuine. DADA2 included an *rpoB* sequence that does not have any associated UMIs (sequence 41), and appears to be a chimera of two more abundant sequences (sequence 3/4/5 and sequence 11) (Table 2.4). MAUI-seq counts UMIs, not individual reads, and the default setting is to require that the primary sequence has at least two more reads than the next most frequent sequence (if any) that has the same UMI. This enriches for genuine sequences, which are generally more abundant than errors, but it means, of course, that the number of counts is much lower than the number of reads. In fact, for this particular set of data, the number of UMIs is orders of magnitude smaller than either the raw reads or the DADA2 count, although still sufficient to provide good estimates of the relative abundance of the sequences that make up the bulk of the population. The main reason for the low UMI count is that the number of reads per UMI was suboptimal in these data for the *rpoB* gene: only 18% of the UMIs had more than one read, and MAUI-seq discards single-read UMIs by default. By contrast, in the equivalent data for the *recA* gene in the same study (Additional File 2: Field-Samples-2-phusion-recA), 37.5% of UMIs had more than one read, making more effective use of the available sequence reads.

Table 2.2 Total number of detected sequence clusters in root nodule samples (Field-Samples-1) using MAUI-seq, DADA2, and UNOISE3 clustering and genetic differentiation between populations. *Output sequences were classified into **reference** (100% identity in at least 1 of 196 *Rhizobium leguminosarum* symbiovar *trifolii* genomes (Cavassim *et al.*, 2019)), **exact BLAST** (100% query coverage and 100% identity against the whole-genome shotgun contigs BLAST database), **single nt** (one nt difference from either reference or exact BLAST match), and **other**. **Total reference** is the total number of detected sequences in the 196 *Rhizobium leguminosarum* symbiovar *trifolii* genomes. † The population global F_{ST} (fixation index) is an estimate of genetic differentiation among populations based on relative allele abundance.

		Detected sequence clusters*					**F_{ST}**†
Gene	**Method**	**Total**	**Reference**	**Exact BLAST**	**Single nt**	**Other**	
rpoB	**MAUI-seq**	12	7	3	1	1	0.032
	DADA2	15	7	3	3	2	0.032
	UNOISE3	30	7	2	7	14	0.012
	Total Reference*	**13**	-	-	-	-	-
recA	**MAUI-seq**	8	6	2	-	-	0.110
	DADA2	13	8	2	3	-	0.090
	UNOISE3	14	5	2	2	5	0.028
	Total Reference	**17**	-	-	-	-	-
nodA	**MAUI-seq**	9	8	-	1	-	0.369
	DADA2	18	12	1	1	4	0.191
	UNOISE3	43	13	-	5	25	0.061
	Total Reference	**14**	-	-	-	-	-
nodD	**MAUI-seq**	18	11	1	2	4	0.139
	DADA2	22	11	1	3	7	0.124
	UNOISE3	57	11	1	4	41	0.031
	Total Reference	**16**	-	-	-	-	-
All genes	**MAUI-seq**	47	32	6	4	5	0.139
	DADA2	68	38	7	10	13	0.105
	UNOISE3	144	36	5	18	85	0.032

Table 2.3 The effect of polymerase choice. Total number of detected sequence clusters in root nodule samples (Field-Samples-2) amplified using Phusion (proofreading) or Platinum (non-proofreading) polymerases. Sequences were clustered using MAUI-seq, DADA2, and UNOISE3. *Output sequences were classified into **reference** (100% identity in at least 1 of 196 *Rhizobium leguminosarum* symbiovar *trifolii* genomes (Cavassim *et al.*, 2019)), **exact BLAST** (100% query coverage and 100% identity against the whole-genome shotgun contigs BLAST database), **single nt** (one nt difference from either reference or exact BLAST match), and **other**.

		Platinum			**Phusion**		
Gene		**MAUI-seq**	**DADA2**	**UNOISE3**	**MAUI-seq**	**DADA2**	**UNOISE3**
rpoB	**Total**	**16**	**24**	**26**	**15**	**15**	**20**
	Reference*	9	9	7	8	9	7
	Exact BLAST*	3	3	2	3	3	2
	Single nt*	3	7	8	3	2	5
	Other*	1	5	9	1	1	6
recA	**Total**	**9**	**10**	**12**	**8**	**9**	**10**
	Reference	5	5	4	5	5	4
	Exact BLAST	0	1	1	0	1	1
	Single nt	3	3	3	3	2	3
	Other	1	1	4	0	1	2
nodA	**Total**	**18**	**14**	**35**	**17**	**11**	**34**
	Reference	7	10	8	9	9	9
	Exact BLAST	0	1	0	0	0	0
	Single nt	6	1	4	6	1	4
	Other	5	2	22	2	1	21
nodD	**Total**	**20**	**17**	**46**	**27**	**24**	**71**
	Reference	10	12	12	16	16	15
	Exact BLAST	0	0	0	0	0	0
	Single nt	6	3	6	5	4	6
	Other	4	2	28	6	3	50

Table 2.4 A comparison between DADA2 and MAUI-seq for a subset of the Field-Samples-2 data summarised in **Table 2.3**: the *rpoB* sequences from samples amplified by Phusion (proofreading) polymerase. Red cells refer to rejected sequences. Green cells refer to sequences, which are accepted by MAUI-seq, while DADA2 rejects them as potential chimeras. Yellow cells refer to sequences filtered out due to low UMI count by MAUI-seq.

Raw reads		**MAUI-seq**				**DADA2**		
Rank	**count**	**rank**	**UMI primary count**	**UMI secondary count**	**accepted**	**rank**	**count**	**accepted**
1	99431	1	7459	197	yes	1	54758	yes
2	86751	2	7067	155	yes	2	48402	yes
3	70318	3	3668	95	yes	3	44412	yes
4	47337	4	1898	106	yes	4	28339	yes
5	13190	8	229	71	yes	5	7854	yes
6	11786	9	118	103	no	none	NA	no
7	10490	5	489	19	yes	6	6009	yes
8	9630	6	362	13	yes	7	5414	yes
9	4738	7	270	15	yes	8	2757	yes
10	4290	12	62	15	yes	none	NA	no
11	3223	11	90	3	yes	9	2041	yes
20	1950	10	96	6	yes	10	981	yes
29	1504	13	42	10	yes	none	NA	no
39	1063	14	35	2	yes	12	618	yes
41	946	none	0	0		11	721	yes
43	826	15	34	0	yes	13	434	yes
51	567	16	22	3	yes	14	341	yes
63	415	24	7	0	(yes)	15	208	yes

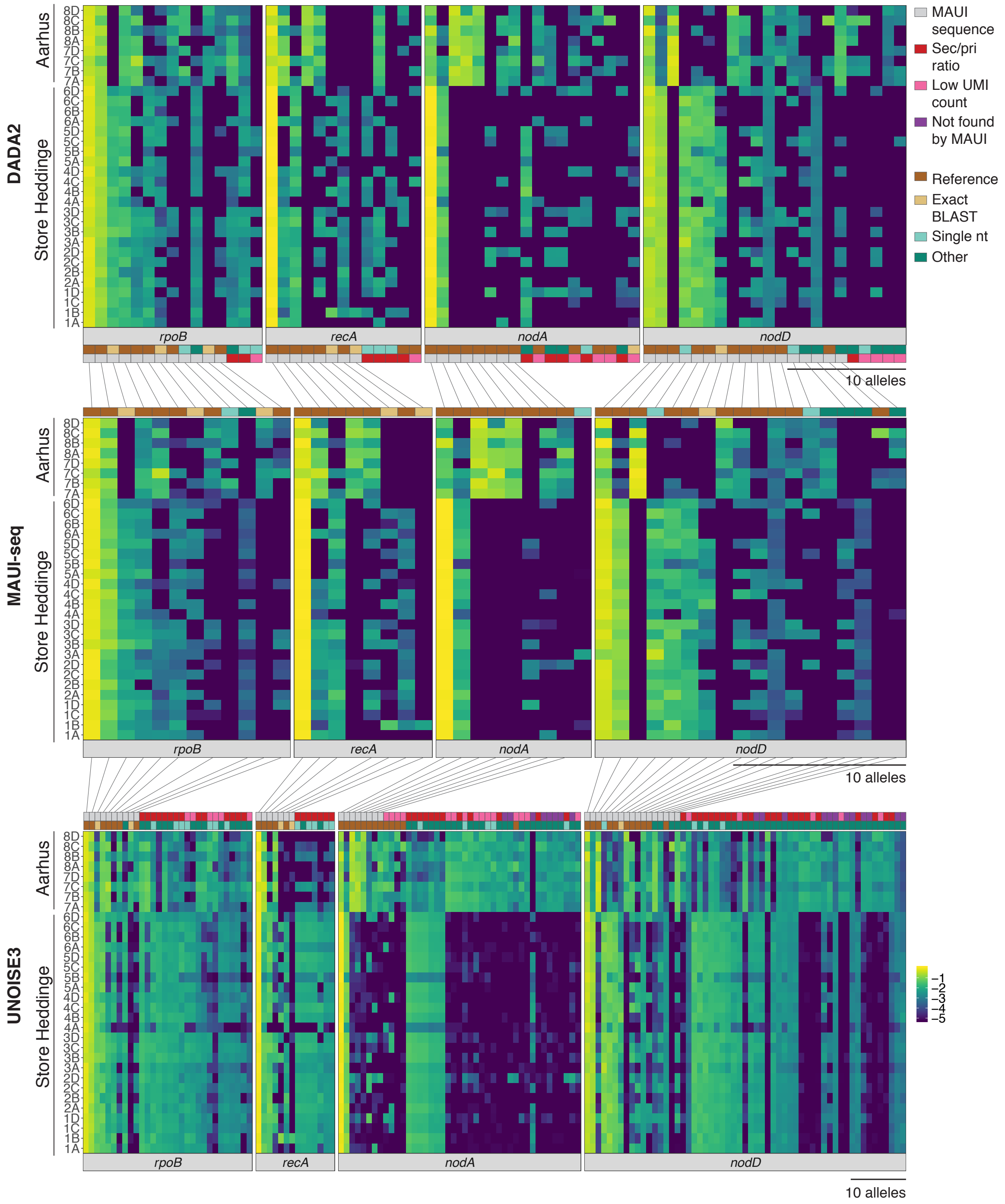


Figure 2.3 Amplicon diversity reported by MAUI-seq compared with the DADA2 and UNOISE3 analysis pipelines. Data are for four genes from nodule samples from two geographic locations, Store Heddinge (1-6) and Aarhus (7-8). Letters A-D denote the replicates within each plot (**Appendix Figure A.1**). Heatmap of the $\log_{10}$ transformed relative allele abundance of sequence clusters for individual genes. Lines connect identical sequences found by different clustering methods. Evidence that sequences are likely to be genuine is denoted by classifying them as **reference** (100% identity in at least 1 of 196 *Rhizobium leguminosarum* symbiovar *trifolii* genomes (Cavassim *et al.*, 2019)), **exact BLAST** (100% query coverage and 100% identity against the whole-genome shotgun contigs BLAST database), **single nt** (one nt difference from either reference or exact BLAST match), and **other**. Sequences not reported by MAUI were classified as **sec/pri ratio** (rejected as erroneous because of a high secondary to primary ratio), **low UMI count** (not reported because too rare), **not found by MAUI** (no accepted UMIs).

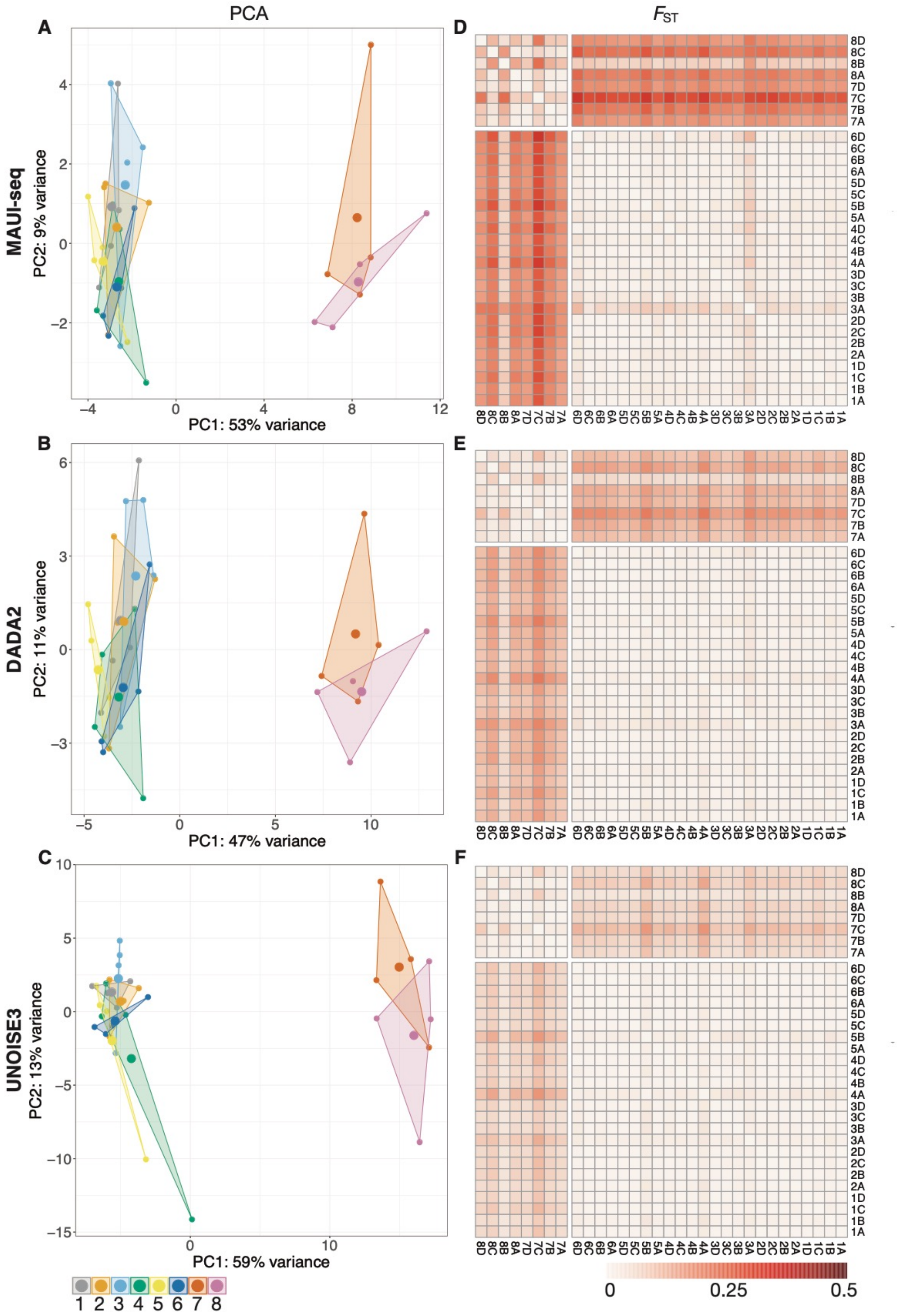


Figure 2.4 Genetic differentiation between populations visualised by Principal Component Analysis (**a-c**) and F_{ST} (**d-f**) of *Rlt* diversity in root nodule samples (8 sites, 4 replicates). Three analysis pipelines are compared: MAUI-seq (**a,d**), DADA2 (**b,e**), UNOISE3 (**c,f**). The PCA analysis was based on log10 transformed relative allele abundance. F_{ST} analysis was based on relative allele abundance. Data from all four genes (*rpoB*, *recA*, *nodA*, and *nodD*) were included in the analysis.

2.5. Discussion

We propose a new HTAS method (MAUI-seq) designed to assess genetic diversity within or across species. It uses global UMI-based errors rates to detect potential PCR artefacts such as chimeras and single-base substitutions more robustly than the widely-used ASV clustering methods, DADA2 and UNOISE3. The approach is potentially applicable to any study of amplicon diversity, including community diversity estimates based on 16S rRNA and other metabarcoding surveys using environmental DNA.

2.5.1. Using UMIs to filter out chimeras and other errors

In the MAUI-seq approach, UMIs are used to reduce errors in two distinct ways. Since all reads with the same UMI should, in principle, be derived from the same initial template copy, any variation among them reflects errors. In some implementations, a consensus sequence is calculated (Kou *et al.*, 2016), but we adopt the simpler approach of accepting the most abundant sequence, which will usually give the same result. Requiring more than one identical read before accepting a UMI creates an important quality filter that greatly reduces the number of rare (and usually erroneous) sequences, but as more reads are required, an increasing number of the original reads are discarded and the number of accepted counts declines. To strike a balance between quantity and quality, we chose to count a sequence provided it had at least two more reads than the next most frequent sequence with the same UMI, but this threshold could be adjusted if, for example, a markedly larger number of reads were available.

While the most abundant sequence associated with a UMI will usually be the correct one, it will sometimes happen that an erroneous sequence will predominate among the small number of reads actually sequenced, leading to these sequences being included among the recorded counts. These errors can be detected, though, by aggregating information across the whole set of samples. When a UMI is associated with more than one sequence, the secondary sequences are most often erroneous, so sequences that are relatively more abundant as secondary sequences than as the primary sequences associated with UMIs are likely to be erroneous. We recorded the number of times each sequence was found as the second sequence associated with a

UMI, and found empirically that a suitable threshold for accepting sequences as genuine was that they occurred less than 0.7 times as often as secondary sequences as they occurred as primary sequences. This threshold can, however, be adjusted to reflect the error distribution observed in a particular study. We found that this approach was very effective in identifying known errors, particularly chimeras, which were generally the most abundant errors. Chimeras were rejected more effectively by MAUI-seq than by the two established ASV clustering methods, DADA2 and UNOISE3. Both of these rely on *de novo* rejection of sequences that could be constructed as recombinants of other sequences that are more abundant in the sample (Edgar, 2016a). This method risks rejecting sequences that appear to be recombinant but are genuine alleles, which may not be uncommon, particularly in intraspecific samples. Our approach, by contrast, uses information on the observed error rates in the data (detected using UMIs) to decide whether a sequence is likely to be genuine, regardless of its actual sequence and relationship to other sequences. Sequences that could be generated as chimeras, or that differ by a single nucleotide from a more abundant sequence, may be accepted as genuine if they are more abundant than expected from their rate of occurrence as minor sequences associated with UMIs. In our study, this approach eliminated many known errors and substantially improved our confidence in the remaining data, providing a powerful additional reason for using UMIs in metabarcoding studies of all kinds. While we found that a simple empirical threshold was effective, we noticed that the proportion of secondary sequences varied markedly across studies and genes, suggesting that an adjustable threshold might give further improvement. A useful future development might be to use the abundance of minor sequences associated with UMIs to generate a statistical model of error processes that would provide a firmer theoretical basis for the classification of sequences.

2.5.2. Using UMIs to reduce amplification bias

One motivation for the use of UMIs is to obtain more accurate relative abundance data by eliminating possible sequence-specific bias in the PCR amplification, which may be introduced by variation in polymerase and primer affinity for some DNA templates. Indeed, we observed that the Platinum polymerase preferentially amplified the SM170C *rpoB* allele, whereas the Phusion enzyme did not have this bias (Table 2.1 and Appendix Figure A.2a-c). Allele variant bias was also shown for other target genes, although the ranking of the two enzymes was not always the same (Table 2.1 and Appendix Figure A.2-Appendix Figure A.5). However, in our study, the use of UMIs did not correct the allele bias. This suggests that the bias was present in the initial round of copying using the target-specific primer, rather than in the subsequent amplification rounds. For our case study, at least, the choice of polymerase was much more important for accurate relative abundance data than the use of UMIs. The main advantage of UMIs was, rather, the ability to remove most sequencing errors, as discussed in the preceding section.

2.5.3. Advantages of multiplexing several amplicons

Increasing the number of monitored amplicons to four increased our ability to robustly distinguish samples from two locations (Figure 2.3, Figure 2.4 and Appendix Figure A.6-Appendix Figure A.11). Multiplexing could be used in other ways, for example to monitor several organisms in the same environment, or to increase read coverage profiling of single genetic markers such as 16S (Fuks *et al.*, 2018). In addition, there is a technical benefit in sequencing multiple different targets together, because a lack of sequence diversity can cause Illumina base-calling issues (Krueger, Andrews and Osborne, 2011).

2.5.4. Optimization of the protocol

As with any metabarcoding project, the first important step is to design the primers carefully to amplify the entire target community with minimum bias, and we used a large database of known gene sequences to achieve this. Another consideration that is shared with other approaches is the choice of polymerase for PCR. For the samples studied here, with abundant template DNA, the proofreading enzyme was clearly superior in performance, although more costly. On the other hand, this enzyme may provide less robust amplification when the template is weak, as we have observed in another project aimed at rhizobial DNA in soil (Boivin *et al.*, 2020). The use of UMIs introduces other design considerations. We used twelve random nucleotides (with some constraints), giving over four million potential UMI sequences, which was sufficient for the scale of our studies, but it would be simple to increase the UMI length if greater sequencing depth was planned. In any metabarcoding study, the choice of sequencing depth is, to some degree, made blindly because the diversity of templates is not known in advance, but UMI-based approaches need greater depth because it is UMIs that are counted, not reads, and the aim is to have several reads per UMI. There are many factors that affect the average number of reads per UMI, but our study is encouraging in that, without separate optimization, all of our target genes in all of our samples gave usable data. In fact, the number of reads per UMI were suboptimal in most cases. Given a fixed sequencing effort, reads per UMI could, if necessary, be increased by reducing the concentration of the forward UMI-bearing primer and/or of the sample DNA so that fewer distinct UMIs were initiated. With our parameters, at least two reads are needed before a UMI is counted, and a sufficient fraction of the UMIs need at least four reads so that some will have a secondary sequence as well as the primary sequence (with at least two reads more than the secondary).

2.5.5. Future directions for MAUI-seq

HTAS is a valuable and widely-used approach for the study of microbial community diversity, but handling erroneous sequences introduced by the amplification and sequencing procedures has always been challenging. The use of UMIs allows MAUI-seq to greatly reduce the incidence of errors through two mechanisms. Firstly, the requirement that a UMI is associated with at least two identical reads eliminates many rare sequences that are predominantly erroneous. Secondly, sequences that are frequently generated as errors can be identified and removed because they occur unexpectedly often as minor components associated with UMIs that are assigned to more abundant sequences. These mechanisms are independent of any reference database and can recognise and retain genuine alleles that differ by a single nucleotide or match a potential chimera. This makes MAUI-seq particularly suited to studies of intraspecific variation, where the range of sequence divergence may be limited and not fully known in advance. However, the efficient elimination of erroneous sequences is also important in community studies such as those based on widely-used 16S primers, and MAUI-seq should be readily adaptable to this field. The analysis pipeline is very fast because no sequence alignment or database searching is involved; only the accepted final sequences would need to be characterised by comparison to a reference database.

Most HTAS studies report the relative proportions of the taxa in a community, but it would sometimes be valuable to estimate the absolute abundance of the microbes in the environmental sample. UMIs can potentially provide such information, if the initial template copying is carefully controlled so that the total number of distinct UMIs reflects the number of templates (Kivioja *et al.*, 2011; Hoshino and Inagaki, 2017). While this would necessitate some additional steps at the start of the experimental protocol, it should still be possible to analyse the resulting sequences using the error-removal approaches provided by MAUI-seq. Alternatively, absolute abundance can be estimated by adding a spike of a known quantity of a recognisable target sequence to the sample before processing (Kebschull and Zador, 2015; Edgar, 2017; Palmer *et al.*, 2018).

The addition of a UMI shortens the maximum length of target sequence that can be read, and the counting of UMIs rather than reads requires a higher depth of sequencing, but these limitations are increasingly unimportant as improvements in sequencing technology lead to increasing length, enabling long-read amplicon sequencing (Karst *et al.*, 2019; Kumar *et al.*, 2019), and numbers of reads. As implemented in MAUI-seq, UMIs are very effective in reducing the errors inherent in HTAS, and have the potential to improve the quality of any amplicon-based study of diversity. There are several parameters (minimum difference between primary and secondary reads of a UMI, ratio of secondary to primary reads of a sequence, minimum relative abundance) that are user-specified and can be adjusted to suit each study. In principle, it should be possible to optimize these using a statistical model of mutational errors, like that implemented in DADA2 (Callahan *et al.*, 2016) and of chimera formation, which is not modelled in detail by DADA2. The UMIs provide an additional source of information to parameterize the model, linking sequences that have a common origin. Such a model would be complex, however, and parameterizing and testing it would need a dataset that was optimized for the purpose. At the same time, it would also be interesting to explore the use of UMIs at both ends of the amplicon, which would provide an additional means to identify and eliminate chimeras (Burke and Darling, 2016).

2.5.6. Conclusions

Some potential advantages of incorporating UMIs in amplicon diversity studies have been explored previously, but here we propose a new way to use the extra information that they provide. Error processes lead to more than one sequence being associated with the same UMI, and this can be used to identify erroneous sequences regardless of their relative abundance or their relationship to other sequences in the sample. The method is experimentally and computationally straightforward, and we demonstrate its effectiveness using known strain mixtures and real environmental samples. It allows decontamination of amplicon sequence data by flagging chimeras and other errors, and can readily be adapted to any target gene of interest in microbiome studies.

Chapter 3. *Rhizobium* nodule diversity is determined by both clover host genotype and local growth conditions

3.1. Abstract

Background: Plant species impose a diverse range of selection pressures on bacterial communities, both within the soil and their own microbiomes. Shaping of plant microbiome composition through 'host-filtering' is well documented in legume-rhizobia symbioses where different plant species can disproportionally change the microbiome composition by interacting with specific symbionts. However, it is less clear how much individual varieties of plant species differentially influence the intraspecies diversity of their symbionts, especially under complex field conditions.
Results: This study compared how host legume genotype affects rhizobium population diversity in root nodules under conventional field conditions in Denmark. Five *Trifolium repens* (white clover) genotypes were grown in a conventional field trial, and differences in root nodule *Rhizobium leguminosarum* symbiovar *trifolii (Rlt)* genotype diversity were compared using MAUI-seq high-throughput amplicon sequencing of two chromosomal housekeeping genes, *rpoB* and *recA*, and two auxiliary plasmid-bound symbiosis genes, *nodA* and *nodD*. It was found that *Rlt rpoB* and *recA* nodule diversities significantly differed between clover genotypes, and *rpoB* and *recA* allele frequencies could be further used to infer differences in the proportions of *Rlt* sub-species (genospecies) between some clover genotypes. Diversity of *rpoB* and *recA* was significantly associated with geographic distance within fields, suggesting that in addition to host genotype, local differences in soil physicochemical properties and microbiota composition also likely influenced nodule diversity. *nodA* and *nodD* diversities were not significantly attributed to host genotype or geographic distance, indicating that intraspecies symbiotic specificity might not be associated with these genes.
Conclusions: These results suggest that variation in local growth conditions and host genotype together influence white clover *Rlt* nodule diversity under agricultural conditions.

3.2. Introduction

Plant species are well documented to influence their surrounding microbial composition by attracting and repelling specific microorganisms to their productive benefit (Burns *et al.*, 2015; Quiza, St-Arnaud and Yergeau, 2015; Fitzpatrick *et al.*, 2018; Schmid *et al.*, 2018; Zhalnina *et al.*, 2018; Jones *et al.*, 2019; Veach *et al.*, 2019). This plant host-filtering of the soil microbial community is mediated through multiple mechanisms, such as secretion of various root exudates enabling microbe-plant signalling and microbe recognition systems to differentiate pathogenic and non-pathogenic bacteria (Jones *et al.*, 2019). In practice, plant host-filtering is exploited agriculturally to improve soil fertility by altering the soil microbial composition between crop rotations of different plant species (Ashworth *et al.*, 2017; Song *et al.*, 2018; Zhang *et al.*, 2019). The effects of distinct varieties of the same plant species on microbial soil diversity has been studied in non-legumes; plant genotypes have been found to significantly shape rhizosphere bacterial diversity under both greenhouse and field conditions, but soil type can influence bacterial diversity to an even greater extent (Inceoğlu *et al.*, 2010; Bulgarelli *et al.*, 2012, 2013, 2015; Lundberg *et al.*, 2012; Peiffer *et al.*, 2013). Intraspecies diversity clearly warrants further investigation, as soil bacterial species can show huge within-species taxonomic and functional diversity, that is undetectable with 16S sequencing, and which can have profound effects on plant-microbe interactions and ecosystem functioning (Gaunt *et al.*, 2001; Case *et al.*, 2007; Adékambi, Drancourt and Raoult, 2009; Vos *et al.*, 2012; Zhang *et al.*, 2012, 2017; Li *et al.*, 2013; Burns *et al.*, 2015; Miranda-Sánchez, Rivera and Vinuesa, 2016; Wang *et al.*, 2018).

Symbiotic rhizobia bacteria are one such highly diverse group of soil microbes regularly exposed to host-filtering by plants. Legumes form root nodule symbioses with particular nitrogen-fixing rhizobia species to increase their nitrogen uptake for subsequent growth. Compatible rhizobia selection for symbiosis is initially mediated by highly specific interactions between plant flavonoid exudates and expressed rhizobial symbiosis genes (Wang *et al.*, 2012; Clúa *et al.*, 2018). Therefore, legume crops are commonly inoculated with rhizobia strains in the field as a sustainable alternative to chemical fertilization. In order to optimise the efficiency and success of the inoculation treatment, inoculant developers aim to select highly genetically compatible strains for the host plant. Consequently, optimisation of host-symbiont partner selection requires further understanding of 1) the degree to which intraspecies genetic specificity is linked with effective symbiosis, 2) the potential competing interactions with native soil rhizobia and 3) rhizobium survival under various abiotic soil conditions (Irisarri *et al.*, 2019; Jones *et al.*, 2019).

While it is known that different legume species select for certain rhizobia species (Bromfield, Barran and Wheatcroft, 1995; Laguerre *et al.*, 2003; Wang, Liu and Zhu, 2018), it is less clear how much distinct varieties of the same legume species differ in their manipulation of rhizobium populations and how this variation affects the intraspecies rhizobium diversity. Plants can influence symbiont community compositions through multiple host-filtering mechanisms including by engaging in symbiotic specificity for nodule occupancy or by influencing rhizosphere community composition through interaction with specific root exudates (Jones *et al.*, 2019). However, it is still disputed whether *Rhizobium* genotype composition in nodules is mainly due to initial strain abundance in a natural soil population or a strain's competitive ability for nodulation (ZéZé, Mutch and Young, 2001). Studies that have aimed to elucidate the extent of symbiotic specificity between legume cultivars have predominantly used a restricted number of strains or introduced a synthetic inoculum community, with limited field crop applicability (Russell and Jones, 1975; Jones and Hardarson, 1979; C. Yang *et al.*, 2017). Under greenhouse conditions, clover varieties have been found to display preferences for specific rhizobial genotypes (Russell and Jones, 1975; Jones and Hardarson, 1979; C. Yang *et al.*, 2017). Genotypes of the same clover variety have also been shown to display significant variations in rhizobial selectivity (Russell and Jones, 1975; Jones and Hardarson, 1979). Conversely, another study observed that *Trifolium repens* (white clover) cultivars did not display preferences for different *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) genotypes under greenhouse conditions (Harrison, Young and Jones, 1987). However, allozyme variants were used to detect differences in population structure, which likely had a reduced association to the determinants of strain selectivity for symbiotic establishment (Harrison, Young and Jones, 1987). Using 'real' crop systems could be advantageous to help identify how genetically specific rhizobial inoculums need to be to achieve fit-for-purpose compatibility with legume host crops (Wadhwa, Dudeja and Yadav, 2011). For example, *nodD* genotype preferences have previously been indistinguishable between *Trifolium* species hosts and soil types when grown across different sites (McGinn *et al.*, 2016). Conversely, inoculation of four *Rlt* strains at various field sites showed competitive ability for *Trifolium subterraneum* cultivar nodule occupancy was associated with host genotype, field site, and bacterial strain compatibility (Roughley, Blowes and Hurridge, 1976). Therefore, while intra-species symbiont selectivity by different legume genotypes have been found previously, they are less often observed and studied in field conditions.

The compatibility between legume genotypes and rhizobia strains has previously been attributed to rhizobium differences in both the symbiosis plasmid and chromosome (Brewin, Wood and Young, 1983; Paffetti *et al.*, 1996). A variety of different gene markers have been used for determining rhizobia population diversity in both nodule and soil samples (Bromfield, Barran and Wheatcroft, 1995; Laguerre *et al.*, 2003; McGinn *et al.*, 2016). *rpoB* (RNA polymerase B subunit) and *recA* (recombinase A) have both been effectively used as robust chromosomal markers for observing intraspecies diversity and as phylogenetic determinants (Xiong *et al.*, 2017; Wang *et al.*, 2018). Several symbiosis genes have also been used to determine population structure based on symbiotic selection. *nodD* (transcriptional regulator of nodulation *nod* gene activation) has previously been chosen for analysis of *Rlt* populations as it can display genetic heterogeneity between *Rlt* strains, and is well characterised by its interspecies selectivity for symbiotic establishment (ZéZé, Mutch and Young, 2001; McGinn *et al.*, 2016). It is plausible that *nodD* would show distinction between clover genotypes, as *nodD* is well-known to play a large role in legume-*Rhizobium* interspecies partner compatibility and host range for symbiosis, although not with host specificity (Redmond *et al.*, 1986; Laguerre *et al.*, 1996; Perret, Staehelin and Broughton, 2000; ZéZé, Mutch and Young, 2001; Maj *et al.*, 2010; Hassan and Mathesius, 2012). While *nodA* (N-acyltransferase essential for successful Nod factor production) has been less commonly used, *nodA* diversity could provide insight into host-selection as it has been suggested to be involved in host specificity and interaction, and its regulation has been used to assess strain competitive abilities (Debellé *et al.*, 1996; Ritsema *et al.*, 1996; Maj *et al.*, 2010; Poinsot *et al.*, 2016; Igolkina *et al.*, 2019). Alterations to the molecular structure of different Nod factors has been shown to influence symbiotic specificity between rhizobia and legumes, and therefore the function of *nodA* by association is also suggested to influence clover host specificity (Lupwayi, Clayton and Rice, 2006; Wang, Liu and Zhu, 2018). Allelic differences in *nodA* have been shown to influence NodA specificity for different fatty acid substrates during N-acyl substitution, and consequently has been suggested to act as a component in host-specific nodulation and host range, at least at an interspecies level (Debellé *et al.*, 1996; Ritsema *et al.*, 1996; Perret, Staehelin and Broughton, 2000; Downie, 2014). In addition, *nodA* alleles were suggested to be more related to host plant taxonomy than bacterial taxonomy, further associating *nodA* allelic differences to host specificity (Debellé *et al.*, 1996; Igolkina *et al.*, 2019). Both *nodA* and *nodD* are plasmid-bound and can be transferred to unrelated strains through horizontal gene transfer. On the other hand, *rpoB* and *recA* are chromosomal housekeeping genes, thereby commonly conveyed through vertical transmission. Therefore, a broad reflection of diversity from horizontal and vertical genetic transmission can be perceived from using multiple chromosomal and plasmid genetic markers for analysis of population diversity (Wernegreen and Riley, 1999).

The aim of this study was to determine the extent of symbiotic specificity from an intraspecies perspective. Differences in *Rlt* genotype diversity were investigated between root nodule populations from five white clover genotypes grown under conventional field conditions using high-throughput amplicon sequencing (MAUI-seq). This was undertaken to provide a more representative reflection of the rhizobial diversity present between white clover genotypes in an agricultural application, and to understand the extent that symbiotic specificity was present at the intraspecies level. The relative allelic diversity of two *Rlt* chromosomal housekeeping genes, *rpoB* and *recA*, and two auxiliary plasmid-bound symbiosis genes, *nodA* and *nodD,* were analysed. It was found that housekeeping genes displayed the greatest distinction between clover genotype nodule samples. However, this was not necessarily a consequence of only host genotype, as geographic distance between sampled plants within the field was also found to be additionally associated with housekeeping gene nodule diversity. Symbiosis genes, *nodA* and *nodD*, showed no association with host genotype or geographic distance. Consequently, choice of gene marker highly influenced observed sample diversity and further identified that a combination of local growth conditions and host genotype can influence *Rlt* diversity of clover nodule populations in agricultural field conditions.

3.3. Methods

3.3.1. Plants, nodule sampling and DNA extraction

Four genetically distinct F2 *Trifolium repens* (white clover) variety crosses (Cross 1; Cross 2; Cross 3; Cross 4), and one pure check variety (Klondike) were grown in conventionally managed trial plots at Store Heddinge, Denmark. No F2 cross shared a parent variety with another cross. Klondike was similarly not a parent to any of the crosses, thereby making all of the crosses genetically distinct from one another and are from here on referred to as clover genotypes. The varieties used to generate the five crosses have been made confidential in this study for DLF Trifolium breeder trial developments.

All plots were sown under the same conditions in June 2017 in the same field. Plots were organised into two Blocks, each containing two rows of 18 plots, and plots were sown in dimensions of 8 meters by 1.5 meters (Figure 3.1). Within each Block, a strip of grassland measuring 3.3 meters separated the two rows of 18 plots. A strip of grassland measuring 8 meters separated the two Blocks of plots. Clover genotypes were sown within Blocks in a rectangular Latin plot design (Figure 3.1); a complete randomized block design with a restriction in the randomization, providing complete set of all clover genotypes in both directions within a Block of plots (Figure 3.1). Within Block 1, clover genotype plots were sown in duplicate (Figure 3.1). An additional third replicate Klondike plot was sampled from Block 2 to enable further geographic distance effects analyses (plot c in addition to a and b; Figure 3.1).

A mixture of 6 g clover genotype seed and 20 g of diploid perennial ryegrass varieties, Indiana and Boyn, were sown on each plot. No other plant species or *Rhizobium* inoculation were added to plots previously before sowing. In the establishment year, the trial plots received no fertiliser treatment. In the second year, plots received fertiliser treatment 4 times across the year, which totalled 170 kg N/harvest year across all plots. Around 3-10 clovers were sampled from three points in each plot in October 2018. Therefore, for each plot three independent replicate samples were collected, and six samples were collected in total for each clover genotype from Block 1. From Block 2, only 3 independent replicate samples were collected from one Klondike plot (plot c; Figure 3.1). For each plot point, clovers were washed, and 100 large nodules picked and pooled. Nodules were stored at -20°C, and DNA was extracted for each within-plot replicate using the Qiagen DNeasy PowerLyzer PowerSoil DNA isolation kit following manufacturer's protocol.

3.3.2. DNA sample processing and read processing.

Rhizobium leguminosarum symbiovar *trifolii* (*Rlt*) specific genes *rpoB*, *recA*, *nodA* and *nodD* were individually PCR amplified and processed for each pooled nodule sample using the MAUI-seq high-throughput amplicon sequencing method, as described in detail previously (Chapter 1; Fields *et al.*, 2019). Briefly, in this method genes are amplified in a nested PCR using primers that contain a region of 12 random bases in the forward inner-primer, which generates a unique molecular identifier (UMI) for each initial DNA strand in the first round of PCR amplification. All subsequent daughter DNA strands generated will contain the same UMI as their parent. This consequently means DNA reads with the same UMI can then be grouped and aid identification of erroneous sequences, such as chimeras and PCR mutations, during sequencing read processing. This is carried out with the aim to better reflect true allelic diversity of samples by identifying and further filtering out identified errors across samples.

Initial PCRs were carried out individually for each primer set using non-proofreading Platinum Taq DNA polymerase (Thermofisher Scientific Inc., USA). Equal volumes of the four PCRs produced for each sample were pooled, cleaned (AMPure XP Beads, Beckman Coulter, USA) and indexed for sequencing as previously (Nextera XT DNA Library Preparation Kit v2 set A, Illumina, USA; Phusion High-Fidelity DNA polymerase, Thermofisher Scientific Inc., USA) (Fields *et al.*, 2019). All samples were pooled, and quality checked by Bioanalyzer 2100 (Agilent, USA) before sequencing using Illumina MiSeq (2 x 300 bp paired end reads) by the University of York Technology Facility. Full method protocols including PCR reaction mixtures and programmes are detailed in Fields *et al.*, 2019 (Chapter 1).

Paired-end reads were first merged using the PEAR assembler (Zhang *et al.*, 2014). MiSeq reads were then processed using MAUI-seq python scripts to firstly separate reads into the 4 *Rlt* genes for each sample, and to secondly calculate the abundance of unique UMI reads for each gene in each sample. Raw read counts for each gene analysis were as follows: *nodA* = 1,942,182; *nodD* = 2,117,588; *recA* = 1,225,324; *rpoB* = 2,256,179. To calculate total abundances, reads were grouped by UMI and the most abundant read sequence (primary sequence) was assigned to that UMI, thereby removing PCR errors. Then, the number of UMIs associated to the same primary sequence were counted. Chimeras were detected by comparing the number of times a sequence appears as the most abundant sequence for a UMI compared to appearing as the second most abundant sequence (secondary sequence) in UMI clustering. Additional parameters used to control the stringency of the analysis were all set to default. These include: 1) count a UMI only if the most abundance sequence has 2 more reads than the second most abundant sequence; 2) reject sequences that occur as secondary sequences at least 0.7 times as often as they appear as primary sequences; 3) discard sequences with an overall relative abundance less than 0.001, when sequences are ordered in rank order. Therefore, sequence counts used in downstream analyses were the number of UMIs associated to each identified sequence (UMI sequence counts), rather than the number of reads for a sequence. Scripts can be found at https://github.com/jpwyoung/MAUI.

3.3.3. Sequence analysis

To enable allele abundance comparison across samples from all four genes, UMI sequence counts were converted to relative abundance within each gene for each sample. Sequence presence across clover genotypes was displayed with Venn diagrams made using R package, Venn (v.1.7). To observe relative abundance of allele sequences across samples in a heatmap, relative abundance counts of 0 (occurring when a sequence is not present in a sample but present in other samples) were converted to one decimal place lower than the lowest relative abundance count (1 x10^{-5}) and subsequently $\log_{10}$ transformed. Log transformation was used because most samples were dominated by two or three alleles, and therefore this would skew observed variance towards more abundant alleles within the population.

To assign a genospecies to each *recA* and *rpoB* allele, BLASTn was used to search for sequences in the genome assemblies of 196 *Rlt* full genome sequenced strains (Cavassim *et al.*, 2019), RCR221, TA1 and TA1-MC2010 reference genomes, which are known genospecies strains. Alleles that did not match any of these strains were aligned to the NCBI database using BLASTn (GenBank). If no 100% sequence match with known genome assemblies was found, sequences were classed as an 'unassigned genospecies'.

3.3.4. Statistical analysis

In all analyses, allelic similarity between nodule samples was estimated using Bray-Curtis dissimilarity. Bray-Curtis dissimilarity metric was calculated for all pairwise sample comparisons with vegan R package (v. 2.5-6) using relative abundance UMI sequence count data for all four genes. To compare intraspecies diversity between clover genotypes, non-metric multidimensional scaling (NMDS) was employed on Bray-Curtis dissimilarities using metaMDS in the vegan R package (v.2.5-6). Two dimensions were specified for NMDS analyses of all genes individually and in combination, which all produced an NMDS stress score of less than 0.2. Intrinsic sequence variable vectors were fitted with a default of 999 permutations to NMDS coordinates using the env.fit function in the vegan package, to determine which alleles associated significantly with NMDS dimensions. To determine significant differences in rhizobial allele diversity for clover genotypes, PERMANOVA was undertaken using the adonis R vegan function. Additionally, to further determine which clover genotypes significantly differed in Bray-Curtis dissimilarity, adonis.pairwise function with Bonferroni adjusted p-value correction was used from the pairwiseAdonis R package (v.0.0.1). Principal component analysis (PCA) was carried out using singular value decomposition using prcomp from the Stats R package (v.3.5.1), with $\log_{10}$ relative abundance counts (and 0 counts = 1 x10^{-5}). Global and pairwise empirical Bayes estimator of fixation index (F_{ST}) values between nodule DNA samples for all four genes were calculated using relative allele abundances with FinePop R package (v.1.5.1), as previously described (Kitada, Nakamichi and Kishino, 2017).

To identify significant differences in the relative abundance of genospecies across clover genotypes for *recA* and *rpoB*, two-way ANOVAs were undertaken. Furthermore, TukeyHSD post hoc from the tidyverse R package (v.1.2.1) was used to identify interaction effects between clover genotype and genospecies relative abundances.

To determine if allelic diversity similarity between samples was associated with geographic distance, the Mantel test Pearson's correlation R statistic was calculated between Euclidean geographic distance and allelic (Bray-Curtis) dissimilarity between samples. Geographic distance between sampled plot points was calculated using Euclidean distance calculation based on x-y geographic coordinates. Similarly, the Mantel test was used to calculate between geographic distance and pairwise F_{ST} between samples.

In order to analyse whether allelic similarity between samples was affected by geographic distance and whether samples were the same clover genotype, a maximum likelihood (ML) mixed effect model was generated for each of the four *Rlt* genes individually (*rpoB*, *recA*, *nodA* and *nodD*) with lme4 R package (v.1.1-21). Euclidean geographic distance and a binary metric of whether samples were isolated from clovers that were the same or different genotype (same clover genotype = 1, different = 0) were classed as fixed effects. The clover genotype IDs of both samples (e.g. Cross 1) were categorised as cross random effects. The random effects had a variance > 0 supporting their incorporation in the model. The linear mixed models enabled accounting for variability due to clover genotype and ability to model multiple random effects simultaneously, which was undertaken to reduce error in the models and increase ability to determine significance of fixed effects. LmerTest was used to generate t-values, degrees of freedom and p-values for each fixed effect variable. The four models were generated and analysed for each gene individually using following pipeline. Firstly, the full factorial models were generated whereby both fixed effects (geographic distance and clover genotype binary metric) and their interaction were included, along with the clover genotype of sample pairs as random effects. The importance of the fixed effects interaction was tested using the likelihood ratio (LR) test using anova() by comparing the full interaction model to a reduced model with no interaction. No significance of interaction was found between geographic distance and clover genotype on allelic dissimilarity for any of the four *Rlt* genes. Subsequently, the full model without interaction was used in further analysis (referred to as full model hereon and in the results). Secondly, the importance of geographic distance and whether the samples were isolated from the same clover genotype were determined by LR test. This tested for significant differences in model fit between a model with geographic distance and clover genotype and as fixed effects and a reduced model without either fixed effect in a stepwise manner. Model fits were determined as significantly different if Chi-squared $p < 0.05$.

Additionally, the importance and reliability of the fixed effects were determined by parametric bootstrapping to obtain 95% confidence intervals for the full model with no interaction. bootMer and boot.ci in boot R package was used to obtain 95%

confidence intervals with 1000 bootstraps. 95% confidence intervals that included 0 were considered not reliable effects, as this suggests that there was no clear association of the fixed effect with the allelic dissimilarity between samples. The bootstrapping model displayed warnings of failed model convergence for each gene: *rpoB* = 55 out of 1000 permutations; *recA* = 62 out of 1000 permutations; *nodA* = 68 out of 1000 permutations. However, because the original model converged, and due to developers suggesting increasing the convergence warning message threshold to 0.01 these were classified as false positive convergence warnings (Bolker, 2020). Similarly, warnings regarding singular fits were generated from bootstrapping for the following genes: *rpoB* = 79 out of 1000 permutations; *recA* = 149 out of 1000 permutations; *nodA* = 498 out of 1000 permutations; *nodD* = 1000 out of 1000 permutations. For the *nodD* full model without fixed effects interaction, lme4 generated warnings of a singular fit which was caused by one of the random effects producing a variance and standard deviation of 0. The random effect was maintained in the full model and subsequent reduced models because retaining the parameter made no difference to estimate quantities (apart from AIC/BIC) and also allowed congruence of model formula with the other gene models (Bolker *et al.*, 2009; Bolker, 2020).

3.4.Results

3.4.1. The presence of rhizobia alleles showed only small variation between clover genotypes

Using high-throughput amplicon sequencing (Fields *et al.*, 2019), nodule samples from five white clover genotypes grown in a conventional trial field management (Figure 3.1) were sequenced for four *Rlt* genes (*rpoB*, *recA*, *nodA* and *nodD*). Per clover genotype, six sample replicates were collected across two plots in Block 1 (Figure 3.1). Firstly, the number of unique allele sequences identified were counted for *rpoB*, *recA*, *nodA* and *nodD Rlt* genes across all samples from Block 1 (Appendix Table B.1). A greater number of alleles were identified for symbiosis genes than housekeeping genes: *rpoB* = 16; *recA* = 8; *nodA* = 23; *nodD* = 21 (Appendix Table B.1). To determine whether rhizobial allelic diversity in white clover nodule samples was associated with clover genotype, the total number of unique alleles for each of the five distinct clover genotypes across all genes was counted (Figure 3.2a-b). *rpoB* and *recA* allele presence showed little variation across clover genotypes (Figure 3.2c-d). *rpoB* was the most homogeneous, with all *Rlt* alleles identified in all clover genotypes with the exception of two alleles which were shared by all clover genotypes except Cross 1 (Figure 3.2c). Symbiosis genes, *nodA* and *nodD*, displayed more specificity to specific clover genotypes, with some alleles only present in the nodules of a subset of clover genotypes (Figure 3.2e-f). Even so, only one *nodD* allele was found to be exclusive to a single cross (Figure 3.2f: Cross 1). Otherwise, all *nodD* and *nodA Rlt* alleles were identified in nodules of at least two clover genotypes, and overall alleles were predominantly found in all clover genotypes. On the whole, there was no clear distinction of allele presence between different clover genotype nodules. This suggests that, based on the four genes tested, clover genotypes did not exclusively select for specific *Rlt* alleles.

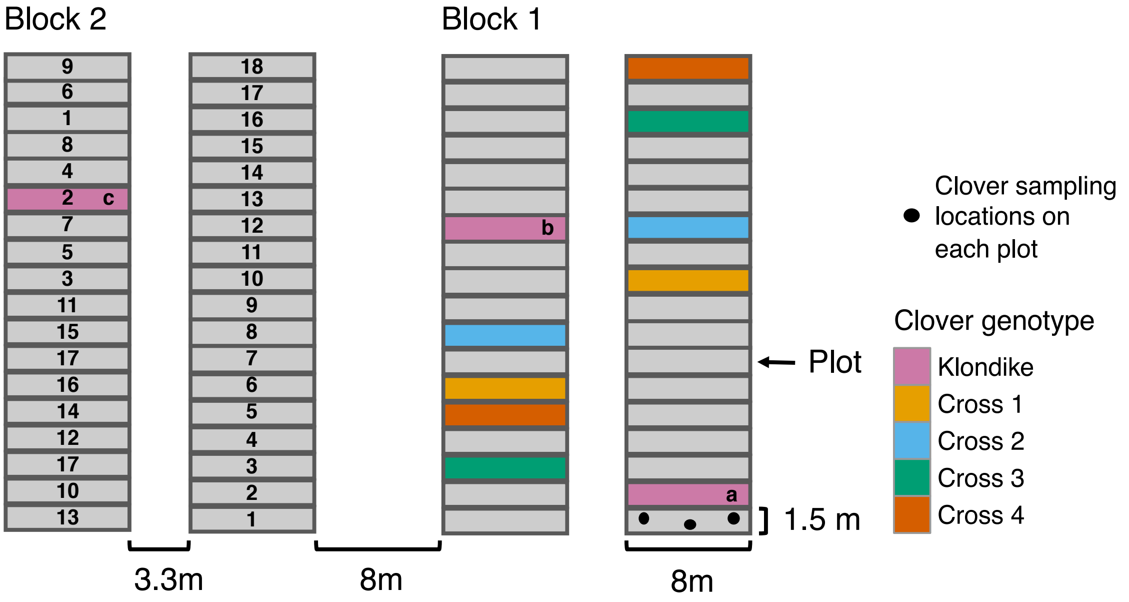


Figure 3.1 Field plot design for sampling. Plots were organised into 2 blocks, each containing 2 rows of 18 plots (grey rectangles), sown in dimensions of 8 meters by 1.5 meters. Blocks were separated by a strip of grassland measuring 8 meters across. Additionally, within Blocks, the two rows of plots were separated by grassland measuring 3.3m across. Plots sampled in this study are coloured respectively by the white clover genotype sown on each plot (see the legend on the right). Two plots were sampled per clover genotype from Block 1, with the exception of Klondike where an additional third plot was sampled from Block 2 to enable further analysis of geographic distance (plot c in addition to plots a and b from Block 1). Clovers were sampled from three locations on each plot (Black dots in the most bottom-right plot). Therefore, 6 samples were collected for each clover genotype from Block 1. 100 nodules were sampled from 3-10 clover plants for each sampling point. Clover genotypes were sown within Blocks in a rectangular Latin plot design, as demonstrated by the numbering system outlined in Block 2, whereby a number represents a clover genotype.

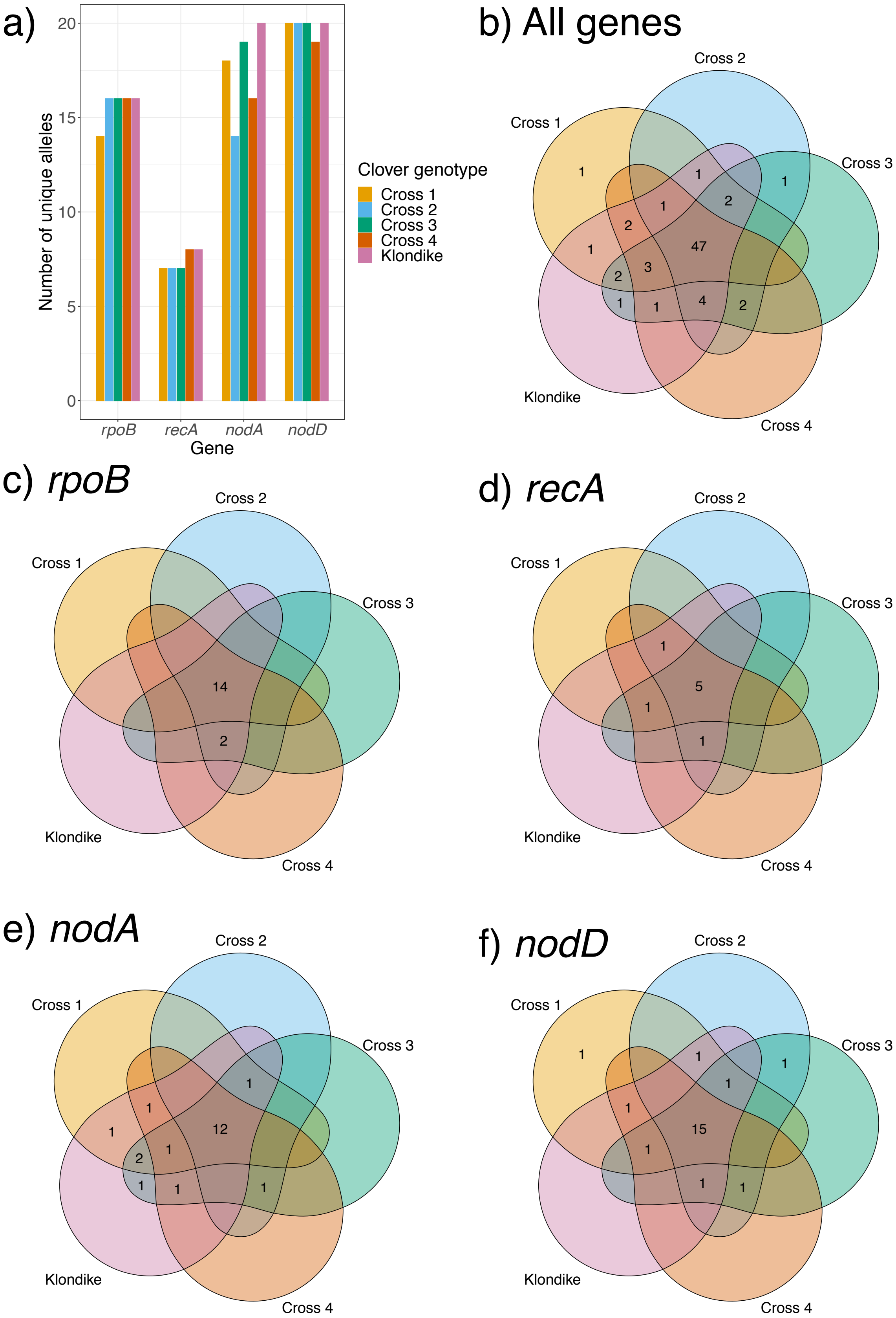


Figure 3.2 a) Unique alleles were identified for four *Rlt* genes (*rpoB*, *recA*, *nodA* and *nodD*) across root nodules from five white clover genotypes. Venn diagram display the overlap of alleles identified in each white clover genotype for, **b)** all genes, **c)** *rpoB*, **d)** *recA*, **e)** *nodA*, and **f)** *nodD* genes. All samples analysed are from Block 1.

3.4.2. Relative allele frequencies of housekeeping genes are more distinct than symbiosis genes between clover genotypes

Due to the homogeneity of allele presence in all clover genotype root nodules, it was investigated whether *Rlt* allele frequencies differed between clover genotypes. Therefore, the relative abundances of *Rlt* alleles for *rpoB*, *recA*, *nodA* and *nodD* were calculated for all samples (Figure 3.3). It was found that for each *Rlt* gene, the relative abundance of alleles varied between clover genotypes (Figure 3.3). However, the most abundant gene alleles seemed to be the same across all clover genotype samples. To further observe the differences in allelic composition between samples, allelic dissimilarity was calculated using Bray-Curtis dissimilarity (referred here on as allelic dissimilarity in all analyses) for all four *Rlt* genes individually and in combination across pairwise sample combinations (Figure 3.4a, Appendix Figure B.1 and Appendix Figure B.2). Non-metric multidimensional scaling (NMDS) analysis of allelic dissimilarities identified some separation between clover genotypes when the relative allelic abundances for all four genes was combined (Figure 3.4b). NMDS suggested a gradient-like separation of clover genotypes across NMDS coordinate 1 (Figure 3.4b). Significant differences in rhizobial allele diversity were identified between clover genotypes (all genes PERMANOVA clover genotype: $F_{4,29} = 3.7036$, $p < 0.001$; Appendix Table B.2). However, only Cross 1 and Cross 4 were identified to have significantly different *Rlt* diversity in post hoc testing (adjusted $p < 0.05$).

Further NMDS analysis of allelic dissimilarities identified that different clover genotypes separated depending on the individual gene of interest (Figure 3.4c-f; *rpoB* PERMANOVA clover genotype, $F_{4,29} = 5.375$, $p < 0.001$; *recA* PERMANOVA clover genotype, $F_{4,29} = 5.247$, $p < 0.01$; *nodA* PERMANOVA clover genotype, $F_{4,29} = 2.678$, $p < 0.05$; *nodD* PERMANOVA clover genotype, $F_{4,29} = 2.123$, $p < 0.05$; Appendix Table B.3-Appendix Table B.6). *rpoB* showed significant differences in allelic diversity of Cross 2 compared with Klondike, Cross 1 and Cross 4 (adjusted $p < 0.05$; Figure 3.4c). On the other hand, while *recA* showed less allelic distinction between clover genotypes, significant differences were identified between Cross 2 and Cross 1, which is congruent with the allelic distinction of *rpoB* alleles (adjusted $p < 0.05$). For *recA* alleles, the variation observed across NMDS coordinate 1 was biased by Klondike samples (Appendix Figure B.3), where half of the Klondike samples predominantly contained the overall most abundant *recA* allele and the other samples mainly contained the second overall most abundant allele. Therefore, when Klondike samples were removed from the analysis, the similar separation of Cross 2 and Cross 1 clover genotypes were clearly observable (Figure 3.4d). On the other hand, for *nodA* and *nodD*, significant allelic differences between clover genotypes was lost after post hoc testing p-value correction. To investigate which allele sequences were driving the sample distribution patterns for each of the four *Rlt* genes, the intrinsic allele sequence variable vectors were fitted to NMDS coordinates. The frequencies of the overall most abundant alleles were shown to significantly drive the separation of clover genotypes (Appendix Figure B.4). These results were further confirmed using principal components analysis that yielded qualitatively similar results to the NMDS analysis, but clover genotypes were not as distinguished (Appendix Figure B.5).

Additionally, in order to further determine which *Rlt* gene most greatly influenced the population structure, fixation index (F_{ST}) was calculated for all four *Rlt* genes individually and as a combined value. No clear pattern was discernible between clover genotypes (Appendix Figure B.6). However, global F_{ST} calculations, which also did not consider the clover genotype or geographic location of samples *a priori*, suggested that population structure was predominantly determined by *nodA* compared to the three other genes (Appendix Table B.1).

Overall, the differences in *Rhizobium* allelic variation could be distinguished between some, but not all clover genotypes. Observed diversity was found to be largely depended on the gene of interest (*rpoB*, *recA*, *nodA* or *nodD*), and *rpoB* and *recA* alleles showed the greatest allelic distinction specifically between Cross 1 and Cross 2 clover genotypes.

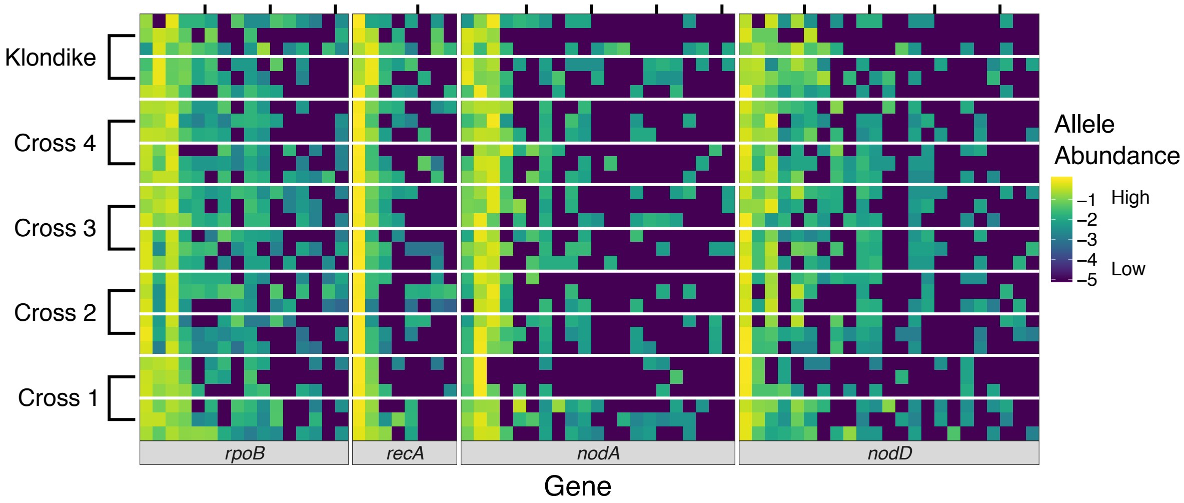


Figure 3.3 Relative abundance of unique alleles identified for four *Rlt* genes (*rpoB*, *recA*, *nodA* and *nodD)* varied in root nodule populations from five white clover genotypes (Cross 1-4, Klondike). Raw UMI sequence counts for each allele were converted to relative abundances (between 0-1) and subsequently $\log_{10}$ transformed for visualisation. Relative abundance counts of 0 were converted to $1x10^{-5}$ (one decimal place lower than the smallest relative abundance value across all samples) before log transformation. Therefore, log transformation produces a negative abundance score, whereby more negative scores denote for a lower allele abundance is (yellow = high abundance, blue = low abundance). Clover genotype samples were collected across two plots (displayed as separate row sections), and 3 locations were sampled within each plot (displayed as three rows within each plot section). Only samples from Block 1 are shown. Tick marks at the top of the heatmap indicate every 5 alleles.

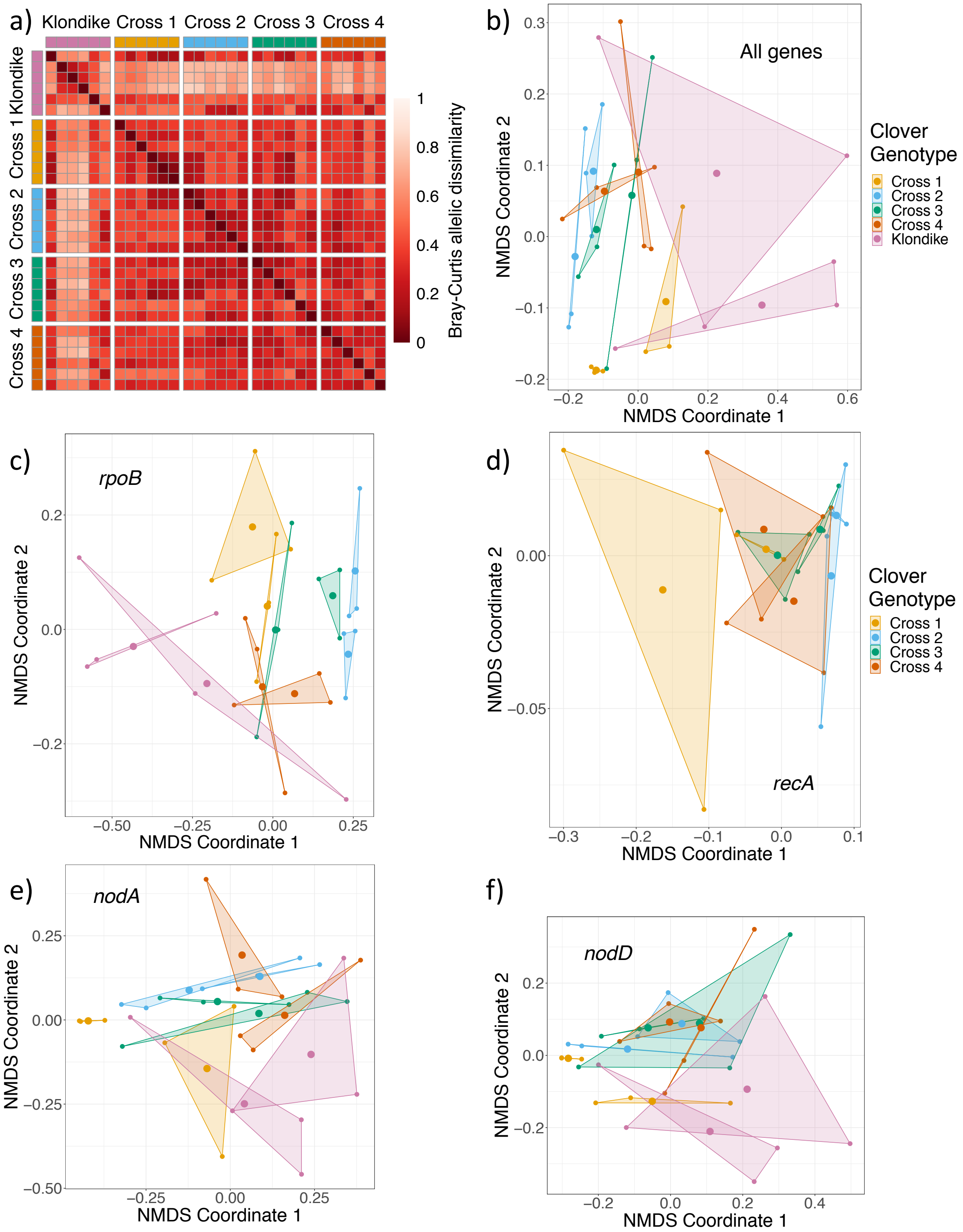


Figure 3.4 The level of observed allelic dissimilarity between clover genotypes differed depending on the *Rlt* gene marker. **a**) Pairwise allelic dissimilarity of four *Rlt* genes combined (*rpoB*, *recA*, *nodA* and *nodD*) between white clover nodule samples. Bray-Curtis dissimilarity is shown on a scale ranging from low (red) to high (white) allelic dissimilarity. Additionally, Non-metric Multi-Dimensional Scaling (NMDS) analysis of the relative abundances of four gene alleles **b)** in combination, and individually; **c)** *rpoB*, **d)** *recA*, **e)** *nodA*, **f)** *nodD* displays the separation of sampled based on their allelic dissimilarity. Allelic dissimilarity is displayed across two dimensions, and samples that are closer are more allelically similar. Samples are from Block 1 and are grouped by their respective plot (n = 3, and 2 plots per clover genotype) and coloured by their clover genotype host.

3.4.3. *Rlt* genospecies frequencies differ between clover genotypes

In order to study whether different white clover genotypes preferentially selected for specific *Rlt* genospecies, *rpoB* and *recA* allele sequences were assigned to a genospecies (gsA-E) (Kumar *et al.*, 2015), and relative genospecies frequencies in nodule samples were calculated (see Methods).

The top four most abundant *rpoB* and *recA* allele sequences could all be assigned to known genospecies using the 196 *Rlt* genome dataset (Cavassim *et al.*, 2019); in total, 9 out of 16 *rpoB* sequences, and 5 out of 8 *recA* sequences, were assigned to genospecies. Other sequences did not largely contribute to the overall relative abundances, but they all exceeded 1% relative ranked abundance in at least one sample. Therefore, the remaining unassigned sequences were searched for against the GenBank database in order to try to assign them at the genospecies level. No remaining allele sequences matched to GenBank whole genome assemblies with known genospecies identities, although some allele sequences matched 100% identity to single GenBank sequences (i.e. not from a full genome assembly). However, genospecies identity could not be determined based on these sequences. Therefore, all remaining allele sequences (7 out of 16 *rpoB* sequences and 3 out of 8 *recA* sequences) were classified as 'unassigned genospecies'.

The genospecies frequencies determined from *rpoB* allele sequences significantly strongly correlated with genospecies frequencies determined from *recA* allele sequences (Pearson's Correlation: R = 0.951, t = 41.193, df = 178, p < 0.001). However, a greater number of *recA* alleles could be associated to gsA, whereas no gsA alleles were identified for *rpoB*. Similarly, *rpoB* alleles were more often associated with gsC, and *recA* alleles associated with gsB, suggesting a potential bias in amplification of some templates depending on the gene of interest.

Noticeably, all clover genotypes were significantly dominated by either gsB or gsC in their nodules (Figure 3.5; *recA* ANOVA genospecies, $F_{5,150}$ = 365.597, p < 0.001; *rpoB* ANOVA genospecies, $F_{5,150}$ = 210.43, p < 0.001; Appendix Table B.7-Appendix Table B.8). Similarly, a significant interaction was identified between genospecies abundances and clover genotype for both *recA* and *rpoB* alleles (*recA* ANOVA genospecies*clover genotype: $F_{20,150}$ = 7.144, p < 0.001; *rpoB* ANOVA genospecies*clover genotype: $F_{20,150}$ = 10.25, p < 0.001). In particular, the check variety, Klondike, contained a significantly lower percentage of gsB *rpoB* and *recA* alleles in nodules compared to all F2 Crosses ($p < 0.05$). Similarly, Klondike nodules also contained a significantly greater percentage of *rpoB* and *recA* gsC alleles, compared to all other F2 Crosses ($p < 0.05$). No significant difference was found in genospecies composition between F2 Crosses for *recA* or *rpoB*, with the exception of Cross 1 *rpoB* alleles containing a significantly greater percentage of gsC, and significantly lower percentage of gsB, compared to Cross 2 ($p < 0.05$), but this was not significantly shown for *recA* alleles. In all samples, gsA, gsD and gsE were present in low abundance compared to gsB and gsC and totalled less than 16% of *recA* and *rpoB* allele representation in nodule samples. Together, these results show that genospecies can differ between clover genotypes, and this was predominantly identified between the check variety, Klondike, compared to the other F2 Crosses.

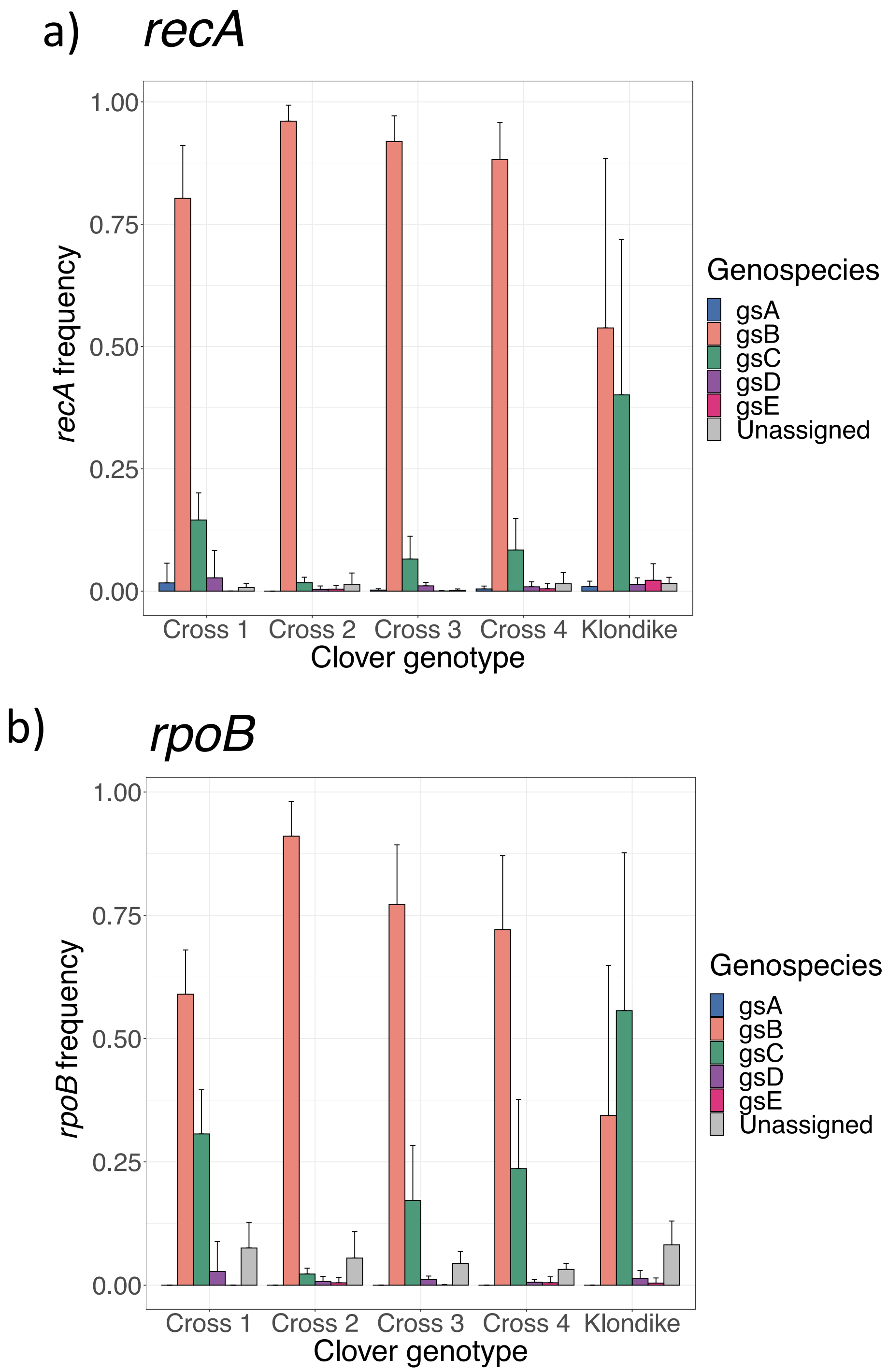


Figure 3.5 The mean relative abundances of genospecies A-E allele sequences within each clover genotype was calculated for **a)** *recA* alleles or **b)** *rpoB* alleles. Klondike samples contained a significantly greater proportion of gsC alleles compared to other crosses, whereas Cross 2 contained a significantly lower proportion of gsC alleles compared to other clover genotypes. Allele sequences which could not be assigned to a genospecies were labelled as ‘Unassigned’. All clover genotypes were sampled in replicates of 6 from Block 1.

3.4.4. Similarity of housekeeping gene allele frequencies is predominantly associated with geographic distance

To test whether allelic similarity between samples was associated with geographic distance or the plant genotype that samples were isolated from, a linear mixed effects model was undertaken for each *Rlt* gene (*rpoB*, *recA*, *nodA* and *nodD*; Appendix Table B.9).

Geographic distance was found to be significantly associated with *rpoB* allelic dissimilarity (Figure 3.6a; *rpoB* $\text{Coeff}_{\text{Geographicdistance}}$ estimate = 0.003, std. error = 0.001, t = 2.894, p < 0.01). Whether samples were isolated from the same clover genotype was also suggested to significantly associate with *rpoB* allelic dissimilarity, with samples of the same clover genotype showing significantly similar *rpoB* allelic diversity than sample pairs from different clover genotypes (*rpoB* $\text{Coeff}_{\text{Genotypedifference}}$ estimate =-0.059, std. error = 0.020, t =-2.967, p < 0.01). However, parametric bootstrapping of confidence interval parameters showed that only geographic distance was a reliable predictor (Table 3.1). Despite this, clover genotype was found to significantly affect model fit (*rpoB*: $X^2_{6,5}$ = 8.653, p < 0.01). Therefore, this suggests that for *rpoB* both geographic distance and clover genotype influences *rpoB* diversity in clover nodules, but geographic distance has a potentially stronger association than clover genotype similarity.

Similarly to *rpoB*, geographic distance was found to be significantly associated with *recA* allelic dissimilarity (Figure 3.6b; *recA* $\text{Coeff}_{\text{Geographicdistance}}$ estimate = 0.003, std. error = 0.001, t = 2.157, p < 0.05). However, *recA* diversity was found to not be significantly influenced by whether samples were isolated from the same clover genotype. LR test and parametric bootstrapping of confidence intervals further confirmed that geographic distance was a reliable effect and improved model fit (Table 3.1; $X^2_{6,5}$ = 4.532, p < 0.05). This suggests that geographic distance is significantly associated with *recA* diversity, whereas the clover genotype from where the strains were isolated had no significant effect.

For symbiosis genes, *nodA* and *nodD*, no significant association was identified between allelic dissimilarity and geographic distance or clover genotype (Appendix Figure B.7). This was confirmed by the following: LR test between full and reduced models removing either clover genotype or geographic distance fixed effects; (*nodA*$_{\text{Genotypedifference}}$ reduced model: $X^2_{6,5} = 3.163$, $p > 0.05$; *nodA*$_{\text{Geographicdistance}}$ reduced model: $X^2_{6,5} = 0.119$, $p > 0.05$; *nodD*$_{\text{Genotypedifference}}$ reduced model: $X^2_{6,5} = 0.478$, $p > 0.05$; *nodD*$_{\text{Geographicdistance}}$ reduced model: $X^2_{6,5} = 0.126$, $p > 0.05$); p-values calculated for fixed effects parameters within models ($p > 0.05$); and parametric bootstrapping for 95% confidence intervals of full model parameters included 0 (Table 3.1).

To further observe the extent to which geographic distance was associated with *Rlt* allelic diversity, the allelic diversity of the six Klondike samples already evaluated from Block 1 (Klondike plots a and b; Figure 3.1) were compared to three additional nodule samples collected from another replicate Klondike plot located within Block 2 which is separated by a larger geographical distance to the other plots (Klondike plot c; Figure 3.1). NMDS analysis identified that the *Rlt* allelic diversity of the Block 2 plot (c) was distinctly different to the other two Klondike plots from Block 1 (a and b) (Figure 3.6c-d). To further support this, PCA showed a similar separation of Klondike plot c from plots a and b (Appendix Figure B.5). Additionally, allelic dissimilarity and pairwise F_{ST} calculated from relative abundances of all four genes combined significantly correlated with geographic distance between samples (Figure 3.6e-f; allelic dissimilarity Pearson's correlation $R = 0.314$, $p < 0.05$; F_{ST} Pearson's correlation $R = 0.348$, $p < 0.05$). This indicated that differences in *Rlt* allelic diversity between samples was likely to be partially driven by distance between sampling points, even for the same clover genotypes (Figure 3.6c-f).

Table 3.1 Parametric bootstrapping of fixed effects 95% confidence intervals for each Rlt gene (*rpoB, recA, nodA and nodD*) mixed effects model.

Model	**Fixed effect**	**95% Confidence Interval**	**Original Beta Estimate**	**Bias**	**Std. error**
rpoB	Geographic distance	0.0012, 0.0055	0.003	-0.00001	0.001
	Clover genotype difference	-0.0986, -0.0196	-0.059	0.0005	0.020
recA	Geographic distance	0.0001, 0.0056	0.003	-0.00003	0.001
	Clover genotype difference	-0.0619, 0.0368	-0.013	-0.00008	0.025
nodA	Geographic distance	-0.0025, 0.0034	0.0005	-0.00002	0.001
	Clover genotype difference	-0.0928, 0.0019	-0.044	-0.0005	0.024
nodD	Geographic distance	-0.0027, 0.0019	-0.0004	-0.00005	0.001
	Clover genotype difference	-0.0575, 0.0258	-0.015	-0.0006	0.021

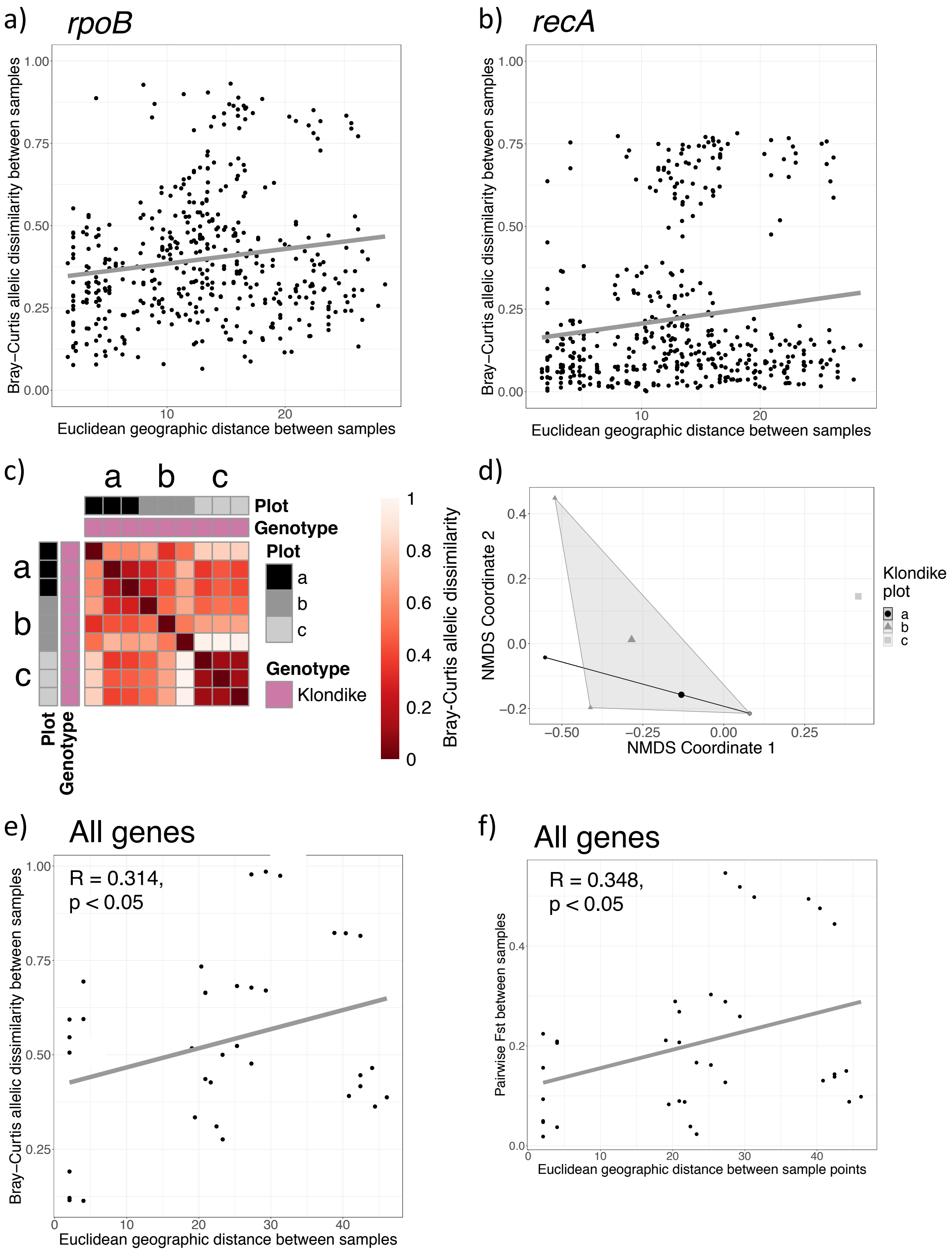


Figure 3.6 Geographic distance was associated with allelic dissimilarity between samples. **a)** Euclidean geographic distance correlated to *rpoB* allelic dissimilarity for all pairwise sample comparisons from Block 1. **b)** Euclidean geographic distance correlated to *recA* allelic dissimilarity for all pairwise sample comparisons from Block 1. **c)** Pairwise allelic dissimilarity was calculated between all Klondike samples across 3 plots (plots a, b from Block 1 and c from Block 2). Allelic dissimilarity is shown on a scale ranging from low (red) to high (white) allelic dissimilarity. **d)** Non-metric Multi-Dimensional Scaling (NMDS) of allelic dissimilarity of all four genes in combination from Klondike nodule samples across 3 plots (plots a, b from Block 1 and c from Block 2). **e)** Euclidean geographic distance correlated to all genes combined allelic dissimilarity for pairwise Klondike sample comparisons across 3 plots (R = 0.314, p < 0.05). **f)** Euclidean geographic distance correlated to all genes combined F_{ST} for pairwise Klondike sample comparisons across 3 plots (R = 0.348, p < 0.05). Allelic (Bray-Curtis) dissimilarity and F_{ST} were calculated using relative abundance UMI sequence counts for each of the four *Rlt* genes (*rpoB*, *recA*, *nodA* and *nodD*).

3.5.Discussion

This study investigated whether clover genotypes were associated with particular symbiont genotypes within one field, and if this pattern was affected depending on the specific gene under investigation or the local growth conditions (distance between sampled plots). *Rlt* nodule populations from five genetically distinct white clover genotypes, grown in conventional trial field conditions, were evaluated for differences in *Rlt* allele composition based on two chromosomal and two symbiosis genes (*rpoB*, *recA*, *nodA* and *nodD*). When the allelic diversity of all genes was considered together, *Rlt* diversity in samples clustered to some extent by clover genotype host. When *Rlt* allelic diversity was evaluated on an individual gene basis, housekeeping genes *rpoB* and *recA* showed a greater distinction between clover genotypes than symbiosis genes. Further analysis identified some clover genotypes displayed significantly different relative proportions of *Rlt* genospecies in nodules. Additionally, diversity of *Rlt* genes was not necessarily only a result of the host clover genotype, and the similarity of *rpoB* and *recA* diversity between samples was significantly associated with geographic distance between sampled plants (higher dissimilarity with increasing sampling distance). This suggests that variation is driven by a combination of clover genotype and local growth conditions in the field, and therefore local microenvironmental variation is also likely important for explaining intraspecific symbiotic diversity in the rhizosphere.

3.5.1. White clover genotype symbiotic selectivity under field conditions

No clear distinction of allele presence or absence was identified between clover genotypes, although a larger number of unique alleles and greater heterogeneity was found for presence of symbiosis gene alleles than housekeeping genes (Figure 3.2). For each of the four *Rlt* genes, all nodule samples were found to predominantly consist of the same few dominating alleles with many additional sequences present at lower abundances, as similarly identified with previous studies (Leung, Wanjage and Bottomley, 1994; ZéZé, Mutch and Young, 2001; Laguerre *et al.*, 2003; Fagerli and Svenning, 2005). It has previously been suggested that isolates from large nodules are usually effective with a wide range of white clover genotypes, which could explain the homogeneity in allele presence across clover genotypes in this study (Mytton, 1975). However, the effect of host genotype on microbial diversity has been demonstrated through variation in abundance of many OTUs, rather than based on the presence of single alleles alone (Bulgarelli *et al.*, 2015). The variation in clover genotype distinction across the four genes suggests that use of multiple genes is essential for providing a better view of differences between clover genotypes, and observed variation is fundamentally dependent on choice of marker genes. This conclusion is further supported by that the compatibility between legume-rhizobium genotypes has been attributed to rhizobial differences in both chromosomal and symbiosis genetic diversity (Brewin, Wood and Young, 1983; Paffetti *et al.*, 1996).

When relative abundances of *rpoB*, *recA*, *nodA* and *nodD* were considered in combination, differences in rhizobial genotype diversity was observed between some clover genotypes, but not all (Figure 3.4b). At the individual gene level, clover genotypes significantly separated based on nodule sample allelic dissimilarity of housekeeping genes, *rpoB* and *recA* alleles, but not *nodA* and *nodD* (Figure 3.4c-f). The *rpoB* and *recA* allelic dissimilarity between samples was assumed to be predominantly driven by changes in allele frequencies rather than allele presence/absence, as all alleles were present in nearly all clover genotypes. Mixed effects models further confirmed that *rpoB* allelic diversity was more similar between samples from the same clover genotype than different clover genotypes. However, for all other genes no significant increase in diversity similarity was found between samples of the same clover genotype. The results of this study are in line with previous findings where it was reported that *Rlt* selection varied between different varieties of clover and also that significant variation in *Rlt* selection was observed even between plants of the same variety (Russell and Jones, 1975; Jones and Hardarson, 1979). Other investigations have similarly found host genotype significantly influenced rhizobia partner choice and nodule populations in multiple legume species (Mytton, 1975; Russell and Jones, 1975; Paffetti *et al.*, 1996; Wadhwa, Dudeja and Yadav, 2011; Bourion *et al.*, 2018). White clover varieties have also been found to differ in their preference for specific *Rlt* strains, but this was only when inoculated with a simplistic two-strain community (Jones and Hardarson, 1979). Additionally, results have been found previously where legume plants were shown to form more productive symbioses with rhizobia isolated from more genetically related plants than with rhizobia isolated from more distantly related plants, suggesting that the preference for rhizobium strains is genetically influenced by the host, and potentially even at the intraspecies level (Mytton, 1975; Jones and Hardarson, 1979).

Contrastingly, some studies have not found any associations between cultivar and rhizobium strains (Bromfield, 1984; Harrison, Young and Jones, 1987; Buttery, Park and van Berkum, 1997; McGinn *et al.*, 2016). The relative diversity of nodule communities is therefore likely effected by both gene marker and experimental design. As a result, a combination of different gene markers should be utilised to better capture this diversity rather than through evaluation of a single gene marker. Furthermore, observing diversity in applicable agricultural experiments likely reduces observed *Rlt* community differences between cultivars due to the influence multiple additional environmental factors compared to investigations in controlled greenhouse conditions.

While housekeeping genes showed the greatest differences between clover genotypes in this study, the symbiosis genes showed no significant distinction between clover genotypes. The interspecies specificity in the legume-rhizobia symbiosis is predominantly determined from the interaction between NodD and legume flavonoids, which if compatible, enable the activation of nodulation (*nod*) genes to begin Nod factor (lipochitooligosaccharide) production for initiation of symbiotic establishment (Redmond *et al.*, 1986; Perret, Staehelin and Broughton, 2000; Maj *et al.*, 2010; Hassan and Mathesius, 2012). Therefore, it was predicted that differences in *nodD* allelic diversity would be observed between clover genotypes due to the importance of *nodD* for determining legume-*Rhizobium* interspecies partner compatibility for interspecies level symbiotic establishment (Redmond *et al.*, 1986; H P Spaink *et al.*, 1987; Laguerre *et al.*, 1996; Perret, Staehelin and Broughton, 2000; ZéZé, Mutch and Young, 2001; Maj *et al.*, 2010; Hassan and Mathesius, 2012). Other studies have also identified that *nodD* genotype preferences were indistinguishable between *Trifolium* species hosts (McGinn *et al.*, 2016), although faba bean cultivars have been suggested to preferentially select for different *nodD* genotypes under greenhouse conditions (Xiong *et al.*, 2017). As this was not observed in this study it suggests that differences in *nodD* nodule diversity may only be evident at the interspecies level for white clover, or this specificity becomes unclear in an agricultural setting. Other known legume-rhizobia symbioses have been used to evaluate cultivar x strain interaction differences utilising around 2-5 strains as an inoculum under sterile greenhouse conditions or using collected soil from different geographical areas (Russell and Jones, 1975; Roughley, Blowes and Hurridge, 1976; Jones and Hardarson, 1979; C. Yang *et al.*, 2017; Bourion *et al.*, 2018). While simplistic rhizobium selection can be identified between cultivars under restricted conditions with synthetic minimal communities, an inoculum with more strains or growth in natural field conditions might better reveal if identified host preferences for rhizobium population genotypes are maintained in application (Wadhwa, Dudeja and Yadav, 2011).

Similarly, for *nodA* no significant differences in diversity were observed between clover genotypes at the intraspecies level. However, *nodA* allele sequences were found to cause the most population structure from global Fixation index (F_{ST}) estimates (Appendix Table B.1). It was predicted that differences in *nodA* diversity would be found between white clover genotypes. This was because in previous studies *nodA* allelic differences have been shown to influence NodA specificity for different fatty acid substrates during N-acyl substitution, and consequently *nodA* has been suggested to act as a component in host-specific nodulation and host range, at least at an interspecies level (Debellé *et al.*, 1996; Ritsema *et al.*, 1996; Perret, Staehelin and Broughton, 2000; Lupwayi, Clayton and Rice, 2006; Downie, 2014; Wang, Liu and Zhu, 2018). However, the contrary was observed at the intraspecies level. *nodA* alleles have also been suggested to be more related to host plant taxonomy than bacterial taxonomy, further associating *nodA* allelic differences to host specificity (Debellé *et al.*, 1996; Igolkina *et al.*, 2019). Despite this, types of rhizobial Nod factors do not strongly correlate to the plants they initiate symbiosis with. For example, major Nod factors secreted by *Rlt* are also secreted by *Rlv* even though they form symbiosis with different legume species (Perret, Staehelin and Broughton, 2000). The majority of studies also agree that a large amount of symbiotic variability and rhizobial genetic diversity is observed at the inter- and intra-species level between legume hosts (Russell and Jones, 1975; Bromfield, 1984; Harrison, Young and Jones, 1987; Wadhwa, Dudeja and Yadav, 2011; McGinn *et al.*, 2016; C. Yang *et al.*, 2017; Kazmierczak *et al.*, 2017; Bourion *et al.*, 2018). As a result, further research into the associations between cultivar host-specificity and *nodA* would aid understanding of its importance for intraspecies symbiotic establishment.

This study also investigated whether the relative abundances of *Rlt* genospecies differed between clover genotypes (Kumar *et al.*, 2015; Cavassim *et al.*, 2019). This was undertaken in order to observe whether *Rlt* variation could be explained by putative functional differences between genospecies when interacting with distinct clover genotypes. In this study, the relative proportions of *Rlt* genospecies were found to differ between clover genotypes (Figure 3.5). Relative frequencies of gsB and gsC, based on *recA* and *rpoB* allele frequencies, were found to significantly differ between pure check variety Klondike nodules and all other F2 variety Crosses, and also between Cross 1 and Cross 2 (Figure 3.5). However, there were no differences in the relative proportions of *Rlt* genospecies between comparisons of most F2 crosses, and this was mostly due to the large amount of variation in genospecies abundances within clover genotypes. The proportion of 'unassigned genospecies' allele sequences may also be able to partially account for the lack of observed genospecies differences. White clover are outbreeders and therefore a large amount of genetic heterogeneity within and between varieties is somewhat expected based on previous findings. For example, within variety heterogeneity has previously been shown through varied nodulation success of inoculated, heterogeneous plant variety populations (Russell and Jones, 1975; Jones and Hardarson, 1979). Nevertheless, the clovers sampled in this study were taken from industrial breeding programme trial plots, and it is assumed that the heterogeneity within these crosses is reduced as much as possible for commercial purposes. The percentage genetic similarity between the white clover F2 variety crosses are unknown. It is also unknown how much the clover genotypes differed in traits associated with host-filtering mechanisms, such as production of various root exudates and microbe recognition systems (Jones *et al.*, 2019). However, it was hypothesised that a clearer distinction of host *Rhizobium* nodule diversity would be observed between pure varieties compared to F2 crosses. Despite this assumption, the pure variety used in this study (Klondike) displayed just as much *Rlt* allelic variation, if not more, as the F2 crosses. Future similar experiments could be undertaken with more genetically distinct and defined clover varieties to confirm whether a greater distinction of rhizobia nodule communities is evident compared to F2 crosses.

3.5.2. Geographical distance contributes to nodule *Rlt* diversity

The genetic dissimilarity in *rpoB* and *recA* housekeeping genes, was significantly associated with geographic distance between sampled plants (Figure 3.6a-b). Conversely, diversity of symbiosis genes, *nodA* and *nodD*, were not significantly associated with geographic distance or whether samples were isolated from the same clover genotype (Appendix Figure B.7). When using samples from a single clover genotype (Klondike), dissimilarity of allelic diversity was found to increase with increased geographic distance between sampling points (Figure 3.6d-f). gsB *rpoB* and *recA* alleles were found to dominate nodule samples from Block 1 plots (Figure 3.5). However, Klondike plot c from Block 2 (Figure 3.6c-d) displays different allelic composition because it is actually dominated by gsC alleles. Additionally, other nodule samples collected from Block 2 plots at the same time but not included in this study also showed gsC alleles dominating nodule populations. This is interesting, as previously gsC has been found to be the most prevalent genospecies in Danish soils (Cavassim *et al.*, 2020). Furthermore, nodule samples used for MAUI-seq method validation were collected from the same site a year earlier (although from different parts of the same site) and these samples were also dominated by gsC alleles. This suggests that intraspecies composition of rhizobia clover symbionts vary considerably within fields between local microenvironments of individual plants. Previous studies have also shown chromosomal genotypes to be strongly associated with geographic origin (Fagerli and Svenning, 2005). Together these results suggest that also other variation associated with local growth conditions likely affected the symbiont diversity in addition to plant genotype identity. For example, differences in soil conditions, such as changes in pH or chemical composition, have been associated with diversity and composition of rhizosphere microbial communities (Wang *et al.*, 2018). This could, in part, determine the initial rhizobial population 'pool' of available genotypes for the plants (Paffetti *et al.*, 1996; Philippot *et al.*, 2013). The influence of environment on population diversity is important, as small differences in *Rhizobium* genotype frequencies between soils have been strongly associated with the distribution of bacterial genotypes in nodules (ZéZé, Mutch and Young, 2001; Laguerre *et al.*, 2003). Unfortunately, the initial rhizosphere soil population for each sample was not determined in this study and could have provided further insight into additional biogeographic patterns between samples.

Moreover, *rpoB* was influenced by both host genotype and geographic distance, although there was no significant interaction between genotype and geographic distance. However, it was observed that even Klondike samples, which showed significant allelic diversity from other clover genotypes, displayed increased genetic dissimilarity with increased geographic distance across the field (Figure 3.6). Therefore, it is possible that interactive effects between the field geography and host genotype could have cooperatively manipulated rhizobia diversity as it has been shown in both soybean and common bean legumes, with soil type predominantly influencing microbiome community and host genotype modifying the selectivity (Aouani *et al.*, 1997; Nleya, Walley and Vandenberg, 2001; Argaw and Muleta, 2017; Liu *et al.*, 2019). This highlights the importance of considering both soil microenvironment and plant genetic variability when applying rhizobial inoculants in the field, as field geography could influence the initial rhizosphere microbial population 'pool' in addition to further symbiotic selectivity imposed by the clover genotype (Liu *et al.*, 2019).

3.5.3. Study limitations and future research

One considerable limitation of this study was that the level of observed diversity within and between host genotypes was shown to be dependent on the gene markers of interest (Figure 3.4). As a result, potential differences in *Rlt* nodule diversity between plant genotypes could have been missed because of marker gene choice. This could have been potentially achieved by analysing the genetic diversity of additional accompanying symbiosis genes which may also influence inter- and intra-species symbiotic specificity. For example, interactions between *nodA* and other nod genes (*nodBC* and *nodEF*) have been show to affect Nod factor production and functionality, and by association, efficiency for symbiotic establishment (Debellé *et al.*, 1996; Ritsema *et al.*, 1996; Duodu *et al.*, 2006; Maj *et al.*, 2010). Additionally, intraspecies host specificity between white clover and *Rlt* may also be regulated by other molecular interactions at later stages of symbiotic establishment, such as from extracellular polysaccharide production, identification of secretion systems and detection of microbe-associated molecular patterns, which could be investigated further (Perret, Staehelin and Broughton, 2000; Simms and Taylor, 2002; Wang, Liu and Zhu, 2018). In the future it would be important to take these also into account by perhaps sequencing for other nodulation genes, or even investigate how differences in legume cultivar gene sequences such as those associated with pathogen recognition systems are associated with observed rhizobium nodule diversity.

Unfortunately, the abiotic soil factors and initial rhizosphere soil community composition for each sample was not evaluated. It would otherwise have been insightful to determine if the initial rhizosphere *Rlt* population pool differed between clover genotype samples. Legume root microbiomes have been found to significantly differ to those of non-legume plants, due to their predominant symbiotic interaction with rhizobia (Turner *et al.*, 2013; Hartman *et al.*, 2017). However, individual genotypes have been reported to only weakly influence rhizosphere microbiome composition depending on the soil properties and plant species (Liu et al., 2019). Additionally, abiotic factors such as nitrogen content, oxygen content, moisture, pH and salinity were also not evaluated which could critically affect rhizobial soil population structure at different sampling points (Harrison, Jones and Young, 1989; Paffetti *et al.*, 1996; Wang *et al.*, 2018). In the future, it would be interesting to replicate the experimental design in more controlled conditions in greenhouse experiments with known initial rhizosphere populations. The link between selection of different rhizobium genotypes and clover yield was also not investigated. In order to achieve this, more controlled greenhouse experiments would be required. This type of experiment would be needed to systematically disentangle interactions between the many different genetic and environmental factors influencing symbiosis.

Finally, in this study clovers were sampled in October which may have resulted in an altered soil microbial community composition compared to if sampling had taken place in the summer months, after nitrogen fertilisation, or at an earlier plant growth stage before the first cut was harvested in the field (Inceoğlu *et al.*, 2010). However, it was previously observed that the same pool of rhizobia genotypes dominated sample populations regardless of the time of year sampling (Duodu *et al.*, 2006), which suggest that the sampling time might have been a lesser problem.

3.5.4. Conclusions

Investigating the extent to which partner-choice is advantageous in mutualistic symbioses is critical to aiding our understanding of the evolutionary dynamics and maintenance of symbiosis and intraspecies diversity (Simms and Taylor, 2002). This study evaluated whether the selection for specific rhizobial genotypes by white clover extended beyond interspecies specificity, and selection was observable at the intraspecies level under genuine field conditions. Overall, some clover genotypes were found to select for different rhizobial genotypes, but a large amount of variation was observed within clover genotypes. Nodule diversity was also largely associated with geographic distance between samples, which was perhaps enhanced by a heterogeneous initial soil rhizobium population pool and other abiotic local growth conditions. The fact that global F_{ST} alternatively identified *nodA* as the major determinant of population structure out of the four *Rlt* genes suggest that additional geographical and environmental factors could determine *Rlt* nodule diversity between samples in addition to clover genotypes. Using all four genes in combination to detect clover genotype differences can generate a greater holistic perspective of *Rlt* genotype differences between clover genotypes. However, observed diversity was found to be largely depended on the gene of interest (*rpoB*, *recA*, *nodA* or *nodD*). *rpoB* and *recA* alleles showed the greatest allelic distinction between clover genotypes, specifically Cross 1 and Cross 2, which was not observed with symbiosis genes. Future work could investigate these other cellular mechanisms of symbiotic selectivity, such as the importance of extracellular polysaccharides and specificity of secretion systems detection by the legume host, wherein intraspecies specificity may also be evident. As other molecular selection processes are involved in *Rlt* genotype x clover genotype compatibility and partner choice, more genes should be included in future diversity analyses. Overall, both soil environment and host genotype are important considerations when choosing compatible inoculants for white clover (Lupwayi, Clayton and Rice, 2006).

Chapter 4. *Rhizobium leguminosarum* symbiovar *trifolii* sub-species display distinct intraspecies transcriptomic variation

4.1.Abstract

Background: Transcriptomic cross-species analyses have identified regions of conserved and divergent gene expression between bacterial species by relying on only a few representative strains. However, within-species transcriptomic comparisons are scarce. The *Rhizobium leguminosarum* species complex contains five genetically distinct genospecies based on an average nucleotide identity < 95%. Here the potential transcriptional differences in the core and accessory genome of different genospecies were compared with their phenotypic differences.

Results: To study how bacterial genetic distance influences gene expression, multiple *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) strains were grown under the same conditions and assessed for core gene expression differences between (3-7 strains) and within (59 strains) genospecies. Genospecies displayed differences in core genome expression profiles and significant differential expression of individual core genes. Core genome expression profiles were less distinct within genospecies, and overall, more genetically diverged strains displayed a higher proportion of differentially expressed core genes. Significant correlations between groups of co-expressed core gene modules and phenotypic growth differences between genospecies were identified. *Rlt* core gene modules enriched with fundamental bacterial metabolism genes were also significantly differentially expressed between genospecies. Additionally, the *Rlt* accessory genome had significantly lower expression compared to the core genome across all genospecies.

Conclusions: Together these results suggest that genospecies can display differences in core genome expression when grown under the same conditions, and this variation can be further associated with growth differences. Furthermore, within genospecies substantial transcriptional variation is also evident which overall demonstrates that within species similar genotypes can show differences in gene expression.

4.2. Introduction

Transcriptome profiling has become an insightful tool for investigating phenotypic differences among organisms (Wang, Gerstein and Snyder, 2009) providing gene expression information for a specific subset of genes up to whole-genome expression. Such studies comparing inter-species expression profiles are commonly called cross-species analyses, and can be used to identify conserved and differential gene regulation between species based on the expression of orthologous genes (Stuart *et al.*, 2003). However, from a prokaryotic perspective it is known that there can be large genetic variation within a species alone, and studies have already shown that significant natural transcriptomic variation is observable between individuals of the same species (Oleksiak, Churchill and Crawford, 2002; Townsend, Cavalieri and Hartl, 2003; Pavey *et al.*, 2010; Madritsch *et al.*, 2019). To identify transcriptomic differences between bacterial species, previous analyses have used between one and four strains as species representatives for direct differential gene expression comparisons (Vital *et al.*, 2015), and up to 51 strains for assessing variation within species regulatory networks (Galardini *et al.*, 2015). In addition, a maximum of three bacterial species have been transcriptionally cross-compared within one study (Hosseinkhan *et al.*, 2015). Therefore, although intraspecies bacterial diversity is known to vary significantly at the genome level, the extent to which genetic differences translate to expression level variation within a species is less well understood.

It has been suggested that evolution of new species could be shaped by gene expression differences within a species promoting adaptive divergence, as well as from genomic changes (Feder and Mitchell-Olds, 2003; Ranz and Machado, 2006; Ng *et al.*, 2019). Ecological speciation arises when gene flow is restricted between populations due to adaptive divergence by, as examples, geographic isolation (allopatric speciation) or niche-specific adaptation (sympatric speciation) (Pavey *et al.*, 2010; Vos, 2011; Friedman, Alm and Shapiro, 2013; Shapiro and Polz, 2015). Gene expression could promote speciation by enabling population persistence through expression variance, and also by directly influencing traits related to reproductive isolation (Pavey *et al.*, 2010). For example, expression differences may lead to colonisation of new environments where the regulation of expression could further become vital for population survival influencing species diversification (Pavey *et al.*, 2010; Ng *et al.*, 2019). Divergence of gene expression has therefore been used as a molecular phenotypic indicator to support the idea species divergence (Pavey *et al.*, 2010; Wolf *et al.*, 2010; Dunning *et al.*, 2016). This is because expression divergence is expected to develop and evolve faster than nucleotide divergence, although it still remains unclear how much this variation influences species divergence (Oleksiak, Churchill and Crawford, 2002; Vicente and Mingorance, 2008; Wolf *et al.*, 2010; González-Torres *et al.*, 2015). Other unanswered questions include to what extent transcriptomic variation can define species differences and whether transcriptome profiles overlap during bacterial speciation (Vital *et al.*, 2015).

While transcriptomic analysis pipelines have been specifically developed for cross-species analyses, they are currently mainly optimised for eukaryotic comparisons (Kuhn, Luthi-Carter and Delorenzi, 2008; Zarrineh *et al.*, 2014; Zhu *et al.*, 2014; LoVerso and Cui, 2015). Typically in cross-species transcriptome comparisons, organisms are exposed to the same environment in 'common garden' experiments, and the transcription levels across orthologous gene regions are compared (Oleksiak, Churchill and Crawford, 2002; Townsend, Cavalieri and Hartl, 2003; Madritsch *et al.*, 2019). There are several ways to identify and compare orthologous genomic regions in cross-species analysis. For example, LoVerso and Cui (2015) use a single reference genome from which to identify orthologous regions in all other genomes, which is more effective the more genetically similar the compared species are. Similarly, 'master' reference transcriptomes have been assembled *de novo* from two species to allow for cross-taxa comparisons (Wolf *et al.*, 2010; Ng *et al.*, 2019). Other pipelines have produced platforms to compare expression levels at the level of single genes, gene sets and gene networks (Langfelder and Horvath, 2008; Chaudhuri *et al.*, 2015). Often, cross-species analysis data are curated from different independent studies, and hence, the environmental conditions used in the studies might confound transcription comparisons as the initial data collected was not intended for cross-species analyses (Stuart *et al.*, 2003; Carrasco, Tan and Duman, 2011; Kristiansson *et al.*, 2013; Hosseinkhan, Mousavian and Masoudi-Nejad, 2018). Therefore, cross-species analyses have been criticised for inconsistencies in data collection and inadequacies in experimental design (Kristiansson *et al.*, 2013).

Bacterial cross-species transcriptomic comparisons have predominantly focused on the most genetically well-characterised species, such as *Escherichia coli*, *Staphylococcus aureus*, *Salmonella enterica* and *Pseudomonas aeruginosa* (Carrasco, Tan and Duman, 2011; Zarrineh *et al.*, 2014; Hosseinkhan *et al.*, 2015; Vital *et al.*, 2015; Hosseinkhan, Mousavian and Masoudi-Nejad, 2018). Cross-species analyses focus mainly on the expression differences in core genes, and disregard accessory genome components (Wolf *et al.*, 2010; Galardini *et al.*, 2015; Young, 2016; McInerney, McNally and O'Connell, 2017; Jiao *et al.*, 2018; Madritsch *et al.*, 2019). Core genes are defined as genes shared amongst all individuals in a distinguished population or species and essential for organism functioning, whereas accessory genes are not present in all individuals and associated with more dispensable, but advantageous functions linked to survival in different environments and are less likely to participate in a species regulatory network (Young *et al.*, 2006; Young, 2016). The pangenome (combination of core and accessory genes within a species) can be very large as horizontal gene transfer and other forms of introgression can contribute substantially to a species genetic variation (Tettelin *et al.*, 2005; McInerney, McNally and O'Connell, 2017). This large genetic diversity within bacterial species can make it hard to identify definitive genetic and phenotypic species differences. For example, while polyphasic taxonomy classifies species based on genetic similarity and by distinctive phenotypic traits (Vandamme *et al.*, 1996; Young, 2016), phenotypic distinction of bacteria can be challenging when a species has a large accessory genome that conveys extensive intraspecies genomic and phenotypic diversity (Vos, 2011; Young, 2016). Because not all strains of a bacterial species share the same accessory genes, fewer studies have focused on accessory gene expression differences relative to core gene expression (Scaria *et al.*, 2013; Vital *et al.*, 2015; Jiao *et al.*, 2018).

Cross-species analyses have not only been utilised to identify differences between bacterial species but also to identify conserved transcriptional regions. Co-expression identified amongst more than one strain or species has been used to provide support that a gene is involved in the same biological process or has a similar function (Hosseinkhan *et al.*, 2015; Hosseinkhan, Mousavian and Masoudi-Nejad, 2018). As a result, cross-species gene expression analyses have enabled identification of larger core transcriptional networks conserved across species (Zarrineh *et al.*, 2014; Hosseinkhan, Mousavian and Masoudi-Nejad, 2018). Similarity of gene expression between strains has previously been found to be more strongly associated with similarity of environmental origins and experimental conditions than phylogenetic relatedness (Vital *et al.*, 2015; Jiao *et al.*, 2018). Previous studies have compared the transcriptomes of strains isolated from different environments in order to account for both the potential genomic diversity and phenotypic diversity associated with living in different niches (Scaria *et al.*, 2013; Vital *et al.*, 2015). In comparison, genetically different strains isolated from the same environment have been compared under various growth conditions to explore how genetic differences contribute to the capacity to express genes in novel environments (Kimes *et al.*, 2014). However, the level to which environment influences conservation of co-expressed genes and their regulatory networks across different phylogenetic distances remains unclear (Zarrineh *et al.*, 2014).

Rhizobia have extensively been used to study variation in bacterial gene expression. Rhizobia are soil bacteria that can live in two physiologies: a motile, free-living soil form or a non-motile bacteroid form within legume root nodules; a physiological transformation associated with significantly altered gene expression (Yoder-Himes *et al.*, 2009; Vercruysse *et al.*, 2011; Lopez-Leal *et al.*, 2014). Rhizobia are a good bacterial model to evaluate gene expression for several reasons. Firstly, sequencing repositories have a large number of strains with fully sequenced *Rhizobium* genomes, and their multipartite genome (Young *et al.*, 2006; diCenzo and Finan, 2017) allows for interesting investigation of regulatory interactions. Secondly, several groups of genes have been heavily studied and their regulatory interactions are known such as in the case of formation of symbiosis. Previous investigations have mainly observed rhizobial transcriptomic responses to induction of known stress conditions (Vercruysse *et al.*, 2011; Liu *et al.*, 2014; Lopez-Leal *et al.*, 2014), altered environmental conditions and symbiosis development (Karunakaran *et al.*, 2009; Ramachandran *et al.*, 2011; Krysciak *et al.*, 2014; Peng *et al.*, 2014; Roux *et al.*, 2014; Perez-Montano *et al.*, 2016), and bacteroid versus free-living physiologies (Yoder-Himes *et al.*, 2009; Vercruysse *et al.*, 2011; Lopez-Leal *et al.*, 2014). However, only a few studies have directly evaluated differences in transcriptome profiles between rhizobia strains with differing organisations of core and accessory genes (Galardini *et al.*, 2015; Rachwal, Matczynska and Janczarek, 2015; Jiao *et al.*, 2018; Green *et al.*, 2019).

This study aimed to investigate how genetic distance influences gene expression by assessing the transcriptome profiles of strains from recently diverged sub-species of *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*), named genospecies (gs) A to E. These genospecies are classed as cryptic individual species based on their genetic distinctiveness (<95% average nucleotide identity) and reduced gene flow (Ravin, 1963; Kumar *et al.*, 2015; Cavassim *et al.*, 2019). However, despite the significant genomic differences between these sibling species, currently no phenotypic traits have been found to be exclusive to a single genospecies (Ravin, 1963; Kumar *et al.*, 2015). In particular, the extent to which transcriptome profiles differed between and within *Rlt* genospecies was assessed. To this end, whole core genome transcriptome expression variation was compared with RNA-seq across 26 *Rlt* strains from 5 genospecies (A-E), and additionally 59 strains from *Rlt* genospecies C, that were isolated from conventional trial managements and organic farm managements across the UK, France and Denmark (Cavassim *et al.*, 2019). This study aimed to answer four key questions regarding whether genetic boundaries can predict gene expression profiles:

i) First, within a simplistic environment, do *Rhizobium leguminosarum* genospecies display basal transcriptional differences of shared genes? And if so, what proportion of shared genes are differentially expressed?
ii) Secondly, does differential gene expression correlate with genetic divergence along with genospecies boundaries?
iii) Thirdly, can transcriptional differences be attributed to distinct phenotypic or metabolic traits that might aid understanding of what drove speciation of these genospecies?
iv) Finally, to what extent does expression of the accessory genome differ within species?

4.3.Methods

4.3.1. Strain metadata, ANI, bacterial growth and RNA sample preparation

Twenty-six *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) strains were selected from the 196 *Rlt* strain NCHAIN collection (Cavassim *et al.*, 2019) and categorised into five genetically distinct *Rlt* sub-species with <95% average nucleotide identity, called genospecies (gs) A-E (gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3) (Kumar *et al.*, 2015). Additionally, a further 59 strains from gsC were also selected from the collection (referred to here as gsC* dataset) and can be categorised into several phylogenetic gsC subbranches (C1 = 17, C3 = 5, C4 = 2, C5 =1, C6 =8, C7 =14, C8 = 4, C9 = 9, C10 = 1). Metadata for strains, including genospecies classification and geographic origin, can be found in the Supplementary Material (Appendix Table C.1).

The genetic relatedness of strains differed within each genospecies. Pairwise Average Nucleotide Identity values (ANI) were calculated based on the proportion of shared single nucleotide polymorphisms (SNPs) in genes that were present in at least 100 strains from the 196 NCHAIN strain dataset (6,529 genes, 441,287 SNPs) (Cavassim *et al.*, 2019) rather than using only genes present in all known *Rlt* strains (282 genes). Using a larger number of genes increased the clustering resolution and genetic distinction between strains (Appendix Figure C.5c). ANI for each genospecies based on the 26 strains ranged from: gsA = 0.9550-0.9751, gsB = 0.9848-0.9966, gsC = 0.9708-0.9999, gsC* (58 strains, without SM132) = 0.9539-0.9999, gsD = 0.9882-0.9998, gsE = 0.9770-0.9928. gsB is most genetically homogeneous genospecies and strains used in this study were isolated from the same UK site.

Strains were revived from glycerol stocks in Tryptone Yeast (TY) broth (5 g Tryptone, 2.5 g Yeast Extract, 1.47 g $CaCl_2$, per litre volume) and grown for 48 h at 28°C, 180 rpm. Optical densities were normalised to 0.1 OD_{600}, in preparation for transfer into inducing conditions. Strains were then individually cultured in TY broth with 1 μM 7,4'-dihydroxyflavone (clover flavonoid stock solubilised in DMSO) at 28°C, 180 rpm, for 48 h. Additionally, 59 strains from gsC were grown under the same conditions in a separate sampling batch (gsC* samples) for processing and sequencing. Within this additional gsC* sample batch, 2 strains (SM158 and SM170C) were also grown in duplicate as biological replicates, bringing the total number of samples to 61. Clover flavonoid, 7,4'-dihydroxyflavone, was also added to the TY broth to model the presence of legume host interaction and because it was found to be the most effective clover flavonoid to induce *Rlt* nod gene expression (Djordjevic et al., 1987). On average, all strains reached 0.3 OD_{600} after 48 h.

Total RNA was isolated using the RNeasy Protect Bacteria Mini kit following manufacturer's instructions (Qiagen). Stabilised pellets were stored at -80°C until total RNA was isolated. Total RNA yield was measured using Bioanalyzer 2100 and for batch 1 (range: 59-271 ng/μl, average yield: 134.82 ng/μl) and gsC* sample batch 2 (range: 30-122 ng/μl, average yield: 69.57 ng/μl). rRNA depletion using Ribo-Zero rRNA Removal kit (Bacteria) (Illumina), clean up with Zymo Clean and Concentrator kit, and RNA-seq paired-end library preparations were carried out by the University of York Technology Facility. Libraries for both sample batches were subject to

Illumina 2 x 150 bp paired end sequencing in independent runs using a HiSeq3000 by the University of Leeds Next Generation Sequencing Facility.

4.3.2. RNA-seq read and count processing

RNA-seq reads were quality checked with FastQC (v.0.11.5) following default parameters. Cutadapt (v.1.15) was used to trim reads of Illumina adapters with the following parameters: maximum error rate = 0.1 (10%), minimum overlap = 5 bp, minimum read length = 15 bp. Trimmed reads for each strain were then mapped to their respective Illumina sequenced, 'Jigome' assembled, whole genome assemblies (Cavassim *et al.*, 2019) using HISAT2 (v. 2.1.0) (Additional File 6: Table S1). All reads passed Samtools flagstat (v. 1.7) and rseqc bamstat (v. 2.6.4) default QC filtering (Additional File 6: Table S1). Total number of reads per sample for the first batch sequencing run of 26 gsA-E samples ranged between 6,793,334 - 15,699,075 reads with an average of 9,183,811 reads. Total number of reads per sample for the second batch sequencing run of 61 gsC* samples ranged between 990,334 – 12,135,188 reads with an average of 4,759,219 reads. All sequences mapped to their individual genomes with an average overall alignment rate of 98.30% (Additional File 6: Table S1). HTSeqCount (v.0.9.1) was used to count reads mapping to each gene feature, with union parameters selected (Additional File 6: Table S1). Orthologous gene groups were previously identified using ProteinOrtho (v.5.16b) and were used to compare ortholog group expression across strains (Cavassim *et al.*, 2019). For functional annotation of genes, Prokka (v.1.12) was previously used to produce gene annotations with equivalent RefSeq accession numbers and protein product information (Cavassim *et al.*, 2019). Overall, only orthologous gene expression was analysed.

4.3.3. Library normalisation methods optimisation

Raw counts of 4,229 orthologous core gene groups were used to produce normalisation scaling factors for each genotype sample. Orthologous core genes were defined as being present in every strain from the 196 *Rlt* strain NCHAIN collection (Cavassim *et al.*, 2019).

Raw read count normalisation was evaluated and optimised by comparing DESeq2 (v.1.22.2), TMM (EdgeR v.3.24.3) and PoissonSeq (v.1.1.2) normalisation methods. The initial count dataset used for optimising normalisation constituted 30 strains from the first sample batch dataset containing strains of all five genospecies (1 gsB strain and 3 gsE strains were subsequently removed to generate the final 26 strain dataset, see below), and 59 strains gsC* strains (61 samples as two strains are in biological duplicates) from the second sample batch dataset.

The core gene raw counts were normalised via the three normalisation methods. In order to test the success of the normalisation methods, the normalised expression counts of three random subsets of 400 core genes (subsets 1-3) across samples were used as a representation of the dataset (approximately 10% of total core genes). Then for each of the three 400-gene subset individually, eigengene values were calculated for each strain using expression count data normalised by either DESeq2, TMM or PoissonSeq methods (Li *et al.*, 2010; Robinson, McCarthy and Smyth, 2010; Love, Huber and Anders, 2014). Eigengene values were calculated using the WGCNA (v.1.66) package in R and equate to the representative normalised gene expression (or first principal component value) of the 400 genes within a subset. These eigengene values were then correlated to samples' raw library sequencing depths and normalised library sequencing depths.

TMM normalisation was found to be the least effective and eigengene values strongly correlated with the raw (Appendix Figure C.1, panel 2) and normalised (Appendix Figure C.2, panel 2) sequencing depths of samples . DESeq2 was much more successful and normalised well to account for the difference in sequencing depth between the two sample batches (Appendix Figure C.1 and Appendix Figure C.2, panel 1). PoissonSeq was the most effective for normalising sequencing depths and no correlation was found to eigengene values (Appendix Figure C.1 and Appendix Figure C.2, panel 3). However, four strain samples were unaffected by any of the three normalisation methods tested (1 gsB strain and 3 gsE strains). The relative distribution of the samples' eigengene values (calculated with PoissonSeq normalised counts) remained consistent regardless of 61 sample gsC* dataset removal from the normalisation step (Appendix Figure C.3 and Appendix Figure C.4, panel 1) and four outliers removal (Appendix Figure C.3 and Appendix Figure C.4, panel 2-4). Therefore, the four outlier samples were removed from the final dataset (Appendix

Figure C.3-Appendix Figure C.4, panel 4). In conclusion, the final multi-genospecies comparison dataset contained 26 samples (gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3), and including the 61 gsC* samples from the second sequencing batch. Raw gene expression count data for both datasets were normalised together using PoissonSeq (Appendix Figure C.2, panel 4).

Additionally, the normalised and log transformed counts from two gsC* strains with biological duplicates were correlated to ensure that replicate samples were comparable (Appendix Figure C.5a-b).

4.3.4. Gene expression analysis

Gene expression was compared using multiple different approaches. First, differential gene expression was compared between different genospecies using DESeq2. Second, the number of differentially expressed genes between individual strains was calculated with GFOLD. Lastly, *Rlt* core gene modules that were differentially expressed between genospecies was determined using WGCNA.

DESeq2 (v.1.22.2) was used to identify differentially expressed orthologous core gene groups (core DEGs) between genospecies using the 26 multi-genospecies sample dataset, and within genospecies using the 61 gsC* sample dataset. Cytoscape (v.3.7.2) was used to create the DEG network figures. pBLAST was used to identify the RefSeq functional annotations (100% identity, 100% query cover) for the amino acid sequences of the top two most significant (FDR corrected) DEGs ($\log_2$Fold Change > ±2).

GFOLD (v1.1.4) was used to determine the number of core DEGs between pairwise sample comparisons (Feng et al., 2012). GFOLD calculates a generalised fold change for ranking DEGs and produces equivalent log fold change values for sample comparisons when no sample replicates are available. Therefore, GFOLD enabled determination of DEGs between all strain comparisons using PoissonSeq normalised read counts. A GFOLD value of >±2 GFOLD was used as a threshold for identifying DEGs.

Co-expressed orthologous core genes were grouped into modules with WGCNA (v.1.66) R package. PoissonSeq normalised, $\log_2(n+1)$ transformed counts from the 26 multi-genospecies sample dataset were used as input for WGCNA. A signed network and signed topological overlap matrix were generated to categorise core genes into modules. A soft threshold power of 7 was used, as it was the lowest power for which the scale-free topology fit index curve flattened out upon reaching 0.90 (Appendix Figure C.6a). The minimum number of genes required to be considered as a distinct module was set to 3. Otherwise, all WGCNA default settings were used. WGCNA calculates an eigengene value for each module for each strain. Module eigengenes are calculated by restricting the gene expression matrix to only the genes within a module and calculating singular value decomposition for that module. For each strain, the first principal component value for that module was used as the eigengene value.

4.3.5. Growth phenotype analysis

For analysis of phenotypic growth differences between genospecies, strains were grown in modified Tryptone Yeast broth (TY) (5 g Tryptone, 2.5 g Yeast Extract, 1.47 g $CaCl_2$, per litre volume) conditions in 96-well plates (Smith, 2018). TY media was altered in the following ways: by modifying pH (pH 4, 5 and 6, 6.68), growth temperature (4, 10, 15, 20, 28 °C), and nutrient concentration (100, 25, 12.5, 6.25, 3.125% TY). Growth measurements (OD_{600}) at 48 h was used as a proxy of rhizobial growth for all measured phenotypic traits. Biofilm formation was also measured (OD_{600}) after 96 hours growth from all the temperature and nutrient altered treatments as follows. Briefly, 20 µl of crystal violet was added to each well and left to stand for 15 minutes. Wells were subsequently rinsed with clean water three times and left to air-dry. Wells were then filled with 225 µl of absolute ethanol and incubated at room temperature for 1 h to dissolve the crystal violet staining.

4.3.6. Gene function annotation and accessory genome analysis

To link differential gene expression with specific functional pathways, Kyoto Encyclopedia of Genes and Genomes (KEGG) BlastKoala was used to identify KEGG K identifiers for core orthologous genes (Kanehisa et al., 2016). KEGG Ontology (KO) identifiers were identified from the 'Prokaryotes, Bacteria' KEGG taxonomy group using the 'Genus, Prokaryotes' KEGG database. KEGG mapper was used to identify KEGG functional pathway classifications of orthologous core genes and genospecies accessory genes. KEGG mapper was also used to identify complete KEGG modules (i.e. containing all genes required for a specified functional unit) present within WGCNA co-expressed *Rlt* core gene modules. KEGG KO IDs were assigned to 2,380 genes out of 4,229 core genes, with 7 of those genes being assigned two KO IDs. Overrepresentation analysis of KEGG functional categories within each module was undertaken with enrich.KEGG ClusterProfiler (v.3.10.1) function in R, using all core genome KO identifiers as a background dataset. For additional pathway enrichment analysis, metacyc IDs were assigned to all possible core genes. In total, metacyc IDs were identified for 2,070 out of 4,229 genes. Metacyc IDs were analysed for significant pathway enrichment using the metacyc pathway enrichment pipeline with Benjamini-Hochberg p-value correction for genes in WGCNA modules and for core DEGs. Additionally, to identify functional associations to PCA principal components, metacyc pathway enrichment analysis and KEGG pathway analysis was undertaken using the genes contributing more than they would on average to each principal component (Appendix Table C.2; Additional File 6: Table S2).

Orthologous accessory genes were identified in each strain as genes not considered a core gene. The number of unique accessory genes was determined by totalling the number of unique orthologous gene groups across all strains in a genospecies that is not a core orthologous gene group. Additionally, the percentage of genes present but not expressed were determined by calculating the expression level for each orthologous gene group in each strain, and orthologous gene groups with expression values of 0 were classed as genes that were present but not expressed. The percentage of core and accessory genes with 0 counts in each strain were used to generate a mean value for each genospecies.

4.3.7. Statistical analyses

For analysis of core genome expression profile differences between strains, Principal Components Analysis (PCA) using singular value decomposition was carried out using R's prcomp function on scaled and centred normalised, $\log_2(n+1)$ transformed core genome count data. Principal components contributing an individual variance of more than 5% and preceding the scree plot inflection point were used for further analysis. Genes contributing more than they would on average if all genes had an equal contribution to a principal component (0.236%; Appendix Table C.2) were selected for pathway enrichment analysis of principal components. Average linkage hierarchical clustering was calculated using R's hclust (method = average) function, using a Euclidean distance matrix of normalised, $\log_2(n+1)$ transformed core gene expression values for each sample.

To measure phenotypic growth differences between strains, 48 h OD_{600} measurements of strains grown under the different TY growth conditions, and OD_{600} biofilm formation measurements were scaled and centred for PCA using singular value decomposition carried out using R's prcomp function. Phenotypic growth traits which contributed to a principal component more than they would on average if all traits had an equal contribution (4.35%) were identified to infer which PCA variables explained principal components 1 and 2.

Eigengene values for each module were correlated to TY growth phenotypes (OD_{600}) with WGCNA using Pearson's correlation coefficient and p-values were Benjamini-Hochberg False Discovery Rate corrected. Strains for transcriptomic analysis were grown under slightly different conditions to the TY growth phenotype conditions that expression levels were correlated against (e.g. 5 ml instead of 200 µl, 180 rpm instead of non-shaking, and 1 µM 7,4'-dihydroxyflavone instead of no clover flavonoid). The correlations between the phenotypic growth data and the expression data were confirmed to be suitable by identifying a strong correlation between the growth of strains grown in 100% TY 28°C conditions and the growth when strains were RNA stabilised for transcriptome analysis (Pearson's correlation: $R = 0.55$, $p < 0.01$; Appendix Figure C.7). To calculate a strain's mean module expression for Modules 16 and Module 9, PoissonSeq normalised, $\mathrm{Log}_2(n+1)$ transformed expression values for genes in the modules were averaged. Significant differences between genospecies in Modules 16 and 9 mean module expression were confirmed with Kruskal-Wallis test and Dunn's post hoc with Benjamini-Hochberg p-value correction.

To determine whether WGCNA modules eigengene expression values significantly differed between genospecies, two-way ANOVA with TukeyHSD post-hoc testing was used. The normalised, log transformed counts for genes within specified modules were displayed in boxplots and heatmaps generated using R's ggplot2 and complex.heatmap.

For analysis of accessory genome sizes between genospecies, Kruskal-Wallis with Dunn's post hoc testing and Benjamini-Hochberg p-value correction was undertaken on the number of genes in accessory genomes between strains grouped by genospecies. To determine the significance of expression level differences between *Rlt* core and accessory genes, a linear mixed effects model was performed using Maximum Likelihood (ML) with lme4 R package (v.1.1-21). The dependent variable was normalised $\log_2(n+1)$ transformed count data, and the gene type (core or accessory) was used as the fixed effect independent variable. Therefore, as input into the model, the number of gene expression values in total for core genes = 367,923, and for accessory genes = 271,061. Individual strain IDs and ortholog gene group IDs were categorised as random effects. A variance of more than 0 supported the incorporation of random effects in the full model. Fixed effect parameter t-values, degrees of freedom and p-values were generated with LmerTest. The significance of the fixed effect was tested with the likelihood ratio (LR) test using anova(), whereby the full model was compared to a reduced model with no fixed effect and a Chi-squared p-values < 0.05 indicated a significant difference in model fit. Additionally, the reliability of the fixed effect was determined by parametric bootstrapping of fixed effect parameter 95% confidence intervals using bootMer and boot.ci with 1000 bootstraps replicates. 95% confidence intervals not including 0 were considered reliable effects. Parametric bootstrapping displayed warnings of failed model convergence for 58 out of 1000 permutations. Due to the original model converging with no warnings, and the arbitrary threshold determination for model convergence warnings, these were classified as false positives (Bolker, 2020).

4.4.Results

4.4.1. Core genome expression differs between *Rlt* genospecies but not within genospecies C

To determine whether there were basal transcriptional differences between *Rlt* genospecies, Principal Components Analysis (PCA) was performed on the expression of 4,229 orthologous core genes present in all 26 strains. These 4,229 core genes constituted 25.84% of the 26 strain ortholog pangenome (16,365 genes) and included 25 symbiosis genes essential for symbiotic establishment with white clover.

PCA of the normalised, $\log_2$ transformed gene expression counts showed that strains' core genome expression profiles clustered by genospecies classification (Figure 4.1a). Out of 26 principal components (PCs), the first five PCs all individually explained greater than 5% of the percentage variance (19.6, 12.7, 11.8, 10.0, 7.5), and cumulatively 66% of the total variance. The first two principal components alone accounted for 32% of the total variation. The first five PCs came before inflection of the scree plot and were thus chosen for further analysis. PC1 and PC2 accounted for 20% and 13% of the total variation, respectively. gsA and gsB were the only genospecies that did not show separation along PC1 and PC2, whereas gsC, gsD and gsE clearly distinguished along these principal components. Additional principal components (PCs 3-5) which accounted for a smaller proportion of the variance distinguished gsA and gsB strains as separate clusters (Appendix Figure C.8a-b). Therefore, these results suggest that core gene expression varies between different *Rlt* sub-species.

The genes contributing most to PC1 were significantly enriched for several functions including aminoacyl-tRNA charging, trehalose biosynthesis, proteinogenic amino acid degradation, amino acid degradation, ATP biosynthesis and transport (Additional File 6: Table S2). KEGG similarly identified PC1 to be overrepresented for ribosome related pathways and additionally quorum sensing pathways (Additional File 6: Table S2). PC2 variation was largely attributed to genes enriched for amine and polyamine degradation and 4-aminobutanoate degradation (Additional File 6: Table S2). Complete KEGG gene sets (KEGG modules) were also searched for these two principal components. KEGG identified seven modules for PC1 including: PRPP biosynthesis involved in central carbon metabolism; production of cytochrome c oxidase and F-type ATPase for ATP synthesis; acyl-CoA and phosphatidylcholine (PC) biosynthesis involved in lipid metabolism; amino acids metabolism such as arginine to putrescine (polyamine biosynthesis); and glutathione biosynthesis. KEGG conversely showed PC2 was overrepresented for ABC transporters, nitrogen metabolism, and ascorbate and aldarate metabolism (Additional File 6: Table S2). Therefore, genospecies were found to vary most notably in expression of core genes associated with basic bacterial metabolism.

To determine how many genes were significantly differentially expressed (DEGs) between genospecies, Differential Gene Expression analysis was carried out using DESeq2. DEGs were identified for all pairwise genospecies comparisons (Figure 4.1b; Table 4.1; Log_2 Fold Change > ±2, FDR adjusted $p < 0.001$). Comparison of gsB and gsD showed the largest number of differentially expressed core genes (1.4%). On the other hand, gsD-gsE and gsA-gsB were found to have the smallest number of DEGs (0.28% and 0.31%, respectively, Figure 4.1b). DEGs identified were largely unique to each genospecies comparison with a maximum of seven of the same DEGs being shared across pairwise genospecies comparisons at an adjusted $p < 0.05$ (Table 4.2). While some DEGs (Log_2 Fold Change > ±2, FDR adjusted $p < 0.05$) were found to have an associated regulatory function, the majority of DEGs had unidentified functions (Table 4.1). Several DEGs with regulatory functions were the top two most significantly DEGs, such as LuxR and XRE transcriptional regulators and a histidine kinase (Table 4.1). As a result, no significant metabolic pathway enrichment could be found for the DEGs between genospecies using metacyc. This is likely because resolution of functional relevance can be low when only a few genes are considered together. For example, gsB-gsD comparison had the largest number of DEGs and was the only exception where pathway enrichment was observed. gsB-gsD DEGs were overrepresented with alanine, aspartate and glutamate metabolism, butanoate metabolism, and starch and sucrose metabolism. Therefore, by analysing only DEGs the context of the expression profile can be lost by omitting genes which do not exactly fit within DEG parameter thresholds.

In addition, average linkage hierarchical clustering was performed on Euclidean distances of core gene expression values (Figure 4.1c). While gsC and gsD were distinct in their branching, one strain each from gsA, gsB and gsE were found within the gsC branch. When 61 additional gsC samples were included in the analysis, gsC strains formed their own branch, gsA and gsB clustered, and gsD and gsE clustered (Appendix Figure C.8c).

The core genome expression of 59 gsC strains (two of which were biologically duplicated; Appendix Figure C.5a-b) were similarly assessed by PCA. Out of 61 PCs, the first four PCs individually explained more than 5% of the percentage variance (22.5, 11.8, 8.3, 7.2), and cumulatively 49.8% of the total variation. The first four PCs preceded the scree plot inflection point, supporting their maintenance in the analysis. PC1 and PC2 accounted for 34% of the total variation. gsC strains were categorised by their phylogenetic subbranch based on the maximum-likelihood phylogeny of 196 *Rlt* strains (Cavassim *et al.*, 2019). Phylogenetic subbranches of gsC were not clearly distinct in their core genome expression, in comparison to genospecies differences (Figure 4.1d). However, PC1 and PC2 explained a comparable amount of the total variation as the genospecies expression PCAs. PC1 and PC2 were identified to be enriched with ABC transporters, nitrogen metabolism, quorum sensing and valine, leucine and isoleucine degradation (Additional File 6: Table S2). Metacyc additionally identified pathways involved in the respiratory electron transport chain (e.g. substrates to cytochrome bo oxidase electron transfer) and amine and sugar derivative degradation for PC2 (Additional File 6: Table S2). While the percentage variance for PC5 and PC6 were less than 5% (4.79, 4.59), these principal components separated gsC strains into distinct phylogenetic subbranch groups, which were enriched for genes involved in secondary metabolite, sugar, amine, glycine betaine derivative degradation (Appendix Figure C.8e). This could suggest that within gsC, gene expression variation is predominantly explained by factors other than genetic similarity. However, genetic divergence clearly does account for a small proportion of gene expression variation within genospecies.

Differential Gene Expression analysis showed that up to 4.07% of core genes displayed significantly different expression between gsC subbranches (subbranches with >= 3 strains) (Figure 4.1e). The relatively higher proportion of core DEGs observed in within-genospecies comparison is likely due to the larger within-genospecies group sample sizes and reduced genetic variation. This likely allowed for a greater resolution between groups and reduced expression noise, as groups were more closely genetically related.

Therefore, DEG comparisons between individual strains were also calculated to observe expression differences beyond genetically defined strain groupings. The number of DEGs and number of shared ortholog gene groups were compared to average nucleotide identity (ANI) for all possible strain comparisons (Figure 4.1f). A strong linear correlation was observed between the ANI of strains and the number of shared ortholog gene groups (Figure 4.1f). Furthermore, the more genetically distinct strains (lower ANI) displayed a larger number of differentially expressed genes, on average, and as the number of shared orthologous genes increased between strains, the number of differentially expressed genes decreased (Figure 4.1f; Appendix Figure C.5d). However, a stronger correlation was observed between ANI and the number of shared orthologous genes, than ANI and DEG expression distance (Appendix Figure C.5e).

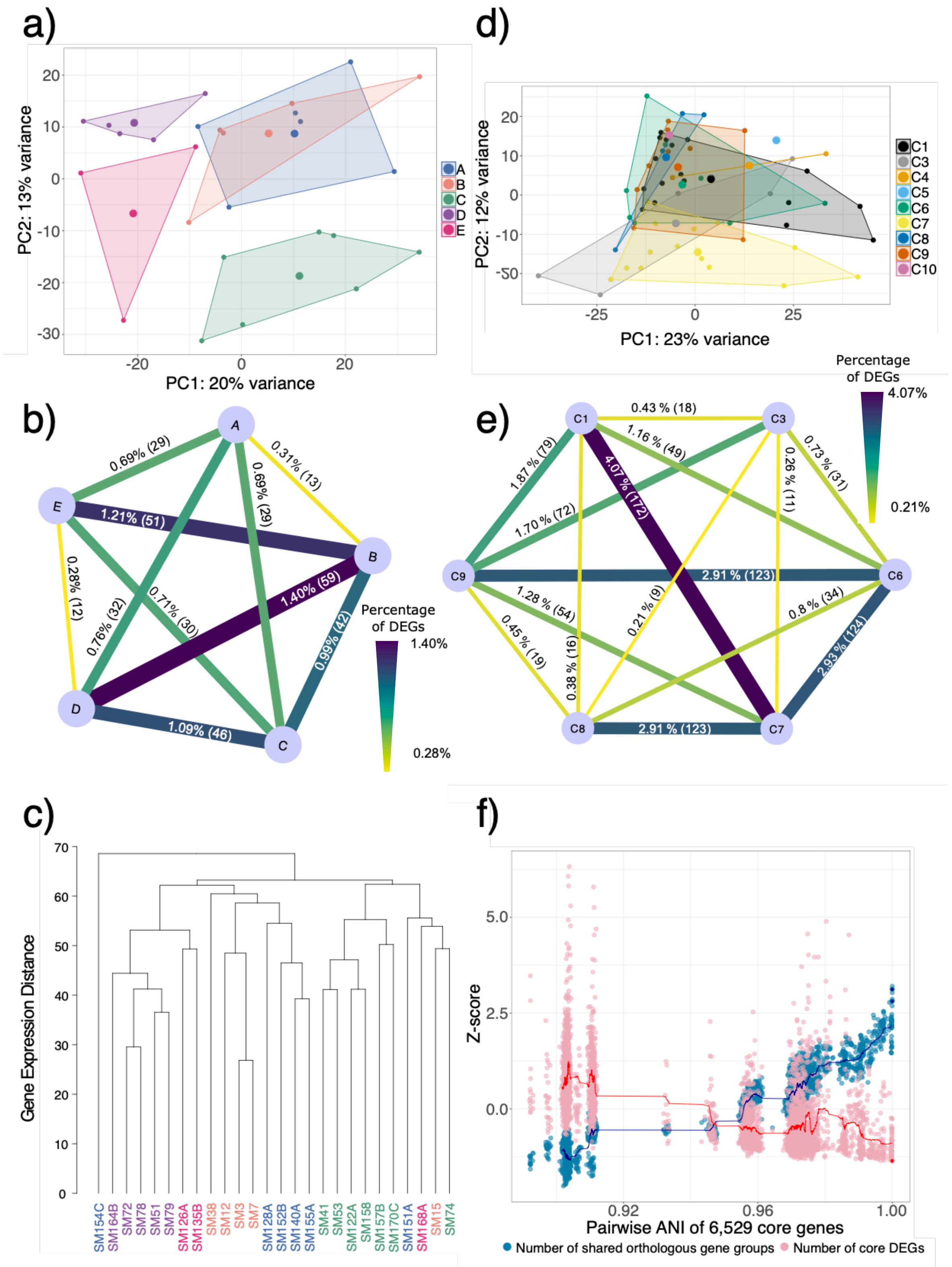


Figure 4.1 Differential gene expression between *Rlt* genospecies and phylogenetic subbranches within genospecies C. **a)** PCA of 4,229 core genes expression for 26 *Rlt* strains coloured and grouped by their genospecies (A-E). **b)** Number (%) of core differentially expressed genes (DEGs) (edges) from all pairwise *Rlt* genospecies (nodes) comparisons (Log$_2$ Fold Change > ±2, FDR adjusted p < 0.001). **c)** Strains cluster by genospecies based on average linkage hierarchical clustering of Euclidean core gene expression distances. **d)** PCA of 4,229 core genes expression for 59 strains of gsC (and additionally 2 strains in duplicate), coloured and grouped by their phylogenetic subbranches (C1-10). **e)** Number (%) of core DEGs (edges) from pairwise *Rlt* gsC phylogenetic subbranch (nodes) comparisons (Log$_2$ Fold Change > ±2, FDR adjusted p < 0.001). **f)** Individual pairwise strain comparisons show that as the average nucleotide identity (ANI) of strains increases, the number of shared orthologous gene groups also increases, but the number of core DEGs decreases. Red and Blue lines display the rolling average (n=100) for DEGs number and shared orthologous gene groups, respectively. Red and dark blue dots highlight strain comparisons which are biological replicates for number of core DEGs and number of shared orthologous gene groups, respectively.

Table 4.1 Number of core differentially expressed genes (DEGs) between pairwise *Rlt* genospecies comparisons. Genes were classed as differentially expressed if they had a Log$_2$ Fold Change > ±2, FDR < 0.05, adjusted p < 0.05. Metabolic pathway enrichment analysis of DEG groups meeting the threshold values of Log$_2$ Fold Change > ±2, FDR adjusted p < 0.05 were evaluated for pathway enrichment. However, not all DEGs DNA sequences could be blasted to a metacyc ID or KEGG K identifier. The amino acid sequence of the top two most significant DEGs were pBLAST to identify the RefSeq functional annotation.

Genospecies comparison	**Number of DEGs, adj. p < 0.05 (% of total core genes)**	**Number of DEGs, adj. p < 0.001 (% of total core genes)**	**DEGs with KEGG K identifier**	**DEGs with metacyc ID**	**DEGS (adj. p < 0.05) with associated regulatory function (% of DEGs)**	**RefSeq Function of top two most significant (adjusted p) DEGs (log$_2$Fold Change > ±2).**
A-B	25 (0.59)	13 (0.31)	10	15	1 (4.00)	1: hypothetical protein 2: ACI57256.1 hypothetical protein
A-C	43 (1.02)	29 (0.69)	20	19	2 (4.65)	1: ACI56088.1 transcriptional regulator LuxR family 2: ACS55582.1 histidine kinase
A-D	42 (0.99)	32 (0.76)	14	18	8 (19.05)	1: ACI57516.1 transcriptional regulator XRE family 2: ACS55582.1 histidine kinase
A-E	58 (1.37)	29 (0.69)	27	19	4 (6.90)	1: ACS56907.1 hypothetical protein 2: ACS55582.1 histidine kinase
B-C	52 (1.23)	42 (0.99)	20	25	1 (1.92)	1: ACI56088.1 transcriptional regulator LuxR family 2: ACS60957.1 3-hydroxybutyrate dehydrogenase
B-D	76 (1.80)	59 (1.40)	27	37	10 (13.16)	1: ACI57516.1 transcriptional regulator XRE family 2: ACS60957.1 3-hydroxybutyrate dehydrogenase
B-E	82 (1.94)	51 (1.21)	40	31	2 (2.44)	1: ACS56366.1 hypothetical protein 2: ACS60957.1 3-hydroxybutyrate dehydrogenase
C-D	65 (1.54)	46 (1.09)	12	32	8 (12.31)	1: ACS56501.1 glutathione-dependent formaldehyde-activating GFA 2: ACS57477.1 autoaggregation protein
C-E	55 (1.30)	30 (0.71)	16	23	2 (3.64)	1: ACS56501.1 glutathione-dependent formaldehyde-activating GFA 2: ACS58639.1 hypothetical protein
D-E	37 (0.87)	12 (0.28)	21	16	8 (21.62)	1: ACS59506.1 hypothetical protein 2: AHF84664.1 hypothetical protein

Table 4.2 Number of differentially expressed orthologous core genes (DEGs) shared across genospecies and gsC subbranch comparisons

Strain grouping comparisons	Number of DEGs shared (adj. p < 0.05)	Number of DEGs shared (adj. p < 0.001)
A-B, A-C, A-D, A-E	0	0
A-B, B-C, B-D, B-E	0	0
A-C, B-C, C-D, C-E	5	0
A-D, B-D, C-D, D-E	7	1
A-E, B-E, C-E, D-E	1	0
C1-3, C1-6, C1-7, C1-8, C1-9	0	0
C1-3, C3-6, C3-7, C3-8, C3-9	1	0
C1-6, C3-6, C6-7, C6-8, C6-9	1	0
C1-7, C3-7, C6-7, C7-8, C7-9	2	1
C1-8, C3-8, C6-8, C7-8, C8-9	1	0
C1-9, C3-9, C6-9, C7-9, C8-9	4	1

4.4.2. Genospecies have distinct growth phenotypes across different Tryptone Yeast broth conditions

In order to understand how the transcriptional differences between genospecies might relate to phenotypic differences, the growth of all strains was measured under different Tryptone Yeast broth (TY) conditions.

PCA revealed the first six PCs displayed individual percentage variances greater than 5% (26.2, 14.7, 12.1, 8.3, 6.8, 5.4), with a cumulative total variance of 73.6%. The first two PCs accounting for the most variance totalled 41% of the total variance. Based on PCA, it was found that strains' growth phenotypes did cluster by genospecies (Figure 4.2a). High growth in 100% and 6.25% TY treatments between pH 5-6.68, 15-28°C has a positive loading on PC1, whereas biofilm formation at 4°C and 10 °C had a positive loading on PC2. When considering only the growth of strains in 100% TY broth at 28°C, gsB strains (which are the most genetically homogenous in the dataset) (Cavassim *et al.*, 2019) were on average the fastest growing genospecies, and gsC were the slowest (Smith, 2018).

Additionally, growth differences were analysed by PCA within gsC using the same growth trait data. The first five PCs were identified with a percentage variance explaining more than 5% (24.4, 13.9, 9.8, 6.8, 6.3), and totalling a cumulative variance of 61.3%. gsC subbranches were found to be less distinct than genospecies groups for PC1 and PC2, which explained 38% of the total variance (Figure 4.2b). The variables contributing to PC1 and PC2 for the between genospecies and within genospecies analysis were predominantly the same, suggesting the largest variation in phenotypic growth traits within gsC were in the same traits that dominated differences between genospecies. In summary, these results suggest that most genospecies (with the exception of gsA) showed clearly distinct separation respective to their growth traits, and that this separation was clearer between than within genospecies.

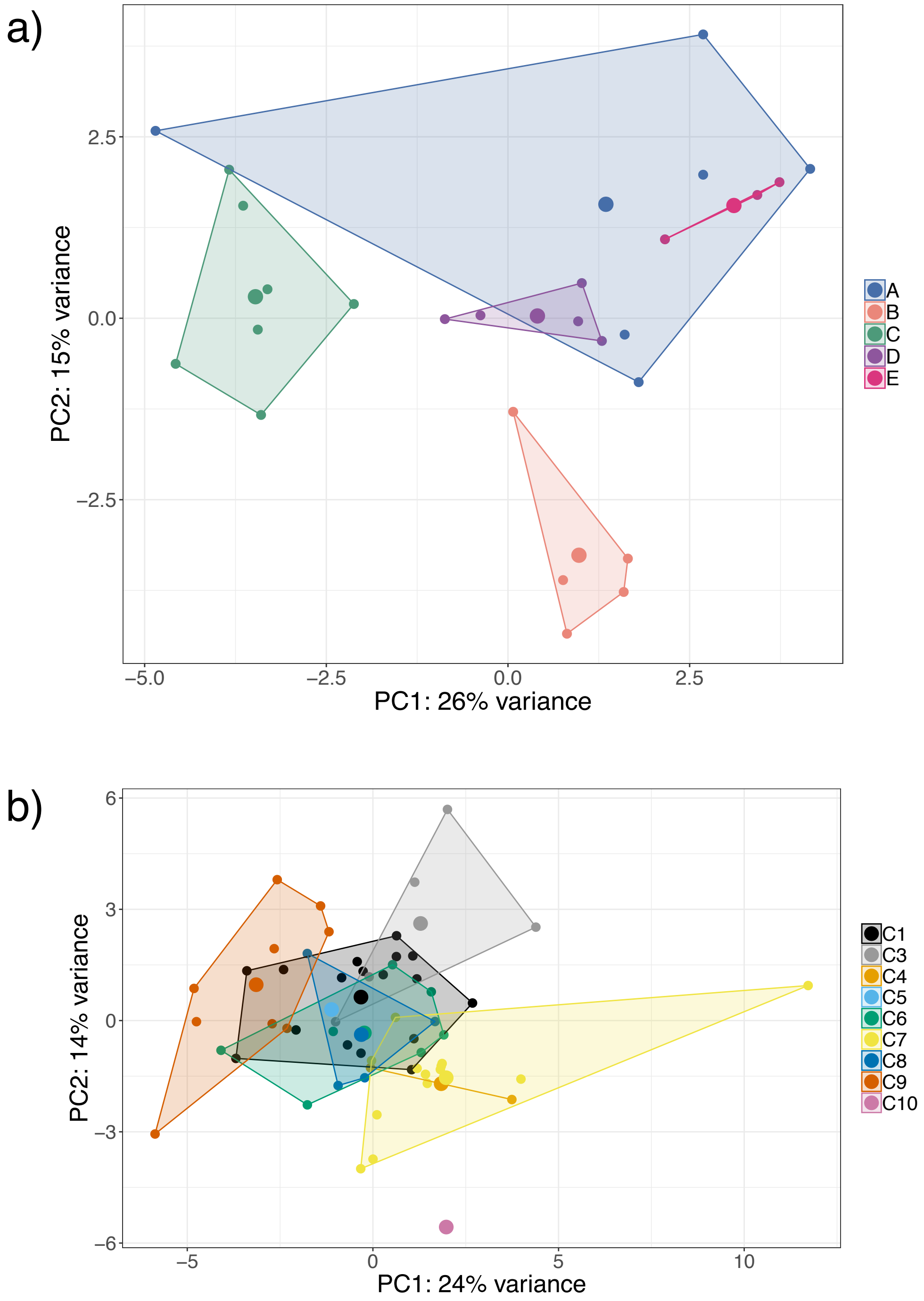


Figure 4.2 PCA of *Rlt* strains grown under different Tryptone Yeast broth (TY) growth conditions, **a)** between genospecies and **b)** between gsC strains. Strains are coloured by their genospecies classification in a), and gsC strains are coloured by their phylogenetic subbranch in b).

4.4.3. Conserved co-expressed gene groups identify genospecies expression differences that correlate to phenotypic growth traits

To compare expression differences at the level of regulatory networks, core genes were grouped by expression similarity into co-expressed gene 'modules' using Weighted Correlation Network Analysis (WGCNA) to simplify the transcriptional organisation of *Rlt* genomes and to generate a representative *Rlt* expression network.

A total of 47 *Rlt* core modules were identified, with modules containing a minimum of 7 genes and a maximum of 603 genes (Appendix Figure C.6b; Additional File 6: Table S3). 128 out of 4,229 core genes were not assigned into any of these modules.

To correlate module expression with strains' phenotypic TY growth traits, an eigengene value was calculated for each module, which is a pseudogene expression value that represents gene expression within a module (Langfelder and Horvath, 2008). Module eigengene values for each strain were then correlated to their respective growth phenotypes in TY, as described previously (Figure 4.2).

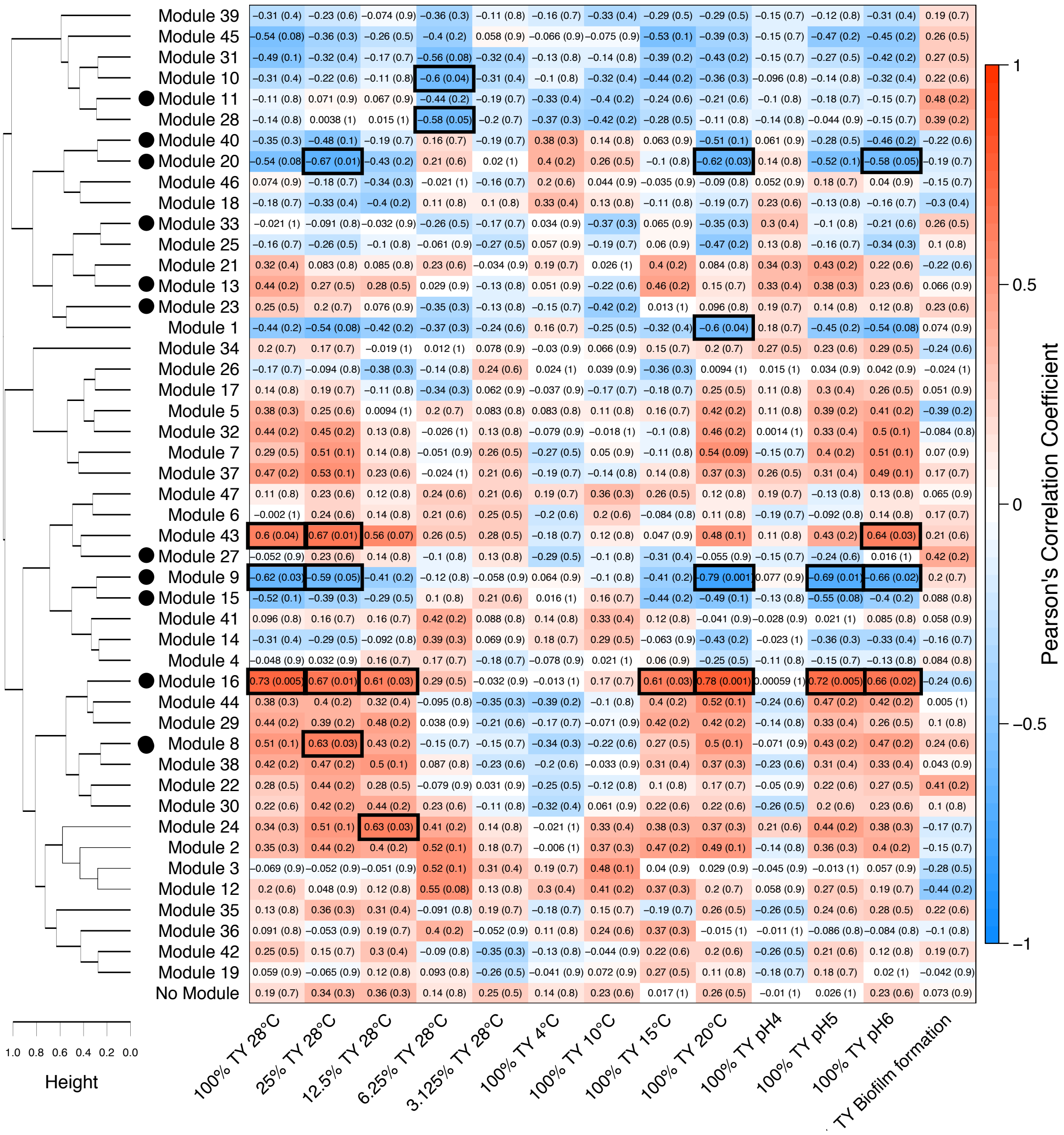
In total, 9 out of 47 module eigengene values were significantly and strongly correlated with at least one phenotypic growth trait (Pearson's correlation R value > ±0.4 and Benjamini-Hochberg corrected $p < 0.05$; Figure 4.3; Appendix Table C.3; Additional File 6: Table S3). Some modules were found to correlate with the same phenotypic traits, generating groups of 'meta-modules', which are displayed in the dendrogram (Figure 4.3).

Modules that showed significant differential expression between genospecies were also identified (Figure 4.4). Module eigengene expression values were compared across strains to identify if genospecies displayed significantly different module expression. In total, 12 of 47 modules significantly differed in module eigengene expression between genospecies (Figure 4.4; Two-way ANOVA genospecies*module interaction: $F_{118,1008} = 5.23$, $p < 0.001$; Appendix Table C.4; Appendix Table C.5). Genospecies comparisons gsA-gsC, gsB-gsC and gsB-gsD had the largest number of differentially expressed modules (3 modules). GsE had no differentially expressed modules to any other genospecies, and this is likely due to the small sample size in the analysis (gsE = 3 strains). Furthermore, no modules where all genospecies were significantly differentially expressed from one another were identified. Therefore, it is likely that a combination of module expression differences contributes to overall differences in genospecies core transcriptome profiles.

Functional pathways associated with 5 out of 12 of the modules differing in genospecies eigengene expression were identified (Additional File 6: Table S3). Module 3 was significantly differentially expressed between gsB-gsD, and was functionally associated with L-arginine, L-ornithine, putrescine and 4-

aminobutanoate metabolism. On the other hand, Module 13 and Module 33 were found to be differentially expressed between gsA-gsD and were functionally associated with two-component systems, drug metabolism and beta-lactam resistance. Module 40, containing just 16 genes, contained genes associated with glycerolipid metabolism, and was differentially expressed between gsA-gsB. Therefore, transcriptional modules associated with basic bacterial metabolism which were also differentially regulated between several *Rlt* genospecies were identified, supporting the earlier finding that genospecies showed most variance in principal metabolism genes.

Two modules displayed differences in genospecies expression and significant association with TY growth (100% TY, 28°C); Module 9 and Module 16 (Figure 4.5; Appendix Figure C.9). Mean expression of genes within these two modules also showed an association with growth for the 59 gsC strain dataset (gsC* in Figure 4.5), providing independent evidence that the association is likely correlated with growth differences (Appendix Figure C.10; Module 16 Pearson's Correlation R =0.26, $p< 0.05$; Module 9 Pearson's Correlation R = -0.42 , $p < 0.001$). Module 9 and 16 were large modules that contained 106 and 64 genes, respectively. Both sample groups of gsC (C = 6 strains and C* = 59 strains) on average expressed genes in Module 16 to a significantly lower level compared to all other genospecies (Figure 4.5a; Appendix Figure C.9a; Kruskal-Wallis: X^2 = 44.482; d.f. = 5; $p < 0.001$; Dunn's post hoc: adjusted $p < 0.05$; Appendix Table C.6). Conversely, genes in Module 9, which were associated with glycine betaine degradation, diacylglyceryl-N,N,N-trimethylhomoserine biosynthesis and quorum sensing functions were expressed to a significantly higher level in gsC (C = 6 strains and C* = 59 strains) compared to other genospecies (Figure 4.5b; Appendix Figure C.9b; Kruskal-Wallis: X^2 = 41.936; d.f. = 5; $p < 0.001$; Dunn's post hoc: adjusted $p < 0.05$; Appendix Table C.7). Taken together, strain growth was found to be significantly correlated with the expression of two gene modules which are suggested to contain genes that have putative functional associations with growth in *Rhizobium leguminosarum*.



	100% TY 28°C	25% TY 28°C	12.5% TY 28°C	6.25% TY 28°C	3.125% TY 28°C	100% TY 4°C	100% TY 10°C	100% TY 15°C	100% TY 20°C	100% TY pH4	100% TY pH5	100% TY pH6	TY Biofilm formation
Module 39	−0.31 (0.4)	−0.23 (0.6)	−0.074 (0.9)	−0.36 (0.3)	−0.11 (0.8)	−0.16 (0.7)	−0.33 (0.4)	−0.29 (0.5)	−0.29 (0.5)	−0.15 (0.7)	−0.12 (0.8)	−0.31 (0.4)	0.19 (0.7)
Module 45	−0.54 (0.08)	−0.36 (0.3)	−0.26 (0.5)	−0.4 (0.2)	0.058 (0.9)	−0.066 (0.9)	−0.075 (0.9)	−0.53 (0.1)	−0.39 (0.3)	−0.15 (0.7)	−0.47 (0.2)	−0.45 (0.2)	0.26 (0.5)
Module 31	−0.49 (0.1)	−0.32 (0.4)	−0.17 (0.7)	−0.56 (0.08)	−0.32 (0.4)	−0.13 (0.8)	−0.14 (0.8)	−0.39 (0.2)	−0.43 (0.2)	−0.15 (0.7)	−0.27 (0.5)	−0.42 (0.2)	0.27 (0.5)
Module 10	−0.31 (0.4)	−0.22 (0.6)	−0.11 (0.8)	−0.6 (0.04)	−0.31 (0.4)	−0.1 (0.8)	−0.32 (0.4)	−0.44 (0.2)	−0.36 (0.3)	−0.096 (0.8)	−0.14 (0.8)	−0.32 (0.4)	0.22 (0.6)
● Module 11	−0.11 (0.8)	0.071 (0.9)	0.067 (0.9)	−0.44 (0.2)	−0.19 (0.7)	−0.33 (0.4)	−0.4 (0.2)	−0.24 (0.6)	−0.21 (0.6)	−0.1 (0.8)	−0.18 (0.7)	−0.15 (0.7)	0.48 (0.2)
Module 28	−0.14 (0.8)	0.0038 (1)	0.015 (1)	−0.58 (0.05)	−0.2 (0.7)	−0.37 (0.3)	−0.42 (0.2)	−0.28 (0.5)	−0.11 (0.8)	−0.14 (0.8)	−0.044 (0.9)	−0.15 (0.7)	0.39 (0.2)
● Module 40	−0.35 (0.3)	−0.48 (0.1)	−0.19 (0.7)	0.16 (0.7)	−0.19 (0.7)	0.38 (0.3)	0.14 (0.8)	0.063 (0.9)	−0.51 (0.1)	0.061 (0.9)	−0.28 (0.5)	−0.46 (0.2)	−0.22 (0.6)
● Module 20	−0.54 (0.08	−0.67 (0.01)	−0.43 (0.2)	0.21 (0.6)	0.02 (1)	0.4 (0.2)	0.26 (0.5)	−0.1 (0.8)	−0.62 (0.03)	0.14 (0.8)	−0.52 (0.1)	−0.58 (0.05)	−0.19 (0.7)
Module 46	0.074 (0.9)	−0.18 (0.7)	−0.34 (0.3)	−0.021 (1)	−0.16 (0.7)	0.2 (0.6)	0.044 (0.9)	−0.035 (0.9)	−0.09 (0.8)	0.052 (0.9)	0.18 (0.7)	0.04 (0.9)	−0.15 (0.7)
Module 18	−0.18 (0.7)	−0.33 (0.4)	−0.4 (0.2)	0.11 (0.8)	0.1 (0.8)	0.33 (0.4)	0.13 (0.8)	−0.11 (0.8)	−0.19 (0.7)	0.23 (0.6)	−0.13 (0.8)	−0.16 (0.7)	−0.3 (0.4)
● Module 33	−0.021 (1)	−0.091 (0.8)	−0.032 (0.9)	−0.26 (0.5)	−0.17 (0.7)	0.034 (0.9)	−0.37 (0.3)	0.065 (0.9)	−0.35 (0.3)	0.3 (0.4)	−0.1 (0.8)	−0.21 (0.6)	0.26 (0.5)
Module 25	−0.16 (0.7)	−0.26 (0.5)	−0.1 (0.8)	−0.061 (0.9)	−0.27 (0.5)	0.057 (0.9)	−0.19 (0.7)	0.06 (0.9)	−0.47 (0.2)	0.13 (0.8)	−0.16 (0.7)	−0.34 (0.3)	0.1 (0.8)
Module 21	0.32 (0.4)	0.083 (0.8)	0.085 (0.8)	0.23 (0.6)	−0.034 (0.9)	0.19 (0.7)	0.026 (1)	0.4 (0.2)	0.084 (0.8)	0.34 (0.3)	0.43 (0.2)	0.22 (0.6)	−0.22 (0.6)
● Module 13	0.44 (0.2)	0.27 (0.5)	0.28 (0.5)	0.029 (0.9)	−0.13 (0.8)	0.051 (0.9)	−0.22 (0.6)	0.46 (0.2)	0.15 (0.7)	0.33 (0.4)	0.38 (0.3)	0.23 (0.6)	0.066 (0.9)
● Module 23	0.25 (0.5)	0.2 (0.7)	0.076 (0.9)	−0.35 (0.3)	−0.13 (0.8)	−0.15 (0.7)	−0.42 (0.2)	0.013 (1)	0.096 (0.8)	0.19 (0.7)	0.14 (0.8)	0.12 (0.8)	0.23 (0.6)
Module 1	−0.44 (0.2)	−0.54 (0.08)	−0.42 (0.2)	−0.37 (0.3)	−0.24 (0.6)	0.16 (0.7)	−0.25 (0.5)	−0.32 (0.4)	−0.6 (0.04)	0.18 (0.7)	−0.45 (0.2)	−0.54 (0.08)	0.074 (0.9)
Module 34	0.2 (0.7)	0.17 (0.7)	−0.019 (1)	0.012 (1)	0.078 (0.9)	−0.03 (0.9)	0.066 (0.9)	0.15 (0.7)	0.2 (0.7)	0.27 (0.5)	0.23 (0.6)	0.29 (0.5)	−0.24 (0.6)
Module 26	−0.17 (0.7)	−0.094 (0.8)	−0.38 (0.3)	−0.14 (0.8)	0.24 (0.6)	0.024 (1)	0.039 (0.9)	−0.36 (0.3)	0.0094 (1)	0.015 (1)	0.034 (0.9)	0.042 (0.9)	−0.024 (1)
Module 17	0.14 (0.8)	0.19 (0.7)	−0.11 (0.8)	−0.34 (0.3)	0.062 (0.9)	−0.037 (0.9)	−0.17 (0.7)	−0.18 (0.7)	0.25 (0.5)	0.11 (0.8)	0.3 (0.4)	0.26 (0.5)	0.051 (0.9)
Module 5	0.38 (0.3)	0.25 (0.6)	0.0094 (1)	0.2 (0.7)	0.083 (0.8)	0.083 (0.8)	0.11 (0.8)	0.16 (0.7)	0.42 (0.2)	0.11 (0.8)	0.39 (0.2)	0.41 (0.2)	−0.39 (0.2)
Module 32	0.44 (0.2)	0.45 (0.2)	0.13 (0.8)	−0.026 (1)	0.13 (0.8)	−0.079 (0.9)	−0.018 (1)	−0.1 (0.8)	0.46 (0.2)	0.0014 (1)	0.33 (0.4)	0.5 (0.1)	−0.084 (0.8)
Module 7	0.29 (0.5)	0.51 (0.1)	0.14 (0.8)	−0.051 (0.9)	0.26 (0.5)	−0.27 (0.5)	0.05 (0.9)	−0.11 (0.8)	0.54 (0.09)	−0.15 (0.7)	0.4 (0.2)	0.51 (0.1)	0.07 (0.9)
Module 37	0.47 (0.2)	0.53 (0.1)	0.23 (0.6)	−0.024 (1)	0.21 (0.6)	−0.19 (0.7)	−0.14 (0.8)	0.14 (0.8)	0.37 (0.3)	0.26 (0.5)	0.31 (0.4)	0.49 (0.1)	0.17 (0.7)
Module 47	0.11 (0.8)	0.23 (0.6)	0.12 (0.8)	0.24 (0.6)	0.21 (0.6)	0.19 (0.7)	0.36 (0.3)	0.26 (0.5)	0.12 (0.8)	0.19 (0.7)	−0.13 (0.8)	0.13 (0.8)	0.065 (0.9)
Module 6	−0.002 (1)	0.24 (0.6)	0.14 (0.8)	0.21 (0.6)	0.25 (0.5)	−0.2 (0.6)	0.2 (0.6)	−0.084 (0.8)	0.11 (0.8)	−0.19 (0.7)	−0.092 (0.8)	0.14 (0.8)	0.17 (0.7)
Module 43	0.6 (0.04)	0.67 (0.01)	0.56 (0.07)	0.26 (0.5)	0.28 (0.5)	−0.18 (0.7)	0.12 (0.8)	0.047 (0.9)	0.48 (0.1)	0.11 (0.8)	0.43 (0.2)	0.64 (0.03)	0.21 (0.6)
● Module 27	−0.052 (0.9)	0.23 (0.6)	0.14 (0.8)	−0.1 (0.8)	0.13 (0.8)	−0.29 (0.5)	−0.1 (0.8)	−0.31 (0.4)	−0.055 (0.9)	−0.15 (0.7)	−0.24 (0.6)	0.016 (1)	0.42 (0.2)
● Module 9	−0.62 (0.03)	−0.59 (0.05)	−0.41 (0.2)	−0.12 (0.8)	−0.058 (0.9)	0.064 (0.9)	−0.1 (0.8)	−0.41 (0.2)	−0.79 (0.001)	0.077 (0.9)	−0.69 (0.01)	−0.66 (0.02)	0.2 (0.7)
● Module 15	−0.52 (0.1)	−0.39 (0.3)	−0.29 (0.5)	0.1 (0.8)	0.21 (0.6)	0.016 (1)	0.16 (0.7)	−0.44 (0.2)	−0.49 (0.1)	−0.13 (0.8)	−0.55 (0.08)	−0.4 (0.2)	0.088 (0.8)
Module 41	0.096 (0.8)	0.16 (0.7)	0.16 (0.7)	0.42 (0.2)	0.088 (0.8)	0.14 (0.8)	0.33 (0.4)	0.12 (0.8)	−0.041 (0.9)	−0.028 (0.9)	0.021 (1)	0.085 (0.8)	0.058 (0.9)
Module 14	−0.31 (0.4)	−0.29 (0.5)	−0.092 (0.8)	0.39 (0.3)	0.069 (0.9)	0.18 (0.7)	0.29 (0.5)	−0.063 (0.9)	−0.43 (0.2)	−0.023 (1)	−0.36 (0.3)	−0.33 (0.4)	−0.16 (0.7)
Module 4	−0.048 (0.9)	0.032 (0.9)	0.16 (0.7)	0.17 (0.7)	−0.18 (0.7)	−0.078 (0.9)	0.021 (1)	0.06 (0.9)	−0.25 (0.5)	−0.11 (0.8)	−0.15 (0.7)	−0.13 (0.8)	0.084 (0.8)
● Module 16	0.73 (0.005)	0.67 (0.01)	0.61 (0.03)	0.29 (0.5)	−0.032 (0.9)	−0.013 (1)	0.17 (0.7)	0.61 (0.03)	0.78 (0.001)	0.00059 (1)	0.72 (0.005)	0.66 (0.02)	−0.24 (0.6)
Module 44	0.38 (0.3)	0.4 (0.2)	0.32 (0.4)	−0.095 (0.8)	−0.35 (0.3)	−0.39 (0.2)	−0.1 (0.8)	0.4 (0.2)	0.52 (0.1)	−0.24 (0.6)	0.47 (0.2)	0.42 (0.2)	0.005 (1)
Module 29	0.44 (0.2)	0.39 (0.2)	0.48 (0.2)	0.038 (0.9)	−0.21 (0.6)	−0.17 (0.7)	−0.071 (0.9)	0.42 (0.2)	0.42 (0.2)	−0.14 (0.8)	0.33 (0.4)	0.26 (0.5)	0.1 (0.8)
● Module 8	0.51 (0.1)	0.63 (0.03)	0.43 (0.2)	−0.15 (0.7)	−0.15 (0.7)	−0.34 (0.3)	−0.22 (0.6)	0.27 (0.5)	0.5 (0.1)	−0.071 (0.9)	0.43 (0.2)	0.47 (0.2)	0.24 (0.6)
Module 38	0.42 (0.2)	0.47 (0.2)	0.5 (0.1)	0.087 (0.8)	−0.23 (0.6)	−0.2 (0.6)	−0.033 (0.9)	0.31 (0.4)	0.37 (0.3)	−0.23 (0.6)	0.31 (0.4)	0.33 (0.4)	0.043 (0.9)
Module 22	0.28 (0.5)	0.44 (0.2)	0.28 (0.5)	−0.079 (0.9)	0.031 (0.9)	−0.25 (0.5)	−0.12 (0.8)	0.1 (0.8)	0.17 (0.7)	−0.05 (0.9)	0.22 (0.6)	0.27 (0.5)	0.41 (0.2)
Module 30	0.22 (0.6)	0.42 (0.2)	0.44 (0.2)	0.23 (0.6)	−0.11 (0.8)	−0.32 (0.4)	0.061 (0.9)	0.22 (0.6)	0.22 (0.6)	−0.26 (0.5)	0.2 (0.6)	0.23 (0.6)	0.1 (0.8)
Module 24	0.34 (0.3)	0.51 (0.1)	0.63 (0.03)	0.41 (0.2)	0.14 (0.8)	−0.021 (1)	0.33 (0.4)	0.38 (0.3)	0.37 (0.3)	0.21 (0.6)	0.44 (0.2)	0.38 (0.3)	−0.17 (0.7)
Module 2	0.35 (0.3)	0.44 (0.2)	0.4 (0.2)	0.52 (0.1)	0.18 (0.7)	−0.006 (1)	0.37 (0.3)	0.47 (0.2)	0.49 (0.1)	−0.14 (0.8)	0.36 (0.3)	0.4 (0.2)	−0.15 (0.7)
Module 3	−0.069 (0.9)	−0.052 (0.9)	−0.051 (0.9)	0.52 (0.1)	0.31 (0.4)	0.19 (0.7)	0.48 (0.1)	0.04 (0.9)	0.029 (0.9)	−0.045 (0.9)	−0.013 (1)	0.057 (0.9)	−0.28 (0.5)
Module 12	0.2 (0.6)	0.048 (0.9)	0.12 (0.8)	0.55 (0.08)	0.13 (0.8)	0.3 (0.4)	0.41 (0.2)	0.37 (0.3)	0.2 (0.7)	0.058 (0.9)	0.27 (0.5)	0.19 (0.7)	−0.44 (0.2)
Module 35	0.13 (0.8)	0.36 (0.3)	0.31 (0.4)	−0.091 (0.8)	0.19 (0.7)	−0.18 (0.7)	0.15 (0.7)	−0.19 (0.7)	0.26 (0.5)	−0.26 (0.5)	0.24 (0.6)	0.28 (0.5)	0.22 (0.6)
Module 36	0.091 (0.8)	−0.053 (0.9)	0.19 (0.7)	0.4 (0.2)	−0.052 (0.9)	0.11 (0.8)	0.24 (0.6)	0.37 (0.3)	−0.015 (1)	−0.011 (1)	−0.086 (0.8)	−0.084 (0.8)	−0.1 (0.8)
Module 42	0.25 (0.5)	0.15 (0.7)	0.3 (0.4)	−0.09 (0.8)	−0.35 (0.3)	−0.13 (0.8)	−0.044 (0.9)	0.22 (0.6)	0.2 (0.6)	−0.26 (0.5)	0.21 (0.6)	0.12 (0.8)	0.19 (0.7)
Module 19	0.059 (0.9)	−0.065 (0.9)	0.12 (0.8)	0.093 (0.8)	−0.26 (0.5)	−0.041 (0.9)	0.072 (0.9)	0.27 (0.5)	0.11 (0.8)	−0.18 (0.7)	0.18 (0.7)	0.02 (1)	−0.042 (0.9)
No Module	0.19 (0.7)	0.34 (0.3)	0.36 (0.3)	0.14 (0.8)	0.25 (0.5)	0.14 (0.8)	0.23 (0.6)	0.017 (1)	0.26 (0.5)	−0.01 (1)	0.026 (1)	0.23 (0.6)	0.073 (0.9)

Figure 4.3 Groups of co-expressed *Rlt* core gene modules correlated with phenotypic growth differences between strains. A total of 47 of co-expressed gene modules were identified from 26 *Rlt* strains. Eigengene module expression values were correlated with growth of strains in various Tryptone Yeast broth conditions shown in X-axis. The heatmap is coloured by Pearson's R correlation values, which are displayed along with bracketed Benjamini-Hochberg corrected p-values. Pearson's R > 0.4 and with an adjusted p-value < 0.05 are highlighted with bold black outlines. Modules are grouped into meta-modules using hierarchical clustering based on module eigengene value correlations. Black dots on Y-axis show modules with significant differences between genospecies groups.

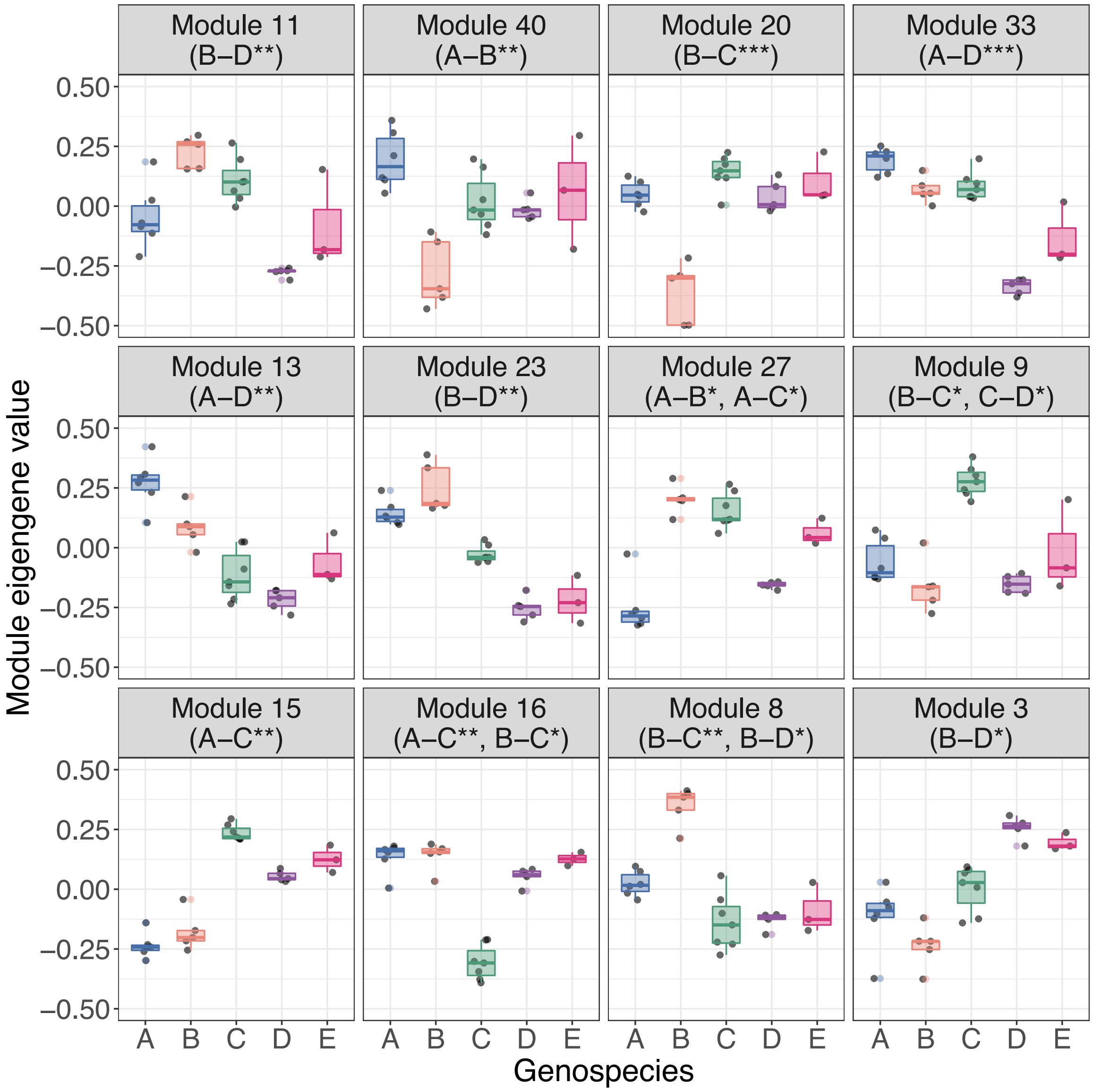


Figure 4.4 Twelve co-expressed *Rlt* core gene modules that showed representative eigengene expression values, which significantly differed between genospecies (A-E). Strains are coloured by their genospecies classification and significances between genospecies are shown at the top of each panel with significance stars equating to; adjusted p < 0.05 = *, < 0.01 = **, < 0.001 = ***. Individual strain eigengene expression values for each module are plotted within each genospecies boxplot.

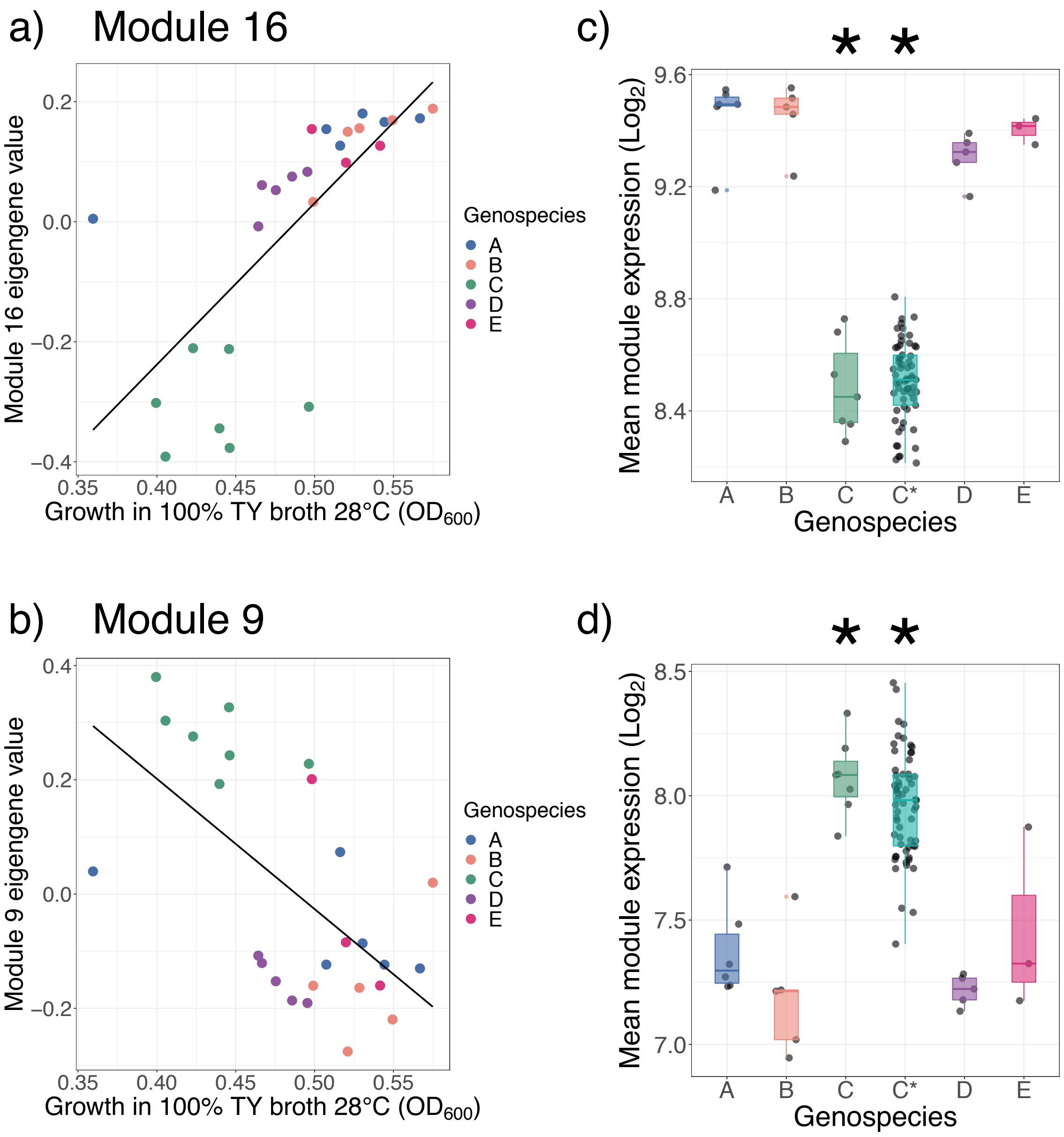


Figure 4.5 Expression of two *Rlt* core gene modules significantly correlated with growth differences between *Rlt* genospecies. **a-b)** Growth of strains (OD_{600}) correlated to eigengene expression values for Module 16 and Module 9, respectively. Pearson's correlation coefficients and adjusted p-values are provided in **Figure 4.3**. **c-d)** gsC strains displayed significantly different mean module expression to other genospecies for Module 16 and 9, respectively. Strains are coloured and grouped by their genospecies classification. * adjusted p-value < 0.05 against all other genospecies comparisons. gsC* = 59 strains (2 in duplicate) utilised for the within-genospecies analyses (see methods).

4.4.4. The expression of accessory genes is relatively lower across genospecies

As only some of the strains within each genospecies shared a small fraction of accessory genome, the average expression levels of core and accessory genomes were instead compared between genospecies. The accessory genome included 12,136 out of 16,365 genes (74.16%) in the 26 *Rlt* ortholog group pangenome. In order to compare how representative expression levels were based on the subset of strains evaluated for each genospecies group, the gene expression levels of 59 gsC strains

(Figure 4.1d) were also included in the analysis (named gsC* strains) for comparison (Figure 4.6). Core genes were found to have significantly higher levels of expression on average than accessory genes across all strains (Figure 4.6a; Table 4.3; Appendix Table C.8; Coeff$_{\text{genetypeCore}}$: estimate = 2.482, std. error = 0.0382, t = 64.97, p < 0.001; LR test: $X^2_{1,5}$= 3822.1, p < 0.001). Furthermore, parametric bootstrap testing further confirmed the difference between expression of core and accessory genes was a reliable effect (95% percentile$_{\text{genetypeCore}}$ (2.410, 2.564): original = 2.482, bias = 0.0009, std. error = 0.0383). Core genes were nearly always expressed, whereas accessory genomes contained more genes that were present but not expressed (Figure 4.6b). Further analysis showed that the higher the ortholog frequency (presence in 196 *Rlt* genomes) the higher the level at which the gene was expressed (Pearson's Correlation: R = 0.415, p < 0.0001; Appendix Figure C.5) (Vital *et al.*, 2015; Jiao *et al.*, 2018). The distributions of core and accessory genome expression levels were similar across genospecies even though genome sizes differed between genospecies; gsC strains were shown to have significantly larger genome sizes than gsD and gsE (Figure 4.7a; Kruskal-Wallis: X^2= 32.424; d.f. = 5; p < 0.001; Dunn's post hoc: p < 0.05; Appendix Table C.9). The average accessory genome size was also found to not significantly differ between a representative subset of six gsC strains compared to the 59 gsC strains used for the within-genospecies analysis (gsC*; Figure 4.7a; Appendix Table C.9). All genospecies accessory genomes displayed a similar representation of KEGG functional categories (Figure 4.7b). Overall, these results show that the accessory genes are expressed at lower levels than core genes across *Rlt* strains, and this expression level difference between core and accessory genomes was also maintained when a larger set of strains was considered (gsC*).

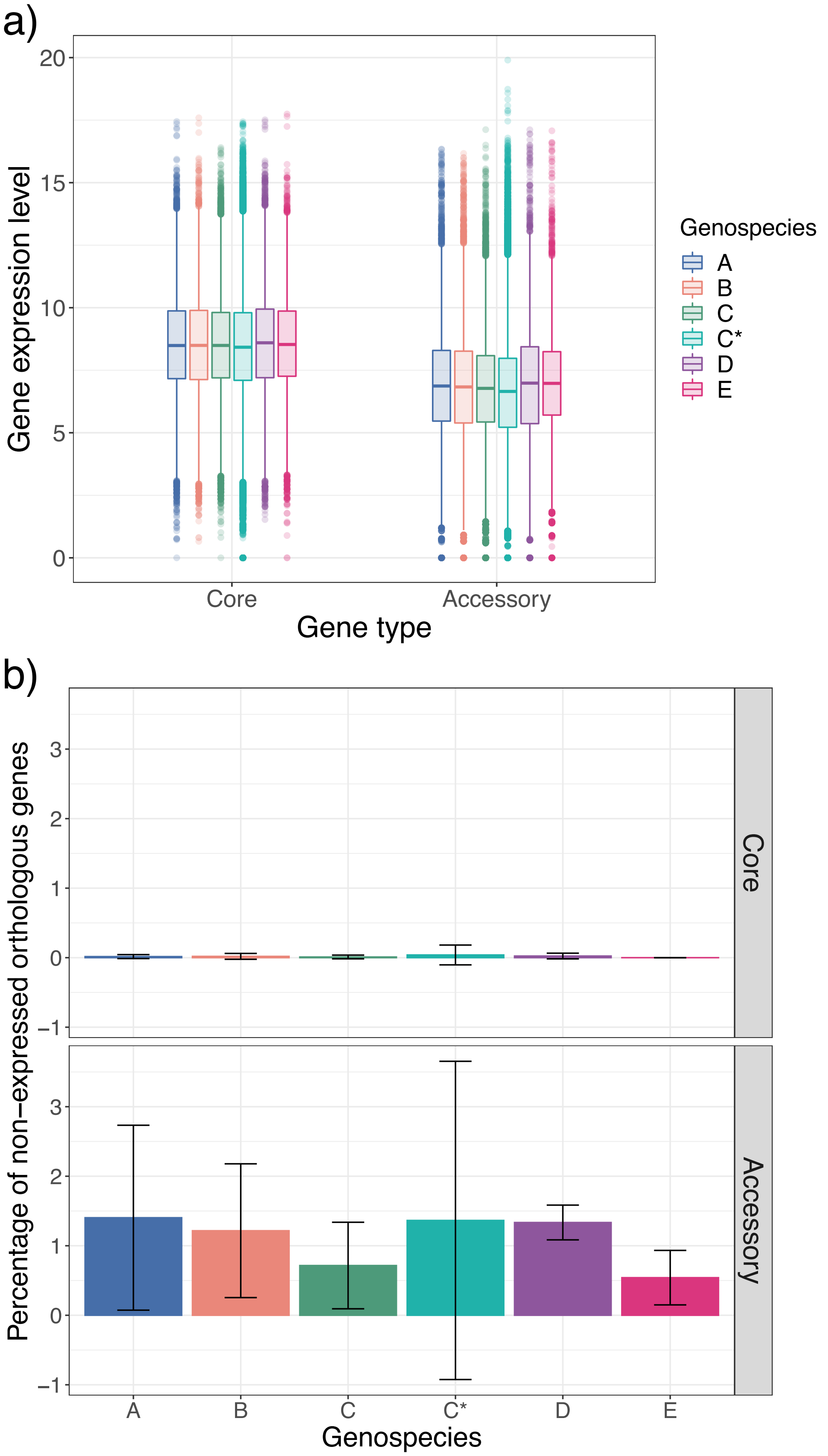


Figure 4.6 Core and accessory genome expression differences across *Rlt*. **a)** Expression levels for core and accessory genes within each strain across *Rlt* genospecies. Gene expression counts were normalised by PoissonSeq and transformed using $\log_2(n+1)$. **b)** Percentage of genes present but not expressed (0 counts) in the core and accessory genomes of strains from each *Rlt* genospecies. Error bars display the standard deviation. gsC* = 59 (2 in duplicate) strains utilised for the within-genospecies analyses (see Methods).

Table 4.3 Mean and median expression levels of *Rlt* genospecies A-E. Expression levels were calculated by normalising raw counts using PoissonSeq and transforming normalised counts by $\log_2(n+1)$ transformation. *Rlt* strains used to calculate descriptive statistics include: gsA = 6, gsB = 5, gsC = 7, gsC* = 59 (2 in duplicate) strains utilised for the within-genospecies analyses (see Methods), gsD = 5, gsE = 3.

		Genospecies					
		A	**B**	**C**	**C***	**D**	**E**
Core	Mean	8.498	8.481	8.505	8.448	8.562	8.546
	Median	8.489	8.493	8.491	8.421	8.597	8.530
Accessory	Mean	6.871	6.843	6.732	6.556	6.921	6.970
	Median	6.870	6.833	6.775	6.653	6.985	6.976

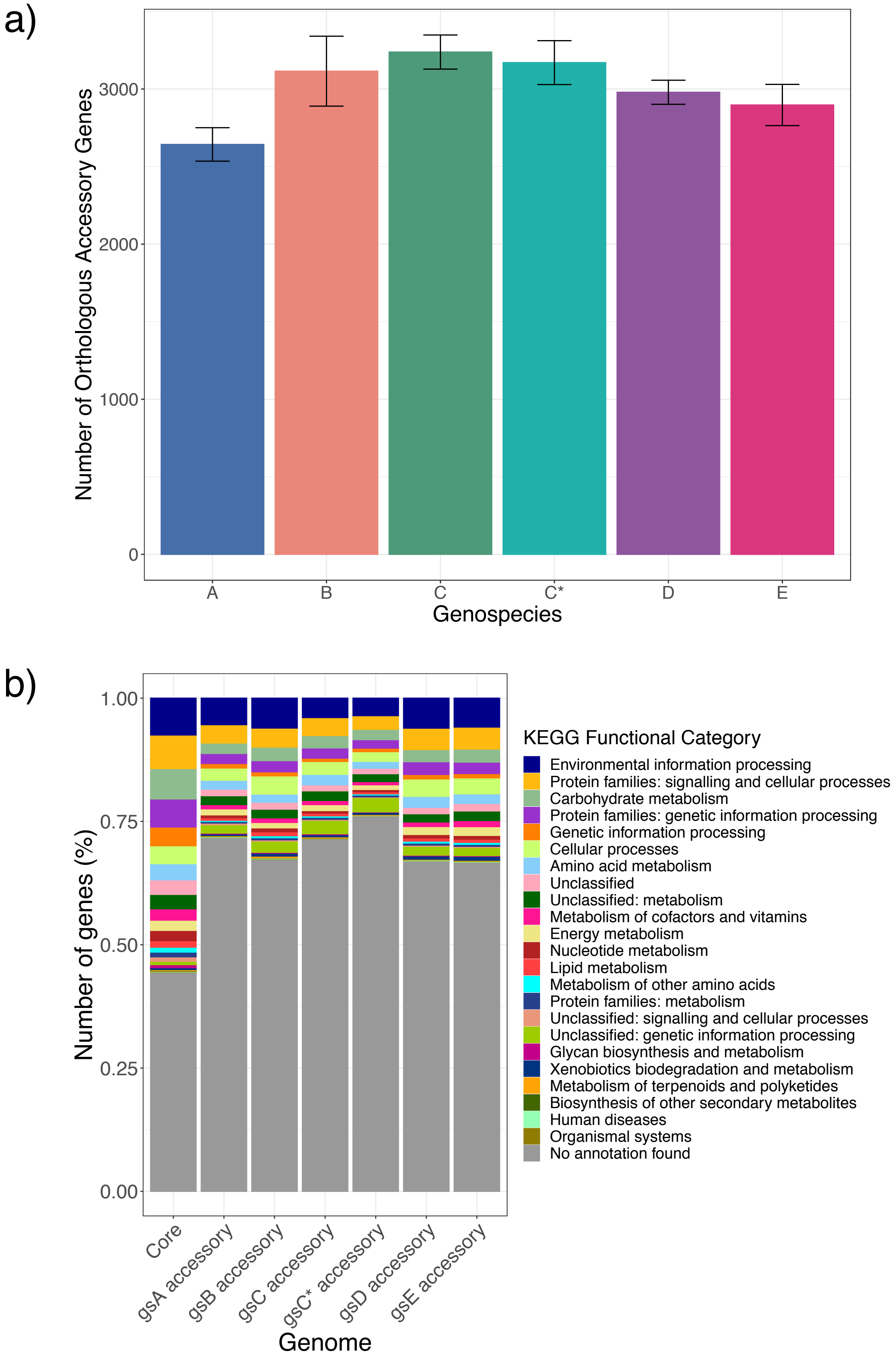


Figure 4.7 Accessory genome size and functional annotation differences between genospecies. **a)** The number of accessory genes for each strain, grouped by genospecies. Error bars display the standard deviation. **b)** Percentage number of genes from *Rlt* core genome, and genospecies accessory genomes, assigned to KEGG functional categories. gsC* = 59 (2 in duplicate) strains utilised for the within-genospecies analyses (see Methods).

4.5. Discussion

This study investigated whether genetically distinct *Rhizobium leguminosarum* genospecies display basal transcriptional core genome differences when grown in the same environment. In addition, gene expression differences were correlated with metabolic traits in order to identify the functional phenotypic differences between genospecies. To accomplish this, RNA-seq whole-genome gene expression variation was compared among 26 *Rhizobium leguminosarum* symbiovar *trifolii* strains from five genospecies (A-E), and an additional 59 strains from genospecies C. Individual strains grown under the same conditions were used to represent observable expression variation between (3-7 strains) and within (59 strains) genospecies, which to our knowledge is an experimental approach that has not been undertaken for cross-species comparisons. Genospecies boundaries contributed to observable gene expression differences at the core genome level (Figure 4.1) and strains that were more genetically diverged tended to have a higher proportion of differentially expressed core genes (Figure 4.1f). When co-expressed core genes were grouped into modules, significant correlations between transcriptomic and phenotypic differences between genospecies were observed (Figure 4.3 and Figure 4.4). These *Rlt* core gene modules were associated with growth, amino acid metabolism and two component signalling systems. Additionally, the accessory genome had significantly lower expression and a greater number of non-expressed genes compared to the core genome, and this was independent of genospecies classification (Figure 4.6). Together these results suggest that genospecies displayed differences in core genome expression when grown in the same environment. Furthermore, this transcriptomic variation was associated with phenotypic differences conserved within genospecies boundaries. However, transcriptional differences still exist within each genospecies demonstrating that substantial variation can occur within a species as similar genotypes can show differences in their gene expression.

4.5.1. Genospecies display divergence in core genome expression

Rlt genospecies, were found to display expression differences at the level of overall core transcriptome profiles and at the level of individual genes. Genospecies also displayed distinct core transcriptome profiles, as indicated by *Rlt* strains clustering into genospecies groups (Figure 4.1a and e). The overlapping expression profiles of gsA and gsB was somewhat unexpected because gsB is the most genetically homogeneous genospecies and all the strains in the collection originated from the UK, which was clearly geographically separated from the origin of other genospecies (Appendix Table C.1). One potential explanation for this is that PC1 and PC2 variances could display the stronger overriding expression differences amongst gsC, gsD and gsE, and the fundamental expression differences between gsA and gsB are explained by other principal components where gsA and gsB clusters are found to separate (Appendix Figure C.8a-b). The overlapping of *Rlt* genospecies expression profiles could indicate incomplete species divergence in their core genome expression (Vital *et al.*, 2015). However, while transcriptomes have been considered a molecular phenotype capable of identifying initial species divergence, the amount to which gene expression corresponds to definitive bacterial species difference is still disputed (Pavey *et al.*, 2010; Wolf *et al.*, 2010; Vital *et al.*, 2015; Dunning *et al.*, 2016)

At the individual gene level, DEGs (Log_2 Fold Change > ±2, adjusted $p < 0.001$) were also identified between genospecies groups (Figure 4.1b; Table 4.1). However, the number of core DEGs observed between genospecies was less than 2% of the core genome (Figure 4.1b). More genetically distant strains were shown to differ more in orthologous core gene expression (Figure 4.1), and this clustering based on genomic similarity was observed both between, and to a lesser extent within, genospecies levels (Figure 4.1b, Figure 4.1f). These results are in line with a previous study, where core genes were found to be differentially expressed between strains of the same species when grown in the same environment (Scaria *et al.*, 2013). However, gene expression variation between natural isolates has also been shown to occur predominantly on a smaller scale, with most gene expression differences displaying a Log_2 Fold Change below 2 (Townsend, Cavalieri and Hartl, 2003). Despite small variations in expression, many genes have still been observed to produce distinct phenotypic variation between isolates (Townsend, Cavalieri and Hartl, 2003). It has also been suggested that gene expression variation could be coordinated by a combined effect of several small genetic changes (Townsend, Cavalieri and Hartl, 2003). Therefore, perhaps the threshold of Log_2 Fold Change used in this study only focused on the very distinct DEGs. Consequently, these strict DEG parameters may exclude some of the phenotypic variation which only require relatively small changes in gene expression levels. Therefore, one potential reason for the relatively low number of core DEGs between genospecies (Figure 4.1b) yet clear distinction of core genome expression profiles (Figure 4.1a) could be that a combination of a few regulatory gene expression differences between genospecies possibly causes small alterations in the expression of multiple genes downstream, thereby creating distinct core genome expression profiles for different genospecies.

4.5.2. Genospecies display distinct core genome transcriptome profiles linked to expression differences in basic metabolism

Genospecies showed differences in phenotypic growth traits, such as ability to grow and biofilm formation, when exposed to abiotic stresses such as temperature, nutrient concentration and pH (Figure 4.2a)(Smith, 2018). Previously, no metabolic phenotypic traits, such as single substrate carbon utilisation, were found to be exclusive to a single genospecies (Ravin, 1963; Kumar *et al.*, 2015; Smith, 2018; Cavassim *et al.*, 2019). To further associate potential transcriptomic differences to phenotypic differences between genospecies groups, the putative functional associations of differentially expressed genes and modules were determined.

Many of the top two most significantly DEGs between genospecies were attributed to a regulatory function (Table 4.1). For example, genes matching a LuxR transcriptional regulator (RefSeq Accession: ACI56088.1) and an XRE transcriptional regulator (RefSeq Accession: ACI57516.1) were found to be most significantly differentially expressed between some genospecies comparisons (Table 4.1). LuxR transcriptional regulators are well known to be important for quorum sensing function (Wisniewski-Dyé and Downie, 2002), and correspondingly PC1 was found to be overrepresented in gene orthologs involved in quorum sensing. There are fewer examples of the XRE transcriptional regulator function, but they have associations with oxidative and high temperature stress tolerance and virulence (Gerstmeir *et al.*, 2004; Hu *et al.*, 2019). The identification of transcriptional regulators as DEGs further supports the theory that expression differences in transcriptional regulators could influence larger regulatory network differences in transcription between genospecies.

By generating *Rlt* modules of co-expressed genes and searching for enriched functional pathways, novel *Rlt* transcriptional modules associated with fundamental bacterial metabolism were identified that differed between genospecies (Additional File 6: Table S3). Grouping core *Rlt* genes into co-expressed modules provided more functional context to *Rlt* transcriptional regulation using a “guilt by association rule” whereby the functions of unannotated genes could be inferred based on their co-expression with annotated genes (Langfelder and Horvath, 2008; Hosseinkhan *et al.*, 2015). For example, gsB and gsD differed in their expression of genes associated with amino acid metabolism (including alanine, aspartate, glutamate and butanoate metabolism), which was also identified from their PCA separation (Figure 4.1a; Additional File 6: Table S2). Additionally, differential gsB-gsD eigengene expression of Module 3 was also enriched for genes associated with L-arginine, putrescine, and 4-aminobutanoate degradation (Figure 4.4; Additional File 6: Table S3). Arginine and putrescine are utilised as a precursors for many compounds in bacteria, and are crucial 'branch point' metabolites in cell functioning (Dunn, 2015). Similarly, putrescine degradation links to the 4-aminobutanoate (GABA) production pathway (Dunn, 2015), which was also found to be associated with Module 3 expression and PC2 variance. Degradation of these substrates have been suggested to contribute to amino acid cycling, central carbon metabolism, and could potentially play a role in ammonia assimilation and energy generation in mature bacteroids (Miller, 1991; Prell *et al.*, 2002; Lodwig *et al.*, 2003; Prell and Poole, 2006; White *et al.*, 2009). Therefore, expression differences were found in central amino acid metabolism between gsB-gsD, and it is tentatively speculated that these expression differences may affect genospecies ability to efficiently grow and colonise soil rhizospheres and could have implications for symbiosis.

In addition, a significantly upregulated histidine kinase family protein (RefSeq Accession: ACS55582.1) was identified in gsA compared to gsD, gsC and gsE (Table 4.1). This gene is present in WGCNA Module 13, which again shows differential expression between gsA-gsD, and is overrepresented with two-component signal transduction system genes. As a part of two-component signalling systems, membrane-bound histidine kinases can sense external stimuli and transmit signal responses to a cytoplasmic response regulator to facilitate bacterial cell changes (Borland, Prigent-Combaret and Wisniewski-Dyé, 2016). Two component systems are a key mechanism for bacteria to sense and respond to changing environments and are involved in chemotaxis response to plant root exudates (Borland, Prigent-Combaret and Wisniewski-Dyé, 2016). In *Rhizobium*, histidine kinases have been shown to be involved in amino acid metabolism and membrane stability, and loss of function can result in defective nodulation (Vanderlinde and Yost, 2012). gsA and gsD also showed differentially expressed Module 33 eigengene values which was associated with drug metabolism and beta-lactam resistance (Figure 4.4). Similarly, two-component systems and histidine kinases have also been shown to be used in beta-lactam resistance (Demanèche *et al.*, 2008; Lingzhi *et al.*, 2018). Beta-lactam resistance is widespread in soil bacteria (Demanèche *et al.*, 2008) and the genes identified within Module 33 are specifically involved in peptidoglycan cell wall recycling, and LysR regulation of beta-lactamase production. The ability to express beta-lactam resistance could be highly advantageous for colonisation of *Rlt* strains in the rhizosphere environment. Taken together, the expression of this histidine kinase ortholog and its many potential roles could be important for rhizosphere colonisation and competitiveness for nodulation between some genospecies.

Furthermore, novel co-expressed *Rlt* core gene modules highly correlated with growth traits were identified between genospecies (Figure 4.5). For example, Module 9 was enriched for pathways involved in glycine betaine degradation and diacylglyceryl-N,N,N-trimethylhomoserine biosynthesis, which is assumed to be related to carbon and nitrogen catabolism and cell membrane production, further suggesting upregulation of these pathways is linked to increased growth (Boncompagni *et al.*, 1999; Geiger *et al.*, 1999; Brhada *et al.*, 2001). Yet, upregulation of Module 9 was associated with reduced growth in genospecies comparisons (Figure 4.5b). However, Module 9 contains 106 genes and approximately half have no associated KEGG or metacyc function. As a result, there will likely be other functions associated with the module which could not be identified in this study.

Overall, these results suggest that species can differ phenotypically even when they have highly similar core genomes, as shown by transcriptomic differences between genospecies at both the level of individual genes and overall core transcriptome profiles. Therefore, the core genome similarity alone does not necessary indicate if the behaviour of bacterial strains is similar. It is possible that observed genospecies variation could be linked with strain fitness and potentially explained by *Rlt* strains' adaptation to specific resource or other niches within the plant rhizosphere.

4.5.3. Accessory gene expression levels are associated with their frequency across strains

The frequency of an orthologous gene across *Rlt* strains was found to positively correlate with increased expression levels (Appendix Figure C.5f). Gene frequencies were determined based on their presence across 196 *Rlt* genomes (Cavassim *et al.*, 2019) and correlated to the mean expression levels across up to 79 *Rlt* strains. Other studies using *E. coli* and *S. fredii* species have similarly identified core genes based on larger available species genome datasets and analysis of the expression data with a smaller number of strains showed a similar trend between orthologous gene frequency and expression levels (Vital *et al.*, 2015; Jiao *et al.*, 2018). It has been suggested that the reason for increased expression with increased genome distribution is due to more frequent orthologs having a higher level of gene connectivity in the transcriptome regulatory network (Jiao *et al.*, 2018). This suggests that accessory genes have lower expression due to their reduced integration into the core regulation network. In line with this, accessory genes were on average expressed to a lower level than core genes (Figure 4.6a). Conversely, it could simply be that the majority of accessory genes are regulated by factors not investigated in this study (Vital *et al.*, 2015). For example, large species accessory genomes can convey a multitude of diverse phenotypes (Young, 2016), and in the complex rhizosphere where strains are exposed to both other organism and heterogeneous abiotic conditions the niche-adaptive capacity provided by the accessory genome is more likely to be utilised then rather than in an axenic laboratory environment.

4.5.4. Study Limitations and future research

The accessory genome was not incorporated into the differential or co-expression analyses due to the biasing background influence of gene presence/absence population structure on the data. In some cases, the accessory genome is very large, and in this study accounts for 79.25% of the 79 strain pangenome. The complications of accessory genome incorporation have been a long-standing technical challenge of cross-species expression analyses. However, it was suggested that differential regulation of shared genes, rather than differential accessory genome content, is the greater influence of species diversification (Vital *et al.*, 2015). On the other hand, accessory genome regulation in coordination with the core genome could influence the observed genospecies transcriptional profiles and potential expression patterns are undoubtedly missed from their exclusion. Subsequently, future research could aim to understand the influence of accessory genomes on the transcription of core genes.

Furthermore, the large number of unannotated genes within the genomes limited the classification of the functional relevance of both the core and accessory genome expression patterns (Figure 4.6b). The lack of functional annotation for genes can make the relevance of observed expression patterns difficult to interpret, as also found with other studies (Ramachandran *et al.*, 2011; Vital *et al.*, 2015; McInerney, McNally and O'Connell, 2017). For that reason, curation of addition functional annotation data for non-model organisms would aid future investigations aiming to understand the functional relevance of genomic regions.

The limited number of biological replicates for validating individual strain expression patterns was an evident shortcoming of this study, and therefore future analyses would aim to include additional strain replicates within each genospecies group. Additionally, inclusion of more strains for each genospecies from across multiple geographic regions and continents could provide further insight into the potential global diversity of gene expression within different genospecies. In future, gene expression patterns that correlated to phenotypic growth traits in this study could be validated through direct lab experiments.

Ultimately, it would be interesting to observe transcriptional differences between strains under more natural environmental conditions, such as in the soil rhizosphere, or in plant root nodules in bacteroid physiology. This may aid further understanding of what has influenced genospecies divergence and how core genome variation changes depending on the environmental context.

4.5.5. Conclusions

Rlt core genome expression variation was associated with distinct phenotypic differences that were conserved within genospecies boundaries. This suggests that core genome similarity does not necessarily predict transcriptomic or phenotypic similarity, and consequently expression levels are an important indicator of species ecological characteristics. Considering the wider scope of understanding species variation, it could be proposed that the major concern for prokaryotic cross-species analysis is the lack of representation of potential variation of expression within a species. Predominantly, prokaryotic cross-species analyses have used one or two isolates to represent a species when undertaking direct species comparisons (Scaria *et al.*, 2013; Kimes *et al.*, 2014; González-Torres *et al.*, 2015; Vital *et al.*, 2015).

Therefore, experimental design of this study, using several strains within a species, is proposed as an alternative approach for cross-species analysis that considers the likely variation observed within species for comparison. By using a multi-strain comparison approach, novel co-expressed gene modules associated with bacterial metabolism were identified, which were associated with certain genospecies differences. These differences in core genome expression could potentially be adaptive and it is tempting to suggest that core expression differences between genospecies may have evolved to provide differing competitive advantages to colonisation and persistence within soil rhizospheres. However, it is noteworthy that the accessory genome is highly likely to play a significant role in the niche-adaptive phenotypic traits observed between strains, which was not fully explored here. Future investigations into the co-expression of *Rlt* modules may shed more light on the functional importance of expression pattern differences between genospecies under more ecologically realistic multi-trophic or rhizosphere-based systems.

Chapter 5. Identifying conserved operonic transcriptional units in *Rhizobium leguminosarum* symbiovar *trifolii* genospecies

5.1.Abstract

Background: Many bacterial operon prediction software have been developed. Most of these are however based on the genomes of model organisms, such as *Escherichia coli*, *Bacillus subtilis* and *Staphylococcus aureus*, and using only few representative bacterial strains per species. As a result, these software are often limited in predicting operons of more distantly related bacterial species.

Results: In this study, genomic and single-replicate transcriptomic data collected from 26 strains of *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) were used to identify transcriptional units conserved within and across five *Rlt* subspecies (genospecies A-E). A combination of ortholog identification, adjacent gene pair identification, mean intergenic distance calculations, and detection of gene co-expression were used to generate transcriptional units equating to putative operons. Calculation of deviance in expression between adjacent genes across multiple strains also supported operon predictions. Furthermore, the well-characterised *nodABCIJ* symbiosis gene operon and adjacent genes were utilised to verify the suitability of the determined parameters. Approximately 1000 transcriptional units that contained both core and accessory genes were identified for each individual *Rlt* genospecies. Additionally, transcriptional units containing genospecies-specific genes were identified including a bacterial efflux pump system and a rhizosphere-induced gene operon. In total, 94 conserved transcriptional units were found across all five genospecies.

Conclusions: The developed operon prediction pipeline utilises the variation in genomic organisation and gene expression levels across multiple strains grown under the same conditions to determine species conserved operons, which offers a further validation to operon prediction pipelines. The use of multiple strains to characterise *Rlt* species and genospecies transcriptional units additionally highlights that substantial variation in the expression of putative operons is evident within bacterial species. The generated database of putative operons for *Rhizobium leguminosarum* can be utilised to further study functional and transcriptional variation with rhizobia.

5.2.Introduction

An operon is a group of genes that are arranged consecutively in the genome and co-directionally transcribed by a common promoter and terminator region. This structure enables a set of genes to be co-transcribed into a single stand of polycistronic mRNA and it is thought that many of the genes within the genomes of prokaryotes are organised into operons. The well-studied *lac* operon in *E. coli* was the first defined classical operon, and it is now hypothesised that many genes involved in related functional pathways are organised into non-random operon structures to enable efficient co-expression (Jacob and Monod, 1961; Wolf *et al.*, 2001; De Hoon *et al.*, 2004; Koonin, 2009; Osbourn and Field, 2009). It has also been suggested that operons (predominantly newly formed) may contain genes that are in different functional pathways but are required under the same environmental conditions (Osbourn and Field, 2009). With these functional dependencies in mind, operon structures have been found to be largely dynamic and can be significantly altered by environmental influences (Okuda *et al.*, 2007; Osbourn and Field, 2009; Fortino *et al.*, 2014). The extent of genomic 'operonization' largely differs between different bacterial species and significant operon structure differences are apparent even between the strains of one bacterial species (Wolf *et al.*, 2001; Wang *et al.*, 2004; Koonin, 2009). Due to this variation, it is challenging to utilise generalised methods to predict operons within different bacterial species using solely genomic information.

The Selfish Operon Model is currently the most accepted model theorizing the persistence of conserved operons across diverse bacteria (Lawrence, 1999). The Selfish Operon Model reasons that operonic gene organisation is maintained in bacteria because it enables all genes required for a selectable phenotype to be propagated through horizontal co-transfer as well as vertical transmission as a result of their close proximity (Lawrence, 1999; Koonin, 2009; Osbourn and Field, 2009). Consequently, the model suggests that gene clustering is beneficial to the constituent genes themselves within the host, rather than purely because of the importance of functional coregulation to the host organism (Lawrence, 1999).

Genes with a conserved order across several or more bacterial genomes have a high probability of belonging to an operon (Ermolaeva, White and Salzberg, 2001; Wolf *et al.*, 2001; Edwards *et al.*, 2005; Junier and Rivoire, 2016). Conservation of gene order across species could be explained by horizontal gene transfer of a block of genes, recent divergence preventing disruption of gene order, or stable maintenance of the gene order due to positive effect on organism fitness (Tamames *et al.*, 1997; Tamames, 2001). The size of conserved genome regions across bacterial species can vary from small 2-gene operons up to big syntenic gene blocks (uber-operons) (Tamames *et al.*, 1997; Tamames, 2001; Junier and Rivoire, 2016). On the other hand, the reduction of operon conservation has been linked with genome divergence distance, evolutionary lineage, and operon complexity (Itoh *et al.*, 1999; Okuda *et al.*, 2007). Previous comparative analyses have found that few operons are fully conserved across species and the 'operome' of individual strains is primarily comprised of unique operons (Salgado *et al.*, 2000; Wolf *et al.*, 2001; Koonin, 2009). Even closely related strains can have regions with no conserved gene order presumably indicative of regions of active rearrangement or constitution of predominantly unique genes (Tamames, 2001). Within operons, gene order could be reorganised frequently during evolution and therefore destruction of operon gene order is seen as an essentially neutral process in the long-term evolution of genome structure (Itoh *et al.*, 1999). For example, there could be flexibility in gene order within operons if it is not essential for operon-product functioning. Addition or insertion of genes into pre-existing operons also can alter the transcriptional structure and generate new operons (Price, Arkin and Alm, 2006). However, this instability of gene order within operons can also lead to difficulties in identification of conserved operon structures (Price, Arkin and Alm, 2006).

Operon prediction using genomic and transcriptomic data has seen extensive development in prokaryotes (examples of software currently available for operon prediction are outlined in Table 5.1). For example, it has recently been shown that a combination of gene expression data with intergenic gene distances provides an additional crucial determinant of successful operon detection (De Hoon *et al.*, 2004). Moreover, genes within and outside operons have been shown to display significant overlap in their intergenic gene distances, and therefore co-expression data is vital for identifying co-regulated gene sets and potentially functional operons (De Hoon *et al.*, 2004). Previous operon prediction studies have largely focused on the genomes of model prokaryotes such as the *Escherichia coli*, *Bacillus subtilis* and *Staphylococcus aureus*, and in some cases only the core chromosomal genes have been considered (Huerta *et al.*, 1998; Salgado *et al.*, 2000; De Hoon *et al.*, 2004; Wang *et al.*, 2004;

Bergman *et al.*, 2007; Chuang *et al.*, 2012; Conway *et al.*, 2014; Fortino *et al.*, 2014; Taboada *et al.*, 2018). As a consequence, most of the operon databases rely on these model organisms and the applicability of these databases in context of more distantly related bacterial species have been questioned due to the lack of experimental evidence supporting the operon predictions.

Table 5.1 Operon prediction software and their references.

Operon Software (Reference)	**Application**
DOOR/DOOR2 (Mao *et al.*, 2009, 2014)	• DOOR: Open access database • DOOR2: operon prediction algorithm using genomic information and not transcriptomic data. • Calculates level of similarity between related operons in different organisms from database.
OperonDB (Pertea *et al.*, 2009)	• Identify gene pairs located on the same DNA strand across different bacterial genomes.
ProOpDB (Taboada *et al.*, 2012)	• Neural networking to predict operons stored in ProOpDB database.
Rockhopper (McClure *et al.*, 2013)	• Algorithm to predict operon structure and transcriptional start and stop sites using RNA-seq mapping data, intergenic distance and expression correlations across experiments. • Run on single RNA-seq dataset at a time.
RegulonDB (Huerta *et al.*, 1998)	• Database for *E. coli* K-12 putative operons identified across different growth treatments.
MicrobesOnline (Dehal *et al.*, 2009)	• Utilises microarray data and genomic information to combine phylogenetic analysis of genes and correlation of expression profiles to identify conserved putative operons.
OperomeDB (Chetal and Janga, 2015)	• Uses RNA-seq data to identify condition-specific putative operon structures. • Uses Rockhopper software to operon prediction and iBrowse to visualise predicted operons. • Analysis carried out individually for each bacterial genome. • Suggested for comparative operomics analysis.
REMap (Pelly *et al.*, 2016)	• Utilises BAM file transcriptomic data, gff file and user-determined expression parameters. • Algorithm evaluates transcription coverage within genes and intergenic regions (intergenic region length not limited). Does not rely heavily on gene structure or functional annotations. • Algorithm can be modified for specific bacterial species. • Alternative to Rockhopper, it does not split putative operons if ORF are identified on the complementary strand.
CONDOP (Fortino, Tagliaferri and Greco, 2016)	• R package • Determines operon pairs and non-operon pairs with genomic (genome sequence, gff and DOOR files) and raw count transcriptomic data • Uses three machine learning approaches (neural networks, support vector machines, random forests) to identify adjacent similarly identified genes and link them into operon groups.
SeqTU (Chen *et al.*, 2017)	• Developed by the creators of DOOR. Part of the DOOR2 package. • Uses RNA-seq expression level continuity and variance. • Gene functional relatedness evaluated with KEGG and GO terms. • Calibrate organism-specific predictor parameters. • Uber-operon predictor.
Operon-Mapper (Taboada *et al.*, 2018)	• Uses only genomic sequences to calculate intergenic distances and relationships between gene functions to generate an artificial neural network for operon prediction. • Transcriptomic data is not considered. • Provides a score of how likely gene pairs are in the same operon.

Operon prediction studies have outlined several common parameters that must be maintained between gene pairs if they are to be considered part of a transcriptional unit or operon (Chuang *et al.*, 2012). These include:

1) genes must be less than a specific threshold of base pairs apart and based on previous studies a threshold between 20 bp to 300 bp is commonly used (Salgado *et al.*, 2000; Ermolaeva, White and Salzberg, 2001; De Hoon *et al.*, 2004; Wang *et al.*, 2004; Price, Arkin and Alm, 2006; Bergman *et al.*, 2007; Fortino *et al.*, 2014; Wang, MacKenzie and White, 2015; Taboada *et al.*, 2018);

2) gene pairs must be adjacent on the genome (Salgado *et al.*, 2000; Price, Arkin and Alm, 2006; ten Broeke-Smits *et al.*, 2010);

3) pairs must be on the same DNA strand (Eyre-Walker, 1995; Salgado *et al.*, 2000; Ermolaeva, White and Salzberg, 2001; Wang *et al.*, 2004; Price, Arkin and Alm, 2006);

4) pairs must be co-expressed (Eyre-Walker, 1995; De Hoon *et al.*, 2004; Price, Arkin and Alm, 2006; ten Broeke-Smits *et al.*, 2010; Fortino *et al.*, 2014; Wang, MacKenzie and White, 2015; Slager, Aprianto and Veening, 2018);

5) identified orthologs in other strains must also be neighbouring/adjacent genes or have a conserved order (Wolf *et al.*, 2001; Wang *et al.*, 2004; Price, Arkin and Alm, 2006; Bergman *et al.*, 2007).

Other additional parameters used in previous studies include confirmation of operons using reference genomes and by considering the functional relationships between adjacent genes (Salgado *et al.*, 2000; De Hoon *et al.*, 2004; Fortino *et al.*, 2014; Taboada *et al.*, 2018). Other investigations have also identified operons via other methods, such as using Hidden Markov models (Yada *et al.*, 1999; Bergman *et al.*, 2007), and identifying ribosomal binding sites and shine-dalgarno sequences (Yada *et al.*, 1999), termination sites (Wang *et al.*, 2004; Conway *et al.*, 2014; Wang, MacKenzie and White, 2015), codon-usage patterns, intergenic region expression patterns (Fortino *et al.*, 2014), RNA-seq read coverage (Conway *et al.*, 2014) and using log-likelihood methods (Ermolaeva, White and Salzberg, 2001).

Some operons can be organised into multiple transcriptional units whereby the expression of groups of adjacent genes within the operon can differ (Okuda *et al.*, 2007). This expression difference can depend on external stimuli that alter gene regulation or the presence of internal promoters and terminators between operon-genes (Okuda *et al.*, 2007). It is well known that genes within operons have shorter intergenic distances than genes outside an operon, or at the borders of transcriptional units within operons (Salgado *et al.*, 2000). This is likely because genes that have shorter intergenic distances also commonly display higher levels of coregulation (Okuda *et al.*, 2007). However, some intergenic regions within an operon can be longer than others due to incorporation of internal regulatory promoter and terminator elements (Okuda *et al.*, 2007). Intergenic distance can also indicate operon age, as newer operons have less-optimal spacing between genes compared to older operons as a result of deletions within intergenic regions over generations (Price, Arkin and Alm, 2006). Relatedly, due to fine-tuning of gene regulation, highly expressed operons are more likely to have relatively larger intergenic regions in order to avoid over-transcribing unnecessary proteins (Price, Arkin and Alm, 2006).

Here we studied operon prediction by using genetic information and gene expression data from 26 strains of *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) grown under the same conditions to build conserved 'operomes' for five genetically distinct *Rlt* sub-species (genospecies A-E). The term 'transcriptional unit' was used as opposed to operon because regulatory promoter or terminator regions around putative operons were not identified, and differential expression has been observed between polycistronic genes that occur within the same operon (Okuda *et al.*, 2007; Conway *et al.*, 2014). Transcriptional units were determined by evaluating the mean intergenic distances and expression level correlations between adjacent orthologous genes across *Rlt* strains. Calculating the deviance in expression between adjacent gene pairs across strains also provided additional supporting evidence to transcriptional unit classification, despite its previously limited application in operon prediction (Fortino *et al.*, 2014). The well-characterised *Rhizobium leguminosarum nodABCIJ* symbiosis gene operon and neighbouring genes were used to confirm the suitability of the determined transcriptional unit parameters (Hong, Burn and Johnston, 1987b). Expression of this operon was ensured by the addition of clover flavonoid to the strain growth media (Djordjevic *et al.*, 1987). Genospecies-conserved putative operons were further compared to identify transcriptional unit structures that were maintained holistically across the *Rlt* species. The utilisation of multiple strains to characterise species conserved operons by taking into account gene expression variation across strains with differing genomic contents offered a further valuable verification to operon prediction pipelines. Findings demonstrate that genomic and single-replicate transcriptomic data generated from a collection of strains can be used to identify transcriptional units that are conserved at the species level. This investigation has subsequently created a database of putative operons for *Rhizobium leguminosarum* that can be utilised for further functional testing and investigations concerning transcriptomic regulation.

5.3.Methods

5.3.1. Genome and transcriptome data sources and strain metadata

Three to seven *Rhizobium leguminosarum* symbiovar *trifolii* strains from each genospecies (gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3: Total = 26) were selected from the 196 *Rlt* strain NCHAIN collection for this study (Cavassim *et al.*, 2020). Only the orthologous genes identified from the 196 *Rlt* strain collection were utilised for the operon prediction (Cavassim *et al.*, 2020). These orthologous gene groups were previously identified using ProteinOrtho (v.5.16b) and were functionally annotated using Prokka (v.1.12), to provide putative RefSeq accession numbers and protein product information (Cavassim *et al.*, 2020).

Transcriptome data for the 26 strains was obtained and processed as described previously (Chapter 4). In brief, to generate the transcriptome data the strains were cultured individually in 5 ml of Tryptone Yeast (TY) broth (5 g Tryptone, 2.5 g Yeast Extract, 1.47 g $CaCl_2$, per litre volume) with 1µM 7,4'-dihydroxyflavone (clover flavonoid stock concentration solubilised in DMSO) for 48 hours, 28°C, 180 rpm. 7,4'-dihydroxyflavone was added to the TY broth to induce expression of *Rlt* nodulation gene operon, *nodABCIJ*, which was used to validate the operon prediction threshold parameters in this study (Djordjevic et al., 1987). Raw gene expression count data was normalised based on expression of 4,229 *Rlt* core genes using PoissonSeq (v.1.1.2), and further log transformed, as described previously (Chapter 4). Metadata for strains can be found in Appendix C: Chapter 4 (Appendix Table C.1).

5.3.2. Transcriptional unit generation

The following analyses were undertaken for each genospecies individually in order to generate a dataset of putative operons for each genospecies. For generation of transcriptional units per genospecies, only ortholog group gene pairs that were present in at least three strains, and adjacent in one strain, of a genospecies were considered. Additionally, genes were only considered adjacent if they were on the same DNA strand (i.e. transcribed in the same direction). For gsE, only gene pairs that were present in all three gsE strains were considered for the transcriptional unit analysis.

For each adjacent gene pair, the mean number of base pairs between adjacent orthologous genes (intergenic distance) across strains was calculated. Pearson's correlation between adjacent gene pairs expression was calculated from the PoissonSeq normalised, $\log_2$ transformed gene expression data using the cor.test function from psych R package (v.1.8.12). Additionally, the deviance in gene expression of adjacent gene pairs across strains (i.e. mean difference in gene expression) was calculated using the PoissonSeq normalised, $\log_2$ transformed gene expression data by the following equation:

$$\frac{\sum\left((S_xG_i - S_xG_j)^2\right)}{\text{number of strains}} = \text{Deviance score}$$

Where the expression difference between adjacent genes, gene *i* (G_i) and gene *j* (G_j), for each strain (S_x) is totalled and normalised by the number of strains with the adjacent gene pair.

Subsequently, adjacent genes within each genospecies were filtered to make transcriptional units for each *Rlt* genospecies. Transcriptional units were identified by stringing together adjacent gene pairs that met the following criteria: 1) intergenic distance must be less than 200 base pairs between adjacent gene pairs; 2) adjacent gene pairs must have correlated expression with a Pearson's correlation R statistic > 0.8; 3) adjacent genes must have a deviance score < 3.

Adjacent gene pairs that did not meet these criteria signalled the end of the transcriptional unit in the string of adjacent genes that met these criteria. Intergenic distance and Pearson's correlation parameters stated above were chosen with consideration to prediction parameters used by previous studies (Dam *et al.*, 2007; ten Broeke-Smits *et al.*, 2010; Wang, MacKenzie and White, 2015; Chen *et al.*, 2017; Slager, Aprianto and Veening, 2018), and from correlation coefficient distribution patterns of adjacent genes using gsB genomic and transcriptomic data. The deviance score parameter was also based on the deviance score distribution patterns of adjacent genes using gsB genomic and transcriptomic data. gsB was used because strains in this genospecies were the most genetically homogeneous (Cavassim *et al.*, 2020). Suitability of the determined transcriptional unit parameters were confirmed based on the known *nodABCIJ* symbiosis gene operon and surrounding *Rlt* SM3 strain symbiosis plasmid genes as it is a well-characterised, known *Rlt* operon, which was actively expressed due to the addition of clover flavonoid in the strain growth media (Djordjevic et al., 1987).

5.3.3. Core and accessory gene classification

Genes were classed as core or accessory based on their frequency in 196 *Rlt* strains (Cavassim *et al.*, 2020). If genes were present in all 196 strains, they were considered core genes, which resulted in 4,229 core genes including 25 symbiosis genes. All other genes were considered as accessory. Genospecies enriched genes were classed as genes present in at least 90% of strains in a genospecies from the 196-strain dataset, and absent in at least 90% of the four other genospecies.

5.3.4. Transcriptional unit validation with known *Rlt* operons and WGCNA *Rlt* core gene modules

Genospecies operon predictions were evaluated by searching for known rhizobia operons within the identified transcriptional units. Known operon gene reference sequences were used to search for orthologous gene groups in strain genomes using nBLAST. If matched operons in genomes did not contain the full set of orthologous BLASTn gene groups, then the intergenic distances, correlation coefficients and deviance scores of matched ortholog gene groups were assessed to determine why genes were excluded from known operons.

To calculate how many operons were maintained when intersected into previously calculated weighted gene correlation network analysis (WGCNA) *Rlt* core gene modules, only the core genes within transcriptional units were considered. Previously (Chapter 4), co-expressed core ortholog genes were grouped into modules with WGCNA R package (v.1.66). PoissonSeq normalised, $\log_2(n+1)$ transformed counts were used as input for WGCNA. All WGCNA default settings were used with the exception that a soft threshold power of 7 was used and the minimum number of genes to form a distinct module was set to 3. In order to observe how core genes in modules intersected, only transcriptional units that contained 2 or more core genes were included in the analysis. This was to determine whether core genes within a transcriptional unit were either split or maintained by a WGCNA module.

5.4.Results

5.4.1. Validation of transcriptional unit parameters

A dataset of transcriptional units was generated for each genospecies individually, based on genomic and transcriptomic information collected for three to seven strains of each genospecies. To classify orthologous gene groups into conserved transcriptional units for each *Rlt* genospecies separately (gsA-E), gene pairs present in at least three strains and adjacent in at least one strain were considered for transcriptional unit assignment. Genes were only considered adjacent if they were located next to each other on the same strand of DNA (i.e. transcribed in the same direction). With consideration to previous parameters used for operon identification, the following thresholds were applied to gene pairs based on their genomic location and expression within a genospecies: 1) pairs must have a mean intergenic region of less than 200 base pairs (intergenic distance measure); 2) comparison of adjacent gene pairs expression must have a Pearson's correlation R statistic > 0.8; 3) adjacent genes must have a gene expression deviance score < 3. The genomic information and gene expression patterns in five strains from gsB were utilised to test the suitability of the chosen transcriptional unit parameters (Figure 5.1 and Figure 5.2). Furthermore, the known symbiosis nodulation gene operon, *nodABCIJ,* and neighbouring symbiosis plasmid genes were additionally used to confirm the suitability of the parameter thresholds (Hong, Burn and Johnston, 1987b), as the expression of this operon was ensured by the addition of clover flavonoid to the growth media (Djordjevic *et al.*, 1987). Using these parameters, 1,122 transcriptional units containing 3,097 (44.65%) ortholog gene groups were identified from 6,936 ortholog gene groups in gsB.

For all genospecies, the majority of gene pairs were found to have a mean intergenic distance of around 40 bp, and a second smaller peak between 50-100 bp (Figure 5.1a). Gene pairs with intergenic distances of 40 bp or less have a high possibility of being in the same operon or transcriptional unit. However, for this pipeline gene pairs were required to have a mean intergenic distance of 200 bp or less to be considered transcriptional unit gene pairs, as the intergenic distribution tails off after 200 bp and this maximum distance threshold is similar to previously identified operons. In order to assess how clear transcriptional units can be observed based on intergenic distances along a genomic region, the gsB mean intergenic distances were plotted between adjacent genes ordered by the positive and negative strand of the *Rlt* strain SM3 (gsB) symbiosis plasmid (Figure 5.1b). *nodABCIJ* operon genes showed that intergenic distances within the operon were below 200 bp and intergenic regions to adjacent genes outside of the operon were greater than 200 bp (Figure 5.1b). This highlighted the group of *nod* genes as a distinct operon based on intergenic distance alone.

The correlation in expression of adjacent gene pairs was observed to tail off below a correlation coefficient of 0.8 (Figure 5.2a). Additionally, the expression correlations for all gene pair combinations in relation to gene order across a genome was assessed using the SM3 symbiosis plasmid gene region as a reference for the potential gene order. Gene pair combinations on opposite strands, and pairs separated by many genes along the same strand, were found to display high expression correlations (Figure 5.2b). For example, the *nodABCIJ* operon was strongly correlated with expression of gsB-enriched genes and symbiosis genes on the opposite strand (Figure 5.2b). However, when the correlations only between adjacent genes across the SM3 symbiosis gene region were observed, it was clearer that correlations in expression were stronger at the start of the operon region and tailed off below 0.8 at the end of the operon region, as demonstrated by the *nodABCIJ* operon and surrounding genes (Figure 5.2c). It was further identified that the majority of adjacent gene pairs with an intergenic distance of approximately 245 bp or less seemed to predominantly display co-expression correlation coefficient values above 0.75 (Appendix Figure D.1). This suggests that the majority of tightly clustered genes showed high co-expression. However, there were also a larger number of instances where genes with small intergenic distances did not display strong co-expression correlations.

In addition, the deviance in expression between adjacent genes was considered. Analysis of the distribution of deviance scores for all adjacent genes in gsB identified that the majority of adjacent genes had an expression deviance score below 3, and deviance scores levelled off after a score of 5 (Figure 5.2d). This suggests that most genes within an operon will have a deviance of less than 5, and most notably less than 3, where the number of gene pairs drastically increases (Figure 5.2d). However, gsB expression deviance scores across the SM3 symbiosis gene region identified that genes at separate parts of the genome also displayed deviance scores below 3, similarly to expression correlations (Figure 5.2e). Additionally, tracking of the deviance score across the SM3 symbiosis gene region highlighted fluctuations in deviance score that seemed to associate with genes clustering by transcriptional units (Figure 5.2f). Estimates of expression deviance below 3 are supported by the *nodABCIJ* operon, where deviance between *nod* genes is close to 0. Due to the use of expression information from different strains with differing genomic contents, a less conservative deviance score of 3 was used to indicate a potential transcriptional unit, in order to account for possible strain-dependent expression variation. Similarly, deviance scores were found to correlate with intergenic distance up to approximately 245 bp ($3x10^2$ on logged intergenic distance axis), after which deviance scores between adjacent gene pairs with larger intergenic distances appeared to become random (Appendix Figure D.2).

Additionally, approximately 12% of adjacent gene pairs had a negative mean intergenic distance, whereby gene pair protein coding regions overlapped (Table 5.2). A negative mean intergenic distance of -4 bp was the most common overlapping gene distance, and this pattern was observed across all five genospecies (Appendix Figure D.3). An adjacent gene pair overlap of -4 bp has been shown to be a standard operon organisation for bacteria, which allows for a frame shift in the stop codon of one gene and the start codon of the consecutive gene (Johnson and Chisholm, 2004; Price, Arkin and Alm, 2006; Sabath, Graur and Landan, 2008; Huvet and Stumpf, 2014). For gsB, 607 out of 7103 adjacent gene pairs (8.55%) had a mean intergenic distance overlap of -4 bp, and 422 of those 607 gene pairs (69.52%) were tagged as transcriptional unit pairs after further filtering with consideration to co-expression correlation and deviance scores. All gene pairs with negative mean intergenic distances were found to be very highly conserved across all strains within a genospecies, to the extent that all negative gene pairs displayed the exact same number of base pair overlaps in all strains with the exception of one to two gene pairs per genospecies. Overall, the distributions of adjacent gene intergenic distances and expression similarities across multiple strains (Figure 5.1a, Figure 5.2a and Figure 5.2d), and further observing the profiles of these parameters across a genomic region

(Figure 5.1b, Figure 5.2c Figure 5.2f), can be used to infer the appropriateness of parameter thresholds for operon prediction.

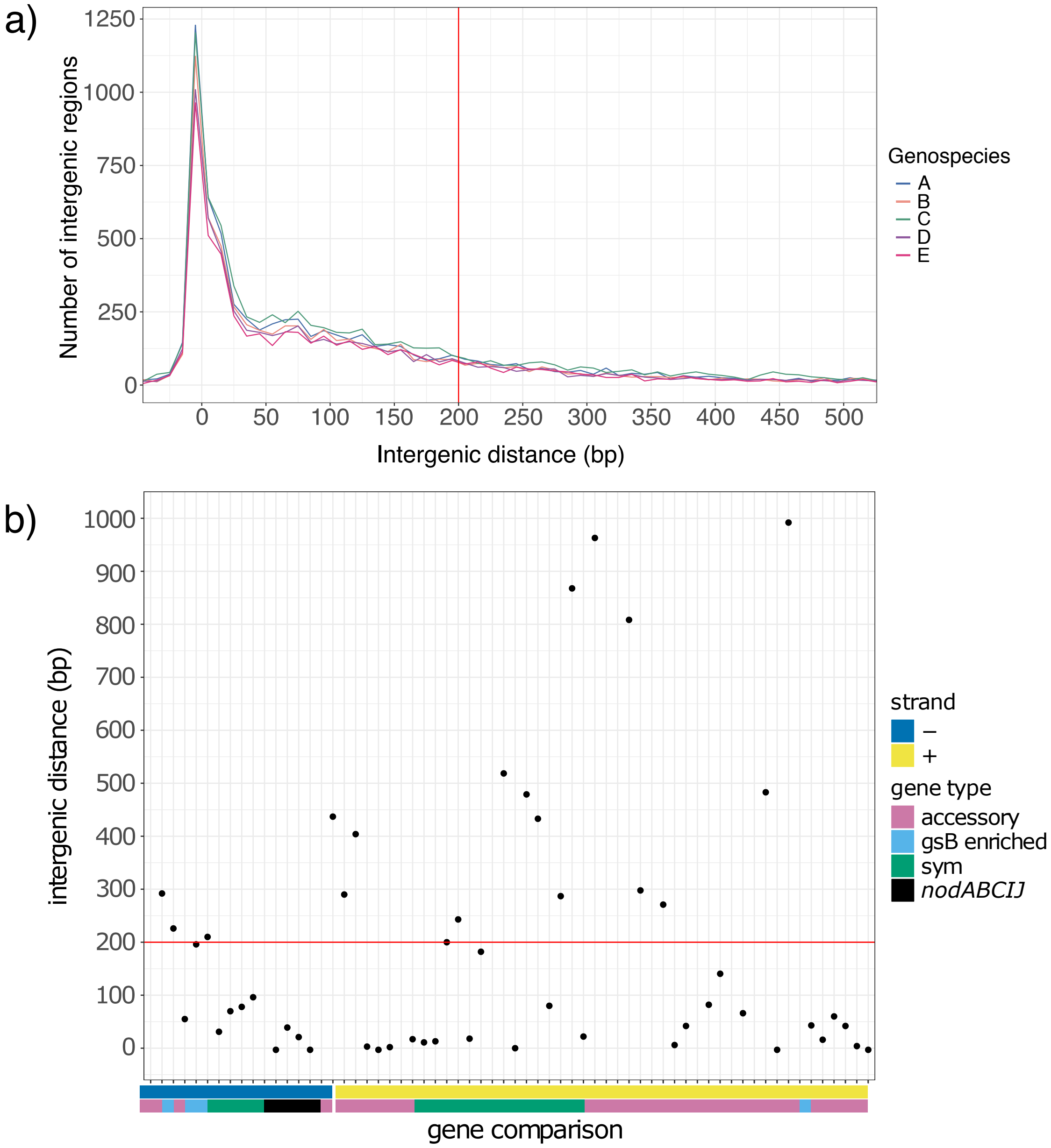


Figure 5.1 Validation of intergenic distance thresholds to determine genospecies transcriptional units. **a)** Distribution of intergenic region lengths (base pairs) between adjacent genes for genospecies A-E. Red line indicates 200 bp**. b**) gsB mean intergenic distance (base pairs) for adjacent genes arranged by their order along a region of the symbiosis plasmid in gsB strain, SM3. Red line indicates an intergenic distance of 200 bp.

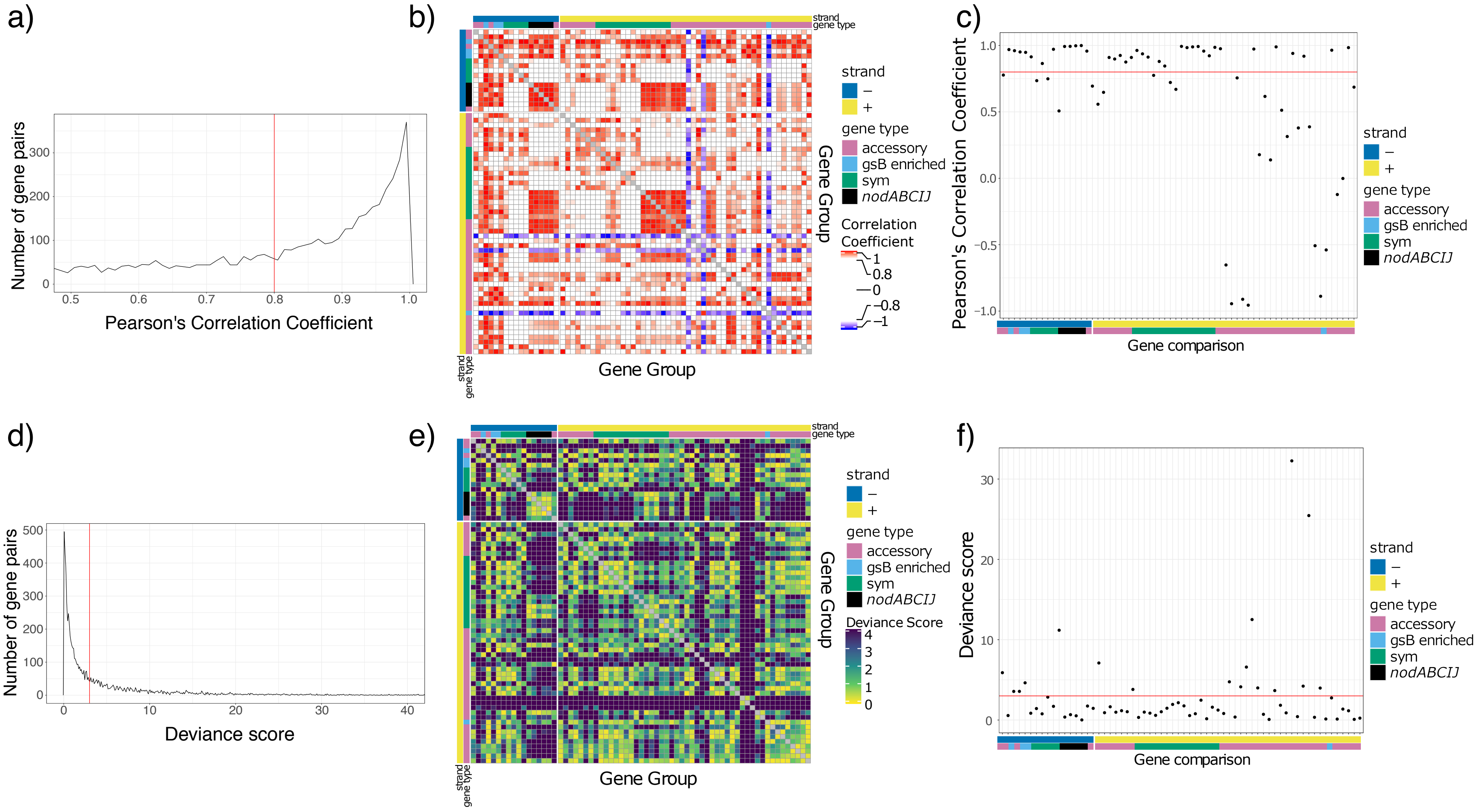


Figure 5.2 Validation of gene expression thresholds to determine genospecies transcriptional units. (Continued on following page).

Figure 5.2 continued. Validation of gene expression thresholds to determine genospecies transcriptional units. Panels **b,e,c**, and **f** are based on genes within a region of the symbiosis plasmid, with genes arranged by their genome order in gsB strain, SM3. **a)** Distribution of adjacent gene pairs gene expression Pearson's correlation coefficients for gsB. Red line indicates a gene expression correlation coefficient of 0.8. **b)** gsB Pearson's correlation coefficients for gene expression correlations of all pairwise gene combinations. Gene expression correlation coefficients are shown on a scale ranging from positive (red) to negative (blue). **c)** Pearson's correlation coefficients for gsB adjacent genes expression correlation. Red line indicates a Pearson's correlation coefficient of 0.8. **d)** Distribution of adjacent gene pair deviance scores for gsB. Red line indicates a deviance score of 3. **e)** gsB deviance scores for all pairwise gene combinations within a region of the SM3 symbiosis plasmid. Deviance scores are shown on a scale ranging from 0 (yellow) to above 3 (dark blue). **f)** Deviance scores for gsB adjacent genes. Red line indicates a deviance score of 3. All gene expression correlation coefficients and deviance scores are calculated from gsB genomic and transcriptomic data.

5.4.2. Differences in transcriptional units are evident between genospecies

In order to evaluate the differences in transcriptional unit generation between genospecies, the number of genes assigned to transcriptional units for each genospecies was calculated. Percentage of genes assigned to a transcriptional unit ranged between 42.17%-48.19% across genospecies pangenomes, with the exception of gsD for which only 34.24% of genes were assigned to a transcriptional unit (Table 5.2; Figure 5.3a). gsD's lower percentage could be because on average adjacent genes across gsD strains displayed lower levels of co-expression with lower correlation coefficients (gsD Pearson's correlation R statistic average: 0.33; genospecies range: 0.40-0.47) and higher expression deviance scores (gsD deviance score average: 5.30; genospecies range: 4.53-5.03) compared to other genospecies. The distribution of transcriptional unit size was found to be similar across all genospecies, with the majority of transcriptional units containing only 2 genes (Figure 5.3c; Table 5.3). For this study, only polycistronic operons were considered, and did not analyse the abundance of single gene operons. Therefore, the average lengths of transcriptional units for the genospecies groups totalled between 2-3 genes (Table 5.3).

Transcriptional units were predominantly comprised of core genes, and this trend was shown across all genospecies (Figure 5.3b; Table 5.4). This is because many accessory genes are filtered out of transcriptional unit assignment as they are present in fewer than 3 strains in a genospecies. For all genospecies, transcriptional units containing predominantly genospecies enriched genes were also identified (genes that are highly abundant in one genospecies and highly absent from all other genospecies) (Figure 5.3; Appendix Figure D.4-Appendix Figure D.7). gsB was found to have the most genospecies enriched genes included in transcriptional units (Figure 5.3b; Table 5.4). For example, in gsB a genospecies enriched gene operon contained five genes, of which four matched BLAST hits to 'efflux transporter, RND family, MFP subunit', 'putative ABC transporter permease', 'ABC transporter' and 'ligand-binding protein SH3' functions (Figure 5.4). This suggests that gsB strains have acquired a bacterial efflux pump system commonly used for active transport of antibiotics, heavy metals, or nodulation factors, that is rare in other genospecies.

Only 94 transcriptional units were shared across all genospecies groups (Table 5.5). These 94 transcriptional units were comprised of 224 genes, of which 191 were core, 26 were accessory and 7 were symbiosis genes (Additional File 6: Table S4). Most accessory genes formed operons with other core genes. Four operons (two genes in length) were found to constitute only accessory genes, however these accessory genes were present in a high frequency of strains (179-195 out of the 196 *Rlt* strains collection). Additionally, the only symbiosis gene operons conserved across all genospecies were the *nodABCIJ* and *nodEF* operons. As expected, the *nodABCIJ* operon was identified as a transcriptional unit in all genospecies based on the chosen parameters, due to the operon being used initially to assess the suitability of the transcriptional unit threshold parameters (Figure 5.4; Appendix Figure D.4-Appendix Figure D.7). *nodMN* are present in all genospecies except for gsE, where a deviance score of 3.1 narrowly rejects them from transcriptional unit acceptance. Nitrogen fixation (*nif* and *fix*) genes, which are used for nitrogen fixation when rhizobia are in their nodule bacteroid physiology, were hypothesised to not be identified effectively by the operon prediction strategy because under the TY broth-clover flavonoid growth conditions these genes were expected to be expressed only at extremely low levels, if at all. As a result, the nitrogen fixation genes were found to be less effectively identified as conserved transcriptional units. For the *fixABCX* operon, *fixA* and *fixX* were not included in the transcriptional units across some genospecies due to a correlation coefficient < 0.8 or a deviance score just above 3. Therefore, the altered content of the *fixABCX* operon across genospecies classes them as different transcriptional units and therefore would be considered as not conserved across genospecies. Similarly, omission of *nifH*, *nifK* and *nifD* genes from the *nifHDKEN* operon due to correlation and deviance scores outside parameter thresholds meant these transcriptional units were also considered un-conserved across genospecies. gsC was found to assign the most symbiosis genes to transcriptional units (22 of 25 symbiosis genes), whereas gsD assigned only 15 of 25 symbiosis genes to a transcriptional unit. Taken together, it was challenging to adequately identify operons consistently between strains for genes whose expression rely on specific environmental conditions.

The largest cross-species conserved operon contained 10 ribosomal protein genes (Appendix Figure D.8). The main functions of the 94 operons followed a similar pattern of essential cellular metabolic functions with 'putative ribosomal protein', 'transporter/efflux pump' and 'ATPase' functions. gsB and gsC shared the most transcriptional units, whereas gsA and gsD shared the fewest transcriptional units (Table 5.5), which could suggest that gsA and gsD had the most differences in regulatory network structures. Gene order was also identified to be highly conserved in transcriptional units shared across genospecies, as when gene order within transcriptional units was not considered, the number of shared transcriptional units remained almost exactly the same (Table 5.5).

Conversely, there were cases where a transcriptional unit was identified in two or more genospecies but the genes within the unit differed between genospecies by the presence or absence of a single gene. Based on the developed pipeline, these would be considered as separate transcriptional units as they differed in overall gene content. For example, 'Transcriptional unit A' in gsB consists of 9 genes encoding the components of a Type IV conjugal transfer system (Figure 5.4a). 'Transcriptional unit A' has two different end genes (group5766 or group7857) because the gsB strains contained either group5766 or group7857 (Figure 5.4a). This disparity between strains was due to both genes encoding the same 'conjugal transfer protein TrbI' function but being identified as different orthologous gene groups by ProteinOrtho. gsC also had the exact same operon but without the end group5766/group7857 genes. This is because in gsC, group5766 was not found adjacent to group5767 but instead adjacent to another set of genes which encoded for a different putative type IV conjugal transfer system. Additionally, group7857 was excluded from gsC 'Transcriptional unit A' because it had a deviance score of 3.25 with its adjacent gene group5767, which consequently narrowly excluded it from the operon (Figure 5.4a). As such, in this analysis the gsB and gsC transcriptional units were classified as different transcriptional units based on the overall difference in gene content. Therefore, minor alterations to the regulatory structure of the operons must be considered when comparing the similarity of putative polycistronic regions between genospecies.

Alterations to the structure of transcriptional units within a genospecies must also be considered when generating transcriptional units from a consensus of adjacent gene pairs from multiple strains. Infrequently, some transcriptional units produced for a genospecies were not linear (Figure 5.4). This was found to be commonly caused by some strains containing one version of the transcriptional unit, and other strains containing an altered version of the transcriptional unit where a different gene ortholog group had been replaced with another gene with a different ortholog group name in the middle of the transcriptional unit. Non-linear transcriptional units can occur by both 'redundant' gene versions being present in at least 3 strains in the genospecies and also being located in the same position within the unit in at least 1 strain. In most cases, the 'redundantly' located genes in the transcriptional unit were again found to be homologs with the same putative function but labelled under different ortholog gene names by ProteinOrtho.

Table 5.2 The number of genes for consideration of transcriptional unit generation before and after operon filtering parameters. Before filtering, gene pairs must be present in at least 3 strains of a genospecies and located adjacently in at least 1 strain. Filtering parameters include intergenic distance < 200 bp, Pearson's correlation coefficient > 0.8, deviance score < 3.

Genospecies	Number of genes before filtering	Number of genes after filtering (%)	Number of unique adjacent gene pairs before filtering	Number of unique adjacent gene pairs after filtering (%)	Number of adjacent gene pairs with negative intergenic distances before filtering (%)
A	6530	2754 (42.17)	6876	1743 (25.35)	844 (12.27)
B	6936	3097 (44.65)	7103	1976 (27.82)	961 (13.53)
C	7634	3392 (44.43)	8198	2212 (26.98)	982 (11.98)
D	6988	2393 (34.24)	7191	1437 (19.98)	913 (12.70)
E	6172	2974 (48.19)	6067	1911 (31.50)	799 (13.17)

Table 5.3 The number of transcriptional units with a specific number of genes. The total number of transcriptional units identified for each genospecies is noted. Additionally, the number of transcriptional units conserved within WGCNA modules (i.e. not split across several WGCNA modules) is noted. *Only core genes were considered for the conservation calculation, and therefore the number of transcriptional units considered are only those that contain 2 or more core genes.

Genospecies	**Number of transcriptional units**	**Average gene number per transcriptional unit**	**Transcriptional units maintained by WGCNA modules***	**Count of transcriptional units with gene number of n**													
				2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	1012	2.73	306/616 (49.66%)	631	211	76	48	24	12	6	2	1	1	-	-	-	-
B	1122	2.77	342/639 (53.52%)	680	239	101	55	22	11	3	6	3	1	-	1	-	-
C	1184	2.86	310/602 (51.49%)	719	230	94	67	32	13	11	10	3	1	2	1	1	-
D	956	2.49	292/542 (53.87%)	684	172	55	22	10	2	3	3	2	2	1	-	-	-
E	1063	2.81	310/710 (43.66%)	640	220	99	46	31	15	2	6	2	-	1	-	-	1

Table 5.4 The total number of genes in transcriptional units that are classified as core, genospecies enriched, accessory and symbiosis genes across all five genospecies A-E. Gene group classification is based on gene frequency in 196 *Rlt* strains.

Genospecies	Total no. of genes	No. core genes (%)	No. genospecies enriched genes (%)	No. accessory genes (%)	No. symbiosis genes (%)
A	2754	1730 (62.82)	83 (3.01)	923 (33.51)	18 (0.65)
B	3097	1782 (57.54)	158 (5.10)	1138 (36.75)	19 (0.61)
C	3392	1757 (51.80)	73 (2.15)	1540 (45.40)	22 (0.65)
D	2393	1448 (60.51)	27 (1.13)	903 (37.73)	15 (0.63)
E	2974	2021 (67.96)	44 (1.48)	893 (30.02)	16 (0.54)

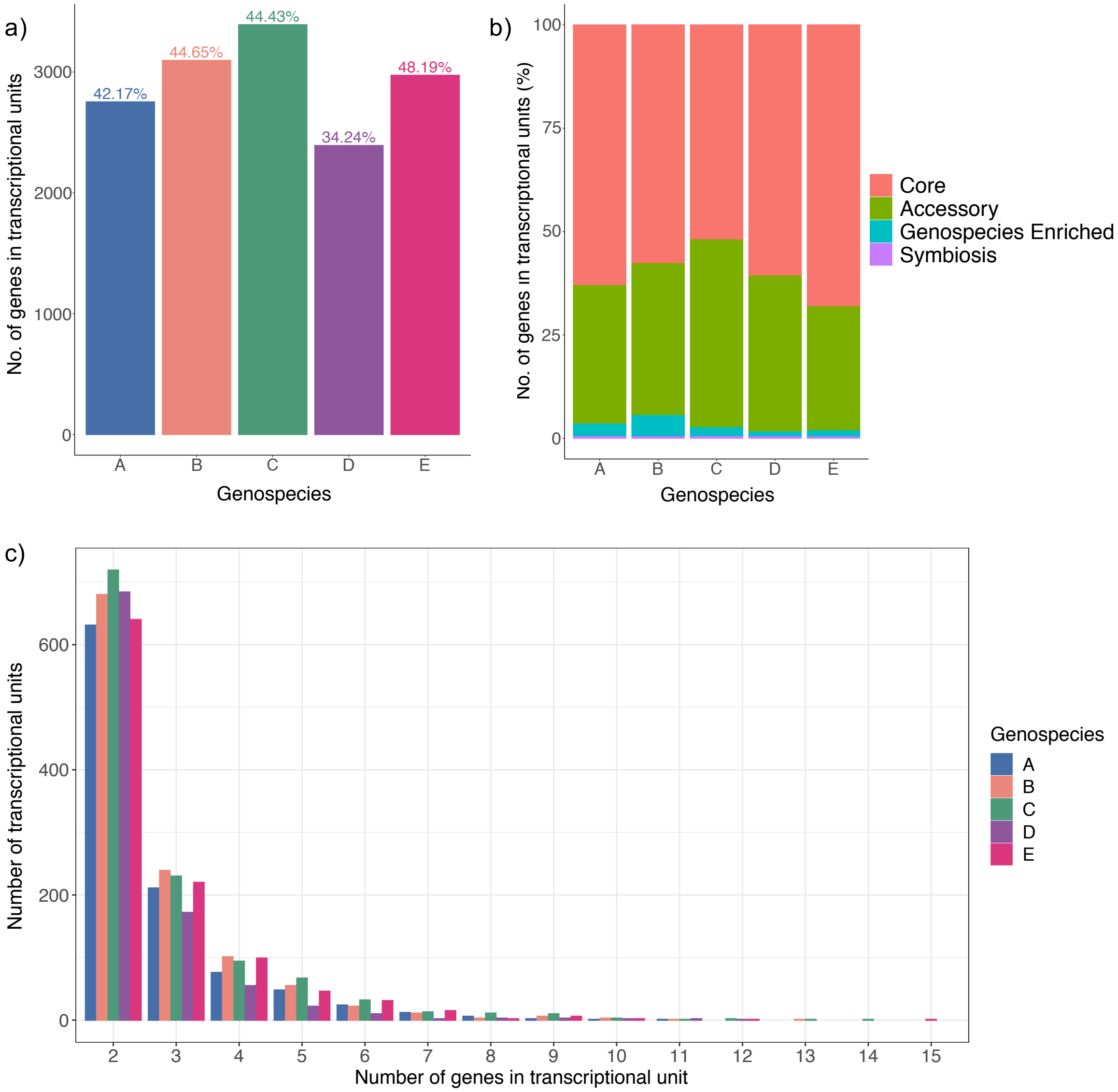


Figure 5.3 Transcriptional unit abundance and size across genospecies. **a)** Number of genes in transcriptional units for each genospecies A-E. Percentages of genes in transcriptional units are displayed. **b)** Percentage of genes in transcriptional units for each genospecies that are core, accessory, genospecies enriched or symbiosis genes. Gene types are classified based on their frequency in 196 *Rlt* strains. **c)** The number of transcriptional units made up of n number of genes across all 5 genospecies.

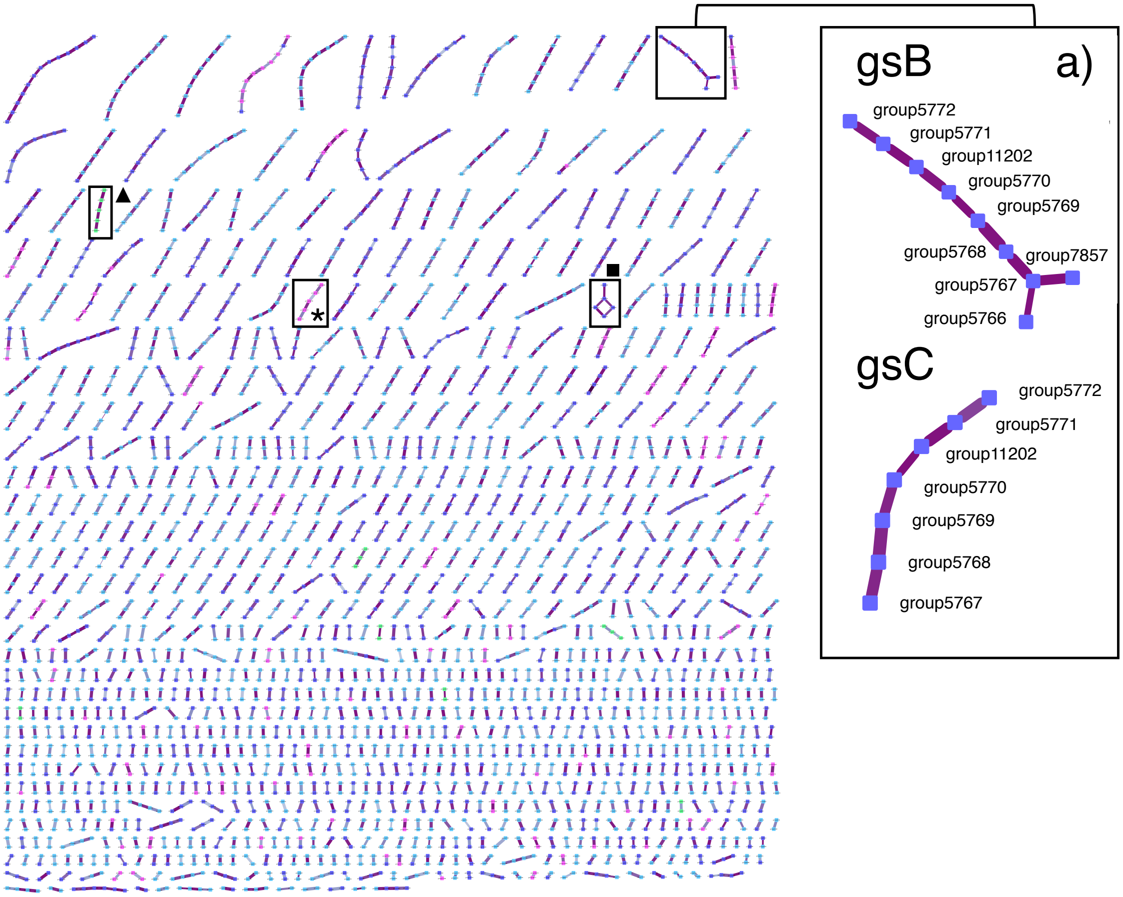


Figure 5.4 gsB transcriptional units generated using the following filtering parameters: R correlation > 0.8, deviance < 3, intergenic distance < 200 bp, must be adjacent gene pair in at least 3 strains. Nodes are genes colour coded by: Blue = core, Purple = accessory, Pink = genospecies enriched, Green = symbiosis. Edge colour increases from blue to purple with increased gene expression correlation between adjacent pairs. Edge thickness increases with decreasing deviance score. **a)** Shows the comparison of 'Transcriptional unit A' gene content between genospecies B and genospecies C. ***** indicates a unit which contains only genospecies enriched genes and has functional associations to an efflux pump system. ▲ Indicates the *nodABCIJ* nodulation gene operon. ■ Indicates a non-linear transcriptional unit.

Table 5.5 The number of transcriptional units shared across genospecies groups.

Comparisons	Number of shared transcriptional units (gene order conserved)	Number of additional shared transcriptional units (gene order not conserved)
AB	368	0
AC	372	1
AD	297	0
AE	352	0
BC	386	0
BD	315	0
BE	335	0
CD	302	0
CE	368	2
DE	299	0
All species	94	0
BCDE	108	0
ACDE	123	0
ABDE	120	0
ABCE	142	0
ABCD	129	0

5.4.3. Other known operons are identifiable as transcriptional units

To further evaluate the genospecies operon predictions, other known rhizobia operons were identified from the generated transcriptional units. Orthologous gene groups were identified that matched the genes from known operons using a BLAST search of the *Rlt* genomes.

The rhizosphere induced operon (*rhiABC*) was identified in gsC, which had previously only been found in *Rhizobium leguminosarum* symbiovar *viciae* strains (Ramachandran *et al.*, 2011) (Appendix Figure D.5). The *cycHJKL* operon was also searched for, which is involved in iron acquisition for maturation of cytochrome c and mutations in this operon can result in loss of nitrogen fixing ability in rhizobia (Yeoman *et al.*, 1997). *cycHJKL* was present as a transcriptional unit in genospecies B, C and D but excluding the *cycL* gene. This was because the *cycL* gene did not meet operon prediction threshold parameters (gsB: correlation = 0.17, deviance = 7.0; gsC: correlation = 0.51, deviance = 7.4; gsD: correlation = 0.67, deviance = 7.2). In gsA, *cycJK* was identified as a two gene operon, but *cycH* and *cycL* were excluded from the operon due to co-expression parameters not being met despite displaying negative intergenic distances (gsA *cycH*-*cycJ* correlation = 0.54, deviance = 11.50; gsA *cycK*-*cycL* correlation = 0.95, deviance = 6.32). The correlation coefficients and deviance scores seemed to indicate the *cycL* was not transcribed in the same way as the other genes. However, the intergenic distance for *cycL* is highly conserved between all strains regardless of genospecies and has a consistent negative base pair distance with *cycK* of -4 bp, which would indicate that *cycL* is part of the operon. Further investigation identified that co-expression correlation coefficients and deviance scores were found to tail off at the end of many transcriptional units (Figure 5.2c, Figure 5.2f). A staircase-like expression behaviour of genes within operons could explain the reduced correlation and increased deviance scores towards the ends of operons, as transcription is often higher at the 5' end of operons with transcription efficiency reducing towards the 3' end resulting in incomplete mRNA transcription (Güell *et al.*, 2009; Maier *et al.*, 2011; Schmidt *et al.*, 2011; Arike *et al.*, 2012). Consequently, the threshold for operon prediction in the samples might have been too conservative. However, stringent parameter values also enabled confident identification of putative operons with highly conserved expression patterns across genospecies.

In addition, transcriptional units were also validated by confirming whether they were maintained within WGCNA modules of co-expressed core genes that were generated previously (Chapter 4). Calculation of WGCNA modules did not consider intergenic distances or genospecies strain-grouping, and instead relied on converting expression correlations of all genes into a signed and weighted expression network for hierarchical clustering based on expression across all *Rlt* strains. 43.66-53.87% of transcriptional units were maintained when intersected by WGCNA modules (Table 5.3). gsD had the highest percentage of conserved transcriptional units in *Rlt* WGCNA core gene modules, whereas gsE had the lowest percentage of maintained transcriptional units (Table 5.3; Additional File 6: Table S5). Similarly, several transcriptional units were identified within the grey WGCNA group (containing genes which were not assigned to a WGCNA module), but these transcriptional units contained less than 3 genes, and therefore would have been removed due to the requirement of 3 genes minimum per WGCNA module. Only one transcriptional unit containing 3 genes encoding a tripartite tricarboxylate membrane transporter was found in the grey WGCNA group, and it is speculated this was identified because it was only classed as a transcriptional unit in gsB and gsE, suggesting its co-expression is not tightly conserved across all genospecies. Together, this suggests that while known operons could be identified as transcriptional units within this study, choosing parameters is also challenging as operon expression can substantially vary across strains which can potentially lead to undetected or truncated transcriptional units.

5.5.Discussion

Genome annotation and single replicate transcriptome data from multiple strains was utilised to build generalised 'operomes' for five *Rlt* genospecies (Additional File 6: Table S6). Approximately 1000 transcriptional units were identified for each genospecies individually, and 94 of these transcriptional units were conserved across all five genospecies A-E. A combination of ortholog identification, intergenic distance measures, gene expression correlation and expression deviance were used to generate transcriptional units equating to putative operons. Expression deviance also provided an additional supporting metric for operon generation. Furthermore, the genomic and transcriptomic traits of the *Rlt* nodulation *nodABCIJ* operon were used as a control to validate the appropriateness of the chosen operon prediction parameters. This method exploits the variation in genomic architecture and expression levels across multiple strains grown under the same conditions to determine species conserved operons. Overall, the use of multiple different strains to characterise *Rlt* species and genospecies transcriptional units also highlighted that substantial variation in the expression of putative operons is evident across strains within the same species.

5.5.1. Optimisation of genetic and transcriptomic parameters using multiple strains

For this study, intergenic distance distributions suggested that most adjacent genes had an intergenic distance <200 bp (Figure 5.1a), and subsequently gene pairs with an intergenic distance below this threshold were chosen as potential transcriptional unit gene pairs for further evaluation. Additionally, these intergenic distance distributions were also in agreement with distributions identified in previous studies (Ermolaeva, White and Salzberg, 2001; De Hoon *et al.*, 2004; Dam *et al.*, 2007; Wang, MacKenzie and White, 2015). Intergenic distance has been suggested to be the most reliable indicator for operon prediction and is used as a parameter in nearly all operon identification studies (Dam *et al.*, 2007). However, this pipeline's intergenic distance calculations differ from previous studies that commonly just use intergenic distance measures determined from a single strain. For analysis of intergenic distance, the average distance was calculated between orthologous gene pairs when they were adjacent in at least 1 strain in a genospecies, but the genes must be present in a minimum of three genomes in order to evaluate co-expression across strains. Therefore, the number of replicates for each adjacent gene intergenic distance calculation is not necessarily the same each time, especially when accessory genes are considered. This enabled identification of genospecies transcriptional units which have more flexible gene arrangements across strains.

A marginally bimodal distribution was identified for intergenic distances between adjacent genes, with a major sharp peak under 40 bp and another smaller shallow peak at around 70-100 bp. This intergenic distance distribution is also observed in genomes of other bacterial species (Salgado *et al.*, 2000; Ermolaeva, White and Salzberg, 2001; De Hoon *et al.*, 2004; Price, Arkin and Alm, 2006; Okuda *et al.*, 2007). The smaller second intergenic distance peak has been suggested to represent the intergenic distances of genes at the borders of transcriptional units in *E. coli* (Salgado *et al.*, 2000). On the other hand, these intergenic distances could also be generated from operons that are highly, but inconsistently, expressed because they have wider intergenic regions than other operons (Eyre-Walker, 1995; Price, Arkin and Alm, 2006). Genome-specific distance models have also shown that operon structures in different bacterial species can differ from the model *E. coli* operon structures (Price, Arkin and Alm, 2006). For example, the maximum accepted intergenic distance between operon-genes varies between studies, ranging from 20bp to 300bp (Salgado *et al.*, 2000; Ermolaeva, White and Salzberg, 2001; De Hoon *et al.*, 2004; Wang *et al.*, 2004; Price, Arkin and Alm, 2006). Predominantly, genes with intergenic distances greater than 200 bp have been considered to not be in the same operon (Ermolaeva, White and Salzberg, 2001; Wang, MacKenzie and White, 2015). The variation in accepted intergenic distance length is understandable as it cannot be assumed that the genomic architecture of all bacterial genomes, and between all operons within the same genome, are consistent (Wang *et al.*, 2004; Edwards *et al.*, 2005; Price, Arkin and Alm, 2006). For these reasons, and due to this analysis calculating the average intergenic distance from multiple genomes, a more relaxed intergenic distance threshold was opted for compared to other studies to account for the genomic variability between strains of the same genospecies (De Hoon *et al.*, 2004; Price *et al.*, 2005; Brinza *et al.*, 2010).

In addition to intergenic distance between adjacent genes, gene pairs were specified to have an expression correlation coefficient > 0.8 and an expression deviance < 3 (Figure 5.2). Expression correlation thresholds were based on the distribution of adjacent gene pair expression correlation coefficients (Figure 5.2a-c) with consideration to the thresholds used in previous operon prediction studies (Dam *et al.*, 2007; ten Broeke-Smits *et al.*, 2010). Using a deviance score further enabled evaluation of transcriptional co-expression between genes, and the threshold was determined based on distribution of deviance scores across adjacent gene groups with additional consideration to the deviance scores calculated for the known symbiosis nodulation gene operon, *nodABCIJ* (Figure 5.2d-f). The distribution of adjacent gene pairs expression correlation coefficients was observed to be similar to those in previous studies using *E. coli* and *B. subtillis* (Okuda *et al.*, 2007). Expression parameters were quite stringent and were chosen in order to confidently identify only the gene pairs which are highly likely to be in operons, as comparing expression across multiple genomes can incur substantial noise. However, identifying putative operons using only expression data also can incur false-positive operons. This is because adjacent genes can be considered co-expressed by having common co-regulation but still be in separate operons (Westover *et al.*, 2005; Dam *et al.*, 2007). Similarly, genes within the same operon can be differentially expressed if there are multiple dynamically expressed transcriptional units within the operon that are dependent on certain environmental signals (Oliver *et al.*, 2009). Therefore, it is vital for operon prediction to be determined based on both genomic and transcriptomic information between adjacent gene pairs from multiple strains in order to fully consider the flexibility of operon structures across species.

For each genospecies, 42.17-48.19% of orthologous genes were assigned into a transcriptional unit (except gsD at 34.24%; Table 5.2; Figure 5.3a), which is lower than the 62% of genes reported for *Staphylococcus aureus* (ten Broeke-Smits *et al.*, 2010). However, the percentage calculated for *S. aureus* was based on the genes in one strain, with transcriptomic data collected from several growth time-points, whereas the data for this pipeline was collected from multiple strains grown under the same conditions (ten Broeke-Smits *et al.*, 2010). Additionally, all *Rlt* genospecies produced transcriptional units with an average size of 2-3 genes (Table 5.3; Figure 5.3c). The average number of genes per transcriptional unit were similar to the operon size distributions of other species (2-4 genes average per operon), but perhaps on the slightly smaller side compared to *Bacillus* (4.1 average) and *E. coli* (3 – 3.5 average) (Itoh *et al.*, 1999; Zheng *et al.*, 2002; Koonin, 2009; ten Broeke-Smits *et al.*, 2010). Increased average length of operons has been associated with genomes displaying high modularity in the genomic organisation of their biochemical pathways, as observed with *E. coli*, *B. subtilis* and *Buchnera* (Zheng *et al.*, 2002). Genomes with smaller average operon lengths have been suggested to have undergone more frequent translocation (Zheng *et al.*, 2002). This could be the reason the genospecies have such a small average operon size, because operon prediction was based on gene-pair presence across at least three genomes. However, the requirement for the conservation of gene pairs across genomes in this study also likely biases selection towards 2-gene operons, which in previous cross-species operon predictions between *Haemophilus influenzae* and *Escherichia coli* were most highly conserved between bacterial species (Tamames *et al.*, 1997).

5.5.2. Species-conserved putative operons contain core and accessory genes

Transcriptional units were identified that contained purely core genes, purely accessory genes, and those containing a mixture of core and accessory genes (Figure 5.4; Appendix Figure D.4-Appendix Figure D.7; Table 5.4). Fewer accessory genes were found to be incorporated into operons compared to core genes (Table 5.4). This could be due to the large number of years required for horizontally transferred genes, such as some accessory genes, to be incorporated into the strain regulatory network (Lercher and Pal, 2008). Similarly, it is speculated that the reduced frequency of accessory genes, their less-essential functional associations, and reduced expression levels slows their integration into the species regulatory network (Galardini *et al.*, 2015). However, 94 operons were found to be conserved across all genospecies, some of which contained both core and accessory gene components (Additional File 6: Table S4). The functions of the cross-species conserved operons were predominantly associated with vital cellular functioning mechanisms such as ribosomal protein assembly and various cellular transport mechanisms, which explains their high conservation across *Rlt* genospecies. The majority of these mixed core-accessory gene operons constituted 2-3 genes. Therefore, introgression of accessory genes into a core operon seems to have enabled regulation of accessory genes by core genome regulatory components in some cases (Galardini *et al.*, 2015).

Additionally, operons containing purely genospecies enriched genes (i.e. genes that are highly present in one genospecies and highly absent in all others) were identified for all genospecies. gsB was found to have the most operons containing only genospecies enriched genes and this is likely because gsB is the most genetically homogeneous genospecies from the 196 *Rlt* strain dataset (Cavassim *et al.*, 2019). For example, in gsB a genospecies enriched gene operon containing five genes was identified that constituted the components of an efflux pump system (Figure 5.4). This suggests that gsB strains have acquired a bacterial efflux pump system potentially for active transport of antibiotics, heavy metals, or nodulation factors (Nikaido, 2018). Additionally, the known rhizosphere-induced operon, *rhiABC*, was identified only in gsC (Appendix Figure D.5) (Cubo *et al.*, 1992; Rodelas *et al.*, 1999). These transcriptional units containing genospecies-enriched genes could be used to identify specific functional attributes that are highly associated to a particular genospecies, and which may provide some competitive advantage in the soil rhizosphere over other genospecies.

Out of the symbiosis nodulation gene operons, only *nodABCIJ* and *nodEF* were conserved across all genospecies as transcriptional units (Hong, Burn and Johnston, 1987a). Nodulation gene transcriptional units were expected and observed to be the most easily identified symbiosis gene groups, because strains were grown in Tryptone Yeast (TY) broth with 1 µM 7,4’-dihydroxyflavone (clover flavonoid) which activates the NodD transcriptional activator of *nod* genes (Djordjevic *et al.*, 1987). However, *nif* and *fix* gene transcriptional units were not conserved across all genospecies, and this is likely because *nif* and *fix* genes are only activated and consistently expressed in clover nodules when strains differentiate into their bacteroid form (Herman P. Spaink *et al.*, 1987). Therefore, variable and low *nif* and *fix* gene expression across strains contributed to their lack of transcriptional unit conservation across genospecies.

Taken together, it is unlikely that the majority of *Rlt* transcriptional units have been identified. This is because rare accessory gene operons will have been prone to removal due to their low frequency across strains excluding them from the analysis. Nevertheless, utilising multiple genomes has been able to identify conserved transcriptional units that do contain accessory genes which were expressed consistently across strains under particular environmental conditions. Environmental differences are also important for identifying core transcriptional units that are expressed under specific conditions. However, these transcriptional units may have not been detected in this study due to the environmental conditions being inadequate to induce their expression.

5.5.3. Study limitations and future research

Parameters included in other studies, which were not considered in this study, are predominantly based on using large confirmed-operon reference databases for *E. coli* or *B. subtilis* (De Hoon *et al.*, 2004; Fortino *et al.*, 2014). For example, optimal operon length parameters were determined for *B. subtilis* and *E. coli* studies based on the distributions from previously curated operon datasets (De Hoon *et al.*, 2004). The use of reference operon databases, such as RegulonDB (Fortino *et al.*, 2014), has also enabled use of Bayesian models to predict operons on other strains (Brinza *et al.*, 2010; Chen *et al.*, 2017). However, operon predictors trained on *E. coli* and *B. subtilis* do not necessarily apply well to other genomes, which are known to be largely diverse in genomic structure and gene content (Wolf *et al.*, 2001; Romero and Karp, 2004; Dam *et al.*, 2007; Koonin, 2009; Osbourn and Field, 2009). It would therefore be insightful to test whether this pipeline can identify operons for other bacterial species where alternative known operons can be used to validate chosen parameter thresholds.

Another previously used determinant of operon prediction that was not used is the identification of Transcriptional Start and Terminator Sites (TSS and TTS) (Brinza *et al.*, 2010; Wang, MacKenzie and White, 2015; Chen *et al.*, 2017; Slager, Aprianto and Veening, 2018). However, studies that searched for TSS and TTS only utilised one strain genome under varying environmental conditions for operon prediction (with the exception Brinza *et al.*'s (2010) use of RegulonDB), whereas in this study three to seven strains grown under the same conditions were used for each genospecies. Future investigations could therefore try to identify conserved TSS and TTS's across genomes and would be a useful addition to operon identification.

Additionally, the operon prediction pipeline used in this study is especially dependent on conservation of gene order because genomic information is utilised from multiple strains to predict genospecies-conserved operons. Gene order conservation across multiple genomes (either from recent vertical or horizontal transmission) increases the probability of gene pairs being part of an operon (Ermolaeva, White and Salzberg, 2001; Tamames, 2001; Wolf *et al.*, 2001; Edwards *et al.*, 2005; Junier and Rivoire, 2016). For the operon prediction pipeline in this study, this suggests that many of the transcriptional units are likely to be real, as gene pairs must be present across at least three strains (although only required to be adjacent in at least 1 strain) to be included in the analysis. At short phylogenetic distances, gene order is more conserved due to recent divergence. Therefore, this criterion worked well for this study, which contained strains within the same sub-species of *Rlt* (Tamames *et al.*, 1997; Tamames, 2001). However, the shortcoming of relying on gene order for operon prediction is that operons with reordered but conserved genes are not recognised (Itoh *et al.*, 1999; Wolf *et al.*, 2001). For example, many of the 94 cross-genospecies conserved transcriptional units have a transporter-associated or protein-subunit function whereby conservation of gene order within the operon is important for function (Additional File 6: Table S4). However, when gene order within transcriptional units is not considered the number of shared transcriptional units between genospecies was found to remain almost exactly the same (Table 5.5). This highlighted the strong conservation of gene order within the transcriptional units in this investigation. Operons are also not identified when genes are appended onto the end of an existing operon (Price, Arkin and Alm, 2006), which for this operon prediction pipeline classifies them as different operons (Figure 5.4a).

If gene pairs had a negative intergenic distance or less than 200 bp distance, but a correlation coefficient below 0.8 and deviance more than 3, then the gene pair would not be considered a transcriptional unit. This is regardless of the fact negative and small intergenic distances would suggest gene pairs are highly likely to part of the same operon (Salgado *et al.*, 2000; De Hoon *et al.*, 2004; Price, Arkin and Alm, 2006). Due to this restrictive criteria, gene pairs that pass all criteria are labelled as a 'transcriptional unit' instead of operon, because even if the gene pairs are genomically close enough to be considered an operon, if no evidence of strong co-expression is provided the gene pair will be excluded. Further development of the operon prediction could include acceptance of some gene pairs as transcriptional unit pairs based on intergenic distance alone, regardless of correlation and deviance scores, if the distance is acceptably small enough that the genes are almost guaranteed to be in the same operon (such as overlapping gene pairs).

The functional relationship between genes was also not considered (Salgado *et al.*, 2000; Taboada *et al.*, 2018). Then again, this criterion maybe more of a limitation than an asset because previous studies have already shown that it is not necessary for genes within an operon to have a related function (Osbourn and Field, 2009; Fortino *et al.*, 2014).

5.5.4. Conclusions

This study provides a new resource of putative transcriptional units in *Rlt,* identified across five genospecies using 26 strains. Additionally, calculating the expression deviance between genes was shown to provide an additional effective metric to determining operons. Parameters for operon detection can be determined from known operons within the species or from looking at the distribution of parameters across genomic regions, as this study has shown. Identifying operons will further aid identification of conserved of regulatory networks and gene order within *Rlt* species and can be used for future functional association studies.

Chapter 6. Competitive rhizobial intraspecies interactions are genospecies specific

6.1.Abstract

Background: Symbiotic interactions between rhizobia bacteria and legume plants, are vital for ecosystem functioning and soil nutrient balance. To form symbiosis with the plant, rhizobia first need to compete for nutrients and space with other symbiont strains in the plant rhizosphere. This study investigated competition between rhizobia genotypes, with the aim to understand to what extent facilitative, inhibitory or neutral intraspecies interactions exist between and within different genospecies and what are the potential underlying mechanisms.

Results: Twenty-four genetically diverse *Rhizobium leguminosarum* sv. *trifolii* (*Rlt*) strains were selected from 3 genospecies (< 95% average nucleotide identity) and pairwise competitive interactions were determined *in vitro*: 1) indirectly, mediated via secreted compounds in cell-free supernatants; and 2) directly, mediated by growth inhibition in the same environment. Significant facilitative and inhibitory interactions were observed through indirect competition. One strong inhibitory genospecies level interaction was detected where genospecies E strains consistently suppressed the growth of genospecies A strains. On average, genospecies E also displayed facilitated growth in other genospecies' supernatants. However, indirect interactions mostly varied largely at the genotype level, and in general, strains that produced more inhibitory supernatants were likely to grow better in supernatants of other strains. This indicated a positive trait correlation between the inhibitory capacity of a strain, and its resistance to inhibition by other strains. Clear genospecies effects were observed also when in direct competition. Overall, genospecies A demonstrated a high susceptibility to direct inhibition, while genospecies E were the most inhibitory strains. Mechanistically, increased genospecies A susceptibility was associated with potential regulatory effects of multiple quorum sensing pathways, while genospecies E strains exclusively contained a Vicibactin siderophore gene cluster and displayed the relatively highest metabolic capacity that may have increased its competitive ability to sequester and deplete resources.

Conclusions: Together, these results demonstrate that *Rlt* shows high intraspecies phenotypic diversity which is linked to variation in resource and interference competition. The outcome of competitive interactions within rhizobial communities could thus affect which strains get to establish symbiosis with the plant.

6.2.Introduction

Rhizobia are bacterial symbionts capable of providing accessible nitrogen to legumes in return for carbon and can be exploited agriculturally to increase crop yield by improving soil nutrient balance (Lupwayi, Clayton and Rice, 2006; Mishra *et al.*, 2013). Natural rhizobia soil populations are very diverse (Kumar *et al.*, 2015), and this diversity could be driven for example by inter- and intra-plant species variation (Kiers and Denison, 2008; Miranda-Sánchez, Rivera and Vinuesa, 2016; Kroll, Agler and Kemen, 2017; Vuong, Thrall and Barrett, 2017; Clúa *et al.*, 2018), by abiotic soil factors (Rice, Penney and Nyborg, 1977; Harrison, Jones and Young, 1989; Xiong *et al.*, 2017; Igiehon and Babalola, 2018; Liu *et al.*, 2019), or by agricultural management practices (Kiers, West and Denison, 2002; Lupwayi, Clayton and Rice, 2006; Shu *et al.*, 2012; Weese *et al.*, 2015). Moreover, variation in rhizobial diversity can be driven by competition with other species of soil bacteria (Pugashetti, Angle and Wagner, 1982; Villacieros *et al.*, 2003; Hibbing *et al.*, 2010; Teng *et al.*, 2015; Lu *et al.*, 2017) and other rhizobia strains for plant nodulation (Denison and Kiers, 2004; Kiers and Denison, 2008; Blanco, Sicardi and Frioni, 2010; Wielbo *et al.*, 2011; Barrett *et al.*, 2015). For example, rhizobia inoculant success has previously been shown to be limited by competition with native soil rhizobia, which are often able to outcompete inoculant strains for nodule occupancy (Berg *et al.*, 1988; Triplett and Sadowsky, 1992; Blanco, Sicardi and Frioni, 2010). Understanding the role of intraspecies rhizobial diversity is thus important as it could affect productivity of the legume-rhizobia symbiosis by determining which strains get to form the symbiosis with the plant (Barrett *et al.*, 2015; Pahua *et al.*, 2018; Liu *et al.*, 2019).

Intraspecific competition between rhizobia strains could be mediated via two main mechanisms. Firstly, strains can compete in an exploitative (indirect) manner (Checcucci *et al.*, 2017). Indirect competitive interactions, such as when a strain more effectively metabolises a resource so that it becomes limited for other strains to utilise, could suppress strain growth in communities and subsequently reduce the symbiont population diversity in the rhizosphere of legume hosts (Ramachandran *et al.*, 2011; Becker *et al.*, 2012). Strains that are more genetically related are suggested to have a stronger competition for shared resources due to higher niche overlap of metabolic capabilities (Griffin, West and Buckling, 2004). Additionally, indirect competition can be mediated by siderophores, which are used by rhizobia to sequester iron and subsequently inhibit growth of competitor strains (Joshi *et al.*, 2008; diCenzo *et al.*, 2014; Kramer, Özkaya and Kümmerli, 2019). Indirect growth suppression has previously been evaluated by observing the interaction between strains and competitor strain supernatants. Supernatant interactions consider the resource consumption and secreted metabolites by one strain into growth media and both mechanisms can restrict growth of other strains. While *Rhizobium* supernatant interactions have been studied at the interspecies level with other microbes (Plazinski and Rolfe, 1985; Abd-Alla *et al.*, 2014), there exists only one study where the supernatant effects between two *Rhizobium leguminosarum* symbiovar *viciae* (*Rlv*) strains was shown to affect strain nodulation and nitrogen fixation efficiency (Bladergroen, Badelt and Spaink, 2003). Rhizobia can also interact via interference (direct) competition, whereby strains actively prevent one another's growth (Ghoul and Mitri, 2016; Checcucci *et al.*, 2017). One mechanism by which bacteria can inhibit growth of neighbouring strains is through secretion of quorum sensing chemical signalling molecules that increase in concentration in a cell density-dependent manner and lead to altered regulation of gene expression of sensitive strains (Miller and Bassler, 2001; Wisniewski-Dyé and Downie, 2002). Quorum sensing molecules have been predominantly identified as *N*-acyl homoserine lactones (AHLs), and *Rhizobium* are known to produce the greatest diversity of these quorum sensing molecules among soil bacteria (Cha *et al.*, 1998; Wisniewski-Dyé and Downie, 2002). Rhizobia can use AHLs to regulate growth inhibition and surface polysaccharide production of susceptible neighbouring strains, in addition to other physiological activities and plant interactions (Schwinghamer and Brockwell, 1978; Miller and Bassler, 2001; Wisniewski-Dyé and Downie, 2002; Downie, 2010). These bacterial quorum sensing systems have been suggested to be advantageous in crowded rhizospheres of nodulated legumes where strains can influence colonization functions such as cell motility, root adhesion and growth of rhizosphere populations (Wisniewski-Dyé and Downie, 2002; He *et al.*, 2003). Previous research into quorum sensing in rhizobia has used *Rlv* as a model from which four main LuxI-type AHL synthase genes (*cinI*, *rhiI*, *raiI*, and *traI*) and their regulators were identified, some of which are homologous to quorum sensing systems in other bacterial species (Miller and Bassler, 2001; Wisniewski-Dyé and Downie, 2002). However, the variation in quorum sensing system-mediated competition has not been associated with observed intraspecific genetic variation.

In addition to quorum sensing AHLs, some stains of *Rhizobium* can also produce bacteriocins, which are narrow-spectrum growth inhibitory agents and active only against closely related strains (Hirsch, 1979). *Rhizobium leguminosarum* strains have been shown to produce three common types of bacteriocins called *small*, *medium* and *large* due to their suggested molecular weights and diffusion properties (Hirsch, 1979; Sanchez-Contreras *et al.*, 2007). *Small* was later found to be a quorum sensing AHL, and *medium* was found to be an RTX-like protein (Hirsch, 1979; Schripsema *et al.*, 1996; Oresnik, Twelker and Hynes, 1999; Lithgow *et al.*, 2000). Bacteriocin-producing strains were found to strongly inhibit growth of sensitive strains, alter strain community composition in liquid media, peat cultures and natural soil, and additionally influence competition for nodule occupancy (Schwinghamer and Brockwell, 1978; Hirsch, 1979; Bosworth, Breil and Triplett, 1993; Wilson, Handley and Beringer, 1998; Oresnik, Twelker and Hynes, 1999). One bacteriocin called trifolitoxin can be produced by some *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) strains and it can induce bacteriostatic properties against *Rhizobium leguminosarum* symbiovars and other *Rhizobium* species as well (Triplett and Barta, 1987; Bosworth, Breil and Triplett, 1993; Robleto, Borneman and Triplett, 1998). Additionally, rhizobia strains can carry temperate phage which have also been shown to suppress growth of sensitive strains (Schwinghamer and Brockwell, 1978; Harrison and Brockhurst, 2017).

Rhizobial strains can also interact in a cooperative manner, and this is more likely to occur between closely related individuals within a community (Zee and Bever, 2014; Barrett *et al.*, 2015). The reasons for maintenance of facilitative interactions is debated, but it has been suggested that: 1) spatially structured environments can maintain cooperation because strains are likely to be located close to their progenitors and genetically similar strains; and 2) cooperation can be beneficial by increasing the invasion resistance of communities (Bruno, Stachowicz and Bertness, 2003; Griffin, West and Buckling, 2004; Hibbing *et al.*, 2010; Zee and Bever, 2014). Syntrophic interactions between rhizobia (where waste products from one strain can be metabolised by another strain) are another cooperative mechanism. This cross-feeding interaction has been observed between rhizobia and other bacterial species (Silva *et al.*, 2019) and it could facilitate the coexistence of rhizobia strains and thereby increase their chances in initiating symbiosis (Bruno, Stachowicz and Bertness, 2003; Silva *et al.*, 2019). In support for this, nodulation and nitrogen fixation has been shown to increase local resources at the legume root, which could also benefit free-living rhizobia in close proximity to the nodule (Zee and Bever, 2014; Teng *et al.*, 2015). Refining facilitative interactions further, some rhizobia can stimulate legume hosts to produce nutrients (e.g. rhizopines) intended only for genetically similar strains close to the root (Zee and Bever, 2014; Barrett *et al.*, 2015). In addition, rhizobial quorum sensing can also instigate facilitative interactions by increasing nodulation efficiency of related strains, and by inducing transfer of symbiotic plasmids and islands and therefore increasing their symbiotic capacity (Miller and Bassler, 2001; Wisniewski-Dyé and Downie, 2002; Downie, 2010; Miao *et al.*, 2018).

This study focused on investigating direct and indirect intraspecies competitive interactions between *Rhizobium leguminosarum* strains capable of forming symbiosis with clover; an agriculturally important forage legume. Specifically, we focused on pairwise interactions between 24 genetically diverse *Rhizobium leguminosarum* strains belonging to three distinct subspecies (genospecies A, C and E; < 95% average nucleotide identity) (Kumar *et al.*, 2015) and two farming treatments (organic and conventional) with the aim to understand whether neutral, facilitative or inhibitory intraspecies interactions are linked with genetic background and agricultural practices (Portella *et al.*, 2009). To achieve this, indirect facilitative and inhibitory interactions were characterised by comparing the growth of strains in their own supernatant (accounting for nutrient consumption and production of secondary metabolites) compared to their growth in a different strain's supernatant. Additionally, inhibition zones produced in direct contact on soft agar plates were quantified as evidence of direct growth repression by specific strains. To understand the potential underlying mechanisms of competition, strains were compared regarding their metabolic capacity and the absence and presence of genes associated with quorum sensing, bacteriocins, secondary metabolites and prophages using comparative genomics.

6.3.Methods

6.3.1. Rhizobia strains

Twenty-four *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) strains isolated from organic and conventional trial (conventional from hereon) farm treatments across Denmark were selected from the NCHAIN *Rlt* isolate collection (Cavassim *et al.*, 2020). Strains were genetically characterised based on their *Rlt* subspecies classification as genospecies (gs) A, C and E (Kumar *et al.*, 2015) and further labelled into four categories based on environmental origin: organic gsA (OA, n=6), organic gsC (OC, n=7), organic gsE (OE, n=5) and conventional gsC (CC, n=6) (Table 6.1). An Average Nucleotide Identity (ANI) value greater than 95% is accepted to equate to a DNA-DNA hybridisation value of 70%, and therefore would indicate strains to be genetically distinct species (Goris *et al.*, 2007). Within genospecies ANI averaged 98.2% and ranged between 96.8-99.9% (ANI based on 441,287 shared single nucleotide polymorphisms in 6,529 genes present in at least 100 strains) (Cavassim *et al.*, 2020). Between genospecies ANI values averaged 91.6% and ranged between 90.2-97.7% (Cavassim *et al.*, 2020). Strains were routinely cultured on Tryptone Yeast (TY) agar or liquid media.

Table 6.1 Twenty-four *Rhizobium leguminosarum* symbiovar *trifolii* strains isolated from *Trifolium repens* nodules across Danish farm sites. OA = organic genospecies A, OC = organic genospecies C, OE = organic genospecies E, and CC = conventional genospecies C.

Genospecies category	**Strain names**						
OA	SM152B	SM137B	SM152A	SM145B	SM154C	SM144A	
OC	SM147A	SM158	SM170C	SM157B	SM165A	SM122A	SM126B
OE	SM149A	SM135B	SM135A	SM159	SM168A		
CC	SM41	SM53	SM57	SM77	SM74	SM67	

6.3.2. Measuring competitive and facilitative pairwise interactions between rhizobia strains

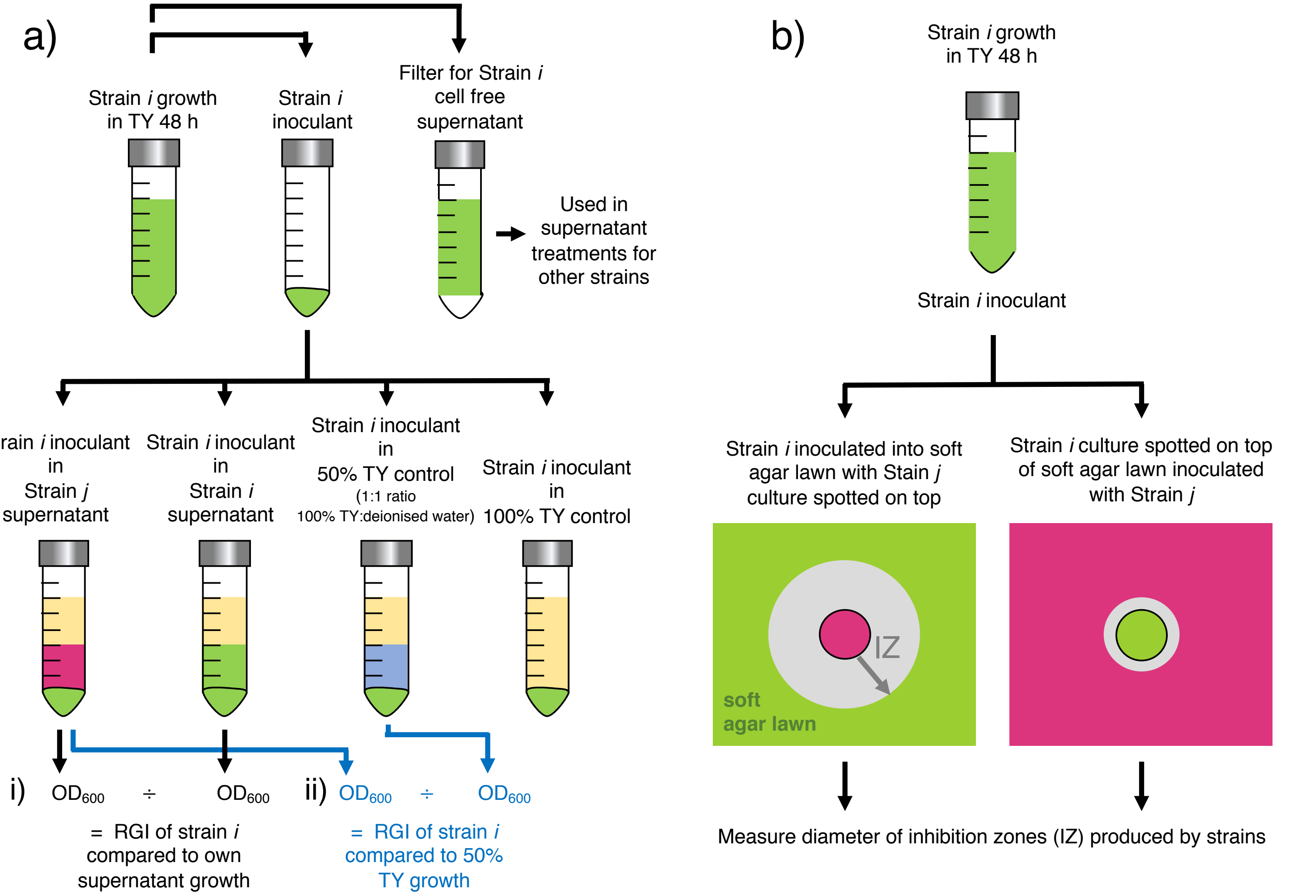
6.3.2.1. Determining rhizobial interactions indirectly using TY supernatant assay

Potential facilitative and inhibitory pairwise strain interactions were determined indirectly based on each strain's growth in the supernatant of every other strain. Strains were revived from frozen glycerol stocks in 40 ml of Tryptone Yeast broth (TY broth: 5 g tryptone, 2.5 g yeast extract, 1.47 g $CaCl_2$ per litre volume) for 48 h (28°C, 180 rpm). 600 µl from each 40 ml culture was saved for later use as an inoculum. The remaining culture was centrifuged (10 minutes, 4000 rpm) and the supernatant was filtered through a 0.2 µm syringe filter. An equal volume of fresh 100% TY broth was added to produce a 50:50 supernatant-broth mixture (supernatant treatment) for each strain. Supernatant treatments account for the resource consumption and secreted metabolites by one strain, both of which could affect the growth of other target strains grown in said supernatant treatments. Additionally, strains were grown in 100% TY and 50% TY (50:50 of 100% TY and deionised water) broth control treatments. The 50% TY control treatment was used to determine a strain's minimum expected growth from a supernatant treatment if the supernatant invoked no inhibitory effects on growth. This is because the amount of added TY broth is the same in 50% control and supernatant treatments. A 100% TY control treatment was used to ensure strains grew well in rich nutrient medium and any observed reductions in growth were either due to lower nutrient broth concentrations (50% TY control) or inhibitory metabolites within the supernatant treatments.

To start the growth assays, 200 µl of each 50:50 supernatant-broth mixtures were added to 96 well plates with 5 replicates per strain. Supernatant treatments were inoculated with the initial inocula (~0.2 µl) using a sterilised microplate pin replicator (Boekel). One well of each supernatant treatment was inoculated with water as a no growth control. Strains were grown at 28°C and OD_{600} measurements were taken, as an indicator of growth, at 0 hours, 24 h, 39 h, 48 h and 62 h after strain inoculation. A total of 624 inoculant-supernatant combinations were analysed, including 100% and 50% TY treatments. Relative growth indices (RGIs) were calculated for all strains in all supernatant treatments after 62 h growth (Appendix Figure E.1). RGI's were calculated by comparing the growth of strain *i* in strain *j*'s cell-free supernatant in relation to growth in its own supernatant for each combination (Figure 6.1a) using the following equation:

$$\frac{\text{Growth } (\text{OD}_{600}) \text{ of strain } i \text{ in supernatant } j}{\text{Growth } (\text{OD}_{600}) \text{ of strain } i \text{ in supernatant } i} = \text{RGI of strain } i$$

Therefore, a value of 1 indicates strain *i* grows equally well in the supernatant of strain *j* as in its own supernatant. If a strain grows to a higher optical density in another strain's supernatant than its own, it receives a score of >1, (i.e. displays facilitated growth from strain *j* supernatant). The converse is assumed if strain *i* grows to a lower optical density in strain *j* supernatant, than when grown in its own. An RGI more than 1 suggests that nutrient resources are left behind by a strain, or that metabolites excreted by a strain can be used for additional growth by another strain. An RGI less than 1 could suggest that the majority of nutrient resources available have been consumed by a strain, therefore offering no additional nutrients than a 50% TY control treatment. Alternatively, it could suggest that a strain has secreted inhibitory metabolites that prevent growth of other strains. Therefore, to only evaluate the effect of resource consumption on rhizobial growth, RGIs were calculated comparing a strains' growth in each treatment compared to when grown in the 50% TY control treatment (Figure 6.1a).



a)
Strain *i* growth in TY 48 h
Strain *i* inoculant
Filter for Strain *i* cell free supernatant
Used in supernatant treatments for other strains
Strain *i* inoculant in Strain *j* supernatant
Strain *i* inoculant in Strain *i* supernatant
Strain *i* inoculant in 50% TY control (1:1 ratio 100% TY:deionised water)
Strain *i* inoculant in 100% TY control
i) OD600 ÷ OD600
= RGI of strain *i* compared to own supernatant growth
ii) OD600 ÷ OD600
= RGI of strain *i* compared to 50% TY growth
b)
Strain *i* growth in TY 48 h
Strain *i* inoculant
Strain *i* inoculated into soft agar lawn with Stain *j* culture spotted on top
Strain *i* culture spotted on top of soft agar lawn inoculated with Strain *j*
IZ
soft agar lawn
Measure diameter of inhibition zones (IZ) produced by strains

Figure 6.1 Experimental designs for the **a)** indirect interaction supernatant assay and **b)** direct interaction spot plating assay. **i)** calculates the relative growth index (RGI) of strain *i* in the strain *j* supernatant compared to growth in its own supernatant, and **ii)** calculates the RGI of strain *i* in the strain *j* supernatant compared to growth in the 50% Tryptone Yeast broth control treatment.

6.3.2.2. Determining rhizobial interactions directly using spot assays on TY agar plates

Direct inhibitory (interference) pairwise interactions between rhizobia strains were determined by spotting a liquid culture of each strain on a bacterial lawn of every other strain (Figure 6.1b). Level of inhibition was determined as the inhibition halo diameter of the bacterial lawn around the spotted bacterial colony. For the assay, strains were grown in 5 ml TY broth for 48 h at 28°C. Optical density of cultures showed strains had grown between 0.055-0.09 at OD_{600}. OA strains had significantly higher initial culture spot inoculum ODs compared to other genospecies groups (Kruskal-Wallis: chi-squared = 20.036, df = 3, $p < 0.001$). However, a simple linear regression confirmed that no significant association was observed between the size of the inhibition zone around the culture spot and the optical density of the inoculum used for the assay (Appendix Figure E.2; $Coeff_{inoculumOD}$: -6.47, $p > 0.05$).

400 µl of culture was then mixed with 40 ml of soft TY agar (7.5% grams of agar per volume) and plated in square petri-dishes and left to cool. Two plates of soft agar were made for each strain and 12 strain were spotted on each plate. Additionally, uninoculated 100% TY soft agar plates were used as a control. 2 µl of each rhizobia culture was spotted onto the plates. Also, a control uninoculated TY broth spot was placed in the corner of each plate as a control to ensure no inhibition was observed from the sterile broth alone. Plates were incubated at 28°C and imaged with digital camera at 24, 48 and 72 h.

The inhibition zone diameters and culture spot diameters were compared to identify if the growth of the spotted strain correlated with the level of inhibition. A very weak positive correlation was identified (Pearson's Correlation R statistic = 0.34, $p < 0.001$) and a simple linear regression found inhibition zone diameters increased by 2.187 mm on average for every 1 mm increase in culture spot diameter ($Coeff_{spotdiameter}$: 2.187, $p < 0.001$). However, this relationship was heavily biased by a few strains producing small inhibition zones (Appendix Figure E.3a). When these samples were removed the resulting correlation weakened (Pearson's Correlation R statistic = 0.17, $p < 0.01$; Appendix Figure E.3b) and a simple linear regression identified that association with inhibition zone diameter reduced to 1.288 mm on average for every unit increase in spot diameter ($Coeff_{spotdiameter}$: 1.288, $p < 0.01$). To control for the growth of the spotted bacterium, the diameter of the spotted bacterial culture was subtracted from the diameter of the inhibition zone in all analyses. Inhibition zone diameters were calculated by subtracting the Feret diameter of the culture spot by the Feret diameter of the inhibition zone using ImageJ (v.1.52k). All strain lawn and spot combinations were measured in 3 technical replicates (with the exception of strains SM137B and SM122A where two replicates were used, and strain SM152B where one spotting replicate was used). All replicates were used to calculate average mean inhibition zones for all pairwise strain combinations.

6.3.3. Characterising the metabolic capacity of different rhizobia strains

To determine differences in metabolic capacity, all strains were grown on 31 single substrates using EcoPlates (Biolog Hayward, CA, USA) (Smith, 2018). The 31 single substrates were defined into the following resource type groups; amines, amino acids, carbohydrates, carboxylic acids, complex carbons, and phosphate carbon and water as a control (Table 6.2). All amino acids are assumed to be available in TY broth as subcomponents, and although other single substrates may not likely be present within TY broth, they provided additional understanding of a strain's metabolic potential. One replicate was generated for each strain, therefore strains were grouped by genospecies and substrates were grouped into above mentioned resource type groups (Table 6.2), to enable statistical analysis.

Before measurements, all strains were grown in 10 ml TY broth for 48 h (28°C, 180 rpm), centrifuged to form a pellet and re-suspended into 10 ml PBS buffer, and incubated for 2 h at room temperature (Smith, 2018). 120 µl of bacterial suspension was added to each of the 31 Ecoplate carbon sources and water control wells (Smith, 2018). Reduction of tetrazolium dye within each carbon source well occurs when microbes can metabolise the resource and subsequently respire. Plates were incubated at 28°C and OD_{590} measurements of the developed dye coloration were taken at 72 h (Smith, 2018). ODs were normalized by subtracting the control water well OD from the substrate well ODs for each strain (Appendix Figure E.4). Strains generating OD values greater than 0 for a substrate well were considered being able to metabolise that particular substrate (Appendix Figure E.4). To indicate whether a strain has predominantly resource generalist or specialist characteristics, the catabolic range of genospecies was determined by totalling the number of substrates metabolised by each strain and was further used to calculate the mean number of substrates metabolised per genospecies group. Metabolically generalist strains (capable of metabolising many substrates) might be able to better deplete the TY broth of resources resulting in supernatants providing fewer available remaining nutrients to other strains. Similarly, metabolically specialist strains (metabolise few resources) would leave behind resources in their supernatants for generalists to use, making their supernatants more facilitative.

Additionally, the average well colour development (AWCD) of each strain calculates an overall metabolic capability of each genospecies, and how efficient strains are at depleting resources. Therefore, AWCD was used as a measure of strain's average metabolic capacity under different substrate treatments. AWCD was calculated using 72 h OD_{590} measurements of each well (Garland and Mills, 1991; Garland, 2006):

$$AWCD = [\Sigma(S - C)]/n$$

S is the substrate well OD_{590} value, C is the control well OD_{590} value and n is the number of substrates (i.e. 31 for AWCD across all substrate treatments). AWCD values were also calculated for substrates grouped into 6 resource type groups by molecular characteristics (Table 6.2; Smith, 2018). Strain SM159 (OE) was removed from the analyses due to abnormally high OD values likely due to contamination.

Table 6.2 Ecoplate single substrate growth treatments grouped into 6 resource type groups, as previously (Smith, 2018).

Resource type groups	**Ecoplate Substrates**
Amines	Phenylethyl-Amine Putrescine
Amino acids	Glycyl-L-Glytamic Acid, L-Arginine L-Asparagine L-Phenylalanine L-Serine L-Threonine
Carbohydrates	D-Cellobiose D-Mannitol D-Xylose i-Erythrithol N-Acetyl-D-Glucosamine α-D-Lactose β-Methyl-D-Glucoside
Carboxylic acids	2-Hydroxy Benzoic Acid 4-Hydroxy Benzoic Acid D-Galactonic Acid γ-Lactone D-Galacturonic Acid D-Glucosaminic acid D-malic acid Itaconic Acid Pyruvic Acid Methyl Ester α-Ketobutyric Acid γ-Hydroxybutyric Acid
Complex carbon sources	Glycogen Tween 40 Tween 80 α-Cyclodextrin
Phosphate carbon	D,L- α-Glycerol Phosphate Glucose-1-Phosphate

6.3.4. Comparative genomic analyses of candidate gene clusters linked with competition

To investigate genetic differences in metabolism linked to the presence of genes linked with quorum-sensing, secondary metabolites and prophages, we undertook comparative genomics analyses using various online platforms.

To compare the ability of strains to produce different bacteriocins and quorum-sensing responses, we used BLASTn to search for genes within full genome assemblies of the 24 strains that are known to be associated with these pathways in *Rhizobium leguminosarum* (Table 6.3) (Schripsema *et al.*, 1996; Wisniewski-Dyé and Downie, 2002; Gonzalez and Marketon, 2003; McAnulla *et al.*, 2007). Additionally, to compare other known mechanisms of strain interactions using bioactive compounds, such as siderophores, secondary metabolite biosynthesis gene clusters were searched for using antiSMASH 5.0 with default settings (Blin *et al.*, 2019). All identified secondary metabolite clusters containing 2 or more genes were counted for analysis, even if the cluster had no sequence similarity to specific known clusters. To compare the number of potential prophage regions across genospecies groups, putative prophage regions were searched for in each strain and identified as either 'intact', 'incomplete' or 'questionable' with PHASTER using default parameter settings (Arndt *et al.*, 2016). Only intact prophage regions were considered to be likely active.

Table 6.3 GenBank accessions of known quorum sensing and bacteriocin associated gene sequences. QS refers for quorum sensing.

GenBank Accession	Gene name	Gene type	Literature reference
L06719.1 (RHMTFXA2G)	*Rhizobium leguminosarum trifolii* trifolitoxin (*tfxA*) gene and *tfxB*, *tfxC*, *tfxD*, *tfxE*, *tfxF*, *tfxG* genes	Bacteriocin	-
AJ001518.1	*medium* bacteriocin	Bacteriocin RTX-like protein	Oresnik et al., 1999
AAF89990.1	*cinI*	QS LuxI-type AHL synthases	-
AAF89989.1	*cinR*	QS LuxR-type regulator	-
CBI71465.1	*cinS*	QS regulator	-
RWX40560.1	*raiI*	QS LuxI-type AHL synthases	-
AAC38173.1	*raiR*	QS LuxR-type regulator	-
AAO21111.1	*bisR*	QS LuxR-type regulator	-
AAO18654.1	*traI*	QS LuxI-type AHL synthases	-
AAO21112.1	*traR*	QS LuxR-type regulator	-
CAK10388.1	*rhiI*	QS LuxI-type AHL synthases	-
CEG06613.1	*rhiR*	QS LuxR-type regulator	-
CAX32456.1	*expR*	QS LuxR-type regulator	-

6.3.5. Statistical Analysis

Multiple statistical approaches were used to analyse the data, including mixed effects models, likelihood ratio (LR) tests and parametric bootstrapping of 95% confidence intervals. Supernatant interactions where a strain was grown in its own supernatant (therefore generating an RGI of 1) were excluded from the analysis, to avoid biasing data distributions. To analyse the effects of genospecies and farm treatment group on direct and indirect inhibition between strains, maximum likelihood (ML) mixed effect models were produced for both supernatant assays, and spot plating assays, with lme4 R package (v.1.1-21). For supernatant assays, inoculant genospecies group (OA, OC, OE, CC) and supernatant genospecies group were included as fixed effects, while individual inoculant strain IDs and supernatant strain IDs were categorised as crossed random effects. For spot plating assays, liquid culture spot genospecies group (OA, OC, OE, CC) and soft agar lawn genospecies group were included as fixed effects, and similarly, individual culture spot strain IDs and soft agar lawn strain IDs were classed as crossed random effects. Random effects accounted for pseudo-replication and their variance of > 0 supported their incorporation in the full models. LmerTest generated t-values, degrees of freedom and p-values for fixed effect parameters in the models. To identify genospecies group differences for each variable, genospecies groups were ranked by average value and the genospecies group with the lowest value set as the intercept. Therefore, for all variables in all models the intercept was set to the OA genospecies group. Firstly, the full models were generated whereby fixed effects included an interaction. The significance of the fixed effects interaction was tested by the likelihood ratio (LR) test using anova() by comparing the full interaction model with a reduced model with no interaction. If Chi-squared p-values were < 0.05, model fits were determined as significantly different. In addition, the reliability of the fixed effects was determined by parametric bootstrapping of fixed effects as 95% confidence intervals in the final models (bootMer and boot.ci with 1000 bootstrap replicates). Fixed effect parameters with 95% confidence intervals that contained 0 were considered as non-reliable effects. The bootstrapping model displayed warnings of failed model convergence for the supernatant assay model (30 out of 1000 permutations) and the spot plating model (52 out of 1000 permutations). Due to the original model converging with no warnings, and the arbitrary nature of the threshold for model convergence warnings, these warnings were classified as false positive convergence warnings (Bolker, 2020). In order to test whether some strains influenced the observed interaction effects, specific strains were removed, and the models were rerun to confirm fixed effects parameters remained significant. Similarly, bootstrapping displayed warnings of failed convergence for some permutations; supernatant assay model without SM154C and SM168A = 44 out of 1000 permutations; spot plating assay model without SM144A, SM154C and SM145B = 38 out of 1000 permutations. Additionally, warnings regarding singular fits were generated from bootstrapping for the following models: spot plating assay model = 24 out of 1000 permutations; spot plating assay model without SM144A, SM154C and SM145B = 665 out of 1000 permutations. To further determine if OC and CC strains displayed significant differences in inhibitory activity when acting as the inoculant or supernatant/soft agar strain, the estimated marginal means of interactions from the mixed effects model were compared using emmeans package in R with Tukey adjusted p-value correction applied. Furthermore, to identify the overall trend across supernatant interactions, Pearson's correlation coefficient and simple linear regression (lm() in R) was used. To determine whether more genetically similar strains displayed more neutral indirect interactions, RGI as an absolute value was correlated to ANI using a linear regression with White's robust standard errors using R's sandwich package to correct for homoscedasticity.

To determine whether a genospecies and farm treatment group displayed generalist or specialist traits, the number of single substrates each strain was able to metabolise ($OD_{590} > 0$) were used to calculate the mean number of metabolised substrates for each genospecies group. Non-parametric Kruskal Wallis test was used to compare metabolic capacities as a measure of AWCD (across 31 single substrate treatments) and to determine whether genospecies groups could metabolise a significantly different number of single substrates. Dunn's post-hoc test was used to identify direct differences between groups from non-parametric Kruskal-Wallis tests with Bonferroni adjusted p-values. Pearson's R correlation coefficient was used to determine the correlation between strain's RGI in supernatant treatments and metabolic capacity as a measure of AWCD across all 31 single substrate treatments. Principal component analysis (PCA) was calculated with R prcomp using singular value decomposition to calculate principal components for explaining: 1) metabolic capacity (AWCD) of *Rlt* strains across six resource type groups; and 2) the strain metabolic capacity across single substrate treatments (i.e. not grouped). PERMANOVA using adonis() in the R vegan package was used to test for significance of PCA clustering and significant pairwise genospecies interactions were identified with post hoc testing using pairwise.adonis() and Bonferroni p-value correction.

6.4. Results

6.4.1. Facilitative and inhibitory rhizobial interactions were observed at both genospecies and genotype level

6.4.1.1. Supernatant growth assays

To assess resource- and metabolite-mediated competitive interactions in pairwise *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) strain interactions, the growth of *Rlt* strains in the supernatants of other *Rlt* strains was compared after 48 hours of initial growth resulting in a total of 576 pairwise combinations (Figure 6.2a). Strain growth in the supernatants of other *Rlt* strains was compared to when the strain was grown in its own supernatant (relative growth index: RGI). Low relative growth in another strain's supernatant suggests inhibitory interactions (RGI < 1), whereas high relative growth indicates more facilitative interactions (RGI > 1). Both facilitative and inhibitory interactions were identified (Figure 6.2a), but overall indirect interactions were predominantly neutral (mean RGI: 1.006; Appendix Figure E.5a). Some strain inoculant and supernatant combinations showed extreme facilitative or inhibitory interactions. 55 combinations (9.55%) had RGI's < 0.75 suggesting they grew worse in other strains supernatants compared to their own. Also, 55 combinations (9.55%) had RGI's > 1.25 suggesting growth in other strain supernatants caused increased growth. Additionally, strains that were genetically more similar were likely to show a more neutral interaction (i.e. an RGI of 1; Appendix Figure E.5b).

Genospecies effects predominantly drove the interaction between inoculant and supernatant groups (Figure 6.3a; Appendix Table E.1; $X^2_{19,9} = 102.5$, $p < 0.0001$). On average, genospecies E (OE) strains were the most facilitated in the supernatants of other strains (Figure 6.3b), and on average their supernatants consistently overly suppressed genospecies A (OA) growth (compared to OA inoculants in OA supernatants as the model intercept reference level; Figure 6.2a; $\text{Coeff}_{\text{supOE}}$: estimate = -0.382, std. error = 0.056, t = -6.791, $p < 0.001$). Parametric bootstrapping of 95% confidence intervals further confirmed that the growth inhibition of OA inoculants in OE supernatants was a reliable effect (Parametric bootstrapping 95% $\text{percentile}_{\text{supOE}}$ (-0.4915, -0.2797): original = -0.382, bias = -0.00124, std. error = 0.0547). Additionally, OE inoculants displayed facilitated growth in OA supernatants, and this facilitation of OE inoculants in OA supernatants (compared to the reference level) was also confirmed to be a reliable effect ($\text{Coeff}_{\text{inocOE}}$: estimate = 0.209, std. error = 0.068, t = 3.08, $p < 0.01$; Parametric bootstrapping 95% $\text{percentile}_{\text{inocOE}}$ (0.0721, 0.3275): original = 0.209, bias = -0.00322, std. error = 0.0668).

OA strains grew to lower densities in OE supernatants than in 50% TY control treatments. Supernatant treatments are composed of a 1:1 ratio of strain supernatant and 100% TY, and so the amount of added TY in the supernatant treatment equates to a 50% TY treatment. Therefore, if strains grow better in supernatant treatments compared to the 50% TY control, it is assumed additional nutrients are provided by remaining resources in the supernatant. Due to OA strains growing worse in OE supernatants compared to 50% TY, this suggests that OA growth inhibition cannot be purely due to nutrient resource depletion in OE supernatants and was likely associated with other inhibitory processes that are preventing growth up to densities expected from 50% TY treatments (Appendix Figure E.6).

Furthermore, genospecies C strains isolated from either organic or conventional farming treatments (OC and CC respectively) were compared to observe if strain environmental origin influenced strain interactions, and to control for genospecies effects. On average, there were no significant differences in genospecies interactions depending on whether genospecies C strains originated from organic or conventional farm treatments (Figure 6.3a: Appendix Table E.2; Appendix Table E.3). OC and CC strains grew to relatively similar densities in other strain supernatants compared to when grown in their own supernatants (Figure 6.3b; Appendix Table E.2). Similarly, the average growth densities of other genospecies did not significantly differ between OC and CC supernatant treatments (Figure 6.3c; Appendix Table E.3).

Rlt interactions were also evaluated at the strain level for genotype-specific effects. SM168A (OE) grew better on average in other supernatant treatments than its own when acting as the inoculum (Figure 6.2a; $\text{RGI}_{\text{inoculant}}$ = 1.381, 95% conf. int = 1.316 – 1.446), and its supernatant highly suppressed the growth of other strains on average (Figure 6.4; $\text{RGI}_{\text{supernatant}}$ = 0.883, 95% conf. int = 0.858 – 0.908). Interestingly, SM168A grew worse in supernatant treatments compared to the 50% TY control on average, despite displaying one of the highest average inoculant RGI's out of the 24 strains. Conversely, SM154C (OA) grew significantly worse in other supernatant treatments on average (Figure 6.2a; $\text{RGI}_{\text{inoculant}}$ = 0.612, 95% conf. int = 0.580 – 0.644), and its supernatant facilitated growth of other strains (Figure 6.3; $\text{RGI}_{\text{supernatant}}$ = 1.269, 95% conf. int = 1.223 – 1.315). To ensure group interactions were not influenced by individual strains, SM168A and SM154C were omitted and the model was re-calculated. Despite strain exclusion, the behaviours of OA and OE groups remained similar and statistically significant, indicative that the genospecies-level strain interactions were maintained (Appendix Table E.4; $\text{Coeff}_{\text{supOE}}$: estimate = -0.376, std. error = 0.0505, t = -7.446, $p < 0.001$; $\text{Coeff}_{\text{inocOE}}$: estimate = 0.122, std. error = 0.0547, t = 2.225, $p < 0.05$; 95% $\text{percentile}_{\text{supOE}}$ (-0.4797, -0.2740): original = -0.376, bias = -0.0005, std. error = 0.0529; 95% $\text{percentile}_{\text{inocOE}}$ (0.0113, 0.2298): original = 0.122, bias = -0.0006, std. error = 0.0545).

To evaluate how facilitative a strain's supernatant was for the growth of other strains in comparison to whether its own growth was facilitated by the supernatants of other strains, the average RGI of all strains grown in strain *i*'s supernatant (suppressiveness as supernatant) was correlated with the average RGI of strain *i* in all supernatant treatments (growth as inoculant). Overall, a positive correlation was observed between the suppressiveness of a *Rlt* strain's supernatant and its growth as an inoculant (Figure 6.4; Simple linear regression: $\text{Coeff}_{\text{RGIsup}}$ = -0.8387, $p < 0.0001$; Pearson's Correlation R statistic = -0.701, t = 4.611, $p < 0.001$) indicative of positive relationship between growth and inhibition.

Together, these results demonstrate that genospecies effects are significantly associated with the facilitative and inhibitory indirect interactions observed between strain pair combinations. In particular, OE strains grew well in the supernatants of other strains, and produced supernatants that were suppressive to other strains, especially OA strain growth. OA strains grew comparatively poorly in supernatants of other strains, but their supernatants were mainly facilitative for other strains, especially for the growth of OE strains. Moreover, variation was also observed at the individual strain level, as demonstrated by relatively strong effects of SM168A and SM154C strains.

6.4.1.2. Direct inhibition assay

To assess whether strains could directly inhibit each other's growth, strains were grown in soft agar lawns and liquid cultures of other strains were spotted on top to observe whether spotted strains induced zones of inhibition. Inhibition zones were visible after 2 days growth, and after 72 hours of growth, 92 out of 576 possible strain combinations (15.97%) formed inhibition zones of varying sizes (Figure 6.2b). Similar to the indirect supernatant growth assay, there was a significant interaction between the genospecies groups of liquid culture spots and soft agar lawns, suggesting the genospecies group significantly determined whether inhibition zones were formed between strains (Figure 6.5a; Appendix Table E.5; $X^2_{19,9} = 95.933$, $p < 0.0001$).

Lawns of OA strains were most susceptible to inhibition zones formed by other strains out of all genospecies groups (Figure 6.5a). Of the other genospecies, OE strains seemed to be the most capable of producing inhibition zones on OA agar lawns (compared to OA inoculants on OA lawns as the model reference level; $\text{Coeff}_{\text{spotOE}}$: estimate = 4.203, std. error = 0.429, t = 9.80, $p < 0.001$; parametric bootstrapping of 95% $\text{percentile}_{\text{spotOE}}$ (3.355, 5.062): original = 4.203, bias -0.006, std. error = 0.445). Similarly, culture spots of OE strains were able to produce inhibition zones on agar lawns of at least one strain in each genospecies group, with the exception of OC strains (Figure 6.2b), which were resistant to inhibition by all other strains (Figure 6.2b; Figure 6.5c).

For strains isolated from different farm treatments, there were no significant differences between the inhibitory interactions of OC and CC strains with other genospecies groups or each other (Figure 6.5a; Appendix Table E.6; Appendix Table E.7). The exception to this was that OC strains were able to produce on average significantly slightly larger inhibition zones than CC strains on OA soft agar lawns (Appendix Table E.6). OC and CC strains produced inhibition zones on the same three OA strains and one OE strain (SM149A) (Figure 6.2b). Additionally, both OC and CC strains were not susceptible to clear inhibition zone formation by other strains, with the exception of one CC strain (SM53) which was susceptible to inhibition zones by all OE strains (Figure 6.2b).

OA susceptibility was largely driven by individual strains SM145B, SM154C and SM144A, which were the only strains to induce inhibition zones of > 5 mm (Figure 6.2b). However, when these three strains were removed from the model, OA strains were still found to be significantly susceptible to inhibition zone production by OE strains, indicating the genospecies level effects were still maintained (Appendix Table E.8; $\text{Coeff}_{\text{spotOE}}$: estimate = 2.118, std. error = 0.236, t = 8.968, $p < 0.001$; 95% $\text{percentile}_{\text{spotOE}}$ (1.654, 2.597): original = 2.118, bias = 0.0123, std. error = 0.238).

Furthermore, strains that grew better in the supernatants of other strains than in their own (high RGIs as inoculants) did not necessarily produce inhibition zones (Figure 6.6). However, if strains showed signs of direct interference competition, the size of the inhibition zone positively correlated with the strains' RGI in supernatant (Figure 6.6). The slope of this association was additionally found to differ depending on whether inhibition zones had a diameter more than 5 mm (Simple linear regression: $\text{Coeff}_{\text{Inhibitionzone}} = 0.040$, $p < 0.001$) or less than 5 mm (Simple linear regression: $\text{Coeff}_{\text{Inhibitionzone}} = 0.0698$, $p < 0.05$). This suggest that the inhibition seen in supernatant assays was not always driven by the same mechanisms observed in direct competition assays.

Together, these results show that differences in direct interference competition effects are evident between genospecies. This difference of interaction was predominantly observed between OE and OA stains with OE strains proving the most capable of producing inhibition zones, to which OA strains were the most susceptible. Additionally, direct interference competition effects correlated to varying degrees with the negative inhibitory effects observed in the supernatant assays. However, this was not always the case and strain combinations displayed suppressive interactions in both supernatant and the soft agar environments.

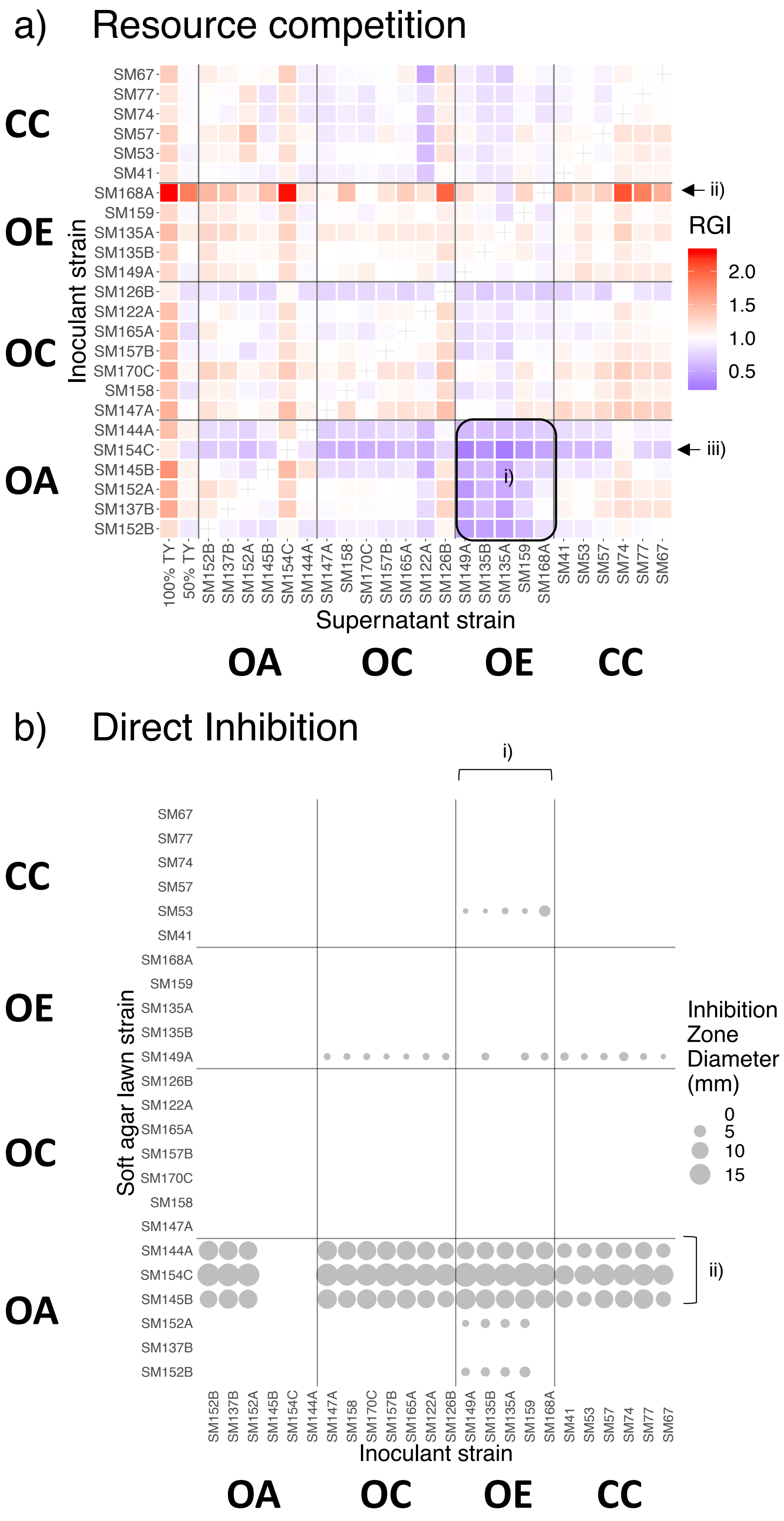
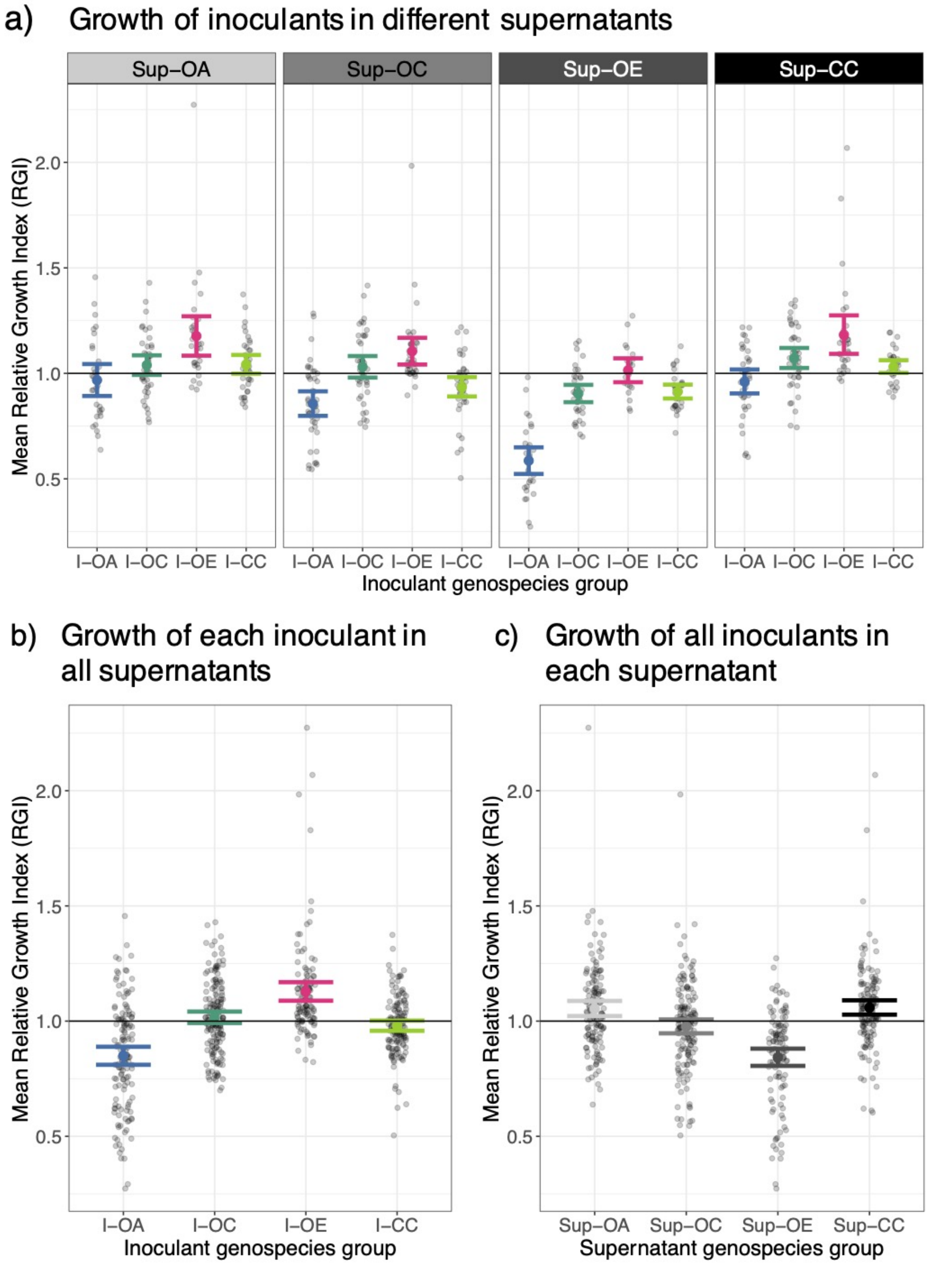


Figure 6.2 Growth of 24 *Rhizobium leguminosarum* symbiovar *trifolii* strains grown in Tryptone Yeast broth depleted by other strains (supernatant) and on soft agar lawns of other strains. **a)** Indirect interaction of strains measured by calculating the relative growth indices (RGIs) of strains inoculated into each other's supernatants (n=5). **i)** OA strains inoculated into OE strain supernatant growth treatments. (continued on following page).

Figure 6.2. continued. a) ii) Strain SM168A inoculated into all other 24 strains' supernatants and control treatments of 100% TY and 50% TY broth. **iii)** SM154C similarly inoculated into 24 supernatant treatments and controls. **b)** The mean diameter of inhibition zones (mm) produced by strains spotted onto soft agar lawns of other strains. The size of the circles indicates the diameter of the inhibition zones. **i)** diameter of inhibition zones produced by OE strain spots on all soft agar strain lawns. **ii)** particularly susceptible OA strains to inhibition zones.



a) Growth of inoculants in different supernatants

b) Growth of each inoculant in all supernatants

c) Growth of all inoculants in each supernatant

Figure 6.3 Average genospecies inoculant growth under different supernatant treatments**. a)** Mean relative growth indices (RGIs) of *Rhizobium leguminosarum* genospecies groups (OA, OC, OE, CC) inoculated (e.g. I-OA) into the supernatants of other genospecies groups (e.g. Sup-OA). **b)** Mean RGIs of each genospecies group inoculants in all other strain supernatants. **c)** Mean RGIs of all genospecies group inoculants in each genospecies group supernatant. Error bars display 95% confidence intervals. Rhizobia strain combinations were grouped by genospecies inoculant group and genospecies supernatant group.

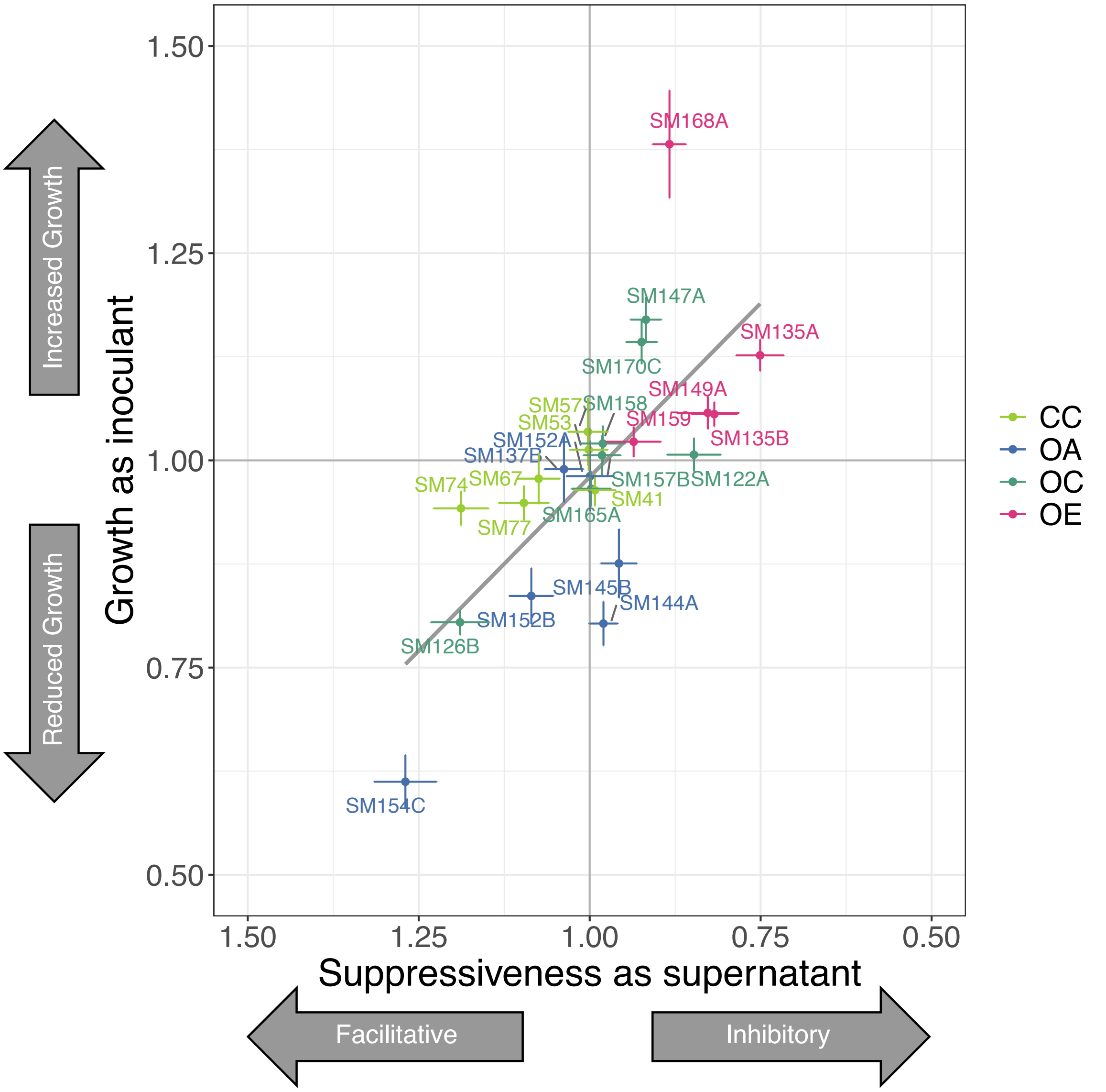


Figure 6.4 The growth of *Rlt* strains as inoculants correlate with the suppressiveness of their supernatants. Growth of inoculant was calculated by averaging the Relative Growth Index (RGI) of strain *i* when grown in all other supernatant treatments (excluding control TY treatments). Suppressiveness of supernatant was calculated by averaging the RGI of all other strains grown in the supernatant of strain *i*. Grey line displays the regression line fit by linear model, and error bars display 95% confidence intervals. RGIs calculation is displayed in the methods.

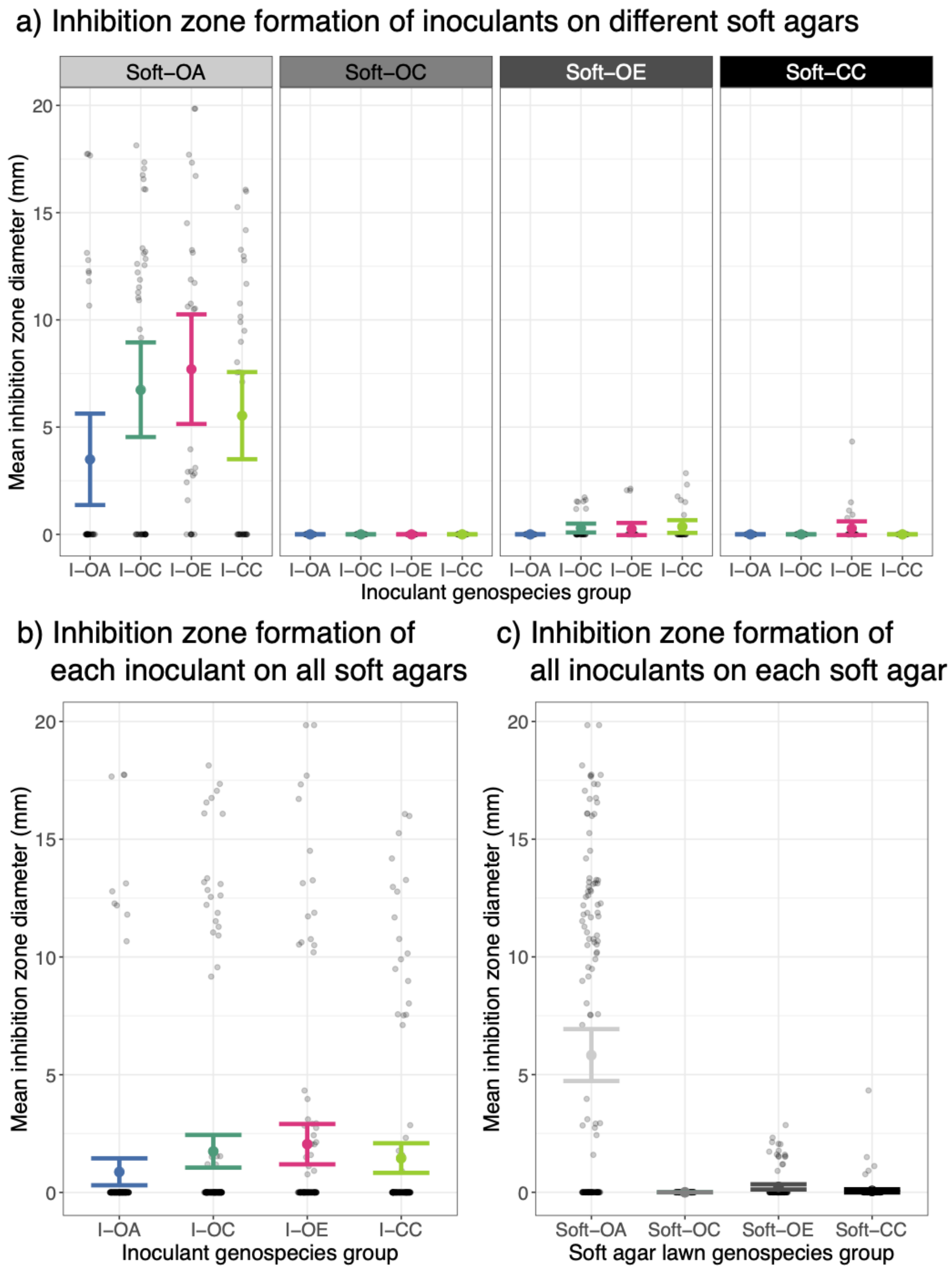


Figure 6.5 Average genospecies inoculant inhibition zone formation on different soft agar treatments. **a)** Mean inhibition zone diameter (mm) of *Rlt* genospecies groups (OA, OC, OE, CC) when liquid cultures (e.g. I-OA) are spotted onto soft agar lawns of other genospecies groups strains (e.g. Sup-OA). **b)** Mean inhibition zone diameter around each genospecies group strains inoculated on soft agar lawns of all other strains. **c)** Mean inhibition zone diameter of all strain inoculants on soft agar lawns of each genospecies group. Error bars display 95% confidence intervals. Rhizobia strain combinations were grouped by genospecies inoculant group and genospecies soft agar group.

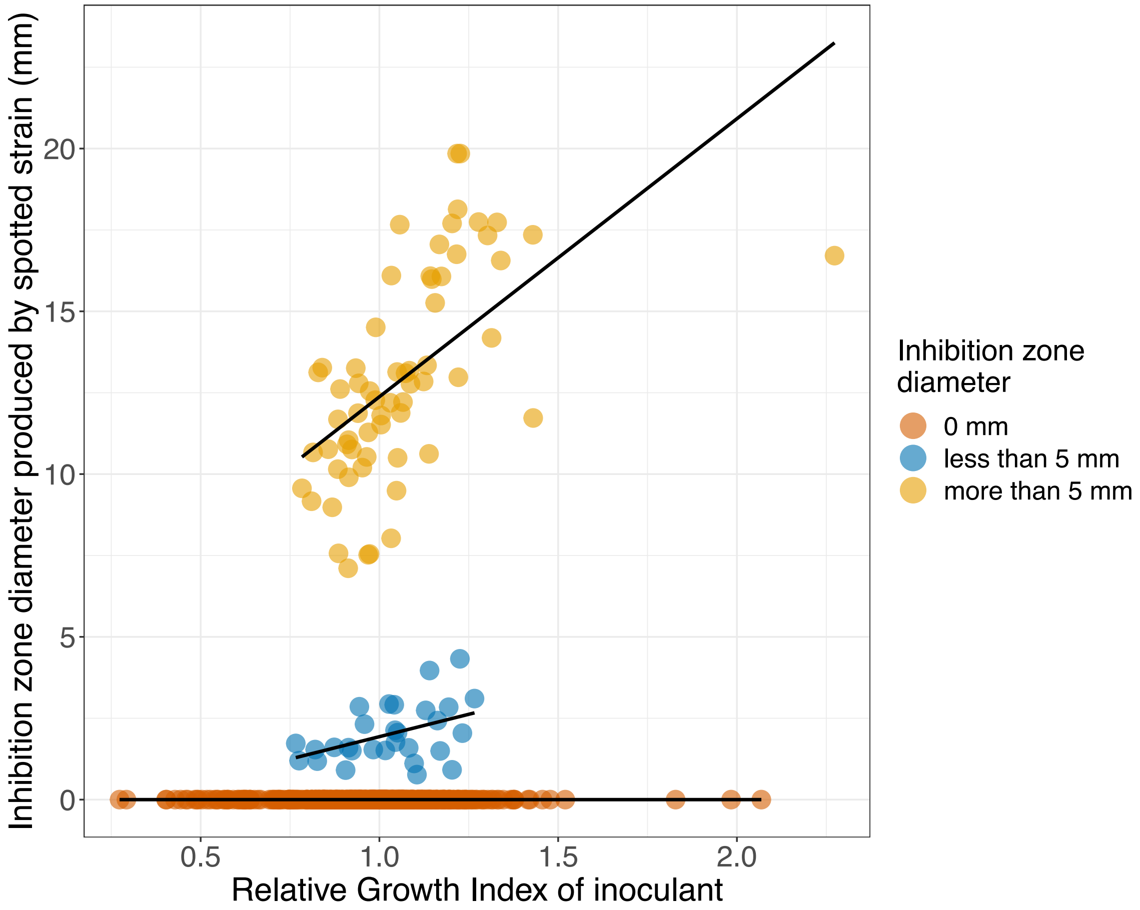


Figure 6.6 Growth in supernatant correlated to size of inhibition zone on soft agar. If strains are able to produce an inhibition zone, it is more likely that they will have a larger inhibition zone on a strain's agar lawn if they grew well in that same strain's supernatant. Mean Relative Growth Index of strain *i* as inoculant in the supernatant of strain *j* (n=5) correlate to the mean inhibition zone diameter (mm) produced by strain *i* on soft agar lawns of strain *j* (n<=3). Regression lines are fit by linear model.

6.4.2. OE strains display a greater metabolic capacity than other genospecies

To compare the overall metabolic capacity of different genospecies, all strains were cultured under 31 single substrate treatments (Appendix Figure E.4) (Smith, 2018). The degree of specialist versus generalist traits were determined as the mean number of substrates that genospecies were able to metabolise and the mean metabolic capacity was measured by calculating the Average Well Colour Development (AWCD) across substrate wells for each genospecies.

On average OE strains were found to metabolise significantly more single substrates than OA strains (Kruskal-Wallis X^2 = 7.940, df = 3, p-value < 0.05; Dunn's post hoc p < 0.05). Furthermore, OE strains also displayed higher metabolic capacity than OA and CC strains (Figure 6.7a; Kruskal Wallis: X^2 = 11.152, df = 3, p < 0.05; Dunn's post hoc:

adjusted $p < 0.05$), while no significant difference was observed between overall metabolic capacity of CC and OC strains. However, metabolic similarity (calculated by Euclidean distance of 31 single substrate treatment OD values) did not strongly correlate to genetic similarity (Mantel R statistic = -0.2747, $p > 0.05$).

To further determine whether strain metabolism profiles clustered by genospecies groups, the metabolic capacities for single substrate treatments for each strain was averaged into 6 resource type groups and assessed by principal components analysis (PCA): amines, amino-acids, carbohydrates, carboxylic acids, complex carbons and phosphate carbons (Table 6.2). Genospecies and farm treatment groups were found to overlap across the first two principal components, explaining 76% of the total variance, however differences between genospecies groups were identified (Figure 6.7c; Appendix Table E.9; PERMANOVA: $F_{3,22} = 3.8293$, $p < 0.01$). Specifically, OE was found to significantly differ in metabolic capacity to CC (PERMANOVA post hoc adjusted $p < 0.05$). The separation of OE strains corresponds to their increased metabolic capacity across the single substrate treatments compared to other strains, particularly for amino acid metabolism, carbohydrates and phosphate carbon (Figure 6.7d). Similarly, PCA of individual substrates did not separate genospecies groups, and Glycyl-L-Glytamic Acid (amino acid group) and Tween 40 (Complex carbon sources) contributed most to variance of PC1 and PC2, respectively (Appendix Figure E.7).

As supernatant contents were largely associated with nutrient depletion, the relatively high metabolic capacity of OE strains would suggest that these strains act as generalists and were therefore able to access a greater proportion of nutrients left behind in supernatant, to facilitate their growth. However, no significant correlation was observed between growth of strains grown in 100% TY broth after 62 h growth and metabolic capacity for any of the 31 single substrate treatments (Pearson's Correlation Coefficient R < ±0.34 , $p > 0.05$) or 6 resource type groups (Pearson's Correlation Coefficient R < ±0.22, $p > 0.05$). The difference in observed metabolic capacity of the sampled OA and OE strains substantially influenced the resulting positive correlation between RGI and resource utilization and therefore should be noted with caution (Figure 6.7b; Pearson's correlation R = 0.55, $p > 0.01$). Consequently, while resource competition between generalists and specialists may have contributed to the observed indirect competitive interactions (Figure 6.2a), it is likely that other inhibitory mechanisms are driving interactions (e.g. direct inhibition competition).

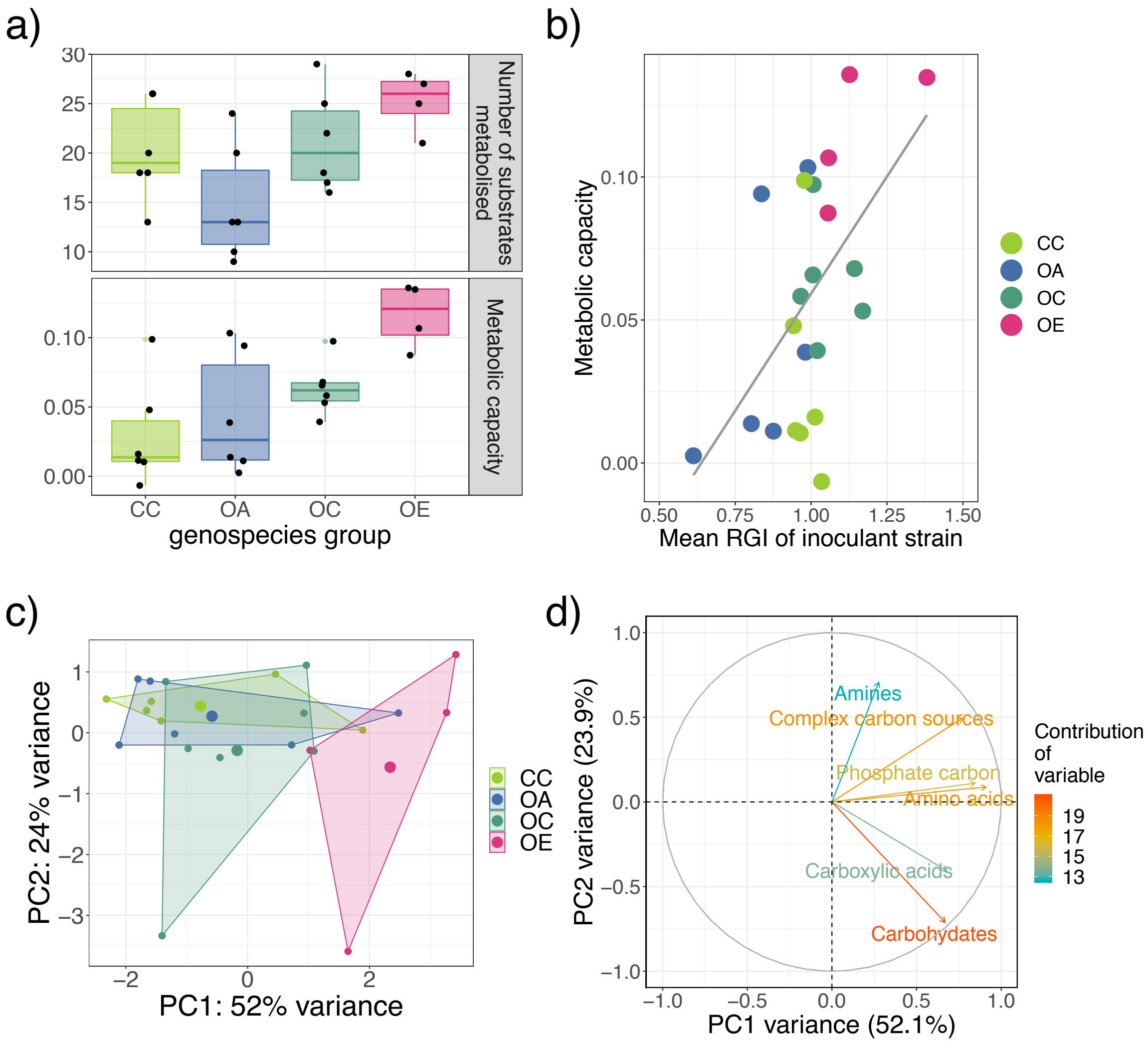


Figure 6.7 Metabolic differences between genospecies groups. **a)** Number of substrates metabolised and metabolic capacity based on Average Well Colour Development was calculated across 31 single substrate growth treatments for 23 *Rlt* strains grouped by their genospecies and environmental origin. * $p < 0.05$. **b)** Average Relative Growth Index (RGI) of inoculant strain grown in all other supernatant treatments correlated to the strain's ability to metabolise the 31 carbon substrates (metabolic capacity as a measure of Average Well Colour Development). (Pearson's correlation coefficient R = 0.55, $p > 0.01$). Regression line is fit by linear model. **c)** Principal Components Analysis for metabolic capacity of 31 single substrate treatments averaged across 6 resource type groups showed OE strains separated from the other genospecies. Points represent *Rlt* strains and are coloured by genospecies and environmental origin. Spread of the strains indicates phenotypic variation amongst resource type groups. **d)** The association of the 6 resource type group variables to the first two principal components. Resource type groups are coloured by their percentage contribution of the total variance for principal components 1 and 2. Individual substrates within each resource type group can be found in Table 2, methods section.

6.4.3. Using comparative genomics to identify potential mechanisms underlying competitive differences between rhizobial strains

To explore the underlying mechanisms behind indirect and direct competitive interactions, the presence of genes associated to 1) bacteriocins, 2) secondary metabolite clusters, and 3) prophages were searched for. Additionally, the presence of genes associated with 4) quorum sensing pathways were also analysed, as quorum sensing signals can be linked to both facilitative and inhibitory interactions between strains.

The presence of known bacteriocins were searched for in the 24 strains (Figure 6.8; Additional File 6: Table S7). BLASTn was used to search for the *medium Rhizobium* bacteriocin. The *medium* bacteriocin was identified in all 24 strains with a percentage sequence identity between 92.76% - 95.17% to the *medium* bacteriocin reference sequence. OA strains were found to have a slightly higher percentage sequence similarity to the reference sequence compared to OE strains (Figure 6.8; Additional File 6 Table S7). Additionally, bacteriocin trifolitoxin genes were only found in OE strain SM135B, but only the putative immunity protein *tfxG* was identified. This suggests that while SM135B might be immune to the effects of trifolitoxin, the strain was unlikely to produce the toxin.

Other known *Rhizobium leguminosarum* quorum sensing AHL synthase genes (*cinI*, *raiI*, *rhiI*, and *traI*) and their related regulatory genes were searched for in the 24 strains (Figure 6.8; Additional File 6: Table S7). Not all quorum sensing pathways were present in the 24 strains. However, *cinI* and *cinR*, which encode the *small* bacteriocin AHL (3OH-$C_{14:1}$-HSL) synthase and its transcriptional regulator, were both found in all 24 strains with high percentage identity to the reference sequence (Figure 6.8). The *raiI/raiR* quorum sensing system was found to be the second most common (in 50% of strains). Presence of both *raiI* and *raiR* together were only found in genospecies C strains (OC/CC), which also contained *rhiI/rhiR* system, which was absent from other strains. The *raiI/raiR* pathway has been suggested to be involved in generation of short chain AHLs and to be some extent functionally redundant with *rhiI/rhiR* pathways (Wisniewski-Dyé and Downie, 2002). Genes *expR* and *cinS*, required for *raiR* expression, were present in all 24 strains (Figure 6.8). Additionally, a greater variation in gene content and percentage identity was observed between organic isolates than conventional isolates (Figure 6.8). Furthermore, three inhibition zone susceptible OA strains (SM144A, SM154C, SM145B) contained *traI, traR and bisR* genes that were only identified together in these strains (Figure 6.8). Presence of *traI*, *traR* and *bisR* have been shown to greatly increase strain sensitivity to the *small* bacteriocin (3OH-$C_{14:1}$-HSL), as the AHL products of the pathway in combination with detection of *small* can further mediate growth sensitivity (Wilkinson *et al.*, 2002).

All genomes were also screened for the presence of secondary metabolite biosynthesis gene clusters (Figure 6.8). For each strain, the number of gene clusters were identified and totalled for Type III polyketide synthases (T3PKS), Bacteriocin, Terpene, Arylopolyene, Ectoine, Homoserine lactones, Proteusin, Pheganomycin-style protein ligase-containing cluster, Non-ribosomal peptide synthetase (NRPS) and NRPS-like clusters (Appendix Table E.10). Homoserine lactones were the most abundant secondary metabolite biosynthesis gene clusters in all strains. Bacteriocin and Proteusin gene clusters were not found in CC strains but were present in at least one strain of all other genospecies groups. Additionally, OA strains lacked NRPS clusters, which were identified in all other genospecies groups. All OE strains contained an NRPS cluster with 100% identity to the siderophore, Vicibactin (NCBI GenBank: CP000138.1), that can be used by strains to sequester iron in the rhizosphere (Heemstra, Walsh and Sattely, 2009; Wright *et al.*, 2013). On the other hand, all but one CC strain contained an NRPS cluster for the rhizosphere-expressed *rhiABC* operon of undetermined function, which is regulated by the OC/CC strain exclusive *rhiI/rhiR* pathway (NCBI GenBank: NC_014718.1) (Cubo *et al.*, 1992; Rodelas *et al.*, 1999).

Finally, putative prophage regions were identified in genomes (Figure 6.8). In total, an intact prophage region was detected in two OA strains, along with one strain each from OC, OE and CC groups. Additionally, multiple prophage regions were identified as questionable or incomplete within each genospecies group, with CC strain genomes containing the largest number of totalled questionable and incomplete prophage regions. However, overall the number prophage regions did not correlate to suppressive or facilitative ability (Appendix Table E.11).

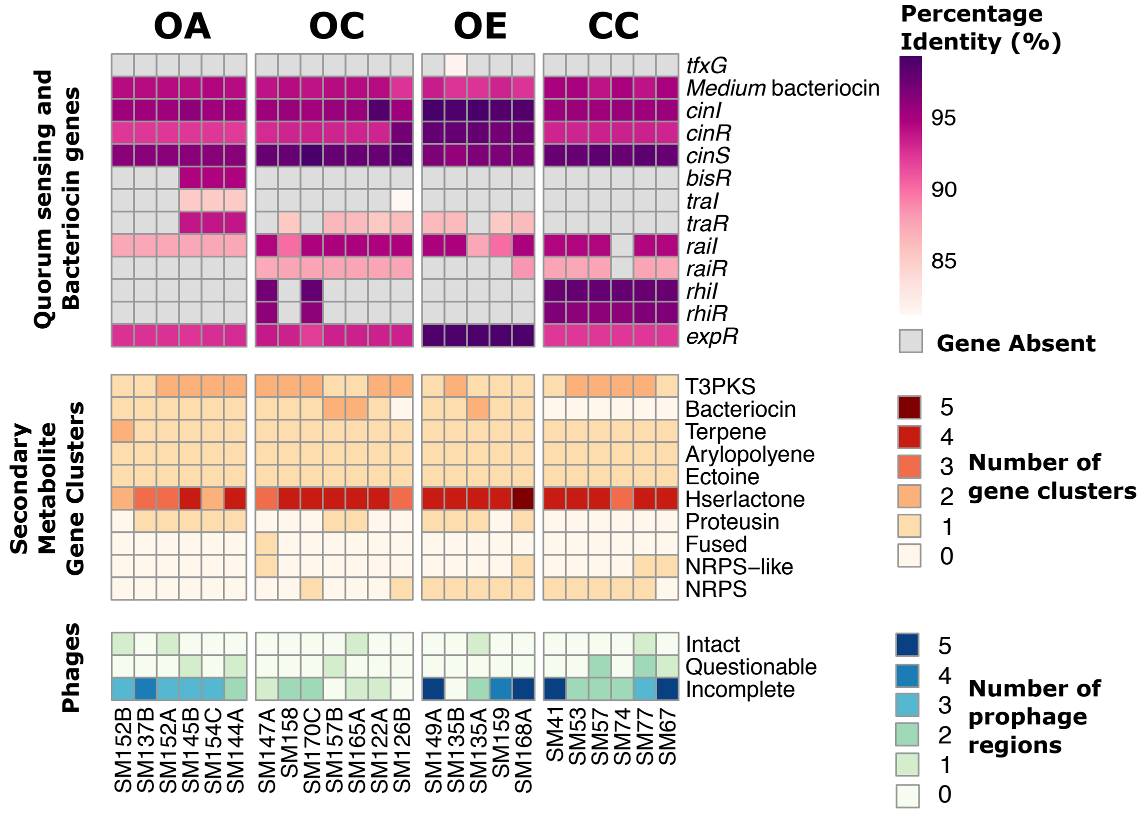


Figure 6.8 Percentage identity of *Rhizobium leguminosarum* bacteriocins and quorum sensing associated genes, and the number of secondary metabolite gene clusters and phages, found in 24 *Rlt* strains. Quorum sensing and bacteriocin genes heatmap colours correspond to increasing percentage identity of quorum sensing gene reference sequences to identified regions in each genome. Grey boxes highlight genes that were not present in a specific genome. Secondary metabolite gene cluster heatmap colours correspond to the number of gene clusters identified for each type of gene cluster. Phage heatmap colours correspond to the number of prophage regions of either intact, questionable or incomplete quality identified in each genome. Strains are clustered according to their genospecies environmental origin; OA = organic genospecies A, OC = organic genospecies C, OE = organic genospecies E, and CC = conventional genospecies C. Accession numbers for quorum sensing associated gene reference sequences can be found in **Table 6.3**.

6.5.Discussion

This study aimed to investigate the intraspecific facilitative and inhibitory interactions of *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) strains in terms of indirect exploitative (resource) competition and direct interference competition (growth inhibition). Significant variation in competitive ability was observed between genospecies. When strains were grown in each other's cell-free supernatants, OE strains disproportionately negatively inhibited OA strain growth, and significantly influenced specific strain interactions. While both facilitative and inhibitory interactions were identified, strains that produced more inhibitory supernatants tended to grow well in other strain supernatants on average. Conversely, strains that had particularly facilitative supernatants grew less well in the supernatants of the other strains. This association could also potentially be a result of resource competition between specialists and generalists. The direct inhibitory activation and growth of OE strains could also be triggered by quorum sensing interactions with other strains, as all strains were found to contain multiple homoserine lactone biosynthesis gene clusters. The synergistic effects of multiple quorum sensing signals could contribute to the observed variation of interactions, as suggested by the identification of the *traI*/*traR*/*bisR* pathway in highly inhibited OA strains. Additionally, OE strains were found to contain Vicibactin siderophore synthesis genes which could provide some competitive advantage to sequester resources from neighbouring strains. Together these results suggest that, intraspecific competitive abilities vary largely between rhizobia strains, and this was observed at both genospecies and genotype levels. In the field, intraspecific competition is a potentially important factor shaping symbiotic specificity, in addition to plant-mediated selection, where the most competitive strains have a greater chance of establishing symbiosis with the plant.

6.5.1. Supernatant-mediated interactions can be facilitative and inhibitory

Indirect inhibitory and facilitative supernatant interactions were significantly influenced by genospecies effects and was predominantly driven by the inhibition of OA strains by OE strains. This is in line with previous studies that have associated *Rhizobium* signalling molecules found in cell-free supernatant with growth inhibition and symbiosis establishment, although the presence of these interactions were not investigated between genospecies (Bladergroen, Badelt and Spaink, 2003; Sanchez-Contreras *et al.*, 2007; Checcucci *et al.*, 2017). For example *Rlv* strain RBL5523 supernatant was found to suppress nodulation and nitrogen fixation of strain RBL5787, through secretion of temperature-sensitive proteins responsible for infection thread formation (Bladergroen, Badelt and Spaink, 2003). Interaction differences between genospecies groups observed in this study are interesting, as exclusive phenotypic distinction between genospecies has not yet been observed (Kumar *et al.*, 2015). These results suggest that maintenance of intraspecies diversity and genospecies groups could be linked with intraspecific strain interactions. Despite genospecies interaction effects, overall the competitiveness and cooperativeness of interactions varied largely across all pairwise strain combinations and depended on interactions between specific genotypes (Figure 6.2). Additionally, more closely related species were not observed to show more inhibitory interactions towards one another (Appendix Figure E.5), as shown in previous studies using multiple different bacterial species (Becker *et al.*, 2012). It is thus possible that *Rlt* competitive interactions are largely driven by variation in accessory genome content, that extensively varies between and within genospecies (Crossman *et al.*, 2008; Kumar *et al.*, 2015; Cavassim *et al.*, 2019).

Farming practice (gsC organic or conventional farm isolates) had only very small effects on indirect competitive interactions. This is in contrast with previous studies suggesting that industrialised farming managements could influence rhizobial population sizes, diversity and subsequently legume root nodulation (Graham and Vance, 2000). Another study found that soybeans inoculated with soil from conventional industrial farm sites had a lower biological nitrogen fixation turnover compared to when inoculated with soil from organic treatments (Schmidt, Weese and Lau, 2017). These soils could have also contained a higher number of other rhizosphere microbes, which is another added complexity to the functionality of community interactions influencing symbiotic productivity.

6.5.2. Growth inhibition was observed in direct interaction

Clear patterns of direct inhibition were observed at both genospecies and genotype levels. OE was the most capable of producing inhibition zones across strains and produced inhibition zones on all but one OA strain, which further demonstrated their ability to inhibit OA strain growth in a direct capacity as well as indirectly through supernatants. On the other hand, gsC seemed to be the most resistant to inhibition zones, but their susceptibility to inhibition zones and their ability to produce inhibition zones on average did not differ between conventional and organic farming treatments (Figure 6.2b). Furthermore, the ability to produce inhibition zones, and the size of those inhibition zones, was found to vary at the level of individual strains. While OA was found to be the most susceptible to inhibition zone formation by all other genospecies, this effect was predominantly caused by three OA strains (SM145B, SM154C and SM144A) that had the largest inhibition zones with diameters ranging from 7.52 mm to 19.85 mm. Direct inhibitory rhizobia interactions displayed through inhibition zone production have also been observed in previous studies (Hirsch, 1979; Oresnik, Twelker and Hynes, 1999; Lithgow *et al.*, 2000; Wilkinson *et al.*, 2002; McAnulla *et al.*, 2007). Rhizobia have similarly been shown to produce a range inhibition zone sizes when plated on agar lawns of other rhizobial strains, and this has been suggested to be a consequence of the production of different bacteriocins and activity of quorum sensing associated mechanisms (see section 4.3.2 below) (Hirsch, 1979; Joseph, Desai and Desai, 1983; Schripsema *et al.*, 1996; Wilkinson *et al.*, 2002; Joshi *et al.*, 2008). However, there has otherwise been limited investigation of direct pairwise competition between natural rhizobia isolates on such a scale (Hirsch, 1979), with previous comparisons predominantly focusing on less than 10 strains.

Together, genospecies groups were found to differ in competitiveness, which was largely driven by both direct and indirect suppressive effects of OE supernatants on OA strain growth (Figure 6.2). Furthermore, OA strains were consistently suppressed by OE strains even though all strains were originally isolated from across different Danish farm sites. However, overall the correlation between negative inhibitory effects observed in the supernatant assays and direct interference competition in the soft agar assays varied largely between *Rlt* strains. Strain combinations that displayed indirect suppressive interactions in supernatant did not necessarily produce inhibition zones when in direct competition, such as with strain SM126B. This further supports the theory that both exploitative (indirect) and interference (direct) competition occur between rhizobia strain combinations, and the extent to which either or both are utilised is determined at the strain level. While it is well known that rhizobia strains, and bacterial species in general, interact through both direct and indirect mechanisms (Hibbing *et al.*, 2010; Checcucci *et al.*, 2017), there has been limited investigation into whether strains that are successful indirect competitors are also good direct competitors.

6.5.3. Underlying mechanisms behind intraspecific competitive interactions

6.5.3.1. Indirect competition could be partially mediated by resource competition

No correlation between metabolic similarity and genetic relatedness was found in this study, and despite some genospecies level patterns, even closely related individual genotypes varied considerably in their competitive effects within genospecies (Figure 6.2a). Previous research has suggested that more genetically related strains will exert relatively stronger competition towards each other through resource competition, because they are more likely to share similar metabolic pathways, and exhibit a resource niche overlap (Russel *et al.*, 2017). Such relationship was not found in this study, potentially because a complex TY medium was used to grow strains, which may have masked niche differences that are visible when growing strains in a more resource restricted media.

Additionally, and as also shown in other studies, no metabolite was found to be exclusively metabolised by a single genospecies; this supports the overall influence of individual strain interactions being the dominant signal in this analysis (Appendix Figure E.4) (Kumar *et al.*, 2015). Similar to previous research, all strains used in this study showed the greatest phenotypic variance in amino acid metabolism and general carbohydrate metabolism, such as sugars (Wielbo *et al.*, 2010), but also phosphate carbons which has not been shown previously (Figure 6.7c-d). OE strains were able to metabolise the greatest number of single substrates and displayed a high metabolic capacity across 31 single substrate treatments. This result supports the hypothesis that OE strains can efficiently deplete nutrients in the complex TY media, acting as generalists, and subsequently produce supernatants that are poor at supporting the growth of other strains. Correspondingly, OA strains showed much lower metabolic capacity and metabolised fewer substrates in general, suggesting they may display more specialist characteristics and leave more nutrients unutilised in the supernatant, which could facilitate growth of other strains. However, OA strains grew even worse in OE supernatants (1:1 supernatant:100% TY) than 50% TY controls (Appendix Figure E.6), which suggests that inhibitory mechanisms other than purely resource competition are being observed. This is because theoretically strains should grow better in supernatant treatments than 50% TY controls as supernatant treatments contain an equal amount of nutrients to 50% TY control treatments plus any additional nutrients left behind in the supernatant.

All five OE strains were found to contain NRPS cluster orthologs with 100% identity to Vicibactin siderophore production, whereas NRPS clusters were absent in OA strain genomes (Appendix Table E.10). Vicibactin can be used by *R. leguminosarum* strains to sequester iron from rhizosphere environments and is associated with productive symbioses, as iron is vital for successful nitrogenase function (Heemstra, Walsh and Sattely, 2009; Geetha and Joshi, 2013; Wright *et al.*, 2013). This sequestering of iron from other non-producer *Rlt* strains provides siderophores with antimicrobial-like qualities, as it prevents the growth of other strains without the ability to compete with their own siderophores (Kramer, Özkaya and Kümmerli, 2019). This could further suggest that OE strains are capable of repressing strain growth through secondary metabolite secretion which enables resource competition for essential molecules such as iron, thereby increasing their competitive advantage.

6.5.3.2. Direct competition is likely mediated by secondary metabolites

Rhizobia are also known to produce bacteriocins, antibiotics and lysogens that can inhibit growth and nodulation of other rhizobia (Triplett and Sadowsky, 1992; Jousset *et al.*, 2011). *cinI/R* are said to be the 'master regulators' of downstream AHL quorum sensing pathways in rhizobia, including regulating production of the *small* bacteriocin (3OH-$C_{14:1}$-HSL), *rhiI/R, traI/R* and *raiI/R* pathways (Lithgow *et al.*, 2000; Wisniewski-Dyé and Downie, 2002; Sanchez-Contreras *et al.*, 2007). AHL quorum sensing pathways are found across all *R. leguminosarum* symbiovars, and producers are immune to its effects (Hirsch, 1979). All 24 strains in this study contained *cinI/R* (Lithgow *et al.*, 2000), but similar to previous studies large variation in the presence of downstream inducer and regulator quorum sensing genes, such as *rhiI/rhiR, raiI/raiR* and *traI/traR* likely contributed to the differences in facilitative and inhibitory interactions (Wisniewski-Dyé and Downie, 2002).

Both inducer and regulator genes for *rhi* and *rai* quorum sensing pathways were found to only be present in gsC strains (CC and OC) (Figure 6.8). These two pathways display some level of redundancy as they can produce some of the same AHLs (Rodelas *et al.*, 1999). While these pathways previously have mainly been found in *Rhizobium etli, Rhizobium leguminosarum* symbiovar *viciae* and *phaseoli* strains, in this study these pathways were also found in *Rlt* gsC strains (Wisniewski-Dyé et al., 2002; Edwards et al., 2009; Downie, 2010). The function of the *raiI/raiR* pathway is unknown, but *raiI* can be activated by other quorum sensing AHLs determined from symbiosis plasmid-bound genes, suggesting a potential association to symbiotic capability (Wisniewski-Dyé et al., 2002). On the other hand, the *rhiI/rhiR* pathway can be induced by the *small* bacteriocin (3OH-$C_{14:1}$-HSL), and regulates the *rhiABC* operon which was also found in the *Rlt* gsC strains in this study (and previously only found in

Rhizobium leguminosarum symbiovar *viciae* strains) and has been suggested to influence nodulation efficiency (Cubo *et al.*, 1992; Rodelas *et al.*, 1999; Wisniewski-Dyé and Downie, 2002). The *raiI/raiR* pathway AHLs have also been suggested to function as redundant regulators of the *rhiABC* operon, along with the *rhiI/rhiR* pathway (Rodelas *et al.*, 1999). The reason for the redundancy of quorum sensing systems in rhizobia remains unclear but could potentially provide some resilience to inhibitory quorum sensing mechanisms imposed by other rhizosphere bacteria.

The greater susceptibility of OA strains (SM144A, SM154C, SM145B) to inhibition by other strains was likely due to the presence of quorum-sensing *traI*, *traR* and *bisR* genes, which increases strain sensitivity to 3OH-$C_{14:1}$-HSL (*small* bacteriocin) as the *traI* AHL products in combination with detection of *small* can further mediate growth sensitivity (Wilkinson *et al.*, 2002; McAnulla *et al.*, 2007). The extensively studied *small* bacteriocin, which was later discovered to be an AHL, is known to produce inhibition zones more than 10 mm and on average 25 mm, with no bacteriocin activity detected in cell-free culture supernatants of producer strains (Hirsch, 1979; van Brussel *et al.*, 1985; Gray *et al.*, 1996; Schripsema *et al.*, 1996; Wilkinson *et al.*, 2002; Wisniewski-Dyé and Downie, 2002). This is similar to the inhibition zones observed in this study, whereas *medium* bacteriocin-producing strains would display smaller inhibition zones < 10 mm with bacteriocin activity and would also be detectable in supernatants (Hirsch, 1979). The *traI*/*traR*/*bisR* genes are also involved in regulating recipient-induced symbiosis plasmid transfer through quorum sensing (Wilkinson *et al.*, 2002; Danino *et al.*, 2003; McAnulla *et al.*, 2007). *traI* is a LuxI-type protein, induced by *traR*, encoding an AHL synthase catalysing the synthesis of AHLs that act as diffusible quorum sensing signals (Hwang *et al.*, 1994; Wilkinson *et al.*, 2002; He *et al.*, 2003; McAnulla *et al.*, 2007; Lang and Faure, 2014). Primarily, *traR* expression is induced by BisR in donor strains containing the symbiosis plasmid in response to CinI AHLs made by recipient strains, such as the *small* bacteriocin (3OH-$C_{14:1}$-HSL) (Wisniewski-Dyé and Downie, 2002; Danino *et al.*, 2003; McAnulla *et al.*, 2007). In addition, the transcriptional regulator *traR* can also be activated once a threshold concentration of *traI*-produced AHLs is reached (Wilkinson *et al.*, 2002; McAnulla *et al.*, 2007; Lang and Faure, 2014). Therefore, strains carrying the full *bisR*/*traR*/*traI* pathway are likely to confer greater sensitivity to 3OH-$C_{14:1}$-HSL producing rhizobial strains. This demonstrates that growth sensitivity can be mediated by a combination of quorum sensing pathways (Wilkinson *et al.*, 2002) and could be advantageous for regulating growth in particularly crowded rhizospheres of nodulated legumes and for increasing symbiotic capacity of communities through conjugal plasmid transfer (He *et al.*, 2003; Downie, 2010).

The combination of potentially indirect and direct competition displayed by OE strains could make them strong rhizosphere competitors. The most competitive strains in the rhizosphere are likely those capable of both outcompeting other strains for resources and also producing direct inhibitory metabolites to regulate growth of closely related neighbouring strains, which could provide some selective advantage for rhizosphere persistence and symbiotic establishment (Schwinghamer and Brockwell, 1978). Therefore, in addition to resource competition between strains, it could be suggested that additional direct repression of strain growth through quorum sensing AHLs and bacteriocin activity also plays a role in the variation of *Rlt* strain interactions (Schripsema *et al.*, 1996; Wisniewski-Dyé and Downie, 2002).

6.5.4. Study limitations and future research

Observing interactions between *Rlt* strains in conventional TY media is not necessarily applicable to the soil rhizosphere where strains would naturally interact. Bacterial signalling is achieved through multiple mechanisms, some of which were discussed here, and all of which are strongly influenced by the environmental context and associated microbial community structure (Checcucci *et al.*, 2017). For example, gsC strains were identified as the only genospecies in this study to contain both *rhiI* and *rhiR* quorum sensing gene orthologs, which have previously only been found in *Rlv* strains and may provide some unknown competitive advantage in the rhizosphere (Gray *et al.*, 1996; Wisniewski-Dyé and Downie, 2002; Sanchez-Contreras *et al.*, 2007).

Furthermore, all strains were originally isolated from farm sites across Denmark reflecting only a small fraction of rhizobial strain variation within one country. Therefore, it would be interesting to see whether these overall genospecies interactions would be maintained at an international level, with strains isolated from other continents. Additionally, larger sample sizes for genospecies and farm treatment groups would enable more robust and meaningful comparisons.

Similarly, no differentiation was made between metabolites or resource concentrations in the supernatant treatments, which would have otherwise provided additional insight to understanding the exact mechanisms causing the observed interactions. In order to also confidently confirm individual strain utilisation profiles, additional replicates would be required in future (Appendix Figure E.4). Future investigations could use mass spectrometry analyses and transcriptomics to measure gene activity and to identify different molecular compounds secreted into supernatants. Another crucial area for future investigation would be to induce targeted mutations in the quorum sensing pathway genes that are putatively involved in the direct interactions observed within this study. This could then confirm whether the specific quorum sensing mechanisms are influencing these direct strain interactions.

Strains were not grown in co-cultured environments due to the inability to distinguish strain densities from one another based on selective plating. Recent advancements in sequencing technologies and utilising unique strains for individual identification (Chapter 1; Fields *et al.*, 2019) or strain ID tagging (Mendoza-Suárez *et al.*, 2020) will be crucial for transferring study of these interactions into larger microbial communities with a more applicable *in planta* context. Future investigations hope to use these interactions to understand how diversity within pairwise communities can affect symbiotic effectiveness at a scaled-up multi-strain community level. A key theory would also be to test if more competitive rhizobia are more likely to form symbiosis when in direct competition. However, this may prove challenging due to the complexity of higher-order effects of multi-strain interactions (Barrett *et al.*, 2015).

6.5.5. Conclusions

Host interaction with the abiotic soil environment has been suggested to maintain functionally diverse *Rhizobium* communities composed of both generalist and specialist strain types (Vuong, Thrall and Barrett, 2017). However, intraspecific interactions between strains within microbial communities can also shape species diversity. In this study, significant variation was found in the competitive ability of *Rlt* strains at both the individual strain level, and between genospecies. These interactions could be partially explained by resource competition between specialists and generalists as supported by genospecies and genotype-specific metabolic profiles. However, secreted compounds also likely played an important role in addition to resource utilisation, including secretion of quorum sensing molecules, bacteriocins and growth-inhibiting secondary metabolites. The inhibitory interactions between rhizobia strains could potentially have a detrimental effect on symbiosis if it leads to a reduced likelihood of symbiotic establishment. It is therefore vital to consider these interactions when considering compatible strain combinations for agricultural inoculants in order to avoid conflict with other co-inoculant strains or with the existing native rhizobial strains.

Chapter 7. General Discussion

7.1. Introduction

This chapter provides an overview and synthesis of all the thesis results and discusses them in the context of the three central research questions, and their significance and contribution to broader knowledge in the rhizobia research field.

The overall purpose of this PhD project was to determine the extent of *Rhizobium leguminosarum* symbiovar *trifolii* (*Rlt*) intraspecies diversity at the genetic and phenotypic levels, with particular focus on identifying differences between *Rhizobium leguminosarum* genospecies. The three specific questions were to:

1) Determine if the diversity of *Rlt* populations can be explained by the selective differences of white clover genotypes
2) understand if *Rlt* genetic diversity manifests itself in the gene expression profiles and growth phenotypes of strains between and within genospecies;
3) identify whether intraspecific *Rlt* interactions can be determined by genetic differences between genospecies and environmental origins of strains.

The wider implications of the findings and avenues for future research are highlighted. The chapter ends the thesis with a general conclusion of the research provided.

7.2. Influence of white clover genotype selectivity on Rlt populations

It was first posed whether different white clover cultivars symbiotically select for different rhizobia strains in the field. If so, this could highlight that symbiotic specificity extends beyond the interspecies level and is also important at the intraspecies level of symbiotic interactions. Additionally, differences in clover genotype symbiotic selectivity could partially explain the large intraspecies diversity of *Rlt* as populations evolve by adaptive evolution to symbiotic engagement with different cultivars. This knowledge could then be utilised to aid development of rhizobial inoculant x clover cultivar combinations that are optimally genetically compatible for improved agricultural use.

To answer this question, the MAUI-seq multiplexed high throughput amplicon sequencing (HTAS) method was developed to enable confident evaluation of intraspecies diversity from environmental DNA samples based on two core and two accessory genes. This was achieved by decontaminating amplicon sequence data of chimeras and other amplification and sequencing errors that were identified using unique molecular identifiers (UMIs). Few studies have used HTAS for intraspecies ecological diversity studies (Kinoti *et al.*, 2017; Poirier *et al.*, 2018), and this is likely because there are significant concerns that genuine allelic sequence variation within highly genetically similar DNA samples cannot be confidently distinguished from incurred sequencing PCR errors. MAUI-seq implements global UMI-based error rates to detect and correct for chimeras and other erroneous PCR artefacts. Multiple amplicons (housekeeping genes, *rpoB* and *recA*, and symbiosis genes, *nodA* and *nodD*) were used to discern intraspecies diversity as 16S rDNA amplicons are too highly conserved to sufficiently determine intraspecies sequence diversity (Gaunt *et al.*, 2001; Case *et al.*, 2007; Adékambi, Drancourt and Raoult, 2009; Vos *et al.*, 2012; Poirier *et al.*, 2018). The method was validated using known synthetic rhizobial DNA mixtures and environmental white clover nodule samples and was found to perform more robustly compared to established amplicon sequence variant clustering methods, DADA2 and UNOISE3 (Figure 2.3 and Figure 2.4).

The validated MAUI-seq method was then used to aid determination of whether five white clover genotypes contained significantly different *Rlt* nodule populations based on *rpoB, recA, nodA and nodD* allele frequencies when grown under field conditions. Several clover genotypes were found to display significantly different *Rlt* diversity (Figure 3.4), however the level of observed intraspecies diversity was influenced by the candidate gene and whether diversity was evaluated at the level of individual genes or their combinations. Furthermore, a large amount of sequence variation was observed within clover genotypes for all four genes (Figure 3.3). *rpoB* and *recA* alleles displayed the greatest distinction between clover genotypes but diversity was also associated with geographic distance between samples in the field rather than solely resulting from host-filtering by the plant. The combined effect of plant genotype and geospatial variation in allele frequencies has similarly been shown by other studies

(Aouani et al., 1997; Nleya, Walley and Vandenberg, 2001; Fagerli and Svenning, 2005; Argaw and Muleta, 2017; Liu et al., 2019), as well as with the samples originally used to validate MAUI-seq (Figure 2.4). This is likely due to differences in local geographic conditions influencing the initial microbial rhizosphere community which is then further selected by the legume genotype (Vuong, Thrall and Barrett, 2017; Liu *et al.*, 2019). Symbiosis genes on the other hand showed no significant distinction between clover genotypes and similarly showed no association to geographic distance between samples. This was surprising as influence of potential nodulation-based symbiotic selection was expected to reflect in differing diversity of *Rlt* nodulation genes between clover genotypes. However, as the genotypic differences between clover F2 crosses were unknown, this could potentially be due to a lack of variation in the mechanisms determining symbiotic selectivity between crosses. Furthermore, without determining the intraspecies diversity of the rhizospheres it remains unclear how much the diversity of *Rlt* nodule populations differ from the initial rhizosphere pools for each sample.

Rhizobia from soil, rhizosphere and nodules of different legume species have been characterised using PCR-restriction fragment length polymorphism (RFLP) of 16S-23S ribosomal DNA intergenic spacers, *nodD* amplicon sequencing, and insertion sequence typing (Bromfield, Barran and Wheatcroft, 1995; Laguerre *et al.*, 2003; McGinn *et al.*, 2016). These methods are useful to detect sub-species and strain level differences, along with repetitive extragenic palindromic polymerase chain reaction (rep-PCR) fingerprinting using enterobacterial repetitive intergenic consensus (ERIC) primers (McGinn *et al.*, 2016). HTAS is also a popular method for microbial community diversity analyses, however sequencing errors introduced during amplification and sequencing poses challenges even for interspecies studies. A main advantage of MAUI-seq over other established HTAS clustering methods (DADA2 and UNOISE3) is that sequences are determined as genuine using UMI-based error rates, rather than rejecting sequences based on their similarity to other sequences in the dataset. This makes the MAUI-seq method different to DADA2 and UNOISE3, where sequences are classed as chimeras if they can be created as recombinants of other sequences. Especially for intraspecies analysis, this risks generating false-positive rejections of genuine alleles in datasets containing sequences with high sequence similarity (Edgar, 2016a). Future improvements to MAUI-seq could involve: 1) utilisation of a statistical model to determine the appropriate secondary/primary sequences ratio threshold for detecting chimeras rather than using a predefined threshold; 2) using recognisable target sequence spike-in controls during initial sample processing in order to determine absolute abundance of sequence alleles; 3) utilising longer amplicon sequences in line with the development of sequencing technologies; 4) implementing UMIs at both ends of the amplicon region as an additional confirmation of chimera detection (Burke and Darling, 2016).

Another advantage of MAUI-seq is that the approach allows assessment of intraspecies diversity using multiple gene amplicons as it was shown that observed intraspecies diversity can be influenced by choice of gene candidates (Figure 3.4). Using multiple amplicons to determine *Rlt* diversity enabled assessment of intraspecies diversity from the perspective of both horizontal (plasmid-bound *nodA* and *nodD*) and vertical (chromosomal-bound *rpoB* and *recA*) gene transmission. However, this also emphasised how the gene markers must be chosen with careful consideration to the research questions in mind. For example, symbiotic specificity can also be influenced by extracellular polysaccharide production, plant-identification of rhizobial secretion systems and detection of microbe-associated molecular patterns. Genes associated with these molecular interactions were not evaluated in this study and could have potentially displayed differences in selection by clover genotypes (Perret, Staehelin and Broughton, 2000; Simms and Taylor, 2002; Wang, Liu and Zhu, 2018). These genes could be tested in future to evaluate their association with intraspecies symbiotic specificity. Additionally, as the MAUI-seq method is applicable to any type of environmental sample, it would have been insightful to evaluate the rhizosphere soil community compositions to determine if *Rlt* rhizosphere populations differed between clover genotype samples. Furthermore, MAUI-seq has the potential to be used for other types of amplicon diversity studies, and in this case could be further used to monitor the general soil community diversity based on 16S and ITS gene regions.

Together, these studies aimed to evaluate whether different clover cultivars could select for significantly different *Rlt* genotypes and the results could be used to aid development of more productive clover inoculants that are matched with plant genotypes. For example, future analyses could focus in pure check clover varieties (and other agriculturally important cultivated legumes (Stagnari *et al.*, 2017)) to determine if specific clover varieties preferentially select for different *Rlt* genotypes.

If so, this would suggest there is significant benefit to developing highly genetically compatible rhizobia inoculants based on the legume genotype.

7.3. Transcriptomic and phenotypic intraspecies diversity of *Rlt*

The species of *Rhizobium leguminosarum* is highly genetically diverse and contains a species complex including at least five genetically distinct genospecies (Kumar *et al.*, 2015; Cavassim *et al.*, 2020). Despite the significant genomic differences between genospecies, no phenotypic traits have been exclusively associated to a single genospecies (Ravin, 1963; Kumar *et al.*, 2015; Smith, 2018). It is unclear to what extent transcriptional variation between and within genospecies is also evident. If bacterial genetic distance evidently influences gene expression, analysis of transcriptional differences between genospecies could highlight the advantages of using transcriptional variation as a phenotypic parameter for taxonomic species distinction. Therefore, this study used a multi-strain approach to determine whether *Rlt* genetic diversity is reflected in the transcriptomes of strains, and furthermore how this transcriptomic diversity can be linked to phenotypic traits.

Although genospecies share the same core genome, both transcriptomic and phenotypic differences were identified between genospecies. Genospecies displayed differences in core genome transcriptome profiles and showed significant expression differences at the level of individual core genes (Figure 4.1a). Within genospecies, core genome transcriptome profiles were less distinct (Figure 4.1d). Following this trend, increased genetic divergence between *Rlt* strains was found to correlate with an increased number of differentially expressed core genes (Figure 4.1f). This suggests that *Rlt* genetic diversity is evident at the gene expression level of the core genome which contains essential genes shared by all strains of the species. To understand the potential functional implications of the observed transcriptional diversity, co-expressed core genes were grouped into modules and expression was correlated to growth of the phenotypes. Significant correlations were identified between the expression of several modules and phenotypic growth differences between genospecies. Core gene modules enriched with gene functions related to fundamental bacterial metabolism were found to also significantly differ in expression between genospecies (Figure 4.3). Taken together, using a multi-strain experimental design to capture the extent of species level expression variation enabled identification of groups of species-conserved co-expressed genes associated with genospecies differences in bacterial metabolism. However, as a significant number of the core genes have unknown functions, future work could further investigate the functional relevance of differing genospecies expression patterns, and also observe whether these differences are reproducible under more applicable rhizosphere-based conditions.

Variation in genome architecture and gene expression was additionally utilised to identify transcriptional units that were conserved across genospecies. This was achieved to further understand the extent genetic diversity causes differential regulation of transcription between genospecies that is observable at the operon level. Overall, 94 transcriptional units were found to be conserved across all five *Rlt* genospecies, with approximately 1000 transcriptional units identified for each genospecies individually. Therefore, differences in genome organisation and gene content have the potential to substantially change the regulatory organisation of genomes and consequent phenotypes within a single bacterial species. The use of genomic and transcriptomic data from multiple strains with differing genomic compositions offered additional verification for predicting operons and additionally highlighted the variation of operon architecture within *Rlt* which likely contributes to the substantial transcriptomic and phenotypic diversity observed across strains. This investigation has generated a new resource of putative operons from five *Rlt* genospecies using 26 strains. It would be interesting to evaluate differences in whole operon expression between genospecies at the operon level. Therefore, future work could aim to test if the 94 conserved operons show differences in expression between genospecies, which would also confirm suggested regulatory differences between genospecies. Moreover, the functional associations of operons found exclusively to specific genospecies could be further explored as well to identify enriched pathways in different genospecies.

Similar to previous studies (Kumar *et al.*, 2015), *Rlt* strains displayed large variation in growth phenotypes and resource metabolism (Figure 4.2; Figure 6.2; Figure 6.7), however no identified phenotype was exclusive to a single genospecies. Metabolic versatility has been associated with replicon diversity which could explain the lack of metabolic traits exclusive to individual genospecies and also the significant variation in metabolic capacity across strains, as plasmids containing different metabolic pathways can transfer between strains within and between bacterial species (Mazur *et al.*, 2013; Ormeño-Orrillo and Martínez-Romero, 2013; Cavassim *et al.*, 2020). While no genospecies-exclusive phenotypes have been identified, it was observed that some genospecies showed clearly distinct growth traits and varied in their ability to metabolise specific substrates (Figure 4.2; Appendix Figure E.4). For example, a selection of representative gsC strains were collectively the slowest (or equally the slowest) growing genospecies in 100% TY broth at 28°C (Figure 4.2; Appendix Figure E.1). However, growth phenotypes did not necessarily correlate with the metabolic capability. Although gsA strains collectively grew to the highest densities in 100% TY (compared to gsC and gsE), gsA strains were collectively also the least metabolically diverse when grown in single substrates (Figure 6.7a). Similarly, gsE strains on average were able to metabolise the most substrates and displayed the greatest ability to metabolise substrates but did not reach the highest densities in TY media compared to other genospecies. Moreover, transcriptomic and phenotypic distinctions between genospecies might also be evident under different environments other than the TY media conditions used extensively in this project. The advantages of growing rhizobia in complex TY media are that: 1) the bacteria will be in a metabolically active free-living physiology; 2) it avoids complications of removing plant material from samples for sequencing; and 3) it is easier to ensure growth conditions are consistent between strains for a more confident comparison of phenotypes. It has been noted that using growth conditions with a variety of substrates could reduce the potential of niche overlap between strains, and encourage transcriptional diversity through utilisation of different substrates (Vital *et al.*, 2015). While TY media may not simulate natural soil conditions, the complex medium was chosen to encourage a more active transcriptomic state that is not limited by lack of nutrients. Previous cross-species transcriptome comparisons have utilised gene expression data from various open repositories where experiments differed in experimental conditions (Stuart *et al.*, 2003; Carrasco, Tan and Duman, 2011; Kristiansson *et al.*, 2013; Hosseinkhan, Mousavian and Masoudi-Nejad, 2018). However, different environmental conditions can greatly influence transcriptome expression (Vital *et al.*, 2015; Jiao *et al.*, 2018). Therefore, the conditions strains are grown under must be comparable when doing cross-species or cross-strain analyses, as it is crucial that growth conditions are consistent in order to evaluate relevant expression patterns. As the phenotypic diversity of strains under free-living physiology were evaluated in this project, future research could use the same multi-strain approach to characterise genospecies transcriptomic and phenotypic traits under bacteroid physiologies to observe *Rlt* diversity of additional relevant phenotypes such as nitrogen fixing abilities or bacteroid metabolism.

With consideration to the wider perspective of understanding species phenotypic variation, a significant challenge for cross-species analyses is the limited representation of variation within species, particularly for transcriptome studies where direct species comparisons have commonly used only one or two isolates to represent a species (Scaria *et al.,* 2013; Kimes *et al*., 2014; González-Torres *et al.*, 2015; Vital *et al.,* 2015). The multi-strain experimental design used in this study aimed to rectify this by considering the likely variation observed within species in order to identify true transcriptional differences between genetically distinct groups of strains. For example, utilisation of this approach identified that *Rlt* genospecies have differentially expressed core genomes (Figure 4.1a-c). Transcriptomes have been considered a molecular phenotype capable of identifying initial species divergence, however the amount to which gene expression corresponds to definitive bacterial species difference is still disputed (Pavey *et al.*, 2010; Wolf *et al.*, 2010; Vital *et al.*, 2015; Dunning *et al*., 2016). Polyphasic taxonomy predominantly classes strains into species groups through genetic similarity and also with consideration of expected characteristic phenotypes of the species (Vandamme et al., 1996; Young, 2016). However, this idea does not necessarily align well to bacterial species where large species accessory genomes can convey a multitude of diverse phenotypes (Young, 2016). The high genetic diversity within bacteria species is largely accounted for by various forms of introgression, which can make identification of definitive species traits challenging (Tettelin *et al.*, 2005; McInerney, McNally and O'Connell, 2017). Subsequently, it has been suggested that bacterial taxonomy should be defined by core gene relationships, and that morphological and metabolic phenotypic similarity should not be a strict requirement of species classification (Chan *et al*., 2012; Kumar *et al*., 2015). Transcriptomic data provide the informative link between genomic and phenotypic variation and could confirm the genomic influence on phenotypic and regulatory divergence of strains while also identifying species phenotypes that do not necessarily result in a physiological trait. For example, gene expression profiles have been able to emphasise regulatory differences in strain physiology to a greater extent than by phylogenetic differences alone (Vital *et al*., 2015). From observing transcriptome differences between *Rlt* genospecies, transcriptome profiles from multiple strains could be used to identify species-specific regulatory expression differences and transcriptional phenotypes in combination with genomic and phenotypic data.

7.4. Genetic diversity and environmental origins of strains as determinants of *Rlt* interactions

This thesis has explored how *Rlt* intraspecies genetic diversity is potentially influenced by plant-mediated interactions and local growth conditions. The large genetic diversity of *Rlt* was additionally shown to result in significant transcriptomic and phenotypic intraspecies variation. The final aim was to investigate how this genetic and phenotypic diversity might influence intraspecific interactions between strains. Variation in rhizobial community diversity can be driven by intraspecific competition between strains for nodule occupancy (Denison and Kiers, 2004; Kiers and Denison, 2008; Blanco, Sicardi and Frioni, 2010; Wielbo *et al.*, 2011; Barrett *et al.*, 2015), or interactions with other microbial species in the soil (Pugashetti, Angle and Wagner, 1982; Villacieros *et al.*, 2003; Hibbing *et al.*, 2010; Teng *et al.*, 2015; Lu *et al.*, 2017). For example, it is well documented that commercial rhizobia inoculants are commonly unable to compete with the diverse native soil rhizobia in the field for nodule occupancy (Berg *et al.*, 1988; Denton *et al.*, 2003; Blanco, Sicardi and Frioni, 2010; Batista *et al.*, 2015; Checcucci *et al.*, 2017; Irisarri *et al.*, 2019; Tartaglia *et al.*, 2019). Understanding the diversity of strain interactions could be used in future work to indicate potential strain compatibilities or conflicts which may affect the productivity of the legume symbiosis, such as by selecting inoculant strains that can compete against native rhizobia and persist in the soil (Barrett *et al.*, 2015; Pahua *et al.*, 2018; Liu *et al.*, 2019). However, it is unclear to what extent intraspecific competitive interactions differ within rhizobia communities. Therefore, this project investigated the intraspecific indirect exploitative competition and direct interference competition between *Rlt* strains to see if pairwise *Rlt* interactions could be determined based on the genetic properties (genospecies) and environmental origin (conventional or organic farming) of strains.

Significant variation was observed in both direct and indirect competitive ability of *Rlt* strains at the level of individual strains and between genospecies (Figure 6.1). Genospecies were hypothesised to interact differently because they are genetically distinct (Cavassim *et al.*, 2020) and shown to differ in the expression of core genes which was associated with differences in growth phenotypes and putative metabolic differences (Figure 4.3 and Figure 4.4). Direct and indirect competitive interactions associated with genospecies differences were largely driven by the overall inhibitory action of gsE strains towards the growth of gsA strains, however ultimately competitive ability varied within genospecies too (Figure 6.2 and Figure 6.4). Interactions were suggested to be explained by resource competition between specialist and generalist strains through investigation of Ecoplate data inferring metabolic capabilities of strains (Figure 6.7). Genetically similar strains did not necessarily display more competitive interactions, despite previous research suggesting that more genetically related strains will display stronger resource competition towards one another (Russel *et al.*, 2017). Genetic differences between strains were identified that were suggested to potentially increase the susceptibility of strains to other rhizobia strains, and additionally other properties were also identified that could enable strains to outcompete and inhibit other rhizobia strains. For example, secreted compounds were suggested to influence competitive ability in addition to resource utilisation, including secretion of quorum sensing signalling molecules, bacteriocins and secondary metabolites (Figure 6.8). Comparative genomic analysis also identified large variation in the presence of inducer and regulator genes for other downstream quorum sensing pathways, such as *rhiI/rhiR, traI/traR* and *raiI/raiR*, similarly to previous investigations (Wisniewski-Dyé and Downie, 2002). This could have contributed to the large variation in observed facilitative and inhibitory interactions across strains. Further exploration of the presence of quorum sensing pathway genes identified that the greater susceptibility of specific gsA strains (SM144A, SM154C, SM145B) to direct growth inhibition was likely the result of the presence of quorum-sensing *traI*, *traR* and *bisR* genes, which mediate strain growth via increased sensitivity to the quorum-sensing AHL, 3OH-$C_{14:1}$-HSL (*small* bacteriocin) (Hwang *et al.*, 1994; Wilkinson *et al.*, 2002; Danino *et al.*, 2003; He *et al.*, 2003; McAnulla *et al.*, 2007; Lang and Faure, 2014). The potential variation in quorum sensing capabilities between genospecies was further supported by the identification of differentially expressed quorum sensing associated genes, and the enrichment of quorum sensing genes contributing to the distinction of genospecies core genome expression profiles (Figure 4.1). Additionally, a search of secondary metabolite biosynthesis gene clusters in strain genomes identified that all gsE strains contained a non-ribosomal peptide synthetase (NRPS) gene cluster for

Vicibactin siderophore production. This siderophore could have increased the competitive ability of gsE strains to repress strain growth through antimicrobial activity and increased resource competition for crucial molecules like iron (Geetha and Joshi, 2013; Wright *et al.*, 2013; Kramer, Özkaya and Kümmerli, 2019).

Overall, this analysis highlighted that *Rlt* strains display a wide range of competitive abilities at the genotype level with the potential to influence the growth of other strains utilising both direct and indirect mechanisms. While this analysis explored multiple mechanisms to potentially explain the patterns of *Rlt* strain interactions, they were by no means extensive. Therefore, a key area for future study would be to conduct additional comparative genomic analyses to investigate metabolic pathway differences between strains, which could aid identification of mechanisms by which strains differ in their competitive resource utilisation. Future work could also utilise mass spectrometry to identify molecular differences between strain supernatants which could prove insightful for understanding why some strain supernatants were particularly inhibitory or facilitative to strain growth.

It was additionally hypothesised that *Rlt* intraspecies interactions may differ between strains isolated from different environmental origins. This is because *Rlt* strains were shown to cluster by genospecies based on gene content similarity, but also cluster by geographic origin as an underlying substructure within genospecies clustering (Cavassim *et al.*, 2020). This could suggest that within genospecies strains have adapted to different environments which may influence their competitive ability. When competitive interactions of gsC strains originating from conventional or organic farm managements were compared, it was found that on average environmental origin was not significantly associated with either direct or indirect competitive interactions (Figure 6.3 and Figure 6.5). Similarly, growth of gsA strains was consistently directly and indirectly suppressed by gsE strains despite all strains being originally isolated from multiple different organic farm sites across Denmark, suggesting that this competitive ability is likely to have adapted on a larger geographic scale and could be replicable across farm managements. However, the reduced influence of environmental origin is contradicted by previous studies which have suggested that differing farming practices can influence rhizobial population size, diversity, nodulation and fixation ability (Graham and Vance, 2000; Schmidt, Weese and Lau, 2017). Therefore, the influence of different farming treatments on the competitive ability of strains may only manifest under larger scale rhizosphere communities or when strains are introduced into foreign agricultural conditions. For example, gsC strains (OC and CC) were the only strain to contain both inducer and regulator orthologs for either *rhi* or *rai* quorum sensing pathways which may infer a currently unknown competitive advantage in the rhizosphere (Gray *et al.*, 1996; Wisniewski-Dyé and Downie, 2002; Sanchez-Contreras *et al.*, 2007). As all strains used in the competition assays were originally isolated from farm sites across Denmark, it would also be interesting for future research to investigate whether observed genospecies-associated patterns are replicable across other countries and continents.

While the intraspecific competition assays are unable to answer how intraspecies competition affects symbiotic success, they do highlight the wide diversity of interactions between *Rlt* strains and the potential importance of considering these interactions when optimising symbiosis. Specifically, inhibitory interactions between rhizobia could have a detrimental effect on symbiotic efficiency if strains with strong nitrogen fixing abilities are subsequently unable to form symbiosis with the plant host (Kiers and Denison, 2008; Blanco, Sicardi and Frioni, 2010; Barrett *et al.*, 2015; Pahua *et al.*, 2018). Therefore, it is crucial that the influence of intraspecies interactions are considered when developing agricultural inoculants to: 1) avoid conflicting interactions between co-inoculant strains; and 2) circumvent incompatibility with indigenous rhizosphere communities which may reduce inoculant effectiveness (Berg *et al.*, 1988; Triplett and Sadowsky, 1992; Blanco, Sicardi and Frioni, 2010). From these competition experiments, it could be suggested that gsA and gsE should not be used as co-inoculants together, as gsE may inhibit gsA growth in the inoculum mixture. Relatedly, it could be suggested to avoid using gsA inoculants on soils where gsE is highly abundant, due to the possibility that gsA inoculants would not be able to outcompete the native soil rhizobia. However, despite the highly competitive phenotype displayed by *Rlt* gsE strains, they have so far only been identified in low abundance in clover nodules compared to other genospecies, although they show higher frequencies in *Rlv* populations from pea and faba bean nodules (Kumar *et al.*, 2015; Boivin *et al.*, 2020; Cavassim *et al.*, 2020). Therefore, these interactions are unlikely to be directly applicable to the field environment where strains must deal with additional abiotic stressors and higher order effects of community interactions with other plant hosts and soil microbes (Barrett *et al.*, 2015). A central theory would be to investigate whether highly competitive rhizobia in pairwise competition assays were also more likely to form symbiosis with the plant, and recent advancements in sequencing technologies able to differentiate individual strains in plant nodules will be crucial for scaling-up these interactions into multi-strain communities *in planta* (Fields *et al.*, 2019; Mendoza-Suárez *et al.*, 2020). Future research also aims to utilise these pairwise interactions as a basis for investigating how variations in rhizobial intraspecies diversity influence clover yield in greenhouse studies.

7.5. Final remarks

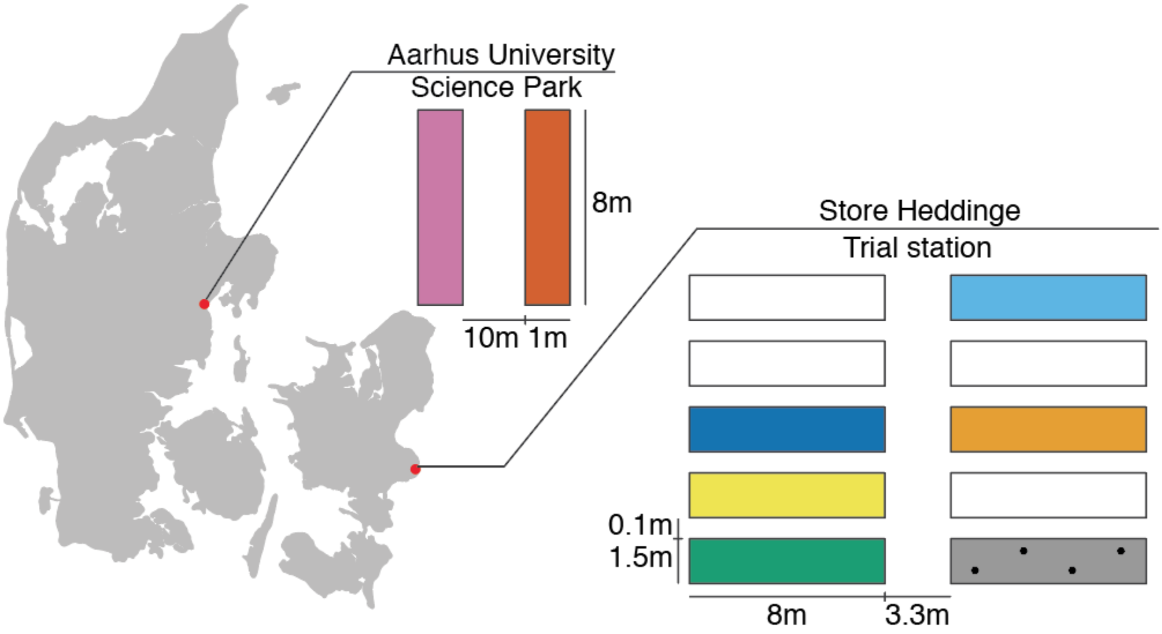
It is well established that significant genetic and phenotypic diversity is present in populations of *Rhizobium leguminosarum* to the extent that a species complex of at least five genetically distinct genospecies have been identified and which display diverse metabolic phenotypes that are not genospecies-exclusive (Kumar *et al.*, 2015; Boivin *et al.*, 2020; Cavassim *et al.*, 2020). Therefore, the relevance of this vast genetic diversity of *Rlt* and the extent of its influence on the phenotypes and rhizosphere interactions associated with symbiosis have remained unclear. In this thesis, the extent of *Rlt* intraspecies diversity was investigated at the genetic and phenotypic levels, with a specific focus to identify functional differences between *Rhizobium leguminosarum* genospecies which might have indicated towards their maintained genetic distinction. This thesis identified that extensive genetic diversity can manifest in significant transcriptional and phenotypic variation across *Rlt* strains, and at the intraspecies level this diversity can influence symbiont-selectivity by different clover hosts and also the competitive interactions among strains. The novel development of the MAUI-seq high throughput amplicon sequencing approach described in this thesis has the potential to further impact future intraspecies investigation by providing a comprehensive chimera and erroneous sequence detection pipeline and by expanding the use of multiple amplicons to confidently characterise diversity at levels of low sequence divergence. This method is also applicable to the wider use of intraspecies studies with other microbial species using alternative environmental samples or research requiring monitoring of multiple species within one sample. Multi-strain transcriptomics, operon prediction, and phenotype experiments reported in this thesis have also demonstrated that multiple strains can be used to capture an improved representation of the level of intraspecies diversity and phenotypic variation of a species. These experiments add weight to the arguments that the use of characteristics displayed by a single strain is not an adequate representation of species-specific traits. As the number of whole genome sequences and development of high throughput phenotypic assays increases, the level of observable intraspecies diversity for many bacterial species will also continue to expand. While large scale pair-wise competition assays between *Rlt* strains identified that significant variation in interactions could be partially associated with genospecies differences, the extent to which these trends could be replicated in an agricultural setting remain unclear. The research in this thesis clearly illustrates that there are still no exclusive phenotypes identified from genospecies genetic divergence based on the representative strains used in this project, but some genospecies on average have been demonstrated to display some phenotypes to a greater extent compared to others. However, this also raises the question of whether there remain any genospecies-exclusive phenotypes, and what has influenced and maintained this genetic separation of genospecies groups. To better understand the implications of these results, future investigations could address the level of phenotypic variation observed between *Rlt* strains and genospecies groups under bacteroid physiologies, as this thesis has singularly focused on diversity of strains in free-living physiologies. Beyond the scope of the findings from these experiments, this thesis has laid the groundwork for future investigations into the significance of intraspecies diversity for symbiotic effectiveness in the rhizobia-legume symbiosis. The significant functional diversity of *Rlt* strains will likely have implications not only for symbiosis but for general ecosystem functioning as well. Utilising this understanding of intraspecies diversity to implement optimised precision farming techniques will support the progression to securing more sustainable global food security for future generations (Sessitsch *et al.*, 2002; Lupwayi, Clayton and Rice, 2006; Checcucci *et al.*, 2017).

Appendices

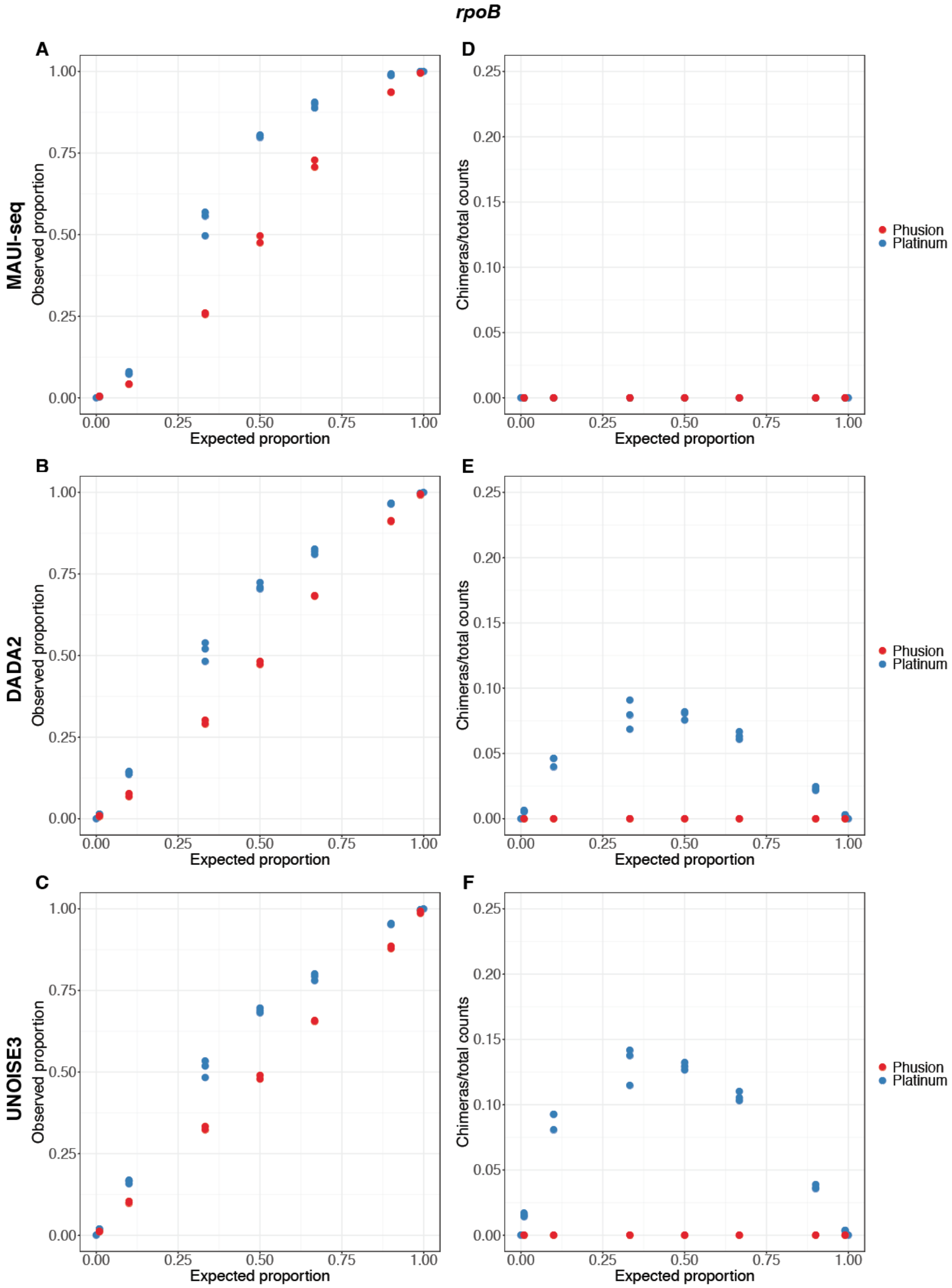
Appendix A. Chapter 2

Appendix Table A.1 Synthetic community mixes sample design.

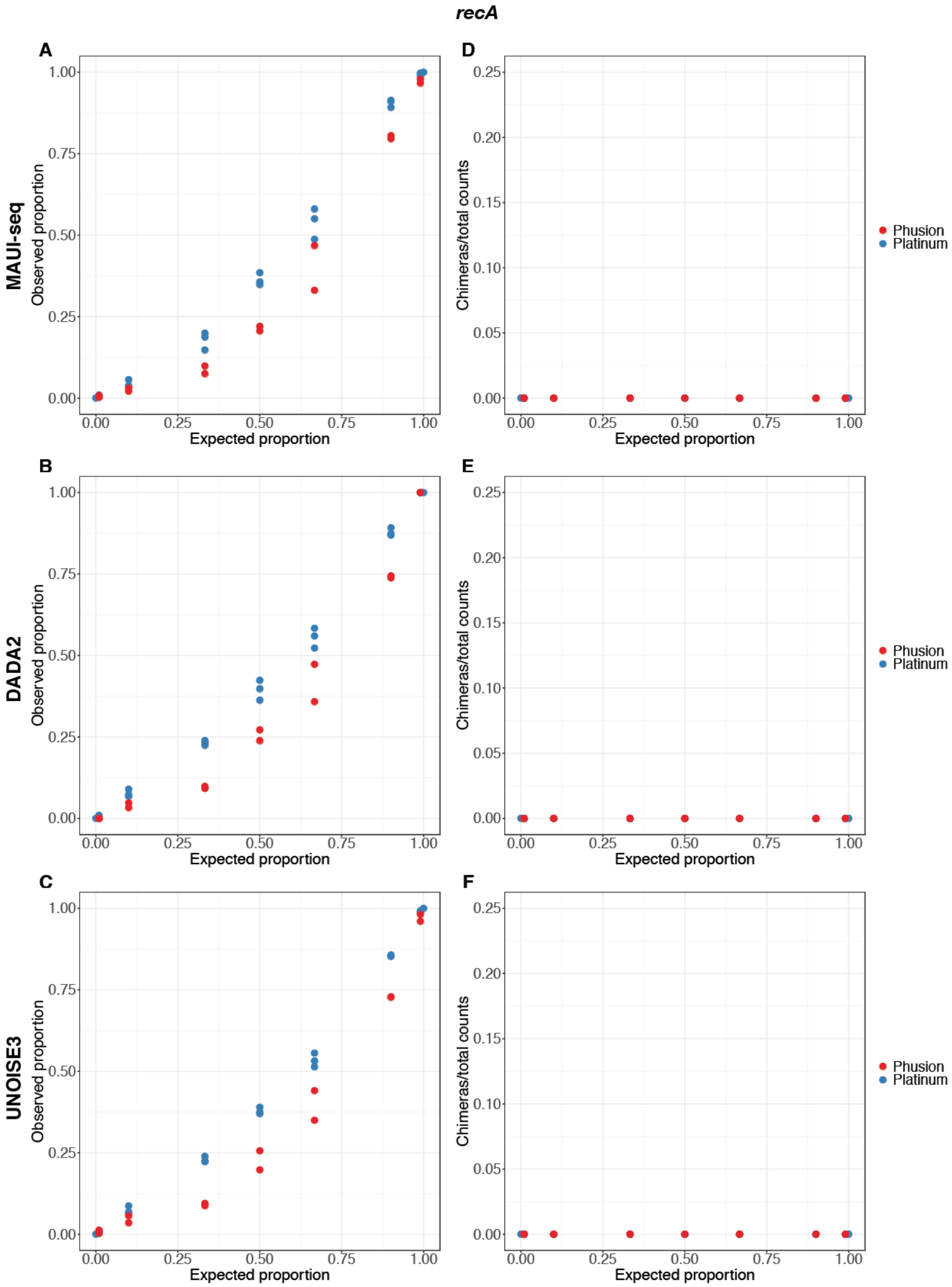
Sample ID	Strain	Community percentage (%)	Strain	Community percentage (%)	Amplified with Platinum	Amplified with Phusion
A1	SM3	100	SM170C	0	1	0
A2	SM170C	100	SM3	0	1	0
B1	SM3	50	SM170C	50	1	1
B2	SM3	50	SM170C	50	1	1
B3	SM3	50	SM170C	50	1	0
C1	SM3	66.6	SM170C	33.3	1	1
C2	SM3	66.6	SM170C	33.3	1	1
C3	SM3	66.6	SM170C	33.3	1	0
D1	SM3	90	SM170C	10	1	1
D2	SM3	90	SM170C	10	1	1
D3	SM3	90	SM170C	10	1	0
E1	SM3	99	SM170C	1	1	1
E2	SM3	99	SM170C	1	1	1
E3	SM3	99	SM170C	1	1	0
F1	SM170C	66.6	SM3	33.3	1	1
F2	SM170C	66.6	SM3	33.3	1	1
F3	SM170C	66.6	SM3	33.3	1	0
G1	SM170C	90	SM3	10	1	1
G2	SM170C	90	SM3	10	1	1
G3	SM170C	90	SM3	10	1	0
H1	SM170C	99	SM3	1	1	1
H2	SM170C	99	SM3	1	1	1
H3	SM170C	99	SM3	1	1	0



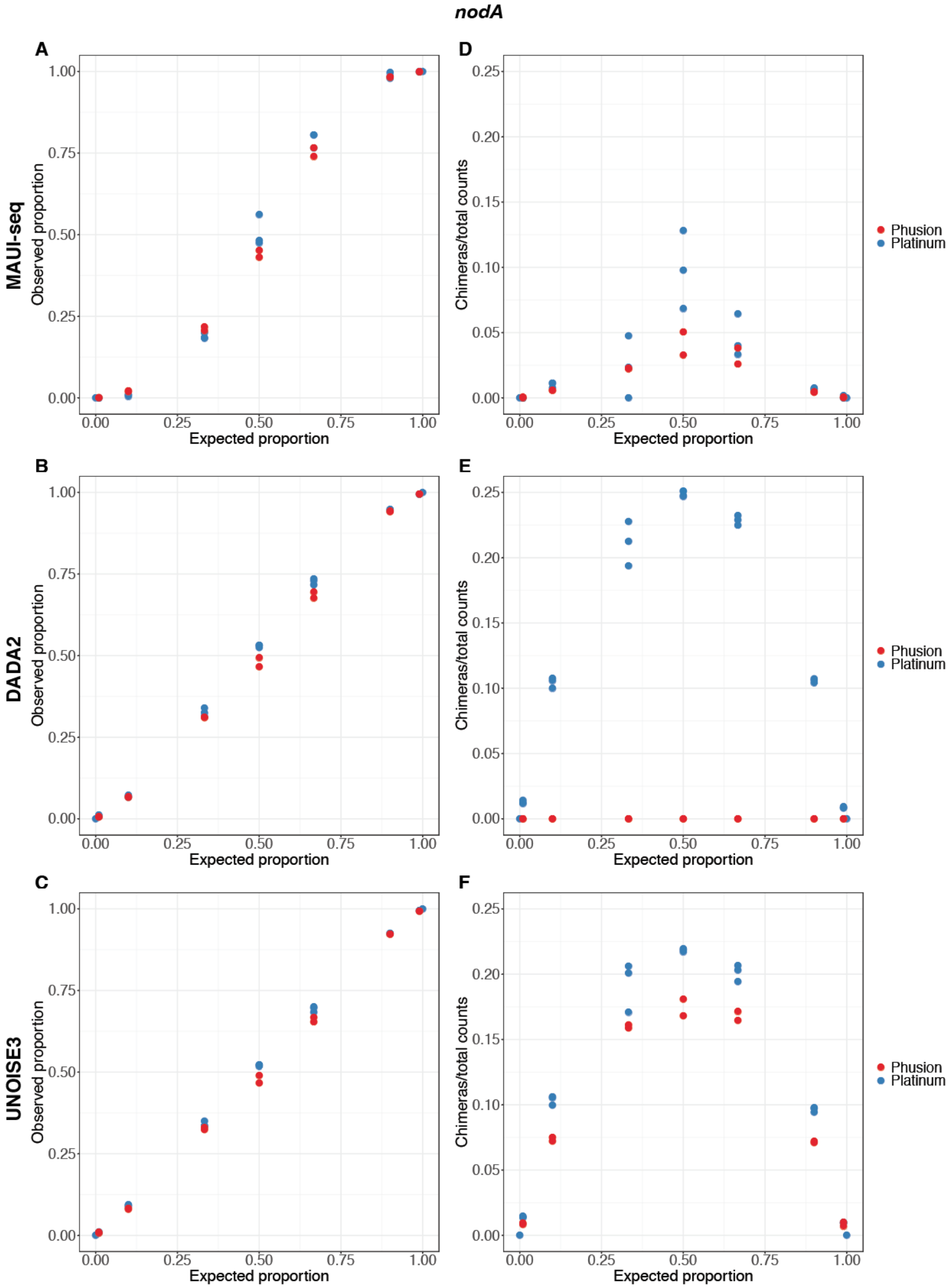
Appendix Figure A.1 Sampling sites for assessment of *Rhizobium leguminosarum* symbiovar *trifolii* diversity in root nodule samples. Sampling locations in the Aarhus University Science Park and a clover trial station in Store Heddinge. DNA was isolated from 100 nodules from four points (black dots) on each individual plot.



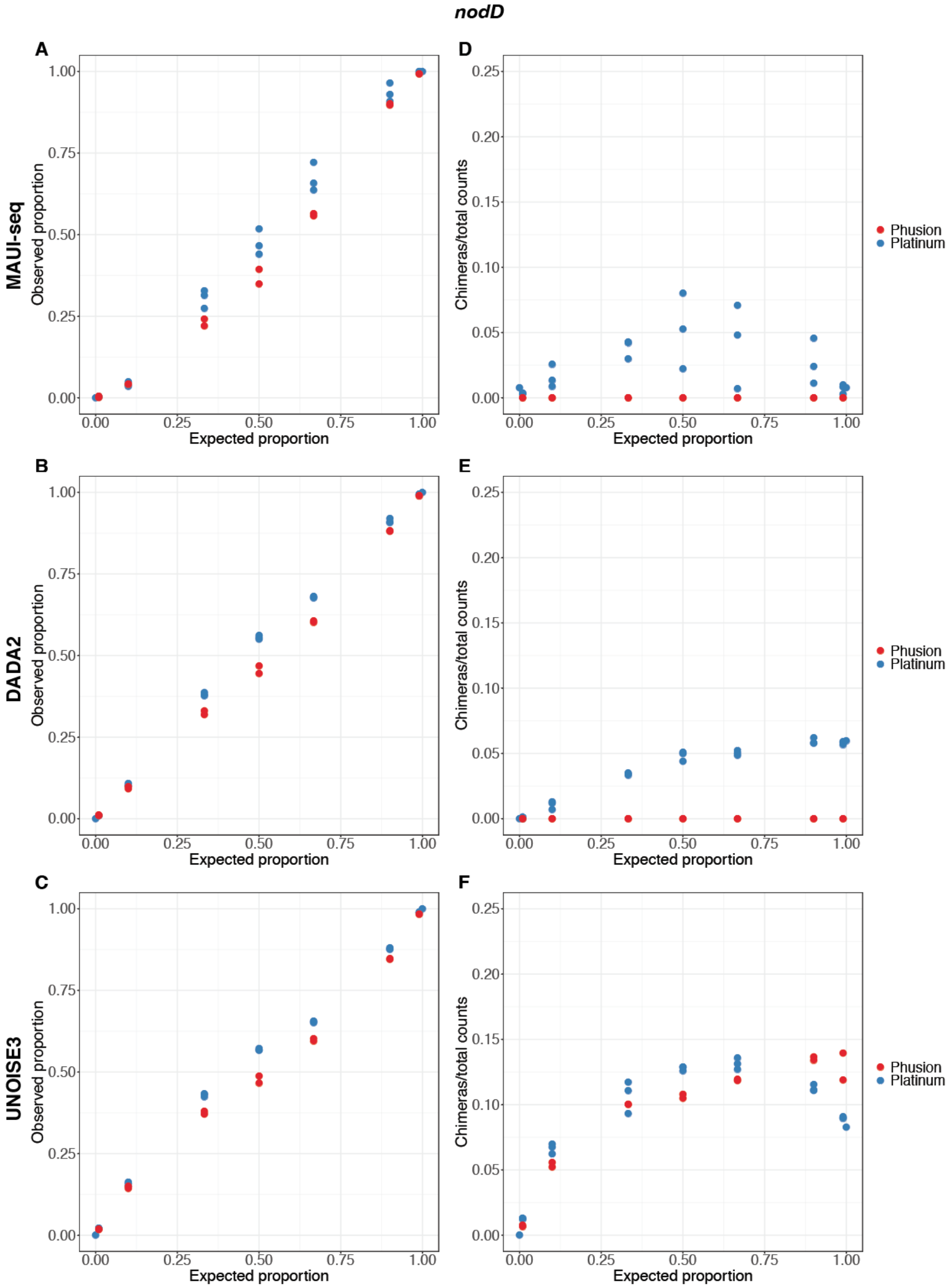
Appendix Figure A.2 Performance on DNA mixtures for *rpoB*. **A and D)** MAUI-seq. **A)** Observed proportion of SM170C *rpoB* reads versus the expected proportion. Pearson correlation for Phusion=0.996 and Platinum=0.956. **D)** Proportion of chimeras compared to total reads. **B and E)** DADA2. **B)** Observed proportion of SM170C *rpoB* reads versus the expected proportion. Pearson correlation for Phusion=0.999 and Platinum=0.977. **E)** Proportion of chimeras compared to total counts. **C and F)** UNOISE3. **C)** Observed proportion of SM170C *rpoB* reads versus the expected proportion. Pearson correlation for Phusion=0.9998 and Platinum=0.981. **F)** Proportion of chimeras compared to total counts.



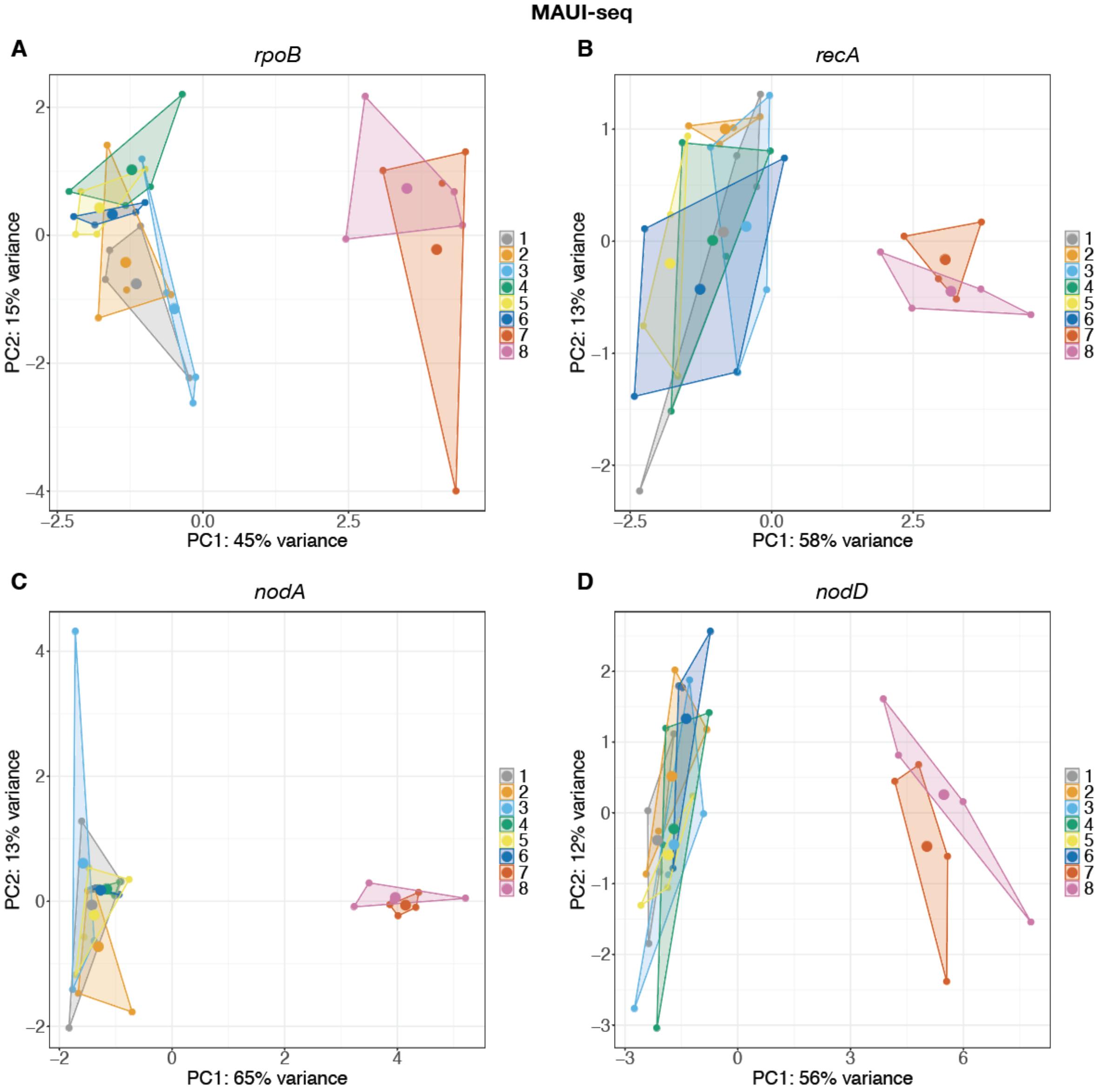
Appendix Figure A.3 Performance on DNA mixtures for *recA*. **A and D)** MAUI-seq. **A)** Observed proportion of SM170C *recA* reads versus the expected proportion. Pearson correlation for Phusion=0.948 and Platinum=0.984. **D)** Proportion of chimeras compared to total reads. **B and E)** DADA2. **B)** Observed proportion of SM170C *recA* reads versus the expected proportion. Pearson correlation for Phusion=0.952 and Platinum=0.991. **E)** Proportion of chimeras compared to total counts. **C and F)** UNOISE3. **C)** Observed proportion of SM170C *recA* reads versus the expected proportion. Pearson correlation for Phusion=0.947 and Platinum=0.989. **F)** Proportion of chimeras compared to total counts.



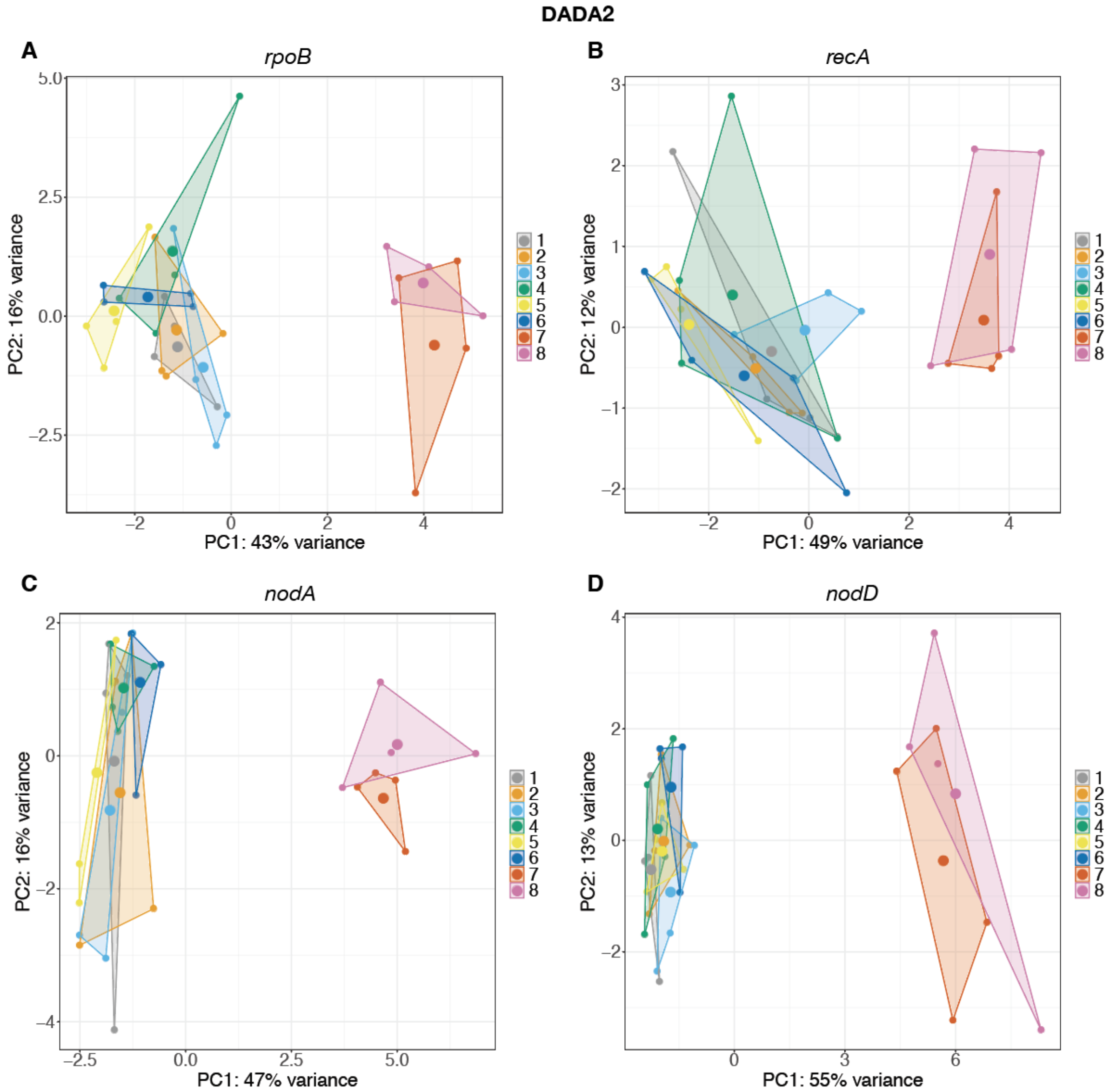
Appendix Figure A.4 Performance on DNA mixtures for *nodA*. **A and D)** MAUI-seq. **A)** Observed proportion of SM170C *nodA* reads versus the expected proportion. Pearson correlation for Phusion=0.989 and Platinum=0.985. D: Proportion of chimeras compared to total reads. **B and E)** DADA2. **B)** Observed proportion of SM170C *nodA* reads versus the expected proportion. Pearson correlation for Phusion=0.999 and Platinum=0.998. **E)** Proportion of chimeras compared to total counts. **C and F)** UNOISE3. **C)** Observed proportion of SM170C *nodA* reads versus the expected proportion. Pearson correlation for Phusion=0.999 and Platinum=0.999. **F)** Proportion of chimeras compared to total counts.



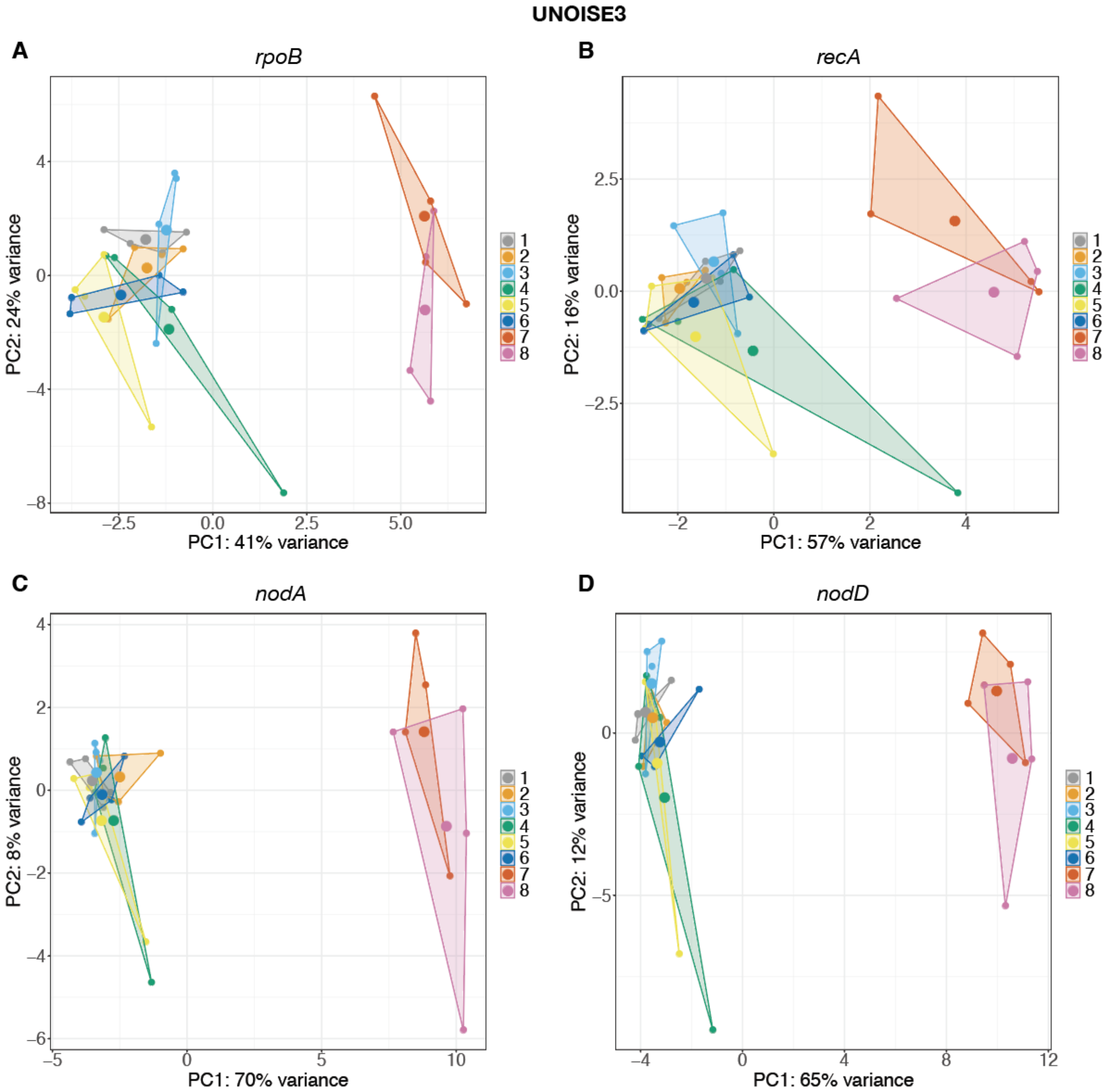
Appendix Figure A.5 Performance on DNA mixtures for *nodD*. **A and D)** MAUI-seq. **A)** Observed proportion of SM170C *nodD* reads versus the expected proportion. Pearson correlation for Phusion=0.990 and Platinum=0.998. **D)** Proportion of chimeras compared to total reads. **B and E)** DADA2. **B)** Observed proportion of SM170C *nodD* reads versus the expected proportion. Pearson correlation for Phusion=0.998 and Platinum=0.998. **E)** Proportion of chimeras compared to total counts. **C and F)** UNOISE3. **C)** Observed proportion of SM170C *nodD* reads versus the expected proportion. Pearson correlation for Phusion=0.995 and Platinum=0.995. **F)** Proportion of chimeras compared to total counts.



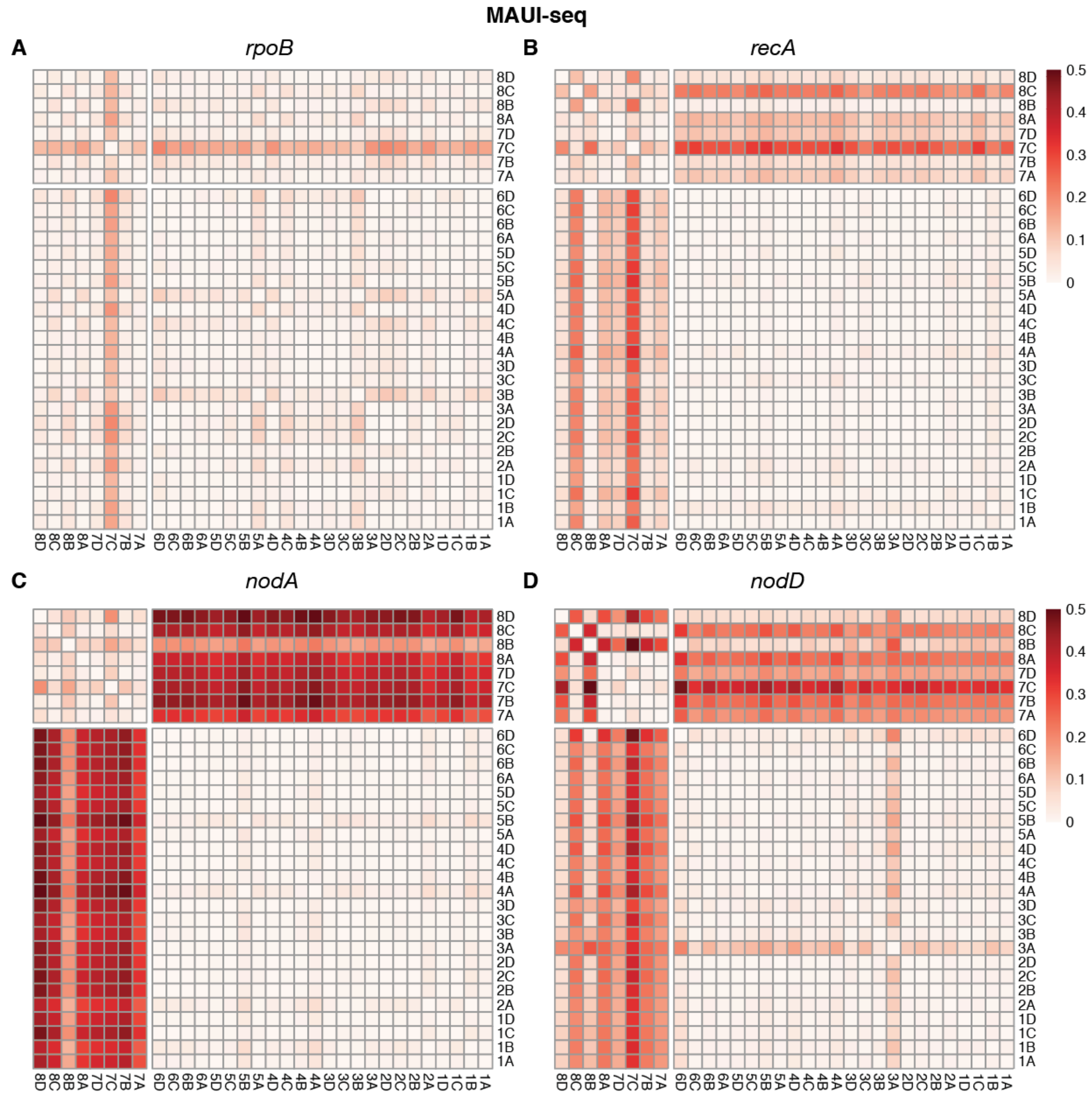
Appendix Figure A.6 Individual Principal Components Analysis of four *Rhizobium leguminosarum* symbiovar *trifolii* genes clustered by MAUI-seq. DNA was isolated from 100 nodules from four points on each individual plot. **A)** *rpoB,* **B)** *recA,* **C)** *nodA* and **D)** *nodD*.



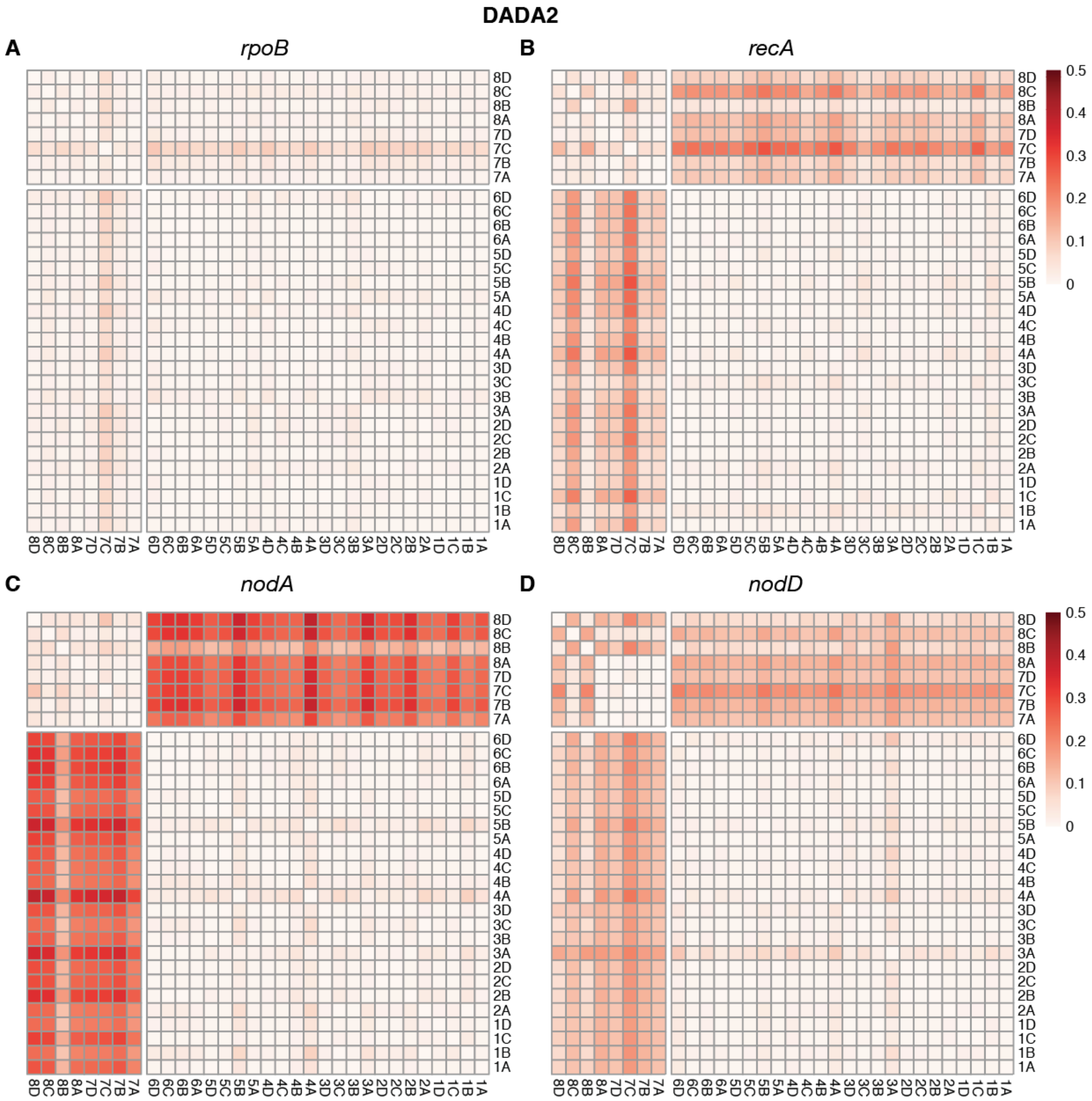
Appendix Figure A.7 Individual Principal Components Analysis of four *Rhizobium leguminosarum* symbiovar *trifolii* genes clustered by DADA2. DNA was isolated from 100 nodules from four points on each individual plot. **A)** *rpoB,* **B)** *recA,* **C)** *nodA* and **D)** *nodD*.



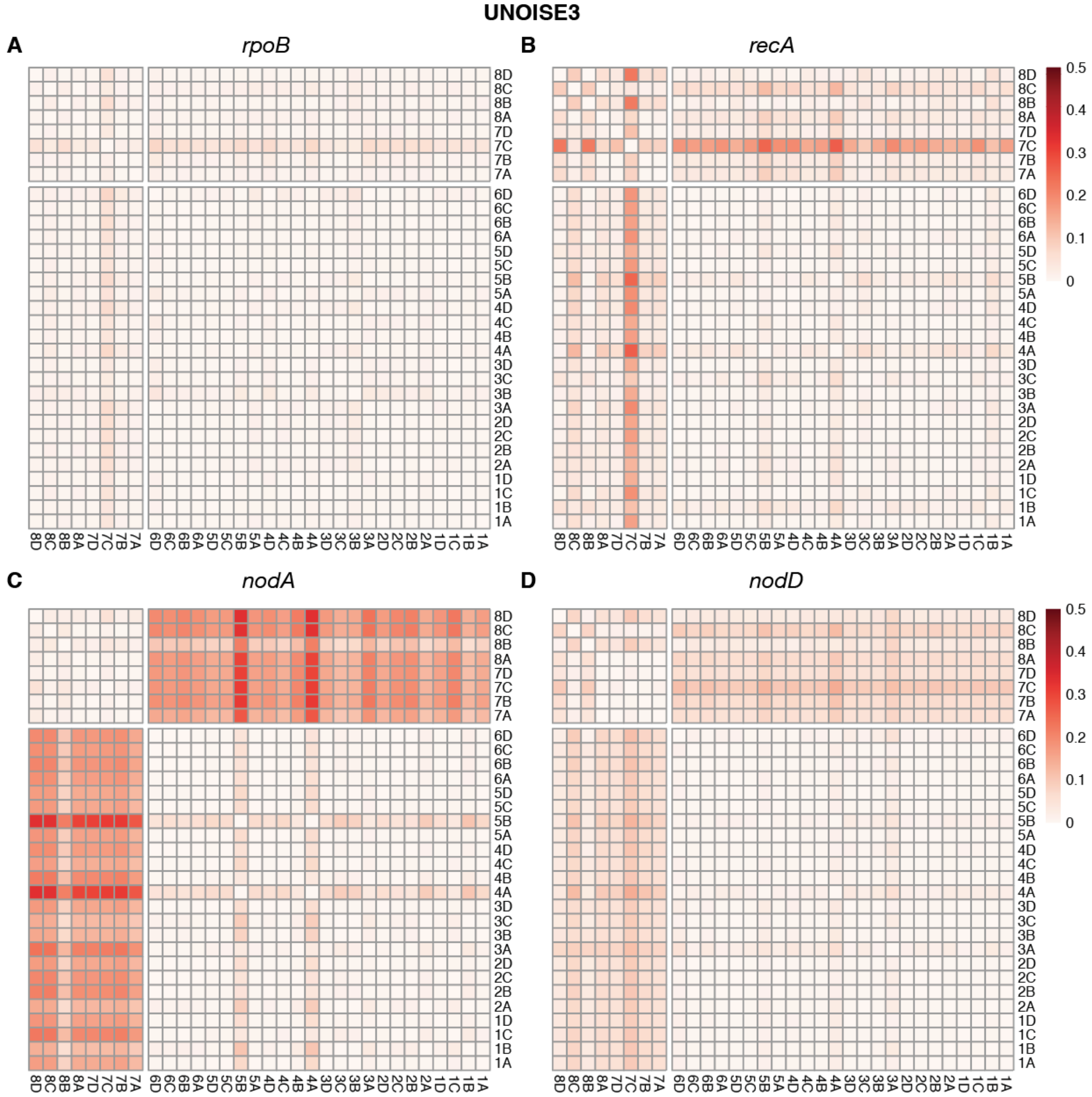
Appendix Figure A.8 Individual Principal Components Analysis of four *Rhizobium leguminosarum* symbiovar *trifolii* genes clustered by UNOISE3. DNA was isolated from 100 nodules from four points on each individual plot. **A)** *rpoB,* **B)** *recA,* **C)** *nodA* and **D)** *nodD*.



Appendix Figure A.9 F_{ST} calculated between samples for individual genes on reads clustered by MAUI-seq. **A)** *rpoB,* **B)** *recA,* **C)** *nodA,* and **D)** *nodD*.



Appendix Figure A.10 F_{ST} calculated between samples for individual genes on reads clustered by DADA2. **A)** *rpoB,* **B)** *recA,* **C)** *nodA,* and **D)** *nodD*.



Appendix Figure A.11 F_{ST} calculated between samples for individual genes on reads clustered by UNOISE3. **A)** *rpoB*, **B)** *recA,* **C)** *nodA,* and **D)** *nodD*.

Appendix B. Chapter 3

Appendix Table B.1 Number of unique allele sequences identified with MAUI-seq UMI clustering method for *Rlt* genes *rpoB*, *recA*, *nodA* and *nodD* across all 39 field samples. Global Fixation Index (F_{ST}) values for each gene individually and in combination was calculated. Global F_{ST} values were calculated using the FinePop package in R.

gene	Total number of detected unique sequences	Global F_{ST}
rpoB	16	0.06738746
recA	8	0.09951743
nodA	23	0.1093755
nodD	21	0.07441329
all genes	68	0.08355223

Appendix Table B.2 PERMANOVA results for individual allele abundances for all four *Rlt* genes combined. Pairwise allelic (Bray-Curtis) dissimilarity data was used as input into the analysis. Permutation: Free. Number of Permutations: 999.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Clover genotype	4	0.93701	0.234254	3.7036	0.37429	0.001
Plot	5	0.30141	0.060283	0.9531	0.1204	0.517
Residuals	20	1.26501	0.06325		0.50531	
Total	29	2.50343			1	

Appendix Table B.3 PERMANOVA results for individual allele abundances for *rpoB Rlt* genes combined. Pairwise allelic (Bray-Curtis) dissimilarity data was used as input into the analysis. Permutation: Free. Number of Permutations: 999.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Clover genotype	4	1.30291	0.32573	5.375	0.46206	0.001
Plot	5	0.30483	0.06097	1.006	0.10811	0.43
Residuals	20	1.21201	0.0606		0.42983	
Total	29	2.81975			1	

Appendix Table B.4 PERMANOVA results for individual allele abundances for *recA Rlt* genes combined. Pairwise allelic (Bray-Curtis) dissimilarity data was used as input into the analysis. Permutation: Free. Number of Permutations: 999.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Clover genotype	4	0.70817	0.177043	5.2474	0.47824	0.006
Plot	5	0.09784	0.019568	0.58	0.06607	0.756
Residuals	20	0.67478	0.033739		0.45569	
Total	29	1.48079			1	

Appendix Table B.5 PERMANOVA results for individual allele abundances for *nodA Rlt* genes combined. Pairwise allelic (Bray-Curtis) dissimilarity data was used as input into the analysis. Permutation: Free. Number of Permutations: 999.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Clover genotype	4	0.79647	0.199118	2.12257	0.25796	0.022
Plot	5	0.4149	0.082979	0.88455	0.13438	0.56
Residuals	20	1.8762	0.09381		0.60766	
Total	29	3.08756			1	

Appendix Table B.6 PERMANOVA results for individual allele abundances for *nodD Rlt* genes combined. Pairwise allelic (Bray-Curtis) dissimilarity data was used as input into the analysis. Permutation: Free. Number of Permutations: 999.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Clover genotype	4	1.1594	0.28986	2.678	0.29114	0.012
Plot	5	0.6582	0.13164	1.2162	0.16527	0.295
Residuals	20	2.1648	0.10824		0.54359	
Total	29	3.9824			1	

Appendix Table B.7 Two-way ANOVA results for clover genotype association with differences in *recA* genospecies relative nodule abundance.

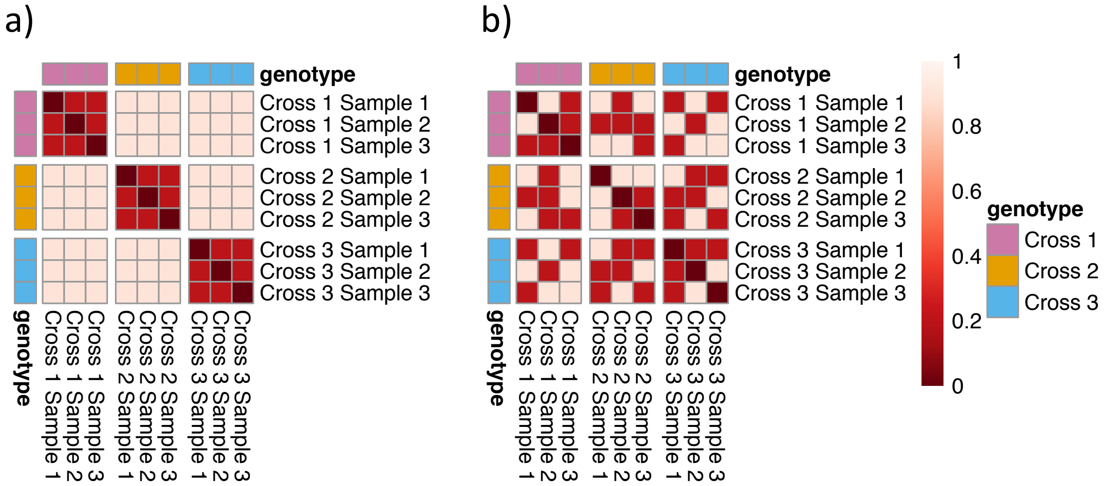
	Df	SumsOfSqs	MeanSqs	F value	Pr(>F)
Clover genotype	4	0	0	0	1
Genospecies	5	15.821	3.164	365.597	0.001
Clover genotype:Genospecies	20	1.237	0.062	7.144	0.001
Residuals	150	1.298	0.009		

Appendix Table B.8 Two-way ANOVA results for clover genotype association with differences in *rpoB* genospecies relative nodule abundance.

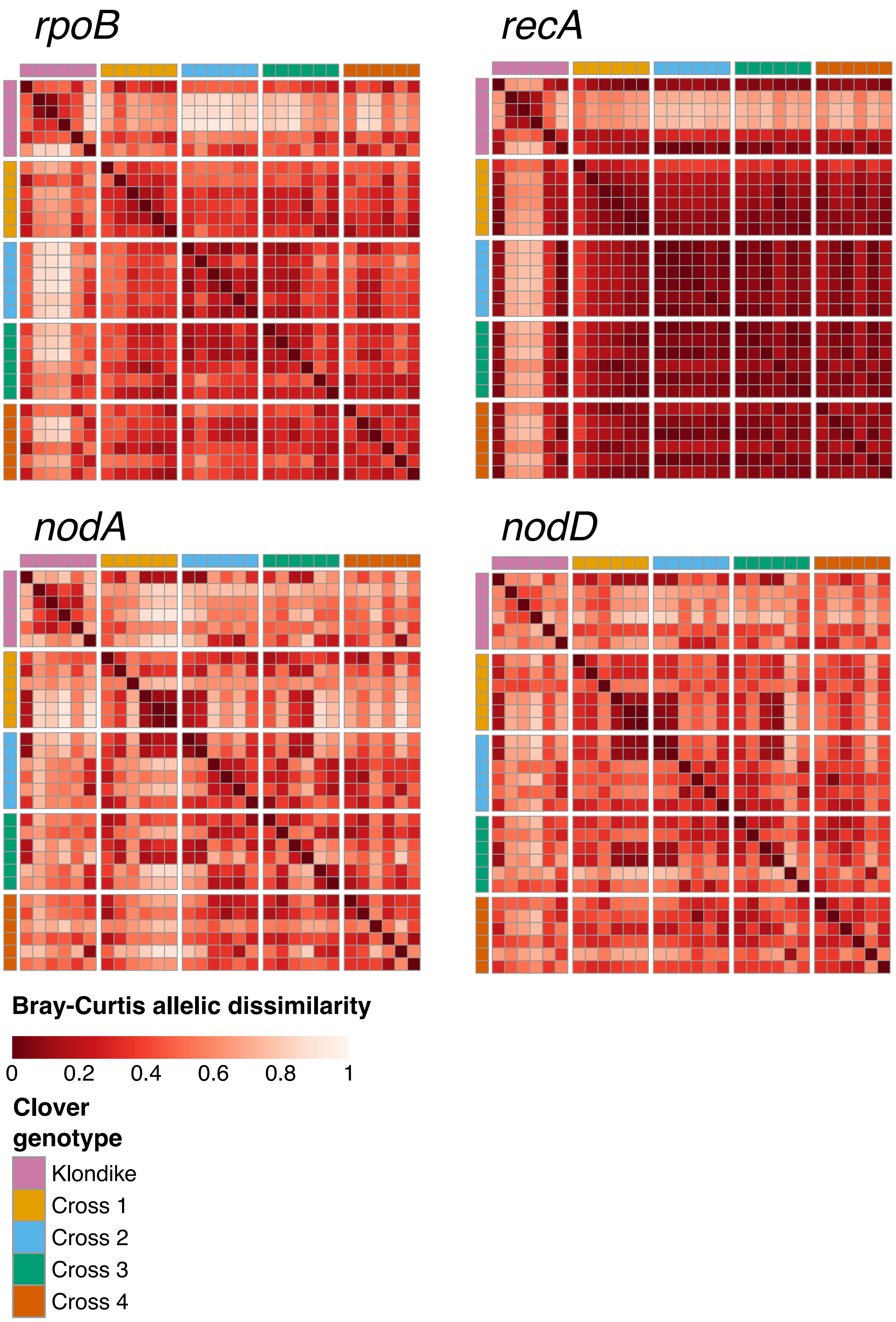
	Df	SumsOfSqs	MeanSqs	F value	Pr(>F)
Clover genotype	4	0	0	0	1
Genospecies	5	10.479	2.0957	210.43	0.001
Clover genotype:Genospecies	20	2.042	0.1021	10.25	0.001
Residuals	150	1.494	0.01		

Appendix Table B.9 Linear mixed effects models for *rpoB*, *recA*, *nodA* and *nodD*. Model formula: allelic dissimilarity ~ geographicdistance + genotypedifference + (1|sample1_genotypeID) + (1|sample2_genotypeID). [1]sample1_genotypeID causes a singular fit but was still included in the model because it did not affect estimate quantities (see methods).

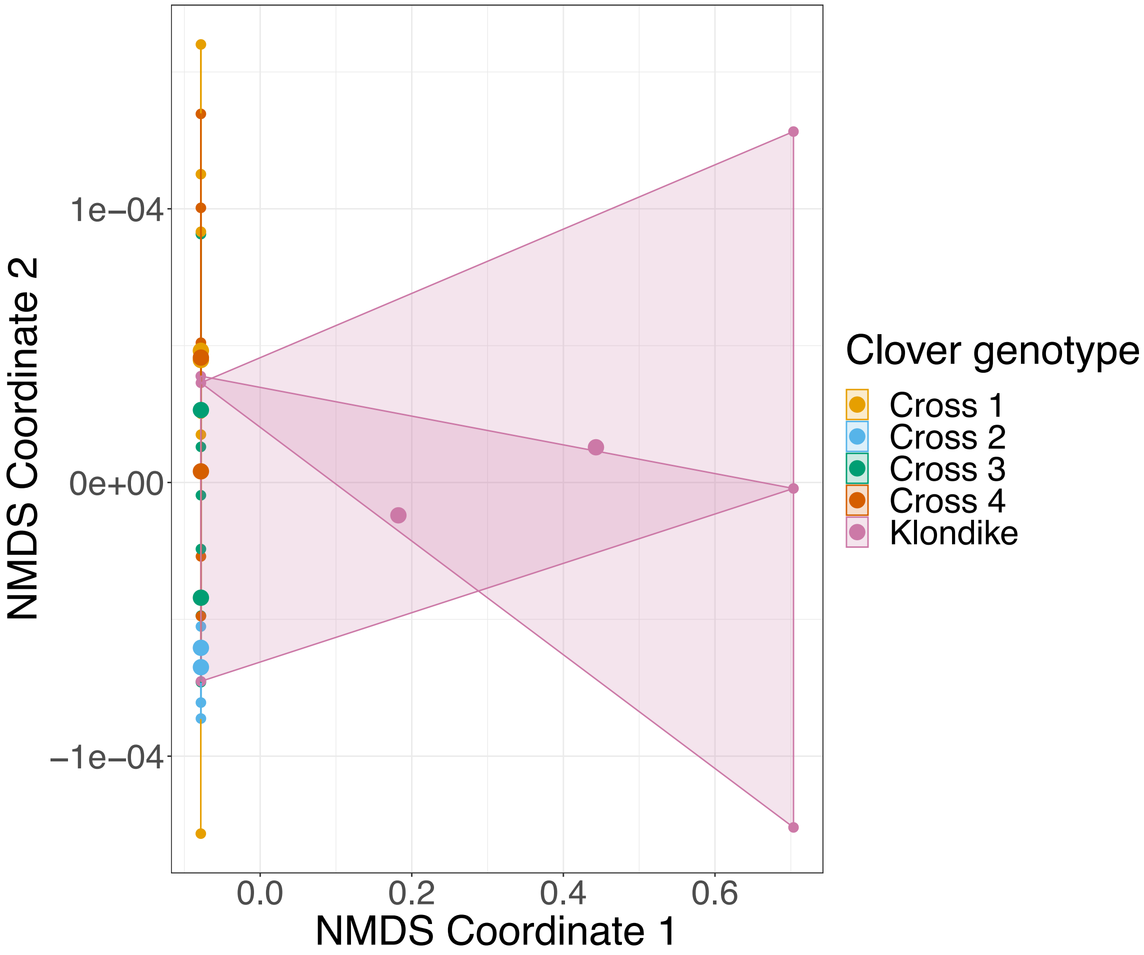
Model	AIC	BIC	Effect	Variable	Variance	Std. Dev	Estimate	Std. Error	df	t value	p-value
rpoB	-397.3	-372.9	random	sample1_genotypeID (Intercept)	0.001119	0.03345	/	/	/	/	/
			random	sample2_genotypeID (Intercept)	0.008612	0.09280	/	/	/	/	/
			random	residual	0.021630	0.14707	/	/	/	/	/
			fixed	intercept	/	/	0.327651	0.047662	7.891664	6.875	0.000136
			fixed	geographicdistance	/	/	0.003264	0.001128	431.720717	2.894	0.003997
			fixed	genotypedifference_1	/	/	-0.059448	0.020038	291.731632	-2.967	0.003259
recA	-233.2	-208.8	random	sample1_genotypeID (Intercept)	0.0008145	0.02854	/	/	/	/	/
			random	sample2_genotypeID (Intercept)	0.0140704	0.11862	/	/	/	/	/
			random	residual	0.0316484	0.17790	/	/	/	/	/
			fixed	intercept	/	/	0.138631	0.058538	7.017869	2.368	0.0496
			fixed	geographicdistance	/	/	0.002929	0.001358	422.511258	2.157	0.0316
			fixed	genotypedifference_1	/	/	-0.013370	0.023754	286.884128	-0.563	0.5740
nodA	-184.4	-159.9	random	sample1_genotypeID (Intercept)	0.0001536	0.01239	/	/	/	/	/
			random	sample2_genotypeID (Intercept)	0.0031531	0.05615	/	/	/	/	/
			random	residual	0.0362925	0.19051	/	/	/	/	/
			fixed	intercept	/	/	0.4709	0.03352	0.1145	14.046	1.44x10^{-8}
			fixed	geographicdistance	/	/	0.0004928	0.001425	0.03609	0.346	0.7297
			fixed	genotypedifference_1	/	/	-0.04404	0.02470	0.03682	-1.783	0.0754
nodD[1]	-316.3	-291.9	random	sample1_genotypeID (Intercept)	2.737x 10^{-12}	1.654x10^{-6}	/	/	/	/	/
			random	sample2_genotypeID (Intercept)	0.002871	0.05358	/	/	/	/	/
			random	residual	0.02682	0.1638	/	/	/	/	/
			fixed	intercept	/	/	0.4155	0.03004	0.1041	13.831	4.94x10^{-8}
			fixed	geographicdistance	/	/	-0.0004326	0.001213	0.04347	-0.357	0.721
			fixed	genotypedifference_1	/	/	-0.01459	0.02107	0.04320	-0.692	0.489



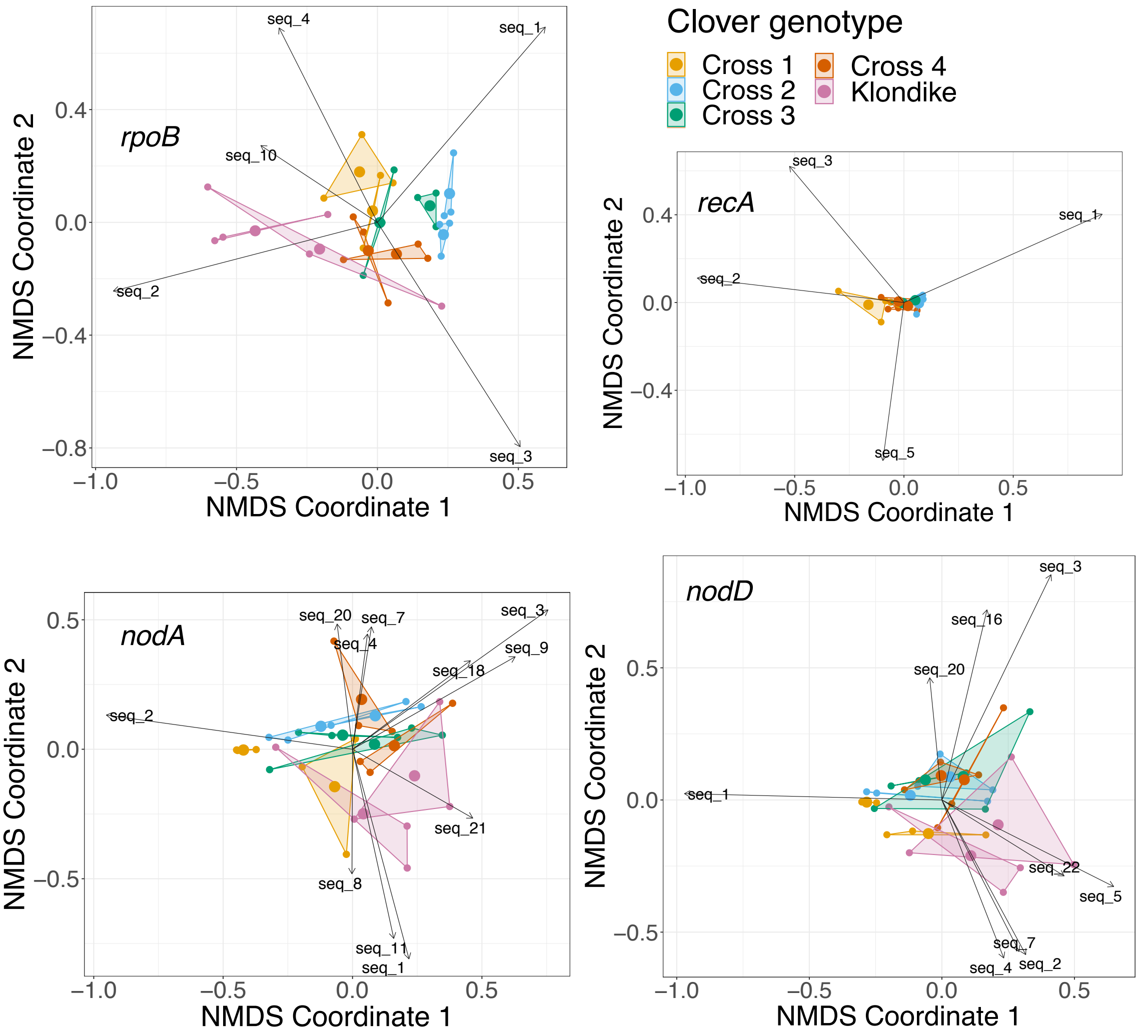
Appendix Figure B.1 Explanation example of allelic (Bray-Curtis) dissimilarity heatmaps. **a)** Samples with *Rlt* diversity that is more genetically similar within clover genotypes than between, **b)** Samples with *Rlt* diversity that is not associated to clover genotype differences. Bray-Curtis dissimilarity is shown on a scale ranging from low (red) to high (white) allelic dissimilarity.



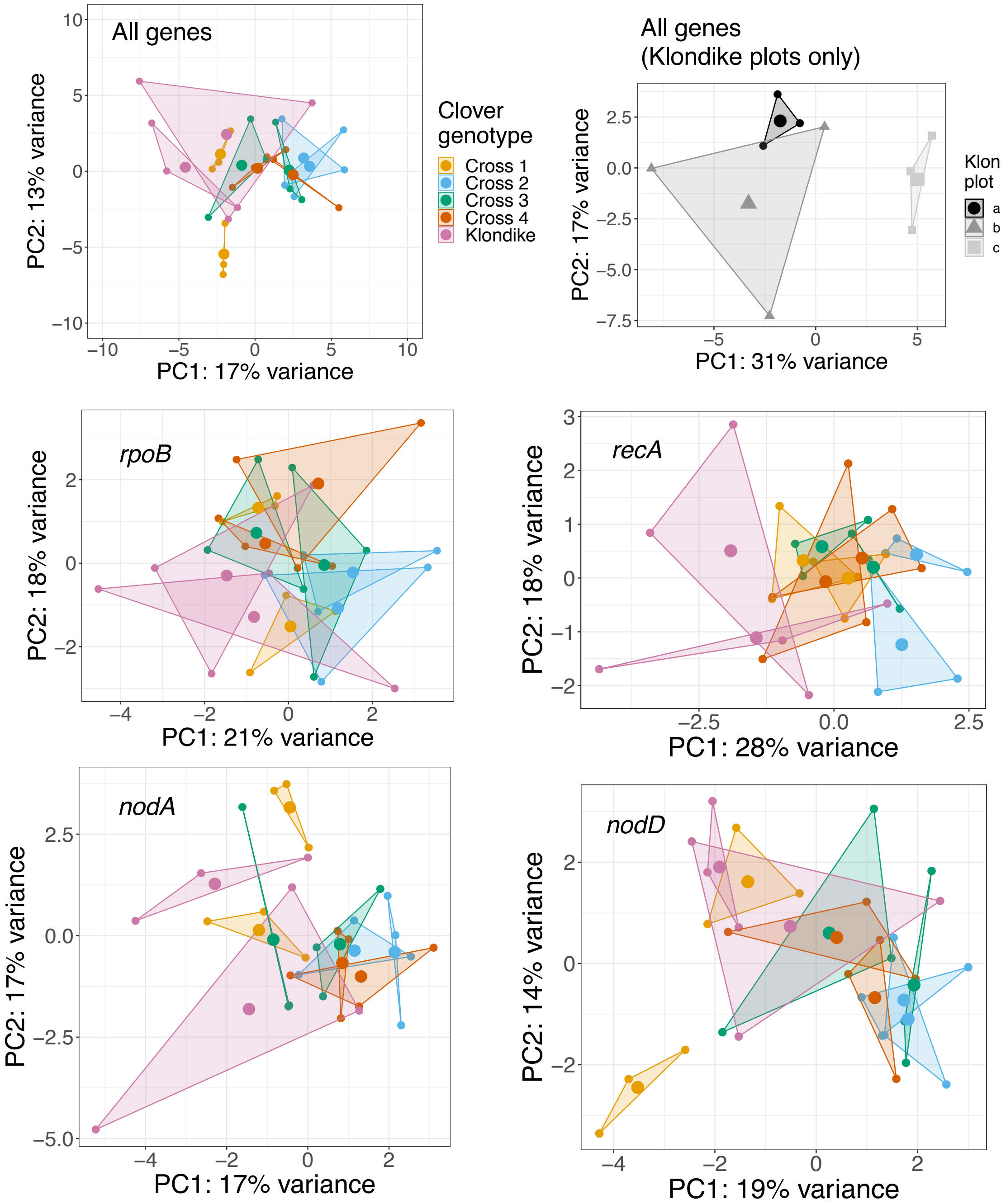
Appendix Figure B.2 Pairwise allelic (Bray-Curtis) dissimilarity of 4 *Rlt* genes. The four genes analysed individually were housekeeping genes *rpoB* and *recA*, and symbiosis genes *nodA* and *nodD*. Samples are coloured by the clover genotype host they were isolated from. Bray-Curtis dissimilarity is shown on a scale ranging from low (red) to high (white) allelic dissimilarity. Each clover genotype was sampled from 2 plots and 3 points were sampled within each plot. Samples are grouped and coloured by their respective clover genotype host.



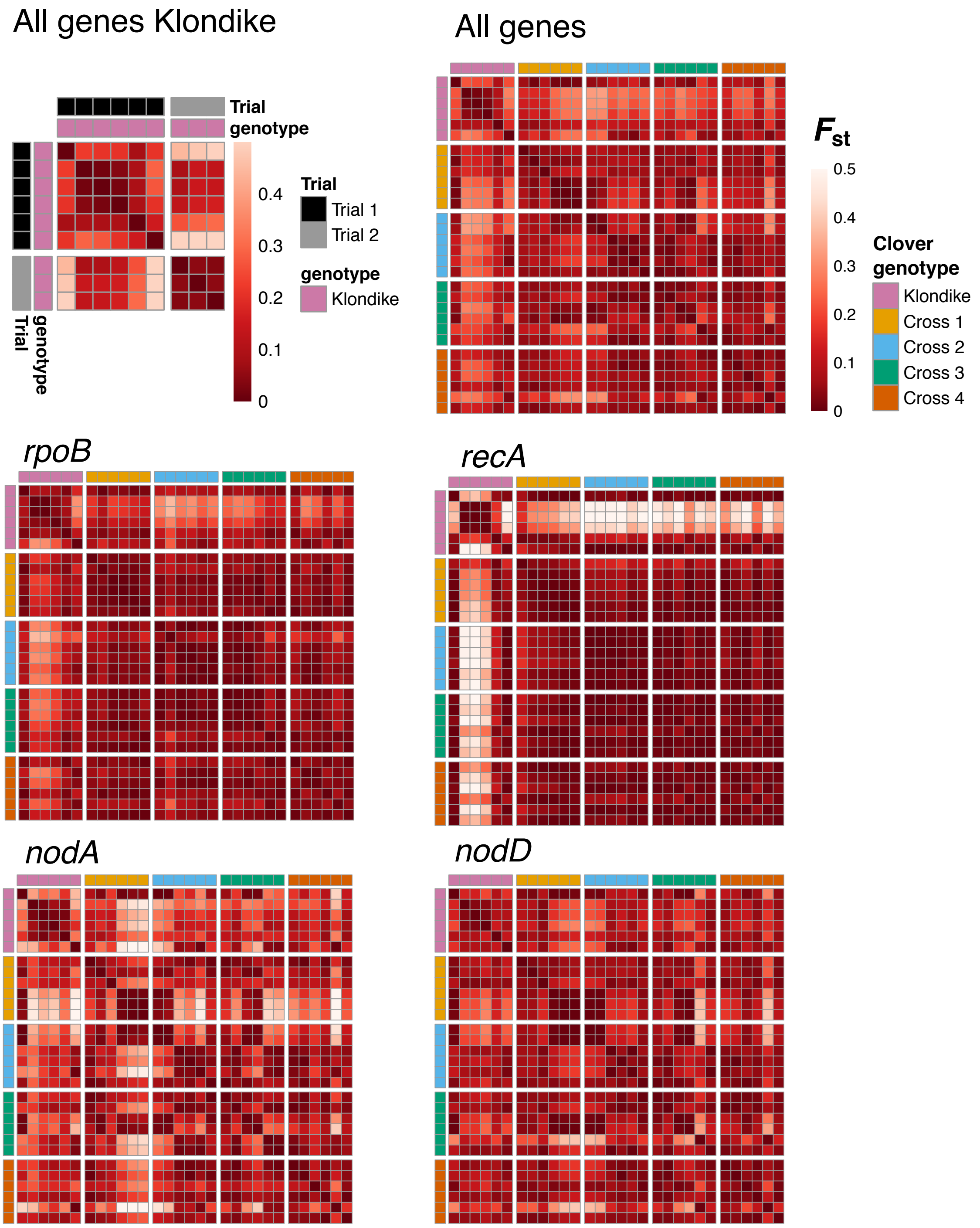
Appendix Figure B.3 Non-metric Multidimensional Scaling analysis of *recA* allelic dissimilarity with Klondike samples included. Two dimensions were specified for the analysis. Samples are grouped by their field plot and coloured by the clover genotype they were isolated from.



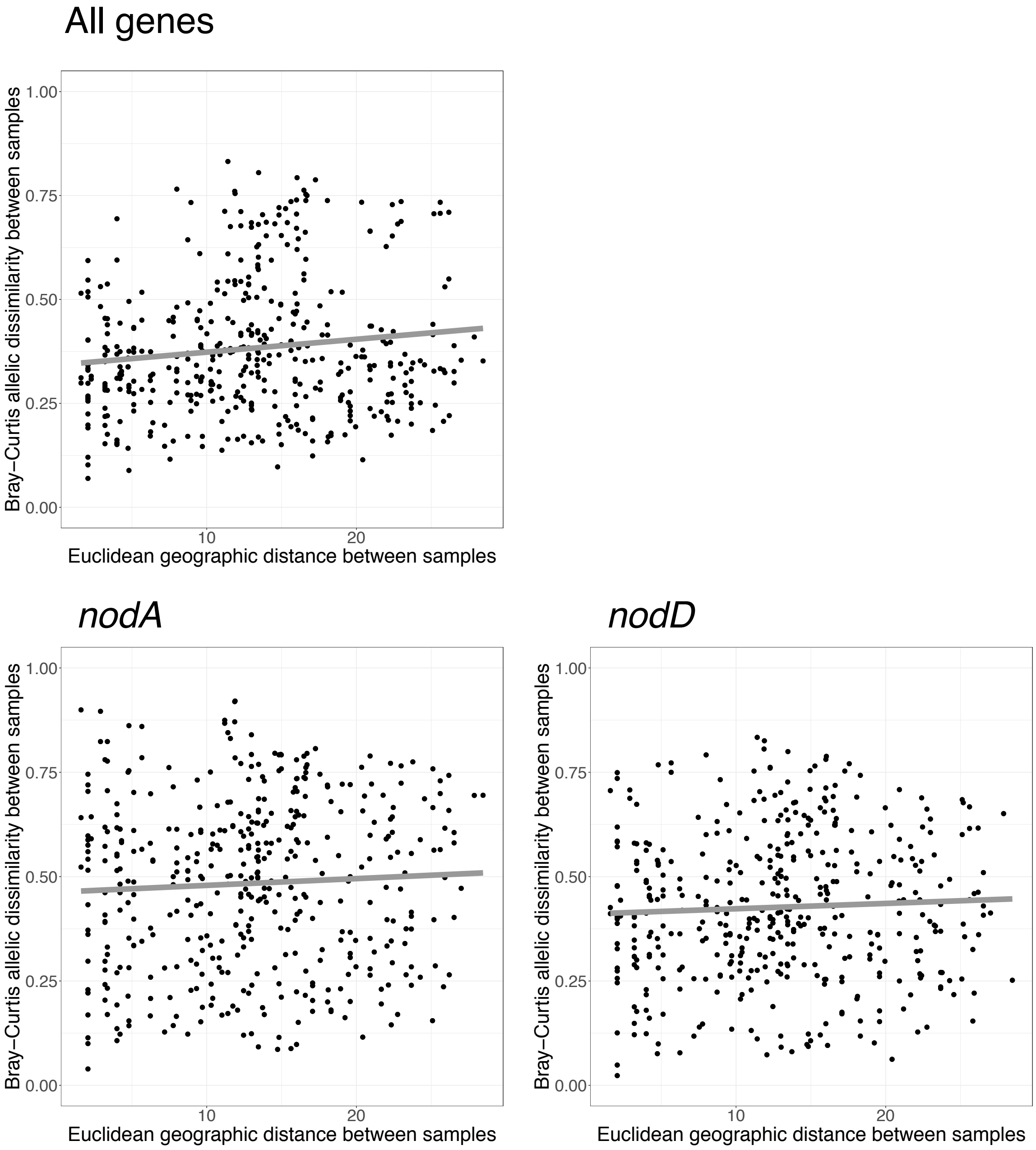
Appendix Figure B.4 Non-metric multidimensional scaling (NMDS) with intrinsic sequence vector variables fitted to NMDS coordinates. Allele sequences are numbered from greatest total abundant in the dataset to smallest, and therefore seq_1 in each analysis corresponds to the sequence with the greatest total abundance across all samples. Sequences with a fit p < 0.05 are displayed. Length of arrows and direction corresponds to scaled correlation coefficient (strong predictors have longer arrows) of fitted vectors and direction of vector correlation with NMDS coordinates.



Appendix Figure B.5 Principal Components Analysis (PCA) of relative allele abundance of *Rlt* genes *rpoB, recA*, *nodD, nodA* from nodule samples; and all genes in combination (all genes). Additionally, all Klondike samples from 3 plots were analysed by PCA using relative abundance of all 4 genes in combination. Samples are grouped by their respective plot (n = 3, 2 plots per clover genotype) and coloured by their clover genotype host.



Appendix Figure B.6 Pairwise Fixation index (F_{ST}) of 4 *Rlt* genes; *rpoB*, *recA*, *nodA* and *nodD*, and all genes in combination across samples from Block 1. Additionally, F_{ST} was calculated for all pairwise Klondike samples from three plots using all 4 genes in combination. F_{ST} is shown on a scale ranging from similar (red) to different (white) allelic diversity score. Samples are grouped and coloured by their clover genotype host.



Appendix Figure B.7 Euclidean geographic distance correlated to allelic (Bray-Curtis) dissimilarity. Relative abundance UMI sequence counts for *nodA* and *nodD* were considered individually, and all four genes were considered in combination (*rpoB*, *recA*, *nodA* and *nodD*). Correlation was calculated using Mantel's R statistic. All samples from Block 1 were used for pairwise comparisons.

Appendix C. Chapter 4

Appendix Table C.1 *Rlt* strain metadata. Strain names are provided, along with genospecies classification, phylogenetic branch grouping, strain country and farm management origin and experiment sequencing batch. DKO = Denmark organic management, DKC = Denmark conventional management, FRC = French conventional management, UKC = United Kingdom conventional management. C* = genospecies C strains in experimental batch 2 were treated as their own individual group. gsC* strains SM158 and SM170C in batch 2 were in duplicate (rep1, rep2).

Strain	Genospecies	Phylogenetic branch	Origin	Sequencing Batch
SM128A	A	A1	DKO	1
SM140A	A	A2	DKO	1
SM151A	A	A1	DKO	1
SM152B	A	A3	DKO	1
SM154C	A	A3	DKO	1
SM155A	A	A2	DKO	1
SM12	B	B3	UKC	1
SM15	B	B1	UKC	1
SM3	B	B2	UKC	1
SM38	B	B1	UKC	1
SM7	B	B2	UKC	1
SM122A	C	C7	DKO	1
SM157B	C	C7	DKO	1
SM158	C	C7	DKO	1
SM170C	C	C6	DKO	1
SM41	C	C1	DKC	1
SM53	C	C1	DKC	1
SM74	C	C1	DKC	1
SM101	C*	C5	FRC	2
SM105	C*	C3	FRC	2
SM107	C*	C6	FRC	2
SM111	C*	C6	FRC	2
SM112	C*	C3	FRC	2
SM113	C*	C9	FRC	2
SM114	C*	C9	FRC	2
SM115	C*	C9	FRC	2
SM116	C*	C9	FRC	2
SM118	C*	C1	FRC	2
SM119	C*	C3	FRC	2
SM121A	C*	C6	DKO	2
SM122A	C*	C7	DKO	2
SM125	C*	C1	DKO	2
SM126B	C*	C7	DKO	2
SM127	C*	C1	DKO	2
SM132	C*	C10	DKO	2
SM134A	C*	C7	DKO	2

SM143	C*	C4	DKO	2
SM147A	C*	C8	DKO	2
SM148A	C*	C7	DKO	2
SM148B	C*	C7	DKO	2
SM149B	C*	C7	DKO	2
SM149C	C*	C7	DKO	2
SM151C	C*	C8	DKO	2
SM153A	C*	C7	DKO	2
SM153C	C*	C7	DKO	2
SM153D	C*	C7	DKO	2
SM157B	C*	C7	DKO	2
SM158 rep1	C*	C7	DKO	2
SM158 rep2	C*	C7	DKO	2
SM164A	C*	C4	DKO	2
SM165A	C*	C7	DKO	2
SM166A	C*	C8	DKO	2
SM168C	C*	C8	DKO	2
SM170A	C*	C6	DKO	2
SM170C rep1	C*	C6	DKO	2
SM170C rep2	C*	C6	DKO	2
SM41	C*	C1	DKC	2
SM42	C*	C1	DKC	2
SM43	C*	C1	DKC	2
SM44	C*	C1	DKC	2
SM46	C*	C1	DKC	2
SM48	C*	C6	DKC	2
SM50	C*	C1	DKC	2
SM53	C*	C1	DKC	2
SM54	C*	C1	DKC	2
SM55	C*	C1	DKC	2
SM59	C*	C1	DKC	2
SM66	C*	C1	DKC	2
SM70	C*	C1	DKC	2
SM71	C*	C1	DKC	2
SM80	C*	C1	DKC	2
SM88	C*	C3	FRC	2
SM89	C*	C6	FRC	2
SM90	C*	C9	FRC	2
SM91	C*	C9	FRC	2
SM94	C*	C9	FRC	2
SM95	C*	C3	FRC	2
SM96	C*	C9	FRC	2
SM97	C*	C9	FRC	2
SM164B	D	D1	DKO	1

SM51	D	D1	DKC	1
SM72	D	D2	DKC	1
SM78	D	D2	DKC	1
SM79	D	D1	DKC	1
SM126A	E	E2	DKO	1
SM135B	E	E3	DKO	1
SM168A	E	E2	DKO	1

Appendix Table C.2 Number of genes contributing more than they would on average (if all genes contributed equally) to Principal Component Analysis (PCA) PCs. If all genes contributed equally to a PC, they would each contribute 0.0236% to a PC. PCs were generated based on PoissonSeq normalised Log_2 transformed read counts for 4,229 core genes.

	Between genospecies PCA		**Within gsC PCA**	
4,229 core gene PCA Principal component	**Number of genes contributing more than average (average contribution would be 0.0236%)**	**Number of genes contributing more than average with a metacyc ID**	**Number of genes contributing more than average (average contribution would be 0.0236%)**	**Number of genes contributing more than average with a metacyc ID**
PC1	1043	484	842	375
PC2	1092	498	1104	470
PC3	1064	496	1121	535
PC4	902	399	923	440
PC5	1068	505	1006	484
PC6	1008	475	770	358

Appendix Table C.3 Pearson's correlation coefficient (R) between WGCNA *Rlt* core gene module and Tryptone Yeast (TY) broth growth condition correlations that remain significant after Benjamini-Hochberg correction (adjusted p-value). R correlations and adjusted p-values are shown in **Figure 4.3**.

Module	**TY broth growth condition**	**R**	**p-value**	**adjusted p-value**
Module 24	12.5% TY 28°C	0.628	0.000596	0.03008639
Module 43	100% TY pH6	0.639	0.000441	0.02503724
Module 43	100% TY 28°C	0.602	0.00113	0.03930164
Module 43	25% TY 28°C	0.671	0.000176	0.01371122
Module 16	100% TY pH5	0.719	0.0000346	0.00539527
Module 16	100% TY pH6	0.664	0.000218	0.01509472
Module 16	100% TY 28°C	0.727	0.0000255	0.00531337
Module 16	12.5% TY 28°C	0.611	0.000914	0.03356538
Module 16	100% TY 15 °C	0.614	0.000855	0.03335323
Module 16	100% TY 20°C	0.777	0.00000305	0.00095036
Module 16	25% TY 28°C	0.672	0.000171	0.01371122
Module 9	100% TY pH5	-0.687	0.000107	0.01338136
Module 9	100% TY pH6	-0.657	0.000265	0.01654357
Module 9	100% TY 28°C	-0.617	0.000796	0.03309899

Module 9	100% TY 20°C	-0.789	0.00000167	0.00095036
Module 9	25% TY 28°C	-0.587	0.00162	0.0480863
Module 8	25% TY 28°C	0.625	0.000639	0.03008639
Module 10	6.25% TY 28°C	-0.596	0.00132	0.04106391
Module 20	100% TY pH6	-0.583	0.00178	0.0488946
Module 20	100% TY 20°C	-0.623	0.000675	0.03008639
Module 20	25% TY 28°C	-0.673	0.000166	0.01371122
Module 28	6.25% TY 28°C	-0.582	0.00180	0.0488946
Module 1	100% TY 20°C	-0.597	0.00129	0.04106391

Appendix Table C.4 Two-way ANOVA for genospecies eigengene value differences for different modules.

	Df	Sum Sq	Mean Sq	F value	P-value
genospecies	4	0.201	0.05016	2.089	0.0802
module	47	0	0	0	1
genospecies*module	188	23.602	0.12554	5.23	<2e-16
Residuals	1008	24.197	0.02401		

Appendix Table C.5 TukeyHSD post hoc for genospecies eigengene value differences for different modules.

Module	Genospecies comparison	diff	lwr	upr	p adj
3	D - B	0.49323784	0.03391945	0.95255622	0.01138355
8	C - B	-0.485699	-0.9109452	-0.0604528	0.00244247
8	D - B	-0.4776748	-0.9369932	-0.0183564	0.02303441
9	C – B	0.43835275	0.01310656	0.86359894	0.02765643
9	D – C	-0.4300967	-0.8553429	-0.0048505	0.04033378
11	D - B	-0.5039173	-0.9632357	-0.0445989	0.00685657
15	C - A	0.47219861	0.06815257	0.87624466	0.00129313
16	C – A	-0.4406899	-0.8447359	-0.0366438	0.00789866
16	C - B	-0.4458662	-0.8711124	-0.02062	0.0193678
20	C – B	0.5023904	0.07714421	0.92763659	0.00094602
23	D – B	-0.5015916	-0.9609099	-0.0422732	0.00766878
27	B – A	0.45271656	0.01295269	0.89248043	0.02841528
27	C – A	0.40498416	0.00093811	0.8090302	0.04786713
33	D – A	-0.5293631	-0.969127	-0.0895992	0.000541
40	B - A	-0.4761462	-0.9159101	-0.0363823	0.00939912
no module group	E - A	0.54254786	0.0290143	1.05608142	0.01641593
no module group	E - D	0.55155966	0.02118447	1.08193484	0.02304465

Appendix Table C.6 Dunn's post hoc for Module 16 mean expression difference between genospecies. P-value correction using Benjamini-Hochberg.

Genospecies Comparison	Z	Unadjusted p	Adjusted p
A - B	6.32E-02	9.50E-01	0.94960567
A - C	3.52E+00	4.33E-04	0.00216518
B - C	3.28E+00	1.04E-03	0.00391648
A - C*	4.29E+00	1.80E-05	0.00026972
B - C*	3.86E+00	1.12E-04	0.00084276
C - C*	-3.08E-01	7.58E-01	1
A - D	5.34E-01	5.93E-01	0.98896279
B - D	4.51E-01	6.52E-01	0.97830351
C - D	-2.7914693	5.25E-03	0.01311733
C* - D	-3.2494306	1.16E-03	0.00346909
A - E	0.27061471	7.87E-01	0.98335922
B - E	0.20961726	8.34E-01	0.96226894
C - E	-2.55988	1.05E-02	0.01963281
C* - E	-2.7791773	5.45E-03	0.01167788
D - E	-0.1807045	8.57E-01	0.91778518

Appendix Table C.7 Dunn's post hoc for Module 9 mean expression difference between genospecies. P-value correction using Benjamini-Hochberg.

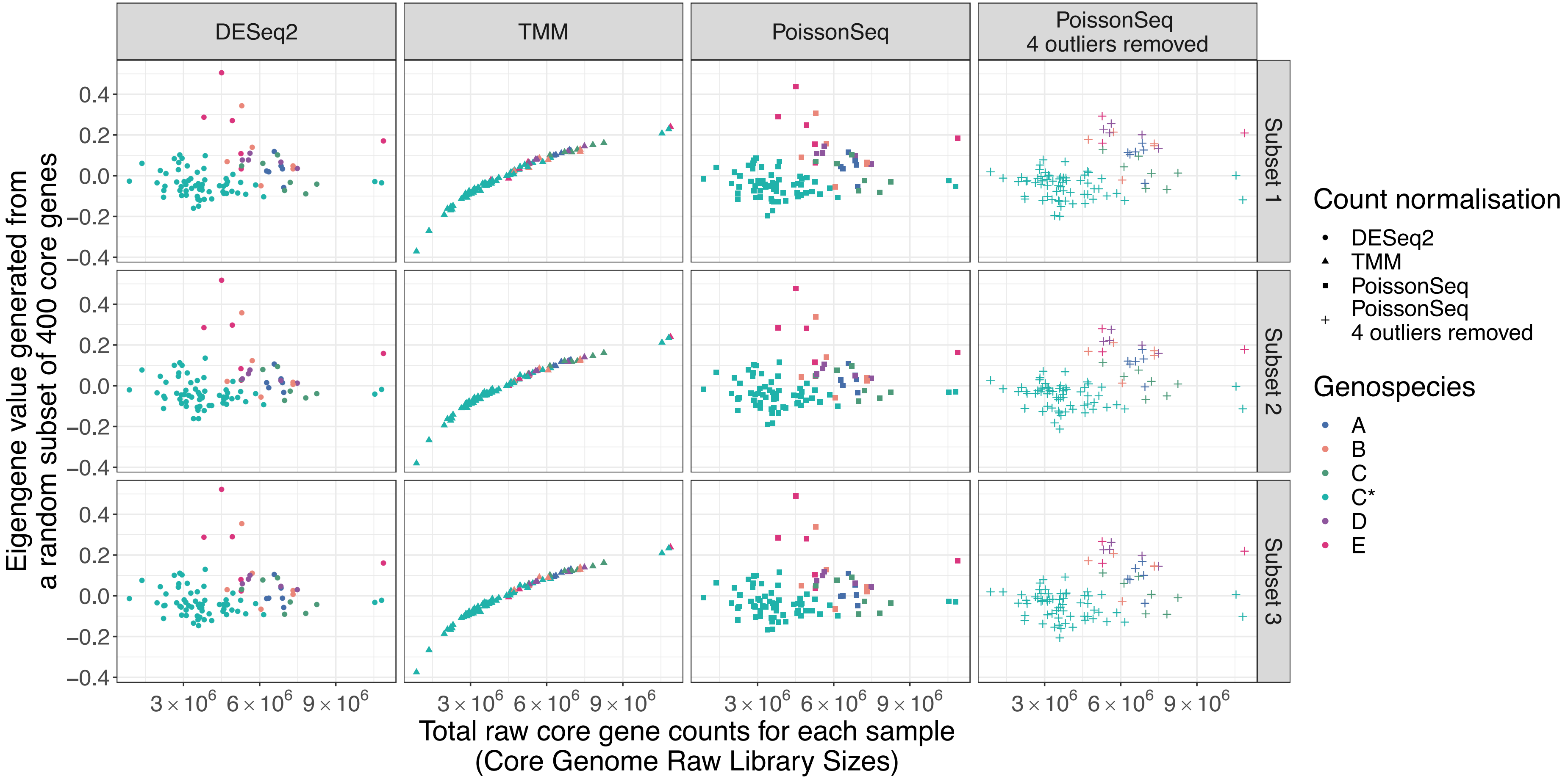
Genospecies Comparison	Z	Unadjusted p	Adjusted p
A - B	4.55E-01	0.64875666	0.81094583
A - C	-3.55E+00	0.00039084	0.00117251
B - C	-3.84E+00	0.00012281	0.00184218
A - C*	-3.48E+00	0.00050223	0.00125558
B - C*	-3.79E+00	0.00014867	0.00111502
C - C*	1.21E+00	0.22502589	0.37504315
A - D	4.03E-01	0.6868121	0.7924755
B - D	-5.01E-02	0.96005999	0.96005999
C - D	3.79E+00	0.00015287	0.00076437
C* - D	3.73E+00	0.00019516	0.00073187
A - E	-3.45E-01	0.72989362	0.78202888
B - E	-7.12E-01	0.47647973	0.71471959
C - E	2.51E+00	0.01223683	0.02622178
C* - E	2.10E+00	0.03532659	0.06623736
D - E	-6.69E-01	0.50374636	0.68692685

Appendix Table C.8 Linear mixed effects model for gene expression level of core and accessory genes. Expression levels were log transformed and PoissonSeq normalised read counts. Ortholog gene group ID and strain ID were classed as random effects. Model formula: expression_level ~ gene_type + (1|geneID) + (1|strainID).

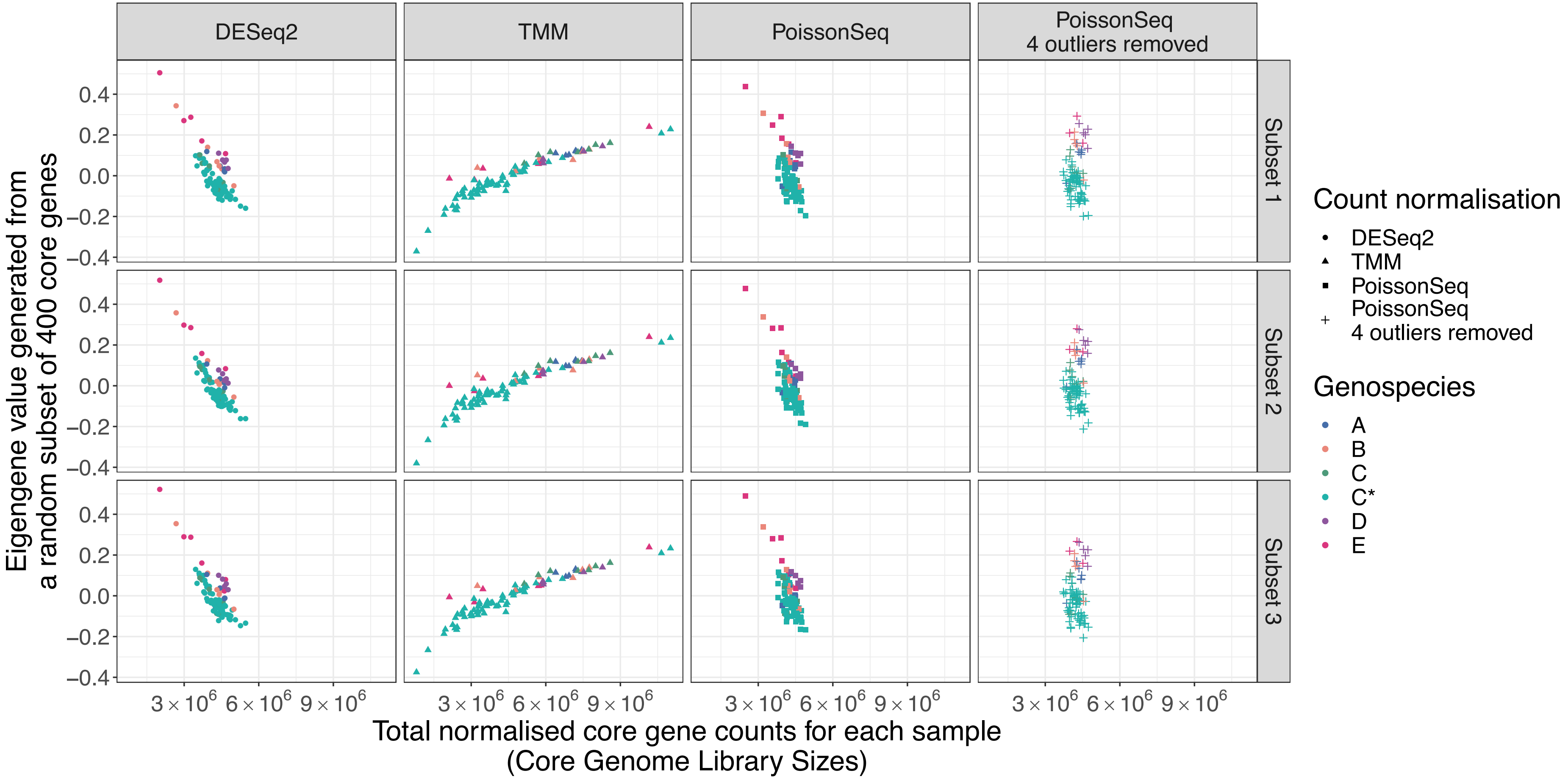
AIC	BIC	Effect	Variable	Variance	Std. Dev	Estimate	Std. Error	df	t value	p-value
1638349.3	1638406.1	random	geneID (Intercept)	4.83723	2.1994	/	/	/	/	/
		random	strainID (Intercept)	0.05617	0.2370	/	/	/	/	/
		random	Residual	0.66190	0.8136	/	/	/	/	/
		fixed	Intercept	/	/	5.986	0.03098	0.01914	193.19	<0.0001
		fixed	genetypeCore	/	/	2.482	0.03821	0.0001905	64.97	<0.0001

Appendix Table C.9 Dunn's post hoc test for accessory genome size differences between genospecies. P-value correction using Benjamini-Hochberg.

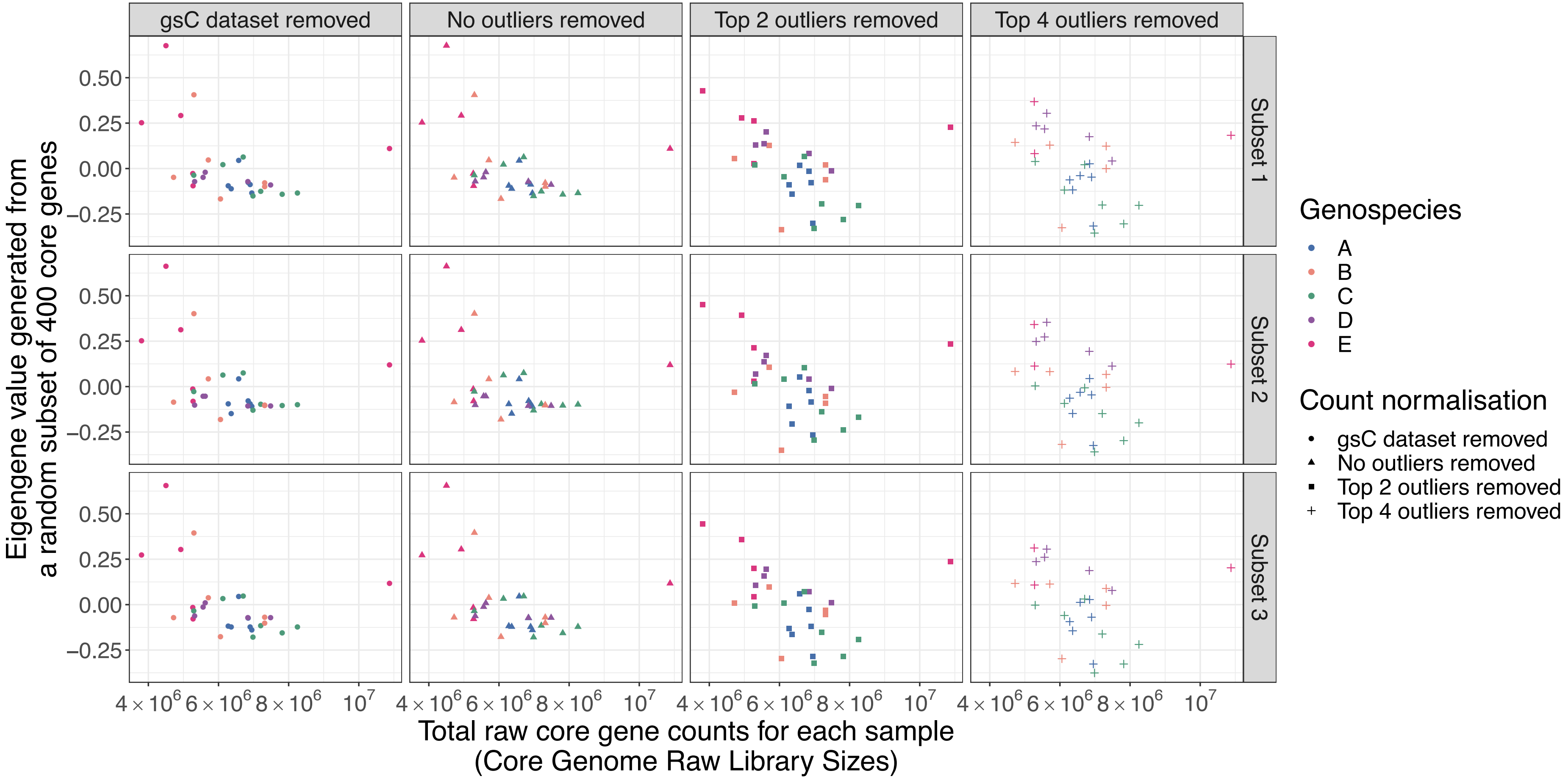
Genospecies Comparison	**Z**	**Unadjusted p**	**Adjusted p**
A – B	-2.6690262	7.61E-03	0.01901788
A – C	-4.1982506	2.69E-05	0.00020174
B – C	-1.2288065	2.19E-01	0.29883325
A – C*	-4.2554381	2.09E-05	0.00031296
B – C*	-0.439691	6.60E-01	0.70731531
C – C*	1.2904744	1.97E-01	0.29532899
A – D	-0.8511676	3.95E-01	0.49334533
B – D	1.7404671	8.18E-02	0.15333194
C – D	3.1087258	1.88E-03	0.0093948
C* - D	2.8060125	5.02E-03	0.01504762
A – E	-0.5459723	5.85E-01	0.67509803
B – E	1.6844042	9.21E-02	0.15350598
C – E	2.8252827	4.72E-03	0.0177146
C* - E	2.4259495	1.53E-02	0.03271799
D – E	0.1771155	8.59E-01	0.8594177



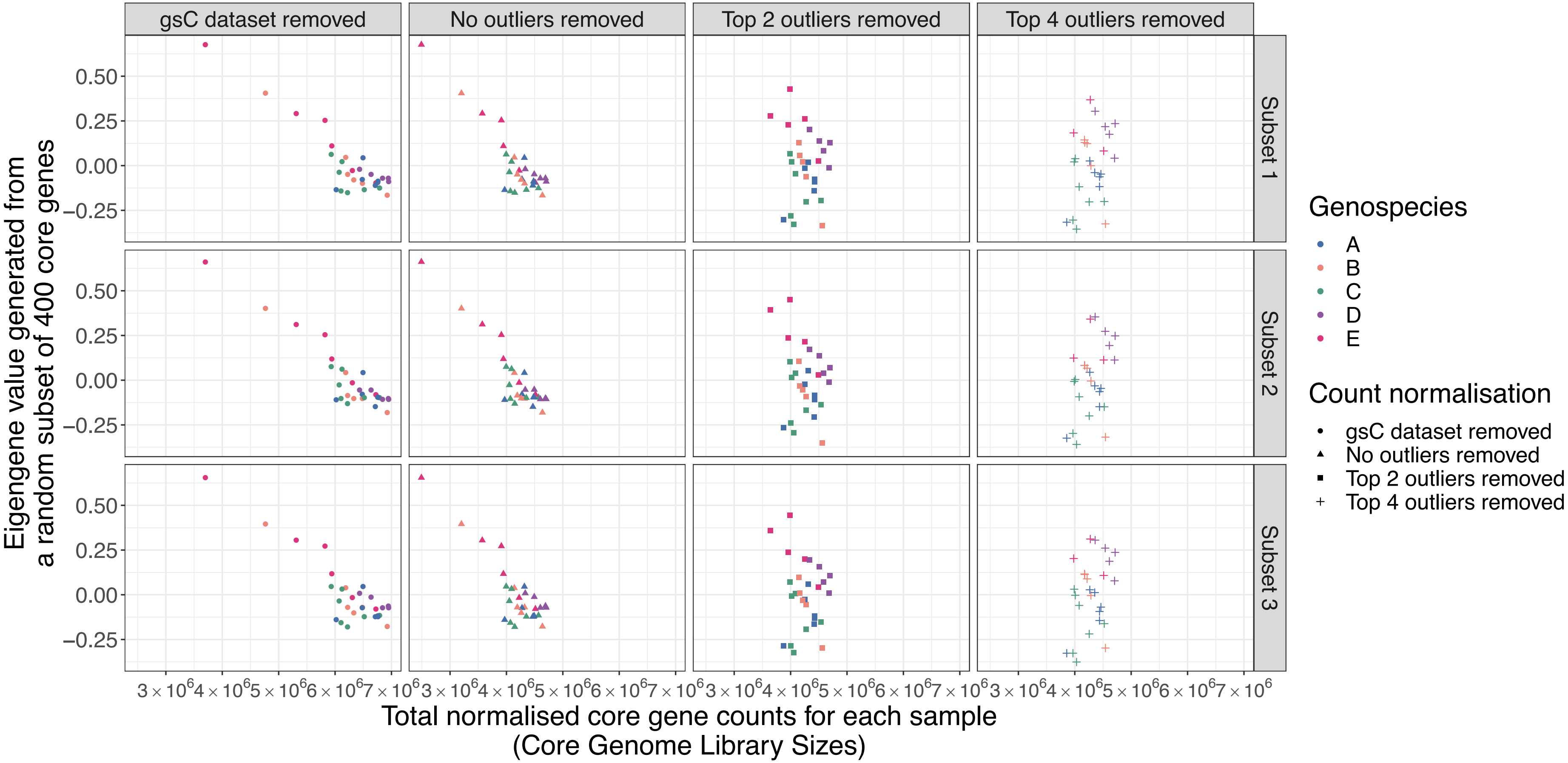
Appendix Figure C.1 Optimisation of raw gene expression count normalisation methods. DESeq2, TMM and PoissonSeq were tested. Total raw core gene counts for each sample correlated to eigengene values calculated using normalised DESeq2/TMM/PoissonSeq normalised, log transformed counts. The final column shows eigengene values calculated using PoissonSeq normalised counts once 4 identified outliers are removed. 3 subsets of 400 random core genes were selected to represent the normalisation across the 4,229 core gene expression data. Eigengenes for each subset were calculated for each strain. Eigengenes were calculated using the expression data normalised by one of the three methods, and then log transformed. Strains are coloured by their genospecies. The dataset contained two sample batches. The first batch has the following number of samples; gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3. The second batch contains the following number of samples; gsC* = 59 + 2 biological duplicates.



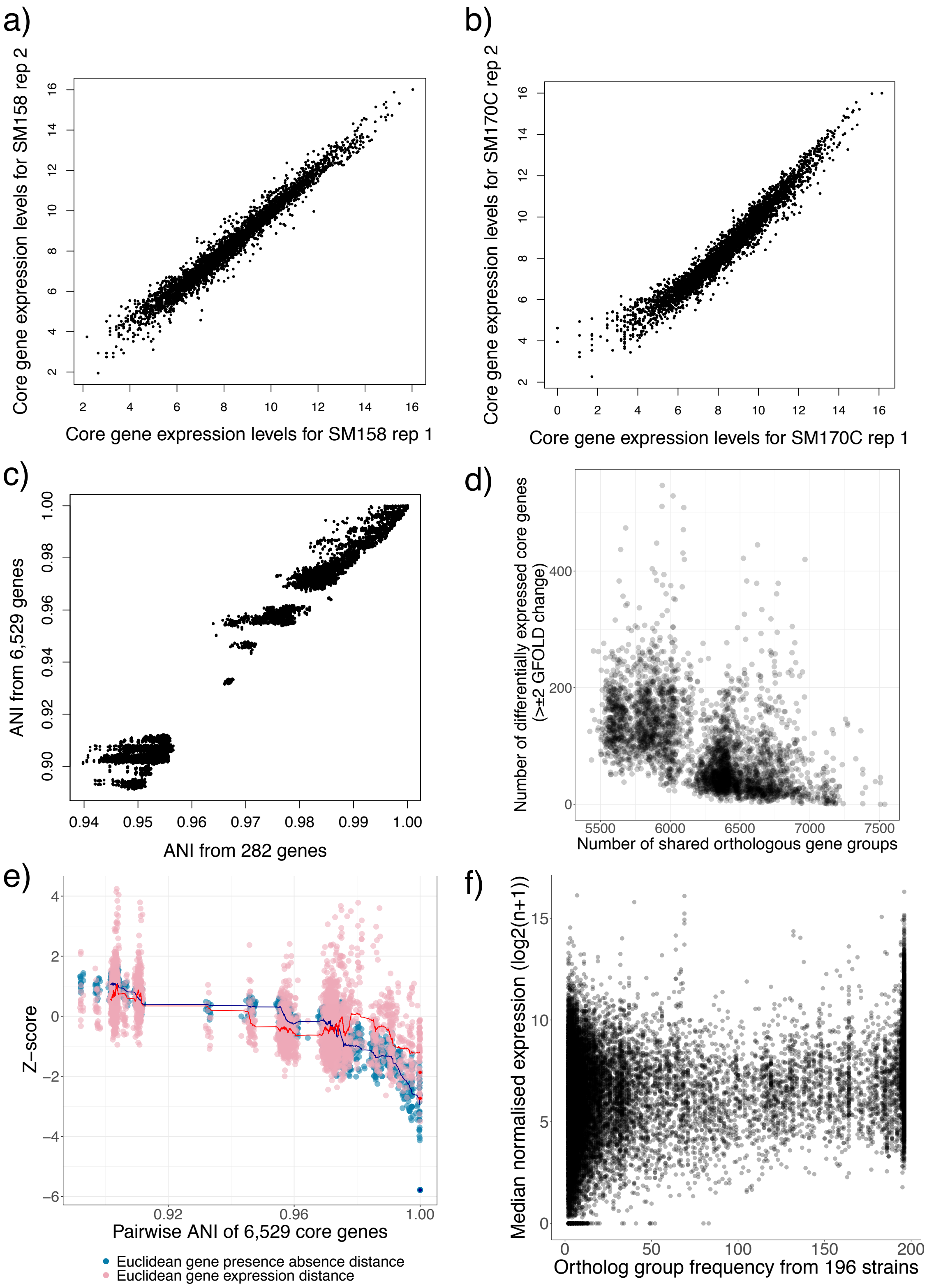
Appendix Figure C.2 Optimisation of raw gene expression count normalisation methods. DESeq2, TMM and PoissonSeq were tested. Total normalised core gene counts for each sample correlated to eigengene values calculated using normalised DESeq2/TMM/PoissonSeq normalised, log transformed counts. The final column shows eigengene values calculated using PoissonSeq normalised counts once 4 identified outliers are removed. 3 subsets of 400 random core genes were selected to assess the normalisation across the 4229 core gene expression data. Eigengenes for each subset were calculated for each strain. Eigengenes were calculated using the expression data that was normalised using one of the three methods, and then log transformed. Strains are coloured by their genospecies. The dataset contained two sample batches. The first batch has the following number of samples; gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3. The second batch contains the following number of samples; gsC* = 59 + 2 biological duplicates.



Appendix Figure C.3 Evaluation of PoissonSeq normalisation after sample removal. Total raw core gene counts for each sample correlated to eigengene values calculated using PoissonSeq normalised, log transformed counts. PoissonSeq normalisation was further tested to evaluate how removal of gsC samples and 2-4 outliers affected normalisation and therefore distribution of samples from the first dataset containing multiple genospecies samples. 3 subsets of 400 random core genes were selected to assess the normalisation across the 4229 core gene expression data. Eigengenes for each subset were calculated for each strain. Eigengenes were calculated using the expression data that was PoissonSeq normalised after the second batch of gsC samples were removed, no outliers were removed (control) and 2-4 outliers were removed. Strains are coloured by their genospecies. The dataset contained two sample batches. The first batch has the following number of samples; gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3. The second batch contains the following number of samples; gsC* = 59 + 2 biological duplicates.

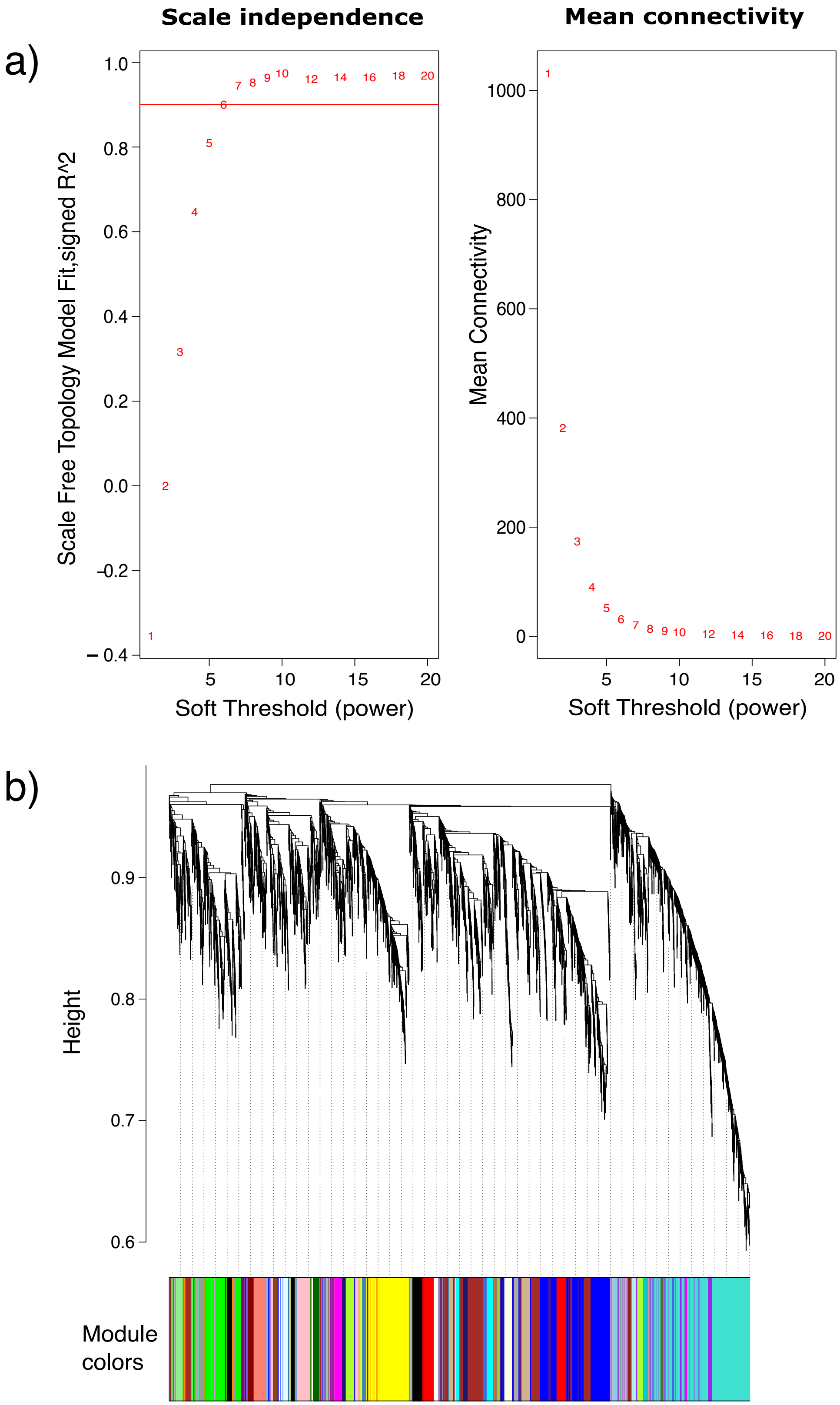


Appendix Figure C.4 Evaluation of PoissonSeq normalisation after sample removal. Total normalised core gene counts for each sample correlated to eigengene values calculated using PoissonSeq normalised, log transformed counts. PoissonSeq normalisation was further tested to evaluate how removal of gsC samples and 2-4 outliers affected normalisation and therefore distribution of samples from the first dataset containing multiple genospecies samples. Eigengenes were calculated using the expression data that was PoissonSeq normalised after the second batch of gsC samples were removed, no outliers were removed (control) and 2-4 outliers were removed. 3 subsets of 400 random core genes were selected to assess the normalisation across the 4229 core gene expression data. Eigengenes for each subset were calculated for each strain. Strains are coloured by their genospecies. The dataset contained two sample batches. The first batch has the following number of samples; gsA = 6, gsB = 5, gsC = 7, gsD = 5, gsE = 3. The second batch contains the following number of samples; gsC = 59 + 2 biological duplicate.

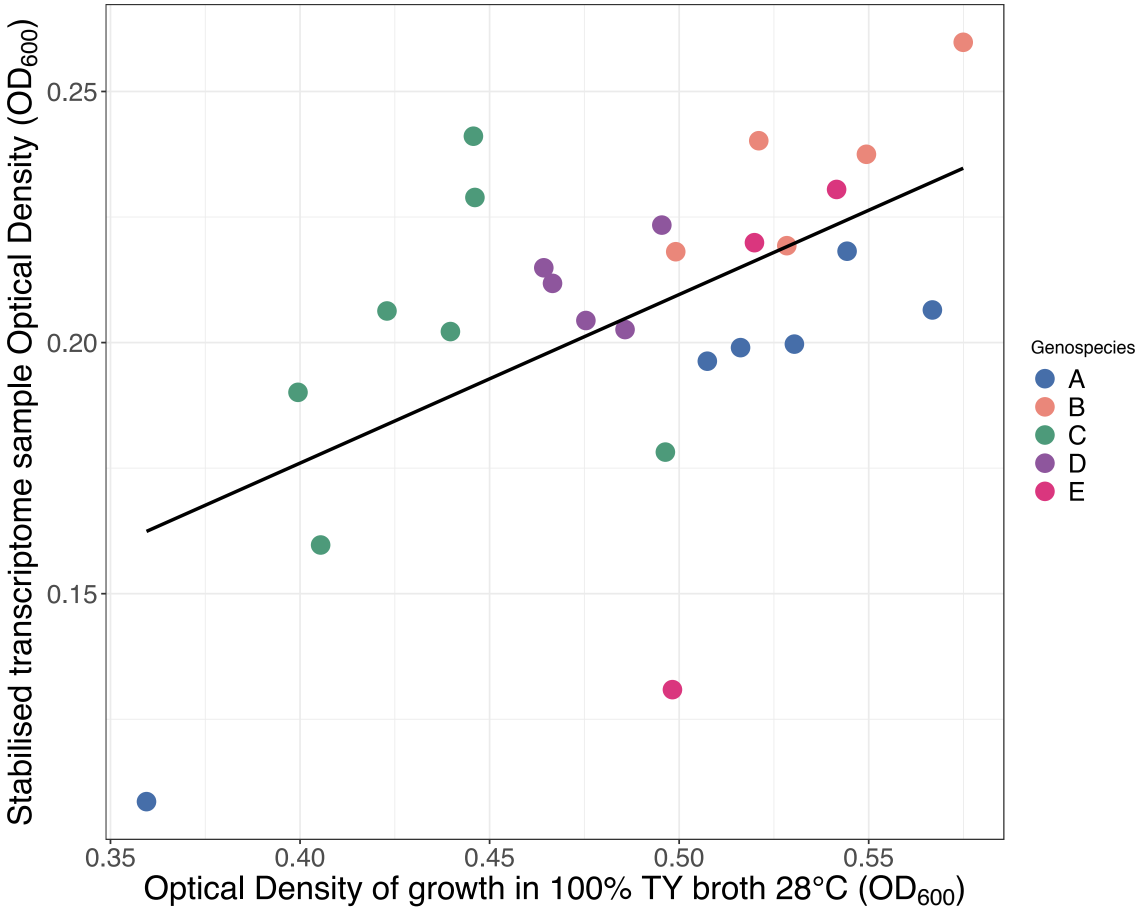


Appendix Figure C.5 Correlation of PoissonSeq normalised log transformed core gene counts for biological replicates of *Rlt* strains **a)** SM158 (gsC) and **b)** SM170C (gsC). **c)** Correlation of pairwise average nucleotide identity (ANI) values generated from 282 core genes from all 196 NCHAIN strains and 6529 genes present in at least 100 out of 196 NCHAIN strains. **d)** The number of shared orthologous genes between pairwise strain comparisons correlated to number of differentially expressed genes between pairwise strain comparisons. Genes were identified as differentially expressed if they had a GFOLD change of more than ±2. (Continued on following page).

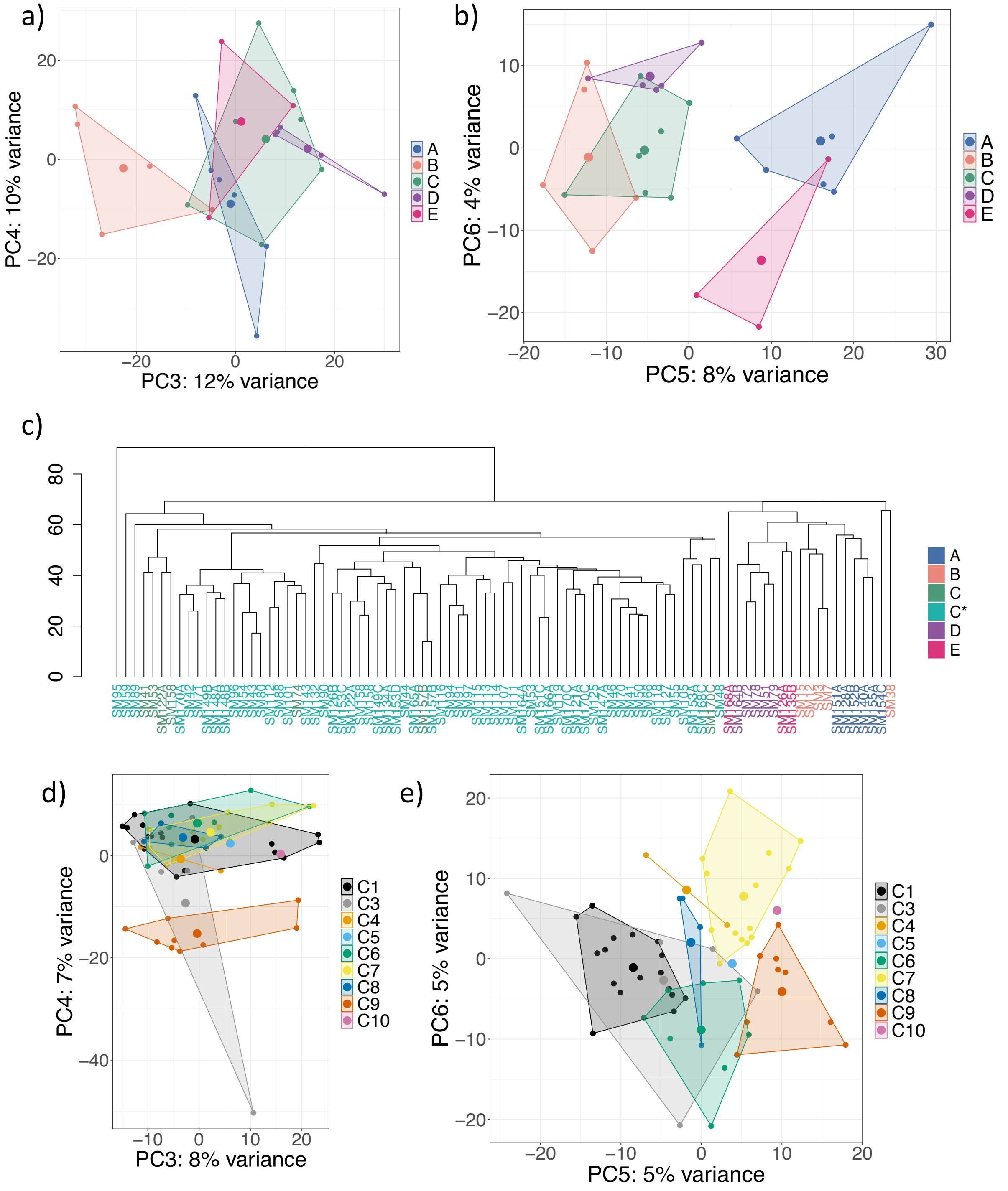
Appendix Figure C.5 continued. e) Correlating average nucleotide identity (ANI) based on 6,529 genes against z-scores for the Euclidean distance of number of shared orthologous gene groups (genes shared between strains) and z-scores for Euclidean distance of gene expression distance. Expression distance was calculated using only genes that were identified with a GFOLD > ±2 in at least one pairwise comparison. Red line displays the rolling average (n=100) for DEGs number. Blue line displays the rolling average (n=100) for number of shared orthologous gene groups. Red and dark blue dots highlight the strain comparisons which are biological replicates for number of core DEGs and number of shared orthologous gene groups, respectively. **f)** The number of *Rlt* NCHAIN genomes containing gene (196 maximum) correlated to median PoissonSeq normalised log transformed expression counts for gene in 85 strain samples. Pearson's Correlation Coefficient: R = 0.415, $p < 0.0001$.



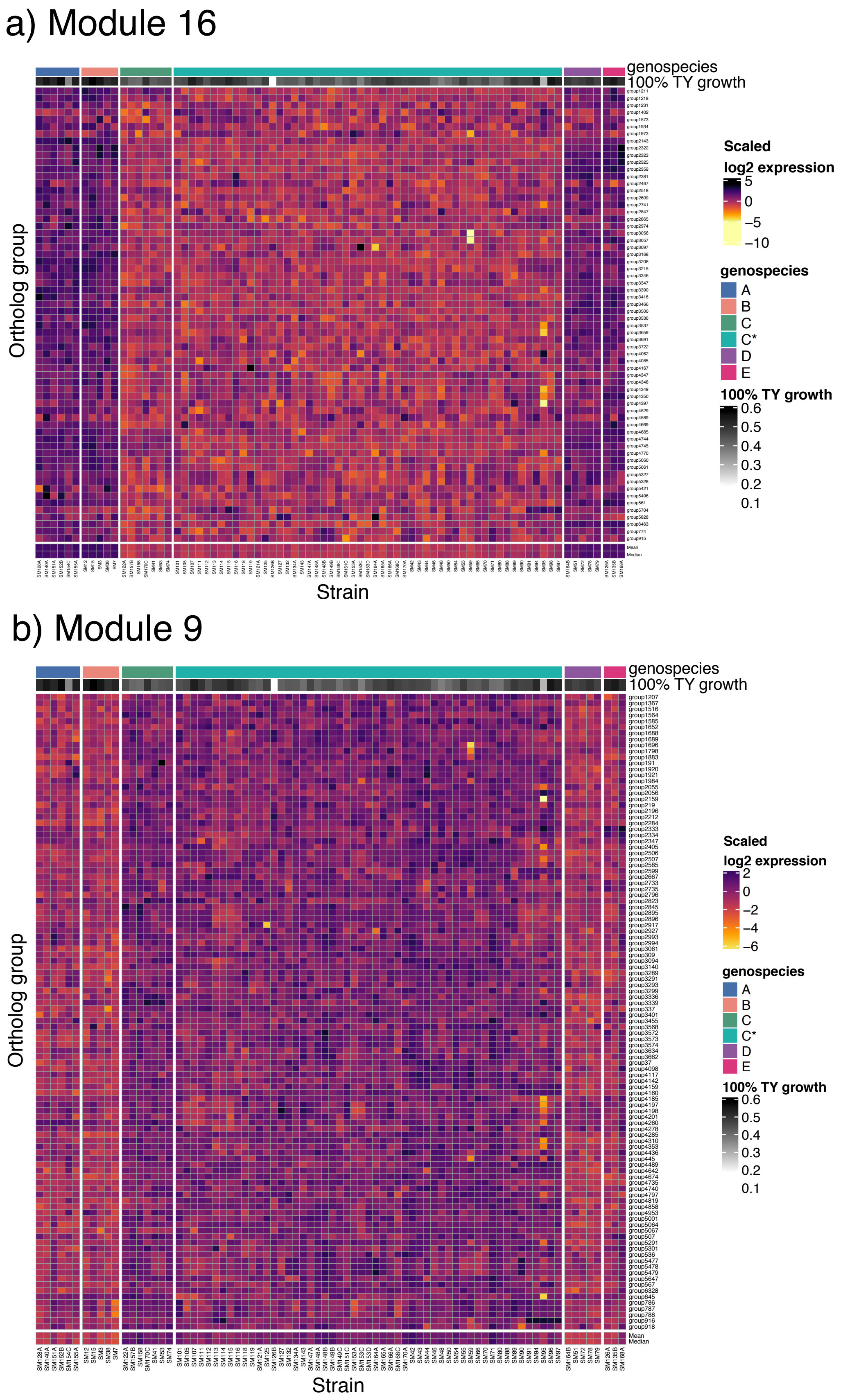
Appendix Figure C.6 WGCNA module and meta-module detection. **a)** Soft threshold power optimisation using scale independence and mean connectivity. **b)** WGCNA module detection using gene connectivity cluster dendrogram and height dendrogram tree cutting threshold. Clustering dendrogram. **c)** Meta-module detection using module eigengene clustering by correlation.



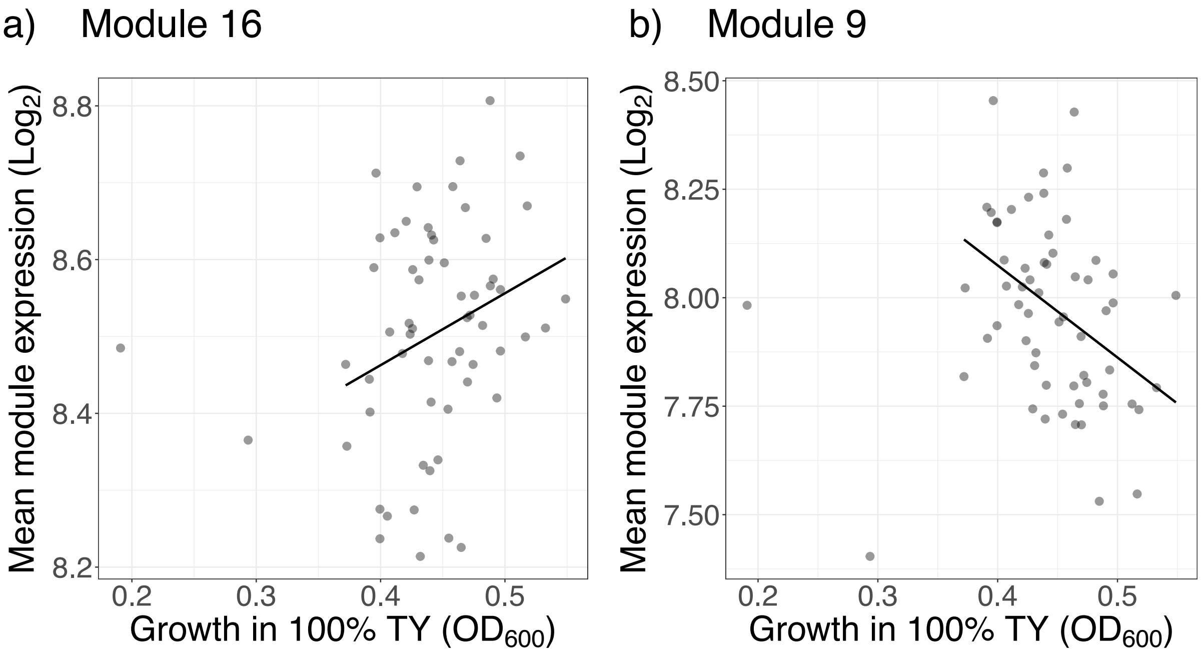
Appendix Figure C.7 Growth (OD_{600}) 26 *Rlt* strains in 200µl 100% Tryptone Yeast (TY) broth at 28°C after 48 hours correlated to growth (OD_{600}) of strains when grown for transcriptome analysis at the point of RNA stabilisation (48 hours). For transcriptome analysis, strains were grown in 5ml 100% TY broth + 1µM 7,4'-dihydroxyflavone (clover flavonoid stock concentration solubilised in DMSO) for 48 hours, 28°C, 180 rpm. Pearson's Correlation Coefficient: R = 0.55, p-value < 0.01.



Appendix Figure C.8 PCA of 4,229 core genes expression for 26 *Rlt* strains displaying **a)** principal components 3 and 4 and **b)** principal components 5 and 6. **c)** Core gene expression average linkage hierarchical clustering based on Euclidean distances calculated from Log$_2$(n+1) transformed normalised core gene expression values for all 85 *Rlt* samples. Additionally, PCA analysis of 4,229 core genes expression for 59 gsC strains (plus 2 strains in duplicate) displaying **d)** principal components 3 and 4 and **e)** principal components 5 and 6. Strains are coloured by their genospecies grouping in **a)**, **b)** and **c)** and by their gsC phylogenetic subbranch for **d)** to **e)**.

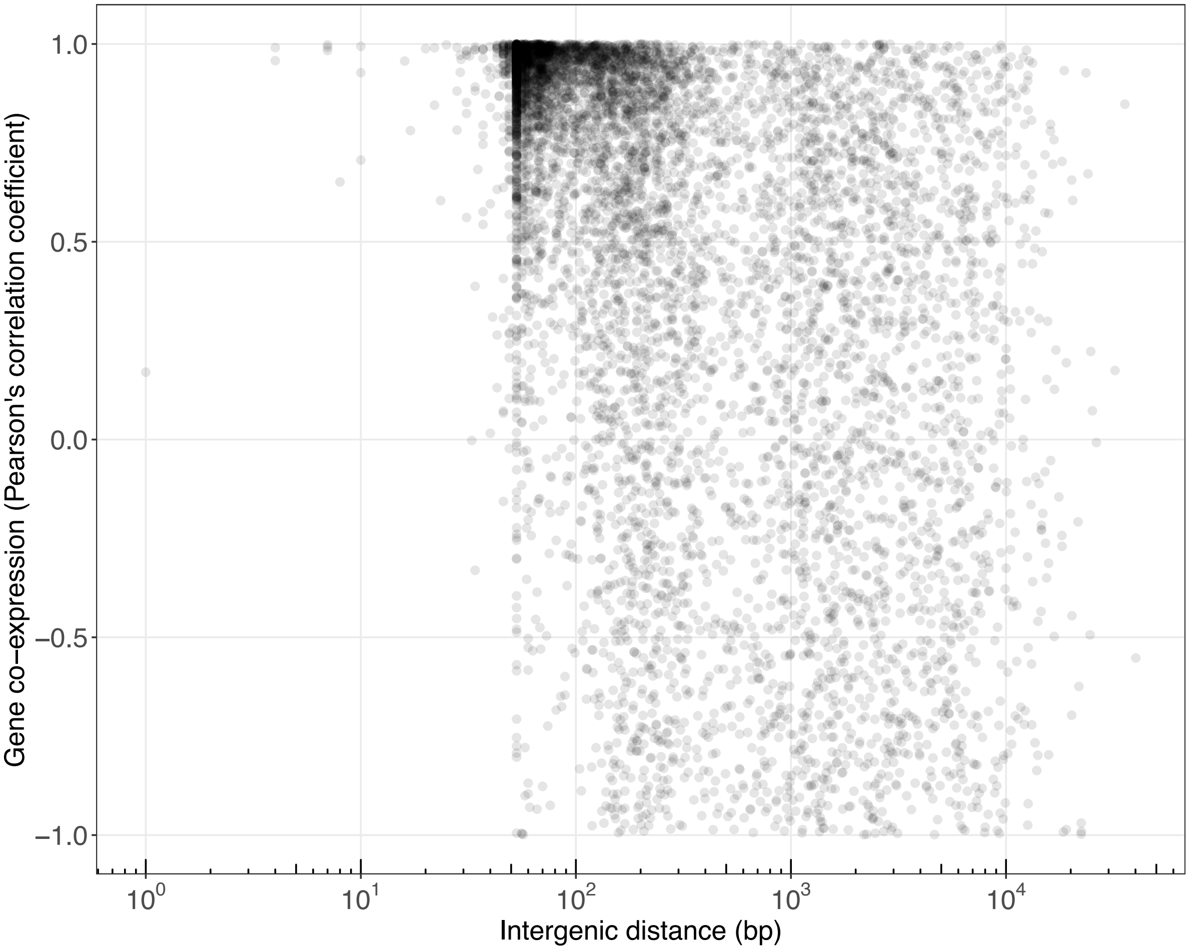


Appendix Figure C.9 Scaled, normalised Log$_2$(n+1) transformed gene expression counts for genes in **a)** Module 16 and **b)** Module 9. Strains are ordered and coloured by genospecies. gsC* = 59 strains. The growth of strains in 100% TY broth (OD$_{600}$) is displayed by increasing grey-scale colour intensity.

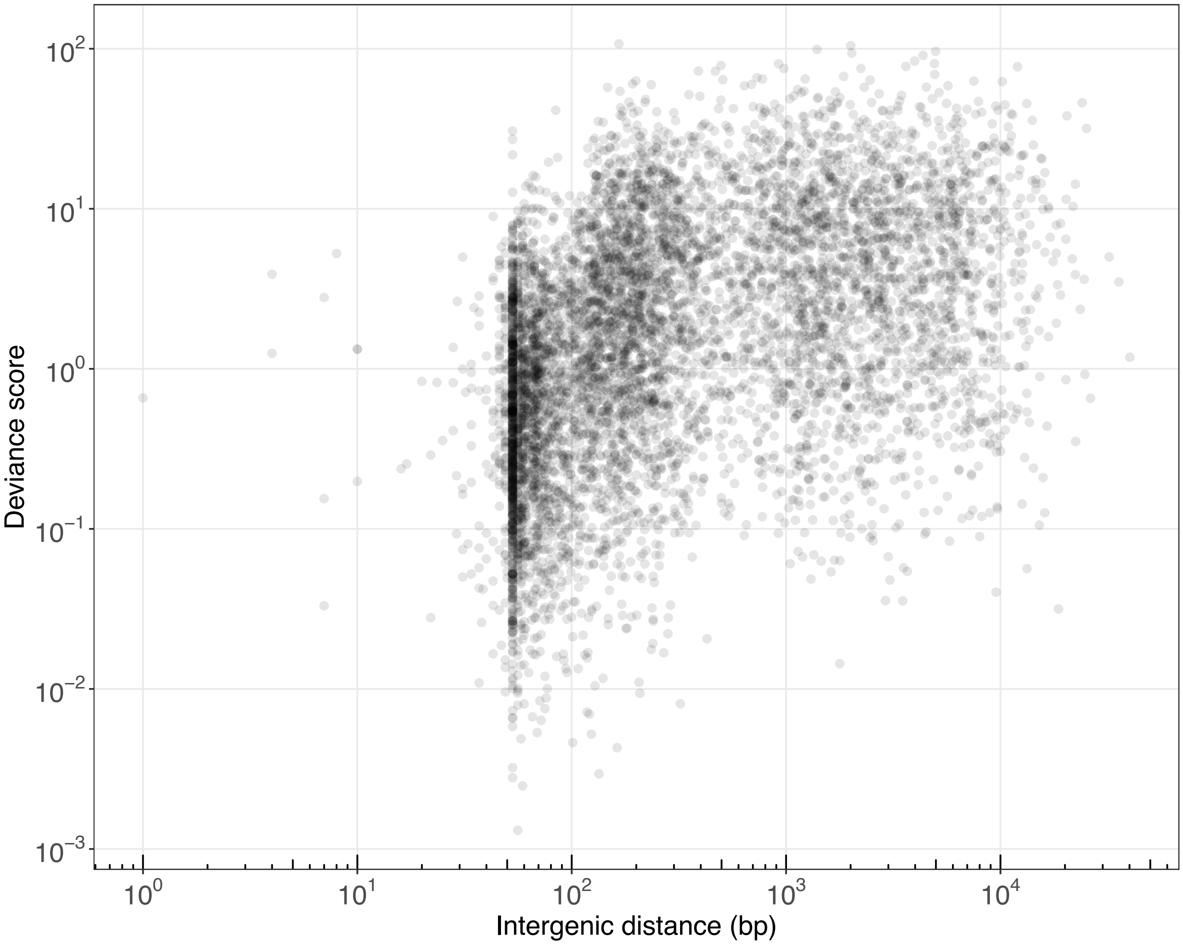


Appendix Figure C.10 a) A significant positive correlation was observed between growth of strains in TY broth and the mean expression of genes in Module 16 (Pearson's correlation: R =0.26, p< 0.05). **b)** A significant negative correlation observed between growth of strains in TY broth and mean expression of genes in Module 9 (Pearson's correlation R = -0.42, p < 0.001). Linear model line and Pearson's Correlation Coefficient excludes outliers identified as growth less than 0.3 OD$_{600}$.

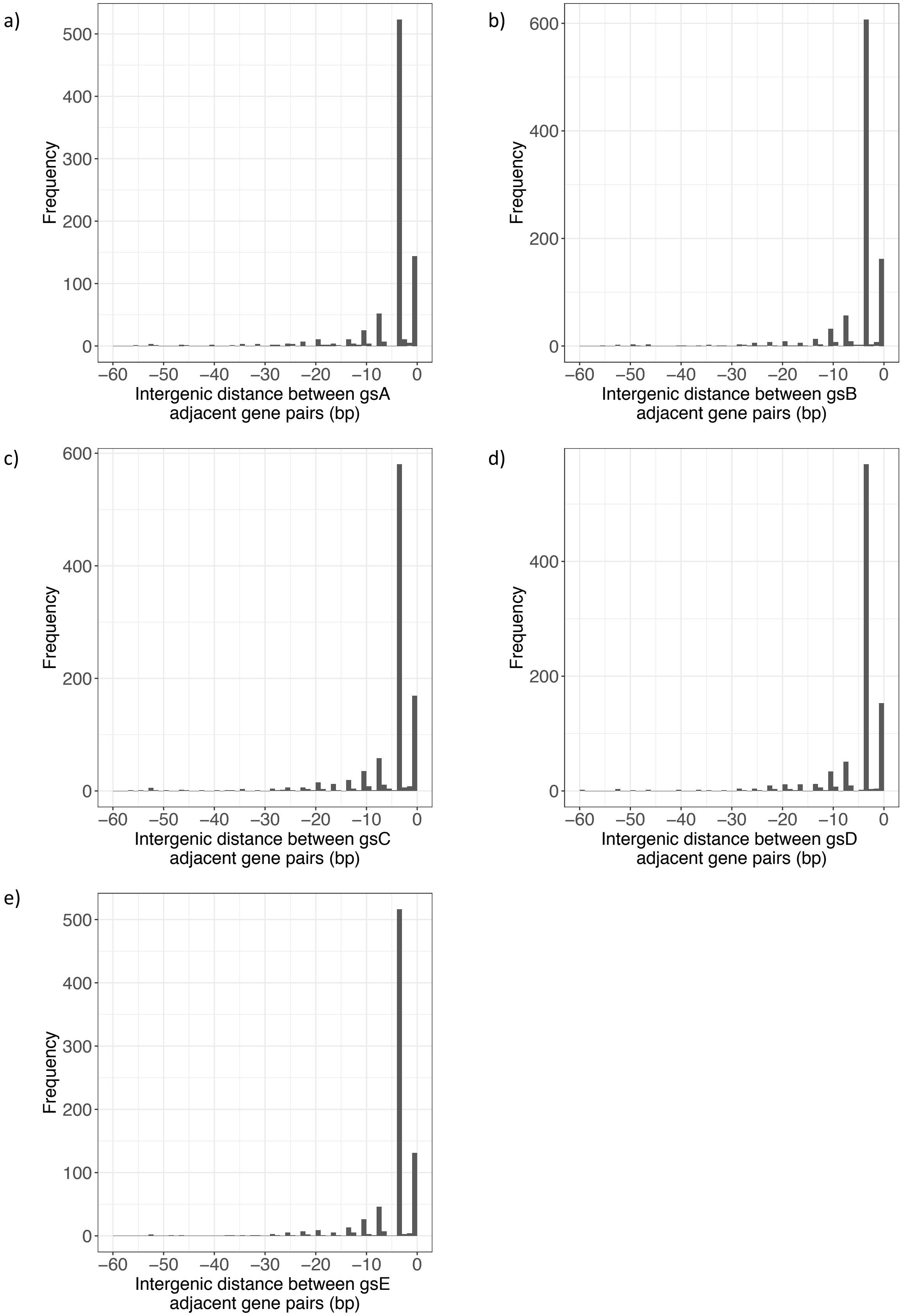
Appendix D. Chapter 5



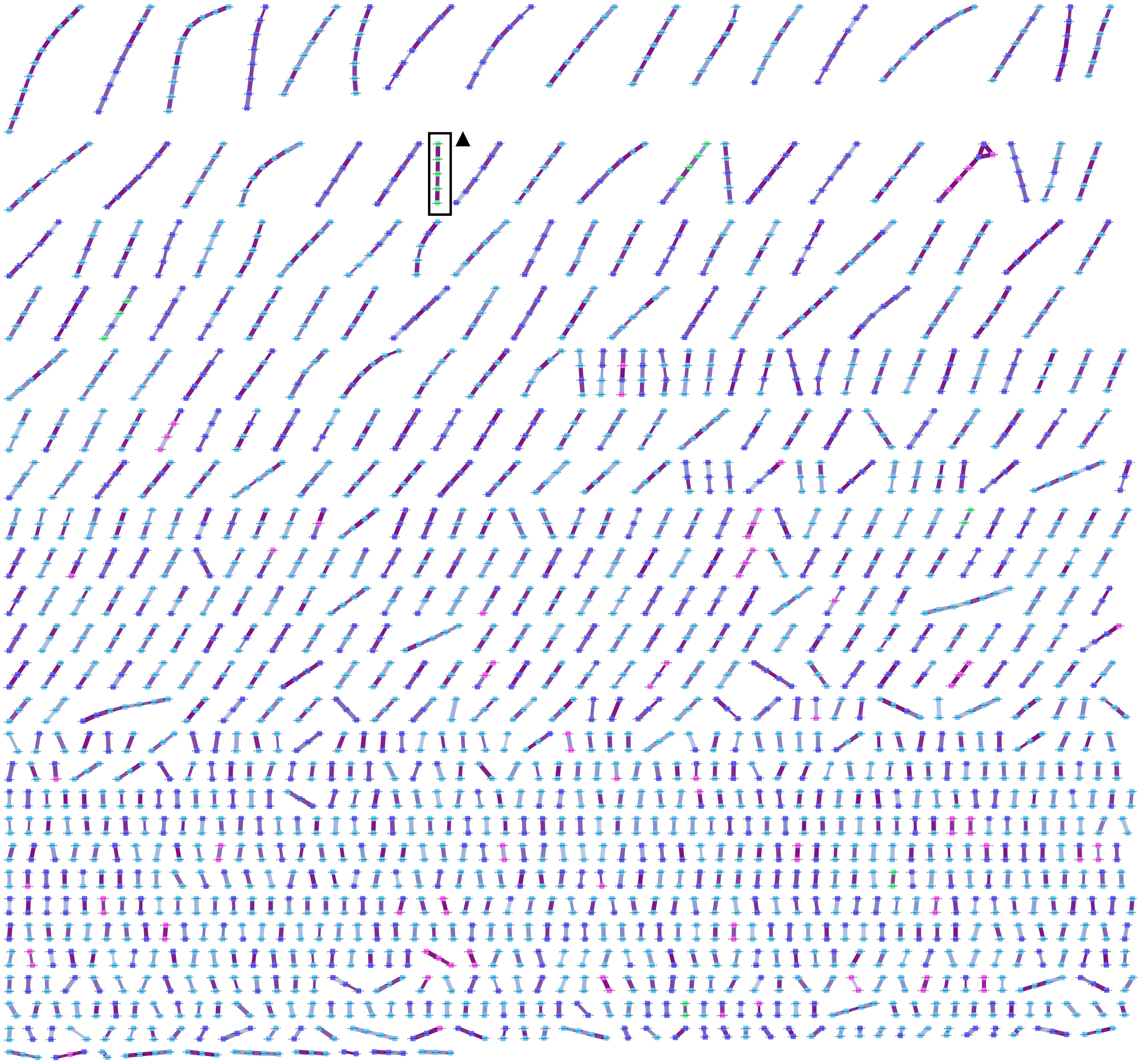
Appendix Figure D.1 Correlation between gsB adjacent genes intergenic distance and Pearson's correlation R statistic value. Intergenic distance axis is presented on a $\log_{10}$ scale. To log the intergenic distance axes, first a constant was added to all intergenic distance values of the absolute value of the most negative intergenic distance measure (for gsB this was 55 bp) plus 1, so that all distances were above 1 for log transformation of the axis.



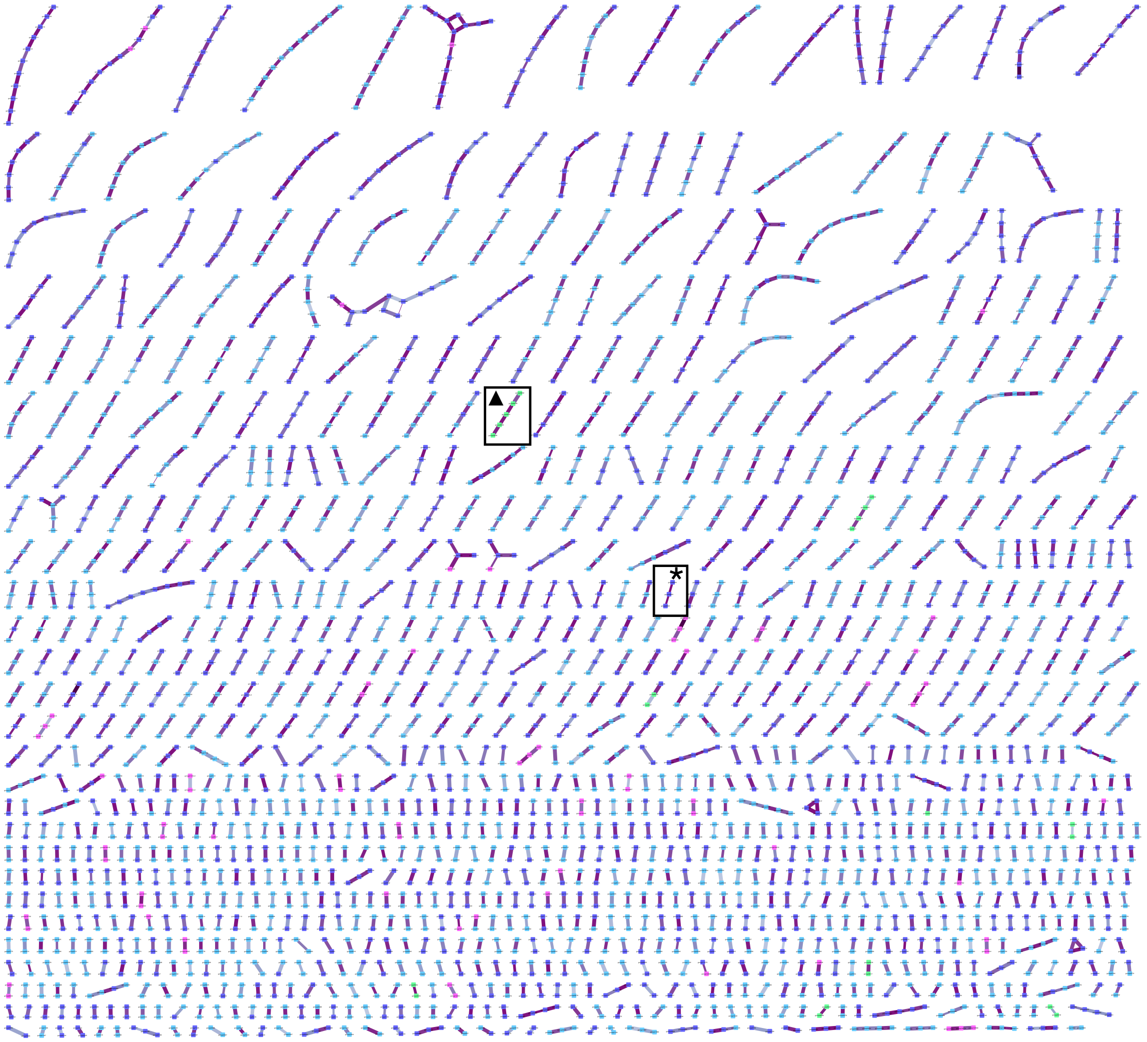
Appendix Figure D.2 Correlation between gsB adjacent genes intergenic distance and deviance score. Intergenic distance and deviance score axes are presented on a $\log_{10}$ scale. To log the intergenic distance axes, first a constant was added to all intergenic distance values of the absolute value of the most negative intergenic distance measure (for gsB this was 55 bp) plus 1, so that all distances were above 1 for log transformation of the axis.



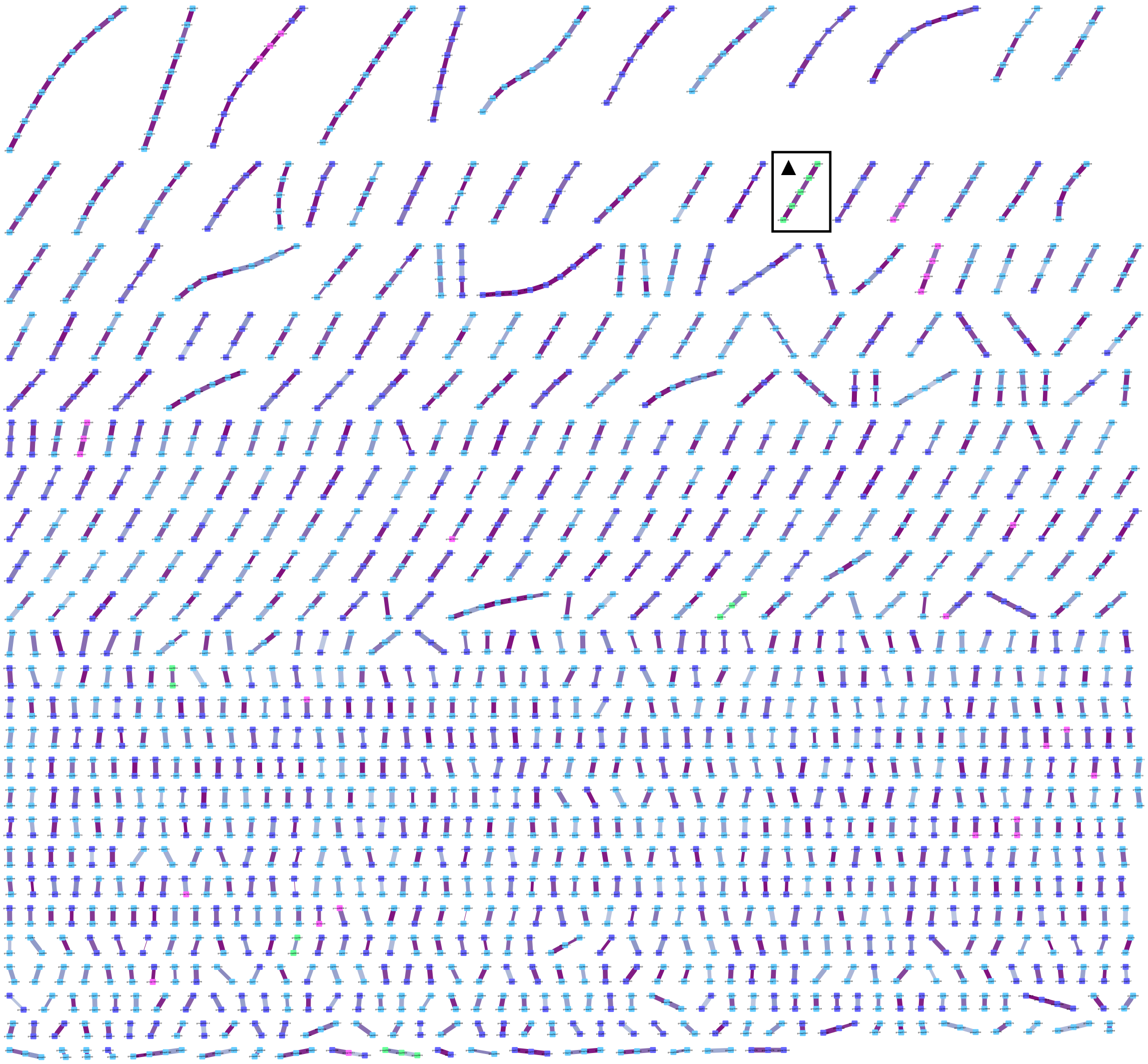
Appendix Figure D.3 Distribution of gene pairs with negative intergenic distances (overlapping gene pairs) for, **a)** genospecies A, **b)** genospecies B, **c)** genospecies C, **d)** genospecies D, **e)** genospecies E.



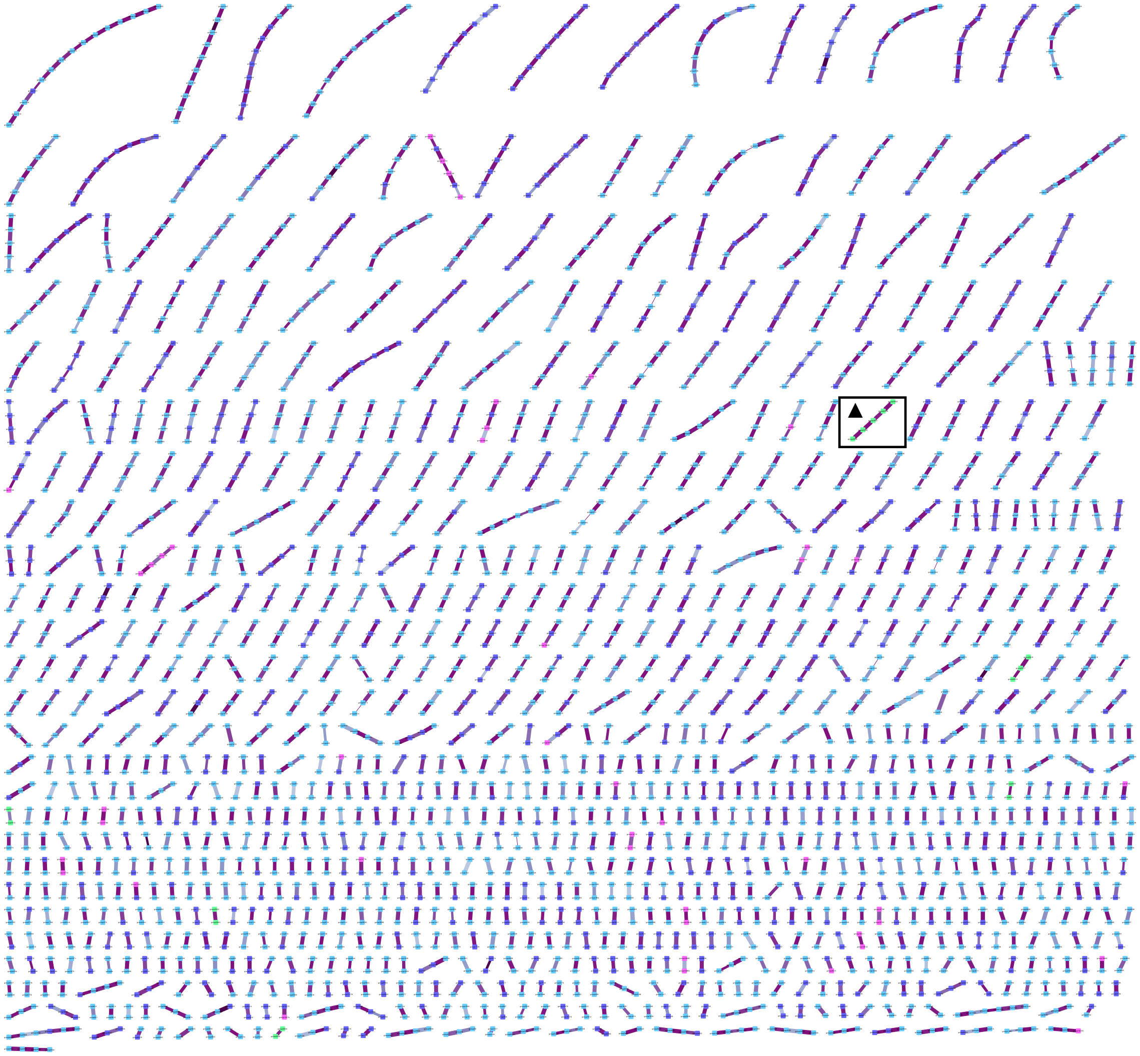
Appendix Figure D.4 Genospecies A transcriptional units generated using the following filtering parameters: R correlation > 0.8, deviance < 3, intergenic distance < 200 bp, must be adjacent gene pair in at least 3 strains. Nodes are genes colour coded by: Blue = core, Purple = accessory, Pink = genospecies enriched, Green = symbiosis. Edge colour increases from blue to purple with increased gene expression correlation between adjacent pairs. Edge thickness increases with decreasing deviance score. ▲ Indicates the *nodABCIJ* operon.



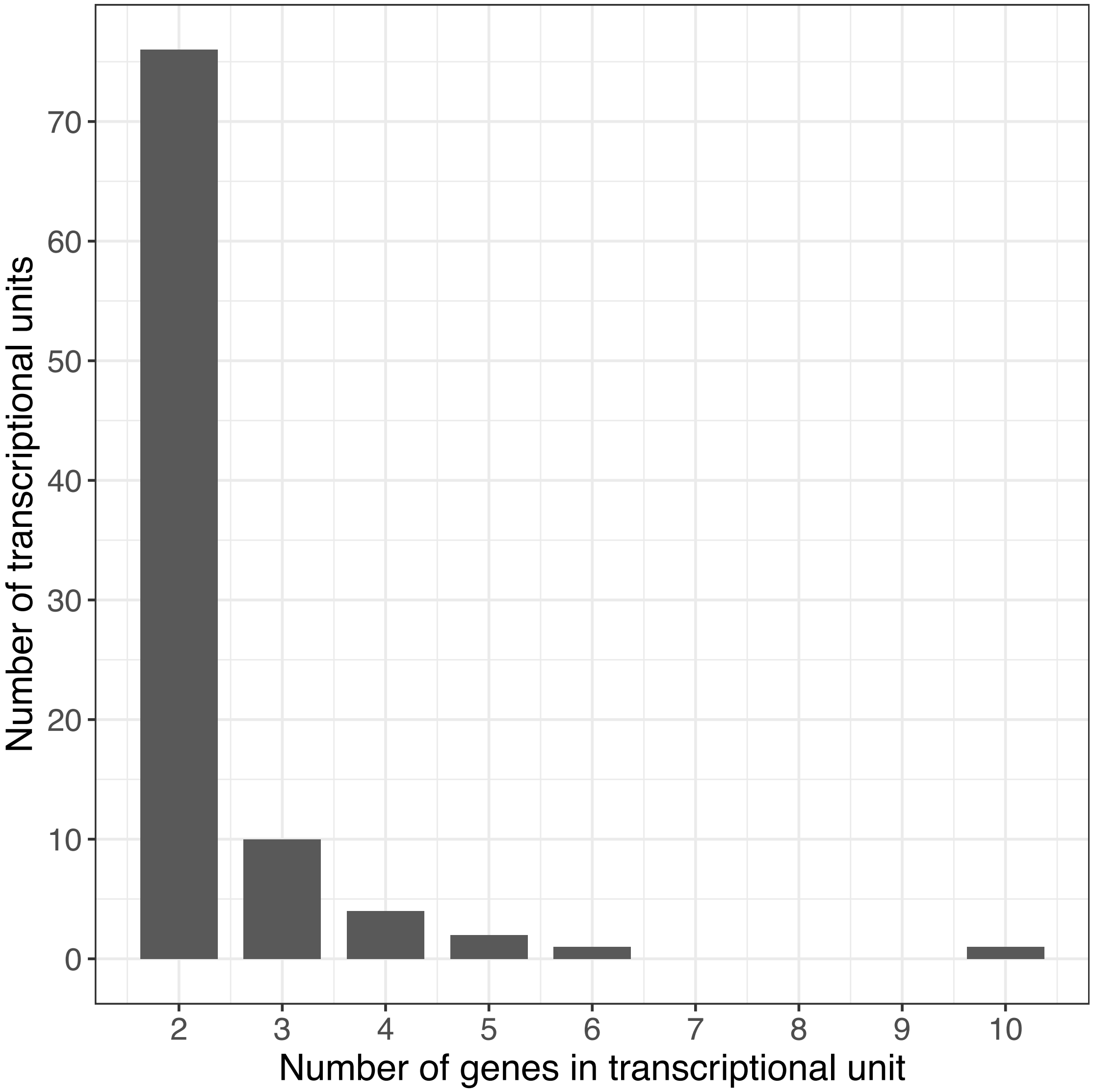
Appendix Figure D.5 Genospecies C transcriptional units generated using the following filtering parameters: R correlation > 0.8, deviance < 3, intergenic distance < 200 bp, must be adjacent gene pair in at least 3 strains. Nodes are genes colour coded by: Blue = core, Purple = accessory, Pink = genospecies enriched, Green = symbiosis. Edge colour increases from blue to purple with increased gene expression correlation between adjacent pairs. Edge thickness increases with decreasing deviance score. *identifies the rhizosphere induced operon (*rhiABC*). ▲ Indicates the *nodABCIJ* operon.



Appendix Figure D.6 Genospecies D transcriptional units generated using the following filtering parameters: R correlation > 0.8, deviance < 3, intergenic distance < 200 bp, must be adjacent gene pair in at least 3 strains. Nodes are genes colour coded by: Blue = core, Purple = accessory, Pink = genospecies enriched, Green = symbiosis. Edge colour increases from blue to purple with increased gene expression correlation between adjacent pairs. Edge thickness increases with decreasing deviance score. ▲ Indicates the *nodABCIJ* operon.

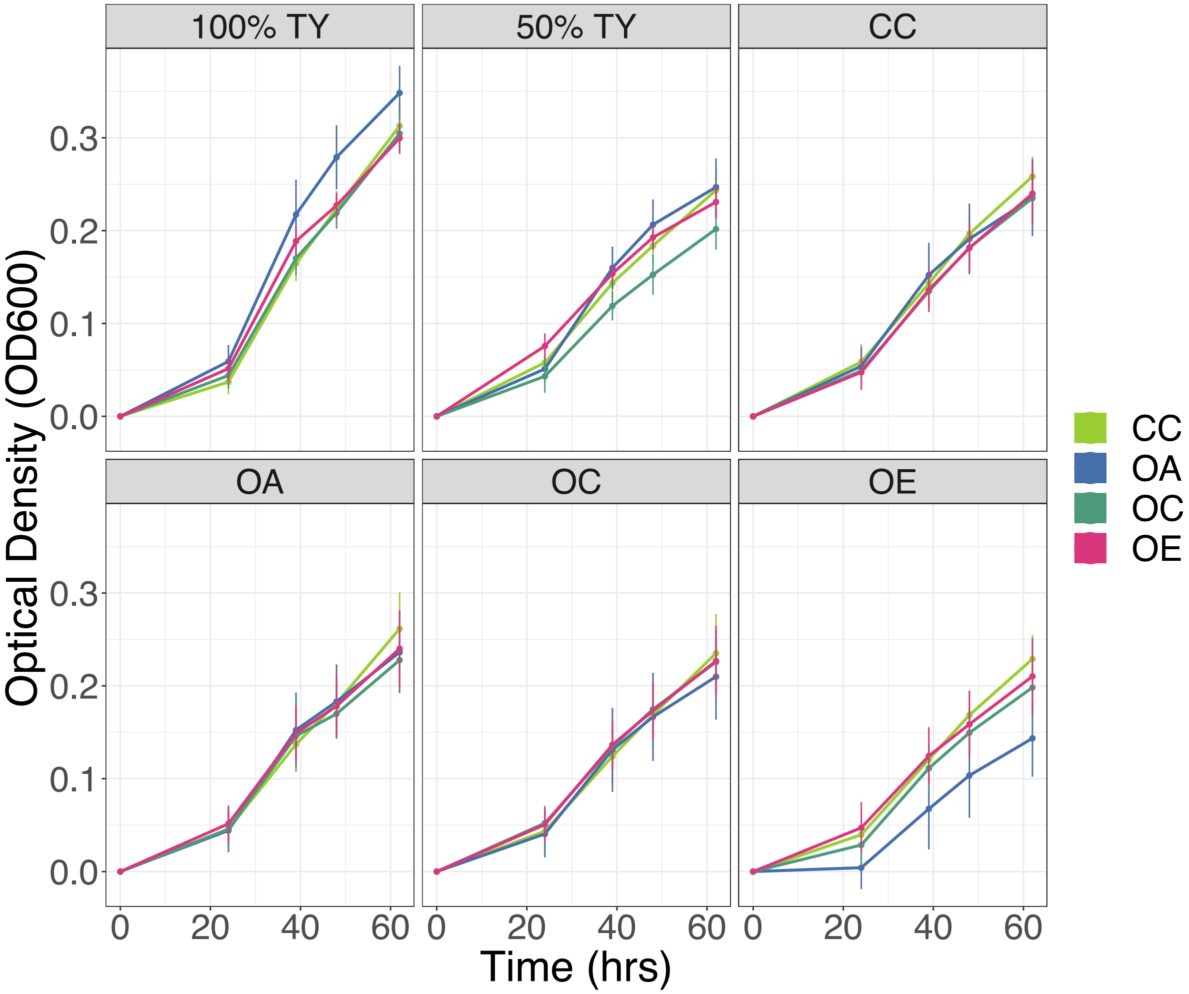


Appendix Figure D.7 Genospecies E transcriptional units generated using the following filtering parameters: R correlation > 0.8, deviance < 3, intergenic distance < 200 bp, must be adjacent gene pair in at least 3 strains. Nodes are genes colour coded by: Blue = core, Purple = accessory, Pink = genospecies enriched, Green = symbiosis. Edge colour increases from blue to purple with increased gene expression correlation between adjacent pairs. Edge thickness increases with decreasing deviance score. ▲ Indicates the *nodABCIJ* operon.

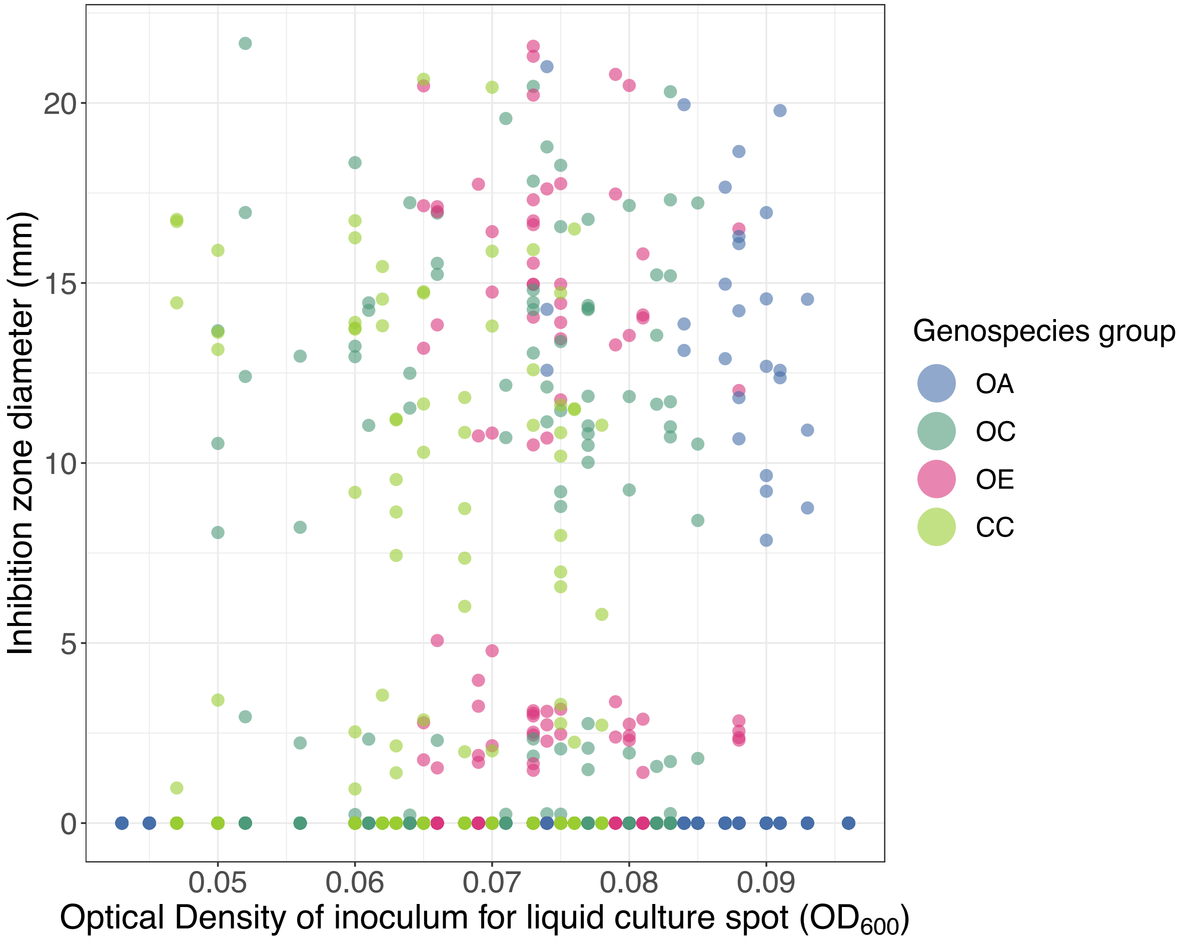


Appendix Figure D.8 Distribution of transcriptional unit size for 94 cross-genospecies conserved transcriptional units. The number of transcriptional units made up of n number of genes across all 5 genospecies.

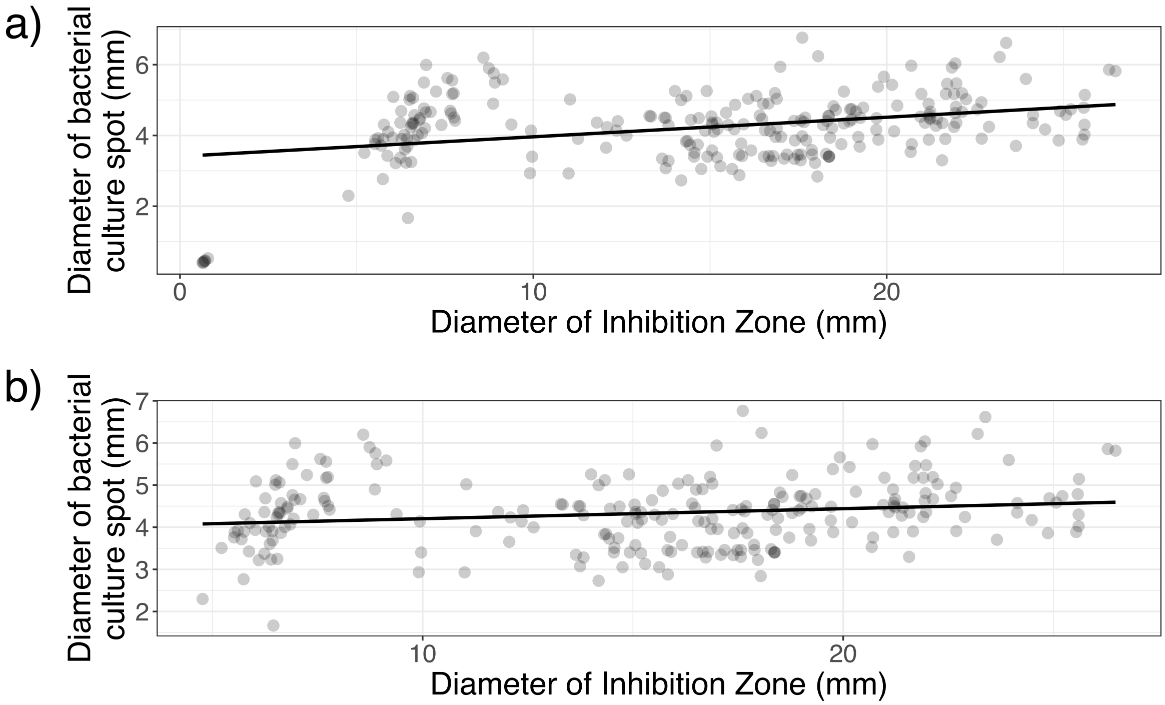
Appendix E. Chapter 6



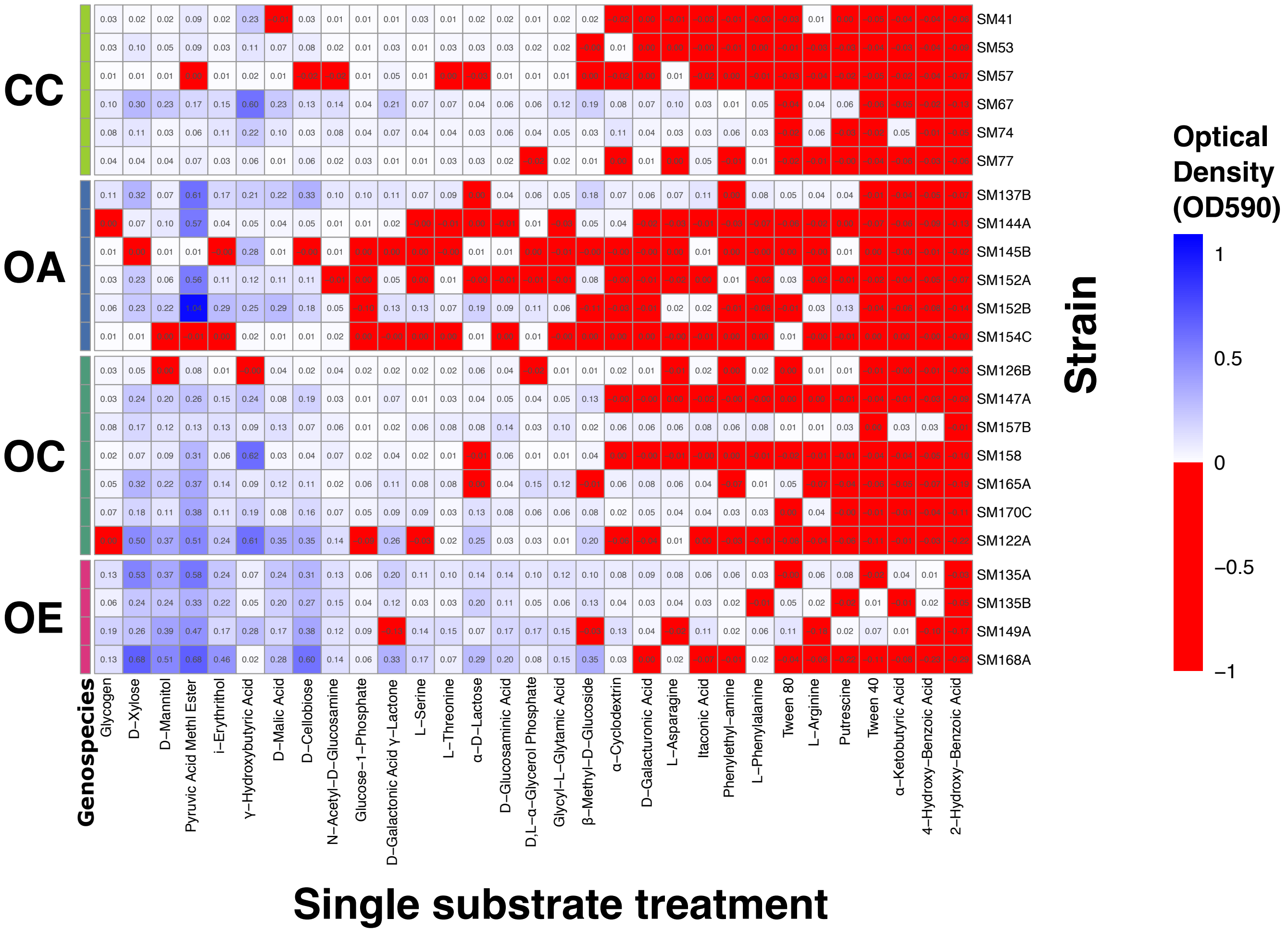
Appendix Figure E.1 Growth curves of strains 24 *Rhizobium leguminosarum* symbiovar *trifolii* strains grown in each other's supernatants in pairwise combinations. Data was grouped by genospecies and farm treatment categories. Strains were additionally grown in 100% and 50% Tryptone Yeast (TY) broth as controls. Optical density (OD_{600}) of strains was measured for 62 hours, and values were normalized by subtracting the 0 h time point optical density. Error bars represent one standard error of the mean. OA = organic genospecies A (n = 6, blue), OC = organic genospecies C (n =7, dark green), OE = organic genospecies E (n = 4, pink), and CC = conventional genospecies C (n = 6, light green).



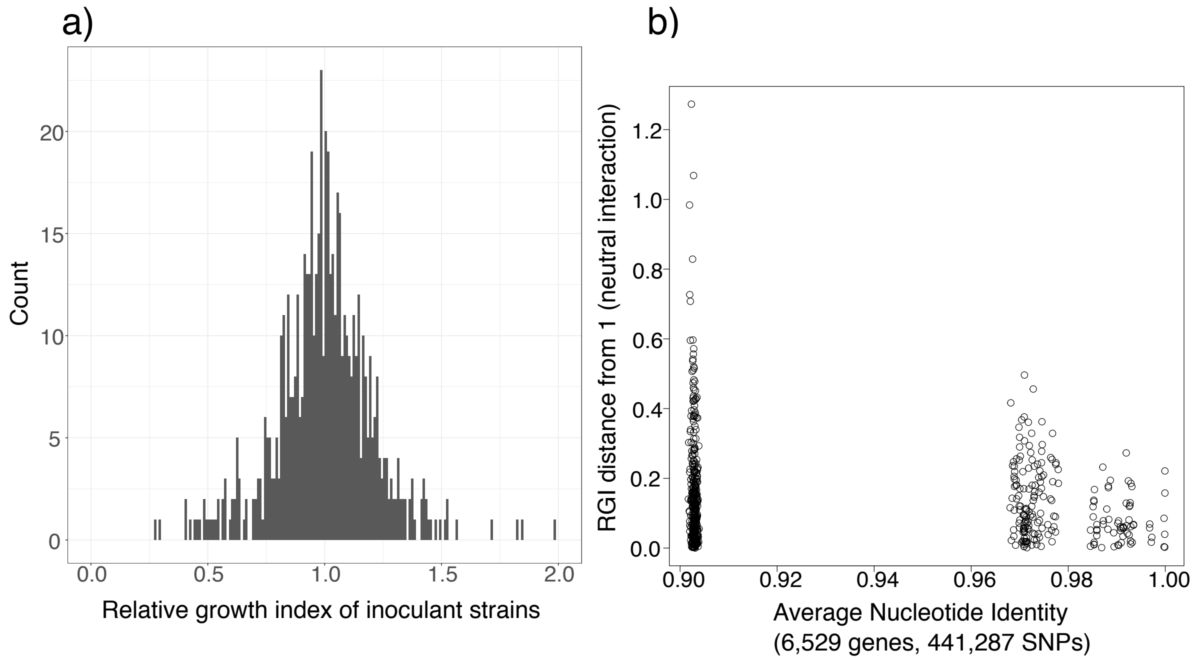
Appendix Figure E.2 Optical density of strain inocula used for liquid culture spotting compared to inhibition zone diameter produced from culture spot on soft lawns of all other strains. Inhibition zone diameter was calculated by deducting the Feret diameter of the culture spot from the Feret diameter of the inhibition zone. Points are coloured by the genospecies grouping of the inoculum strain.



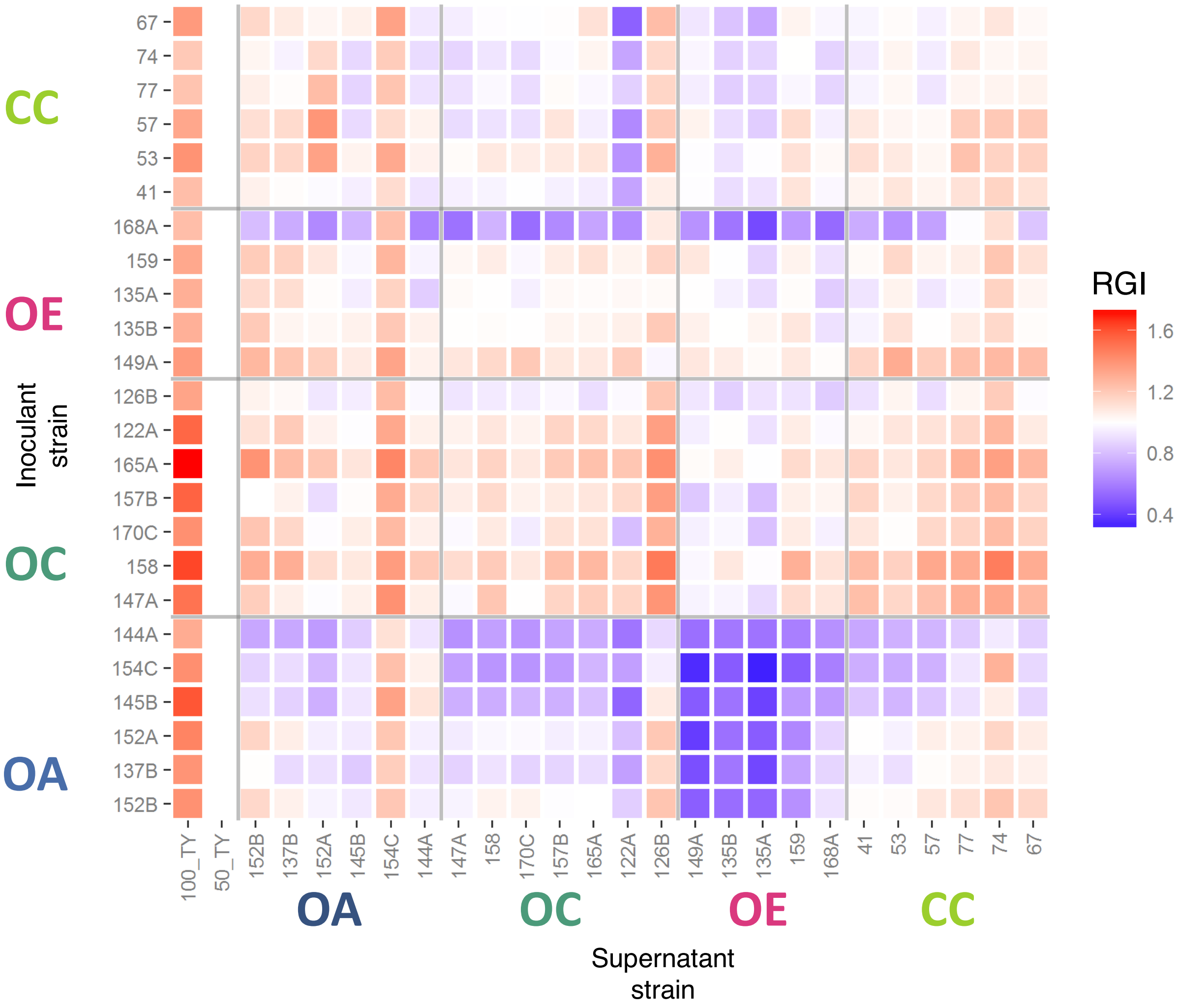
Appendix Figure E.3 a) Correlation between the diameter of the inhibition zone (mm) and the diameter of the liquid culture spot (mm) after 72 hours growth. **b)** Correlation between the diameter of the inhibition zone (mm) and the diameter of the liquid culture spot (mm) after 72 hours growth after removal of liquid culture spots with a diameter of less than 1 mm.



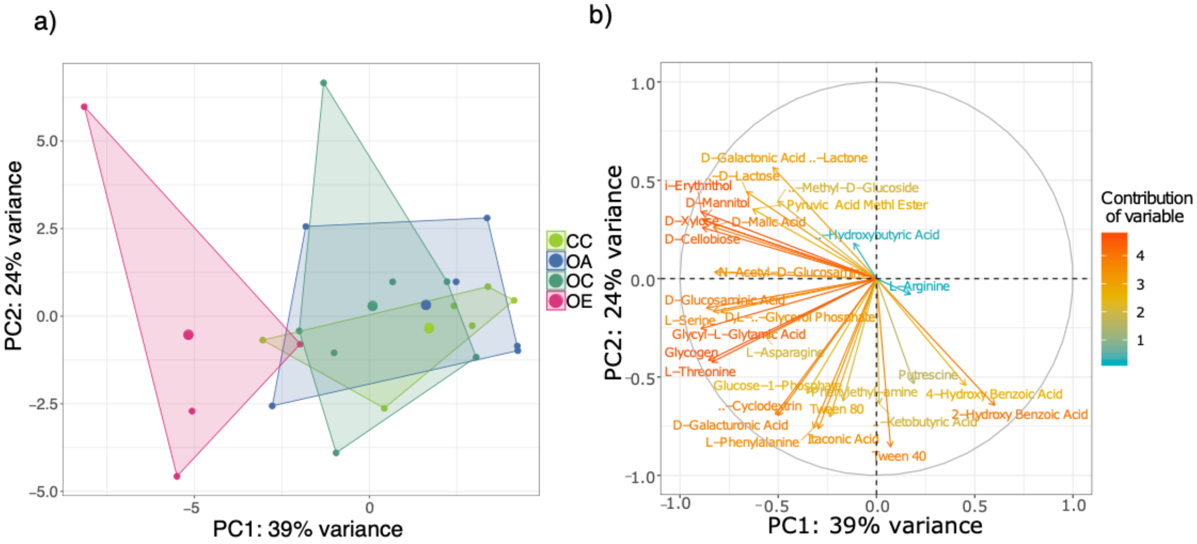
Appendix Figure E.4 Metabolic capacity of 23 *Rhizobium leguminosarum* symbiovar *trifolii* strains on 31 single substrate growth treatments. Ability to metabolise substrates was determined using Biolog Ecoplates and measuring OD_{590} nm of tetrazolium dye in each well. OD_{590} nm values were normalised by subtracting control well OD (water) from the substrate well OD after 72 hours growth. Values of 0.00 OD_{590} nm or less were identified as no observable substrate metabolism (red), values of more than 0.00 OD_{590} nm were considered to have putative capacity to metabolise the substrate (blue).



Appendix Figure E.5 a) Distribution of the Relative Growth Index of all strains grown in each other's supernatants. Relative growth index is calculated as described in the methods. **b)** ANI of interacting strains correlated to relative growth index (RGI) distance from 1 (neutral interaction) of strain grown in other strain's supernatant (Simple linear model with robust standard errors: Coeff$_{ANI}$ = -0.64755, std. error = 0.14621, t = -4.4291, p < .001)



Appendix Figure E.6 Relative growth indices (RGIs) of *Rhizobium leguminosarum* symbiovar *trifolii* strains inoculated into each other's supernatants. RGIs calculation is displayed in the methods by comparing a strains growth in another strain's supernatant relative to its growth in 50% Tryptone Yeast (TY) broth. Supernatant growth treatments constituted of *Rhizobium* strain supernatant and an equal volume of 100% Tryptone Yeast Broth (TY). OA = organic genospecies A, OC = organic genospecies C, OE = organic genospecies E, and CC = conventional genospecies C.



Appendix Figure E.7 Principal Components analysis for metabolic utilization of each individual single substrate Ecoplate treatments by 23 *Rlt* strains. **a)** points each represent a *Rhizobium* strain and is coloured and grouped by genospecies group. PC1 accounted for 38.9% of the variance, and PC2 explained for 23.6% of the variance. **b)** The association of the variables to the first two principle components, which are coloured by percentage contribution of each specific variable to the first two principle components. OA = organic genospecies A, OC = organic genospecies C, OE = organic genospecies E, and CC = conventional genospecies C.

Appendix Table E.1 Linear mixed effects models for supernatant indirect inhibition assay. Model formula: Mean Relative Growth Index ~ inoculant group * supernatant group + (1|inoculant strain) + (1|supernatant strain).

Model	AIC	BIC	Effect	Variable	Variance	Std. Dev	Estimate	Std. Error	df	t value	p-value
Supernatant full model	-858.3	-776.3	random	Inoculant strain (Intercept)	0.010893	0.10437	/	/	/	/	/
			random	Supernatant strain (Intercept)	0.007039	0.08390	/	/	/	/	/
			random	Residual	0.008800	0.09381	/	/	/	/	/
			fixed	Intercept	/	/	0.968512	0.057290	50.202794	16.905	< 2e-16 ***
			fixed	InocOC	/	/	0.070689	0.062247	29.375171	1.136	0.265290
			fixed	InocOE	/	/	0.208752	0.067683	29.260421	3.084	0.004426 **
			fixed	InocCC	/	/	0.074302	0.064567	29.322551	1.151	0.259127
			fixed	SupOC	/	/	-0.111758	0.051786	32.171426	-2.158	0.038488 *
			fixed	SupOE	/	/	-0.382209	0.056283	31.990536	-6.791	1.13e-07 ***
			fixed	SupCC	/	/	-0.006766	0.053705	32.088460	-0.126	0.900528
			fixed	InocOC:SupOC	/	/	0.103648	0.030363	505.043471	3.414	0.000693 ***
			fixed	InocOE:SupOC	/	/	0.039788	0.032367	505.043471	1.229	0.219540
			fixed	InocCC:SupOC	/	/	0.005379	0.030933	505.043471	0.174	0.862029
			fixed	InocOC:SupOE	/	/	0.247622	0.032367	505.043471	7.650	1.02e-13 ***
			fixed	InocOE:SupOE	/	/	0.219678	0.036332	505.043471	6.046	2.88e-09 ***
			fixed	InocCC:SupOE	/	/	0.253220	0.033533	505.043471	7.551	2.03e-13 ***
			fixed	InocOC:SupCC	/	/	0.040557	0.030933	505.043471	1.311	0.190414
			fixed	InocOE:SupCC	/	/	0.013217	0.033533	505.043471	0.394	0.693638
			fixed	InocCC:SupCC	/	/	-0.003805	0.032796	505.043471	-0.116	0.907690

Appendix Table E.2 Estimated marginal means of genospecies supernatant effects on genospecies inoculant growth. Estimates are calculated based on the full model

Filter by sup-genospecies	**contrast**	**estimate**	**SE**	**df**	**lower.CL**	**upper.CL**	**t.ratio**	**p.value**
OA	Sup-OA - Sup-OC	0.11176	0.0555	37.2	-0.03745	0.261	2.014	0.2012
	Sup-OA - Sup-OE	0.38221	0.0603	37	0.21997	0.5444	6.336	<.0001
	Sup-OA - Sup-CC	0.00677	0.0576	37.1	-0.14801	0.1615	0.118	0.9994
	Sup-OC-Sup-OE	0.27045	0.0579	35.9	0.11442	0.4265	4.669	0.0002
	Sup-OC - Sup-CC	-0.10499	0.055	35.9	-0.25325	0.0433	-1.908	0.243
	Sup-OE - Sup-CC	-0.37544	0.0599	35.9	-0.5368	-0.2141	-6.267	<.0001
OC	Sup-OA - Sup-OC	0.00811	0.0547	35	-0.13946	0.1557	0.148	0.9988
	Sup-OA - Sup-OE	0.13459	0.0592	34.2	-0.02539	0.2946	2.272	0.1248
	Sup-OA - Sup-CC	-0.03379	0.0565	34.2	-0.18632	0.1187	-0.598	0.9319
	Sup-OC - Sup-OE	0.12648	0.0576	34.9	-0.02878	0.2817	2.197	0.1439
	Sup-OC - Sup-CC	-0.0419	0.0547	35	-0.18947	0.1057	-0.766	0.8693
	Sup-OE - Sup-CC	-0.16838	0.0592	34.2	-0.32836	-0.0084	-2.842	0.036
OE	Sup-OA - Sup-OC	0.07197	0.0559	38.4	-0.07806	0.222	1.288	0.5759
	Sup-OA - Sup-OE	0.16253	0.0615	40.5	-0.00235	0.3274	2.641	0.0546
	Sup-OA - Sup-CC	-0.00645	0.058	38.4	-0.16215	0.1492	-0.111	0.9995
	Sup-OC - Sup-OE	0.09056	0.0596	40.6	-0.06898	0.2501	1.52	0.435
	Sup-OC - Sup-CC	-0.07842	0.0559	38.4	-0.22845	0.0716	-1.404	0.5049
	Sup-OE - Sup-CC	-0.16898	0.0615	40.5	-0.33386	-0.0041	-2.746	0.0427
CC	Sup-OA - Sup-OC	0.10638	0.055	35.9	-0.04187	0.2546	1.933	0.2327
	Sup-OA - Sup-OE	0.12899	0.0599	35.9	-0.03237	0.2903	2.153	0.1561
	Sup-OA - Sup-CC	0.01057	0.0576	37.1	-0.1442	0.1653	0.184	0.9978
	Sup-OC - Sup-OE	0.02261	0.0579	35.9	-0.13342	0.1786	0.39	0.9795
	Sup-OC - Sup-CC	-0.09581	0.0555	37.2	-0.24502	0.0534	-1.727	0.3247
	Sup-OE - Sup-CC	-0.11842	0.0603	37	-0.28066	0.0438	-1.963	0.2203

Appendix Table E.3 Estimated marginal means of genospecies inoculant growth under different genospecies supernatant treatments. Estimates are calculated based on the full model

Filter by inoc-genospecies	contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
OA	Sup-OA - Sup-OC	0.11176	0.0555	37.2	-0.03745	0.261	2.014	0.2012
	Sup-OA - Sup-OE	0.38221	0.0603	37	0.21997	0.5444	6.336	<.0001
	Sup-OA - Sup-CC	0.00677	0.0576	37.1	-0.14801	0.1615	0.118	0.9994
	Sup-OC - Sup-OE	0.27045	0.0579	35.9	0.11442	0.4265	4.669	0.0002
	Sup-OC - Sup-CC	-0.10499	0.055	35.9	-0.25325	0.0433	-1.908	0.243
	Sup-OE - Sup-CC	-0.37544	0.0599	35.9	-0.5368	-0.2141	-6.267	<.0001
OC	Sup-OA - Sup-OC	0.00811	0.0547	35	-0.13946	0.1557	0.148	0.9988
	Sup-OA - Sup-OE	0.13459	0.0592	34.2	-0.02539	0.2946	2.272	0.1248
	Sup-OA - Sup-CC	-0.03379	0.0565	34.2	-0.18632	0.1187	-0.598	0.9319
	Sup-OC - Sup-OE	0.12648	0.0576	34.9	-0.02878	0.2817	2.197	0.1439
	Sup-OC - Sup-CC	-0.0419	0.0547	35	-0.18947	0.1057	-0.766	0.8693
	Sup-OE - Sup-CC	-0.16838	0.0592	34.2	-0.32836	-0.0084	-2.842	0.036
OE	Sup-OA - Sup-OC	0.07197	0.0559	38.4	-0.07806	0.222	1.288	0.5759
	Sup-OA - Sup-OE	0.16253	0.0615	40.5	-0.00235	0.3274	2.641	0.0546
	Sup-OA - Sup-CC	-0.00645	0.058	38.4	-0.16215	0.1492	-0.111	0.9995
	Sup-OC - Sup-OE	0.09056	0.0596	40.6	-0.06898	0.2501	1.52	0.435
	Sup-OC - Sup-CC	-0.07842	0.0559	38.4	-0.22845	0.0716	-1.404	0.5049
	Sup-OE - Sup-CC	-0.16898	0.0615	40.5	-0.33386	-0.0041	-2.746	0.0427
CC	Sup-OA - Sup-OC	0.10638	0.055	35.9	-0.04187	0.2546	1.933	0.2327
	Sup-OA - Sup-OE	0.12899	0.0599	35.9	-0.03237	0.2903	2.153	0.1561
	Sup-OA - Sup-CC	0.01057	0.0576	37.1	-0.1442	0.1653	0.184	0.9978
	Sup-OC - Sup-OE	0.02261	0.0579	35.9	-0.13342	0.1786	0.39	0.9795
	Sup-OC - Sup-CC	-0.09581	0.0555	37.2	-0.24502	0.0534	-1.727	0.3247
	Sup-OE - Sup-CC	-0.11842	0.0603	37	-0.28066	0.0438	-1.963	0.2203

Appendix Table E.4 Linear mixed effects models for supernatant indirect inhibition assay, with strains SM154C and SM168A removed. Model formula: Mean Relative Growth Index ~ inoculant group * supernatant group + (1|inoculant strain) + (1|supernatant strain).

Model	AIC	BIC	Effect	Variable	Variance	Std. Dev	Estimate	Std. Error	df	t value	p-value
Supernatant full model without strains SM154C and SM168A	-934.8	-856.2	random	Inoculant strain (Intercept)	0.005460	0.07389	/	/	/	/	/
			random	Supernatant strain (Intercept)	0.004472	0.06687	/	/	/	/	/
			random	Residual	0.005380	0.07335	/	/	/	/	/
			fixed	Intercept	/	/	0.95298	0.04749	49.78721	20.067	< 2e-16 ***
			fixed	InocOC	/	/	0.04985	0.04790	29.84159	1.041	0.30640
			fixed	InocOE	/	/	0.12176	0.05473	29.51277	2.225	0.03389 *
			fixed	InocCC	/	/	0.06182	0.04950	29.75175	1.249	0.22147
			fixed	SupOC	/	/	-0.04500	0.04422	31.52945	-1.017	0.31670
			fixed	SupOE	/	/	-0.37603	0.05050	31.12300	-7.446	2.13e-08 ***
			fixed	SupCC	/	/	0.05627	0.04569	31.41836	1.231	0.22728
			fixed	InocOC:SupOC	/	/	0.07326	0.02654	419.05229	2.760	0.00603 **
			fixed	InocOE:SupOC	/	/	0.02403	0.02973	419.05229	0.808	0.41934
			fixed	InocCC:SupOC	/	/	-0.03337	0.02702	419.05229	-1.235	0.21757
			fixed	InocOC:SupOE	/	/	0.27146	0.02973	419.05229	9.131	< 2e-16 ***
			fixed	InocOE:SupOE	/	/	0.32202	0.03543	419.05229	9.089	< 2e-16 ***
			fixed	InocCC:SupOE	/	/	0.27555	0.03068	419.05229	8.980	< 2e-16 ***
			fixed	InocOC:SupCC	/	/	0.01389	0.02702	419.05229	0.514	0.60742
			fixed	InocOE:SupCC	/	/	-0.04022	0.03068	419.05229	-1.311	0.19061
			fixed	InocCC:SupCC	/	/	-0.03882	0.02841	419.05229	-1.367	0.17246

Appendix Table E.5 Linear mixed effects models for spot plating direct inhibition assay. Model formula: Mean inhibition zone diameter ~ inoculant group * supernatant group + (1|inoculant strain) + (1|supernatant strain).

Model	AIC	BIC	Effect	Variable	Variance	Std. Dev	Estimate	Std. Error	df	t value	p-value
Spot plating full model	2274.8	2357.6	random	Spot strain (Intercept)	0.1207	0.3474	/	/	/	/	/
			random	Soft strain (Intercept)	8.2418	2.8709	/	/	/	/	/
			random	Residual	2.2852	1.5117	/	/	/	/	/
			fixed	Intercept	/	/	3.4994	1.2072	26.3757	2.899	0.00745 **
			fixed	SpotOC	/	/	3.2443	0.3940	115.5787	8.234	3.12e-13 ***
			fixed	SpotOE	/	/	4.2026	0.4288	115.5787	9.800	< 2e-16 ***
			fixed	SpotCC	/	/	2.0357	0.4089	115.5787	4.979	2.26e-06 ***
			fixed	SoftOC	/	/	-3.4994	1.6337	25.6712	-2.142	0.04185 *
			fixed	SoftOE	/	/	-3.4994	1.7781	25.6712	-1.968	0.05995
			fixed	SoftCC	/	/	-3.4994	1.6954	25.6712	-2.064	0.04925 *
			fixed	SpotOC:SoftOC	/	/	-3.2443	0.4679	528.9775	-6.934	1.20e-11 ***
			fixed	SpotOE:SoftOC	/	/	-4.2026	0.5093	528.9775	-8.252	1.25e-15 ***
			fixed	SpotCC:SoftOC	/	/	-2.0357	0.4856	528.9775	-4.192	3.24e-05 ***
			fixed	SpotOC:SoftOE	/	/	-2.9500	0.5093	528.9775	-5.793	1.19e-08 ***
			fixed	SpotOE:SoftOE	/	/	-3.9534	0.5543	528.9775	-7.132	3.26e-12 ***
			fixed	SpotCC:SoftOE	/	/	-1.6708	0.5285	528.9775	-3.161	0.00166 **
			fixed	SpotOC:SoftCC	/	/	-3.2443	0.4856	528.9775	-6.681	6.01e-11 ***
			fixed	SpotOE:SoftCC	/	/	-3.9154	0.5285	528.9775	-7.408	5.08e-13 ***
			fixed	SpotCC:SoftCC	/	/	-2.0357	0.5039	528.9775	-4.040	6.14e-05 ***

Appendix Table E.6 Estimated marginal means of genospecies soft agar treatment effect on inhibition zone formation by different genospecies spotted inoculants. Estimates are calculated based on the full model

Filter by soft-genospecies	contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
OA	I-OA - I-OC	-3.2443	0.409	132	-4.307	-2.181	-7.942	<.0001
	I-OA - I-OE	-4.2026	0.445	132	-5.36	-3.046	-9.452	<.0001
	I-OA - I-CC	-2.0357	0.424	132	-3.139	-0.933	-4.802	<.0001
	I-OC - I-OE	-0.9583	0.43	132	-2.077	0.16	-2.229	0.1208
	I-OC - I-CC	1.2086	0.409	132	0.146	2.272	2.958	0.019
	I-OE - I-CC	2.1669	0.445	132	1.01	3.324	4.874	<.0001
OC	I-OA - I-OC	0	0.387	108	-1.01	1.01	0	1
	I-OA - I-OE	0	0.421	108	-1.099	1.099	0	1
	I-OA - I-CC	0	0.402	108	-1.048	1.048	0	1
	I-OC - I-OE	0	0.407	108	-1.063	1.063	0	1
	I-OC - I-CC	0	0.387	108	-1.01	1.01	0	1
	I-OE - I-CC	0	0.421	108	-1.099	1.099	0	1
OE	I-OA - I-OC	-0.2942	0.437	167	-1.428	0.84	-0.673	0.907
	I-OA - I-OE	-0.2492	0.476	167	-1.483	0.985	-0.524	0.9532
	I-OA - I-CC	-0.3649	0.453	167	-1.541	0.812	-0.805	0.852
	I-OC - I-OE	0.045	0.46	167	-1.148	1.238	0.098	0.9997
	I-OC - I-CC	-0.0707	0.437	167	-1.204	1.063	-0.162	0.9985
	I-OE - I-CC	-0.1157	0.476	167	-1.35	1.118	-0.243	0.9949
CC	I-OA - I-OC	0	0.409	132	-1.063	1.063	0	1
	I-OA - I-OE	-0.2872	0.445	132	-1.444	0.87	-0.646	0.9168
	I-OA - I-CC	0	0.424	132	-1.103	1.103	0	1
	I-OC - I-OE	-0.2872	0.43	132	-1.406	0.832	-0.668	0.9089
	I-OC - I-CC	0	0.409	132	-1.063	1.063	0	1
	I-OE - I-CC	0.2872	0.445	132	-0.87	1.444	0.646	0.9168

Appendix Table E.7 Estimated marginal means of genospecies spotted inoculant inhibition zone formation on different genospecies soft agar treatments. Estimates are calculated based on the full model

Filter by inoc-genospecies	contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
OA	Soft-OA - Soft-OC	3.499	1.78	30.8	-1.346	8.34	1.961	0.2247
	Soft-OA - Soft-OE	3.499	1.94	30.8	-1.774	8.77	1.802	0.2919
	Soft-OA - Soft-CC	3.499	1.85	30.8	-1.529	8.53	1.89	0.2533
	Soft-OC - Soft-OE	0	1.88	30.8	-5.1	5.1	0	1
	Soft-OC - Soft-CC	0	1.78	30.8	-4.846	4.85	0	1
	Soft-OE - Soft-CC	0	1.94	30.8	-5.274	5.27	0	1
OC	Soft-OA - Soft-OC	6.744	1.78	30.4	1.908	11.58	3.789	0.0036
	Soft-OA - Soft-OE	6.449	1.94	30.4	1.186	11.71	3.329	0.0116
	Soft-OA - Soft-CC	6.744	1.85	30.4	1.725	11.76	3.651	0.0051
	Soft-OC - Soft-OE	-0.294	1.87	30.4	-5.384	4.8	-0.157	0.9986
	Soft-OC - Soft-CC	0	1.78	30.4	-4.836	4.84	0	1
	Soft-OE - Soft-CC	0.294	1.94	30.4	-4.969	5.56	0.152	0.9987
OE	Soft-OA - Soft-OC	7.702	1.79	31.3	2.843	12.56	4.3	0.0009
	Soft-OA - Soft-OE	7.453	1.95	31.3	2.164	12.74	3.823	0.0031
	Soft-OA - Soft-CC	7.415	1.86	31.3	2.372	12.46	3.989	0.002
	Soft-OC - Soft-OE	-0.249	1.89	31.3	-5.363	4.86	-0.132	0.9992
	Soft-OC - Soft-CC	-0.287	1.79	31.3	-5.146	4.57	-0.16	0.9985
	Soft-OE - Soft-CC	-0.038	1.95	31.3	-5.327	5.25	-0.019	1
CC	Soft-OA - Soft-OC	5.535	1.78	30.8	0.69	10.38	3.102	0.0203
	Soft-OA - Soft-OE	5.17	1.94	30.8	-0.104	10.44	2.662	0.0563
	Soft-OA - Soft-CC	5.535	1.85	30.8	0.507	10.56	2.989	0.0266
	Soft-OC - Soft-OE	-0.365	1.88	30.8	-5.465	4.73	-0.194	0.9973
	Soft-OC - Soft-CC	0	1.78	30.8	-4.846	4.85	0	1
	Soft-OE - Soft-CC	0.365	1.94	30.8	-4.909	5.64	0.188	0.9976

Appendix Table E.8 Linear mixed effects models for spot plating direct inhibition assay with OA strains SM144A, SM154C and SM145B removed. Model formula: Mean inhibition zone diameter ~ inoculant group * supernatant group + (1|inoculant strain) + (1|supernatant strain).

Model	AIC	BIC	Effect	Variable	Variance	Std. Dev	Estimate	Std. Error	df	t value	p-value
Spot plating full model without SM144A, SM154C and SM145B	1186.0	1265.5	random	Spot strain (Intercept)	0.0000	0.0000	/	/	/	/	/
			random	Soft strain (Intercept)	3.9822	1.9955	/	/	/	/	/
			random	Residual	0.4959	0.7042	/	/	/	/	/
			fixed	Intercept	/	/	2.23290	1.01318	23.12783	2.204	0.03776 *
			fixed	SpotOC	/	/	0.71154	0.22069	462.00000	3.224	0.00135 **
			fixed	SpotOE	/	/	2.11815	0.23619	462.00000	8.968	< 2e-16 ***
			fixed	SpotCC	/	/	0.31570	0.22728	462.00000	1.389	0.16548
			fixed	SoftOC	/	/	-2.23290	1.27009	23.12784	-1.758	0.09197
			fixed	SoftOE	/	/	-2.23290	1.35932	23.12784	-1.643	0.11398
			fixed	SoftCC	/	/	-2.23290	1.30801	23.12784	-1.707	0.10120
			fixed	SpotOC:SoftOC	/	/	-0.71154	0.27664	462.00000	-2.572	0.01042 *
			fixed	SpotOE:SoftOC	/	/	-2.11815	0.29608	462.00000	-7.154	3.33e-12 ***
			fixed	SpotCC:SoftOC	/	/	-0.31570	0.28490	462.00000	-1.108	0.26840
			fixed	SpotOC:SoftOE	/	/	-0.41731	0.29608	462.00000	-1.409	0.15938
			fixed	SpotOE:SoftOE	/	/	-1.86894	0.31688	462.00000	-5.898	7.12e-09 ***
			fixed	SpotCC:SoftOE	/	/	0.04923	0.30492	462.00000	0.161	0.87180
			fixed	SpotOC:SoftCC	/	/	-0.71154	0.28490	462.00000	-2.497	0.01285 *
			fixed	SpotOE:SoftCC	/	/	-1.83093	0.30492	462.00000	-6.005	3.89e-09 ***
			fixed	SpotCC:SoftCC	/	/	-0.31570	0.29341	462.00000	-1.076	0.28250

Appendix Table E.9 PERMANOVA of strains' metabolic capacity for 6 resource type groups based on average well colour development of 31 single substrate growth treatments. Genospecies groups correspond to OA, OC, OE and CC.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
Genospecies group	3	0.12932	0.043108	3.8293	0.3768	0.006
Residuals	19	0.21389	0.011257	0.6232		
Total	22	0.34321	1			

Appendix Table E.10 The number of secondary metabolite biosynthesis gene clusters identified in 24 *Rhizobium leguminosarum* symbiovar *trifolii* strains. antiSMASH was used to identify the gene clusters, and putative clusters were only counted if gene cluster regions contained at least 2 identifiable metabolite biosynthesis related genes. The strain can be divided into 4 genospecies/environmental origin groups; OA = organic genospecies A, OC = organic genospecies C, OE = organic genospecies E, and CC = conventional genospecies C. T3PKS = Type III polyketide synthases, Hserlactone = Homoserine lactone cluster, Fused = Pheganomycin-style protein ligase-containing cluster, NRPS = Non-ribosomal peptide synthetase cluster.

		Secondary metabolite biosynthesis gene clusters									
Genospecies	Strain	T3PKS	Bacteriocin	Terpene	Arylopolyene	Ectoine	Hserlactone	Proteusin	Fused	NRPS-like	NRPS
CC	SM67	1	0	1	1	1	4	0	0	1	0
	SM77	2	0	1	1	1	4	0	0	1	1
	SM74	2	0	1	1	1	3	0	0	0	1
	SM57	2	0	1	1	1	4	0	0	0	1
	SM53	2	0	1	1	1	4	0	0	0	1
	SM41	1	0	1	1	1	4	0	0	0	1
OA	SM144A	2	1	1	1	1	4	1	0	0	0
	SM154C	2	1	1	1	1	2	1	0	0	0
	SM145B	2	1	1	1	1	4	1	0	0	0
	SM152A	2	1	1	1	1	3	1	0	0	0
	SM137B	1	1	1	1	1	3	1	0	0	0
	SM152B	1	1	2	1	1	2	0	0	0	0
OC	SM126B	2	0	1	1	1	3	0	0	0	1
	SM122A	2	1	1	1	1	4	0	0	0	0
	SM165A	1	2	1	1	1	4	1	0	0	0
	SM157B	1	2	1	1	1	4	1	0	0	0
	SM170C	2	1	1	1	1	4	0	0	0	1
	SM158	2	1	1	1	1	4	0	0	0	0
	SM147A	2	1	1	1	1	3	0	1	1	0
OE	SM168A	1	1	1	1	1	5	1	0	1	1
	SM159	1	1	1	1	1	4	0	0	0	1
	SM135A	1	2	1	1	1	4	1	0	0	1
	SM135B	2	1	1	1	1	4	1	0	0	1
	SM149A	1	1	1	1	1	4	1	0	0	1

Appendix Table E.11 The number of putative prophage regions identified in 24 *Rhizobium leguminosarum* symbiovar *trifolii* strains. Prophage regions were identified using PHASTER, which assigns a completeness score to identified regions based on the proportion of phage-related genes within the region; 'intact', 'questionable' or 'incomplete'. The strain can be divided into 4 genospecies/environmental origin groups; OA = organic genospecies A, OC = organic genospecies C, OE = organic genospecies E, and CC = conventional genospecies C.

		Prophage region		
Genospecies	Strain	Intact	Questionable	Incomplete
CC	SM67	0	1	5
	SM77	1	2	3
	SM74	0	0	2
	SM57	0	2	2
	SM53	0	0	2
	SM41	0	0	5
OA	SM144A	0	1	2
	SM154C	0	0	3
	SM145B	0	1	3
	SM152A	1	0	3
	SM137B	0	0	4
	SM152B	1	0	3
OC	SM126B	0	0	0
	SM122A	0	0	1
	SM165A	1	0	1
	SM157B	0	1	0
	SM170C	0	0	2
	SM158	0	0	2
	SM147A	0	0	1
OE	SM168A	0	0	5
	SM159	0	0	4
	SM135A	1	0	2
	SM135B	0	0	0
	SM149A	0	0	5

References

Abd-Alla, M. H. *et al.* (2014) 'Synergistic interaction of Rhizobium leguminosarum bv. viciae and arbuscular mycorrhizal fungi as a plant growth promoting biofertilizers for faba bean (Vicia faba L.) in alkaline soil', *Microbiol. Res.*, 169(1), pp. 49–58. doi: 10.1016/j.micres.2013.07.007.

Adékambi, T., Drancourt, M. and Raoult, D. (2009) 'The rpoB gene as a tool for clinical microbiologists', *Trends in Microbiology*, pp. 37–45. doi: 10.1016/j.tim.2008.09.008.

Andrews, M. *et al.* (2007) 'Use of white clover as an alternative to nitrogen fertiliser for dairy pastures in nitrate vulnerable zones in the UK: productivity, environmental impact and economic considerations', *Annals of Applied Biology*, 151(1), pp. 11–23. doi: 10.1111/j.1744-7348.2007.00137.x.

Annicchiarico, P. *et al.* (2015) 'Achievements and Challenges in Improving Temperate Perennial Forage Legumes', *Critical Reviews in Plant Sciences*, 34(1–3), pp. 327–380. doi: 10.1080/07352689.2014.898462.

Aouani, M. *et al.* (1997) 'Potential for inoculation of common bean by effective rhizobia in Tunisian soils', *Agronomie*, 17, pp. 445–454.

Argaw, A. and Muleta, D. (2017) 'Effect of genotypes-Rhizobium-environment interaction on nodulation and productivity of common bean (Phaseolus vulgaris L.) in eastern Ethiopia', *Environmental Systems Research*, 6(1), p. 14. doi: 10.1186/s40068-017-0091-8.

Arike, L. *et al.* (2012) 'Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli', *Journal of Proteomics*, 75(17), pp. 5437–5448. doi: 10.1016/j.jprot.2012.06.020.

Arndt, D. *et al.* (2016) 'PHASTER: a better, faster version of the PHAST phage search tool', *Nucleic Acids Research*, 44(W1), pp. W16–W21. doi: 10.1093/nar/gkw387.

Ashworth, A. J. *et al.* (2017) 'Microbial community structure is affected by cropping sequences and poultry litter under long-term no-tillage', *Soil Biology and Biochemistry*, 114, pp. 210–219. doi: 10.1016/j.soilbio.2017.07.019.

Asner, G. P. *et al.* (2004) 'Grazing systems, ecosystem responses, and global change', *Annual Review of Environment and Resources*, 29(1), pp. 261–299. doi: 10.1146/annurev.energy.29.062403.102142.

Barrett, L. G. *et al.* (2015) 'Partner diversity and identity impacts on plant productivity in Acacia-rhizobial interactions', *Journal of Ecology*, 103(1), pp. 130–142. doi: 10.1111/1365-2745.12336.

Batista, L. *et al.* (2015) 'Nodulation competitiveness as a requisite for improved rhizobial inoculants of Trifolium pratense', *Biology and Fertility of Soils*, 51(1), pp. 11–20. doi: 10.1007/s00374-014-0946-3.

Becker, J. *et al.* (2012) 'Increasing antagonistic interactions cause bacterial communities to collapse at high diversity', *Ecol. Lett.*, 15(5), pp. 468–474. doi: 10.1111/j.1461-0248.2012.01759.x.

Berg, R. K. *et al.* (1988) 'Nodule Occupancy by Introduced Bradyrhizobium japonicum in Iowa Soils', *Agron. J.*, 80, pp. 876–881. doi: 10.2134/agronj1988.00021962008000060007x.

Bergman, N. H. *et al.* (2007) 'Operon prediction for sequenced bacterial genomes without experimental information', *Appl. Environ. Microbiol.*, 73(3), pp. 846–854. doi: 10.1128/AEM.01686-06.

Birtel, J. *et al.* (2015) 'Estimating bacterial diversity for ecological studies: Methods, metrics, and assumptions', *PLoS ONE*, 10(4). doi: 10.1371/journal.pone.0125356.

Black, M. *et al.* (2012) 'The genetics of symbiotic nitrogen fixation: Comparative genomics of 14 rhizobia strains by resolution of protein clusters', *Genes*, 3(1), pp. 138–166. doi: 10.3390/genes3010138.

Bladergroen, M. R., Badelt, K. and Spaink, H. P. (2003) 'Infection-blocking genes of a symbiotic Rhizobium leguminosarum strain that are involved in temperature-dependent protein secretion', *Mol. Plant. Microbe. Interact.*, 16(1), pp. 53–64. doi: 10.1094/MPMI.2003.16.1.53.

Blanco, A. R., Sicardi, M. and Frioni, L. (2010) 'Competition for nodule occupancy between introduced and native strains of Rhizobium leguminosarum biovar trifolii', *Biol. Fertil. Soils*, 46(4), pp. 419–425. doi: 10.1007/s00374-010-0439-y.

Blin, K. *et al.* (2019) 'antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline', *Nucleic Acids Research*, 47(W1), pp. W81–W87. doi: 10.1093/nar/gkz310.

Bohlool, B. B. *et al.* (1992) 'Biological nitrogen fixation for sustainable agriculture: A perspective', *Plant and Soil*, pp. 1–11. doi: 10.1007/BF00011307.

Boivin, S. *et al.* (2020) 'Host-specific competitiveness to form nodules in Rhizobium leguminosarum symbiovar viciae', *New Phytologist*, 226(2), pp. 555–568. doi: 10.1111/nph.16392.

Bokulich, N. A. *et al.* (2018) 'Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin', *Microbiome*, 6(1), p. 90. doi: 10.1186/s40168-018-0470-z.

Bolker, B. (2020) *GLMM FAQ*. Available at: http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html (Accessed: 11 March 2020).

Bolker, B. M. *et al.* (2009) 'Generalized linear mixed models: a practical guide for ecology and evolution', *Trends in Ecology and Evolution*, pp. 127–135. doi: 10.1016/j.tree.2008.10.008.

Bolnick, D. I. *et al.* (2011) 'Why intraspecific trait variation matters in community ecology', *Trends in Ecology and Evolution*, pp. 183–192. doi: 10.1016/j.tree.2011.01.009.

Boncompagni, E. *et al.* (1999) 'Occurrence of choline and glycine betaine uptake and metabolism in the family Rhizobiaceae and their roles in osmoprotection', *Applied and Environmental Microbiology*, 65(5), pp. 2072–2077.

Borland, S., Prigent-Combaret, C. and Wisniewski-Dyé, F. (2016) 'Bacterial hybrid histidine kinases in plant-bacteria interactions', *Microbiology (United Kingdom)*, pp. 1715–1734. doi: 10.1099/mic.0.000370.

Bosworth, A. H., Breil, B. T. and Triplett, E. W. (1993) 'Production of the anti-rhizobial peptide, trifolitoxin, in sterile soils by Rhizobium leguminosarum bv. trifolii T24', *Soil Biology and Biochemistry*, 25(7), pp. 829–832. doi: 10.1016/0038-0717(93)90083-N.

Bourion, V. *et al.* (2018) 'Co-inoculation of a Pea Core-Collection with Diverse Rhizobial Strains Shows Competitiveness for Nodulation and Efficiency of Nitrogen Fixation Are Distinct traits in the Interaction', *Frontiers in Plant Science*, 8, p. 2249. doi: 10.3389/fpls.2017.02249.

Brewin, N., Wood, E. and Young, J. P. W. (1983) 'Contribution of the Symbiotic Plasmid to the Competitiveness of Rhizobium leguminosavum', *Journal of General Microbiology*, 12, pp. 2973–2977.

Brhada, F. *et al.* (2001) 'Osmoprotection mechanisms in rhizobia isolated from Vicia faba var. major and Cicer arietinum', *Agronomie, EDP Sciences*, 21(7), pp. 583–590. doi: 10.1051/agro:2001148ï.

Brinza, L. *et al.* (2010) 'Structure and dynamics of the operon map of Buchnera aphidicola sp. strain APS', *BMC Genomics*, 11, p. 666. doi: 10.1186/1471-2164-11-666.

Brockhurst, M. A. *et al.* (2019) 'The Ecology and Evolution of Pangenomes', *Current Biology*, pp. R1094–R1103. doi: 10.1016/j.cub.2019.08.012.

Brockwell, J. and Bottomley, P. J. (1995) 'Recent advances in inoculant technology and prospects for the future', *Soil Biology and Biochemistry*, 27(4–5), pp. 683–697. doi: 10.1016/0038-0717(95)98649-9.

Brockwell, J., McIlroy, R. A. and Hebb, D. M. (1998) *The Australian Collection of Rhizobium Strains for Temperate Legumes. Catalogue 1998*. Available at: https://publications.csiro.au/publications/#publication/PIprocite:310e178d-726c-4bba-a5c2-116ddf88fc8c (Accessed: 6 May 2020).

ten Broeke-Smits, N. J. P. *et al.* (2010) 'Operon structure of Staphylococcus aureus', *Nucleic Acids Res.*, 38(10), pp. 3263–3274. doi: 10.1093/nar/gkq058.

Bromfield, E. S. P. (1984) 'Variation in Preference for Rhizobium meliloti Within and Between Medicago sativa Cultivars Grown in Soil', *Applied and Environmental Microbiology*, 48(6), pp. 1231–1236. doi: 10.1128/aem.48.6.1231-1236.1984.

Bromfield, E. S. P., Barran, L. R. and Wheatcroft, R. (1995) 'Relative genetic structure of a population of Rhizobium meliloti isolated directly from soil and from nodules of alfalfa (Medicago sativa) and sweet clover (Melilotus alba)', *Molecular Ecology*, 4(2), pp. 183–188. doi: 10.1111/j.1365-294X.1995.tb00207.x.

Bruno, J. F., Stachowicz, J. J. and Bertness, M. D. (2003) 'Inclusion of facilitation into ecological theory', *Trends in Ecology and Evolution*, 18(3), pp. 119–125.

van Brussel, A. A. *et al.* (1985) 'Bacteriocin small of fast-growing rhizobia is chloroform soluble and is not required for effective nodulation', *J. Bacteriol.*, 162(3), pp. 1079–1082.

Bulgarelli, D. *et al.* (2012) 'Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota', *Nature*, 488. doi: 10.1038/nature11336.

Bulgarelli, D. *et al.* (2013) 'Structure and functions of the bacterial microbiota of plants', *Annu Rev Plant Biol*, 64. doi: 10.1146/annurev-arplant-050312-120106.

Bulgarelli, D. *et al.* (2015) 'Structure and function of the bacterial root microbiota in wild and domesticated barley', *Cell Host Microbe*, 17. doi: 10.1016/j.chom.2015.01.011.

Burke, C. M. and Darling, A. E. (2016) 'A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq', *PeerJ*, 2016(9), p. e2492. doi: 10.7717/peerj.2492.

Burns, J. H. *et al.* (2015) 'Soil microbial community variation correlates most strongly with plant species identity, followed by soil chemistry, spatial location and plant genus', *The open-access journal for plant sciences*, 5(plv030). doi: 10.1093/aobpla/plv030.

Busby, P. E. *et al.* (2017) 'Research priorities for harnessing plant microbiomes in sustainable agriculture', *PLOS Biology*. Public Library of Science, 15(3), p. e2001793. doi: 10.1371/journal.pbio.2001793.

Buttery, B., Park, S. and van Berkum, P. (1997) 'Effects of common bean (Phaseolus vulgaris L.) cultivar and rhizobium strain on plant growth, seed yield and nitrogen content', *Canadian Journal of Plant Science*, 77:, pp. 347–351.

Callahan, B. J. *et al.* (2016) 'DADA2: High-resolution sample inference from Illumina amplicon data', *Nat. Methods*, 13(7), pp. 581–583. doi: 10.1038/nmeth.3869.

Cameron, K. C., Di, H. J. and Moir, J. L. (2013) 'Nitrogen losses from the soil/plant system: a review', *Annals of Applied Biology*, 162(2), pp. 145–173. doi: 10.1111/aab.12014.

Carlsen, S. C. K. *et al.* (2012) 'Fate in Soil of Flavonoids Released from White Clover (Trifolium repens L.)', *Applied and Environmental Soil Science*, 2012. doi: 10.1155/2012/743413.

Carrasco, M. A., Tan, J. C. and Duman, J. G. (2011) 'A cross-species compendium of proteins/gene products related to cold stress identified by bioinformatic approaches', *Journal of Insect Physiology*, 57(8), pp. 1127–1135. doi: 10.1016/j.jinsphys.2011.04.021.

Case, R. J. *et al.* (2007) 'Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies', *Applied and Environmental Microbiology*, 73(1), pp. 278–288. doi: 10.1128/AEM.01177-06.

Catroux, G., Hartmann, A. and Revellin, C. (2001) *'Trends in rhizobial inoculant production and use'*, *Plant and Soil*.

Cavassim, M. I. A. *et al.* (2019) 'The genomic architecture of introgression among sibling species of bacteria', *bioRxiv*. doi: 10.1101/526707.

Cavassim, M. I. A. *et al.* (2020) 'Symbiosis genes show a unique pattern of introgression and selection within a Rhizobium leguminosarum species complex', *Microbial Genomics*, 6. doi: 10.1099/mgen.0.000351.

Cha, C. *et al.* (1998) *Production of Acyl-Homoserine Lactone Quorum-Sensing Signals by Gram-Negative Plant-Associated Bacteria*, */ 1119 MPMI*.

Chan, J. Z.-M. *et al.* (2012) 'Defining bacterial species in the genomic era: insights from the genus Acinetobacter', *BMC Microbiology*, 12(1), p. 302. doi: 10.1186/1471-2180-12-302.

Chaudhuri, R. *et al.* (2015) 'Cross-species gene expression analysis identifies a novel set of genes implicated in human insulin sensitivity', *npj Systems Biology and Applications*, 1. doi: 10.1038/npjsba.2015.10.

Checcucci, A. *et al.* (2017) 'Trade, Diplomacy, and Warfare: The Quest for Elite Rhizobia Inoculant Strains', *Front. Microbiol.*, 8, p. 2207. doi: 10.3389/fmicb.2017.02207.

Chen, X. *et al.* (2017) 'SeqTU: A Web Server for Identification of Bacterial Transcription Units', *Scientific Reports*, 7. doi: 10.1038/srep43925.

Chetal, K. and Janga, S. C. (2015) 'OperomeDB: A Database of Condition-Specific Transcription Units in Prokaryotic Genomes', *Biomed Res. Int.*, 2015, p. 318217. doi: 10.1155/2015/318217.

Chuang, L.-Y. *et al.* (2012) 'Features for computational operon prediction in prokaryotes', *Brief. Funct. Genomics*, 11(4), pp. 291–299. doi: 10.1093/bfgp/els024.

Clúa, J. *et al.* (2018) 'Compatibility between Legumes and Rhizobia for the Establishment of a Successful Nitrogen-Fixing Symbiosis', *Genes*, 9(125), p. 125. doi: 10.3390/genes9030125.

Connolly, J. P. R. *et al.* (2019) 'Distinct intraspecies virulence mechanisms regulated by a conserved transcription factor', *Proceedings of the National Academy of Sciences of the United States of America*, 116(39), pp. 19695–19704. doi: 10.1073/pnas.1903461116.

Conway, T. *et al.* (2014) 'Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing', *MBio*, 5(4), pp. e01442-14. doi: 10.1128/mBio.01442-14.

Crossman, L. C. *et al.* (2008) 'Correction: A Common Genomic Framework for a Diverse Assembly of Plasmids in the Symbiotic Nitrogen Fixing Bacteria', *PLoS One*, 3(8), pp. 10.1371/annotation/4a58a9bd-7531-41d0-b101-53039a8. doi: 10.1371/annotation/4a58a9bd-7531-41d0-b101-53039a88c25b.

Cubo, M. T. *et al.* (1992) 'Molecular characterization and regulation of the rhizosphere-expressed genes rhiABCR that can influence nodulation by Rhizobium leguminosarum biovar viciae', *Journal of Bacteriology*, 174(12), pp. 4026–4035. doi: 10.1128/jb.174.12.4026-4035.1992.

Dam, P. *et al.* (2007) 'Operon prediction using both genome-specific and general genomic information', *Nucleic Acids Res.*, 35(1), pp. 288–298. doi: 10.1093/nar/gkl1018.

Danino, V. E. *et al.* (2003) 'Recipient-induced transfer of the symbiotic plasmid pRL1JI in Rhizobium leguminosarum bv. viciae is regulated by a quorum-sensing relay', *Molecular Microbiology*, 50(2), pp. 511–525. doi: 10.1046/j.1365-2958.2003.03699.x.

Deakin, W. J. and Broughton, W. J. (2009) 'Symbiotic use of pathogenic strategies: rhizobial protein secretion systems', *Nat. Rev. Microbiol.*, 7(4), pp. 312–320. doi: 10.1038/nrmicro2091.

Debellé, F. *et al.* (1996) 'The NodA proteins of Rhizobium meliloti and Rhizobium tropici specify the N - acylation of Nod factors by different fatty acids', *Molecular Microbiology*, 22(2), pp. 303–314. doi: 10.1046/j.1365-2958.1996.00069.x.

Dehal, P. S. *et al.* (2009) 'MicrobesOnline: An integrated portal for comparative and functional genomics', *Nucleic Acids Research*, 38(SUPPL.1). doi: 10.1093/nar/gkp919.

Deiner, K. *et al.* (2017) 'Environmental DNA metabarcoding: Transforming how we survey animal and plant communities', *Molecular Ecology*, pp. 5872–5895. doi: 10.1111/mec.14350.

Delestre, C. *et al.* (2015) 'Genome sequence of the clover symbiont Rhizobium leguminosarum bv. trifolii strain CC275e', *Stand. Genomic Sci.*, 10. doi: 10.1186/s40793-015-0110-1.

Demanèche, S. *et al.* (2008) 'Antibiotic-resistant soil bacteria in transgenic plant fields', *Proceedings of the National Academy of Sciences of the United States of America*, 105(10), pp. 3957–3962. doi: 10.1073/pnas.0800072105.

Dénarié, J., Debelle, F. and Rosenberg, C. (1992) 'Signaling and Host Range Variation in Nodulation', *Annual*

Review of Microbiology, 46(1), pp. 497–531. doi: 10.1146/annurev.mi.46.100192.002433.

Denison, R. F. (2000) 'Legume Sanctions and the Evolution of Symbiotic Cooperation by Rhizobia', *Am. Nat.*, 156(6), pp. 567–576. doi: 10.1086/316994.

Denison, R. F. and Kiers, E. T. (2004) 'Lifestyle alternatives for rhizobia: mutualism, parasitism, and forgoing symbiosis', *FEMS Microbiol. Lett.*, 237(2), pp. 187–193. doi: 10.1111/j.1574-6968.2004.tb09695.x.

Denton, M. D. *et al.* (2003) 'Competitive abilities of common field isolates and a commercial strain of Rhizobium leguminosarum bv. trifolii for clover nodule occupancy', *Soil Biology and Biochemistry*, 35(8), pp. 1039–1048. doi: 10.1016/S0038-0717(03)00146-9.

diCenzo, G. C. *et al.* (2014) 'Examination of Prokaryotic Multipartite Genome Evolution through Experimental Genome Reduction', *PLoS Genetics*, 10(10). doi: 10.1371/journal.pgen.1004742.

diCenzo, G. C. *et al.* (2019) 'Multidisciplinary approaches for studying rhizobium–legume symbioses', *Canadian Journal of Microbiology*, 65(1), pp. 1–33. doi: 10.1139/cjm-2018-0377.

diCenzo, G. C. and Finan, T. M. (2017) 'The Divided Bacterial Genome: Structure, Function, and Evolution', *Microbiology and Molecular Biology Reviews*, 81(3). doi: 10.1128/mmbr.00019-17.

Djordjevic, M. A. *et al.* (1987) 'Clovers secrete specific phenolic-compounds which either stimulate or repress nod gene-expression in Rhizobium-trifolii', *EMBO J.*, 6(5), pp. 1173–1179.

Downie, J. A. *et al.* (1985) 'Identification of genes and gene products involved in the nodulation of peas by Rhizobium leguminosarum', *MGG Molecular & General Genetics*, 198(2), pp. 255–262. doi: 10.1007/BF00383003.

Downie, J. A. (2010) 'The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots', *FEMS Microbiology Reviews*, pp. 150–170. doi: 10.1111/j.1574-6976.2009.00205.x.

Downie, J. A. (2014) 'Legume nodulation', *Current Biology*, 24(5), pp. R184–R190. doi: 10.1016/j.cub.2014.01.028.

Dunn, M. F. (2015) 'Key roles of microsymbiont amino acid metabolism in rhizobia-legume interactions', *Critical Reviews in Microbiology*, 41(4), pp. 411–451. doi: 10.3109/1040841X.2013.856854.

Dunning, L. T. *et al.* (2016) 'Ecological speciation in sympatric palms: 1. Gene expression, selection and pleiotropy', *Journal of evolutionary biology*, 29(8), pp. 1472–1487. doi: 10.1111/jeb.12895.

Duodu, S. *et al.* (2006) 'Genetic diversity of a natural population of Rhizobium leguminosarum biovar trifolii analysed from field nodules and by a plant infection technique', *Soil Biol. Biochem.*, 38(5), pp. 1162–1165. doi: 10.1016/j.soilbio.2005.07.015.

Edgar, R. C. *et al.* (2011) 'UCHIME improves sensitivity and speed of chimera detection', *Bioinformatics*, 27(16), pp. 2194–2200. doi: 10.1093/bioinformatics/btr381.

Edgar, R. C. (2013) 'UPARSE: highly accurate OTU sequences from microbial amplicon reads', *Nat. Methods*, 10(10), pp. 996–998. doi: 10.1038/nmeth.2604.

Edgar, R. C. (2016a) 'UCHIME2: improved chimera prediction for amplicon sequencing', *bioRxiv*, p. 074252. doi: 10.1101/074252.

Edgar, R. C. (2016b) 'UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing', *bioRxiv*, p. 081257. doi: 10.1101/081257.

Edgar, R. C. (2017) 'UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success', *bioRxiv*, p. 124149. doi: 10.1101/124149.

Edwards, A. *et al.* (2009) 'The cin and rai quorum-sensing regulatory systems in Rhizobium leguminosarum are coordinated by ExpR and CinS, a small regulatory protein coexpressed with CinI', *Journal of Bacteriology*, 191(9), pp. 3059–3067. doi: 10.1128/JB.01650-08.

Edwards, M. T. *et al.* (2005) 'A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context', *Nucleic Acids Res.*, 33(10), pp. 3253–3262. doi: 10.1093/nar/gki634.

Elbrecht, V. and Leese, F. (2015) 'Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass--Sequence Relationships with an Innovative Metabarcoding Protocol', *PLoS One*, 10(7), p. e0130324. doi: 10.1371/journal.pone.0130324.

Erisman, J. W. *et al.* (2008) 'How a century of ammonia synthesis changed the world', *Nature Geoscience*, 1(10), pp. 636–639. doi: 10.1038/ngeo325.

Ermolaeva, M. D., White, O. and Salzberg, S. L. (2001) 'Prediction of operons in microbial genomes', *Nucleic Acids Res.*, 29(5), pp. 1216–1221.

Eyre-Walker, A. (1995) 'The distance between Escherichia coli genes is related to gene expression levels', *J. Bacteriol.*, 177(18), pp. 5368–5369.

Fageria, N. K., Baligar, V. C. and Bailey, B. A. (2005) 'Role of Cover Crops in Improving Soil and Row Crop Productivity', *Communications in Soil Science and Plant Analysis*, 36(19–20), pp. 2733–2757. doi: 10.1080/00103620500303939.

Fagerli, I. L. and Svenning, M. M. (2005) 'Arctic and subarctic soil populations of Rhizobium leguminosarum biovar trifolii nodulating three different clover species: Characterisation by diversity at chromosomal and symbiosis loci', in *Plant and Soil*, pp. 371–381. doi: 10.1007/s11104-005-3103-9.

Faith, J. J. *et al.* (2013) 'The long-term stability of the human gut microbiota', *Science*, 341(6141), p. 1237439. doi: 10.1126/science.1237439.

Fauvart, M. and Michiels, J. (2008) 'Rhizobial secreted proteins as determinants of host specificity in the rhizobium-legume symbiosis', *FEMS Microbiology Letters*, pp. 1–9. doi: 10.1111/j.1574-6968.2008.01254.x.

Feder, M. E. and Mitchell-Olds, T. (2003) 'Evolutionary and ecological functional genomics', *Nature Reviews Genetics*, 4(8), pp. 649–655. doi: 10.1038/nrg1128.

Fields, B. *et al.* (2019) 'MAUI-seq: Multiplexed, high-throughput amplicon diversity profiling using unique molecular identifiers', *bioRxiv*, p. 538587. doi: 10.1101/538587.

Fierer, N., Brewer, T. and Choudoir, M. (2017) 'Lumping versus splitting – is it time for microbial ecologists to abandon OTUs? - Fierer Lab'. Available at: http://fiererlab.org/2017/05/02/lumping-versus-splitting-is-it-time-for-microbial-ecologists-to-abandon-otus/.

Filiatrault, M. J. (2011) 'Progress in prokaryotic transcriptomics', *Curr. Opin. Microbiol.*, 14(5), pp. 579–586. doi: 10.1016/j.mib.2011.07.023.

Fischer, H. M. (1994) 'Genetic regulation of nitrogen fixation in rhizobia.', *Microbiological Reviews*, 58(3), p. 352.

Fitzpatrick, C. R. *et al.* (2018) 'Assembly and ecological function of the root microbiome across angiosperm plant species', *PNAS*, 115(6), pp. E1157–E1165. doi: 10.5061/dryad.5p414.

Fonseca, V. G. (2018) 'Pitfalls in relative abundance estimation using eDNA metabarcoding', *Mol. Ecol. Resour.*, 18(5), pp. 923–926. doi: 10.1111/1755-0998.12902.

Fortino, V. *et al.* (2014) 'Transcriptome dynamics-based operon prediction in prokaryotes', *BMC Bioinformatics*, 15, p. 145. doi: 10.1186/1471-2105-15-145.

Fortino, V., Tagliaferri, R. and Greco, D. (2016) 'CONDOP: an R package for CONdition-Dependent Operon Predictions', *Bioinformatics*, 32(20), pp. 3199–3200. doi: 10.1093/bioinformatics/btw330.

Fraysse, N., Couderc, F. and Poinsot, V. (2003) 'Surface polysaccharide involvement in establishing the rhizobium-legume symbiosis', *European Journal of Biochemistry*, pp. 1365–1380. doi: 10.1046/j.1432-1033.2003.03492.x.

Friedman, J., Alm, E. J. and Shapiro, B. J. (2013) 'Sympatric Speciation: When Is It Possible in Bacteria?', *PLoS ONE*, 8(1). doi: 10.1371/journal.pone.0053539.

Fuks, G. *et al.* (2018) 'Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling', *Microbiome*, 6(1). doi: 10.1186/s40168-017-0396-x.

Galardini, M. *et al.* (2015) 'Evolution of Intra-specific Regulatory Networks in a Multipartite Bacterial Genome', *PLOS Computational Biology*, 11(9), p. e1004478. doi: 10.1371/journal.pcbi.1004478.

Galloway, J. N. *et al.* (2004) 'Nitrogen cycles: Past, present, and future', *Biogeochemistry*, 70(2), pp. 153–226. doi: 10.1007/s10533-004-0370-0.

Garland, J. L. (2006) 'Analysis and interpretation of community-level physiological profiles in microbial ecology', *FEMS Microbiol. Ecol.*, 24(4), pp. 289–300. doi: 10.1111/j.1574-6941.1997.tb00446.x.

Garland, J. L. and Mills, A. L. (1991) 'Classification and Characterization of Heterotrophic Microbial Communities on the Basis of Patterns of Community-Level Sole-Carbon-Source Utilization', *Applied and Environmental Microbiology*, 57(8), pp. 2351–2359.

Gaunt, M. W. *et al.* (2001) 'Phylogenies of atpD and recA support the small subunit rRNA-based classification of rhizobia', *International Journal of Systematic and Evolutionary Microbiology*, 51, pp. 2037–2048.

Geetha, S. J. and Joshi, S. J. (2013) 'Engineering Rhizobial Bioinoculants: A Strategy to Improve Iron Nutrition', *The Scientific World Journal*, 2013. doi: 10.1155/2013/315890.

Geiger, O. *et al.* (1999) 'The regulator gene phoB mediates phosphate stress-controlled synthesis of the membrane lipid diacylglyceryl-N,N,N-trimethylhomoserine in Rhizobium (Sinorhizobium) meliloti', *Molecular Microbiology*, 32(1), pp. 63–73. doi: 10.1046/j.1365-2958.1999.01325.x.

Gerstmeir, R. *et al.* (2004) 'RamB, a Novel Transcriptional Regulator of Genes Involved in Acetate Metabolism of Corynebacterium glutamicum', *Journal of Bacteriology*, 186(9), pp. 2798–2809. doi: 10.1128/JB.186.9.2798-2809.2004.

Ghosh, P. K. and Maiti, T. K. (2016) 'Structure of Extracellular Polysaccharides (EPS) Produced by Rhizobia and their Functions in Legume–Bacteria Symbiosis: — A Review', *Achievements in the Life Sciences*. Elsevier BV, 10(2), pp. 136–143. doi: 10.1016/j.als.2016.11.003.

Ghoul, M. and Mitri, S. (2016) 'The Ecology and Evolution of Microbial Competition', *Trends Microbiol.*, 24(10), pp. 833–845. doi: 10.1016/j.tim.2016.06.011.

Godfray, H. C. J. *et al.* (2010) 'Food security: The challenge of feeding 9 billion people', *Science*, pp. 812–818. doi: 10.1126/science.1185383.

Gohl, D. M. *et al.* (2016) ‘An optimized protocol for high-throughput amplicon-based microbiome profiling’. doi: 10.1038/protex.2016.030.

González-Torres, P. *et al.* (2015) ‘Interactions between closely related bacterial strains are revealed by deep transcriptome sequencing’, *Applied and Environmental Microbiology*, 81(24), pp. 8445–8456. doi: 10.1128/AEM.02690-15.

Gonzalez, J. E. and Marketon, M. M. (2003) ‘Quorum Sensing in Nitrogen-Fixing Rhizobia’, *Microbiology and Molecular Biology Reviews*, 67(4), pp. 574–592. doi: 10.1128/mmbr.67.4.574-592.2003.

González, V. *et al.* (2019) ‘Phylogenomic Rhizobium Species Are Structured by a Continuum of Diversity and Genomic Clusters’, *Frontiers in Microbiology*, 10(APR), p. 910. doi: 10.3389/fmicb.2019.00910.

Goris, J. *et al.* (2007) ‘DNA-DNA hybridization values and their relationship to whole-genome sequence similarities’, *Int J Syst Evol Microbiol.*, 57, pp. 81–91. doi: 10.1099/ijs.0.64483-0.

Göttfert, M. (1993) ‘Regulation and function of rhizobial nodulation genes’, *FEMS Microbiology Letters*, 104(1–2), pp. 39–63. doi: 10.1111/j.1574-6968.1993.tb05863.x.

Graham, P. H. and Vance, C. P. (2000) ‘Nitrogen Fixation in perspective: an overview of research and extension needs’, *Field Crops Res.*, 65, pp. 93–106.

Graham, P. H. and Vance, C. P. (2003) ‘Legumes: Importance and constraints to greater use’, *Plant Physiology*, pp. 872–877. doi: 10.1104/pp.017004.

Gray, K. M. *et al.* (1996) ‘Cell-to-cell signaling in the symbiotic nitrogen-fixing bacterium Rhizobium leguminosarum: autoinduction of a stationary phase and rhizosphere-expressed genes’, *J. Bacteriol.*, 178(2), pp. 372–376.

Green, R. T. *et al.* (2019) ‘Transcriptomic analysis of rhizobium leguminosarum bacteroids in determinate and indeterminate nodules’, *Microbial Genomics*, 5(2). doi: 10.1099/mgen.0.000254.

Griffin, A. S., West, S. A. and Buckling, A. (2004) ‘Cooperation and competition in pathogenic bacteria’, *Nature*, 430(7003), pp. 1024–1027. doi: 10.1038/nature02744.

Güell, M. *et al.* (2009) ‘Transcriptome complexity in a genome-reduced bacterium’, *Science*, 326(5957), pp. 1268–1271. doi: 10.1126/science.1176951.

Haag, A. F. *et al.* (2013) ‘Molecular insights into bacteroid development during Rhizobium-legume symbiosis’, *FEMS Microbiology Reviews*, pp. 364–383. doi: 10.1111/1574-6976.12003.

Haeze, W. D. ’ and Holsters, M. (2002) ‘Nod factor structures, responses, and perception during initiation of nodule development’, *Glycobiology*, 12(6), pp. 79–105.

Harrison, E. and Brockhurst, M. A. (2012) ‘Plasmid-mediated horizontal gene transfer is a coevolutionary process’, *Trends in Microbiology*, pp. 262–267. doi: 10.1016/j.tim.2012.04.003.

Harrison, E. and Brockhurst, M. A. (2017) ‘Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn’t Kill You Makes You Stronger’, *BioEssays*, 39(12), p. 1700112. doi: 10.1002/bies.201700112.

Harrison, P. W. *et al.* (2010) ‘Introducing the bacterial “chromid”: not a chromosome, not a plasmid’, *Trends Microbiol.*, 18(4), pp. 141–148. doi: 10.1016/j.tim.2009.12.010.

Harrison, S. P., Jones, D. G. and Young, J. P. W. (1989) ‘Rhizobium Population Genetics: Genetic Variation Within and Between Populations from Diverse Locations’, *Microbiology*, 135(5), pp. 1061–1069. doi: 10.1099/00221287-135-5-1061.

Harrison, S. P., Young, J. P. W. and Jones, D. G. (1987) ‘Rhizobium population genetics: Effect of clover variety and inoculum dilution on the genetic diversity sampled from natural populations’, *Plant and Soil*, 103, pp. 147–150.

Hartman, K. *et al.* (2017) ‘Deciphering composition and function of the root microbiome of a legume plant’, *Microbiome*, 5(1), p. 2. doi: 10.1186/s40168-016-0220-z.

Hassan, S. and Mathesius, U. (2012) ‘The role of flavonoids in root-rhizosphere signalling: opportunities and challenges for improving plant-microbe interactions’, *Journal of Experimental Botany*, 63(9), pp. 3429–3444. doi: 10.1093/jxb/err430.

He, X. *et al.* (2003) ‘Quorum sensing in Rhizobium sp. strain NGR234 regulates conjugal transfer (tra) gene expression and influences growth rate.’, *Journal of bacteriology*, 185(3), pp. 809–22. doi: 10.1128/jb.185.3.809-822.2003.

Heath, K. D., Burke, P. V. and Stinchcombe, J. R. (2012) ‘Coevolutionary genetic variation in the legume-rhizobium transcriptome’, *Molecular Ecology*, 21(19), pp. 4735–4747. doi: 10.1111/j.1365-294X.2012.05629.x.

Heemstra, J. R., Walsh, C. T. and Sattely, E. S. (2009) ‘Enzymatic Tailoring of Ornithine in the Biosynthesis of the Rhizobium Cyclic Trihydroxamate Siderophore Vicibactin’, *J. Am. Chem. Soc.*, 131, pp. 15317–15329. doi: 10.1021/ja9056008.

Herridge, D. F., Peoples, M. B. and Boddey, R. M. (2008) ‘Global inputs of biological nitrogen fixation in agricultural systems’, *Plant and Soil*, pp. 1–18. doi: 10.1007/s11104-008-9668-3.

Hibbing, M. E. *et al.* (2010) ‘Bacterial competition: Surviving and thriving in the microbial jungle’, *Nature*

Reviews Microbiology, pp. 15–25. doi: 10.1038/nrmicro2259.

Hirsch, A. M., Lum, M. R. and Downie, J. A. (2001) 'What makes the rhizobia-legume symbiosis so special?', *Plant Physiol.*, 127(4), pp. 1484–1492. doi: 10.1104/pp.010866.

Hirsch, P. R. (1979) 'Plasmid-determined Bacteriocin Production by Rhizobium leguminosarum', *Microbiology*, 113(2), pp. 219–228. doi: 10.1099/00221287-113-2-219.

Hong, G.-F., Burn, J. E. and Johnston, A. W. B. (1987a) '"Evidence that DNA involved in the expression of nodulation (nod) genes in Rhaizobium binds to the product of the regulatory gene nodD"', *Nucleic Acids Research*, 15, pp. 9677–9690.

Hong, G.-F., Burn, J. E. and Johnston, A. W. B. (1987b) 'Evidence that DNA involved in the expression of nodulation (nod) genes in Rhaizobium binds to the product of the regulatory gene nodD', *Nucleic Acids Research*, 15, pp. 9677–9690.

De Hoon, M. J. L. *et al.* (2004) 'Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information', *Pac. Symp. Biocomput.*, pp. 276–287.

Hornischer, K. *et al.* (2019) 'BACTOME-a reference database to explore the sequence-and gene expression-variation landscape of Pseudomonas aeruginosa clinical isolates', *Nucleic Acids Research*, 47. doi: 10.1093/nar/gky895.

Hoshino, T. and Inagaki, F. (2017) 'Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA Genes', *PLoS One*, 12(1), p. e0169431. doi: 10.1371/journal.pone.0169431.

Hosseinkhan, N. *et al.* (2015) 'Co-expressional conservation in virulence and stress related genes of three Gammaproteobacterial species: Escherichia coli, Salmonella enterica and Pseudomonas aeruginosa', *Mol. BioSyst*, 11, p. 3137. doi: 10.1039/c5mb00353a.

Hosseinkhan, N., Mousavian, Z. and Masoudi-Nejad, A. (2018) 'Comparison of gene co-expression networks in Pseudomonas aeruginosa and Staphylococcus aureus reveals conservation in some aspects of virulence', *Gene*, 639, pp. 1–10. doi: 10.1016/j.gene.2017.10.005.

Howieson, J. G. and Dilworth, M. J. (2016) *Working with rhizobia*.

Howieson, J., Malden, J. and Yates, R. (2000) 'Techniques for the Selection and Development of Elite Inoculant Strains of Rhizobium leguminosarum in Southern Australia', *Symbiosis*, 28, pp. 33–48.

Hu, Y. *et al.* (2019) 'The XRE Family Transcriptional Regulator SrtR in Streptococcus suis Is Involved in Oxidant Tolerance and Virulence', *Frontiers in Cellular and Infection Microbiology*, 8. doi: 10.3389/fcimb.2018.00452.

Huerta, A. M. *et al.* (1998) 'RegulonDB: a database on transcriptional regulation in Escherichia coli', *Nucleic Acids Res.*, 26(1), pp. 55–59.

Huse, S. M. *et al.* (2010) 'Ironing out the wrinkles in the rare biosphere through improved OTU clustering', *Environ. Microbiol.*, 12(7), pp. 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x.

Huvet, M. and Stumpf, M. P. H. (2014) 'Overlapping genes: A window on gene evolvability', *BMC Genomics*, 15(1). doi: 10.1186/1471-2164-15-721.

Hwang, I. *et al.* (1994) 'TraI, a luxI homologue, is responsible for production of conjugation factor, the Ti plasmid N-acylhomoserine lactone autoinducer', *Proceedings of the National Academy of Sciences of the United States of America*, 91(11), pp. 4639–4643. doi: 10.1073/pnas.91.11.4639.

Igiehon, N. O. and Babalola, O. O. (2018) 'Rhizosphere microbiome modulators: Contributions of nitrogen fixing bacteria towards sustainable agriculture', *International Journal of Environmental Research and Public Health*. doi: 10.3390/ijerph15040574.

Igolkina, A. A. *et al.* (2019) 'Matching population diversity of rhizobial nod A and legume NFR5 genes in plant–microbe symbiosis', *Ecology and Evolution*, p. ece3.5556. doi: 10.1002/ece3.5556.

Inceoğlu, Ö. *et al.* (2010) 'Effects of plant genotype and growth stage on the betaproteobacterial communities associated with different potato cultivars in two fields', *Applied and Environmental Microbiology*, 76(11), pp. 3675–3684. doi: 10.1128/AEM.00040-10.

Irisarri, P. *et al.* (2019) 'Selection of Competitive and Efficient Rhizobia Strains for White Clover', *Frontiers in Microbiology*, 10, p. 768. doi: 10.3389/fmicb.2019.00768.

Itoh, T. *et al.* (1999) 'Evolutionary Instability of Operon Structures Disclosed by Sequence Comparisons of Complete Microbial Genomes', *Mol. Biol. Evol*, 16(3), pp. 332–346.

Jabara, C. B. *et al.* (2011) 'Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID', *Proc. Natl. Acad. Sci. U. S. A.*, 108(50), pp. 20166–20171. doi: 10.1073/pnas.1110064108.

Jacob, F. and Monod, J. (1961) 'Genetic regulatory mechanisms in the synthesis of proteins', *Journal of Molecular Biology*, pp. 318–356. doi: 10.1016/S0022-2836(61)80072-7.

Jacobs, W., Egelhoff, T. T. and Long, S. R. (1985) 'Physical and Genetic Map of a Rhizobium rneliloti Nodulation Gene Region and Nucleotide Sequence of nodC', *Journal of Bacteriology*, 162(2), pp. 469–476.

Janczarek, M. and Rachwał, K. (2013) 'Mutation in the pssA gene involved in exopolysaccharide synthesis leads to several physiological and symbiotic defects in Rhizobium leguminosarum bv. trifolii', *International Journal of Molecular Sciences*, 14(12), pp. 23711–23735. doi: 10.3390/ijms141223711.

Jensen, E. S. *et al.* (2012) 'Legumes for mitigation of climate change and the provision of feedstock for biofuels and biorefineries. A review', *Agronomy for Sustainable Development*. Springer, pp. 329–364. doi: 10.1007/s13593-011-0056-7.

Jensen, E. S. and Hauggaard-Nielsen, H. (2003) 'How can increased use of biological N2 fixation in agriculture benefit the environment?', *Plant and Soil*. Springer, 252(1), pp. 177–186. doi: 10.1023/A:1024189029226.

Jiao, J. *et al.* (2018) 'Coordinated regulation of core and accessory genes in the multipartite genome of Sinorhizobium fredii', *PLoS Genet.*, 14(5), p. e1007428. doi: 10.1371/journal.pgen.1007428.

Jiménez-Guerrero, I. *et al.* (2017) 'Transcriptomic Studies of the Effect of nod Gene-Inducing Molecules in Rhizobia: Different Weapons, One Purpose', *Genes*, 9(1), p. 1. doi: 10.3390/genes9010001.

Johnson, Z. I. and Chisholm, S. W. (2004) 'Properties of overlapping genes are conserved across microbial genomes', *Genome Research*, 14(11), pp. 2268–2272. doi: 10.1101/gr.2433104.

Jones, D. G. and Hardarson, G. (1979) 'Variation within and between white clover varieties in their preference for strains of Rhizobium trifolii', *Annals of Applied Biology*, 92(2), pp. 221–228. doi: 10.1111/j.1744-7348.1979.tb03867.x.

Jones, P. *et al.* (2019) 'Plant Host-Associated Mechanisms for Microbial Selection', *Frontiers in Plant Science*, 10, p. 862. doi: 10.3389/fpls.2019.00862.

Joseph, M. V, Desai, J. D. and Desai, A. J. (1983) 'Production of Antimicrobial and Bacteriocin-Like Substances by Rhizobium trifolii Downloaded from', *Applied and Environmental Microbiology*, 45(2), pp. 532–535.

Joshi, F. R. *et al.* (2008) 'Siderophore cross-utilization amongst nodule isolates of the cowpea miscellany group and its effect on plant growth in the presence of antagonistic organisms', *Microbiological Research*, 163(5), pp. 564–570. doi: 10.1016/j.micres.2006.08.004.

Jousset, A. *et al.* (2011) 'Genotypic richness and dissimilarity opposingly affect ecosystem functioning', *Ecol. Lett.*, 14(6), pp. 537–545. doi: 10.1111/j.1461-0248.2011.01613.x.

Junier, I. and Rivoire, O. (2016) 'Conserved units of co-expression in bacterial genomes: An evolutionary insight into transcriptional regulation', *PLoS ONE*, 11(5). doi: 10.1371/journal.pone.0155740.

Kannenberg, E. L., Rathbun, E. A. and Brewin, N. J. (1992) 'Molecular dissection of structure and function in the lipopolysaccharide of Rhizobium leguminosarum strain 3841 using monoclonal antibodies and genetic analysis', *Molecular Microbiology*, 6(17), pp. 2477–2487. doi: 10.1111/j.1365-2958.1992.tb01424.x.

Karst, S. M. *et al.* (2019) 'Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers and Nanopore sequencing', *bioRxiv*, p. 645903. doi: 10.1101/645903.

Karunakaran, R. *et al.* (2009) 'Transcriptomic Analysis of Rhizobium leguminosarum Biovar viciae in Symbiosis with Host Plants Pisum sativum and Vicia cracca', *J. Bacteriol.*, 191(12), pp. 4002–4014. doi: 10.1128/jb.00165-09.

Kawaharada, Y. *et al.* (2015) 'Receptor-mediated exopolysaccharide perception controls bacterial infection', *Nature*, 523(7560), pp. 308–312. doi: 10.1038/nature14611.

Kazmierczak, T. *et al.* (2017) 'Specific Host-Responsive Associations Between Medicago truncatula Accessions and Sinorhizobium Strains', *Molecular Plant-Microbe Interactions*, 30(5), pp. 399–409. doi: 10.1094/MPMI-01-17-0009-R.

Kebschull, J. M. and Zador, A. M. (2015) 'Sources of PCR-induced distortions in high-throughput sequencing data sets', *Nucleic Acids Res.*, 43(21), p. e143. doi: 10.1093/nar/gkv717.

Kereszt, A., Mergaert, P. and Kondorosi, E. (2011) 'Bacteroid development in legume nodules: Evolution of mutual benefit or of sacrificial victims?', *Molecular Plant-Microbe Interactions*, pp. 1300–1309. doi: 10.1094/MPMI-06-11-0152.

Kiers, E. T. and Denison, R. F. (2008) 'Sanctions, Cooperation, and the Stability of Plant-Rhizosphere Mutualisms', *Annu. Rev. Ecol. Evol. Syst.*, 39(1), pp. 215–236. doi: 10.1146/annurev.ecolsys.39.110707.173423.

Kiers, E. T., West, S. A. and Denison, R. F. (2002) 'Mediating Mutualisms: Farm Management Practices and Evolutionary Changes in Symbiont Co-operation', *J. Appl. Ecol.*, 39(5), pp. 745–754.

Kimes, N. E. *et al.* (2014) 'RNA sequencing provides evidence for functional variability between naturally co-existing Alteromonas macleodii lineages', *BMC Genomics*, 15(1), p. 938. doi: 10.1186/1471-2164-15-938.

Kinde, I. *et al.* (2011) 'Detection and quantification of rare mutations with massively parallel sequencing', *Proc. Natl. Acad. Sci. U. S. A.*, 108(23), pp. 9530–9535. doi: 10.1073/pnas.1105422108.

Kinoti, W. M. *et al.* (2017) 'Generic Amplicon Deep Sequencing to Determine Ilarvirus Species Diversity in Australian Prunus', *Front. Microbiol.*, 8, p. 1219. doi: 10.3389/fmicb.2017.01219.

Kitada, S., Nakamichi, R. and Kishino, H. (2017) 'The empirical Bayes estimators of fine-scale population structure in high gene flow species', *Molecular Ecology Resources*, 17(6), pp. 1210–1222. doi: 10.1111/1755-0998.12663.

Kivioja, T. *et al.* (2011) 'Counting absolute numbers of molecules using unique molecular identifiers', *Nat. Methods*, 9(1), pp. 72–74. doi: 10.1038/nmeth.1778.

Kloesges, T. *et al.* (2011) 'Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths', *Molecular Biology and Evolution*, 28(2), pp. 1057–1074. doi: 10.1093/molbev/msq297.

Koonin, E. V. (2009) 'Evolution of genome architecture', *International Journal of Biochemistry and Cell Biology*, pp. 298–306. doi: 10.1016/j.biocel.2008.09.015.

Kou, R. *et al.* (2016) 'Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations', *PLOS ONE*, 11(1), p. e0146638. doi: 10.1371/journal.pone.0146638.

Kramer, J., Özkaya, Ö. and Kümmerli, R. (2019) 'Bacterial siderophores in community and host interactions', *Nature Reviews Microbiology*. doi: 10.1038/s41579-019-0284-4.

Krehenwinkel, H. *et al.* (2018) 'Scaling up DNA barcoding - Primer sets for simple and cost efficient arthropod systematics by multiplex PCR and Illumina amplicon sequencing', *Methods Ecol. Evol.*, 9(11), pp. 2181–2193. doi: 10.1111/2041-210X.13064.

Kristiansson, E. *et al.* (2013) 'A novel method for cross-species gene expression analysis', *BMC Bioinformatics*, 14, p. 70. doi: 10.1186/1471-2105-14-70.

Krol, J. E. *et al.* (2008) 'Application of physical and genetic map of Rhizobium leguminosarum bv. trifolii TA1 to comparison of three closely related rhizobial genomes', *Mol. Genet. Genomics*, 279(2), pp. 107–121. doi: 10.1007/s00438-007-0299-9.

Kroll, S., Agler, M. T. and Kemen, E. (2017) 'Genomic dissection of host–microbe and microbe–microbe interactions for advanced plant breeding', *Curr. Opin. Plant Biol.*, 36, pp. 71–78. doi: 10.1016/j.pbi.2017.01.004.

Krueger, F., Andrews, S. R. and Osborne, C. S. (2011) 'Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling', *PLoS One*, 6(1), p. e16607. doi: 10.1371/journal.pone.0016607.

Krysciak, D. *et al.* (2014) 'RNA Sequencing Analysis of the Broad-Host-Range Strain Sinorhizobium fredii NGR234 Identifies a Large Set of Genes Linked to Quorum Sensing-Dependent Regulation in the Background of a traI and ngrI Deletion Mutant', *Appl. Environ. Microbiol.*, 80(18), pp. 5655–5671. doi: 10.1128/aem.01835-14.

Kuhn, A., Luthi-Carter, R. and Delorenzi, M. (2008) 'Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package "annotationTools"', *BMC Bioinformatics*, 9, p. 26. doi: 10.1186/1471-2105-9-26.

Kumar, N. *et al.* (2015) 'Bacterial genospecies that are not ecologically coherent: Population genomics of rhizobium leguminosarum 140133', *Open Biology*, 5(1). doi: 10.1098/rsob.140133.

Kumar, V. *et al.* (2019) 'Long-read amplicon denoising', *Nucleic Acids Research*, 47(18), pp. e104–e104. doi: 10.1093/nar/gkz657.

Laguerre, G. *et al.* (1996) 'Typing of Rhizobia by PCR DNA Fingerprinting and PCR-Restriction Fragment Length Polymorphism Analysis of Chromosomal and Symbiotic Gene Regions: Application to Rhizobium leguminosarum and Its Different Biovars', *Applied and Environmental Microbiology*, 62(6), pp. 2029–2036.

Laguerre, G. *et al.* (2003) 'Compatibility of rhizobial genotypes within natural populations of Rhizobium leguminosarum biovar viciae for nodulation of host legumes.', *Applied and environmental microbiology*. American Society for Microbiology, 69(4), pp. 2276–2283. doi: 10.1128/AEM.69.4.2276-2283.2003.

Lang, J. and Faure, D. (2014) 'Functions and regulation of quorum-sensing in Agrobacterium tumefaciens', *Frontiers in Plant Science*, p. 14. doi: 10.3389/fpls.2014.00014.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, 9, p. 559. doi: 10.1186/1471-2105-9-559.

Laranjo, M., Alexandre, A. and Oliveira, S. (2014) 'Legume growth-promoting rhizobia: An overview on the Mesorhizobium genus', *Microbiol. Res.*, 169(1), pp. 2–17. doi: 10.1016/j.micres.2013.09.012.

Lawrence, J. (1999) 'Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes', *Current Opinion in Genetics and Development*, pp. 642–648. doi: 10.1016/S0959-437X(99)00025-8.

Lazar, I. and Lazar, I. (2012) 'Gel Analyzer 2010a: Freeware 1D gel electrophoresis image analysis software. 2010'.

Ledgard, S. F. (2001) 'Nitrogen cycling in low input legume-based agriculture, with emphasis on legume/grass pastures', *Plant and Soil*. Springer, 228(1), pp. 43–59. doi: 10.1023/A:1004810620983.

Lercher, M. J. and Pal, C. (2008) 'Integration of Horizontally Transferred Genes into Regulatory Interaction

Networks Takes Many Million Years', *Molecular Biology and Evolution*, 25(3), pp. 559–567. doi: 10.1093/molbev/msm283.

Leung, K., Wanjage, F. N. and Bottomley, P. J. (1994) 'Symbiotic Characteristics of Rhizobium leguminosarum bv. trifolii Isolates Which Represent Major and Minor Nodule-Occupying Chromosomal Types of Field-Grown Subclover (Trifolium subterraneum L.)t', *Applied and Environmental Microbiology*, 60(2), pp. 427–433.

Li, J. *et al.* (2010) 'Normalization, testing, and false discovery rate estimation for RNA-sequencing data', *Biostatistics*, 13(3), pp. 1–38. doi: 10.1093/biostatistics/kxr031.

Li, Y. *et al.* (2013) 'High-Resolution Transcriptomic Analyses of Sinorhizobium sp NGR234 Bacteroids in Determinate Nodules of Vigna unguiculata and Indeterminate Nodules of Leucaena leucocephala', *PLoS One*, 8(8). doi: 10.1371/journal.pone.0070531.

Lindahl, B. D. *et al.* (2013) 'Fungal community analysis by high-throughput sequencing of amplified markers--a user's guide', *New Phytol.*, 199(1), pp. 288–299. doi: 10.1111/nph.12243.

Lindström, K. *et al.* (2010) 'The biodiversity of beneficial microbe-host mutualism: The case of rhizobia', *Research in Microbiology*, 161(6), pp. 453–463. doi: 10.1016/j.resmic.2010.05.005.

Lingzhi, L. *et al.* (2018) 'The role of two-component regulatory system in β-lactam antibiotics resistance', *Microbiological Research*, pp. 126–129. doi: 10.1016/j.micres.2018.07.005.

Lithgow, J. K. *et al.* (2000) 'The regulatory locus cinRI in Rhizobium leguminosarum controls a network of quorum-sensing loci', *Molecular Microbiology*, 37(1), pp. 81–97. doi: 10.1046/j.1365-2958.2000.01960.x.

Liu, F. *et al.* (2019) 'Soil indigenous microbiome and plant genotypes cooperatively modify soybean rhizosphere microbiome assembly', *BMC Microbiology*, 19(1). doi: 10.1186/s12866-019-1572-x.

Liu, X. D. *et al.* (2014) 'Global transcriptome analysis of Mesorhizobium alhagi CCNWXJ12-2 under salt stress', *BMC Microbiol.*, 14. doi: 10.1186/s12866-014-0319-y.

Lodwig, E. M. *et al.* (2003) 'Amino-acid cycling drives nitrogen fixation in the legume-Rhizobium symbiosis', *Nature*, 422(6933), pp. 722–726. doi: 10.1038/nature01527.

Lopez-Leal, G. *et al.* (2014) 'RNA-Seq analysis of the multipartite genome of Rhizobium etli CE3 shows different replicon contributions under heat and saline shock', *BMC Genomics*, 15, p. 770. doi: 10.1186/1471-2164-15-770.

Łotocka, B., Kopcińska, J. and Skalniak, M. (2012) 'Review article: The meristem in indeterminate root nodules of Faboideae', in *Symbiosis*, pp. 63–72. doi: 10.1007/s13199-013-0225-3.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biol.*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

LoVerso, P. R. and Cui, F. (2015) 'A Computational Pipeline for Cross-Species Analysis of RNA-seq Data Using R and Bioconductor', *Bioinform. Biol. Insights*, 9, pp. 165–174. doi: 10.4137/BBI.S30884.

Lu, J. *et al.* (2017) 'Co-existence of Rhizobia and Diverse Non-rhizobial Bacteria in the Rhizosphere and Nodules of Dalbergia odorifera Seedlings Inoculated with Bradyrhizobium elkanii, Rhizobium multihospitium–Like and Burkholderia pyrrocinia–Like Strains', *Frontiers in Microbiology*, 8(NOV), p. 2255. doi: 10.3389/fmicb.2017.02255.

Lundberg, D. S. *et al.* (2012) 'Defining the core Arabidopsis thaliana root microbiome', *Nature*, 488. doi: 10.1038/nature11237.

Lundberg, D. S. *et al.* (2013) 'Practical innovations for high-throughput amplicon sequencing', *Nat. Methods*, 10(10), pp. 999–1002. doi: 10.1038/nmeth.2634.

Lupwayi, N. Z., Clayton, G. W. and Rice, W. A. (2006) 'Rhizobial Inoculants for Legume Crops', *Journal of Crop Improvement*, 15(2), pp. 289–321. doi: 10.1300/J411v15n02_09.

Lüscher, A. *et al.* (2014) 'Potential of legume-based grassland-livestock systems in Europe: A review', *Grass and Forage Science*, pp. 206–228. doi: 10.1111/gfs.12124.

MacLean, A. M., Finan, T. M. and Sadowsky, M. J. (2007) 'Genomes of the symbiotic nitrogen-fixing bacteria of legumes', *Plant Physiol.*, 144(2), pp. 615–622. doi: 10.1104/pp.107.101634.

Madritsch, S. *et al.* (2019) 'Elucidating Drought Stress Tolerance in European Oaks Through Cross-Species Transcriptomics'. doi: 10.1534/g3.119.400456.

Maier, T. *et al.* (2011) 'Quantification of mRNA and protein and integration with protein turnover in a bacterium', *Molecular Systems Biology*, 7, p. 511. doi: 10.1038/msb.2011.38.

Maj, D. *et al.* (2010) 'Response to flavonoids as a factor influencing competitiveness and symbiotic activity of Rhizobium leguminosarum', *Microbiological Research*, 165(1), pp. 50–60. doi: 10.1016/j.micres.2008.06.002.

Mao, F. *et al.* (2009) 'DOOR: A database for prokaryotic operons', *Nucleic Acids Research*, 37(SUPPL. 1). doi: 10.1093/nar/gkn757.

Mao, X. *et al.* (2014) 'DOOR 2.0: Presenting operons and their functions through dynamic and integrated views', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1048.

Masson-Boivin, C. *et al.* (2009) ‘Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes?’, *Trends in Microbiology*, pp. 458–466. doi: 10.1016/j.tim.2009.07.004.

Mauchline, T. H. *et al.* (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(2), pp. 238–246. doi: 10.1111/lam.12270.

Mazur, A. *et al.* (2011) ‘Intragenomic diversity of Rhizobium leguminosarum bv. trifolii clover nodule isolates’, *BMC Microbiol.*, 11, p. 123. doi: 10.1186/1471-2180-11-123.

Mazur, A. *et al.* (2013) ‘Phenotype profiling of Rhizobium leguminosarum bv. trifolii clover nodule isolates reveal their both versatile and specialized metabolic capabilities’, *Archives of Microbiology*, 195(4), pp. 255–267. doi: 10.1007/s00203-013-0874-x.

McAnulla, C. *et al.* (2007) ‘Quorum-sensing-regulated transcriptional initiation of plasmid transfer and replication genes in Rhizobium leguminosarum biovar viciae’, *Microbiology*, 153, pp. 2074–2082. doi: 10.1099/mic.0.2007/007153-0.

McClure, R. *et al.* (2013) ‘Computational analysis of bacterial RNA-Seq data’, *Nucleic Acids Res.*, 41(14), p. e140. doi: 10.1093/nar/gkt444.

McGinn, K. J. *et al.* (2016) ‘Trifolium species associate with a similar richness of soil-borne mutualists in their introduced and native ranges’, *Journal of Biogeography*, 43(5), pp. 944–954. doi: 10.1111/jbi.12690.

McInerney, J. O., McNally, A. and O’Connell, M. J. (2017) ‘Why prokaryotes have pangenomes’, *Nature Microbiology*, 2(4), p. 17040. doi: 10.1038/nmicrobiol.2017.40.

Mendoza-Suárez, M. A. *et al.* (2020) ‘Optimizing Rhizobium-legume symbioses by simultaneous measurement of rhizobial competitiveness and N2 fixation in nodules’, *Proceedings of the National Academy of Sciences of the United States of America*, 117(18), pp. 9822–9831. doi: 10.1073/pnas.1921225117.

Mergaert, P. *et al.* (2006) ‘Eukaryotic control on bacterial cell cycle and differentiation in the Rhizobium-legume symbiosis’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(13), pp. 5230–5235. doi: 10.1073/pnas.0600912103.

Miao, J. *et al.* (2018) ‘Soil commensal rhizobia promote Rhizobium etli nodulation efficiency through CinR-mediated quorum sensing’, *Archives of Microbiology*, 200(5), pp. 685–694. doi: 10.1007/s00203-018-1478-2.

Miller, M. B. and Bassler, B. L. (2001) ‘Quorum sensing in bacteria’, 12, p. 48. Available at: www.annualreviews.org (Accessed: 25 March 2020).

Miller, R. W. (1991) ‘Glutamate and γ -Aminobutyrate Metabolism in Isolated Rhizobium meliloti Bacteroids’, *Molecular Plant-Microbe Interactions*, 4(1), p. 37. doi: 10.1094/mpmi-4-037.

Miller, S. H. *et al.* (2007) ‘Host-specific regulation of symbiotic nitrogen fixation in Rhizobium leguminosarum biovar trifolii’, *Microbiology*, 153, pp. 3184–3195. doi: 10.1099/mic.0.2007/006924-0.

Miranda-Sánchez, F., Rivera, J. and Vinuesa, P. (2016) ‘Diversity patterns of Rhizobiaceae communities inhabiting soils, root surfaces and nodules reveal a strong selection of rhizobial partners by legumes’, *Environmental Microbiology*, 18(8), pp. 2375–2391. doi: 10.1111/1462-2920.13061.

Mishra, D. J. *et al.* (2013) ‘Role of bio-fertilizer in organic agriculture: a review’, *Research Journal of Recent Sciences*, 2(1), pp. 39–41.

Munoz Aguilar, J. M. *et al.* (1988) ‘Chemotaxis of Rhizobium leguminosarum biovar phaseoli towards Flavonoid Inducers of the Symbiotic Nodulation Genes’, *Microbiology*, 134(10), pp. 2741–2746. doi: 10.1099/00221287-134-10-2741.

Mutch, L. A. and Young, J. P. W. (2004) ‘Diversity and specificity of Rhizobium leguminosarum biovar viciae on wild and cultivated legumes’, *Mol. Ecol.*, 13(8), pp. 2435–2444. doi: 10.1111/j.1365-294X.2004.02259.x.

Mytton, L. R. (1975) ‘Plant genotype x rhizobium strain interactions in white clover’, *Ann. appl. Biol*, 80(1g75), pp. 103–107.

Nearing, J. T. *et al.* (2018) ‘Denoising the Denoisers: An independent evaluation of microbiome sequence error- correction approaches’, *PeerJ*, 2018(8). doi: 10.7717/peerj.5364.

Ng, W. L. *et al.* (2019) ‘Comparative transcriptomics sheds light on differential adaptation and species diversification between two Melastoma species and their F1 hybrid’, *AoB PLANTS*, 11(2). doi: 10.1093/aobpla/plz019.

Nikaido, H. (2018) ‘RND transporters in the living world’, *Research in Microbiology*, 169(7–8), pp. 363–371. doi: 10.1016/j.resmic.2018.03.001.

Nleya, T., Walley, F. and Vandenberg, A. (2001) ‘Response of four common bean cultivars to granular inoculant in a short-season dryland production system’, *Canadian Journal of Plant Science*, 81, pp. 385–390.

O’Hara, G. W. (1998) ‘The Role of Nitrogen Fixation in Crop Production’, *Journal of Crop Production*, 1(2), pp. 115–138. doi: 10.1300/J144v01n02_05.

Okazaki, S. *et al.* (2013) ‘Hijacking of leguminous nodulation signaling by the rhizobial type III secretion system’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(42), pp. 17131–17136. doi: 10.1073/pnas.1302360110.

Okazaki, S. *et al.* (2016) ‘Rhizobium-legume symbiosis in the absence of Nod factors: Two possible scenarios with or without the T3SS’, *ISME Journal*, 10(1), pp. 64–74. doi: 10.1038/ismej.2015.103.

Okuda, S. *et al.* (2007) ‘Characterization of relationships between transcriptional units and operon structures in Bacillus subtilis and Escherichia coli’, *BMC Genomics*, 8, p. 48. doi: 10.1186/1471-2164-8-48.

Oldroyd, G. E. D. *et al.* (2011) ‘The Rules of Engagement in the Legume-Rhizobial Symbiosis’, *Annu. Rev. Genet. 2011.*, 45, pp. 119–144. doi: 10.1146/annurev-genet-110410-132549.

Oleksiak, M. F., Churchill, G. A. and Crawford, D. L. (2002) ‘Variation in gene expression within and among natural populations’, *Nature Genetics*, 32(2), pp. 261–266. doi: 10.1038/ng983.

Oliver, A. K. *et al.* (2015) ‘Polymerase matters: Non-proofreading enzymes inflate fungal community richness estimates by up to 15%’, *Fungal Ecology*, 15, pp. 86–89. doi: 10.1016/j.funeco.2015.03.003.

Oliver, H. F. *et al.* (2009) ‘Deep RNA sequencing of L. monocytogenes reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs.’, *BMC genomics*, 10, p. 641. doi: 10.1186/1471-2164-10-641.

Oresnik, I. J., Twelker, S. and Hynes, M. F. (1999) ‘Cloning and Characterization of a Rhizobium leguminosarum Gene Encoding a Bacteriocin with Similarities to RTX Toxins’, *Applied and Environmental Microbiology*, 65(7), pp. 2833–2840.

Ormeño-Orrillo, E. *et al.* (2015) ‘Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics’, *Systematic and Applied Microbiology*, pp. 287–291. doi: 10.1016/j.syapm.2014.12.002.

Ormeño-Orrillo, E. and Martínez-Romero, E. (2013) ‘Phenotypic tests in Rhizobium species description: An opinion and (a sympatric speciation) hypothesis’, *Systematic and Applied Microbiology*, 36(3), pp. 145–147. doi: 10.1016/j.syapm.2012.11.009.

Osbourn, A. E. and Field, B. (2009) ‘Operons’, *Cellular and Molecular Life Sciences*, pp. 3755–3775. doi: 10.1007/s00018-009-0114-3.

Ott, T. *et al.* (2005) ‘Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development’, *Current Biology*, 15(6), pp. 531–535. doi: 10.1016/j.cub.2005.01.042.

Paffetti, D. *et al.* (1996) ‘Genetic diversity of an Italian Rhizobium meliloti population from different Medicago sativa varieties.’, *Applied and environmental microbiology*, 62(7), pp. 2279–85.

Pahua, V. J. *et al.* (2018) ‘Fitness variation among host species and the paradox of ineffective rhizobia’, *J. Evol. Biol.*, 31(4), pp. 599–610. doi: 10.1111/jeb.13249.

Palmer, J. M. *et al.* (2018) ‘Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data’, *PeerJ*, 6, p. e4925. doi: 10.7717/peerj.4925.

Pavey, S. A. *et al.* (2010) ‘The role of gene expression in ecological speciation’, *Annals of the New York Academy of Sciences*, pp. 110–129. doi: 10.1111/j.1749-6632.2010.05765.x.

Peiffer, J. A. *et al.* (2013) ‘Diversity and heritability of the maize rhizosphere microbiome under field conditions’, *Proc. Natl. Acad. Sci. U. S. A.*, 110. doi: 10.1073/pnas.1302837110.

Pelly, S. *et al.* (2016) ‘REMap: Operon map of M. tuberculosis based on RNA sequence data’, *Tuberculosis*, 99, pp. 70–80. doi: 10.1016/j.tube.2016.04.010.

Peng, J. L. *et al.* (2014) ‘RNA-Seq and Microarrays Analyses Reveal Global Differential Transcriptomes of Mesorhizobium huakuii 7653R between Bacteroids and Free-Living Cells’, *PLoS One*, 9(4). doi: 10.1371/journal.pone.0093626.

Perez-Montano, F. *et al.* (2016) ‘A transcriptomic analysis of the effect of genistein on Sinorhizobium fredii HH103 reveals novel rhizobial genes putatively involved in symbiosis’, *Sci. Rep.*, 6. doi: 10.1038/srep31592.

Pérez Carrascal, O. M. *et al.* (2016) ‘Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing Rhizobium species associated with Phaseolus vulgaris’, *Environmental Microbiology*, 18(8), pp. 2660–2676. doi: 10.1111/1462-2920.13415.

Perret, X., Staehelin, C. and Broughton, W. J. (2000) ‘Molecular Basis of Symbiotic Promiscuity’, *Microbiology and Molecular Biology Reviews*, 64(1), pp. 180–201.

Pertea, M. *et al.* (2009) ‘OperonDB: A comprehensive database of predicted operons in microbial genomes’, *Nucleic Acids Research*, 37(SUPPL. 1). doi: 10.1093/nar/gkn784.

Phelan, P. *et al.* (2015) ‘Forage Legumes for Grazing and Conserving in Ruminant Production Systems’, *Critical Reviews in Plant Sciences*, 34(1–3), pp. 281–326. doi: 10.1080/07352689.2014.898455.

Philippot, L. *et al.* (2013) ‘Going back to the roots: The microbial ecology of the rhizosphere’, *Nature Reviews Microbiology*, pp. 789–799. doi: 10.1038/nrmicro3109.

Pini, F. *et al.* (2011) ‘Plant-bacteria association and symbiosis: Are there common genomic traits in alphaproteobacteria?’, *Genes*, 2(4), pp. 1017–1032. doi: 10.3390/genes2041017.

Plazinski, J. and Rolfe, B. G. (1985) ‘Interaction of azospirillum and Rhizobium strains leading to inhibition of nodulation’, *Appl. Environ. Microbiol.*, 49(4), pp. 990–993.

Poinsot, V. *et al.* (2016) ‘New insights into Nod factor biosynthesis: Analyses of chitooligomers and lipo-chitooligomers of Rhizobium sp. IRBG74 mutants’, *Carbohydrate Research*, 434, pp. 83–93. doi: 10.1016/j.carres.2016.08.001.

Poirier, S. *et al.* (2018) ‘Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing’, *PLoS One*, 13(9), p. e0204629. doi: 10.1371/journal.pone.0204629.

Poisot, T., Péquin, B. and Gravel, D. (2013) ‘High-Throughput Sequencing: A Roadmap Toward Community Ecology’, *Ecology and Evolution*, 3(4), pp. 1125–1139. doi: 10.1002/ece3.508.

Pongsilip, N. (2012) *Phenotypic and Genotypic Diversity of Rhizobia*. Bentham Science Publishers. doi: 10.2174/97816080546191120101.

Poole, P. *et al.* (1999) ‘Regulation of the *mdh-sucCDAB* operon in *Rhizobium leguminosarum*’, *FEMS Microbiology Letters*, 176(1), pp. 247–255. doi: 10.1111/j.1574-6968.1999.tb13669.x.

Poole, P., Ramachandran, V. and Terpolilli, J. (2018) ‘Rhizobia: from saprophytes to endosymbionts’, *Nat. Rev. Microbiol.*, 16(5), pp. 291–303. doi: 10.1038/nrmicro.2017.171.

Portella, A. C. F. *et al.* (2009) ‘Modelling antagonic effect of lactic acid eacteria supernatants on some pathogenic bacteria’, *Braz. Arch. Biol. Technol.*, 52(SPE), pp. 29–36. doi: 10.1590/S1516-89132009000700004.

Prell, J. *et al.* (2002) ‘The Rhizobium leguminosarum bv. viciae VF39 γ-aminobutyrate (GABA) aminotransferase gene (gabT) is induced by GABA and highly expressed in bacteroids’, *Microbiology*, 148(2), pp. 615–623. doi: 10.1099/00221287-148-2-615.

Prell, J. and Poole, P. (2006) ‘Metabolic changes of rhizobia in legume nodules’, *Trends in Microbiology*, pp. 161–168. doi: 10.1016/j.tim.2006.02.005.

Price, M. N. *et al.* (2005) ‘A novel method for accurate operon predictions in all sequenced prokaryotes’, *Nucleic Acids Research*, 33(3), pp. 880–892. doi: 10.1093/nar/gki232.

Price, M. N., Arkin, A. P. and Alm, E. J. (2006) ‘The life-cycle of operons’, *PLoS Genetics*, 2(6), pp. 0859–0873. doi: 10.1371/journal.pgen.0020096.

Provorov, N. A., Andronov, E. E. and Onishchuk, O. P. (2017) ‘Forms of natural selection controlling the genomic evolution in nodule bacteria’, *Russ. J. Genet.*, 53(4), pp. 411–419. doi: 10.1134/S1022795417040123.

Pugashetti, B. K., Angle, J. S. and Wagner, G. H. (1982) ‘Soil microorganisms antagonistic towards Rhizobium japonicum’, *Soil Biology and Biochemistry*, 14(1), pp. 45–49. doi: 10.1016/0038-0717(82)90075-X.

Quiza, L., St-Arnaud, M. and Yergeau, E. (2015) ‘Harnessing phytomicrobiome signaling for rhizosphere microbiome engineering.’, *Frontiers in plant science*, 6, p. 507. doi: 10.3389/fpls.2015.00507.

Rachwal, K. *et al.* (2016) ‘The Regulatory Protein RosR Affects Rhizobium leguminosarum bv. trifolii Protein Profiles, Cell Surface Properties, and Symbiosis with Clover’, *Front. Microbiol.*, 7. doi: 10.3389/fmicb.2016.01302.

Rachwal, K., Matczynska, E. and Janczarek, M. (2015) ‘Transcriptome profiling of a Rhizobium leguminosarum bv. trifolii rosR mutant reveals the role of the transcriptional regulator RosR in motility, synthesis of cell-surface components, and other cellular processes’, *BMC Genomics*, 16. doi: 10.1186/s12864-015-2332-4.

Ramachandran, V. K. *et al.* (2011) ‘Adaptation of Rhizobium leguminosarum to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics’, *Genome Biol.*, 12(10). doi: 10.1186/gb-2011-12-10-r106.

Ramírez-Babena, M. H. *et al.* (2008) ‘Revision of the taxonomic status of the species Rhizobium leguminosarum (Frank 1879) Frank 1889AL, Rhizobium phaseoli Dangeard 1926AL and Rhizobium trifolii Dangeard 1926AL. R. trifolii is a later synonym of R. leguminosarum.Reclassification of the strain’, *International Journal of Systematic and Evolutionary Microbiology*, 58(11), pp. 2484–2490. doi: 10.1099/ijs.0.65621-0.

Ramírez-Bahena, M.-H. *et al.* (2009) ‘Phenotypic, genotypic, and symbiotic diversities in strains nodulating clover in different soils in Spain’, *Canadian Journal of Microbiology*, 55(10), pp. 1207–1216. doi: 10.1139/W09-074.

Ranz, J. M. and Machado, C. A. (2006) ‘Uncovering evolutionary patterns of gene expression using microarrays’, *Trends in Ecology and Evolution*. doi: 10.1016/j.tree.2005.09.002.

Rashid, M. H. or *et al.* (2015) ‘Average nucleotide identity of genome sequences supports the description of Rhizobium lentis sp. nov., Rhizobium bangladeshense sp. nov. and Rhizobium binae sp. nov. from lentil (Lens culinaris) nodules’, *International Journal of Systematic and Evolutionary Microbiology*, 65(9), pp.

3037–3045. doi: 10.1099/ijs.0.000373.

Ravin, A. W. (1960) 'The origin of bacterial species. Genetic recombination and factors limiting it between bacterial populations.', *Bacteriological reviews*, 24(2), pp. 201–220.

Ravin, A. W. (1963) 'Experimental Approaches to the Study of Bacterial Phylogeny', *The American Naturalist*, 97(896), pp. 307–318. doi: 10.1086/282282.

Redmond, J. W. *et al.* (1986) 'Flavones induce expression of nodulation genes in Rhizobium', *Nature*, 323(6089), pp. 632–635. doi: 10.1038/323632a0.

Reeve, W., O'Hara, G., Chain, P., Ardley, J., Brau, L., Nandesena, K., Tiwari, R., Copeland, A., *et al.* (2010) 'Complete genome sequence of Rhizobium leguminosarum bv. trifolii strain WSM1325, an effective microsymbiont of annual Mediterranean clovers', *Stand. Genomic Sci.*, 2(3), pp. 347–356. doi: 10.4056/sigs.852027.

Reeve, W., O'Hara, G., Chain, P., Ardley, J., Brau, L., Nandesena, K., Tiwari, R., Malfatti, S., *et al.* (2010) 'Complete genome sequence of Rhizobium leguminosarum bv trifolii strain WSM2304, an effective microsymbiont of the South American clover Trifolium polymorphum', *Stand. Genomic Sci.*, 2(1), pp. 66–76. doi: 10.4056/sigs.44642.

Reeve, W. *et al.* (2013) 'Genome sequence of the clover-nodulating Rhizobium leguminosarum bv. trifolii strain TA1', *Stand. Genomic Sci.*, 9(2), pp. 243–253. doi: 10.4056/sigs.4488254.

Reeve, W. *et al.* (2015) 'A Genomic Encyclopedia of the Root Nodule Bacteria: assessing genetic diversity through a systematic biogeographic survey', *Standards in Genomic Sciences*, 10(1), p. 14. doi: 10.1186/1944-3277-10-14.

Ribeiro, R. A. *et al.* (2009) 'Multilocus sequence analysis of Brazilian Rhizobium microsymbionts of common bean (Phaseolus vulgaris L.) reveals unexpected taxonomic diversity', *Research in Microbiology*, 160(4), pp. 297–306. doi: 10.1016/j.resmic.2009.03.009.

Rice, W. A., Penney, D. C. and Nyborg, M. (1977) 'Effects of soil acidity on rhizobia numbers, nodulation and nitrogen fixation by alfalfa and red clover', *Canadian Journal of Soil Science*, 57(2), pp. 197–203. doi: 10.4141/cjss77-024.

Richter, M. and Rosselló-Móra, R. (2009) 'Shifting the genomic gold standard for the prokaryotic species definition', *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), pp. 19126–19131. doi: 10.1073/pnas.0906412106.

Ritsema, T. *et al.* (1996) 'Rhizobium nodulation protein NodA is a host-specific determinant of the transfer of fatty acids in Nod factor biosynthesis', 251, pp. 44–51.

Roberts, R. *et al.* (2017) 'Is there sufficient Ensifer and Rhizobium species diversity in UK farmland soils to support red clover (Trifolium pratense), white clover (T. repens), lucerne (Medicago sativa) and black medic (M. lupulina)?', *Applied Soil Ecology*, 120, pp. 35–43. doi: 10.1016/j.apsoil.2017.06.030.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.

Robleto, E. A., Borneman, J. and Triplett, E. W. (1998) 'Effects of Bacterial Antibiotic Production on Rhizosphere Microbial Communities from a Culture-Independent Perspective', *Applied and Environmental Microbiology*, 64(12), pp. 5020–5022.

Rocha, E. P. C. (2008) 'The Organization of the Bacterial Genome', *Annual Review of Genetics*. Annual Reviews, 42(1), pp. 211–233. doi: 10.1146/annurev.genet.42.110807.091653.

Rodelas, B. *et al.* (1999) 'Analysis of quorum-sensing-dependent control of rhizosphere-expressed (rhi) genes in Rhizobium leguminosarum bv. viciae', *Journal of Bacteriology*, 181(12), pp. 3816–3823. doi: 10.1128/jb.181.12.3816-3823.1999.

Rogel, M. A., Ormeño-Orrillo, E. and Martinez Romero, E. (2011) 'Symbiovars in rhizobia reflect bacterial adaptation to legumes', *Systematic and Applied Microbiology*, pp. 96–104. doi: 10.1016/j.syapm.2010.11.015.

Romero, P. R. and Karp, P. D. (2004) 'Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases', *Bioinformatics*, 20(5), pp. 709–717. doi: 10.1093/bioinformatics/btg471.

Roughley, R. J., Blowes, W. M. and Hurridge, D. F. (1976) 'Nodulation of Trifolium subterraneum by introduced rhizobia in competition with naturalized strains', *Soil Biology and Biochemistry*, 8(5), pp. 403–407. doi: 10.1016/0038-0717(76)90041-9.

Roux, B. *et al.* (2014) 'An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing', *Plant J.*, 77(6), pp. 817–837. doi: 10.1111/tpj.12442.

Russel, J. *et al.* (2017) 'Antagonism correlates with metabolic similarity in diverse bacteria', *PNAS*, 114(40), pp. 10684–10688. doi: 10.1073/pnas.1706016114.

Russell, P. E. and Jones, D. G. (1975) 'Variation in the selection of Rhizobium trifolii by varieties of red and white clover', *Soil Biology and Biochemistry*, 7(1), pp. 15–18. doi: 10.1016/0038-0717(75)90024-3.

Rys, G. J. and Bonish, P. M. (1981) 'New Zealand Journal of Experimental Agriculture Effectiveness of Rhizobium trifolii populations associated with Trifolium species in Taranaki, New Zealand Effectiveness of Rhizobium trifolii populations associated with Trifolium species in Taranaki, New Zealand', *New Zealand Journal of Experimental Agriculture*, 9, pp. 327–335. doi: 10.1080/03015521.1981.10425430.

Sabath, N., Graur, D. and Landan, G. (2008) 'Same-strand overlapping genes in bacteria: Compositional determinants of phase bias', *Biology Direct*, p. 36. doi: 10.1186/1745-6150-3-36.

Salgado, H. *et al.* (2000) 'Operons in Escherichia coli: genomic analyses and predictions', *Proc. Natl. Acad. Sci. U. S. A.*, 97(12), pp. 6652–6657. doi: 10.1073/pnas.110147297.

Sánchez-Cañizares, C. *et al.* (2018) 'Genomic Diversity in the Endosymbiotic Bacterium Rhizobium leguminosarum', *Genes*, 9(2), p. 60. doi: 10.3390/genes9020060.

Sanchez-Contreras, M. *et al.* (2007) 'Quorum-sensing regulation in rhizobia and its role in symbiotic interactions with legumes', *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 362(1483), pp. 1149–1163. doi: 10.1098/rstb.2007.2041.

Santos, M. S., Nogueira, M. A. and Hungria, M. (2019) 'Microbial inoculants: reviewing the past, discussing the present and previewing an outstanding future for the use of beneficial bacteria in agriculture', *AMB Express*. doi: 10.1186/s13568-019-0932-0.

Scaria, J. *et al.* (2013) 'Differential Stress Transcriptome Landscape of Historic and Recently Emerged Hypervirulent Strains of Clostridium difficile Strains Determined Using RNA-seq', *PLoS ONE*, 8(11), p. e78489. doi: 10.1371/journal.pone.0078489.

Schmid, M. W. *et al.* (2018) 'Rhizosphere bacterial community composition depends on plant diversity legacy in soil and plant species identity', *bioRxiv preprint*. doi: 10.1101/287235.

Schmidt, A. *et al.* (2011) 'Absolute quantification of microbial proteomes at different states by directed mass spectrometry', *Molecular Systems Biology*, 7, p. 510. doi: 10.1038/msb.2011.37.

Schmidt, J. E., Weese, D. J. and Lau, J. A. (2017) 'Long-term agricultural management does not alter the evolution of a soybean-rhizobium mutualism', *Ecol. Appl.* doi: 10.1002/eap.1625.

Schripsema, J. *et al.* (1996) 'Bacteriocin small of Rhizobium leguminosarum Belongs to the Class of N-Acyl-L-Homoserine Lactone Molecules, Known as Autoinducers and as Quorum Sensing Co-Transcription Factors', *Journal of Bacteriology*, 178(2), pp. 366–371.

Schwinghamer, E. A. A. and Brockwell, J. (1978) 'Competitive advantage of bacteroicin- and phage-producing strains of Rhizobium trifolii in mixed culture', *Soil. Biol. Biochem.*, 10(5), pp. 383–387. doi: 10.1016/0038-0717(78)90062-7.

Sessitsch, A. *et al.* (2002) 'Advances in Rhizobium research', *Critical Reviews in Plant Sciences*. CRC Press LLC, pp. 323–378. doi: 10.1080/0735-260291044278.

Shamseldin, A., Abdelkhalek, A. and Sadowsky, M. J. (2017) 'Recent changes to the classification of symbiotic, nitrogen-fixing, legume-associating bacteria: a review', *Symbiosis*, 71(2), pp. 91–109. doi: 10.1007/s13199-016-0462-3.

Shapiro, B. J. and Polz, M. F. (2015) 'Microbial speciation', *Cold Spring Harbor Perspectives in Biology*, 7(10). doi: 10.1101/cshperspect.a018143.

Shu, W. *et al.* (2012) 'Abundance and diversity of nitrogen-fixing bacteria in rhizosphere and bulk paddy soil under different duration of organic management', *World J. Microbiol. Biotechnol.*, 28(2), pp. 493–503. doi: 10.1007/s11274-011-0840-1.

Silva, V. M. A. *et al.* (2019) 'Cross-Feeding Among Soil Bacterial Populations: Selection and Characterization of Potential Bio-inoculants', *Journal of Agricultural Science*, 11(5). doi: 10.5539/jas.v11n5p23.

Simms, E. L. and Taylor, D. L. (2002) 'Partner Choice in Nitrogen-Fixation Mutualisms of Legumes and Rhizobia', *Integrative and Comparative Biology*, 42(2), pp. 369–380. doi: 10.1093/icb/42.2.369.

Skorupska, A. *et al.* (2006) 'Rhizobial exopolysaccharides: genetic control and symbiotic functions', *Microb. Cell Fact.*, 5. doi: 10.1186/1475-2859-5-7.

Slager, J., Aprianto, R. and Veening, J.-W. (2018) 'Deep genome annotation of the opportunistic human pathogen Streptococcus pneumoniae D39', *Nucleic Acids Res.*, 46(19), pp. 9971–9989. doi: 10.1093/nar/gky725.

Smith, E. (2018) *Discovering the phenotypic variation within Rhizobium leguminosarum and determining the best strains for soil inoculums. MSc by Research Project.* University of York, UK.

Somasegaran, P. and Hoben, H. (1985) *Handbook for Rhizobia: Methods in legume-Rhizobium technology*. University of Hawaii NifTAL Project and MIRCEN. doi: 10.1007/978-1-4613-8375-8.

Song, X. *et al.* (2018) 'Changes in the Microbial Community Structure and Soil Chemical Properties of Vertisols Under Different Cropping Systems in Northern China', *Frontiers in Environmental Science*, 6, p. 132. doi: 10.3389/fenvs.2018.00132.

Spaink, Herman P. *et al.* (1987) 'Promoters in the nodulation region of the Rhizobium leguminosarum Sym plasmid pRLIJI', *Plant Molecular Biology*, 9, pp. 27–39.

Spaink, H P *et al.* (1987) 'Rhizobium nodulation gene nodD as a determinant of host specificity', *Nature*, 328(6128), pp. 337–340. doi: 10.1038/328337a0.

Sprent, J. I., Sutherland, J. M. and De Faria, S. M. (1987) 'Some aspects of the biology of nitrogen-fixing organisms', *Trans. R. Soc. Lond. B*, 317, pp. 111–129.

Stackebrandt, E., Murray, R. G. E. and Truper, H. G. (1988) 'Proteobacteria classis nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives"', *International Journal of Systematic Bacteriology*, 38(3), pp. 321–325. doi: 10.1099/00207713-38-3-321.

Stagnari, F. *et al.* (2017) 'Multiple benefits of legumes for agriculture sustainability: an overview', *Chemical and Biological Technologies in Agriculture*, 4(1), p. 2. doi: 10.1186/s40538-016-0085-1.

Stefan, A. *et al.* (2018) 'Genetic diversity and structure of Rhizobium leguminosarum populations associated with clover plants are influenced by local environmental variables', *Syst. Appl. Microbiol.*, 41(3), pp. 251–259. doi: 10.1016/j.syapm.2018.01.007.

Stuart, J. M. *et al.* (2003) 'A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules', *Science*, 302, pp. 249–225. Available at: http://science.sciencemag.org/ (Accessed: 26 November 2019).

Taboada, B. *et al.* (2012) 'ProOpDB: Prokaryotic operon database', *Nucleic Acids Research*, 40(D1). doi: 10.1093/nar/gkr1020.

Taboada, B. *et al.* (2018) 'Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes', *Bioinformatics*, 34(23), pp. 4118–4120. doi: 10.1093/bioinformatics/bty496.

Tamames, J. *et al.* (1997) 'Conserved clusters of functionally related genes in two bacterial genomes', *Journal of Molecular Evolution*, 44(1), pp. 66–73. doi: 10.1007/PL00006122.

Tamames, J. (2001) 'Evolution of gene order conservation in prokaryotes.', *Genome biology*, 2(6). doi: 10.1186/gb-2001-2-6-research0020.

Tartaglia, C. *et al.* (2019) 'Phylogenetic relationships among introduced and autochthonous rhizobia nodulating Trifolium spp. in Uruguayan soils', *Applied Soil Ecology*, 139, pp. 40–46. doi: 10.1016/j.apsoil.2019.03.014.

Team, R. C. (2015) *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Teng, Y. *et al.* (2015) 'Rhizobia and their bio-partners as novel drivers for functional remediation in contaminated soils', *Frontiers in Plant Science*. doi: 10.3389/fpls.2015.00032.

Terpolilli, J. *et al.* (2014) 'Genome sequence of Rhizobium leguminosarum bv trifolii strain WSM1689, the microsymbiont of the one flowered clover Trifolium uniflorum', *Stand. Genomic Sci.*, 9(3), pp. 527–539. doi: 10.4056/sigs.4988693.

Tessler, M. *et al.* (2017) 'Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing', *Sci. Rep.*, 7(1), p. 6589. doi: 10.1038/s41598-017-06665-3.

Tettelin, H. *et al.* (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(39), pp. 13950–13955. doi: 10.1073/pnas.0506758102.

Tilman, D. *et al.* (2002) 'Agricultural sustainability and intensive production practices', *Nature*, pp. 671–677. doi: 10.1038/nature01014.

Tilman, D. *et al.* (2011) 'Global food demand and the sustainable intensification of agriculture', *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp. 20260–20264. doi: 10.1073/pnas.1116437108.

Townsend, J. P., Cavalieri, D. and Hartl, D. L. (2003) 'Population Genetic Variation in Genome-Wide Gene Expression', *Molecular Biology and Evolution*, 20(6), pp. 955–963. doi: 10.1093/molbev/msg106.

Triplett, E. W. and Barta, T. M. (1987) 'Trifolitoxin Production and Nodulation Are Necessary for the Expression of Superior Nodulation Competitiveness by Rhizobium leguminosarum bv. trifolii Strain T24 on Clover'', *Plant Physiol*, 85, pp. 335–342.

Triplett, E. W. and Sadowsky, M. J. (1992) 'Genetics of competition for nodulation of legumes', *Annu. Rev. Microbiol.*, 46(1), pp. 399–428. doi: 10.1146/annurev.mi.46.100192.002151.

Turner, T. R. *et al.* (2013) 'Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants', *ISME Journal*, 7(12), pp. 2248–2258. doi: 10.1038/ismej.2013.119.

Vandamme, P. *et al.* (1996) 'Polyphasic taxonomy, a consensus approach to bacterial systematics.', *Microbiological reviews*, 60(2), pp. 407–38.

Vanderlinde, E. M. and Yost, C. K. (2012) 'Mutation of the sensor kinase chvG in Rhizobium leguminosarum negatively impacts cellular metabolism, outer membrane stability, and symbiosis', *Journal of Bacteriology*, 194(4), pp. 768–777. doi: 10.1128/JB.06357-11.

Vanlauwe, B. *et al.* (2019) 'The role of legumes in the sustainable intensification of African smallholder agriculture: Lessons learnt and challenges for the future', *Agriculture, Ecosystems and Environment*. Elsevier B.V., 284, p. 106583. doi: 10.1016/j.agee.2019.106583.

Veach, A. M. *et al.* (2019) 'Rhizosphere microbiomes diverge among Populus trichocarpa plant-host genotypes and chemotypes, but it depends on soil origin', *Microbiome*, 7(1), p. 76. doi: 10.1186/s40168-019-0668-8.

Vercruysse, M. *et al.* (2011) 'A Comparative Transcriptome Analysis of Rhizobium etli Bacteroids: Specific Gene Expression During Symbiotic Nongrowth', *Mol. Plant. Microbe. Interact.*, 24(12), pp. 1553–1561. doi: 10.1094/mpmi-05-11-0140.

Vessey, J. K. (2003) 'Plant growth promoting rhizobacteria as biofertilizers', *Plant and Soil*, 255(2), pp. 571–586. doi: 10.1023/A:1026037216893.

Vicente, M. and Mingorance, J. (2008) 'Microbial evolution: the genome, the regulome and beyond', *Environmental Microbiology*, 10(7), pp. 1663–1667. doi: 10.1111/j.1462-2920.2008.01635.x.

Villacieros, M. *et al.* (2003) 'Colonization behaviour of Pseudomonas fluorescens and Sinorhizobium meliloti in the alfalfa (Medicago sativa) rhizosphere', *Plant and Soil*, 251(1), pp. 47–54. doi: 10.1023/A:1022943708794.

Vital, M. *et al.* (2015) 'Gene expression analysis of E. coli strains provides insights into the role of gene regulation in diversification.', *The ISME journal*, 9(5), pp. 1130–40. doi: 10.1038/ismej.2014.204.

Vos, M. (2011) 'A species concept for bacteria based on adaptive divergence', *Trends in Microbiology*, 19(1), pp. 1–7. doi: 10.1016/j.tim.2010.10.003.

Vos, M. *et al.* (2012) 'A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity', *PLOS ONE*, 7(2), p. e30600. doi: 10.1371/journal.pone.0030600.

Vuong, H. B., Thrall, P. H. and Barrett, L. G. (2017) 'Host species and environmental variation can influence rhizobial community composition', *J. Ecol.*, 105(2), pp. 540–548. doi: 10.1111/1365-2745.12687.

Wadhwa, K., Dudeja, S. S. and Yadav, R. K. (2011) 'Molecular diversity of native rhizobia trapped by five field pea genotypes in Indian soils', *Journal of Basic Microbiology*, 51(1), pp. 89–97. doi: 10.1002/jobm.201000065.

Wang, D. *et al.* (2012) 'Symbiosis specificity in the legume - rhizobial mutualism', *Cellular Microbiology*, 14(3), pp. 334–342. doi: 10.1111/j.1462-5822.2011.01736.x.

Wang, L. *et al.* (2004) 'Genome-wide operon prediction in Staphylococcus aureus', *Nucleic Acids Res.*, 32(12), pp. 3689–3702. doi: 10.1093/nar/gkh694.

Wang, Q., Liu, J. and Zhu, H. (2018) 'Genetic and Molecular Mechanisms Underlying Symbiotic Specificity in Legume-Rhizobium Interactions', *Frontiers in Plant Science*, 9, p. 313. doi: 10.3389/fpls.2018.00313.

Wang, X. L. *et al.* (2018) 'Rhizobia inhabiting nodules and rhizosphere soils of alfalfa: A strong selection of facultative microsymbionts', *Soil Biology and Biochemistry*, 116, pp. 340–350. doi: 10.1016/J.SOILBIO.2017.10.033.

Wang, Y., MacKenzie, K. D. and White, A. P. (2015) 'An empirical strategy to detect bacterial transcript structure from directional RNA-seq transcriptome data', *BMC Genomics*, 16, p. 359. doi: 10.1186/s12864-015-1555-8.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nat. Rev. Genet.*, 10(1), pp. 57–63. doi: 10.1038/nrg2484.

Weese, D. J. *et al.* (2015) 'Long-term nitrogen addition causes the evolution of less-cooperative mutualists', *Evolution*, 69(3), pp. 631–642. doi: 10.1111/evo.12594.

Wernegreen, J. J. and Riley, M. A. (1999) 'Comparison of the evolutionary dynamics of symbiotic and housekeeping loci: A case for the genetic coherence of rhizobial lineages', *Molecular Biology and Evolution*, 16(1), pp. 98–113. doi: 10.1093/oxfordjournals.molbev.a026041.

Westover, B. P. *et al.* (2005) 'Operon prediction without a training set', *Bioinformatics*, 21(7), pp. 880–888. doi: 10.1093/bioinformatics/bti123.

White, J. P. *et al.* (2009) 'Characterization of a γ-aminobutyric acid transport system of Rhizobium leguminosarum bv. viciae 3841', *Journal of Bacteriology*, 191(5), pp. 1547–1555. doi: 10.1128/JB.00926-08.

Wickham, H. (2009) 'ggplot2: Elegant Graphics for Data Analysis Springer-Verlag', *New York*.

Wiedenbeck, J. and Cohan, F. M. (2011) 'Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches', *FEMS Microbiol. Rev.*, 35(5), pp. 957–976. doi: 10.1111/j.1574-6976.2011.00292.x.

Wielbo, J. *et al.* (2010) 'Genetic and Metabolic Divergence within a Rhizobium leguminosarum bv. trifolii Population Recovered from Clover Nodules', *Appl. Environ. Microbiol.*, 76(14), pp. 4593–4600. doi: 10.1128/aem.00667-10.

Wielbo, J. *et al.* (2011) 'Competitiveness of Rhizobium leguminosarum bv. trifolii Strains in Mixed Inoculation of Clover (Trifolium pratense)', *Pol. J. Microbiol.*, 60(1), pp. 43–49.

Wilkinson, A. *et al.* (2002) 'N-acyl-homoserine lactone inhibition of rhizobial growth is mediated by two quorum-sensing genes that regulate plasmid transfer', *J. Bacteriol.*, 184(16), pp. 4510–4519. doi: 10.1128/JB.184.16.4510-4519.2002.

Wilson, R. A., Handley, B. A. and Beringer, J. E. (1998) ‘Bacteriocin production and resistance in a field population of Rhizobium leguminosarum biovar viciae’, *Soil Biology and Biochemistry*, 30(3), pp. 413–417. doi: 10.1016/S0038-0717(97)00123-5.

Wisniewski-Dyé, F. *et al.* (2002) ‘raiIR genes are part of a quorum-sensing network controlled by cinI and cinR in Rhizobium leguminosarum’, *Journal of Bacteriology*, 184(6), pp. 1597–1606. doi: 10.1128/JB.184.6.1597-1606.2002.

Wisniewski-Dyé, F. and Downie, J. A. (2002) ‘Quorum-sensing in Rhizobium’, *Antonie Van Leeuwenhoek*, 81(1–4), pp. 397–407.

Wolf, J. B. W. *et al.* (2010) ‘Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow’, *Molecular Ecology*, 19, pp. 162–175. doi: 10.1111/j.1365-294X.2009.04471.x.

Wolf, Y. I. *et al.* (2001) ‘Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context’, *Genome Research*, 11(3), pp. 356–372. doi: 10.1101/gr.161901.

Wright, W. *et al.* (2013) ‘Isolation and structural identification of the trihydroxamate siderophore vicibactin and its degradative products from Rhizobium leguminosarum ATCC 14479 bv. trifolii’, *BioMetals*, 26(2), pp. 271–283. doi: 10.1007/s10534-013-9609-3.

Xiong, H. Y. *et al.* (2017) ‘The epidemicity of facultative microsymbionts in faba bean rhizosphere soils’, *Soil Biology and Biochemistry*, 115, pp. 243–252.

Yada, T. *et al.* (1999) ‘Modeling and predicting transcriptional units of Escherichia coligenes using hidden Markov models’, *Bioinformatics*, 15(12), pp. 987–993. doi: 10.1093/bioinformatics/15.12.987.

Yang, C. *et al.* (2017) ‘Symbiosis of selected Rhizobium leguminosarum bv. viciae strains with diverse pea genotypes: effects on biological nitrogen fixation’, *Can. J. Microbiol.* doi: 10.1139/cjm-2017-0281.

Yang, S. *et al.* (2017) ‘Microsymbiont discrimination mediated by a host-secreted peptide in Medicago truncatula’, *P*, 114(26), pp. 6848–6853. doi: 10.1073/pnas.1700460114.

Yeoman, K. H. *et al.* (1997) ‘High affinity iron acquisition in Rhizobium leguminosarum requires the cycHJKL operon and the feuPQ gene products, which belong to the family of two-component transcriptional regulators’, *Microbiology*, 143(1), pp. 127–134. doi: 10.1099/00221287-143-1-127.

Yoder-Himes, D. R. *et al.* (2009) ‘Mapping the Burkholderia cenocepacia niche response via high-throughput sequencing’, *Proc. Natl. Acad. Sci. U. S. A.*, 106(10), pp. 3976–3981. doi: 10.1073/pnas.0813403106.

Young, J. and Haukka, K. E. (1996) ‘Diversity and phylogeny of rhizobia’, *New Phytologist*, 133(1), pp. 87–94. doi: 10.1111/j.1469-8137.1996.tb04344.x.

Young, J. P. W. *et al.* (2006) ‘The genome of Rhizobium leguminosarum has recognizable core and accessory components’, *Genome Biology*, 7(4). doi: 10.1186/gb-2006-7-4-r34.

Young, J. P. W. (2016) ‘Bacteria Are Smartphones and Mobile Genes Are Apps’, *Trends in Microbiology*. doi: 10.1016/j.tim.2016.09.002.

Zahran, H. H. (2017) ‘Plasmids impact on rhizobia-legumes symbiosis in diverse environments’, *Symbiosis*, pp. 1–17. doi: 10.1007/s13199-017-0476-5.

Zaidi, A., Khan, M. S. and Musarrat, J. (2017) *Microbes for Legume Improvement*. Springer. Available at: https://market.android.com/details?id=book-gng5DwAAQBAJ.

Zarrineh, P. *et al.* (2014) ‘Genome-Scale Co-Expression Network Comparison across Escherichia coli and Salmonella enterica Serovar Typhimurium Reveals Significant Conservation at the Regulon Level of Local Regulators Despite Their Dissimilar Lifestyles’, *PLoS ONE*, 9(8), p. 102871. doi: 10.1371/journal.pone.0102871.

Zee, P. C. and Bever, J. D. (2014) ‘Joint evolution of kin recognition and cooperation in spatially structured rhizobium populations’, *PLoS ONE*, 9(4). doi: 10.1371/journal.pone.0095141.

ZéZé, A., Mutch, L. A. and Young, J. P. . (2001) ‘Direct amplification of nodD from community DNA reveals the genetic diversity of Rhizobium leguminosarum in soil’, *Environmental Microbiology*, 3(6), pp. 363–370. doi: 10.1046/j.1462-2920.2001.00202.x.

Zhalnina, K. *et al.* (2018) ‘Dynamic root exudate chemistry and microbial substrate preferences drive patterns in rhizosphere microbial community assembly’, *Nature Microbiology*, 3. doi: 10.1038/s41564-018-0129-3.

Zhang, J. *et al.* (2014) ‘PEAR: a fast and accurate Illumina Paired-End reAd mergeR’, *Bioinformatics*, 30(5), pp. 614–620. doi: 10.1093/bioinformatics/btt593.

Zhang, P. *et al.* (2019) ‘Effect of Soybean and Maize Rotation on Soil Microbial Community Structure’, *Agronomy*, 9(2), p. 42. doi: 10.3390/agronomy9020042.

Zhang, X. X. *et al.* (2017) ‘Pyrosequencing of rpoB uncovers a significant biogeographical pattern of rhizobial species in soybean rhizosphere’, *Journal of Biogeography*, 44(7), pp. 1491–1499. doi: 10.1111/jbi.12891.

Zhang, Y. M. *et al.* (2012) ‘Robust Markers Reflecting Phylogeny and Taxonomy of Rhizobia’, *PLoS One*, 7(9), p. 6. doi: 10.1371/journal.pone.0044936.

Zheng, Y. *et al.* (2002) ‘Computational Identification of Operons in Microbial Genomes’, *Genome Research*, 12(8), pp. 1221–1230. doi: 10.1101/gr.200602.

Zhu, Y. *et al.* (2014) ‘XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons’, *BMC Genomics*, 15, p. 343. doi: 10.1186/1471-2164-15-343.

Zipfel, C. and Oldroyd, G. E. D. (2017) ‘Plant signalling in symbiosis and immunity’, *Nature*, pp. 328–336. doi: 10.1038/nature22009.