BACTERIAL MULTIPARTITE GENOME EVOLUTION

DIVIDE AND CONQUER: HOW CONQUERING MULTIPLE NICHES INFLUENCED THE EVOLUTION OF THE DIVIDED BACTERIAL GENOME

By

GEORGE COLIN DICENZO, BSc (Hons)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University

DESCRIPTIVE NOTE

McMaster University DOCTOR OF PHILOSOPHY (2017) Hamilton, Ontario (Biology)

TITLE: Divide and conquer: How conquering multiple niches influenced the evolution of the divided bacterial genome

AUTHOR: George Colin diCenzo, B.Sc. (McMaster University)

SUPERVISOR: Professor Turlough M Finan

NUMBER OF PAGES: xxvi, 357

LAY ABSTRACT

Many bacteria that enter into symbiotic or pathogenic relationships with plants, animals, and humans contain a genome that is divided into multiple chromosome-like molecules. One example is the N_2-fixing legume symbiont *Sinorhizobium meliloti*, whose genome contains three chromosome-sized molecules. Here, the functions associated with each molecule in the *S. meliloti* genome were examined through a combination of experimental genetic analyses and computer based simulations. Results from these approaches suggested that adaptation to unique environments selected for the evolution of secondary chromosome-like molecules, with each predominately contributing to growth in a specific environment, including environments associated with an eukaryotic host. The genes on these replicons are therefore prime targets for manipulation of bacterium-host interactions, and represent reservoirs of valuable genes for use in synthetic biology applications.

Additionally, the genome reduction approach employed in this study laid out a ground work for identification of the minimal N_2-fixing symbiotic genome. This represents a crucial step towards successfully engineering improved nitrogen fixation, and the engineering of synthetic N_2-fixing symbioses involving non-legumes and/or non-rhizobia.

ABSTRACT

Approximately 10% of sequenced bacterial genomes are multipartite, consisting of two or more large chromosome-sized replicons. This genome organization can be found in many plant, animal, and human pathogens and symbionts. However, the advantage of harbouring multiple replicons remains unclear. One species with a multipartite genome is *Sinorhizobium meliloti*, a model rhizobium that enters into N_2-fixing symbioses with various legume crops. In this work, *S. meliloti* derivatives lacking one or both of the secondary replicons (termed pSymA and pSymB) were constructed. Phenotypic characterization of these strains, including growth rate, metabolic capacity, and competitive fitness, provided some of the first experimental evidence that secondary replicons evolved to provide a niche specific advantage, improving fitness in a newly colonized environment. These results were further supported by characterizing the symbiotic phenotypes of 36 large-scale pSymA and pSymB deletion mutants. To further this analysis, an *in silico S. meliloti* genome-scale metabolic network reconstruction was developed and flux balance analysis used to examine the contribution of each replicon to fitness in three niches. These simulations were consistent with the hypothesis that metabolic pathways encoded by pSymB improve fitness specifically during growth in the plant-associated rhizosphere. Phylogenetic analysis of a pSymB region containing two essential genes provided a clean example of how a translocation from the primary chromosome to a secondary replicon can render the secondary replicon essential. Moreover, an experimental analysis of genetic redundancy indicated that 10-15% of

chromosomal genes are functionally redundant with a pSymA or pSymB encoded gene, providing an alternative method for how secondary replicons can become essential and influence the evolution of the primary chromosome. Finally, the work presented here provides a novel framework for forward genetic analysis of N_2-fixing symbiosis and the identification of the minimal N_2-fixing symbiotic genome, which will help facilitate the development of synthetic symbioses.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Dr. Turlough Finan. I was first given the opportunity to work in Dr. Finan's lab as a summer student between my second and third years of undergraduate studies. The wonderful learning and research environment that he fostered in his lab enticed me to continuously return, and now here I am nearly 7 years later finishing up my PhD studies. I sincerely thank him for all the knowledge and skills I have learned from him, his patience answering the many (many) e-mails I have sent him, his continual support, the opportunity to attend numerous international conferences, and the freedom he gave me in all of my projects. I am the scientist I am today because of his guidance, and will forever be grateful for having had to opportunity to learn from him.

I would like to thank my committee members, Dr. Marie Elliot and Dr. Brian Golding, for their interest and support throughout this process, and whose advice and contributions to my work were invaluable. I would also like to thank Dr. Richard Morton, whose support and critical thought process challenged me to view my work from a different angle, and adjust my hypotheses and experiments accordingly. I am also grateful to all current at past members of the Finan lab, particularly Dr. Branislava Milunovic, Dr. Ye Zhang, and Dr. Maryam Zamani, for not only teaching me countless techniques and helping me develop as a scientist, but also for all the support and great times we shared both in the lab and outside of the lab.

I would like to thank Dr. Alessio Mengoni for providing me an opportunity to

spend a four month research period in his lab at the University of Florence in Italy, and for assisting in the never ending bureaucracy necessary to keep my stay in Italy legal. My time there was invaluable and gave me the chance to expand my horizons, both in terms of science and life. I would also like to thank Dr. Marco Fondi from the University of Florence, whose guidance was instrumental in the success of my research stay in Italy, and whose teachings provided me a novel skill set and enhanced interest in computational biology. I am also grateful to everyone who made my stay in Florence so enjoyable.

Last, but not least, I would like to sincerely thank my entire family for supporting me not only throughout the completion of this thesis, but during all my years of education. Even though I may not have visited as much as I should have, and spent a lot of my time working on my laptop when I did visit, they were there for me every step of the way. Their support was instrumental in the completion of my studies and getting me to where I am now.

For the work in Chapter 2, I thank Michael Hynes, Ivan Oresnik, Marie Elliot, and Robin Cameron for kindly providing strains, and Jianping Xu for assistance in isolating and identifying the *Aspergillus* species. I also thank Jianping Xu and Richard Morton for comments. For the work in Chapter 3, I thank David Romero for kindly providing plasmid pTE3Yp028, and Harsh Sharthiya for assistance in testing the Flp/FRT cloning method, and construction of *S. meliloti* RmP3277 and plasmid pTH2914. For the work in Chapter 5, I thank Arlene Sutherland and the students in the Molecular Biology 3V03 course at McMaster University for assistance in isolating many of the mutants used

in that study. I also acknowledge NSERC for providing me with funding through the Alexander Graham Bell Canada Graduate Scholarship and the Michael Smith Foreign Study Supplement programs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND SYMBOLS

~ – Approximately
> – Greater than
< – Less than
°C – Degrees Celsius
Δ – deletion (delta)
µg – Microgram
µL – Microliter
µM – Micromolar
A_{880} – Absorbance at 880 nm
ABC – ATP binding cassette
ala-gly – Alanine-glycine
AMP – Adenosine monophosphate
ATP – Adenosine triphosphate
AV – Activity index
bp – Base pair
BNF – Biological nitrogen fixation
CFU – Colony forming units
Cm – Chloramphenicol
cm – centimeter
COG – Cluster of orthologous genes
ddH_2O – Deionized, distilled water
DHAP – Dihydroxyacetone phosphate
DNA – Deoxyribonucleic acid
E.g. – For example
ED – Entner-Duodoroff
EMP – Embden-Meyerhof-Parnas
ETR – *engA*-tRNA-*rmlC*
FBA – Flux balance analysis
FRT – Flippase recognition target
FVA – Flux variability analysis
g – Gravity
GABA – γ-aminobutyric acid
GC – Guanine-cytosine
GLY – Glycolytic
gly-asp – Glycine-aspartate
gly-glu – Glycine-glutamate
Gm – Gentamicin
gm – Gram
GNG – Gluconeogenic
gly-asp – Glycine-aspartate
GPR – Gene-protein-reaction relationship
HGT – Horizontal gene transfer
HMW – High molecular weight
hr – Hour
I.e. – That is
kb – Kilobase
kg – Kilogram
kGy – Kilogray
Km – Kanamycin
L – Liter
LB – Lysogeny broth
LBmc – LB with $MgSO_4$ and $CaCl_2$
LCA – Lignocellulose agar
LMW – Low molecular weight
LPS – Lipopolysaccharide
Mb – Megabase
mL – Milliliter
mM – Millimolar
mmol – Millimole
MOPS – 4-morpholinepropanesulfonic acid
mRNA – Messenger RNA
N/A – Not applicable
NAG – N-acetylglutamate
NCBI – National Center for Biotechnology information
Nm – Neomycin
nM – Nanomolar
NSERC – Natural Sciences and Engineering Research Council
NT – Not tested
nt – Nucleotide
OD_{600} – optical density measured at a wavelength of 600 nm
PBS – Phosphate-buffered saline
PCA – Protocatechuic acid
PCR – Polymerase chain reaction

PHB – Poly-hydroxybutyrate
PPP – Pentose phosphate pathway
psig – Pounds per square inch gage
Rif – Rifampicin
rRNA – Ribosomal RNA
RNA – Ribonucleic acid
RNA-seq – RNA sequencing
rRNA – Ribosomal RNA
SBML – Systems biology markup language
SD – Standard deviation
Sm – Streptomycin
SNF – Symbiotic nitrogen fixation
Sp – Spectinomycin
Tc – Tetracycline
TCA – Tricarboxylic acid
tRNA – Transfer RNA
TY – Tryptone-yeast
vol – Volume
Vs – Versus
wt – Weight
YPD – Yeast-peptone-dextrose

DECLARATION OF ACADEMIC ACHIEVEMENT

I declare that I (George Colin diCenzo) have prepared this thesis, and that is has not been submitted for a previous degree application. I performed the work that is described in this thesis, and contributions from all other contributors are clearly acknowledge in "authors' contributions and justification for inclusion" section. All sources of information used in the preparation of this thesis are appropriately identified.

RESEARCH CONTRIBUTIONS

*Equal authorship.

14. diCenzo, G.C.*, Zamani, M.*, Ludwig, H. & Finan, T.M. (In press). Heterologous complementation reveals a specialized activity for BacA in the *Medicago* – *Sinorhizobium meliloti* symbiosis. *Mol Plant Microbe Interact*.

13. diCenzo, G.C. & Finan, T.M. (In press). Techniques for large-scale bacterial genome manipulation and characterization of the mutants with respect to *in silico* metabolic reconstructions. In *Metabolic Network Reconstruction and Modelling*. Edited by M. Fondi. Springer.

12. Zamani, M., diCenzo, G.C., Milunovic, B. & Finan, T.M. (2017). A putative 3-hydroxyisobutyryl-CoA is required for efficient symbiotic nitrogen fixation in *Sinorhizobium meliloti* and *Sinorhizobium fredii* NGR234. *Environ Microbiol* **19**, 218-236.

11. Zhang, Y., Smallbone, L.A., diCenzo, G.C., Morton, R. & Finan, T.M. (2016). Loss of malic enzymes leads to metabolic imbalance and altered levels of trehalose and putrescine in the bacterium *Sinorhizobium meliloti*. *BMC Microbiol* **16**, 163.

10. diCenzo, G.C., Checcucci, A., Bazzicalupo, M., Mengoni, A., Viti, C., Dziewit, L., Finan, T.M., Galardini, M. & Fondi, M. (2016). Metabolic modelling reveals the specialization of secondary replicons for niche adaptation. *Nat Commun* **7**, 12219.

9. Dzimitrowicz, A., Jamróz, P., diCenzo, G.C., Sergiel, I., Koźlecki, T. & Pohl, P. (2016). Preparation and characterization of gold nanoparticles using *Melissa officinalis*, *Salvia officinalis*, and *Mentha piperita* aqueous extracts. *Arabian J Chem* doi:

j.arabjc.2016.04.004.

8. diCenzo, G.C., Zamani, M., Milunovic, B. & Finan, T.M. (2016). Genomic resources for identification of the minimal N_2-fixing symbiotic genome. *Environ Microbiol* **18**, 2534-2547.

7. Chen, S.*, White, C.E.*, diCenzo, G.C., Zhang, Y., Stogios, P.J., Savchenko, A. & Finan, T.M. (2016). L-hydroxyproline and D-proline catabolism in *Sinorhizobium meliloti*. *J Bacteriol* **198**, 1171-1181.

6. Fei, F.*, diCenzo, G.C.*, Bowdish, D.M.E., McCarry, B.E. & Finan, T.M. (2016). Effects of synthetic large-scale genome reduction on metabolism and metabolic preferences in a nutritionally complex environment. *Metabolomics* **12**, 23.

5. diCenzo, G.C.*, Zamani, M.*, Cowie, A. & Finan, T.M. (2015). Proline auxotrophy in *Sinorhizobium meliloti* results in a plant-specific symbiotic phenotype. *Microbiology* **161**, 2341-2351.

4. diCenzo, G.C. & Finan, T.M. (2015). Genetic redundancy is prevalent within the 6.7 Mb *Sinorhizobium meliloti* genome. *Mol Genet Genomics* **290**, 1345-1356.

3. diCenzo, G.C., MacLean, A.M., Milunovic, B., Golding, G.B. & Finan, T.M. (2014). Examination of prokaryotic multipartite genome evolution through experimental genome reduction. *PLOS Genet* **10**, e1004742.

2. Milunovic, B., diCenzo, G.C., Morton, R.A. & Finan, T.M. (2014). Cell growth inhibition upon deletion of four toxin-antitoxin loci from the megaplasmids of *Sinorhizobium meliloti*. *J Bacteriol* **196**, 811-824.

1. diCenzo, G.*, Milunovic, B.*, Cheng, J. & Finan, T.M. (2013). $tRNA^{arg}$ and *engA* are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. *J Bacteriol* **195**, 202-212.

AUTHORS' CONTRIBUTIONS AND JUSTIFICATION FOR INCLUSION

Each of the research chapters presented in this thesis consist of published articles that I have published as the first author. Below, the justification for inclusion of these works in the thesis is outlined, as are my contributions and of all co-authors.

Chapter 2

This chapter describes the construction of a set of strains consisting of wild type *S. meliloti* and the ΔpSymA, ΔpSymB, and ΔpSymAB replicon cured derivatives that either directly or indirectly formed the basis of the majority of the work described in this thesis. In this chapter, these strains were phenotypically characterized, specifically examining the growth rate, metabolic capacity, and competitive fitness phenotypes. These data presented in this chapter were interpreted in light of a generalized theory of multipartite genome evolution and biology, and provided the first experimental evidence supporting this theory. The model described in this chapter serves as the prime overarching theme of this thesis.

The majority of the experiments were conceived and designed by George diCenzo and Turlough Finan. The exception was the development of the soil mesocosms and the testing of the pSymB deletion library for growth in the soil mesocosms, which was conceived and designed by Allyson MacLean and Turlough Finan. Most experiments were performed by George diCenzo. However, the growth of the pSymB deletion library in the soil mesocosms was performed by Allyson MacLean, and Branislava Milunovic

contributed to the identification of the siderophore as the causative agent of the wild type's inhibition of the growth of the ΔpSymA strain. Allyson MacLean and Turlough Finan analyzed the data obtained from the growth of the pSymB deletion library in the soil mesocosms, Brian Golding analyzed the Illumina sequencing results, and all other data was analysed by George diCenzo and Turlough Finan. The manuscript was drafted by George diCenzo, and George diCenzo, Allyson MacLean, Brian Golding, and Turlough Finan revised the manuscript. All work performed by George diCenzo was done between late 2011 and 2014, with writing of the manuscript performed between late 2013 and 2014.

Chapter 3

This chapter details the construction of an *in silico* genome-scale metabolic reconstruction of *S. meliloti*, and the use of this metabolic model combined with flux balance analysis to gain study the metabolic reprogramming occurring during niche adaptation and the role of each replicon in colonizing each environment. The ability to simulate the phenotypic consequences of deleting single genes and gene pairs in different nutritional conditions, which is not feasible with an experimental genetics approach, allowed for the testing of the contribution of genes on each replicon to growth in nutritionally diverse environments. Hence, this work was included in this thesis as it provided previously missing data that supports the multipartite genome evolution model described in chapter 2. Specifically, it provided support for a specialized role of a secondary replicon in adaptation of the cell to a particular niche.

The study was conceived by George diCenzo, Marco Fondi, Alessio Mengoni, Marco Galardini, and Turlough Finan. The genome-scale metabolic reconstruction was predominately performed by George diCenzo with guidance from Marco Fondi and input from Lukasz Dziewit, while Marco Galardini and Alice Checcucci assisted with model validation. Flux balance analysis experiments were designed by George diCenzo, Marco Fondi, and Alessio Mengoni, and the flux balance analyses were scripted and performed by George diCenzo and Marco Fondi. Phenotype MicroArray experiments were designed by George diCenzo, Alice Checcucci, and Alessio Mengoni, and performed by George diCenzo, Alice Checcucci, and Carlo Viti. Pangenome and regulon analyses were designed by George diCenzo, Marco Galardini, Marco Fondi, and Alessio Mengoni, and performed by Marco Galardini. The manuscript was drafted by George diCenzo, Marco Fondi, and Alessio Mengoni, and all authors contributed to revising the manuscript. The work and writing was performed from 2015 to 2016.

Chapter 4

This chapter, in part, is an extension of chapter 2 and previous works that I have authored. This chapter describes the construction of new *S. meliloti* strain sets consisting of strains containing either both pSymA and pSymB, just pSymA, just pSymB, or neither pSymA nor pSymB. A major advantage of the new strain sets is that each set has a highly isogenic chromosomal background, which cannot be said for the strains used in chapter 2. Additionally, one of the strain sets contains additional genes integrated into the chromosome, which allows it to better facilitate studies aimed at the identification of the

minimal N_2-fixing symbiotic genome. A phylogenetic analysis of a region of the pSymB replicon provides insights into an evolutionary process through which a secondary replicon can become essential, supporting the model described in chapter 2. Additionally, the pSymA and pSymB deletion libraries were screened for symbiotic defects, with the results both providing insights into the evolutionary dynamics of each replicon, as well as representing the necessary first step in delineating the essential symbiotic genome. Thus, this chapter is included in this thesis as it contributes to understanding the evolutionary process through which multipartite genomes evolve, specifically how secondary replicons gain core functions.

Most experiments were designed by George diCenzo and Turlough Finan, with the exception of the screening of the pSymA and pSymB deletion libraries for symbiotic phenotypes with alfalfa, which was designed by George diCenzo, Maryam Zamani, and Turlough Finan. The screening of the deletion libraries for symbiotic phenotypes was performed primarily by George diCenzo and Maryam Zamani, with Branislava Milunovic contributing as well. All other experiments were performed by George diCenzo. Data analysis was performed by George diCenzo and Turlough M Finan. The manuscript was drafted by George diCenzo, and George diCenzo, Maryam Zamani, and Turlough Finan contributed to revising the manuscript. All work performed by George diCenzo was completed between 2013 and 2015, with writing of the manuscript undertaken in 2015.

Chapter 5

This chapter details an examination of functional redundancy between the *S.*

meliloti chromosome and the pSymA and pSymB replicons. The results suggested that greater than 10% of chromosomal genes are potentially functionally redundant with genes present on pSymA or pSymB. This reason for including this work in this thesis is because it provides insights into a mechanism through which secondary replicons can come to encode core functions, as well as how secondary replicons can influence the evolutionary trajectory of the primary chromosome. It thus sheds light on a different aspect of multipartite genome evolution.

The study was conceived and designed by George diCenzo and Turlough Finan. All experiments were performed by George diCenzo, and data analysis was performed by George diCenzo and Turlough Finan. The manuscript was drafted by George diCenzo, and George diCenzo and Turlough Finan revised the manuscript. The experiments detailed in this chapter were performed between 2013 and 2014, and writing of the manuscript was performed in 2014.

CHAPTER 1. INTRODUCTION

1.1 Bacterial genome organization

Autoradiography work from John Cairns in 1963 provided the first evidence that the *Escherichia coli* genome consists of a single circular chromosome (Cairns, 1963). Together with subsequent studies, this work led to the generally accepted view that all bacterial genomes consist of a single circular chromosome, possibly including some smaller, non-essential, circular plasmids. However, this view had begun to change within 20 years. The identification of the first linear plasmid in *Streptomyces* in 1979 (Hayakawa *et al.*, 1979) and the determination that the *Borrelia burgdorferi* chromosome is linear in 1989 (Baril *et al.*, 1989; Ferdows & Barbour, 1989) illustrated that bacterial DNA molecules need not be circular. Moreover, a *Sinorhizobium meliloti* plasmid with a molecular weight greater than 300×10^6 Da (~ 460 kilobases) that the authors termed a 'megaplasmid' was identified in 1981 (Rosenberg *et al.*, 1981), challenging the notion that nearly all the bacterial genome is located on the chromosome (Bonhoeffer & Messer, 1969). And finally in 1989, the report of a 'second chromosome' in *Rhodobacter sphaeroides* (Suwanto & Kaplan, 1989) illustrated the potential for essential cell functions to be encoded by multiple replicons within the bacterium. The recent explosion in complete genome sequencing has revealed that approximately 10% of bacterial genomes do not fit into the simple '*E. coli* rules' of genome organization and contain several large and potentially essential replicons of either linear or circular nature (see Section 10 and Harrison *et al.*, 2010). The genome architecture consisting of a chromosome plus one or more megaplasmids or chromids is referred to as a divided

genome or a multipartite genome.

1.2 Linear chromosomes

Of the 1,708 bacterial species with a fully sequenced and assembled genome available on NCBI Genome (accessed 21/03/2016), 92 species (5.4%) have at least one strain containing a linear chromosome, while 1,657 (97.0%) have at least one strain with a circular chromosome. The distribution of linear chromosomes with respect to bacterial phylogeny reveals two clear clusters (Figure 1.1). Linear chromosomes are a characteristic feature of all 12 *Borrelia* species, and they are a common but not universal feature of the *Streptomyces* genome. However, the nature of the linear chromosomes in these genera is different. As reviewed elsewhere (Volff & Altenbuchner, 2000), the termini of the *Borrelia* chromosomes consist of covalently closed hairpins, whereas the 5' ends of the termini of the *Streptomyces* linear chromosomes are bound to proteins.

Outside of the *Borrelia* and *Streptomyces*, linear chromosomes are interspersed among circular chromosomes throughout the phylogeny with no clear pattern observed (Figure 1.1). The lack of additional clusters may be due, in part, to the limited number of fully assembled genomes publically available for many genera. For example, the only representatives of the *Luteipulveratus* and *Kineococcus* genera contain a linear chromosome, and sequencing of additional representatives may, or may not, illustrate this to be a general characteristic of these genera. Yet in many cases, the presence of a linear chromosome appears to be a spurious event and this is highlighted by the 41 species, including *E. coli*, *Bacillus subtilis*, and *Streptococcus aureus*, where at least one strain has

a linear chromosome and at least one other has a circular chromosome. However, it is important to keep in mind that some, hopefully few, of the linear chromosomes may be mislabelled in the Genbank file; for example, synteny analysis suggests that the linear and circular chromosomes of *Agrobacterium tumefaciens* S33 are misannotated (not shown).

A spurious appearance of linear chromosomes is consistent with linear chromosomes being easily evolved from circular chromosomes with few genetic changes required. It has been suggested that linear plasmids evolved from linear phage genomes (Casjens, 1999; Chang *et al.*, 1996; Hinnebusch & Tilly, 1993; Volff & Altenbuchner, 2000), and that linear chromosomes evolved through recombination between a circular chromosome and linear plasmid (Casjens, 1999; Chen *et al.*, 2002; Volff & Altenbuchner, 2000). It also seems reasonable that a linear chromosome could evolve directly through recombination between a circular chromosome and a linear phage genome. In support of these hypotheses, (Cui *et al.*, 2007) demonstrated the ease through which linear chromosomes can arise by linearizing the *E. coli* chromosome simply by addition of the *E. coli* phage N15 *tos* site and expression of the N15 *telN* gene. When linearized such that the *oriV* remained in the center of the chromosome, no phenotypic and few transcriptional changes were observed (Cui *et al.*, 2007). However, growth and morphological defects were observed if the *oriV* was off-centered (Cui *et al.*, 2007). Thus, it may be that linear chromosomes arise easily and potentially even often in bacterial evolution, but are generally selected against due to the linearization resulting in the *oriV* being off-centered.

The advantage of, and potential reasons for, maintaining a linear chromosome have been reviewed in depth elsewhere (Galperin, 2007; Hinnebusch & Tilly, 1993; Volff & Altenbuchner, 2000). One suggestion was that linear chromosomes are capable of carrying a greater amount of DNA (Galperin, 2007). However, the identification of equally large circular chromosomes suggested this may not be true (Galperin, 2007). Analysis of the 1,708 bacterial species with a completed genome in NCBI Genbank revealed the average size of linear bacterial chromosomes (~ 4.36 Mb) is larger than the average size of circular bacterial chromosomes (~ 3.61 Mb), although there is little difference in the median (~ 3.55 Mb for linear versus ~3.45 Mb for circular) (Figure 1.2A). However, the size of linear chromosomes is not normally distributed due to their enrichment in the *Streptomyces* (large chromosomes) and *Borrelia* (small chromosomes) (Figure 1.2A). Moreover, the largest fully sequenced and assembled linear and circular chromosomes are close in size, ~12.5 Mb and ~ 14.8 Mb, respectively, suggesting that both organizations are equally capable of carrying large amounts of DNA and therefore not a major selective force behind the evolution of linear chromosomes (Galperin, 2007).

1.3 Bacterial replicon classification

There are several terms to describe the DNA molecules in a multipartite genome.

1.3.1 Replicon and secondary replicon

'Replicon' is a general term used in reference to any DNA molecule regardless of its specific nature, and replicons are further classified based on their specific characteristics as described below. Using the classification scheme proposed by

(Harrison *et al.*, 2010), replicons can be characterized as 'chromosomes', 'plasmids', or 'chromids'. However, as described below, I propose that 'second chromosome' continue to be used as well in specific cases. The term 'secondary replicon' refers to any replicon that is not the primary chromosome.

1.3.2 Chromosome

Chromosome refers to the primary replicon; based on the analysis by Harrison *et al*. (2010), it is always the largest replicon in the genome and contains the vast majority of the core/essential genes. There is nearly a 100-fold size difference between the smallest and largest fully sequenced and assembled bacterial chromosomes, with an average and median size of ~ 3.65 Mb and ~ 3.46 Mb, respectively (Figure 1.2B). Given that the average and median bacterial total genome sizes are ~ 3.87 Mb and ~ 3.65 Mb, respectively (Figure 1.3B), it is clear that the chromosome generally accounts for the large majority of the genomic content of a cell. Indeed, 1,017 (59.5%) of the 1,708 bacterial species with a finished genome available on NCBI Genome contain only a chromosome with no secondary replicons (chromosome, megaplasmid, plasmid), while only 192 (~ 11%) have a chromid and/or megaplasmid (Figure 1.3).

1.3.3 Plasmid and megaplasmid

Plasmids refer to horizontally transferred and non-essential replicons whose genomic signatures, i.e. GC content and dinucleotide composition, differ significantly from the chromosome (Harrison *et al.*, 2010). A subdivision of plasmids are the 'megaplasmids', whose defining feature is size. While any size limitation is essentially

arbitrary, here I suggest a lower cut-off of 350 kb for megaplasmid status as this is equal to ~ 10% of the median bacterial genome. When using this boundary to distinguish megaplasmids from plasmids, the average and median plasmid sizes are ~ 78.9 kb and ~ 46.2 kb, while the average and median megaplasmid sizes are ~ 772 kb and ~ 558 kb. It is interesting that the size distribution of neither the plasmids nor megaplasmids follow a normal distribution but are instead positively skewed (Figures 1.2C, 1.2D)

The largest megaplasmid that I identified in a finished genome was ~ 3.93 Mb, which is larger than the average bacterial genome. This megaplasmid is present in *Burkholderia gladioli* ATCC 10248, and accounted for ~ 44% of the entire genome. However, it is worth pointing out that my classification of a replicon as a megaplasmid instead of a chromid (described in the legend of Figure 1.3) did not take into account potential essential genes on the replicon, and thus it is possible that this replicon should actually be classified as a chromid. To my knowledge, the largest replicon experimentally demonstrated to be non-essential is the ~ 1.35 Mb pSymA megaplasmid of *S. meliloti* Rm2011 (Oresnik *et al.*, 2000).

1.3.4 Chromid

The term chromid itself is a combination of chromosome and plasmid (Harrison *et al.*, 2010), and underscores how chromid refers to a replicon intermediary between a plasmid and a chromosome (Harrison *et al.*, 2010). Chromids have replication systems similar to those of bacterial plasmids, but unlike plasmids, they have at least one gene essential for cell viability and are expected to have genomic signatures similar to the

chromosome (Harrison *et al.*, 2010). It is worth noting that the majority of replicons currently classified as chromids are based solely on bioinformatics support; e.g., an essential gene is expected to be located on the replicon based on homology. However, such expectations are not always correct (Cheng *et al.*, 2007), and so it is important to experimentally validate the essential nature of more chromids to get a better idea of the prevalence of this replicon type. Simply being unable to remove the replicon from the genome should not be considered sufficient to confirm essentiality as other explanations exists, such as the presence of plasmid addiction systems (Gerdes *et al.*, 2005; Hayes, 2003; Van Melderen & Saavedra De Bast, 2009). For example, the pSymA megaplasmid of *S. meliloti* harbours at least three active toxin-antitoxin loci (Milunovic *et al.*, 2014). Despite being a non-essential replicon (Oresnik *et al.*, 2000), it is nearly impossible to chase pSymA from the cell [see for example (MacLellan *et al.*, 2005)] unless the three antitoxin loci are present *in trans* (GC diCenzo, B Milunovic, TM Finan, unpublished).

It has also been suggested that chromids be sub-divided into 'primary' and 'secondary' chromids (Dziewit *et al.*, 2014). In this classification, primary chromids refer to chromids absolutely essential for cell viability. In contrast, secondary chromids refer to chromids that are not essential in the laboratory, but are expected to be essential in the species natural habitat. While a potentially useful classification, secondary chromids would be more difficult to confirm experimentally, and I would add that a secondary chromid should be essential in the organism's primary niche. For example, a secondary replicon in a hypothetical soil bacterium that is also an opportunistic human pathogen

should carry genes essential for growth in the soil to be considered a secondary chromid and not just essential for survival in the human host. For the sake of this review, primary and secondary chromids will not be differentiated.

Using the criteria as defined in the legend of Figure 1.3, putative chromids were identified in 117 (6.9%) of the species with finished genomes available through NCBI Genbank (Figure 1.3). The large majority (91%) of species containing a chromid had a single chromid. At most, five putative chromids were detected in a single species, and all three species containing five chromids belong to the genus *Azospirillum*. The average (~ 1.52 Mb) and median (~ 1.26 Mb) chromid sizes are around two-fold larger than the average and median megaplasmid sizes (~ 0.77 Mb and ~ 0.56 Mb, respectively) (Figure 1.2D). Yet, the maximum chromid size is approximately equal to the largest megaplasmid, ~ 3.90 Mb and ~ 3.93 Mb, respectively. Like the megaplasmids and plasmids, the sizes of the chromids do not follow a normal distribution, which is in contrast to the bell-shaped distribution of bacterial chromosomes (Figures 1.2B, 1.2D).

1.3.5 Second chromosome

The term 'second chromosome' or 'secondary chromosome' has been used in the description of bacterial replicons and its use was roughly equivalent to the term chromid. As nicely described by (Harrison *et al.*, 2010), chromid should almost always be used in place of second chromosome as such replicons have consistently different properties than true bacterial chromosomes and appear to have evolved from megaplasmids and not from chromosomes (Harrison *et al.*, 2010). However, I propose 'second chromosome'

continue to be used to describe the rare example of a secondary replicon formed through the split of an ancestral replicon into two. This appears to be a rare event, yet I was able to detect at least two examples of this in the finished genomes available through NCBI. The *Salmonella enterica* strain NCTC10384 contains an ~ 3.75 Mb chromosome and an ~ 0.73 Mb chromid, while the other 160 *S. enterica* strains lack a chromid and have a chromosome averaging ~ 4.75 Mb. Dot plot analyses between the *S. enterica* NCTC10384 chromid with the *S. enterica* SC-B67 chromosome reveal near complete synteny (Figure 1.4A), while the corresponding region was absent from the *S. enterica* NCTC10384 chromosome (Figure 1.4B). Using similar logic, a second chromosome was also identified in *Nocardia farcinica* NCTC11134 (not shown). To the best of my knowledge, and assuming these are not errors in the genome assemblies, these are the only example of a chromosomal split resulting in two replicons in the literature.

1.4 Proposed mechanisms of chromid formation

Two major hypotheses have been put forth describing the mechanism through which a chromid is formed: the schism hypothesis and the plasmid hypothesis (Choudhary *et al.*, 2012; Egan *et al.*, 2005; Moreno, 1998; Prozorov, 2008).

1.4.1 The schism hypothesis

The schism hypothesis states that a split of an ancestral chromosome resulted in the formation of the chromosome and chromid. If true, it would be expected that the properties of both the chromosome and chromid are highly similar, and that there would be an equal distribution of core genes between both replicons. The ability to produce

viable *E. coli* or *B. subtilis* strains that have had their single chromosome artificially split into two self-replicating chromosomes provides support that such a scenario is possible (Itaya & Tanaka, 1997; Liang *et al.*, 2013). Although I was able to detect two examples of potential secondary chromosome formation by chromosome schism in the NCBI database (see section 1.3.5), there is little evidence to support the schism hypothesis as a dominant method of chromid formation in nature. That said, the schism hypothesis has been put forth to explain the chromid formation of *Brucella suis* (Jumas-Bilak *et al.*, 1998) and *Rhodobacter sphaeroides* (Choudhary *et al.*, 1997), but as described below, this seems unlikely.

Three of four *B. suis* biovars contain a chromosome and chromid, while biovar 3 has a single chromosome equal to the combined size of the chromosome and chromid of the other biovars (Jumas-Bilak *et al.*, 1998). It was originally suggested that the common ancestor of these strains had a single chromosome partly on the basis that the majority of the α-proteobacteria have a single chromosome (Jumas-Bilak *et al.*, 1998). However, phylogenetic evidence is inconsistent with this hypothesis as *B. suis* is not a deep branching lineage within the *Brucella*, and thus it is more likely the *Brucella* ancestor had a chromosome and a chromid and that the one chromosome of *B. suis* biovar 3 is the result of a fusion of these replicons (Moreno, 1998).

In the case of *R. sphaeroides*, it was suggested that the chromid resulted from a split of the ancestral chromosome due to the many genomic features being highly similar between the chromosome and chromid (Choudhary *et al.*, 1997), the large number of gene

duplications between these replicons (Choudhary *et al.*, 2004), and a large number of genes on the chromid predicted to be essential (Choudhary *et al.*, 1994; Mackenzie *et al.*, 2001). However, several observations are inconsistent with a common origin of the chromosome and chromid. These include differences in the trinucleotide composition (Choudhary *et al.*, 1997), the lower coding density of the chromid relative to the chromosome (Mackenzie *et al.*, 2001), and gene functional biases as determined through Cluster of Orthologous Genes (COG) analyses between the chromosome and chromid (Mackenzie *et al.*, 2001). Additionally, if the chromid and chromosome evolved from the same ancestral replicon then both should show synteny with the chromosome of a closely related species. While the *R. sphaeroides* 2.4.1 chromosome shows stretches of synteny with the chromosome of *Ruegeria pomeroyi* DSS-3 (Mackenzie *et al.*, 2007), this is not observed for the *R. sphaeroides* chromid (not shown). Thus, there is little evidence for chromid formation through the schism hypothesis in nature except perhaps very rarely.

1.4.2 The plasmid hypothesis

This hypothesis states that chromids have evolved from megaplasmids, and this hypothesis predicts that the continued existence of a megaplasmid with a chromosome will result in regression of the megaplasmid's genomic signatures to that of the chromosome and the gain of essential genes potentially through transfer from the chromosome. The replication and partitioning machinery of chromids resembles that of megaplasmids, although in many cases experimental evidence for the functionality of these systems is lacking (Egan *et al.*, 2005). That said, replicons of the *repABC* family

that carry essential genes, and are thus chromids, have *repABC* replication/partitioning genes that have a codon usage more similar to that of the chromosome than do *repABC* family members that do not carry essential genes (Castillo-Ramírez *et al.*, 2009), consistent with these replication systems being functional. Additionally, a phylogenetic analysis of the plasmid partitioning protein RepA was consistent with chromids evolving from pre-existing megaplasmids in the α-proteobacteria (Harrison *et al.*, 2010).

Two main observations about chromids can be explained by the plasmid hypothesis that cannot be accounted for by the schism hypothesis. One, although chromids do carry core/essential genes, the percentage of core/essential genes on a chromid is much lower than the percentage of core/essential genes on the chromosome of the same species (Harrison *et al.*, 2010). If the chromid and chromosome resulted from a split of an ancestral replicon (schism hypothesis), it would be expected that the ratio of core/essential genes on the chromosome/chromid would be equal to the ratio of the sizes of these replicons. However, fewer essential core/essential genes are predicted to be on the chromid if the presence of such genes resulted from translocations from the chromosome to a megaplasmid (plasmid hypothesis). Two, there is also a consistently observed bias in the functional annotation of chromid versus chromosome genes as determined via COG analyses (see for example Chain *et al.*, 2006; Goodner *et al.*, 2001; Heidelberg *et al.*, 2000; Mackenzie *et al.*, 2001). If the chromosome and chromid have a common source (schism hypothesis), such biases are unexpected. In contrast, functional biases are expected to be present if the two replicons have different evolutionary histories

(plasmid hypothesis). Hence, *in toto* it appears as though the plasmid hypothesis is likely to explain the formation of most, if not all, chromids studied to date.

1.4.3 Conversion of a megaplasmid to chromid

For a chromid to evolve from an existing megaplasmid, two main conversions must take place: amelioration of the genomic signatures to that of the chromosome and the acquisition of core/essential genes. Genomic signatures such as codon usage and dinucleotide composition are shaped by a variety of factors and can have adaptive advantages (Carbone *et al.*, 2003; Wong *et al.*, 2002). Therefore, the similarity in the genomic signatures between chromosomes and chromids in the same species is presumably a result of evolutionary forces selecting for optimized genome function and can be caused by, for example, selection for improved translational efficiency or mutational biases of the cellular machinery (Carbone *et al.*, 2003; Wong *et al.*, 2002).

The occurrence of essential genes on a secondary replicon is somewhat more surprising than the similarity of the genomic signatures of the chromosome and chromid; if the cell could survive without the chromid in the past, why has this replicon now become an essential component of the genome? I can envisage two mechanisms through which this could occur. The primary method would be through gene transfer from the chromosome to the chromid, which is supported by several studies. Perhaps the clearest example of a gene transfer event conferring essentiality to a secondary replicon is in *S. meliloti* (diCenzo *et al.*, 2013; 2016b). Two essential genes have been experimentally demonstrated to exist on the *S. meliloti* pSymB chromid, *engA* and an arginine tRNA

(diCenzo *et al.*, 2013). Computational analysis of the region around these essential genes demonstrated that their presence on pSymB is the result of an inter-replicon translocation event that transferred a contiguous 69 kilobase fragment, including *engA* and the tRNA, from the chromosome to the chromid in a *S. meliloti* ancestor (diCenzo *et al.*, 2013; 2016b; see also Chapter 4). Additionally, the four putatively essential genes present in two clusters on the *Vibrio cholerae* chromid are present on the chromosome of related *Vibrio* species (Egan *et al.*, 2005), while numerous other clusters of *Vibrio* genes, as well as genes in the *Burkholderia* and the *Rhizobiales* order, are predicted to have moved from the chromosome to a secondary replicon (Slater *et al.*, 2009). Similarly, 25-30% of genes on the *S. meliloti* chromid have orthologs on the *A. tumefaciens* chromosome, suggestive of significant amounts of inter-replicon gene flow (Wong & Golding, 2003). The mechanism through which gene transfer from the chromosome to a secondary replicon occurs hasn't been studied. However, considering that the three replicons of *S. meliloti* may naturally form co-integrants (Guo *et al.*, 2003), it may be that replicon integration followed by imprecise excision results in inter-replicon gene transfer (Ng *et al.*, 1998).

The second putative way through which secondary replicons may come to carry core genes is through genetic redundancy. It was experimentally shown through transposon mutagenesis that numerous (potentially > 10%) of chromosomal genes in *S. meliloti* have a functionally redundant copy on one of the secondary replicons (diCenzo & Finan, 2015; see Chapter 5). Based on sequence similarity, there also appear to be many gene duplications between the chromosome and secondary replicons of *R. sphaeroides*

(Bavishi *et al.*, 2010b; Choudhary *et al.*, 2004), *V. cholerae* (Heidelberg *et al.*, 2000), and *Burkholderia vietnamiensis* (Maida *et al.*, 2014). Genetic redundancy between core genes on the chromosome and chromid could be a result of an inter-replicon duplication of a chromosome gene, or through acquisition of a homologous gene through horizontal gene transfer. In cases where the copy of the gene on the secondary replicon is able to fully complement disruption of the chromosomal version, degeneration of the chromosomal copy would be fitness neutral and the version of the gene on the secondary replicon would become the sole copy, in effect transferring a core gene to the secondary replicon.

1.5 Phylogenetic distribution of multipartite genomes

Large scale analysis of multipartite genomes has largely focused on the distribution of chromids with little attention paid to the distribution of megaplasmids, the evolutionary precursor of chromids. In 2010, (Harrison *et al.*, 2010) reported that ~ 10% of the 1,085 finished bacterial genomes contained a chromid. Organisms containing a chromid are enriched in the proteobacteria including members of the α-, β-, and γ-proteobacteria, but chromids can also be detected in phylogenetically distant genera, including among others the genera *Prevotella*, *Leptospira*, and *Deinococcus* (Harrison *et al.*, 2010; Landeta *et al.*, 2011). The number of finished genomes sequences has drastically increased since 2010, with over 4,500 finished genomes representing 1,708 bacterial species now available through NCBI. Therefore, this dataset was analyzed with respect to the distribution of megaplasmids and chromids (Figure 1.3).

Chromids and megaplasmids were annotated in all finished genomes available

through NCBI as described in the Figure 1.3 legend. Based on these definitions, 3.5% of all the finished genomes contain at least one megaplasmid while 7.4% contain at least one chromid. Of the 1,708 species with a finished genome available, 6.4% include at least one strain with at least one megaplasmid, 7.4% include at least one strain with at least one chromid, and overall 11.2% include at least one strain with a multipartite genome (megaplasmid and/or chromid). Perhaps surprisingly given the frequency of chromids and megaplasmids, 2.5% of the bacterial species examined have both a chromid or megaplasmid (not necessarily in the same strain), and the majority of these are in the *Burkholderia* genus and the *Rhizobiales* order (Figure 1.3). It is also intriguing to observe that chromids are more prevalent than megaplasmids in the bacterial phylogeny (7.4% of species versus 6.4%). Accepting that chromids evolve from existing megaplasmids, it might have been expected that megaplasmids would be more common than chromids. That this does not appear to be true may reflect a 'do or die' nature of megaplasmids; if the megaplasmid is advantageous it may rapidly evolve into a chromid, whereas megaplasmids lacking strong advantages may be quickly lost from the population.

Looking at the phylogenetic distribution of megaplasmids and chromids (Figure 1.3), both can be seen dispersed throughout the phylogeny but clear clusters emerge. Megaplasmids were observed to be common in the *Bacillus*, *Rhodococcus*, *Novosphingobium*, *Burkholderia*, *Sinorhizobium*, *Rhizobium*, *Mesorhizobium*, *Agrobacterium*, and *Methylobacterium*. Chromids were enriched in the *Sinorhizobium*, *Rhizobium*, *Agrobacterium*, *Burkholderia*, *Cupriavidus*, *Ralstonia*, *Deinococcus*, *Vibrio*,

Pseudoaltermonas, and *Prevotella*. It should be noted that the lack of additional clusters may simply reflect an under-representation of the taxa among the sequenced genomes; for example, the *Leptospira* were seen to have chromids, but the lack of genome sequences from the genus precluded a clear cluster being observable in the phylogeny.

It was previously noted that chromids appear to contain genus specific genes and the presence of a chromid may correspond the emergence of a new genus (Harrison *et al.*, 2010). This observation remains largely true in my dataset, although some exceptions were detected where the presence of a chromid is not a defining characterizing of the genus. For example, *R. sphaeroides* contains a chromid whereas *Rhodobacter capsulatas* does not, *Xanthomonas sacchari* contains a putative chromid whereas the other seven *Xanthomonas* species do not, and not all *Deinococcus* species have a chromid. On the other hand, the acquisition of a chromid may also pre-date the emergence of a genus. For example, it has been argued that the chromid of the *Sinorhizobium*, *Rhizobium*, and *Agrobacterium* was acquired by the common ancestor of these genera prior to their divergence (Slater *et al.*, 2009). Nevertheless, it remains largely true that the presence of a chromid corresponds with a taxonomic split in the bacterial phylogeny. In contrast to chromids, megaplasmids are rarely conserved at the genus level, although multiple species in a genus will often contain a megaplasmid. Even in the rare cases where megaplasmids are present in all species of a genus, the different species may have unique megaplasmids. For example, all *Sinorhizobium* species have at least one strain with a megaplasmid, but analysis of the replication and partitioning proteins suggests they do

not share common ancestry (Österman *et al.*, 2014).

1.6 Characteristics of chromosomes, chromids, and megaplasmids.

Chromosomes, chromids, and megaplasmids show consistently different characteristics from each other. In the current section, some of these characteristics are examined for each replicon class, and the results compared and contrasted.

1.6.1 Genomic signatures

Previous work has illustrated that the genomic signatures of the chromosomes, chromids, and megaplasmids show distinct biases, and can thus be used as rough parameters in the classification of bacterial replicons. Biases in codon usage between each replicon in a genome has been detected; although the codon usage of both chromids and plasmids are distinct from that of the chromosome, the difference in codon usage between the chromosome and chromid is smaller than the difference between either the chromosome or chromid and the plasmid (Harrison *et al.*, 2010). For example, in *Burkholderia* species, codon usage bias was greatest on the chromosome, followed by the chromid, and lastly by the megaplasmid (Cooper *et al.*, 2010).

The GC content of each replicon has also been observed to differ, with the GC content of chromids within 1% that of the chromosome, while plasmids have a GC content differing by more than 1% (Harrison *et al.*, 2010). As GC content was used in defining chromids versus megaplasmids in my dataset, differences are expected to be observed. Nevertheless, the average GC content difference between the chromosome and chromid was 0.4% (SD 0.3%), chromosome versus megaplasmid was 2.4% (SD 2.0%),

and was 3.5% (SD 3.1%) for chromosome versus plasmid (Figure 1.5A). It has previously been noted that the GC contents of plasmids are almost always lower than that of the chromosome (Harrison *et al.*, 2010); however, exceptions do exists. One notable example is the megaplasmid of *Thermobaculum terrenum* ATCC BAA-798, which has a GC content 15.7% higher than the chromosome (63.8% versus 48.1%) (Kiss *et al.*, 2010).

The profiles of dinucleotide abundances in a genome, i.e. the frequency that each pair of nucleotides appear next to each other, has been shown to be distinct for each bacterial genome and can differentiate chromids from plasmids from chromosomes (Karlin & Burge, 1995; van Passel *et al.*, 2006; Wong *et al.*, 2002). The dinucleotide relative abundance distance refers to the sum of the differences in the frequency of each dinucleotide pair between two sources of DNA. As with GC content, dinucleotide relative abundance distances were used when defining chromid versus megaplasmid in this study, and so biases are expected when examining the dataset. Nevertheless, we saw that the difference between the chromosome and chromid was 0.21 (SD 0.06), chromosome versus megaplasmid was 0.55 (SD 0.28), and the difference between chromosome and plasmid was 0.98 (SD 0.50) (Figure 1.5B).

It is interesting to note that the dinucleotide relative abundance distances of chromids compared to chromosomes followed a normal distribution (Figure 1.5B), whereas the difference in GC content of chromids compared to chromosomes was positively skewed (Figure 1.5A). This perhaps suggests that whereas the GC content of chromids is continually ameliorated towards that of the chromosome, there is a constraint

on the amelioration of the dinucleotide composition. Whether this is simply a consequence of the starting difference between the two replicons or whether this reflects an adaptive function is unclear, but may reflect that there are different evolutionary forces acting on each replicon in a multipartite genome (Galardini *et al.*, 2013).

Another interesting observation was that both the difference in GC content and the dinucleotide relative abundance distance between chromosomes and megaplasmids were smaller than between chromosomes and plasmids despite both being non-essential replicons (Figure 1.5). This may suggest that the mobility of megaplasmids is lower than plasmids and that their transfer to phylogenetically distant organisms occurs less frequently than it does for plasmids. Although megaplasmids retain conjugative machinery (see for example He *et al.*, 2003; Pérez-Mendoza *et al.*, 2005; Romanchuk *et al.*, 2014; Yang *et al.*, 2007) and the transfer of megaplasmids between related organisms has been observed in nature (Brom *et al.*, 2000; 2002; Herrera-Cervera *et al.*, 1999; Young & Wexler, 1988; among others), it appears as though there may be limits to megaplasmid transfer to phylogenetically distant species (Romanchuk *et al.*, 2014). Another possibility is that the large size of megaplasmids necessitates an increased rate of amelioration to limit the costs of carrying the megaplasmid. Quite likely, a combination of these factors, and perhaps others, contributes to the observed bias. I will also note here that chromids can retain conjugative properties and can be induced to transfer to naïve cells under laboratory conditions (Blanca-Ordóñez *et al.*, 2010) but there is no evidence for the transfer of chromids in nature (Harrison *et al.*, 2010), and *A. tumefaciens* cells

containing the *S. meliloti* pSymB chromid display an obvious fitness decrease relative to wild type *A. tumefaciens* cells (Hynes *et al*., 1985; Finan *et al.*, 1986).

1.6.2 Genetic variability

The level of sequence and gene conservation between related bacterial strains and species is different for chromosomes, chromids, and megaplasmids. Bavishi *et al*. observed that in 7 of 9 species belonging to the genera *Brucella*, *Rhodobacter*, *Burkholderia*, and *Vibrio*, the level of nucleotide identity between the chromosomes of strains belonging to the same species was greater than the level of nucleotide identity between the chromids (Bavishi *et al.*, 2010a). Similarly, these authors observed the same pattern for 9 of 10 inter-genera comparisons of related species (Bavishi *et al.*, 2010a). Greater synteny between the chromosomes of *R. sphaeroides* strains has also been observed than between the chromids of this species (Choudhary *et al.*, 2007). In a population genomics study, Epstein *et al*. observed that similar percentages (~ 95%) of the chromosome and chromid of the reference *S. meliloti* and *Sinorhizobium medicae* genomes were conserved across 32 and 12 strains, respectively, while the conservation of the megaplasmid was much less (< 80%) (Epstein *et al.*, 2012). This finding was supported by a sequencing study where three competed *S. meliloti* genomes were aligned (Galardini *et al.*, 2011), but an earlier population genomics study found the megaplasmid genes were most variably present/absent, followed by the chromid genes, and finally by the chromosome genes (Guo *et al.*, 2009).

As an alternate method of examining replicon conservation, the number of

orthologs between strains/species has been examined. Cooper *et al*. observed a greater fraction of *Vibrio* chromosomal genes were conserved across species than the fraction of chromid genes that were conserved (Cooper *et al.*, 2010). Similarly, a greater percentage of chromosomal *Burkholderia* genes were conserved between species than for the chromid, with the megaplasmid genes being least conserved (Cooper *et al.*, 2010; Holden *et al.*, 2009). This pattern was also observed between strains of *Burkholderia cenocepacia* (Cooper *et al.*, 2010). Interestingly, the number of conserved chromosomal genes in the *B. cenocepacia* strains was similar to the number of conserved chromosomal genes across the *Burkholderia* species, whereas many fewer chromid or megaplasmid genes were conserved across species than between the *B. cenocepacia* strains. Comparison between the related *Burkholderia pseudomallei* and *Ralstonia solanacearum* species revealed greater orthology between chromosomal genes than between the chromid genes (Holden *et al.*, 2004), and similarly, comparison between *S. meliloti* and *S. medicae* detected highest ortholog content on the chromosome, then chromid, then megaplasmid (Epstein *et al.*, 2012). Comparing between *Cupriavidus* species showed greater orthology between chromosomal genes than between chromid genes, whereas comparison between strains belonging to the same *Cupriavidus* species showed only slightly greater ortholog content on the chromosome than the chromid (Janssen *et al.*, 2010; Van Houdt & Mergeay, 2012; Van Houdt *et al.*, 2012). Finally, in my own analysis of the 11 *Bacillus thuringensis* strains with a megaplasmid, 1,984 chromosomal genes were present in all 11 strain (5,153 genes were present in at least five strains), whereas no megaplasmid genes

were present in all strains and only 163 were present in at least five strains (not shown). All considered, these data suggest that chromosomes are the most genetically stable replicons, followed by chromids, and lastly by megaplasmids. Additionally, at the species level, chromids, but not megaplasmids, may be conserved nearly as strongly as chromosomes, but the level of conservation drops off at the genus level.

1.6.3 Functional biases

Functional analyses have repeatedly shown that each replicon in a genome contains functional biases. In *R. sphaeroides* 2.4.1, core processes are enriched on the chromosome, while inorganic ion and amino acid transport/metabolism are enriched on the chromid (Mackenzie *et al.*, 2001). Hypothetical genes and genes of unknown function are also enriched on the *R. sphaeroides* 2.4.1 chromosome, which is in contrast to *V. cholerae* El Tor N16961 where these categories are enriched on the chromid (Heidelberg *et al.*, 2000; Mackenzie *et al.*, 2001). Core functional classes are also enriched on the *V. cholerae* chromosome, while central intermediary metabolism, transport, and regulatory functions are enriched on the chromid (Heidelberg *et al.*, 2000). In *Burkholderia xenovorans* LB400, the core processes are again biased towards the chromosome, the chromid is enriched in transcription, carbohydrate transport and metabolism, and signal transduction, while the chromid and megaplasmid are biased towards energy and secondary metabolism, and inorganic and amino acid transport/metabolism (Chain *et al.*, 2006). In *S. meliloti*, the chromid and megaplasmid have a greater percentage of regulatory genes and genes without a database hit (orphans),

and the chromid is particularly enriched in transport systems (Galibert *et al.*, 2001). Functional biases are also detected between the replicons present in the multipartite genomes of *Paracoccus aminophilus* JCM 7686 (Dziewit *et al.*, 2014), *Cupriavidus metallidurans* CH34 (Janssen *et al.*, 2010), *B. suis* 1330 (Paulsen *et al.*, 2002), *Bacillus cereus* (Zheng *et al.*, 2015), and *Marinovum algicola* DG898 (Frank *et al.*, 2015). Additionally, functions related to metabolism, transcription, and signal transduction appear over-represented on chromids in the proteobacteria (Choudhary *et al.*, 2012).

I wished to examine whether any global biases between the functional annotation of different replicon classes could be observed, or whether the functional biases differed across the bacterial phylogeny. Therefore, a COG analysis (Galperin *et al.*, 2015) of all replicons from a representative genome from each of the 1,708 species with a finished genome available through NCBI was performed (Table 1.1). Indeed, several global biases could be observed and some notable observations are summarized here. Not surprisingly, the chromosome was enriched for core functions, with the largest enrichments in cytoskeleton (COG Z), RNA processing and modification (COG A) and translation (COG J). As expected, chromids were enriched in some core functions compared to megaplasmids. This was most evident in RNA processing and modification (COG A), chromatin structure and dynamics (COG B), motility (COG N), which may not be essential but presumably highly important, and to a lesser extent translation (COG J). Interestingly, there were few biases detected in the categories representing general function prediction only (COG R) or function unknown (COG S). This highlights that

our general functional understanding of bacterial genes is not biased towards any replicon class, and indeed, a third of the essential genes in a synthetic minimal genome based on *Mycoplasma mycoides* are of unknown function (Hutchison *et al.*, 2016).

In general, functions enriched in chromids and megaplasmids were similar, although the extent of enrichment on megaplasmids or chromids was often unique. In general, chromids showed a greater level of functional enrichment than did megaplasmids. Energy production and conversion (COG C), as well as amino acid, carbohydrate, and inorganic ion transport and metabolism (COGs E,G,P) were clearly enriched on chromids with a slight enrichment on megaplasmids, whereas lipid transport and metabolism (COG I), secondary metabolite metabolism (COG Q), transcription (COG K), and extracellular structures (COG W) were similarly enriched on both chromids and megaplasmids. The enrichment in transcription likely represents a greater number of transcription factors present on these replicons allowing gene regulation in response to numerous environmental signals (e.g. carbon availability). Additionally, cell motility (COG N) and signal transduction (COG T) are enriched on chromids but not megaplasmids. Finally, plasmids were enriched in replication related functions (COGs D and L), and intracellular trafficking, secretion, and vesicular transport (COG U), which may be related to resistance to toxic compounds, and the conjugal transfer of the plasmid. Thus, it is clear that global functional biases between each replicon class are present.

This global functional analysis provided some interesting insights into the evolution of multipartite genomes. It was notable that functional biases could be detected

between megaplasmids and plasmids, similar to how the genomic signatures of megaplasmids and plasmids differ (see section 1.6.1), consistent with megaplasmids representing a distinct class of replicons from that of plasmids. The local adaptation hypothesis states that plasmids allow adaptation to conditions/stresses that occur in a narrow time or space range (Eberhard, 1990). Thus, the functional biases between plasmids and megaplasmids may be consistent with plasmids providing a selective advantage to a very specific condition or stress (e.g. the presence of an antibiotic), whereas megaplasmids provide a somewhat more general advantage. Given that chromids are thought to have evolved from megaplasmids (Harrison *et al.*, 2010; section 1.4), it was a bit surprising the number of functional biases that were detected between the chromids and megaplasmids. This suggests that as a secondary replicon is maintained, it continues to acquire functions providing a selective advantage to the cell, but these functions are distinct from those originally on the megaplasmid. Overall, I interpret this functional data as consistent with a hypothesis based on views put forth previously (Chain *et al.*, 2006; diCenzo *et al.*, 2014). Specifically, chromids allow adaptation to a broad niche (e.g. the rhizosphere), megaplasmids improve fitness in a particular environment within that niche (e.g. the alfalfa rhizosphere) and/or allow the colonization of another unique environment (e.g. the legume root nodule), while plasmids provide adaptation to particular stresses/nutrients (e.g. a plant secreted antibiotic).

1.7 Inter-replicon interactions

Despite genes on each replicon in a multipartite genome being physically

separated, in many cases there may be interactions between their gene products. Enzymes involved in the biosynthesis of pantothenate and of lipopolysaccharide are encoded by multiple replicons in *Rhizobium etli* and *Rhizobium leguminosarum* (García-de los Santos & Brom, 1997; Villaseñor *et al.*, 2011). In the case of panthothenate, this may be due to gene transfer form the chromosome to the secondary replicon (Villaseñor *et al.*, 2011). Similarly, complex biological processes, such as the rhizobium – legume symbiosis, can require genes situated on different replicons (Brom *et al.*, 1992; Hynes & McGregor, 1990; Hynes *et al.*, 1986). Additionally, an *in silico* analysis of the seven replicons of *R. etli* predicted functional links between each of the replicons, with the most recently acquired replicons showing the least connections (González *et al.*, 2006).

Interactions between replicons can also occur at a regulatory level. Replication of the chromid of *V. cholerae* is subjected to regulation by mechanisms encoded by the chromosome (Heidelberg *et al.*, 2000). It has also been noted that in *V. cholerae*, the chromosomally encoded RpoS regulates genes on both the chromosome and chromid, *hylA* on the chromid is regulated by HylU encoded by the chromosome, and quorum-sensing genes are split between the chromosome and chromid (Heidelberg *et al.*, 2000). An *in silico* regulon analysis by Galardini *et al.* observed that most *S. meliloti* transcriptional factors included in their dataset regulated genes on the same replicon (Galardini *et al.*, 2015). However, a subset were predicted to regulate genes across multiple replicons, and a bias for chromosomal regulators to modulate chromid/megaplasmid genes was observed compared to the number of

chromid/megaplasmid genes predicted to regulate chromosomal genes (Galardini *et al.*, 2015). Finally, in unpublished data, we have observed that the complete deletion of the *S. meliloti* chromid results in at least a 2-fold change in gene expression in 5-10% of chromosomal genes, whereas no statistically significant changes in chromosomal gene expression were observed when the megaplasmid was absent from the genome (diCenzo GC, Golding GC, Finan TM, unpublished).

1.8 Costs associated with multipartite genomes

As will be described in Section 1.9, megaplasmids and chromids may provide certain advantages to the host cell. However, these large plasmids may come with a cost both to their native cell and to new hosts following horizontal transfer. Transfer of the chromid of *S. meliloti* to *A. tumefaciens* resulted in a lower growth rate, and the *A. tumefaciens* cells spontaneously lost the chromid (Finan *et al.*, 1986). Similarly, when the large megaplasmid of *Pseudomonas syringae* Pla107 is transferred to other pseudomonads, including other *P. syringae* strains, the presence of the megaplasmid decreased growth rate and the competitive fitness of the cell, and the megaplasmid could be spontaneously lost (Romanchuk *et al.*, 2014). It is interesting to note, however, that despite this megaplasmid being recently acquired by Pla107 (Baltrus *et al.*, 2011), it is difficult to construct a Pla107 derivative lacking the megaplasmid (Romanchuk *et al.*, 2014), suggestive of rapid adaptation to accommodate this megaplasmid. Additionally, *S. meliloti* Rm2011 strains that are cured of the megaplasmid appear to show a slightly faster growth rate than wild type *S. meliloti* Rm2011 (diCenzo *et al.*, 2014; 2016b). In

competition experiments, an *A. tumefaciens* C58 strains lacking the At plasmid is able to outcompete the wild type under laboratory conditions (Morton *et al.*, 2013).

The studies summarized in the previous paragraph illustrate how large secondary replicons can impair the fitness of the host cell in particular environments; however, presumably in at least one of the natural niches of the species the benefits of the megaplasmid/chromid outweigh the costs. Why exactly these fitness costs are observed is unclear, although several suggestions have been put forth. Streamlining theory suggests the loss of the replicon could be favoured as it reduces the amount of phosphorus tied up in DNA (Hessen *et al.*, 2010), although others have argued there is little support for this hypothesis (Vieira-Silva *et al.*, 2010). Alternatively, loss of the replicon could be favoured by reducing the energetic demands associated with DNA replication and/or gene expression (transcription and translation), particularly expression of ABC transport systems (Morton *et al.*, 2013; Romanchuk *et al.*, 2014). Decreasing the number of transcripts of non-essential proteins could also free up ribosomes for translation of core proteins, and/or decreasing the number of recently acquired genes whose gene products may be misfolded could also promote loss of a secondary replicon (Romanchuk *et al.*, 2014). Finally, negative interactions between pathways encoded by the chromosome and secondary replicon could promote loss of the secondary replicon, as could negative interactions between these replicons at a transcriptional level (Morton *et al.*, 2013; Romanchuk *et al.*, 2014). With respect to this last point, we have observed that loss of the *S. meliloti* megaplasmid has no statistically significant effects on chromosomal gene

expression (diCenzo GC, Golding GB, Finan TM, unpublished). Likely, a combination of factors explain why secondary replicons confer fitness costs to the host.

1.9 Suggested advantages of multipartite genomes

Several hypotheses have been suggested describing why bacterial multipartite genomes emerged and are maintained. In this section, suggested advantages potentially associated with multipartite genomes are discussed.

1.9.1 Larger genome

It has been suggested that multipartite genomes allow for further genome expansion once the chromosome has reached its maximal size (Slater *et al.*, 2009). In support of this, it was noted that as of 2010, the mean total genome size of genomes lacking a chromid was 3.38 (SD 1.81) Mb, whereas the mean size for genomes with a chromid was 5.73 (SD 1.66) Mb (Harrison *et al.*, 2010). In contrast, it was previously pointed out that some small genomes, like *Brucella melitensis* are multipartite whereas some large genomes like *Myxococcus xanthus* have a single chromosome (Egan *et al.*, 2005). Based on the finished genomes available through NCBI, the median size of genomes containing a megaplasmid and/or a chromid was ~ 2 Mb larger than those lacking such replicons (5.56 Mb versus 3.41 Mb; Figure 1.6A). The difference in genome size between those with and without megaplasmids/chromids was primarily associated with the presence of the secondary replicon and not differences in chromosomal size as the median chromosome size for species with megaplasmids/chromids was 3.38 Mb compared to 3.65 Mb for genomes with megaplasmids/chromids (Figure 1.6B). This

supports that multipartite genome are, on average, larger than genomes lacking large secondary replicons. However, the presence of a multipartite genome is clearly not a pre-requisite for a large genome, and in fact, the majority of large genomes lack chromids/megaplasmids (Table 1.2). Hence, it seems unlikely that multipartite genomes are simply to allow increased gene accumulation as the majority of large genomes are not multipartite. Additionally, causality has not been demonstrated; i.e., it has not been established whether genomes are multipartite to allow increased size, or whether the increased size is a consequence of having chromids/megaplasmids.

1.9.2 More rapid bacterial growth rate

A second consideration is that the multipartite genome organization may allow for faster bacterial division by decreasing the time required to replicate the genome as each replicon can replicate concurrently (Frage *et al.*, 2016; Rasmussen *et al.*, 2007). While perhaps true in some cases, there is little general support for this hypothesis. Indeed, while some of the fastest replicating species have a multipartite genome (e.g. *Vibrio* species), multipartite genomes are also present in several slow growing organisms (e.g. *R. sphaeroides*) (Egan *et al.*, 2005), and no correlation has been detected between genome size and growth rate (Vieira-Silva *et al.*, 2010). Furthermore, if growth rate was the driving force for the division and maintenance of the multipartite genome, it would be expected that each replicon in the genome would be of similar size, which is not what is observed (Couturier & Rocha, 2006). Additionally, this argument assumes DNA replication is the growth limiting step in cell division, and if true, combining the multiple

replicons of a multipartite genome into a single replicon should significantly impair growth rate. Although a slight decrease in growth rate is observed if all three of the *S. meliloti* replicons are integrated into a single chromosome (Guo *et al.*, 2003), the difference is much less than would be expected if DNA replication was the growth rate limiting factor. Hence, it seems unlikely that growth rate is a common driving force for the evolution of multipartite genomes.

1.9.3 Different evolutionary trajectories

Several studies have observed different rates of evolution on each replicon in a multipartite genome. Work by Cooper *et al*. illustrated that the substitution rate of the chromid of the *Vibrio* species is greater than that of the chromosome, whereas purifying selection is weaker on the chromid (Cooper *et al.*, 2010). Similarly, the substitution rate of the chromid in the *Burkholderia* is greater than that of the chromosome but less than the megaplasmid, while purifying selection is greatest on the chromosome, then chromid, and finally the megaplasmid (Cooper *et al.*, 2010). Comparison of the conserved sequences of the chromosomes and chromids of *R. sphaeroides* strains was suggestive of the chromid undergoing more rapid evolution (Choudhary *et al.*, 2007), although only a non-statistically significance increase in evolutionary rates of genes duplicated between the chromosome and chromid was observed compared to the evolutionary rates of duplicated genes both present on the chromosome or both on the chromid (Peters *et al.*, 2012). Comparison of orthologous genes between *Burkholderia* genomes indicated that the percent amino acid identity of the gene products was highest for the chromosome,

intermediate for the chromid, and lowest for the megaplasmid (Chain *et al.*, 2006). In contrast, genes involved in the rhizobium – legume symbiosis encoded by the megaplasmid of *Sinorhizobium* species showed less divergence than those encoded by the chromosome or chromid (Guo *et al.*, 2014). While at first glance this result is conflicting that of the *Vibrio* and *Burkholderia* observations mentioned above, it is perhaps not surprising as the megaplasmid is the primary replicon with respect to the symbiosis. While some of the differences in the evolutionary patterns on each replicon may simply be due to differences in the gene content of each replicon, at least some of the evolutionary biases can be attributed to the location of the genes on a chromosome versus chromid versus megaplasmid (Cooper *et al.*, 2010).

In addition to difference in the rate of evolution on each replicon, data suggests that the overall evolutionary pattern of each replicon in a multipartite genome may differ. Specifically, Galardini *et al*. (2013) demonstrated that the chromosome of *S. meliloti* is structurally stable and primarily vertically transmitted, the chromid was formed by ancient horizontal gene transfer and is under greater positive selection particularly in genes for environmental adaptation, while the megaplasmid is structurally fluid and formed by recent and ongoing horizontal transfer. Differences in evolutionary patterns of each replicon are further supported by the genetic variability associated with a multipartite genome as reviewed in Section 1.6.3. Overall, the data suggest that each replicon in a multipartite genome experiences different evolutionary pressures and display unique rates of evolution.

1.9.4 Co-ordinated gene regulation

A further hypothesis in the literature is that the division of genes between multiple replicons facilitates their co-ordinated regulation. In support of this, an analysis of the regulons of *S. meliloti* transcription factors illustrated a bias for transcription factors to regulate genes on the same replicon (Galardini *et al.*, 2015). Others have suggested that division of genes between replicons provides a mechanism for modulating gene dosage. It has been observed that the earlier initiation of replication of the chromosomes of the *Vibrio* species, compared to the chromid, results in higher average gene dosage and thus expression of the chromosomal genes (Dryselius *et al.*, 2008). However, it is likely that this effect is limited to fast-growing bacteria (Couturier & Rocha, 2006). Recently, an interesting suggestion that localization of genes to different replicons may facilitate their correct sub-cellular positioning was suggested (Frank *et al.*, 2015), although experimental support for this hypothesis is currently lacking. Indeed, the topological organization of multipartite genomes in the cell has received little attention. Very recent work in *V. cholerae* showed that the chromosome and chromid had distinct organization and subcellular localization, and physical interactions between the replicons were detected (Galli *et al*., 2016). It would be interesting in future work to examine whether this is specific to *V. cholerae* or a general characteristic of species with multiple large replicons.

Several studies have reported replicon biases in transcriptional regulation of genes during niche adaptation. Comparison of the *V. cholerae* transcriptome between laboratory growth and intestinal growth illustrated that many more chromid genes are

expressed in the intestine than in lab conditions (Xu *et al.*, 2003). In a comparative transcriptomics study of *B. cenocepacia* J2315, there was an over-representation of the large chromid genes among the genes expressed during soil colonization, whereas during growth during *in vitro* cystic fibrosis conditions, chromosomal genes were over-represented among the expressed genes (Yoder-Himes *et al.*, 2010). Additionally, transcriptional differences between J2315 and strain HI2424 were biased towards the smaller chromid of J2315 (Yoder-Himes *et al.*, 2010). The pRL8 replicon of *R. leguminosarum* is enriched for genes up-regulated during growth in the pea rhizosphere (Ramachandran *et al.*, 2011), and similarly, secondary replicons of *Rhizobium phaseoli* are enriched in genes upregulated during rhizosphere colonization (López-Guerrero *et al.*, 2012). During symbiosis with legumes, genes down-regulated in *S. meliloti* are over-represented on the chromosome compared to the other replicons, whereas genes up-regulated are over-represented on the megaplasmid (Barnett *et al.*, 2004; Becker *et al.*, 2004). These studies clearly demonstrate that replicon specific patterns of gene expression can be observed. However, they do not address whether the division of genes between replicons is directly to facilitate coordinated regulation, or whether it is a by-product of functionally related genes being co-localized to the same replicon.

1.9.5 Niche adaptation

Various aspects of the previously described observations regarding replicon specific regulation patterns, functional biases, and evolutionary patterns have resulted in the suggestion that each replicon in a genome serves a specialized role with secondary

replicons contributing to niche adaptation (Chain *et al.*, 2006; Galardini *et al.*, 2013). Considering the observations summarized above, I propose that the chromosome is a non-differentiated replicon encoding the general requirements for survival without focus on any particular environment, whereas secondary replicons (chromids and megaplasmids) are specialized entities required for colonization of a specific niche, in particular, a niche associated with an eukaryotic host. However, while several lines of evidence are consistent with this hypothesis, no clear evidence had previously existed that directly implicated secondary replicons as contributing to fitness of a bacterial species specifically or predominately in one particular environment.

1.10 The rhizobium – legume symbiosis

Biological nitrogen fixation (BNF) is a process whereby certain prokaryotes convert atmospheric N_2 gas into NH_3 using the nitrogenase enzyme. This can be performed, for example, by free-living diazotrophs living in close association with plant roots, or by bacteria during endosymbiotic relationships with particular plant species. The rhizobia are bacterial species that enter into endosymbiotic relationships with legume species. These bacteria are not monophyletic and consist of both α- and β-proteobacteria. During this symbiosis, the rhizobia are present intra-cellularly in a specialized legume organ known as a nodule where the rhizobia convert (fix) atmospheric N_2 gas into NH_3 for use by the plant as a nitrogen source in exchange for carbon substrate from the plant.

1.10.1 The symbiosis

The symbiotic relationship is initiated following an exchange of signals between

the bacteria in the rhizosphere and the plant. Legume root hairs curl around the detected rhizobial cells and invagination of the plant cell wall leads to entry of the rhizobium into the plant. The rhizobia replicate and travel through the infection thread until they reach a nodule cell, where the rhizobia are released from the infection thread into the nodule cell. Once inside the nodule cell, the bacteria divide and differentiate into nitrogen-fixing bacteroids. Whereas *S. meliloti* cells early in the infection process can leave the plant and begin growing as a free-living cell in the soil, the terminally differentiated bacteroids cannot. The various stages of the symbiotic process are reviewed in depth elsewhere (Gage, 2004; Long, 2001; Maróti & Kondorosi, 2014; Oldroyd *et al.*, 2011; Udvardi & Poole, 2013; Wang *et al.*, 2012).

1.10.2 Synthetic symbioses

An underlying goal of the study of the rhizobium – legume symbiosis has always been to manipulate this process for agricultural gains. This can be seen by the many reviews available through PubMed dating back to the 1960s that comment on this purpose (Bohlool *et al.*, 1992; Brewin & Legocki, 1996; Henzell, 1988; Nutman, 1971; Silver, 1969). The ultimate goal would be to develop a sustainable and biological solution to the nitrogen requirement of modern agriculture that can support the world's growing population. This would include the engineering of 'synthetic symbioses' with non-leguminous crop plants that are currently unable to reap the benefits of symbiotic nitrogen fixation, such as the cereals. Facilitated by recent strides in the field of synthetic biology, many researchers are now working towards this elusive goal although our understanding

of the underlying genetics, biochemistry, and physiology remain incomplete.

There are three prominent approaches to engineering of biological nitrogen fixation into non-legume plants: directly expressing nitrogenase in the plant, engineering new endosymbiotic relationships, or establishing a successful relationship between the plant and free-living diazotrophs (Mus *et al.*, 2016). The potential, the challenges, and the progress of these approaches are reviewed in depth elsewhere (Beatty & Good, 2011; Bhattacharjee *et al.*, 2008; Charpentier & Oldroyd, 2010; Curatti & Rubio, 2014; Geddes *et al.*, 2015; Oldroyd & Dixon, 2014; Rogers & Oldroyd, 2014). Engineering novel bacterial – plant N_2-fixing symbiotic relationships is perhaps the best long-term option, but is also quite difficult due to a lack of an elucidated necessary and sufficient symbiotic gene complement and lack of suitable genomic platforms for testing synthetic constructs.

1.11 *Sinorhizobium meliloti*

Sinorhizobium meliloti is an established model species for the study of several fields including carbon metabolism, multipartite genomes, and the rhizobium – legume symbiosis (Galardini *et al.*, 2013; Geddes & Oresnik, 2014). In addition to colonizing bulk soil environments, *S. meliloti* is also found in the rhizosphere, which is the soil directly surrounding the root of plants, as an endophyte in plant tissues, and within the nodules of legumes belonging to the *Medicago*, *Melilotus*, and *Trigonella* genera.

The genome properties of *S. meliloti* make it a good model species for the study of the multipartite genome (Barnett *et al.*, 2001; Capela *et al.*, 2001; Finan *et al.*, 2001; Galibert *et al.*, 2001). The genome of the model strain, Rm1021, consists of three

replicons. The 3.65 Mb chromosome encodes nearly all the essential cell functions. The 1.68 Mb chromid known as pSymB encodes two essential genes (diCenzo *et al.*, 2013), and is highly enriched in ABC solute transport systems (Finan *et al.*, 2001) and exopolysaccharide biosynthetic genes (Finan *et al.*, 1986; 2001). The 1.35 Mb pSymA megaplasmid is non-essential for free-living growth (Oresnik *et al.*, 2000), but is required for the establishment of a successful symbiosis (Barnett *et al.*, 2001; Rosenberg *et al.*, 1981). All *S. meliloti* strains isolated from soil environments have symbiotic capabilities (Bromfield *et al*., 1995; Hartmann *et al*., 1998). All three of the *S. meliloti* Rm1021 replicons are conserved in all sequenced *S. meliloti* strains although the gene contents and sizes can vary considerably (Galardini *et al.*, 2013; Guo *et al.*, 2009; Sugawara *et al.*, 2013), and most wild type strains may contain additional smaller plasmids (Hartmann *et al*., 1998; Galardini *et al.*, 2011; 2013; Schneiker-Bekel *et al.*, 2011).

The use of *S. meliloti* as a model system is facilitated by the large number of genomic resources and systems level data available for *S. meliloti*. Available genomic resources include a transcriptional fusion library (Cowie *et al.*, 2006), and an ORFeome library (Schroeder *et al.*, 2005), multiple large-scale deletion libraries (Milunovic *et al.*, 2014; Yurgel *et al.*, 2013), a signature-tagged mutant library (Pobigaylo *et al.*, 2006; 2008), and RNA-seq data available through the COLOMBOS database (Moretto *et al.*, 2016). This was expanded in this work to include a collection of strains lacking pSymA and/or pSymB, as well as a metabolic network reconstruction (Chapters 2 and 3). In terms of systems biology data available for *S. meliloti*, numerous transcriptomics (e.g.

Roux *et al.*, 2014; Schlüter *et al.*, 2010), proteomics (e.g. Djordjevic, 2004; Sobrero *et al.*, 2012), metabolomics (e.g. Fei *et al.*, 2016; Gemperline *et al.*, 2015), and phenomics (e.g. Biondi *et al.*, 2009; Spini *et al.*, 2015) datasets are present, as are pangenome analyses (e.g. Galardini *et al.*, 2013; Sugawara *et al.*, 2013), and a regulon analysis (Galardini *et al.*, 2015). These tools and resources greatly facilitate the study of *S. meliloti* biology.

1.12 This work

1.12.1 The aim

Prokaryotic genomes are not simply stochastically organized but instead reflect some functional or regulatory purpose (Junier, 2014; Lawrence, 2003; Rocha, 2008). For example, enzymes encoding each step of a pathway are generally encoded as a single operon, and are often co-localized on the chromosome with their regulator or operons encoding pathways of related function, such as transport and metabolism of a substrate (Rocha, 2008). The chromosomal location of a gene can influence gene expression level (Bryant *et al.*, 2015), and at least in fast replicating species, the copy number of the gene (Dryselius *et al.*, 2008). Additionally, there is a general bias for bacterial genes to be enriched in the leading strand to avoid head-on collisions between the transcriptional and DNA replicative machinery (Rocha, 2004). Considering that prokaryotic genomes are highly structured, the emergence of the multipartite genome organization was presumably shaped by selective pressures to facilitate a particular process.

Understanding what these evolutionary forces are and the advantage of maintaining multiple replicons is particularly important as many important bacteria

contain this genomic architecture (Figure 1.4). These include plant symbionts like many of the rhizobia, plant pathogens such as the *Agrobacterium*, and animal and human pathogens like *Brucella*, *Vibrio*, and *Burkholderia*. I hope that by understanding the emergence and function of this genome structure we can gain generalizable insights into the biology of these diverse organisms that could lead to practical applications in promoting or suppressing these symbiotic and pathogenic relationships. In this work, I use *S. meliloti* as a model organism together with a combination of experimental and *in silico* tools to study the role and evolution of the bacterial multipartite genome.

1.12.2 Key findings

The key outcomes of this thesis can be split into two categories: new genomic and *in silico* resources for the study of *S. meliloti*, and novel insights into the evolution and function of the multipartite bacterial genome.

The new resources presented are: **a)** a procedure for the removal of pSymB; **b)** a set of strains that include wild type *S. meliloti* as well as those lacking either the pSymA megaplasmid, the pSymB chromid, or both; **c)** a strategy for the conjugation of unmarked pSymA and pSymB replicons; **d)** the use of this technique to produce highly isogenic strains with both pSymA and pSymB, just pSymA, just pSymB, or neither; **e)** a large-scale *in vivo* cloning technique for *Sinorhizobium*; **f)** a *S. meliloti* strain lacking both pSymA and pSymB, but with an ancestral 69 kilobase region integrated into the chromosome, which will serve as a platform for identification of the minimal symbiotic genome; and **g)** an *in silico S. meliloti* genome-scale metabolic network reconstruction.

The major new insights derived from this work include: **a)** pSymA provides little advantage to free-living *S. meliloti* cells, although it may be advantageous in unique environments (e.g. during iron limitation through siderophore biosynthesis); **b)** pSymB is non-essential for growth in bulk soil; **c)** the majority of the metabolic capabilities dependent on pSymB seem to be of little relevance in bulk soil and may be specialized for colonization of the rhizosphere; **d)** adaptation to the rhizosphere involves a metabolic refinement with pSymB becoming of greater importance, whereas the transition from a free-living microbe to a nitrogen fixing bacteroid required a major metabolic change; **e)** the essential genes present on pSymB are due to a recombination event that transferred a 69 kilobase region (the ETR region) to pSymB from the chromosome following the split of *S. meliloti* from *Sinorhizobium fredii*; **f)** there is significant functional redundancy between the gene products encoded by the *S. meliloti* chromosome with those encoded by pSymB and less so with pSymA; **g)** adaptation to multiple niches is likely a major driving force behind the evolution of the multipartite bacterial genome, with secondary replicons specialized for a particular environment; **h)** gene flow between the chromosome and other replicons contributes to chromid formation; and **i)** the single copy pSymA and pSymB genes essential for symbiotic N_2-fixation were delineated to < 12% of these replicons.

1.12.3 Overlap between chapters

There is little direct overlap between chapters. There is overlap between the Materials and Methods sections Chapters 2, 4, and 5. Additionally, there is some overlap in the introduction sections of Chapters 2 and 3, and the Sections 2.5.6 and 3.5.

1.13 Tables and Figures

Table 1.1 Global replicon specific COG analysis.

Classification		Replicon Enrichment				
COG	Description	Chromosome	Chromid	Megaplasmid	Plasmid	Total genes
A	RNA processing and modification	0.06	-1.64	-2.61	-2.37	2,212
B	Chromatin structure and dynamics	0.04	-0.48	-1.28	-1.98	1,982
C	Energy production and conversion	0.00	0.22	0.11	-1.24	311,749
D	Cell cycle control, cell division, chromosome partitioning	0.01	-0.83	-0.76	0.95	53,815
E	Amino acid transport and metabolism	0.00	0.46	0.04	-1.42	423,590
F	Nucleotide transport and metabolism	0.05	-0.91	-1.39	-2.08	122,047
G	Carbohydrate transport and metabolism	-0.01	0.50	0.10	-1.07	327,915
H	Coenzyme transport and metabolism	0.03	-0.54	-0.77	-1.46	218,207
I	Lipid transport and metabolism	0.00	0.28	0.22	-1.18	192,881
J	Translation, ribosomal structure and biogenesis	0.06	-1.69	-2.25	-3.20	272,466
K	Transcription	-0.02	0.57	0.39	-0.35	390,373
L	Replication, recombination and repair	0.00	-0.85	0.04	1.11	265,234
M	Cell wall/membrane/envelope biogenesis	0.02	-0.13	-0.53	-0.88	294,750
N	Cell motility	0.01	0.16	-0.53	-0.90	97,783
O	Posttranslational modification, protein turnover, chaperones	0.03	-0.51	-0.59	-1.08	181,774
P	Inorganic ion transport and metabolism	-0.01	0.36	0.08	-0.51	259,353
Q	Secondary metabolites biosynthesis, transport and catabolism	-0.03	0.60	0.65	-0.53	127,189
R	General function prediction only	0.01	0.09	-0.07	-0.85	596,567
S	Function unknown	0.01	0.06	-0.24	-0.64	426,972
T	Signal transduction mechanisms	0.00	0.30	-0.03	-0.92	318,564
U	Intracellular trafficking, secretion, and vesicular transport	0.00	-0.29	-0.16	0.49	121,189
V	Defense mechanisms	0.01	-0.15	-0.13	-0.32	83,369
W	Extracellular structures	-0.08	1.26	1.16	-1.18	324
Y	Nuclear structure	0.08	0.00	0.00	0.00	1
Z	Cytoskeleton	0.06	-2.53	-1.79	-0.54	936

Results of the Cluster of Orthologous Genes (COG) functional analysis is presented. One representative genome from each of the 1,708 species with finished genomes available through NCBI was randomly chosen, all genes from each replicon type were extracted (genes: chromosome – 5,342,421; chromid – 174,984; megaplasmid – 62,606; plasmid – 79,077; total – 5,659,088), and genes annotated with COG categories via WebMGA (Wu *et al.*, 2011). The enrichment ($\log_2$[observed / expected]) for each COG category for each replicon is given, as is the total number of genes annotated with each COG class.

Table 1.2. Frequency of multipartite genomes in large bacterial genomes.

Genome size (Mb)	Total species	Chromid*	Megaplasmid*	Multipartite*†
≥ 5	439	77 (18)	62 (14)	104 (24)
≥ 6	217	52 (24)	45 (21)	67 (31)
≥ 7	126	33 (26)	31 (25)	42 (33)
≥ 8	77	15 (19)	15 (19)	21 (27)
≥ 9	44	5 (11)	5 (11)	7 (16)
≥ 10	16	0 (0)	0 (0)	0 (0)

* The number of species with at least one chromid, at least one megaplasmid, or at least one chromid or megaplasmid is shown, with the percentage of the total genomes shown in parentheses.

† The multipartite column includes species containing either a chromid or a megaplasmid.

Figure 1.1. Phylogenetic distribution of bacterial linear chromosomes.

A phylogenetic distribution of 1,708 bacterial species with a finished genome available through NCBI is shown. The taxa names are coloured based on chromosome structure: red for species with circular chromosomes, green for species with linear chromosomes, and purple for species where at least one strain has a circular chromosome while at least one other strain has a linear chromosome. The *Streptomyces* and the *Borrelia* genera are indicated as these genera are particularly enriched for linear chromosomes. For construction of the phylogeny, the AMPHORA2 pipeline (Wu & Scott, 2012) was used to identify 11 highly conserved ribosomal proteins (RplA, RplC, RplE, RplF, RplN, RplT, RpsC, RpsE, RpsI, RpsK, RpsM) in the proteomes of each of the 1,708 species using hidden Markov models and HMMER 3.1.b2 (hmmer.org). Each of the 11 proteins were aligned with Clustal Omega (Sievers *et al.*, 2011), trimmed with trimAl (Capella-Gutiérrez *et al.*, 2009), concatenated, the phylogeny built using the RAxML BlackBox webserver mirror site on the Cipres Science Gateway (Miller *et al.*, 2010; Stamatakis, 2014; Stamatakis *et al.*, 2008), and visualized with FigTree (tree.bio.ed.ac.uk/software/figtree/). The presence of linear or circular chromosomes was determined based on the annotation in the corresponding GenBank flat file.


Streptomyces
Borrelia
0.4

Figure 1.2. Size distribution of bacterial genomes and replicons.

Histograms displaying the size distribution of (**A**) circular bacterial chromosomes (blue) and linear bacterial chromosomes (red), (**B**) all bacterial genomes (blue) and all bacterial chromosomes (red), (**C**) plasmids, and (**D**) chromids (blue) and megaplasmids (red). The purple colour occurs as a result of overlap between the red and blue bars. Histograms are based on all 1,708 bacterial species with a completed genome available through NCBI. Where more than one genome was available for a species, the species chromosome and genome sizes used were the species average, whereas a random strain was chosen for analysis of the plasmids, megaplasmids, and chromids. Histograms were produced in R (http://www.R-project.org/).

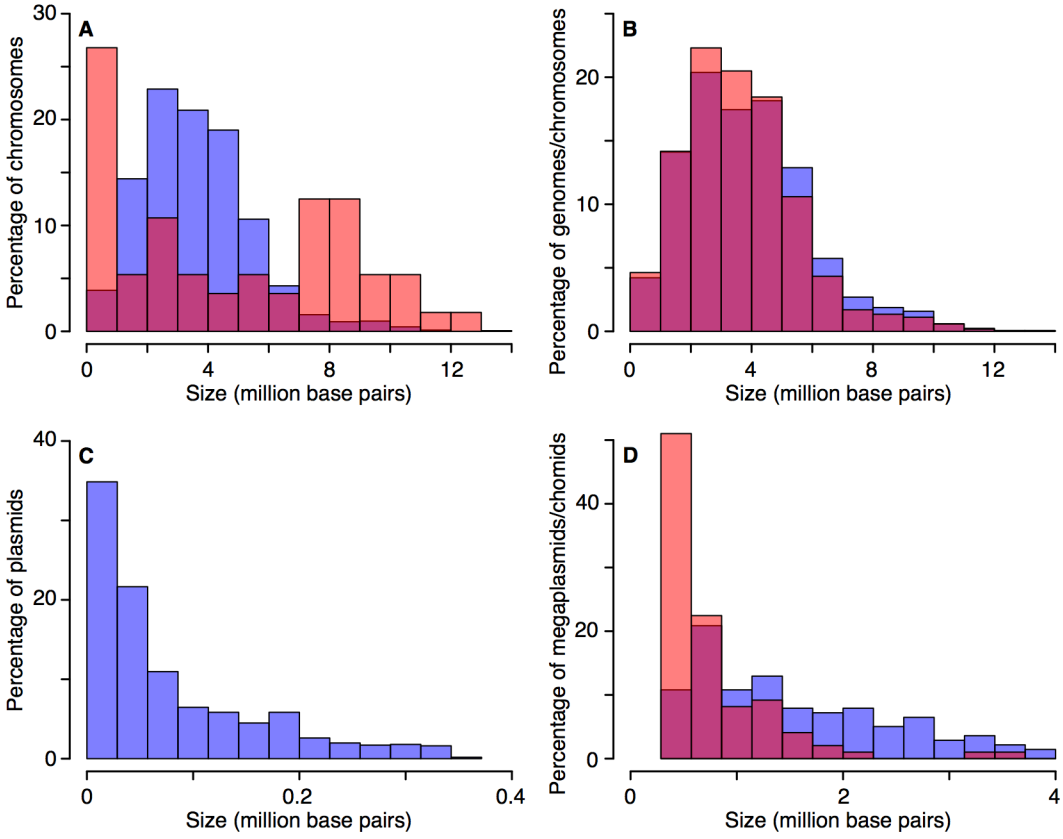

A
Percentage of chromosomes
Size (million base pairs)
B
Percentage of genomes/chromosomes
Size (million base pairs)
C
Percentage of plasmids
Size (million base pairs)
D
Percentage of megaplasmids/chomids
Size (million base pairs)

Figure 1.3. Phylogenetic distribution of bacterial replicon classes.

A phylogenetic distribution of 1,708 bacterial species with a finished genome available through NCBI is shown. The taxa names are coloured based on genome structure: red for species with just a single chromosome, green for species with a megaplasmid(s) but no chromid, blue for species with a chromid(s) but no megaplasmid, and purple for species with both a megaplasmid(s) and chromid(s). Genera enriched for megaplasmids and/or chromids are labelled. For species with more than one completed genome available through NCBI, the species was considered to have a megaplasmid or chromid as long as it was present in at least one strain. The phylogeny was built as described in the Figure 1.1 legend. Classification of replicons was performed as follows. The largest replicon in each genome was considered the chromosome. Any replicon smaller than 350 kb was classified as a plasmid. Non-chromosomal replicons $\geq$ 350 kb were classified as a chromid if the GC content was within 1% that of the chromosome of the same species and if the dinucleotide relative abundance difference was < 0.4 compared to the chromosome, otherwise they were considered megaplasmids.

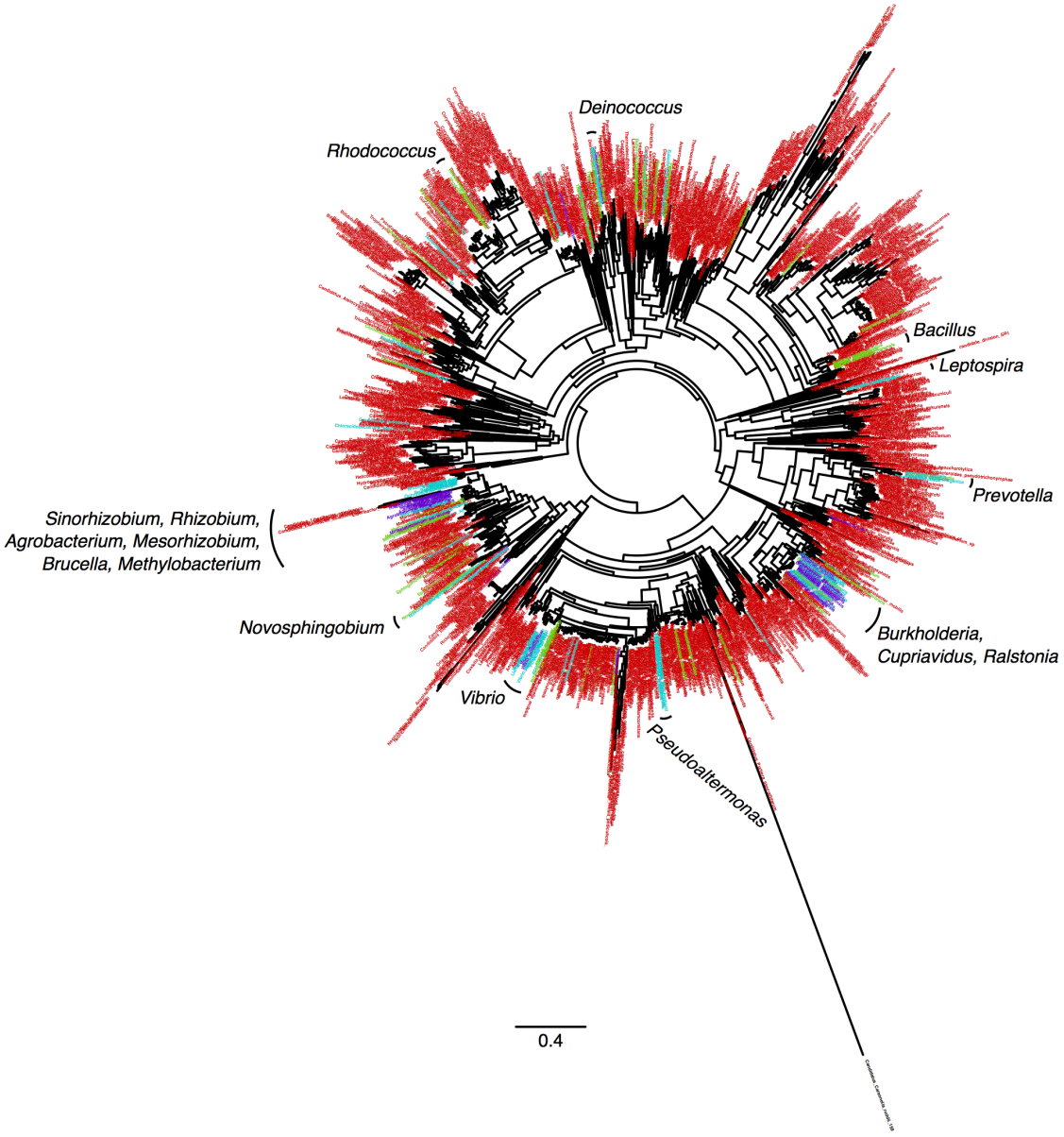

Deinococcus
Rhodococcus
Bacillus
Leptospira
Prevotella
Sinorhizobium, Rhizobium,
Agrobacterium, Mesorhizobium,
Brucella, Methylobacterium
Novosphingobium
Burkholderia,
Cupriavidus, Ralstonia
Vibrio
Pseudoalteromonas
0.4

Figure 1.4. Identification of a secondary chromosome in *S. enterica* NCTC10384.

Dot plots between the (**A**) *S. enterica* NCTC10384 secondary replicon with the *S. enterica* SC-B67 chromosome and (**B**) the NCTC10384 chromosome with the SC-B67 chromosome. The high synteny between the secondary replicon of NCTC10384 with the chromosome of *S. enterica* SC-B67 suggests that the two replicons of NCTC10384 resulted from a split of an ancestral chromosome into two replicons. Dot plots were created using the YASS online tool (Noé & Kucherov, 2005), with an e-value cut-off of 1×10^{-30} and with all other options left at the default settings.

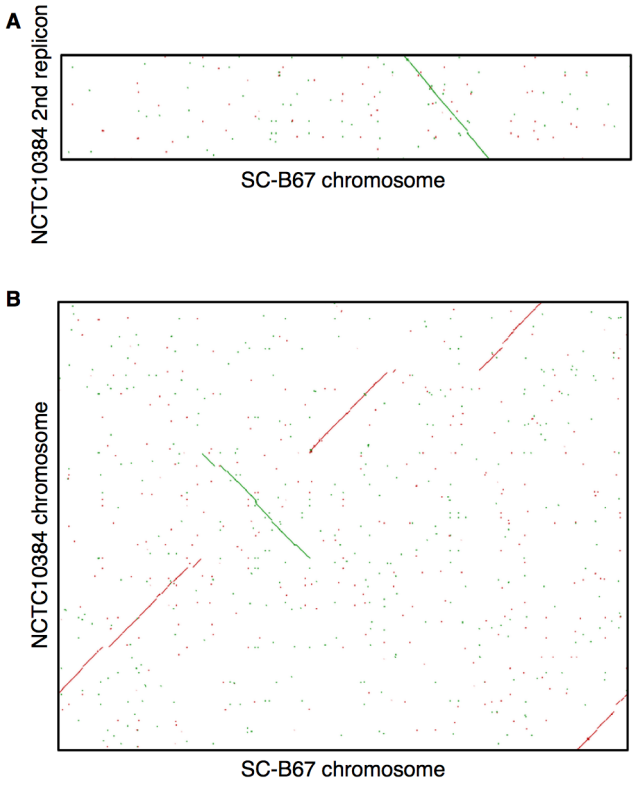

A
NCTC10384 2nd replicon
SC-B67 chromosome
B
NCTC10384 chromosome
SC-B67 chromosome

Figure 1.5. Genomic signatures of bacterial chromids, megaplasmids, and plasmids. Histograms displaying the (**A**) difference in GC content between the chromosome and secondary replicons and (**B**) the dinucleotide relative abundance distance between the chromosome and secondary replicons is shown. Data for chromids are shown in blue, megaplasmids in red, and plasmids in green. Additional colours occur when bars of different colours overlap. Histograms are based on one representative strain for each of the 1,708 bacterial species with a completed genome available through NCBI. Histograms were produced in R (http://www.R-project.org/).

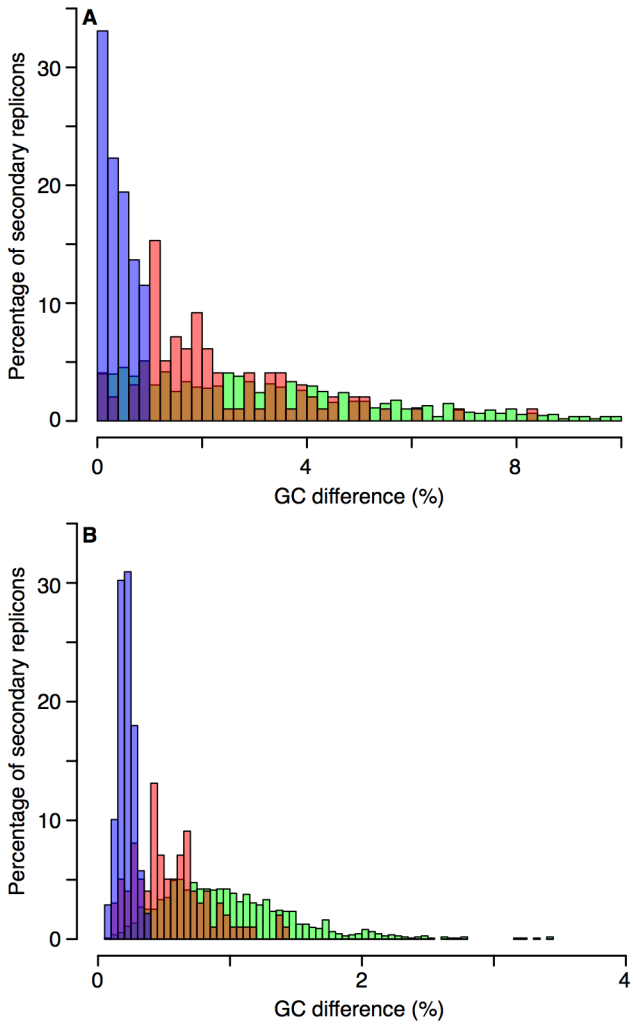
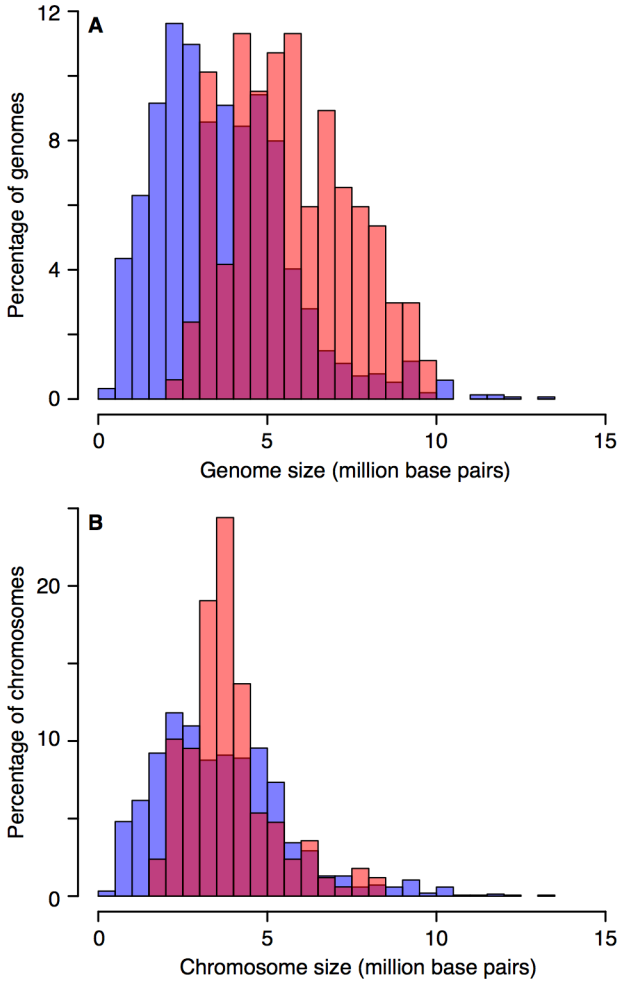

A
Percentage of secondary replicons
30
20
10
0
0
4
8
GC difference (%)
B
Percentage of secondary replicons
30
20
10
0
0
2
4
GC difference (%)

Figure 1.6. Size distribution of genomes and chromosomes for species with and without multipartite genomes.

Histograms displaying the distribution of (**A**) total genome sizes or (**B**) chromosome sizes for multipartite genomes (red) and genomes without a chromid/megaplasmid (blue) are shown. The purple colour occurs as a result of overlap between the red and blue bars. Histograms are based on one randomly chosen strain for each of the 1,708 bacterial species with a completed genome available through NCBI. Histograms were produced in R (http://www.R-project.org/).


A
Percentage of genomes
0
4
8
12
0
5
10
15
Genome size (million base pairs)
B
Percentage of chromosomes
0
10
20
0
5
10
15
Chromosome size (million base pairs)

CHAPTER 2. EXAMINATION OF PROKARYOTIC MULTIPARTITE GENOME EVOLUTION THROUGH EXPERIMENTAL GENOME REDUCTION

Citation: diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. 2014. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet 10(10): e1004742.

2.1 Preface

The work detailed in this chapter either directly or indirectly sets the stage for all work described in the subsequent chapters. In this chapter, a hypothesis for multipartite genome evolution is described, and this theory is used for work described in the later chapters. It is argued that secondary replicons carry the genetic determinants required to inhabit niche not previously inhabited by the species, but contribute less to fitness in the cell's original environment, and as such, the loss of the replicon is favoured in some environments. Over time, secondary replicons becomes more integrated into core cellular metabolism primarily through transfer of genes from the chromosome, resulting in the stable inheritance of the replicon. Inhabiting the new niche leads to a selective pressure to accumulate fitness promoting genes for this new environment. These new genes are primarily gained by the secondary replicon, and this results in the evolution of a large, chromosome like replicon that serves largely as a niche specific entity.

Prior to this work, several of the points presented in this hypothesis had been described previously but had not been synthesized into a single hypothesis as done here. However, the lack of experimental support limited acceptance of these points. The work described in this chapter represents the first real experimental evaluation of the described hypothesis. The genome of *Sinorhizobium meliloti* consists of a primary chromosome, the pSymB chromid, and the pSymA megaplasmid. *S. meliloti* is therefore a good model for studying multipartite genomes as it has a replicon from each of the major replicon classes, with chromids representing 'older' secondary replicons and megaplasmid

referring to 'young' secondary replicons. A set of *S. meliloti* strains consisting of wild type and strains lacking pSymA, pSymB, or both pSymA and pSymB were developed. This allowed us to examine the biological roles of both pSymA and pSymB, and compare/contrast the phenotypic consequences resulting from the removal of a chromid versus a megaplasmid.

Using growth studies in synthetic laboratory media and sterilized bulk soil (the original niche of *S. meliloti*), intra- and inter-species competition assays, and metabolic characterization with the Biolog Phenotype MicroArray^{TM} technology, we provided evidence that pSymB is more integrated into core cellular metabolism than is pSymA, which played little to no role in the tested conditions. This has since been further supported by metabolomics (Fei *et al.*, 2016) and transcriptomics (unpublished) analyses. But at the same time, we noted that while pSymB is required for the utilization of a large number of nutrient sources, the majority of metabolic functions associated with pSymB were of little to no value in a bulk soil environment, consistent with the majority of pSymB encoded functions being niche specialized. These analyses therefore support two tenants of the model described above: that secondary replicons are integrated into the organism's core metabolism over time, and that secondary replicons are primarily specialized for adaptation to particular niches.

In conclusion, this chapter details an overarching model for multipartite genome evolution, describes a set of strains that will facilitate new experimental studies on the evolution of multipartite genomes, and presents experimental evidence consistent with

two tenants of the multipartite genome evolution hypothesis. Additionally, various aspects of the described model serve as hypotheses that will be tested in subsequent chapters.

2.2 Abstract

Many bacteria carry two or more chromosome-like replicons. This occurs in pathogens such as *Vibrio cholerae* and *Brucella abortus* as well as in many N_2-fixing plant symbionts, including all sequenced isolates of the alfalfa root-nodule bacteria *Sinorhizobium meliloti*. Understanding the evolution and role of this multipartite genome organization will provide significant insight into these important organisms; yet this knowledge remains incomplete, in part, because technical challenges of large-scale genome manipulations have limited experimental analyses. The distinct evolutionary histories and characteristics of the three replicons that constitute the *S. meliloti* genome [the chromosome (3.65 Mb), pSymA megaplasmid (1.35 Mb), and pSymB chromid (1.68 Mb)] makes this a good model to examine this topic. We transferred essential genes from pSymB into the chromosome, and constructed strains that lack pSymB as well as both pSymA and pSymB. This is the largest reduction (45.4%, 3.04 Megabases, 2866 genes) of a prokaryotic genome to date and the first removal of an essential chromid. Strikingly, strains lacking pSymA and pSymB (ΔpSymAB) lost the ability to utilize 55 of 74 carbon sources and various sources of nitrogen, phosphorous, and sulfur, yet the ΔpSymAB strain grew well in minimal salts media and in sterile soil. This suggests that the core chromosome is sufficient for growth in a bulk soil environment and that the pSymA and

pSymB replicons carry genes with more specialized functions such as growth in the rhizosphere and interaction with the plant. These experimental data support a generalized evolutionary model, in which non-chromosomal replicons primarily carry genes with more specialized functions. These large secondary replicons increase the organism's niche range, which offsets their metabolic burden on the cell (e.g. pSymA). Subsequent co-evolution with the chromosome then leads to the formation of a chromid through the acquisition of functions core to all niches (e.g. pSymB).

2.3 Author Summary

Rhizobia are free-living bacteria of agricultural and environmental importance that form root-nodules on leguminous plants and provide these plants with fixed nitrogen. Many of the rhizobia have a multipartite genome, as do several plant and animal pathogens. All isolates of the alfalfa symbiont, *Sinorhizobium meliloti*, carry three large replicons, the chromosome (~3.7 Mb), pSymA megaplasmid (~1.4 Mb), and pSymB chromid (~1.7 Mb). To gain insight into the role and evolutionary history of these replicons, we have 'reversed evolution' by constructing a *S. meliloti* strain consisting solely of the chromosome and lacking the pSymB chromid and pSymA megaplasmid. As the resulting strain was viable, we could perform a detailed phenotypic analysis and these data provided significant insight into the biology and metabolism of *S. meliloti*. The data lend direct experimental evidence in understanding the evolution and role of the multipartite genome. Specifically, the large secondary replicons increase the organism's niche range, and this advantage offsets the metabolic burden of these replicons on the

cell. Additionally, the single-chromosome strain offers a useful platform to facilitate future forward genetic approaches to understanding and manipulating the symbiosis and plant-microbe interactions.

2.4 Introduction

While most bacterial genomes have only a single chromosome, many are more complex and consist of two or more large replicons. Depending on their characteristics, these replicons are classified as a chromosome (largest replicon containing most of the core genes), megaplasmid (laterally acquired with a plasmid origin of replication and lacking core genes), or a chromid (displays characteristics of both chromosomes and megaplasmids) (Harrison *et al.*, 2010). While this genome organization is most commonly found in the proteobacteria, it is by no means limited to this class (Landeta *et al.*, 2011). Interestingly, many plant symbionts (e.g. *Sinorhizobium* and *Rhizobium* species) and plant and animal pathogens (e.g. *Agrobacterium*, *Vibrio*, *Burkholderia*, and *Brucella*) contain a multipartite genome (Harrison *et al.*, 2010; Landeta *et al.*, 2011). As such, understanding the general role and evolution of these accessory replicons may provide vital insight into the biology of these organisms and possible strategies to promote or suppress these interactions.

The potential advantages of multipartite genomes imply that this genome architecture is not simply an evolutionary peculiarity. For example, the division of a genome may decrease the time required for genome replication, potentially allowing more rapid growth. Indeed, multipartite genomes are larger on average (Harrison *et al.*, 2010)

and some of the fastest replicating species have divided genomes (Couturier & Rocha, 2006). However, each replicon within a divided genome is not of equal size (Couturier & Rocha, 2006) and there is no correlation between genome size and maximal growth rate (Vieira-Silva *et al.*, 2010). Alternatively, multipartite genomes may provide a method of controlling gene dosage and thus expression, as in *Vibrio* species (Couturier & Rocha, 2006; Dryselius *et al.*, 2008). This can consequently result in weaker purifying selection and greater rates of evolution on the smaller replicon, as observed in *Vibrio* and *Burkholderia* (Chain *et al.*, 2006; Cooper *et al.*, 2010). However, this does not hold true for slow-replicating species with a divided genome (Dryselius *et al.*, 2008). A third hypothesis is that multipartite genomes allow for additional genome expansion once the chromosome reaches its maximal size (Slater *et al.*, 2009). Yet, some species with multipartite genomes have primary chromosomes smaller than 2.5 Mb, while some species with a single chromosome have genomes greater than 9 Mb (Kaneko *et al.*, 2002; Michaux *et al.*, 1993). Moreover, *Brucella* species generally have two chromosome-like replicons, except for *Brucella suis* biovar 3, which has a single chromosome equivalent in size to the total of both replicons in related strains due to integration of one replicon into the other (Jumas-Bilak *et al.*, 1998; Moreno, 1998). While all three of the ideas discussed above may help promote the maintenance of a divided genome architecture once established, the observations inconsistent with each suggest they are unlikely to be general driving forces for multipartite genome evolution.

An alternative hypothesis is that multipartite genomes allow for the functional

division of genes onto separate replicons (Heidelberg *et al.*, 2000). Several lines of evidence are consistent with this idea: uneven COG distribution between each replicon such as in *Burkholderia xenovorans* (Chain *et al.*, 2006) and *Rhizobium etli* (González *et al.*, 2006), replicon-dependent evolution in *Sinorhizobium meliloti* (Galardini *et al.*, 2013), and replicon-dependent gene regulation in *Vibrio cholerae* (Xu *et al.*, 2003) and *S. meliloti* (Becker *et al.*, 2004). Furthermore, an association exists between the presence of a divided genome and an interaction with a host organism (Egan *et al.*, 2005). This hypothesis implies that secondary replicons are over-represented in cellular processes specific to host interaction, which, if true, should focus the genetic analyses of these processes; however, the acceptance of this idea is limited due to a paucity of experimental support (Harrison *et al.*, 2010).

S. meliloti is a N_2-fixing endosymbiont of legumes, and inhabits diverse environments including bulk soil, the rhizosphere, and the legume root nodule. It is an interesting organism to study the evolution of multipartite genomes as the large 6.7 Megabase (Mb) genome of the model strain Rm1021 (and the highly related strain, Rm2011) is divided into a chromosome (~3.7 Mb), an evolutionarily old and conserved chromid (pSymB; ~1.7 Mb), and an evolutionarily recent and variable megaplasmid (pSymA; ~1.4 Mb) (Epstein *et al.*, 2012; Galibert *et al.*, 2001; Guo *et al.*, 2009). Each of these is present in all wild-type isolates (Epstein *et al.*, 2012; Guo *et al.*, 2009), and there is no evidence that pSymA or pSymB are naturally lost by *S. meliloti*. This indicates that each replicon is a stable and indispensible part of the genome in the natural environment.

Both pSymA and pSymB encode major pathways of interaction with the plant symbiont and the environment: exopolysaccharide biosynthesis and many ABC transporters are encoded by pSymB (Finan *et al.*, 2001), and the nodulation and nitrogen fixation genes are present on pSymA (Barnett *et al.*, 2001). The complete removal of pSymA has been described (Oresnik *et al.*, 2000), and we now report the removal of pSymB and the construction of a strain lacking both pSymA and pSymB. This reduced genome provides a novel platform to facilitate forward genetic studies of rhizobium and bacterium-plant interactions, and we employed it here to experimentally test hypotheses surrounding the evolution and role of multipartite genomes.

2.5 Results and Discussion

S. meliloti forms N_2-fixing root nodules on alfalfa and all wild-type *S. meliloti* isolates examined thus far carry replicons equivalent to pSymA and pSymB (Epstein *et al.*, 2012; Guo *et al.*, 2009). Despite intensive investigation of *S. meliloti* over the past 40 years, there have been no reports of the successful removal of the pSymB chromid, with the earliest documented attempts published 25 years ago (Charles & Finan, 1991; Hynes *et al.*, 1989). The removal of pSymB as reported here was made possible through the application of several findings. First, the two essential genes (*engA* and $tRNA^{arg}$) that are located on pSymB were integrated into the chromosome (diCenzo *et al.*, 2013). Second, an active toxin-antitoxin locus (*smb21127/smb21128*) on pSymB was removed through the introduction of a 234 kilobase deletion (ΔB180) (Milunovic *et al.*, 2014). And third, cells that failed to inherit the remaining 1.45 Mb of pSymB were recovered by selecting

for the gain of a small plasmid carrying the *incα* incompatibility gene from the pSymB replication and partitioning *repABC* locus, rendering it incompatible with pSymB (MacLellan *et al.*, 2005). The latter transconjugants were obtained at a frequency of ~ 10^{-4}/recipient on LBmc medium containing excess cobalt, as the major *S. meliloti* cobalt uptake system (*cbtJKL*) is located on pSymB (Cheng *et al.*, 2011). Additionally, using a similar procedure, pSymB was removed from a previously isolated strain lacking pSymA (Oresnik *et al.*, 2000), resulting in cells with a genome consisting solely of the chromosome. For simplicity we refer to strains lacking pSymA, pSymB, or both, as ΔpSymA, ΔpSymB, or ΔpSymAB, respectively, although we accept that the Δ notation is normally reserved for a deletion in genetic notation. The genomes of the parent and three cured strains were sequenced using an Illumina MiSeq and reads were aligned to the previously reported Rm1021 and Rm2011 genome sequences (Galibert *et al.*, 2001; Sallet *et al.*, 2013) to confirm the absence of pSymA and/or pSymB sequences, as appropriate. The removal of pSymB is described in greater detail in the materials and methods.

Several other prokaryotic genome reduction studies have been reported in the past [e.g. *Escherichia coli* (Pósfai *et al.*, 2006), *Bacillus subtilis* (Ara *et al.*, 2007), and *Rhizobium* species (Hynes & McGregor, 1990; Moënne-Loccoz *et al*., 1995)], with the largest being a 38.9% (1.8 Mb) reduction of the *E. coli* genome (Iwadate *et al.*, 2011). The ΔpSymAB strain reported here lacks 3.04 Megabases, 2866 genes, and 45.4% of the *S. meliloti* genome, and thus represents the largest genome reduction reported to date and includes the first complete removal of an essential chromid from a genome. The

ΔpSymAB strain will facilitate new studies within a wide range of fields including refining the minimal symbiotic genome, general plant-microbe interactions, functional and evolutionary genomics, and biotechnology. Here, we detail phenotypic analyses of the *S. meliloti* strains lacking one or two replicons, and relate these observations to a generalized model for multipartite genome evolution.

2.5.1 Nutritional requirements

Optimal growth of the ΔpSymAB strain on complex LB or TY media required cobalt (Cheng *et al.*, 2011) and calcium supplementation, while growth on minimal M9 medium is best with thiamine (Finan *et al.*, 1986) and iron (Yurgel *et al.*, 2013) addition. The effect of calcium could possibly be related to the loss of exopolysaccharide loci on pSymB. No additional nutritional requirements were identified, which was unexpected as the genome sequence indicated the asparagine biosynthesis genes to be located on pSymB (Galibert *et al.*, 2001). Below we describe the growth of *S. meliloti* in sterile bulk soil, and interestingly, we observed that growth of the ΔpSymB and ΔpSymAB strains in this soil did not require thiamine supplementation. This indicates that thiamine biosynthesis, the sole nutrient whose biosynthesis is pSymB-dependent, is not required for growth in *S. meliloti*'s natural environment, although thiamine concentrations may limit growth in the rhizosphere (Streit *et al.*, 1996). Thus, very few fundamental genes are located on these replicons.

2.5.2 Effects on growth

Growth profiles of each strain were examined in complex and minimal media

(Figures 2.1A, 2.1B, 2.S1) by monitoring the change in OD_{600}. The light scattering properties of all strains were the same as in soil mesocosm experiments described below, a 10^{-4} dilution of cell suspensions with an OD_{600} value of 1 repeatedly resulted in viable counts of 4×10^3 CFU gm^{-1} of soil for each strain, indicating that the CFU/OD_{600} in the inoculum was 2×10^9 for all strains (Figures 2.1C, 2.2A, 2.3B). Removal of both replicons led to a surprisingly small growth deficit in minimal medium, with the ΔpSymAB strain showing only a 1.37-fold slower growth rate than that of the wild type. However, a striking pattern emerged when the effect of the removal of pSymA and pSymB was examined independently: loss of the evolutionarily older pSymB resulted in a 1.6-fold slower growth rate, while loss of the evolutionarily younger pSymA led to a 1.18-fold increase in growth rate. Qualitatively similar exponential phase dynamics are observed in complex media, although a large decrease in stationary phase density is observed when the cells lack pSymB.

Others have observed a fitness improvement following the loss of large replicons, such as a megaplasmid from *Agrobacterium tumefaciens* (Morton *et al.*, 2013) or large virulence plasmids from pathogenic *Escherichia coli* (Mellata *et al.*, 2010). Thus, it appears a general characteristic for large replicons is to be metabolically expensive, and that their maintenance indicates they must provide a fitness advantage to the cell not necessarily evident during laboratory growth; the symbiotic nodulation and N_2-fixation loci on pSymA would provide such a fitness benefit. While we also expect pSymB to impose a metabolic burden on growing cells, we postulate the loss of pSymB resulted in a

decreased growth rate because of the acquisition of core genes (by core genes, we mean genes that encode products that are either essential for survival or are involved in central bacterial processes) on pSymB from the chromosome (e.g. *bacA*, *minCDE*, *bdhA*). It has been shown that gene transfer occurs from the primary chromosome to secondary chromosomes and chromids (diCenzo *et al.*, 2013; Slater *et al.*, 2009); indeed, 25–30% of genes located on pSymB that are also present in the related species *A. tumefaciens* are located on the *A. tumefaciens* circular (primary) chromosome (Wong & Golding, 2003). This suggests that since their divergence, there has been significant gene transfer between the primary chromosome and secondary replicons in *S. meliloti* and *A. tumefaciens*. Furthermore, a bioinformatics approach indicated that in *Rhizobium etli* there is a correlation between the evolutionary age of a replicon and the level of functional integration with the chromosome (González *et al.*, 2006). Thus, while gene transfer from the chromosome to pSymA presumably occurs as well, the young evolutionary age of pSymA has so far precluded a significant accumulation of core genes.

2.5.3 Metabolic capacity

The decreased stationary phase density of strains lacking pSymB (Figure 2.1B) prompted an examination of the metabolic capacity of these cells. Accordingly, wild-type *S. meliloti* and the cured derivatives were examined for the ability to grow (increase in OD_{600}) with various sources of carbon, nitrogen, phosphorus, and sulfur. Wild-type *S. meliloti* grew on 73 carbon, 55 nitrogen, 53 phosphorus, and 20 sulfur sources (Table 2.1), and the removal of pSymA and particularly pSymB greatly decreased this potential

(Table 2.1, Data sets 2.S1, 2.S2, 2.S3, 2.S4). This was most evident in carbon metabolism, as 50 of 73 carbon sources required pSymB and/or pSymA (Table 2.2) to be effectively utilized. As pSymA and pSymB account for 45% of the genome (20% and 25%, respectively), if the carbon transport and metabolic genes were randomly distributed throughout the genome, only 45% of the carbon sources metabolized by the wild type (equivalent 33 of the 73) should be dependent on these replicon. Thus, the data show that carbon utilization loci are over-represented on the non-chromosomal replicons (50 vs. 33). Moreover, pSymB is essential for the metabolism of twice the expected number of carbon sources (36 vs. 18), which is consistent with the prevalence of predicted solute ABC transporters on pSymB (Ampomah *et al.*, 2013; Biondi *et al.*, 2009; Charles & Finan, 1991; Ding *et al.*, 2012; Finan *et al.*, 1988; 2001; Gage & Long, 1998; Geddes & Oresnik, 2012a, b; Geddes *et al.*, 2010; Jensen *et al.*, 2002; Kohler *et al.*, 2010; Lambert *et al.*, 2001; MacLean *et al.*, 2009; Mauchline *et al.*, 2006; Poysti *et al.*, 2007; Richardson *et al.*, 2004; Steele *et al.*, 2009; White *et al.*, 2012; Willis & Walker, 1999). Additionally, nitrogen and sulfur transport/metabolism is significantly enhanced by the presence of pSymB, although to a lesser extent than that of carbon metabolism, while phosphorus transport/metabolism is largely dependent on the chromosome (Table 2.1, Data sets 2.S2, 2.S3, 2.S4).

2.5.4 Saprophytic competence

To investigate the environmental significance of pSymA and pSymB, we developed a sterile soil mesocosm system to study the growth of wild-type *S. meliloti* and

the cured derivatives (Figure 2.1C; see materials and methods). In this system, the exponential growth dynamics of each strain were qualitatively similar to that in minimal medium; the loss of pSymA resulted in faster growth, the loss of pSymB impaired growth, and the removal of both resulted in an intermediate phenotype. Additionally, strains lacking pSymB showed a decreased stationary phase cell density similar to that observed in complex medium and consistent with their decreased metabolic capacity.

To identify the region(s) responsible for the growth defect associated with the removal of pSymB, a library of 14 strains in which defined regions of pSymB were deleted (representing ~ 90% of pSymB) (Milunovic *et al.*, 2014) was screened for growth in soil. None of the pSymB deletion strains showed a significant change in exponential growth dynamics, and only the loss of the two regions identified as B116 (pSymB nucleotide (nt) position 1,256,503 to 1,307,752) and B122 (nt 1,529,711–1,572,422) showed a significant and reproducible reduction in the stationary phase density in soil (Figure 2.2A). To investigate whether carbon availability was a growth-limiting factor in the soil, 15 mM glucose was added to stationary phase soil cultures of the wild type, ΔpSymAB strain, and strains with deletions of either the B116 or B122 regions. Viable cell counts following 3 and 5 days of incubation showed that glucose stimulated growth of all four strains, whereas no growth stimulation was observed following supplementation with nitrogen, phosphorus, and sulfur (Figure 2.2B). Thus, the availability of a usable carbon source appears to be a major factor limiting stationary phase growth for all strains in the soil mesocosms.

The deletion of B116 resulted in a 2 fold decrease in viable cell density in bulk soil, and the removal of B122 resulted in a 5–25 fold reduction (Figure 2.2A). While we have not confirmed which genes within these regions are responsible for the observed phenotype, we note that the B122 region includes genes (*bhbA-D*) involved in metabolism of the carbon storage compound poly-3-hydroxybutyrate (Charles & Aneja, 1999), while half of the B116 region spans a DNA fragment known to have translocated to pSymB from the chromosome in a *S. meliloti* ancestor (diCenzo *et al.*, 2013). As the stationary phase defect associated with the loss of both B116 and B122 is related to decreased carbon metabolic abilities (Figure 2.2B), it is reasonable to assume a multiplicative effect if both B116 and B122 are removed simultaneously, which would be a 10–50 fold decrease in stationary phase density. In fact, this is highly consistent with the observed stationary phase reduction of the ΔpSymB and ΔpSymAB strains (Figure 2.1C). Thus, we propose that the stationary phase defect associated with the removal of pSymB may be attributed predominately, if not entirely, to the loss of genes within these two regions.

In summary, the growth rate of *S. meliloti* in soil appears to be positively impacted by the removal of pSymA, but negatively impacted by the removal of pSymB, likely for the reasons discussed previously (see 'effects on growth'). On the other hand, the evidence shows that few pSymA- or pSymB-encoded metabolic capabilities are biologically necessary during growth of *S. meliloti* in sterile bulk soil. Thus, we wondered what evolutionary pressures maintain these metabolic capabilities. Slater et al.

presented strong bioinformatics evidence suggesting the common ancestor of the *Rhizobiales* order contained a single chromosome, and that this species captured a *repABC* plasmid (which they referred to as the ITR) that has evolved into secondary chromosomes or chromids in many modern day *Rhizobiales* (e.g. pSymB in *S. meliloti*, and the second chromosome of *Agrobacterium* species) (Slater *et al.*, 2009). The presence of exopolysaccharide biosynthetic genes, which facilitates a strong plant-microbe interaction (Egan *et al.*, 2005), on pSymB (Finan *et al.*, 2001) and the second chromosome of *Agrobacterium* species (Slater *et al.*, 2009) suggest that these genes may have originated on the ITR. Furthermore, phylogenetic studies have concluded that the evolution of an association with plants was associated with a large increase in solute, and particularly sugar, transporters (Boussau *et al.*, 2004; Pini *et al.*, 2011). Indeed, *S. meliloti* is capable of using a broad range of carbon sources for growth, and these functions are significantly over-represented on pSymB. Consequently, we suspect that an early plasmid derived from the ITR allowed improved colonization of the rhizosphere, leading to a selection for new genes specific to growth in this novel niche. Unlike plasmids, large rearrangements of bacterial chromosomes are generally selected against (Rocha, 2006; Slater *et al.*, 2009), thus the subsequent genome expansion occurred primarily with the ITR-derived plasmid, resulting in a replicon specialized for growth in the rhizosphere. While the fitness advantage provided by pSymB in the rhizosphere was not directly assessed here, we note that many of the carbon sources unable to support growth of the ΔpSymAB strain are indeed present in the rhizosphere (e.g. organic acids,

galactosides, and several polyols and sugars (Bringhurst *et al.*, 2001; Knee *et al.*, 2001; Ramachandran *et al.*, 2011) (Table 2.2).

2.5.5 Competitive phenotype

In natural environments, microorganisms are found as mixed populations and compete with each other for available resources. We therefore wished to examine whether pSymA and pSymB influence the competitive fitness of *S. meliloti*. Interestingly, in an agar plate assay, growth of the wild-type *S. meliloti* was found to inhibit the growth of strains lacking pSymA (Figures 2.3A, S2.2), but not the ΔpSymB strain (Figure S2.2). While such inhibition was observed previously (Perrine-Walker *et al.*, 2009), the nature of the inhibition was not identified. To identify loci responsible for this phenotype, we analyzed a library of strains, in which defined regions of pSymA were deleted (Milunovic *et al.*, 2014), for inhibition by the wild type. This screen identified a 64 kilobase region (A133) whose loss confers sensitivity to the inhibition by the wild type. This region encodes siderophore biosynthetic (*rhbA-F*) and uptake (*rhtA*, *rhtX*) genes (Lynch *et al.*, 2001), and subsequent mutant analysis revealed that simply disrupting the siderophore uptake genes conferred sensitivity to inhibition by the wild type (Figure 2.S2), while disrupting the biosynthetic genes in the wild-type background precluded inhibition of the ΔpSymAB strain (Figure 2.3A). Furthermore, no inhibition was observed in the presence of excess iron (Figure 2.3A). Taken together, these analyses revealed that inhibition of the ΔpSymA and ΔpSymAB strains was mediated through sequestering of environmental iron by the siderophore (Figures 2.3A, 2.S2).

The effect of this siderophore during soil growth was examined through co-inoculation of the wild type and the ΔpSymA strain in the same soil mesocosm (Figure 2.3B). Consistent with carbon being the growth-limiting nutrient and available iron being in excess, the presence of the wild type did not impact the growth of the ΔpSymA strain, and the ΔpSymA strain easily outcompeted the wild type. In line with this result, Loper and Henkels previously reported that the *Pseudomonas fluorescens* siderophore was not expressed during growth in bulk soil (Loper & Henkels, 1997). However, the synthesis/uptake of a siderophore may impact fitness in the rhizosphere (Loper & Henkels, 1997) and possibly affect symbiosis (Lynch *et al.*, 2001).

In addition to intra-species competition, inter-species competition is a major fitness determinant. We assessed the growth of the ΔpSymAB strain in the presence of three competing species: *Pseudomonas syringae*, *Streptomyces coelicolor*, and a soil-isolated *Aspergillus* species (Figure 2.4). The early exponential growth of the *S. meliloti* strain was not adversely impacted by any of the competing species, and the ΔpSymAB strain was able to establish a stable population in the presence of these species over the course of the 26-day assay. However, we observed that the maximum cell density attained by the ΔpSymAB strain was decreased ~10–20 fold when co-inoculated with a competitor, which may be attributed to inter-species competition for common nutrients and energy sources. As a whole, our data nonetheless suggest that neither pSymA nor pSymB are required for *S. meliloti* to effectively establish a long-term population or compete for resources with other species, and their loss does not render *S. meliloti*

susceptible to killing by these species.

2.5.6 Model of multipartite genome evolution

There are two general scenarios for the evolution of multipartite genomes. The schism hypothesis suggests that second chromosomes or chromids result from the split of an ancestral chromosome into two (Egan *et al.*, 2005). This has been suggested to have occurred in *Rhodobacter sphaeroides* (Choudhary *et al.*, 1997). Alternatively, the plasmid hypothesis suggests chromids result from the capture of a megaplasmid that subsequently acquires core genes from the chromosome (Egan *et al.*, 2005). The often-observed bias for essential genes to be located on one chromosome suggests that the plasmid hypothesis is more generally applicable (Egan *et al.*, 2005), and evidence suggests that the plasmid hypothesis is true in the case of *Vibrio*, *Agrobacterium*, *Rhizobium*, and *Sinorhizobium*, among others (Heidelberg *et al.*, 2000; Landeta *et al.*, 2011; Slater *et al.*, 2009; Wong *et al.*, 2002).

Several hypotheses exist about the function of multipartite genomes, and the evolution of multipartite genomes through the plasmid hypothesis; however, little experimental evidence has previously been reported to support these ideas. The presence of two replicons with distinct evolutionary histories (i.e. pSymA was a much more recent addition to the genome than pSymB) and characteristics (i.e. megaplasmid vs. chromid), and the presence of strains lacking one or both of these replicons makes *S. meliloti* an ideal system in which to experimentally develop a model describing the evolution of multipartite genomes. In the proposed model (Figure 2.5), as a first step a host cell

captures a plasmid that encodes genetic determinants allowing the cell to occupy a novel niche. Inhabiting this new environment puts an evolutionary pressure on the cell to obtain additional genetic material that provides a fitness benefit unique to this location. As genetic rearrangements of bacterial chromosomes are generally associated with a fitness cost (Rocha, 2006), this new genetic material is disproportionately acquired by the plasmid, resulting in a plasmid specialized for a specific niche. As plasmids are mobile elements, this enrichment of niche-specific traits is advantageous as it would promote plasmid retention following transfer to a new unichromosomal organism. From the host's view, while this plasmid is valuable in the new niche, its specialized nature means it provides little advantage in the original environment and is in fact a fitness burden due to its metabolic load. In *S. meliloti*, pSymA represents an example of a plasmid that encodes functions essential to a specialized niche (forming N_2-fixing root nodules with legumes) and yet imposes a fitness cost to cells growing in the species original environment (bulk soil). Thus, strains lacking pSymA grow more rapidly and outcompete wild-type *S. meliloti* in bulk soil (Figures 2.1, 2.3B), although ΔpSymA strains are unable to form root nodules (Barnett *et al.*, 2001; Yurgel *et al.*, 2013) and growth of the ΔpSymA strains may be inhibited by the wild type in specific environments (Figure 2.3A).

Over time, random translocations from the chromosome to a resident plasmid would result in the formation of a chromid, leading to an evolutionary pressure to maintain the chromid in all environments, including the species original niche where the loss of the replicon would otherwise be favoured. pSymB has had a long association with

the *S. meliloti* lineage and during this time has acquired core elements from the chromosome (diCenzo *et al.*, 2013; Slater *et al.*, 2009; Wong & Golding, 2003). Loss of this replicon adversely affects the growth of *S. meliloti* in bulk soil (Figure 2.1C) despite the reduced metabolic demand of no longer maintaining the chromid. However, the many metabolic functions dependent on pSymB largely appear to not be necessary for growth in bulk soil, and may be more relevant during growth in the rhizosphere, consistent with a niche-specialized role of this replicon. Overall, the phenotypic data reported here support a model where environmental specialization is a general driving force for multipartite genome evolution, with secondary replicons being enriched for functions unique to the new environment. Indeed, previous comparative genomics analyses (Galardini *et al.*, 2013; Harrison *et al.*, 2010; Slater *et al.*, 2009) presented evidence consistent with many of the core postulates of this model that were derived through experimental examination.

While this model addresses the evolution and primary role of secondary replicons, it is still unclear as to why this genome architecture persists, and why secondary replicons do not integrate into the chromosome. Integration has been postulated to have occurred in *Mesorhizobium* and *Bradyrhizobium* (Slater *et al.*, 2009), which carry a single large chromosome, despite having similar lifestyles to *Sinorhizobium* and *Rhizobium*, which have divided genomes. While it is possible that the presence of a divided genome is an evolutionarily transient event, this seems unlikely. As discussed in the introduction, several advantages have been ascribed to the presence of a multipartite genome that may promote its maintenance. Indeed, *S. meliloti* strains that carry all three replicons

recombined into one show a growth defect (Guo *et al.*, 2003), illustrating how genome structure and not just gene content affects the cell's phenotype. There may also be constraints on the ability of a chromid or megaplasmid to recombine into the chromosome. The origin and terminus of replication separate bacterial chromosomes into subdivisions that tend to be equal in size. Large insertions, such as the integration of a secondary replicon into the primary chromosome, would disrupt this balance and thus be unfavourable (Song *et al.*, 2003). Additionally, it has been suggested that there is an upper size limit of bacterial chromosomes, which could potentially preclude the integration of a large replicon into the chromosome (Slater *et al.*, 2009). Finally, plasmids and even chromids (Banfalvi *et al*., 1985), being mobile elements, can move into naïve cells, leading to further propagation of their DNA. As such, the fitness of the plasmid would be reduced following recombination into the main chromosome.

2.6 Materials and Methods

Except for the strains and plasmids constructed in this study, all other strains and plasmids have been previously described (Cowie *et al.*, 2006; diCenzo *et al.*, 2013; MacLellan *et al.*, 2005; Meade *et al.*, 1982; Milunovic *et al.*, 2014; Yuan *et al.*, 2006; Yurgel *et al.*, 2013) and are listed in Table 2.S1.

2.6.1 Growth conditions

Complex media included LB (10 gm/L tryptone, 5 gm/L yeast extract, 5 gm/L sodium chloride), LBmc (LB with 2.5 mM $MgSO_4$ 2.5 mM $CaCl_2$), and TY (5 gm/L tryptone, 2.5 gm/L yeast extract, 10 mM $CaCl_2$). For growth of *S. meliloti*, complex

media was supplemented with 2 mM $CoCl_2$. Minimal media included M9 (41 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.6 mM NaCl, 18.7 mM NH_4Cl, 4.1 mM biotin, 42 nM $CoCl_2$, 1 mM $MgSO_4$, 0.25 mM $CaCl_2$, 38 mM $FeCl_3$, 5 mM thiamine-HCl, 10 mM sucrose) and a 4-morpholinepropanesulfonic acid (MOPS) buffered medium (M9 with the phosphate buffer replaced with 40 mM MOPS and 20 mM KOH, with 2 mM KH_2PO_4). For the Phenotype MicroArray™ analysis, cultures were grown in M9 medium for the carbon and sulfur analyses, while strains were grown in MOPS medium for the nitrogen and phosphorus analyses. Additionally, the concentration of biotin was reduced to 40 nM for the analysis of sulfur metabolism. Unless stated otherwise, antibiotics were added to the following concentrations (μg/mL) for *S. meliloti* (*E. coli*), when appropriate: streptomycin 200 (N/A), spectinomycin 100 (100), tetracycline 5 (5), gentamicin 60 (20), neomycin 200 (N/A), kanamycin N/A (25), and chloramphenicol N/A (5). Antibiotic concentrations were halved for liquid media. *S. meliloti* was grown at 30°C and *E. coli* was grown at 37°C.

2.6.2 Genetic techniques

Common genetic techniques and manipulations were performed as previously described (Cowie *et al.*, 2006; Finan *et al.*, 1984; Sambrook *et al.*, 1989).

2.6.3 Growth curves

Overnight cultures were washed, resuspended, and diluted in fresh media. One hundred and fifty μL of diluted cultures (OD_{600} ~0.05, measured with a 1 cm wavelength) were inoculated in triplicate wells of 96-well plates. The edges of the 96-well plates were

taped to prevent moisture loss and the 96-well plates were incubated in a Tecan Safire for 48 hours at 30°C (+/- 1°C) with shaking. OD_{600} measurements were taken every 15 minutes. A Perl script was written to calculate averages, standard deviations, and generation times.

2.6.4 Phenotype MicroArray™

Phenotype MicroArray™ experiments were performed in Biolog plates (PM1, PM2A, PM3B, PM4A). Overnight cultures were washed, resuspended, and starved overnight in media free of the appropriate nutrient. Starved cultures were washed, resuspended, and diluted in fresh media, then 100 µL was inoculated into each well of the Biolog plates. Plates were incubated at 30°C for 5–7 days in a SteadyShake 757 Benchtop Incubator Shaker (Amerex Instruments, Inc.), with OD_{600} readings taken every 12–24 hours.

2.6.5 Soil preparation

In 2007, a 40 kg soil sample was obtained from an alfalfa field within a dairy farm near Guelph, Ontario, Canada, which does not apply pesticides, fertilizers, or herbicides. Large materials were manually removed, and following 9 days of drying, the soil was passed through a sieve to remove fragments larger than 2 mm. The soil was heat sealed in polyethylene freezer bags (FoodSaver; Jarden Corporation) as 100–300 gm samples. Soil samples were subjected to γ-irradiation (using ^{60}Co as a source) at the McMaster University Nuclear Reactor with a final dosage of 25.0 kGy (over a period of 54.3 hrs). As subsequent testing revealed the soil was not sterile, a second round of γ-irradiation at a

final dosage of 42.3 kGy was performed, and stored at -20°C until use. As a *Deinococcus* species still remained viable, soil samples were autoclaved once (123°C; 17 psig; 20 minutes) within a few days of beginning each growth assay. A chemical analysis of the soil was performed by the University of Guelph Laboratory Services Agricultural and Food Laboratory (Guelph, Ontario, Canada), and the results are presented in Table 2.S2.

2.6.6 Soil growth protocol

The γ-irradiation soil (47.62 gm, which was equal to 40 gm dry weight) was added to 500 mL screw-capped glass bottles (Gibco), autoclaved, and allowed to cool. Within a few days, *S. meliloti* strains were grown in LBmc or TY and cells were washed once with 0.85% NaCl and three times with de-ionized, autoclaved water (ddH_2O). Cells were resuspended in ddH_2O, adjusted to an OD_{600} of 1 and serial diluted to 10^{-4}, which equals approximately 2×10^5 CFU/mL. One mL of this dilution, together with an additional 1.38 mL ddH_2O, was added to each mesocosm. The resulting mesocosm contained 50 gm soil [40 gm dry weight with 20% moisture (wt/vol)], and $\sim 4 \times 10^3$ CFU gm^{-1}. Soil mesocosms were incubated at room temperature (22°C +/- 2°C) in the dark, and soil moisture content was maintained by the addition of ddH_2O every one to two weeks (at the rate of 48 µL daily).

To determine cell density, 0.62 gm samples were removed from each mesocosm into a 2 mL Eppendorf tube in a sterile environment. One mL of 0.85% NaCl was added to each tube and cells were re-suspended with vigorous vortexing. Soil particles were pelleted by vortexing for 1 minute at 60 *g*. The supernatant was serial diluted and plated

on LB or LBmc to determine CFU gm^{-1}.

Throughout co-inoculation experiments, when plating for CFU gm^{-1}, dilutions were plated on non-selective and selective media. When wild-type *S. meliloti* and the ΔpSymA strain were co-inoculated, each strain was inoculated to ~ 2×10^3 CFU gm^{-1}, and strains were differentiated based on growth with 10 mM trigonelline as the sole carbon source as only the wild type will grow. For co-inoculation of *S. meliloti* with *P. syringae*, *S. meliloti* was inoculated to ~ 2×10^3 CFU gm^{-1} while *P. syringae* was inoculated to ~ 8×10^2 CFU gm^{-1}, and CFU gm^{-1} was determined by plating on LB with streptomycin (*S. meliloti*) or LB with 20 μg/mL rifampicin (*P. syringae*). When co-inoculated with *S. coelicolor*, *S. meliloti* was inoculated to ~ 2×10^3 CFU gm^{-1} while *S. coelicolor* was inoculated with 2×10^3 spores gm^{-1}, and *S. meliloti* was selected for with streptomycin. Co-inoculation with *Aspergillus* was initiated with ~ 4×10^3 CFU gm^{-1} of *S. meliloti* and ~ 8×10^2 spores gm^{-1} of *Aspergillus*, and CFU gm^{-1} of *S. meliloti* was determined on LB with 100 μg/mL cycloheximide.

2.6.7 Isolation of a soil Aspergillus *species*

A 0.45 gm sample of non-sterilized soil was vigorously vortexed in 1 mL 0.85% saline, and dilutions plated on YPD medium (10 gm/L yeast extract, 20 gm/L peptone, 20 gm/L dextrose, 15 gm/L) with 50 μg/mL chloramphenicol. Based on morphology, an *Aspergillus* species was identified and streak purified on YPD. The *Aspergillus* was sporulated on LCA medium (Miura, 1970) for eight days at 30°C, and spores were re-suspended in PBS (0.8% NaCl, 0.02% KCl, 0.144% Na_2HPO_4, 0.024% KH_2PO_4) with

100 µg/mL streptomycin. Spores/mL were determined with a hemacytometer.

2.6.8 Siderophore assay

This assay was performed essentially as described previously for the identification of bacteriocins (Hirsch, 1979; Perrine-Walker *et al.*, 2009). Strains being tested for bacteriocin production were stabbed into TY agar plates and incubated at 30°C overnight. The next day, surface growth of the producer was largely removed using a sterile toothpick. Overnight cultures of strains being tested for bacteriocin sensitivity were diluted to an OD_{600} ~0.01 in TY, and 1 mL was mixed with 5 mL of TY with 240 µg/mL streptomycin or spectinomycin and 0.6% agar (giving a final concentration of Sm^{200} or Sp^{200} and 0.5% agar), and all 6 mL were poured onto the plates with the stabbed producers. The inclusion of streptomycin or spectinomycin was to prevent the producer from growing into the soft agar overlay. Plates were incubated at 30°C for two nights, following which zones of clearance were identified. When applicable, 150 µM $FeCl_3$ was added to the soft agar overlay.

2.6.9 Removal of pSymB

Previous work has indicated that there are only two single copy essential genes outside of the chromosome (*engA* and $tRNA^{arg}$), both located on pSymB (diCenzo *et al.*, 2013; Milunovic *et al.*, 2014). Previously, these two essential genes were integrated into the chromosome (diCenzo *et al.*, 2013). However, this integration included a neomycin resistance marker, and we wished to use neomycin as a selective marker during the process of removing pSymB. Thus, it was necessary to begin by constructing a neomycin

sensitive integration of the essential genes into the chromosome.

Based on how the integration was performed, there were two possible orientations of the genes following integration (Figures 2.S3A, 2.S3B), and using PCR we determined that the construct integrated as seen in Figure 2.S3B (RmP2686). Using the same procedure as previously followed (diCenzo *et al.*, 2013), we isolated a second strain with the orientation illustrated in Figure 2.S3A (RmP2711). The insertion in RmP2711 was transduced into a *metH*::Tn*5*-B20 strain selecting for spectinomycin resistance; *metH* is located ~ 5 kilobases upstream of the *engA*/tRNA insertion site (Figure 2.S3C). The resulting strain was the recipient in a transduction with a phage lysate prepared from RmP2686 (Figure 2.S3D). Colonies were selected for based on a $MetH^+$ phenotype on minimal medium and screened for neomycin resistance. Following the isolation of a neomycin sensitive colony, PCR was used to confirm that the genetic organization of the insertion was as expected (Figure 2.S3E). The insertion in this final strain, RmP2719, was stable and neomycin sensitive.

The two-gene operon, *smb21127*/*smb21128* (pSymB nt: 766,498 – 767,430), functions as an active toxin-antitoxin locus, although it is possible to delete this system with a low frequency (Milunovic *et al.*, 2014). Therefore, the deletion ΔB180 (pSymB nt: 635,940 – 869,645) was transduced into *S. meliloti* strain Rm2011, selecting for neomycin resistance. Subsequently, the chromosomal integration of *engA* and $tRNA^{arg}$ was transduced into this strain, selecting for spectinomycin resistance. The resulting strain, RmP3005, carried the essential pSymB genes on the chromosome as well as a 234

kb deletion that removed the only known active toxin-antitoxin system on this replicon.

The replication and segregation machinery of pSymB is encoded by the *repABC* operon (MacLellan *et al.*, 2005). An incompatibility factor, *incA*, is encoded within the *repB* and *repC* intergenic region; thus, pSymB cannot be stably co-inherited with another replicon carrying an exact copy of *incA* (MacLellan *et al.*, 2005). Thus, pTH1414 (pOT1 carrying the pSymB *incA* region) (MacLellan *et al.*, 2005) was introduced into *S. meliloti* RmP3005, and streptomycin/gentamicin resistant colonies were selected for on LB supplemented with 2 μM cobalt chloride to compensate for the loss of the major *S. meliloti* cobalt uptake ABC transporter (CbtJKL), which is pSymB-encoded (Cheng *et al.*, 2011). Recovered colonies were streak purified, and initially the inability to amplify six pSymB fragments using PCR was evidence that pSymB was indeed lost. One colony was inoculated in LBmc broth, serially diluted and plated on LB, and colonies were screened for loss of pTH1414 by patching for gentamicin sensitivity. A gentamicin sensitive colony was purified and stored as *S. meliloti* RmP3009.

In order to construct the strain lacking both pSymA and pSymB, the same procedure was followed as above, with two modifications. The starting strain lacked pSymA (Oresnik *et al.*, 2000). Additionally, the chromosomal integration of *engA* and $tRNA^{arg}$ from *S. meliloti* RmP2719 was transduced into this strain prior to the transduction of ΔB180. Following the removal of pSymB using incompatibility, and the subsequent loss of pTH1414, the resulting strain was frozen as *S. meliloti* RmP2917, which lacks both pSymA and pSymB.

2.8 Tables and Figures

Table 2.1. Nutrient sources supporting growth of *S. meliloti*.

	Number of substrates supporting growth			
Genotype	Carbon	Nitrogen	Phosphorus	Sulfur
Wild type[*†]	73	55	53	20
ΔpSymA	69[‡]	54[‡]	53	20
ΔpSymB	37[‡]	42[‡]	50	14
ΔpSymAB	23[§]	42	48[§]	10[§]

[*]Includes only those sources supporting strong growth of *S. meliloti*, by which I mean the wild type reached stationary phase during the duration of the experiment and the growth rate was sufficiently high to detect a consistent increase in density at all time points until stationary phase. For example, compounds like acetoacetate and 3-hydroxybutyrate were excluded.

[†]In several cases, the presence of either pSymA or pSymB improved growth.

[‡]For both carbon and nitrogen, the usage of one source required both pSymA and pSymB.

[§]In several cases, no growth was only observed when both pSymA and pSymB were removed.

Table 2.2. Carbon sources supporting growth of *S. meliloti*.[*]

Sugars	
Pentose	α-glucosides
D-Arabinose[†]	Sucrose
D-Ribose	Maltose[‡]
D-Xylose[‡]	Turanose
L-Arabinose	*D-Melezitose*
L-Lyxose	*D-Trehalose*
Hexose	*Maltotriose*
D-Fructose	*Palatinose*
D-Mannose	β-glucosides
L-Rhamnose	Arbutin
α-D-Glucose[‡]	D-Cellobiose
D-Galactose	Gentiobiose
D-Psicose	Salicin
D-Tagatose	β-Methyl-D-Glucoside
L-Fucose	α-galactoside
β-D-Allose	*D-Melibiose*
Glucose analog	*D-Raffinose*
3-Methyl Glucose	*Melibionic Acid*
Sugar phosphate	*α-Methyl-D-Galactoside*
D-Glucose-6-Phosphate	β-galactoside
Polyol	*3-0-β-D-Galactopyranosyl-*
Adonitol	*D-Arabinose*
D-Arabitol	*Lactulose*
D-Mannitol	*α-D-Lactose*
D-Sorbitol	*β-Methyl-D-Galactoside*
L-Arabitol	β-xyloside
Dulcitol	β-Methyl-D-Xyloside
Glycerol	Sugar amine
I-Erythritol	*D-Glucosamine*
Lactitol	Sugar amine derivative
M-Inositol	N-Acetyl-D-Galactosamine
Maltitol	*N-Acetyl-D-Glucosamine*
Organic acids	
Amino acid	Amino acid derivatives
L-Alanine	*Glycyl-L-Proline*
L-Arginine	*Hydroxy-L-Proline*
L-Aspartic Acid	Carboxylic Acids
L-Glutamic Acid	*D-Gluconic Acid*
L-Histidine	*D-Glucosaminic Acid*
L-Leucine	*β-Hydroxy Butyric Acid*
L-Lysine	Dicarboxylic Acids
L-Ornithine	*D,L-Malic Acid*
L-Proline	*Fumaric Acid*
γ-amino acid	*L-Malic Acid*
γ-Amino Butyric Acid	*Succinic Acid*
Other	
Nucleoside	Lactone
Uridine	*D-Ribono-1,4-Lactone*
Alkanolamine	Quarternary Ammonium
2-Aminoethanol	D,L-Carnitine

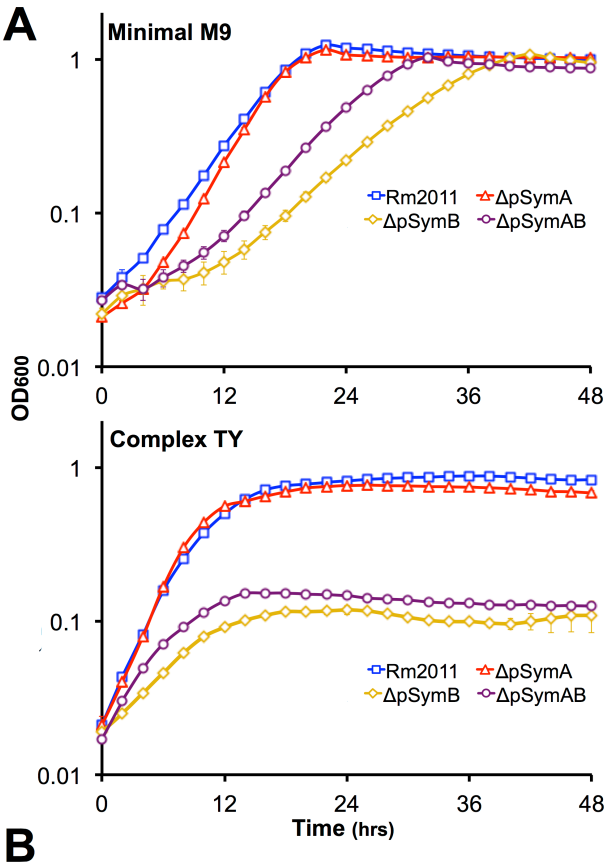
[*]Substrates requiring pSymA and/or pSymB are indicated in italics.

[†]Growth on this substrate is improved by the presence of pSymA.

[‡]Growth on this substrate is improved by the presence of pSymB.

Figure 2.1. The effect of the removal of pSymA and/or pSymB on the growth of *S. meliloti*.

The growth of *S. meliloti* was examined in M9 minimal medium (**A** – top panel), TY complex medium (**A** – bottom panel), or sterile bulk soil mesocosms (**C**). Data points represent averages from triplicate (**A**) or duplicate (**C**) samples. Error bars represent +/- one standard deviation from triplicate samples (**A**) or the range from duplicate samples (**C**). (**B**) Average generation times and standard deviations for each strain grown in M9 medium or TY medium, calculated from a total of six replicates from two independent experiments. Blue – wild type; red – ΔpSymA; yellow –ΔpSymB; purple – ΔpSymAB.



B

Average generation times (standard deviation)

Medium	Rm2011	ΔpSymA	ΔpSymB	ΔpSymAB
M9	3.3 (0.2)	2.8 (0.1)	5.4 (0.4)	4.4 (0.1)
TY	2.2 (0.1)	1.9 (0.1)	4.7 (0.3)	3.2 (0.2)

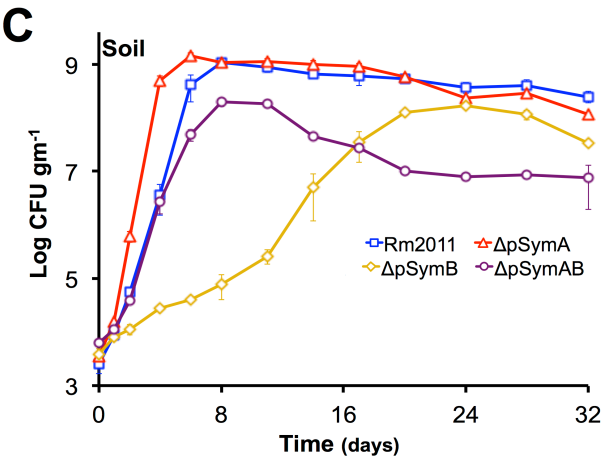


Figure 2.2. The decreased stationary phase density of strains lacking pSymB in bulk soil is due to carbon limitation and can be traced to two loci.

(**A** – top panel) The strain with a deletion of B116 (orange) shows a slight, but repeatable, decrease in stationary phase density relative to the wild type (dark blue). (**A** – bottom panel) The strains with deletions of B123 (dark teal) and B122 (light teal), a sub-region of B123, show a large decrease in stationary phase density relative to the wild type (dark blue). (**B**) Stationary phase soil populations were supplemented with either 15 mM glucose (solid bars) or 5 mM NH_4Cl, 2 mM KH_2PO_4, and 1 mM $MgSO_4$ (striped bars). Only the addition of a carbon source (glucose) stimulated further growth for the wild type (dark blue), the ΔpSymAB strain (purple), and strains with deletions of B123 (dark teal), which includes that entire B122 region, and B116 (orange). (**A** and **B**) Data points represent the average from duplicate experiments, while error bars represent the range from duplicate samples.

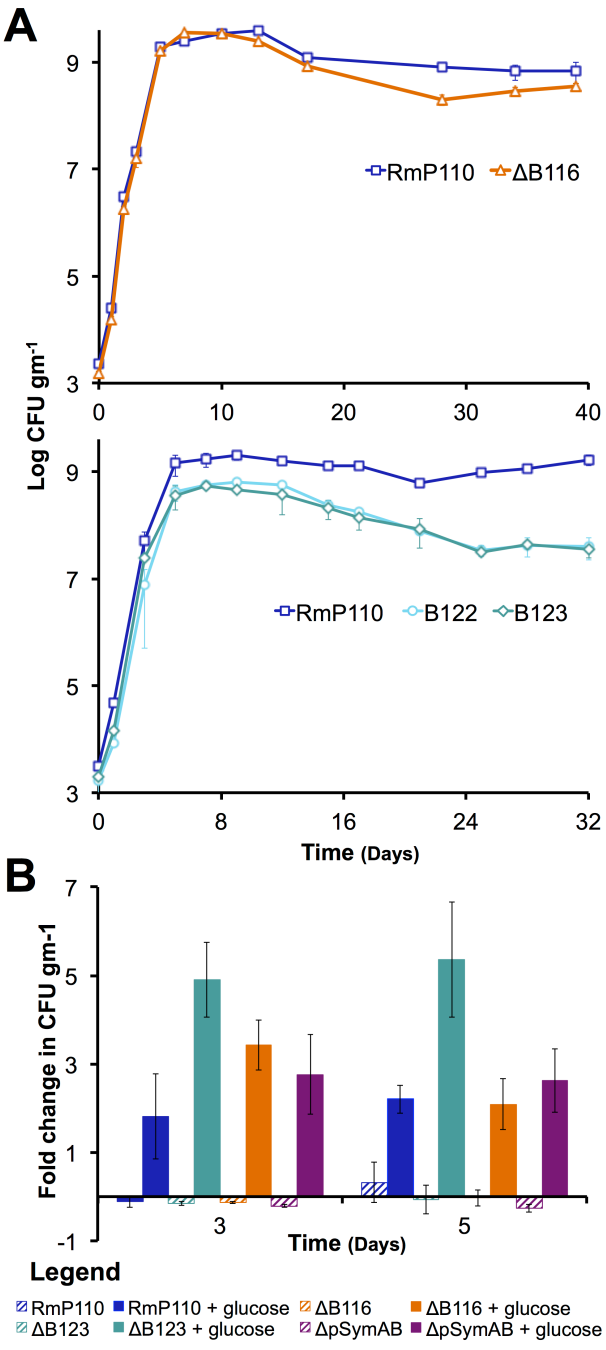

A
Log CFU gm-1
RmP110
ΔB116
RmP110
B122
B123
Time (Days)
B
Fold change in CFU gm-1
Time (Days)
Legend
RmP110
RmP110 + glucose
ΔB116
ΔB116 + glucose
ΔB123
ΔB123 + glucose
ΔpSymAB
ΔpSymAB + glucose

Figure 2.3. Environment specific growth inhibition by a pSymA-encoded siderophore.

Growth of the ΔpSymAB strain is inhibited by a siderophore produced by the wild type (left stab) when grown in TY medium (**A** – left panel), but not if the overlay is supplemented with 150 mM $FeCl_3$ (**A** – right panel). This inhibition fails to occur when the siderophore biosynthesis genes are knocked out in the wild type, as is the case in *S. meliloti* RmFL2950 (Cowie *et al.*, 2006) (right stab). (**B**) When co-inoculated in the same soil mesocosm, the ΔpSymA strain easily outcompetes the wild type, and the wild type fails to inhibit growth of the ΔpSymA strain. Data points represent averages of duplicate samples, while error bars represent the range from duplicate samples. Solid lines indicate growth pattern during co-inoculation, while dotted lines indicate growth pattern when individually inoculated. Blue – wild type; red – ΔpSymA.

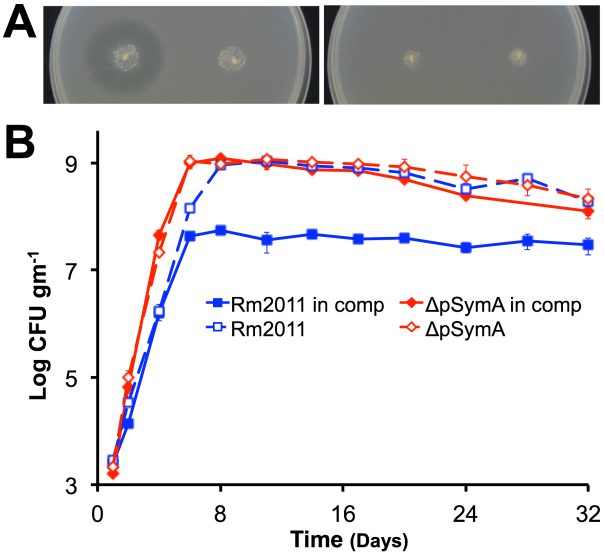

A
B
9
7
5
3
Log CFU gm-1
0
8
16
24
32
Time (Days)
Rm2011 in comp
Rm2011
ΔpSymA in comp
ΔpSymA

Figure 2.4. Effect of competing species on the growth of the *S. meliloti* ΔpSymAB strain in bulk soil mesocosms.

The *S. meliloti* ΔpSymAB strain was grown in bulk soil mesocosms in the presence of either an *Aspergillus* species, *Pseudomonas syringae*, or *Streptomyces coelicolor* and the growth of the ΔpSymAB strain was examined. The decreased stationary phase density during competition is presumably reflective of competition for nutrients and the reduced availability of nutrients due to usage by the competitor. On the other hand, there is a relative lack of effect on the early exponential growth of the ΔpSymAB strain, and it is able to establish a stable stationary phase population in the presence of the competing species. See materials and methods for details on experimental set-up. Purple – ΔpSymAB alone; teal – ΔpSymAB with a soil-isolated *Aspergillus* species; brown – ΔpSymAB with *P. syringae*; orange – ΔpSymAB with *S. coelicolor*. Data points represent single values.

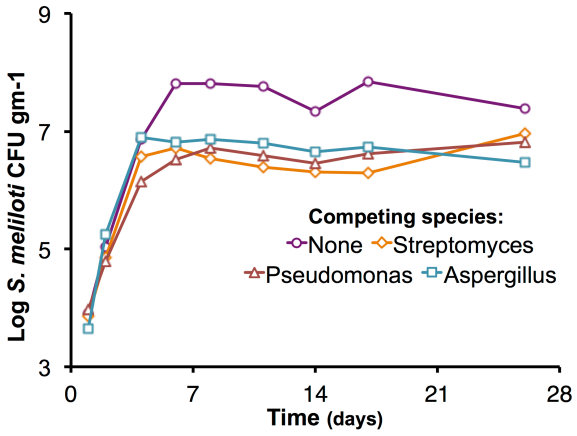
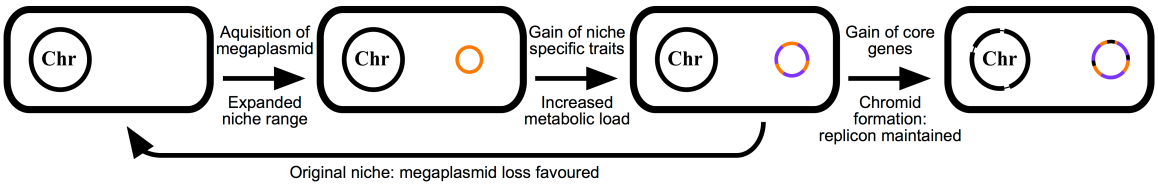

9
7
5
3
Log S. meliloti CFU gm-1
0
7
14
21
28
Time (days)
Competing species:
None
Streptomyces
Pseudomonas
Aspergillus

Figure 2.5. Schematic illustrating the described model of multipartite genome evolution and chromid formation.

The acquisition of a megaplasmid (orange) expands the niche range of the cell. Subsequently, this replicon accumulates horizontally acquired genes that provide a fitness advantage in this novel environment (purple). This results in a large metabolic load being associated with the megaplasmid, and its loss is favoured in the cell's original niche. However, gene transfer from the chromosome (black) renders the megaplasmid (now a chromid) indispensible in all environments. See the text 'model of multipartite genome evolution' for additional details.


Chr
Aquisition of megaplasmid
Expanded niche range
Chr
Gain of niche specific traits
Increased metabolic load
Chr
Gain of core genes
Chromid formation: replicon maintained
Chr
Original niche: megaplasmid loss favoured

2.9 Supplementary Materials

2.9.1 Supplementary Tables and Figures

Table 2.S1. Bacterial strains and plasmids.

Strain or Plasmid	Characteristics	Source
Sinorhizobium meliloti		
Rm1021	Wild type SU47 *str-21*; Sm^R	(Meade *et al.*, 1982)
Rm2011	Wild type SU47 *str-3*; Sm^R	M. Hynes
Rm5000	Wild type SU47 *rif-5*; Rif^R	(Finan *et al.*, 1984)
SmA818	Rm2011 cured of pSymA; Sm^R	(Oresnik *et al.*, 2000)
RmFL2878	RmP110, *rhtA*::pTH1522; Sm^R Gm^R	(Cowie *et al.*, 2006)
RmFL2950	RmP110, *rhbB*::pTH1522; Sm^R Gm^R	(Cowie *et al.*, 2006)
RmP110	Rm1021 with wild type *pstC*; Sm^R	(Yuan *et al.*, 2006)
RmP798	RmP110, ΔB122 (deletion of pSymB nt: 1,529,711 - 1,572,422), pTH1944; Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP801	RmP110, ΔB116 (deletion of pSymB nt: 1,256,503 - 1,307,752), pTH1944; Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP806	RmP110, ΔB123 (deletion of pSymB nt: 1,529,711 - 1,652,588), pTH1944; Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP963	RmP110, ΔA133 (deletion of pSymA nt: 1,281,754 - 1,348,238), pTH1944; Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP1615	*metH*::Tn5; methionine auxotroph; Sm^R Nm^R	Lab Collection
RmP1815	RmP110, ΔB123 from RmP806; Sm^R Nm^R Gm^R	This study
RmP2686	RmP110, *attB*::(pTH2750) *smb20996-engA smb21712* (tRNAarg) via *attP1*; Sm^R Sp^R Nm^R	(diCenzo *et al.*, 2013)
RmP2711	RmP110, *attB*::(pTH2750) *smb20996-engA smb21712* (tRNAarg) via *attP2*; Sm^R Sp^R Nm^R	This study
RmP2719	RmP110, *attR-smb21712-engA-smb20996*-ΩSmSp-*attL* via *attB*; Sm^R Sp^R	This study
RmP2745	RmP110, ΔB180 (635,940-869,642), pTH1944; Sm^R Nm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2778	SmA818, attR-smb21712-engA-smb20996-ΩSmSp-attL from RmP2917; Sm^R Sp^R	This study
RmP2805	RmP2778, ΔB180 from RmP2745; Sm^R Sp^R Nm^R	This study
RmP2917	Rm2011 cured of pSymA and pSymB; Sm^R Sp^R	This study
RmP3004	Rm2011, ΔB180 from RmP2745; Sm^R Nm^R	This study
RmP3005	RmP3004, *attR-smb21712-engA-smb20996*-ΩSmSp-*attL* from RmP2778; Sm^R Sp^R Nm^R	This study
RmP3009	Rm2011 cured of pSymB; Sm^R Sp^R	This study
Other		
Aspergillus species	Isolated from alfalfa farm soil	This study
Pseudomonas syringae pv. tomato DC3000	Wild type; Rif^R	R. Cameron
Streptomyces coelicolor M145	Wild type	M. Elliot
Plasmids		
pRK600	pRK2013 Nm^R::Tn9, RK2 *tra* genes; Cm^R	(Finan *et al.*, 1986)
pTH1414	pOT1 expressing pSymB *incα* incompatibility element; Gm^R	(MacLellan *et al.*, 2005)
pTH1522	A transcriptional reporter vector, pBR322 origin; Gm^R	(Cowie *et al.*, 2006)
pTH1944	A pBBRmcs-3 derivative expressing *flp*; Tc^R	(Milunovic *et al.*, 2014)
pTH2750	A pUX19 derivative with *attP*-ΩSmSp-*smb20996-engA-smb21712-attP*; Sm^R Sp^R Km^R	(diCenzo *et al.*, 2013)

Sm – streptomycin; Sp – spectinomycin; Gm – gentamycin; Nm – neomycin; Km – kanamycin; Tc – tetracycline; Cm – chloramphenicol; Rif – rifampicin.

Table 2.S2. Physiochemical properties of the soil used in this study.

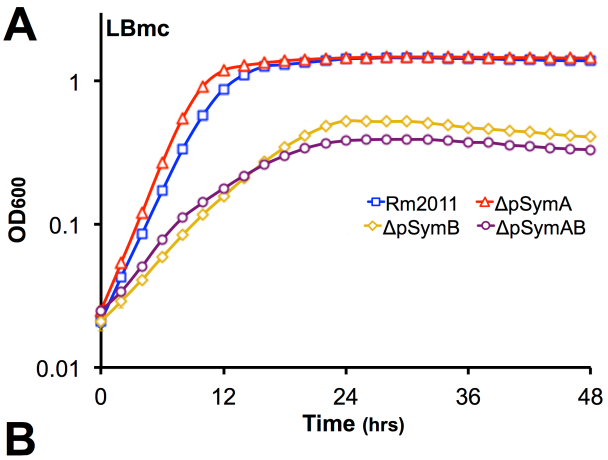
Characteristics	Soil Sample	
	Not autoclaved*	Autoclaved
Total C	4.94 % dry	4.20 % dry
Inorganic C	1.53 % dry	1.38 % dry
Organic C	3.41 % dry	2.82 % dry
pH	7.6	7.5
NH_4-N	46.2 mg/kg dry	51.8 mg/kg dry
NO_3-N	2.16 mg/kg dry	1.22 mg/kg dry
NO_2-N	0.042 mg/kg	0.094 mg/kg
N	0.35%	0.30%
P	146 mg/kg dry	127 mg/kg dry
Mg	494 mg/kg dry	NT†
K	383 mg/kg dry	NT
S	0.05 % dry	0.05 % dry
Ca	2713 mg/kg dry	NT
Fe	36.46 mg/kg dry	NT
Mn	172.5 mg/kg dry	NT
Zn	24.79 mg/kg dry	NT
% soil moisture	20.64%	20.65%
Organic matter	5.9 % dry	5.3 % dry
Very coarse sand	0.8 % w/w	0.5 % w/w
Coarse sand	4.2 % w/w	3.8 % w/w
Medium sand	7.0 % w/w	6.7 % w/w
Fine sand	16.6 % w/w	18.0 % w/w
Very fine sand	21.2 % w/w	22.1 % w/w
Sand	49.8 % w/w	51.1 % w/w
Silt	35.4 % w/w	34.3 % w/w
Clay	14.8 % w/w	14.6 % w/w
Texture	Loam	Loam
Organic matter	5.9 % dry	5.3 % dry
Gravel	0.0 % w/w	0.0 % w/w

*Analysis performed following γ–irradiation of soil samples by the University of Guelph Laboratory Services Agricultural and Food Laboratory.

†NT, not tested.

Figure 2.S1. The effect of the removal of pSymA and/or pSymB on the growth of *S. meliloti*.

(**A**) Growth curves of the wild type and replicon cured strains in LBmc. Data points represent averages from triplicate, and error bars represent +/- one standard deviation from triplicate samples. (**B**) Average generation times and standard deviations for each strain grown in LBmc medium, calculated from a total of six replicates from two independent experiments. Blue – wild type Rm2011; red – ΔpSymA; yellow – ΔpSymB; purple – ΔpSymAB.



B

Average generation times (standard deviation)

Medium	Rm2011	ΔpSymA	ΔpSymB	ΔpSymAB
LBmc	2.2 (0.1)	1.9 (0.1)	4.4 (0.5)	3.5 (0.1)

Figure 2.S2. Images showing the bacteriocin-like effect of the pSymA-encoded siderophore.

In all images, the stabbed strain is the wild-type *S. meliloti* Rm5000 (Cowie *et al*., 2006), which is a rifampicin resistant and streptomycin sensitive derivative of the same nodule isolate of *S. meliloti* Rm2011 (streptomycin resistant). The wild type is able to inhibit the growth of strains lacking pSymA, but not a strain lacking just pSymB. The sensitivity of the pSymA cured strain is conferred by the inability to uptake the siderophore, as is seen by the sensitivity of *S. meliloti* RmFL2878 (*rhtA*::pTH1522), which is a siderophore uptake mutant (Cowie *et al.*, 2006; Lynch *et al.*, 2001).

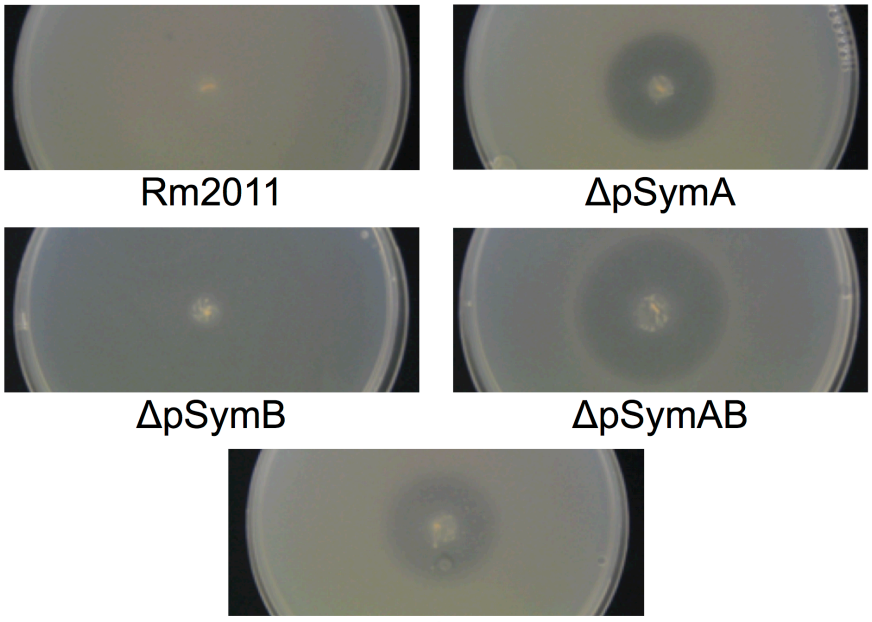
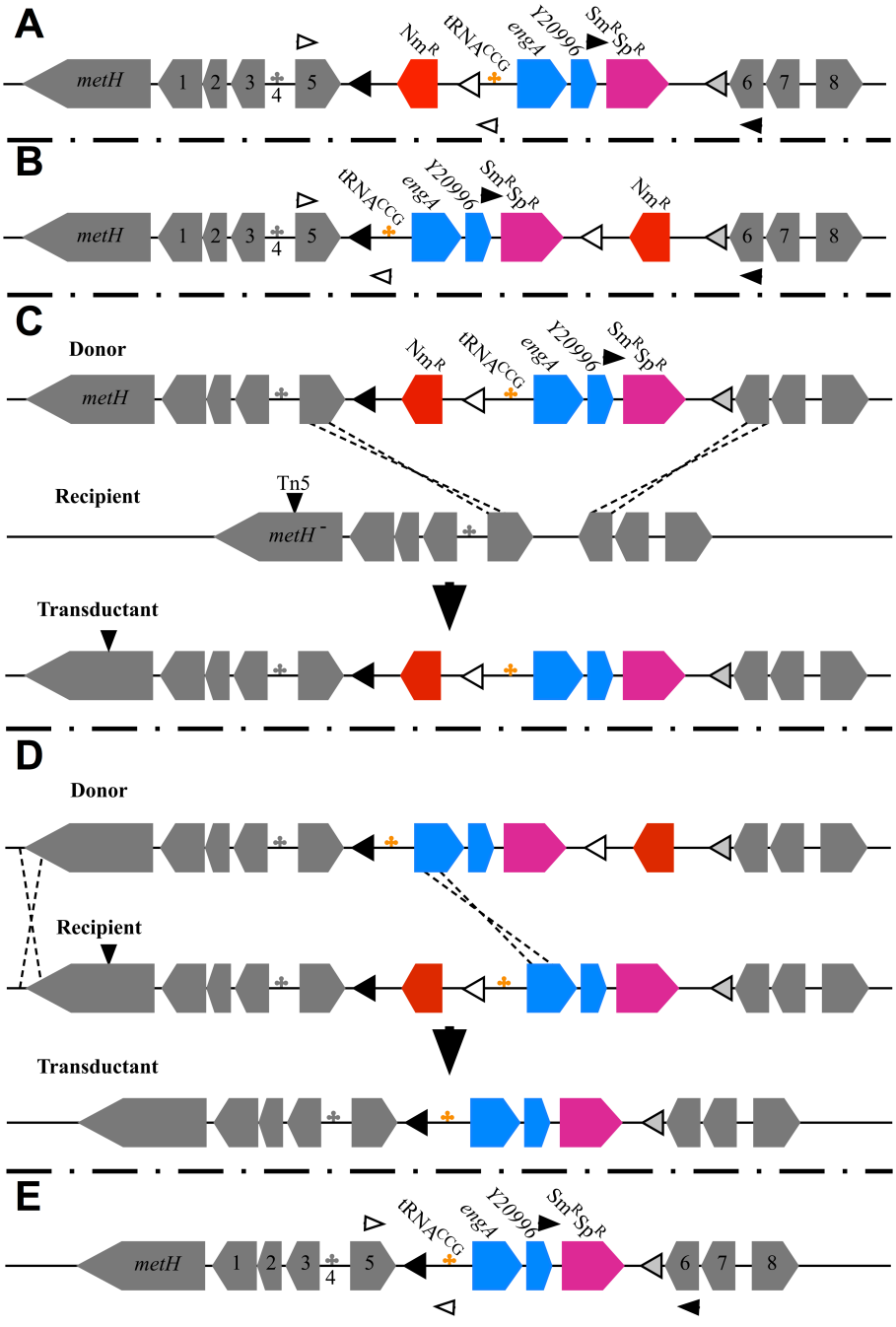

Rm2011
ΔpSymA
ΔpSymB
ΔpSymAB
RmFL2878 (siderophore uptake mutant)

Figure 2.S3. A diagrammatic representation of how the neomycin sensitive integration of the pSymB essential genes into the chromosome was constructed.

(**A**) and (**B**) represent the two possible genetic organizations following integration into the chromosome, while (**E**) represents the final genetic organization. The arrows represent the approximate location of primers able to differentiate between each of the three organizations; the open-ended arrows amplify a product in (**B**) and (**E**), while the close-ended arrows amplify a product in (**A**) and (**E**). (**C**) and (**D**) illustrate the two transductions involved in the creation of a neomycin sensitive integration. The Tn*5* loss-of function insertion in *metH* rendered the strain unable to grow on minimal medium not supplemented with methionine. The *attR* sequence is indicated by the black arrowheads, the *attL* sequence by the light gray arrowheads, and the *attP* sequence by the white arrowheads. 1 – *pmi*; 2 – *Y03111*; 3 – *mak*; 4 – $tRNA^{CCG}$; 5 – *Y03108*; 6 – *Y03107*; 7 – *Y03106*; 8 – *dxr*.


A
metH
1
2
3
4
5
NmR
tRNACCG
engA
Y20996
SmRSpR
6
7
8
B
metH
1
2
3
4
5
tRNACCG
engA
Y20996
SmRSpR
NmR
6
7
8
C
Donor
metH
NmR
tRNACCG
engA
Y20996
SmRSpR
Recipient
Tn5
metH-
Transductant
D
Donor
Recipient
Transductant
E
metH
1
2
3
4
5
tRNACCG
engA
Y20996
SmRSpR
6
7
8

2.9.2 Supplementary Data Sets

All supplementary data sets can be found online at:

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004742#s4

Data set 2.S1. Carbon Phenotype MicroArray™ results. Results of the carbon utilization Phenotype MicroArray™ experiments. The ability of each strain to grow (Yes) or not grow (No) with each tested carbon source is indicated.

Data set 2.S2. Nitrogen Phenotype MicroArray™ results. Results of the nitrogen utilization Phenotype MicroArray™ experiments. The ability of each strain to grow (Yes) or not grow (No) with each tested nitrogen source is indicated.

Data set 2.S3. Phosphorus Phenotype MicroArray™ results. Results of the phosphorus utilization Phenotype MicroArray™ experiments. The ability of each strain to grow (Yes) or not grow (No) with each tested phosphorus source is indicated.

Data set 2.S4. Sulfur Phenotype MicroArray™ results. Results of the sulfur utilization Phenotype MicroArray™ experiments. The ability of each strain to grow (Yes) or not grow (No) with each tested sulfur source is indicated.

CHAPTER 3. METABOLIC MODELLING REVEALS THE SPECIALIZATION OF SECONDARY REPLICONS FOR NICHE ADAPTATION IN *SINORHIZOBIUM MELILOTI*

Citation: diCenzo GC, Checcucci A, Bazzicalupo M, Mengoni A, Viti C, Dziewit L, Finan TM, Galardini M, Fondi M. 2016. Metabolic modeling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. Nat Commun. 7: 12219.

3.1 Preface

The previous chapter described a general model for the function and evolution of multipartite genomes. The data presented in Chapter 2 supported that the loss of young secondary replicons may be favoured in some environments and that secondary replicons become more integrated into core metabolism through co-evolution with the chromosome. The data were also consistent with secondary replicons being specialized for adaptation to particular niches but did not directly support this. It was shown that the metabolic function associated with the *S. meliloti* secondary replicons were not relevant in bulk soil suggesting they are niche specialized; however, it was not actually shown that the functions on these replicons improve fitness in any environment. The work presented in the current chapter was designed to address this outstanding question.

An *in silico* reconstruction of the *S. meliloti* metabolism was performed, resulting in a working metabolic model encompassing ~ 25% of the entire *S. meliloti* proteome. Additionally, *in silico* nutritional approximations of the bulk soil and rhizosphere soil environments were developed, and a nutritional approximation of the nodule adapted from previous work. Following validation of the metabolic model, growth (biomass production) was simulated in the bulk soil and rhizosphere environments, and a N_2-fixing symbiosis simulated in the nodule environment. Differences in the metabolic network of the cell were detected in each environment, showing that different reactions, and hence different genes, are of greater or lesser importance in the different environments. A single and double gene deletion analysis was then performed to examine the contribution

of each gene to fitness of the cell during simulated growth in each environment, allowing for a comprehensive analysis of the contribution of each of the *S. meliloti* replicons (chromosome, pSymB chromid, pSymA, megaplasmid) to growth in each environment.

The results of the gene deletion analysis suggested that the chromosome was equally important for growth in both bulk soil and the rhizosphere, but much less important for N_2-fixation in the nodule. Thus, it was concluded that the chromosome is an undifferentiated replicon that encodes the core metabolic machinery required for soil growth regardless of the specific environment. In contrast, it was observed that the contribution of the pSymB chromid to growth in the rhizosphere was significantly greater than the contribution to growth in bulk soil, consistent with the results in Chapter 2. Hence, it was concluded that the metabolic capabilities associated with pSymB are specialized for adaptation to a rhizosphere environment. Finally, pSymA only contributed functions relevant to the symbiosis, consistent with its specialization for the symbiosis. Together, these data provides novel support for the theory of multipartite genome biology and evolution as outlined in Chapter 2 by showing a niche specialization of the metabolism encoded by each replicon in a multipartite genome.

3.2 Abstract

The genome of about 10% of bacterial species is divided among two or more large chromosome-sized replicons. The contribution of each replicon to the microbial life cycle (e.g. environmental adaptations and/or niche switching) remains unclear. We report a genome-scale metabolic model of the legume symbiont *Sinorhizobium meliloti* that was

integrated with carbon utilization data for 1500 genes with 192 carbon substrates. Growth of *S. meliloti* was modelled in three ecological niches (bulk soil, rhizosphere and nodule) with a focus on the role of each of its three replicons. Clear metabolic differences during growth in the tested ecological niches and an overall reprogramming following niche switching was observed. *In silico* examination of the inferred fitness of gene deletion mutants suggested that secondary replicons evolved to fulfill a specialized function, particularly host-associated niche adaptation. Thus, genes on secondary replicons might potentially be manipulated to promote or suppress host interactions for biotechnological purposes.

3.3 Introduction

Recent years have witnessed a growing attention toward the ecological and evolutionary implication of the multiple replicon bacterial genome (diCenzo *et al.*, 2014; Galardini *et al.*, 2013; Harrison *et al.*, 2010; Landeta *et al.*, 2011) that is present in about 10% of sequenced bacterial genomes (Harrison *et al.*, 2010). This genome architecture is common in the proteobacterial species that interact with a host and are of importance to the human population (Harrison *et al.*, 2010; Landeta *et al.*, 2011), including crop plant symbionts (e.g. *Sinorhizobium* and *Rhizobium*), plant pathogens (e.g. *Agrobacterium*), and animal and human pathogens (e.g. *Brucella*, *Burkholderia*, and *Vibrio*). As the bacterial genome is non-randomly organized (Rocha, 2008), it is proposed that this genome organization was shaped by selective pressures to facilitate improved host interactions and niche adaptation. Though it is well-established that secondary replicons

often carry genetic determinants essential to colonize a novel environment, for example virulence or symbiotic genes, such genes often account for only a small proportion of these replicons (diCenzo *et al.*, 2016b; Johnson & Nolan, 2009). The majority of the genes on a secondary replicon are not directly essential to colonize a specific environment, and the adaptive function of these genes and why they are localized on a secondary replicon remains unclear. Several recent studies have provided evidence consistent with the secondary replicons in a multipartite genome encoding environment-specific fitness promoting but non-essential functions (Becker *et al.*, 2004; diCenzo *et al.*, 2014; Galardini *et al.*, 2013; 2015; Morton *et al.*, 2013; Romanchuk *et al.*, 2014; Xu *et al.*, 2003). However, none of these studies demonstrated that secondary replicons indeed carry environment-specific fitness determinants, thus serving as reservoirs for niche-specific functions.

Sinorhizobium meliloti is a N_2-fixing endosymbiont of legume species that has recently become a model organism for the study of bacterial multipartite genome function and evolution. All sequenced *S. meliloti* genomes contain at least three large replicons (the primary chromosome, the pSymB chromid, and the pSymA megaplasmid), with some strains hosting additional small accessory plasmids (Galardini *et al.*, 2011; 2013; Galibert *et al.*, 2001). *S. meliloti* experiences a complex life cycle and successfully colonizes three distinct niches. Two of these are bulk soil and rhizosphere soil (i.e. the soil directly influenced by the plant root system), which are quite different environments with the rhizosphere generally considered to be a nutritionally richer environment due to

plant root exudates (Hinsinger *et al.*, 2009). The third niche inhabited by *S. meliloti* is the legume root nodule. *S. meliloti* can induce root nodule formation in certain legumes and within nodules the bacteria differentiate into N_2-fixing bacteroids. Manipulation and optimization of this agriculturally and ecologically important symbiosis is an ultimate goal of the rhizobial community (Geddes *et al.*, 2015; Oldroyd & Dixon, 2014; Remigi *et al.*, 2016). Effectively doing so will require a complete understanding of the evolution (Remigi *et al.*, 2016), genetics (diCenzo *et al.*, 2016b), and metabolism of the organism in both rhizosphere and nodule environments, and the corresponding metabolic shifts.

Here, we combined a genome-scale metabolic network reconstruction of the *S. meliloti* genome, flux balance analysis (FBA), and growth phenotype data for 11 large-scale *S. meliloti* deletion mutants to examine the metabolic changes accompanying the shifts between bulk soil, rhizosphere, and nodule environments. We used an *in silico* approach to predict the phenotypes resulting from the deletion of 1,575 *S. meliloti* metabolic genes, estimate the fitness contribution of each replicon within each environment, and thus provide insight into the evolution of multipartite genomes.

3.4 Results

3.4.1 Reconstruction of a S. meliloti *genome-scale metabolic model*

As described in the Supplementary Methods (Section 3.9.2.1), an *in silico* representation of the metabolism of *S. meliloti* was developed, and the final model that was termed iGD1575 contained 1575 genes, 1825 reactions, and 1579 metabolites. iGD1575 accounts for 25.4% of the protein coding genes in the *S. meliloti* genome, and

the other main features of the model are listed in Table 3.1. COG analyses confirmed that the gene functional biases of each replicon are accurately represented in iGD1575 (Supplementary Figure 3.S1) (Galardini *et al.*, 2015; Galibert *et al.*, 2001). The iGD1575 model encompasses 529 of the 565 genes present in iHZ565, a previously reported *S. meliloti* small metabolic model (Zhao *et al.*, 2012). The remaining 31 genes were not added to iGD1575 as experimental data were inconsistent with their annotation, we felt their annotation too general to have high confidence in the enzymes' substrates/products, or the associated reaction involved a metabolite not present in any other reaction in the model and thus the reaction would never carry flux during FBA (Supplementary Table 3.S1). Comparison of the number of genes in iGD1575 to that of other available rhizobial and non-rhizobial models (Resendis-Antonio *et al.*, 2007; 2012; Schellenberger *et al.*, 2010; Zhao *et al.*, 2012) showed that iGD1575 is currently one of the largest metabolic reconstructions of a bacterial genome. Additionally, iGD1575 is the first metabolic model capable of representing metabolism of both a symbiotic and free-living rhizobium.

3.4.2 Quantitative validation of iGD1575

Previous work (Fuhrer *et al.*, 2005) has shown that *S. meliloti* transports glucose into the cell at a rate of 2.41 mmol $hour^{-1}$ (gram cellular dry $weight)^{-1}$. When glucose is provided to the iGD1575 model as the sole carbon source at the experimentally determined rate, a specific growth rate of 0.33 $hour^{-1}$ is predicted, which is consistent with our experimentally derived growth rate of 0.31 $hour^{-1}$ (standard deviation [SD] 0.01) for *S. meliloti* grown with glucose. Similarly, *S. meliloti* transports succinate into the cell

at a rate of 6.25 mmol $hour^{-1}$ $(gram\ protein)^{-1}$ (Yurgel *et al.*, 2000). Providing succinate as the sole source of carbon to iGD1575 at the experimentally determined value led to a predicted specific growth rate of 0.28 $hour^{-1}$, similar to the experimentally derived rate of 0.25 $hour^{-1}$ (SD 0.03) for *S. meliloti* grown with succinate. Measuring the amount of phosphate remaining in the spent growth medium following growth of *S. meliloti* indicated that 63.7 μM (SD 6.7) and 39.8 μM (SD 1.5) of phosphate was used per mM of glucose and succinate, respectively. These experimental values were relatively consistent with the phosphate usage values predicted by iGD1575 of 72.7 μM and 48.5 μM per mM of glucose and succinate, respectively.

Little experimental flux data has been reported for *S. meliloti*; however, flux measurements for 22 central carbon metabolic reactions when *S. meliloti* is grown with glucose as the sole source of carbon have been reported (Fuhrer *et al.*, 2005). Not surprisingly, the experimentally determined fluxes did not match well with the iGD1575 derived values. This is because the specific growth rate of *S. meliloti* was just 0.17 $hour^{-1}$ in the Fuhrer *et al*. study, indicating that *S. meliloti* was grown in sub-optimal conditions that presumably affected the flux distribution. Nevertheless, if the flux through these 22 central carbon metabolic reactions in iGD1575 was set as experimentally determined by Fuhrer *et al*., the predicted specific growth rate was reduced to 0.159 $hour^{-1}$, in line with the 0.17 $hour^{-1}$ reported by Fuhrer *et al*. The good relationship between flux distribution and specific growth rate, and the strong ability of iGD1575 to predict growth rate and phosphate usage when grown with glucose or succinate, suggest that the flux distributions

predicted by iGD1575 should represent quantitatively accurate estimations.

3.4.3 iGD1575 captures the metabolic capacity of S. meliloti

The ability of *S. meliloti* to grow with various carbon and nitrogen sources has been well studied by means of the Phenotype MicroArray™ (Biolog) technology (Biondi *et al.*, 2009; diCenzo *et al.*, 2014; Spini *et al.*, 2015). These previously published studies were used to guide model expansion and refinement during the curation process. Once all of the manual curation of iGD1575 was complete, FBA illustrated that the final model could accurately predict the ability of *S. meliloti* to produce, or not produce, biomass (as defined in Supplementary Table 3.S2) on 85% (138/162) and 75% (64/85) of the tested carbon and nitrogen substrates, respectively, for which the ability of *S. meliloti* to utilize, or not, these compounds is known (Figure 3.1, Supplementary Data 3.1). Most of the discrepancies between the experimental data and the iGD1575 growth prediction were false negatives (71% and 95% for growth with carbon and nitrogen substrates, respectively). These represent compounds that *S. meliloti* can metabolize but the model could not use for the production of biomass, likely representing gene annotation gaps in our knowledge of *S. meliloti* that will serve as targets for future research. The predictive power of bacterial metabolic models reported in previous studies (Bartell *et al.*, 2014; Fondi *et al.*, 2014; Schatschneider *et al.*, 2013) are similar to that reported here for iGD1575. Hence, iGD1575 is at least as good as other current genome scale metabolic reconstructions at representing the organism's metabolic capabilities. This suggests that iGD1575 effectively captures the metabolic capacity of *S. meliloti* and can validly be used

to model metabolism in nutritionally diverse environments.

3.4.4 Carbon growth phenotypes of S. meliloti *deletion mutants*

Carbon utilization phenotypes for a subset of large-scale pSymB deletion mutants (Milunovic *et al*., 2014) that cumulatively remove ~ 1.65 Mb (98%) of pSymB (Supplementary Figure 3.S2) were determined using PM1 and PM2A Biolog plates. This screen effectively generated a carbon utilization dataset for ~ 1500 pSymB genes. Overall, growth was observed with 76 carbon substrates, and a total of 43 growth phenotypes were observed for the deletion mutants (Table 3.2, Supplementary Figures 3.S3 and 3.S4, and Supplementary Data 3.S1 and 3.S2). In the process of developing and validating iGD1575, an *in silico* representation of the same experiment was performed, and where possible, the model was updated to fix discrepancies between the experimental and *in silico* results. Following this integration of the Phenotype MicroArray^{TM} dataset with the metabolic reconstruction, very good agreement between the experimentally observed results and the *in silico* simulations was observed (Table 3.S2 and Supplementary Data 3.S1). *In silico* simulations did not predict any 'no growth' phenotypes that were not experimentally observed, and 23 of the 36 (63.9%) experimentally observed phenotypes for compounds that support growth of iGD1575 were replicated *in silico*. Some of the discrepancies between the experimental and *in silico* data represent gaps in our knowledge of catabolic pathways in *S. meliloti*, while other phenotypes may occur for non-metabolic reasons and therefore not give a phenotype *in silico*. For example, the *S. meliloti* deletion mutant ΔB154 is more sensitive

to cobalt chelation than the wild type (Cheng *et al.*, 2011), and the lack of growth in wells with L-histidine or D-glucosamine may simply reflect cobalt chelation (Lerivrey *et al.*, 1986; Malhotra *et al.*, 1986).

In addition to model refinement, integrating the mutant phenotype data with iGD1575, the DuctApe software (Galardini *et al.*, 2014), the *S. meliloti* genome annotation (Galibert *et al.*, 2001), and an ABC transporter induction study (Mauchline *et al.*, 2006) allowed for the prediction of novel carbon catabolic loci. One example compound is the monosaccharide psicose. Our analysis suggests that psicose is transported by the SupABCD (Smb20484-Smb20487) ABC transporter, and then converted to fructose by an isomerase encoded by *smb20488*. A second example is D-galactosamine which, as elaborated on in Section 3.9.1.1, we predict is transported by the Smb21216, Smb21219-Smb21221 transporter and potentially the Smb21135-Smb21138 transporter, and then metabolized by Smb21217, Smb21218, Smb21373, and Smb21374.

3.4.5 Rhizosphere colonization required a metabolic refinement

The metabolic shifts experienced by *S. meliloti* during transition between bulk soil, the rhizosphere, and the nodule were modelled using *in silico* representations of the nutritional composition of each environment. These took into account the relative ratios of each component in the different environments and the development of these environments are described in the Supplementary Materials (Section 3.9.2.5). In the bulk soil and rhizosphere environments, the model was optimized for the production of biomass as defined in Supplementary Table 3.S2, whereas in the nodule environment the

model was optimized for production of an effective N_2-fixing symbiosis as defined previously (Zhao *et al.*, 2012). The optimal flux patterns in each of the three niches were obtained using FBA and visualized with iPath (Figure 3.2) (Yamada *et al.*, 2011).

The metabolic network appears globally similar in both the bulk soil and rhizosphere environments (Figure 3.2 and Table 3.3), although many subtle differences were present when reaction specific parameters were examined (Figure 3.2, Supplementary Data 3.S3). Despite good correlation between the $\log_{10}$ of the absolute flux through a given reaction that was active in both environments (p-value < 0.01 using a Spearman's Rank Order Correlation test, median [absolute residual / observed] = 0.09; Supplementary Figure 3.S5A), ~ 20% of the reactions showed at least a 50% change in flux between the two environments while an additional 6% switched directions. Similarly, the effect on fitness [defined as the flux through the objective function (biomass formation or symbiosis) in the mutant relative to the flux through the objective function in the wild type] of individual reaction deletions displayed a strong correlation between the two environments (p-value < 0.01 using a Spearman's Rank Order Correlation test, R^2 = 0.95; Supplementary Figure 3.S5B). Nevertheless, ~ 7% had at least a 10% variation in fitness effect between environments and ~ 4% were essential in just one of the two niches. Interestingly, optimal growth in the rhizosphere required a greater repertoire of metabolic reactions as illustrated by the increased number of reactions required for maximal fitness. Additionally, ~ 13% of the active reactions were specific to just one of the environments. The reactions whose fluxes were considered to

change between growth in bulk soil and the rhizosphere were further validated through a procedure involving flux variability analysis as detailed in Section 3.9.1.2.

Few outstanding biases (p-value < 0.01 using a Pearson's Chi Squared test) were seen in the COG annotations of the genes associated with reactions whose flux or fitness contribution was dependent on the soil environment. This indicated that the reactions important in the rhizosphere were biologically similar to, but functionally distinct from, the reactions important in bulk soil. However, coenzyme transport and metabolism (COG H) and cell wall, membrane, and envelope biogenesis (COG M) were more important in the rhizosphere than in bulk soil. This possibly reflects different coenzyme requirements for the metabolic pathways active in the two environments and the increased succinoglycan content of *S. meliloti* in the rhizosphere that is necessary to facilitate root biofilm formation. Lipid transport and metabolism (COG I) was over-represented in the bulk soil, perhaps due to the over-abundance of ketogenic amino acids in bulk soil. At the pathway level, only a few changes could not be explained by differences in the nutritional composition and biomass objective functions (Supplementary Data 3.S3). For example, the importance of various carbon catabolic pathways and amino acid biosynthetic pathways reflected the abundance of the sugars and amino acids in each environment. This analysis also revealed that *S. meliloti* relies more heavily on glycolytic substrate during growth in bulk soil but on gluconeogenic substrate in the rhizosphere, which was consistent with the high concentration of organic acids in the rhizosphere. The increased gluconeogenic flux and the increased flux through the pantothenate and

Coenzyme-A biosynthesis pathways in the rhizosphere was also consistent with an increased sugar demand for the rhizosphere specific Nod factor production and increased exopolysaccharide biosynthesis (Carlson *et al.*, 1994; Schmid *et al.*, 2015). Finally, the urea cycle contributed more to cellular fitness in bulk soil than in the rhizosphere.

3.4.6 Complex metabolic reprogramming is associated with symbiosis

The rhizosphere to nodule transition was accompanied with much more pronounced metabolic changes than the bulk soil to rhizosphere transition (Figure 3.2, Supplementary Data 3.S4). Half as many reactions carried flux in the nodule than in the rhizosphere, with ~ 61% of rhizosphere reactions off in the nodule and ~ 22% of active nodule reactions off in the rhizosphere. This overall decrease in metabolic reactions active in the nodule is consistent with the global transcriptional down-regulation in differentiated bacteroids (Barnett *et al.*, 2004; Capela *et al.*, 2006). For reactions active in both environments, there was a significant correlation (p-value < 0.01 using a Spearman's Rank Order Correlation test; Supplementary Figure 3.S5C) in the $\log_{10}$ of the absolute flux values, but the dispersion of the observed values from the regression line was high (median [residual / observed] = 1.48). Approximately half of the common flux carrying reactions displayed at least 50% more flux in one of the environments and a further 12% switched directions. Additionally, little correlation was observed between the fitness effects of individual reaction deletions in the two environments, (R^2 = 0.03; Supplementary Figure 3.S5D). Of the active reactions, ~ 38% were essential specifically in one environment, while the deletion of another 12% gave fitness effects ≥ 10%

different in the two niches. The reactions whose fluxes were considered to change between growth in the rhizosphere and symbiosis in the nodule were further validated through a procedure involving flux variability analysis as detailed in Section 3.9.1.2.

A clear shift in the functional annotation of genes associated with the variable reactions was observed. Functions associated with generating the large amount of energy required for nitrogen fixation displayed increased importance in the nodule: e.g. energy production and conversion (COG C) and coenzyme transport and metabolism (COG H). On the other hand, the lack of growth of the differentiated bacteroids not surprisingly rendered biomass component biosynthesis (COGs E,F,L,M,I,J) less important. A few additional interesting observations were noted by looking at pathway level changes (Supplementary Data 3.S4). Flux through the Kreb's Cycle and AMP synthesis were increased, presumably to accommodate the high ATP demand of nitrogenase. Glycolysis was less important in the nodule than in free-living cells, consistent with the lack of glycolysis specific enzymes detected in the *S. meliloti* nodule proteome (Djordjevic, 2004). Flux through various pathways producing compounds (including steroids, glutathionine, vitamin B6, and heme) required for a successful symbiosis was observed, and in most cases these changes were supported by previously published proteomic, RNA-seq, or induction studies (Djordjevic, 2004; Prell *et al.*, 2002; Roux *et al.*, 2014). Flux through the non-oxidative pentose phosphate pathway, which is poorly studied in *S. meliloti* (Geddes & Oresnik, 2014), was also increased, consistent with the detection of two enzyme of this pathway in the *S. meliloti* nodule proteome (Djordjevic, 2004) and the

need for *S. meliloti* to synthesize sugars for biosynthesis (Udvardi & Poole, 2013).

3.4.7 S. meliloti *replicons encode niche specific metabolism*

We performed comprehensive, replicon specific *in silico* single and double gene deletion analyses to determine the contribution of the three *S. meliloti* replicons to the overall fitness of *S. meliloti* in each of the tested environments (Supplementary Table 3.S3, Figure 3.3, and Supplementary Figure 3.S6). The use of a double gene deletion analysis was intended to account for functionally redundant gene pairs that would mask phenotypes during the single gene deletion analysis (diCenzo & Finan, 2015; diCenzo *et al.*, 2015). As before, fitness was determined as the flux through the objective function of the mutant relative to the wild type, with the biomass formation (Supplementary Table 3.S2) being the objective function during growth in bulk soil and the rhizosphere, and N_2-fixation (Zhao *et al.*, 2012) being the objective function in the nodule environment.

The mutant analyses revealed that the *S. meliloti* chromosome had a similar contribution to fitness in bulk soil and the rhizosphere; there was little change in the number of essential or fitness contributing chromosomal genes in these two environments (Figure 3.3 and Supplementary Table 3.S3). However, there was a clear reduction in the importance of the chromosome during symbiosis in the nodule, consistent with microarray data showing an over-representation of chromosomal genes amongst the genes with low expression in symbiotic relative to free-living bacteria (Becker *et al.*, 2004).

Similar to the chromosome and consistent with the global *S. meliloti* transcriptional down-regulation in the nodule (Capela *et al.*, 2001), pSymB contributed

more or less only to the fitness of the free-living bacterium with little role detected in the bacteroids (Figure 3.3 and Supplementary Table 3.S3). However unlike the chromosome, pSymB showed a bias in importance between growth in bulk soil and the rhizosphere; the number of fitness promoting genes was ~ 3.5 fold greater in the rhizosphere. Moreover, every pSymB gene that contributed to fitness in bulk soil had a greater fitness contribution in the rhizosphere. This rhizosphere bias was further amplified when considering the origin of fitness promoting genes. Of the five pSymB genes contributing to fitness in bulk soil, four are involved in arabinose transport or catabolism (Poysti *et al.*, 2007). All four of these genes have a chromosomal origin and were transferred to pSymB through an inter-replicon translocation event (diCenzo *et al.*, 2013). We therefore detected only a single gene (*smb20201*) contributing to fitness in bulk soil that originated on pSymB. Similarly, transcriptomics work with the pea symbiont, *Rhizobium leguminosarum*, indicated that one of its plasmids (pRL8) is over-represented in genes up-regulated specifically in the pea rhizosphere (Ramachandran *et al.*, 2011). However, with a few exceptions, the fitness contributions of the pLR8 up-regulated genes in bulk soil versus the rhizosphere were not determined.

Even though these data clearly illustrated that the metabolic capabilities encoded by pSymB were either specific or more important for growth in the rhizosphere than bulk soil, we believe that the observed bias was an under-representation of the actual situation. The succinoglycan biosynthetic genes are classified as essential in both bulk soil and the rhizosphere due to their inclusion in the biomass objective functions; however, they are

not truly essential but likely have greater importance in the rhizosphere through facilitating biofilm formation on the legume root. Furthermore, a more complete formulation of the bulk soil and rhizosphere environment may exaggerate the bias. For example, protocatechuate was not included due to a lack of information on its abundance. However, recent work showed that protocatechuate metabolism improved fitness of *R. leguminosarum* in the rhizosphere (Garcia-Fraile *et al.*, 2015), and 13 pSymB genes are involved in protocatechuate transport and metabolism (MacLean *et al.*, 2006; 2009).

In contrast with the other replicons, the pSymA megaplasmid contributed no fitness promoting genes (Figure 3.3 and Supplementary Table 3.S3). No phenotypes were detected in bulk soil, while the 'essential' genes in the rhizosphere were due to the removal of Nod factor biosynthetic genes. In fact, Nod factor biosynthesis is not essential for growth but is required for the initiation of symbiosis. In the nodule, the essential genes that were identified were required for the synthesis and functioning of the nitrogenase enzyme. The lack of fitness contributing pSymA genes in the nodule was somewhat surprising but consistent with published data (diCenzo *et al.*, 2016b; Yurgel *et al.*, 2013), suggesting that few genes outside of the core symbiotic machinery contribute to the nitrogen fixation process. Indeed, the large rearrangements in the structure of pSymA between wild type *S. meliloti* nodule isolates (Galardini *et al.*, 2013) may reflect low selective constraints on the pSymA megaplasmid, and thus explain the low metabolic contribution and importance of pSymA even during the symbiotic interaction.

The biases observed for the importance of each replicon in the different

environments were confirmed via random permutations, testing up to 1,000 different nutritional compositions as described in detail in Section 3.9.1.3. These environments were created by randomly varying, at each iteration, the maximal allowable uptake of each nutrient with respect to the original value and by also randomly removing two nutrients from the environment. Despite some interesting biological insights being derived from this analysis (see Section 3.9.1.3 and Supplementary Figures 3.S7 and 3.S8), little variation was seen in the number of essential plus fitness contributing genes on each replicon or in each environments (Supplementary Figures 3.S7 and 3.S8). The robustness of these results to environmental variations provides support for the validity of our conclusions and shows that the niche specialization is not unique to the specific environmental composition used throughout this study.

Finally, a comparison of genes differentially contributing to growth in each environment with a recent regulon analysis in *S. meliloti* (Galardini *et al.*, 2015) was not conclusive due to the low overlap of the datasets (Supplementary Table 3.S4, Supplementary Data 3.S5; additional details in Section 3.9.1.4). On the other hand, grouping these genes based on their pangenome classification (Galardini *et al.*, 2015) illustrated that nearly all fitness contributing genes belonged to the core genome, a clear enrichment relative to the percentage of core genes in iGD1575 overall (Supplementary Table 3.S5, Supplementary Data 3.S5; additional details in Section 3.9.1.5).

3.5 Discussion

We have completed a comprehensive, manually and experimentally curated

genome-scale metabolic reconstruction of a model multipartite genome of the N_2-fixing endosymbiont *S. meliloti*, and modelled the metabolic changes associated with niche transition. The switch from bulk soil to the rhizosphere was accompanied by a metabolic fine-tuning, primarily through changes in carbon metabolism and amino acid biosynthesis. In contrast, moving from the rhizosphere to the nodule involved a comprehensive metabolic reprogramming. This involved essentially shutting off production of all biomass compounds and instead synthesizing co-factors necessary for a successful symbiosis, maximizing ATP production, and fixing atmospheric nitrogen.

The analysis of the *in silico* fitness contributions of genes included in iGD1575 revealed that the chromosome is not metabolically specialized for a particular niche, but instead encodes the core metabolic machinery that enables growth of *S. meliloti* as a free living microbe. In contrast, the evidence indicated that pSymB is metabolically specialized for the rhizosphere, helping *S. meliloti* to adapt to this environment and utilize the newly available substrates. The analysis failed to detect any environment where pSymA contributed to improved fitness, but it was seen that pSymA functions were solely relevant to the symbiotic process. Concerning multipartite genome evolution, these observations are consistent with an evolutionary scenario where: (1) the gain of a pSymB ancestor from another bacterium first significantly improved the ability of *S. meliloti* to colonize the rhizosphere as suggested previously (diCenzo *et al.*, 2014; Egan *et al.*, 2005), (2) pSymB gained additional genes, through horizontal gene transfer, encoding metabolic functions that contribute to fitness predominately in the rhizosphere, and (3) pSymA only

contributes metabolic functions relevant for establishing a nitrogen fixing symbiosis.

We speculate that our observations here with *S. meliloti* may be generalizable to other bacteria with a multipartite genome that interact with a eukaryotic host. We hypothesize that secondary replicons might facilitate the start of a host interaction; this is the case for the large *E. coli* virulence plasmids (Johnson & Nolan, 2009) and the rhizobial symbiotic plasmids (Remigi *et al.*, 2016). Once the organism begins inhabiting the host-associated niche, the secondary replicon might acquire genes that improve fitness specifically in this new environment, whereas the chromosome remains largely undifferentiated, carrying the general metabolic pathways required for life and traits specific to the cell's original environment.

The modelling framework we have developed for this work can be adapted to study other types of biological association (e.g. pathogenesis) and the metabolic reprogramming that is needed to operate the switch towards a novel ecological niche. Moreover, by demonstrating here that chromids and megaplasmids carry genes that primarily improve fitness in a specific niche, such as host interaction, this work illustrates secondary replicons as a rich reservoir of genes that have potential in synthetic biology applications. Finally, we anticipate that the iGD1575 model herein reconstructed will represent a valuable platform for future manipulations of *S. meliloti* aimed at its biotechnological exploitation in the context of agricultural procedures.

3.6 Materials and Methods

3.6.1 Metabolic network reconstruction

A draft metabolic model was constructed using the KBase Narrative Interface (www.kbase.us) and then manually and experimentally validated and expanded based on published data as described in the Section 3.9.2.1. The final *S. meliloti* model was termed iGD1575 in accordance with the nomenclature standard (Reed *et al.*, 2003), and includes 1575 genes, 1825 reactions, and 1579 metabolites. The SBML file of the model was validated by the online SBML validator tool (http://sbml.org/Facilities/Validator/), and is available as Supplementary Data 3.S6. Metabolic modeling was performed using Matlab R2015a (Mathworks), using scripts from the Cobra Toolbox (Schellenberger *et al.*, 2010) and the Gurobi 6.0.1 solver (www.gurobi.com). A detailed description of the modeling procedure is reported in Sections 3.9.2.2 through 3.9.2.4. For comparison of iGD1575 with previously published flux data (Fuhrer *et al.*, 2005), the flux through each reaction was constrained by setting the upper and lower bounds to the average plus or minus the error of the experimentally derived values. To facilitate construction of the *in silico* large-scale *S. meliloti* gene deletion mutants in iGD1575, identification of essential model genes was performed as described in Section 3.9.1.6 and the essential iGD1575 genes are listed in Supplementary Table 3.S7.

3.6.2 Biomass composition

No comprehensive description of the macromolecular composition of the *S. meliloti* biomass exists in the literature. However, such data is available for *Rhodobacter*

sphaeroides, a related α-proteobacterium (Imam *et al.*, 2011). We therefore approximated the *S. meliloti* gross biomass composition using that of *R. sphaeroides*. Nevertheless, the specific composition of DNA, RNA, protein, and lipids were adjusted based on the *S. meliloti* GC content (Galibert *et al.*, 2001), codon usage (Nakamura *et al.*, 1999), and lipid composition (Basconcillo *et al.*, 2009; Gao *et al.*, 2004; Weissenmayer *et al.*, 2002; Zavaleta-Pastor *et al.*, 2010). Furthermore, succinoglycan was included in the biomass at 5% of the dry weight, which was estimated based on the literature (Dorken *et al.*, 2012; Glenn *et al.*, 2007; Wang *et al.*, 2007). The complete biomass composition is given in Supplementary Table 3.S2.

3.6.3 Objective function formulation

The objective function for growth in synthetic media and bulk soil was set as a biomass reaction, producing biomass as described in the above section and fully detailed in Supplementary Table 3.S2; this objective function was termed 'biomass_bulk_c0'. The objective function for growth in the rhizosphere ('biomass_rhizo_c0') was the same as for bulk soil except that the amount of succinoglycan was doubled to account for biofilm formation on the plant root, and Nod factor was included (1 mg per gm dry weight) as its production would be stimulated by the legume and is required for the initiation of symbiosis (Supplementary Table 3.S2). Finally, the 'symbiosis_c0' objective function was adapted from a published *S. meliloti* model (Zhao *et al.*, 2012), and was used for modelling symbiosis. In short, the symbiosis objective function involved the synthesis of biomolecules relevant to symbiosis as well as the export of L-alanine, L-

aspartate, and ammonium from fixed N_2.

3.6.4 In silico *environmental representations*

In silico representations of the nutritional composition of the rhizosphere and bulk soil were constructed from data available in the literature (Table 3.3). For both soil representations, ammonium and nitrate were included at a one to one ratio, and sufficient ammonium, nitrate, phosphate, sulphate, metal ions, and gases were included so that these compounds were not growth rate limiting. The relative abundance of the major carbon compounds was derived from the available literature as described in the Section 3.9.2.5. The boundaries of the exchange reactions used to define each environment are listed in Supplementary Table 3.S6, as are the flux rate through all active exchange reactions.

3.6.5 Gene functional analysis

The WebMGA webserver (Wu *et al.*, 2011) was used to provide functional Cluster of Orthologous Gene (COG) annotations (p-value cut off of 0.001) to each gene in the model. Between replicon biases were determined after standardizing by the number of genes from each replicon in iGD1575. To perform the COG analyses of the genes associated with variable reactions during the transition between niches, the COG annotation for each gene associated with the variable reaction classes was extracted from the WebMGA output of the previous COG analysis. Biases were determined after standardizing by the number of genes in each class of variable genes. Statistical significance was determined using Pearson's Chi-squared tests. The complete list of COG annotations is available as Supplementary Data 3.S7.

3.6.6 Phenotype MicroArray™ analysis

Phenotype MicroArray™ experiments using Biolog plates PM1 and PM2A were performed largely as described previously (Biondi *et al.*, 2009) and elaborated on in Section 3.9.2.6. All bacterial strains used in this study were described previously (diCenzo *et al.*, 2013; Milunovic *et al.*, 2014) and are listed in Supplementary Table 3.S8. Of note, whereas most strains were inoculated from agar plates, *S. meliloti* RmP2754 (ΔB180) and a second wild type control were inoculated from liquid M9-glucose cultures as RmP2754 grew poorly when inoculated directly from an agar plate. Data analysis was performed with DuctApe (Galardini *et al.*, 2014). Activity index (AV) values were calculated following subtraction of the blank well from the experimental wells, whereas plots of the growth curves were of the unblanked data. Growth with each compound was considered positive if the AV value was ≥ 4. Negative growth phenotypes of the mutant strains were called if the AV value was ≤ 3, and only following manual inspection of the unblanked curves. However, a growth cut-off of 4 was likely to falsely eliminate some compounds that support slow growth of *S. meliloti* (Galardini *et al.*, 2014), such as β-hydroxybutyrate (AV value = 3) and acetoacetate (AV value = 2) (Charles *et al.*, 1997).

3.6.7 Growth curves and phosphate determination

S. meliloti was grown overnight in LBmc complex medium (diCenzo *et al.*, 2014). These cultures were washed with 0.85% saline and resuspended to an OD_{600} of ~ 0.05 in MM9 minimal medium (diCenzo *et al.*, 2014) with either 7.5 mM glucose or 20 mM succinate as the sole carbon source. Two hundred µL of the cell suspensions were

transferred in triplicate to wells of a 96-well microtitre plates and grown for 24 hours at 30°C with shaking in a BioTek Cytation 3 plate reader. OD_{600} readings were measured every 15 minutes. Growth rates were calculated between OD_{600} readings (not corrected for pathlength) of 0.1-0.5 with a previously developed Perl script (diCenzo *et al.*, 2014).

To measure the amount of phosphate used by *S. meliloti*, the phosphate concentrations in the spent media following the completion of the growth curves were determined via the molybdenate blue – ascorbic acid colorimetric method (Murphy & Riley, 1962). In brief, cultures were centrifuged, 50 µL of supernatant was diluted with 5 mL of phosphate-free water, and 0.8 mL of mixed reagent (Murphy & Riley, 1962) was added to each sample. Following 10-30 minutes of incubation at room temperature, the A_{880} of each sample was measured and compared to a standard curve. The amount of phosphate remaining in spent media was compared to the phosphate present in the bacteria-free cultures to determine the amount of phosphate used by the bacteria. As the carbon source was growth limiting in these media, the used phosphate to carbon source ratio was calculated by dividing the amount of phosphate removed from the medium by the initial concentration of the carbon source (i.e. 7.5 mM glucose or 20 mM succinate).

3.6.8 Data availability.

This article together with the supplementary data include all data that is relevant to the conclusions of this work. Matlab scripts used for generation of the FBA data are available from the authors upon request.

3.8 Tables and Figures

Table 3.1. Summary of the main properties of iGD1575.

S. meliloti 1021 genome	
Total genome size	6 691 694
Size of the chromosome (% total)	3 654 135 (54.6)
Size of pSymB (% total)	1 683 333 (25.2)
Size of pSymA (% total)	1 354 226 (20.2)
Total protein coding genes (PCG)	6 204
Chromosome PCG (% total)	3 341 (53.9)
pSymB PCG (% total)	1 570 (25.3)
pSymA PCG (% total)	1 293 (20.8)
iGD1575 characteristics	
Total genes (% of *S. meliloti* genes)	1 575 (25.4)
Chromosome genes (% total)	944 (69.9)
pSymB genes (% total)	390 (24.8)
pSymA genes (% total)	241 (15.3)
Total reactions (rxns)	1 825
Gene associated rxns (gar) (% total)	1 404 (76.9)
Chromosome dependent (% gar)	898 (64.0)
pSymB dependent (% gar)	205 (14.6)
pSymA dependent (% gar)	73 (5.2)
Multiple replicons (% gar)	228 (16.2)
Unknown metabolic GPR (% total)	63 (3.4)
Unknown transport GPR (% total)	46 (2.5)
Exchange reactions (% total)*	270 (14.8)
Demand reactions (% total)†	22 (1.2)
Diffusion reactions (% total)	8 (0.4)
Spontaneous reactions (% total)	10 (0.5)
Objective functions (% total)	3 (0.2)
Total metabolites	1 579

* Exchange reactions were used to define the medium components. † Demand reactions were used to provide compounds whose synthesis is not represented in the model. Twenty of the demand reactions represented the uncharged tRNA molecules, two were for fatty acids.

Table 3.2. Carbon utilization phenotypes observed for pSymB deletion mutants.

	Phenotypes		
Strain	Biolog	Both	Model
ΔB154	L-histidine D-glucosamine	None	None
ΔB141	Palatinose* Maltitol**	D-trehalose *Trans*-4-hydroxy-L-proline	*Cis*-4-hydroxy-D-proline† D-proline†
ΔB163	Dulcitol* D-glucosamine	D-psicose	None
ΔB180	Uridine* *N*-acetyl-D-galactosamine* Arbutin* D-raffinose Organic acids ø	L-leucine	D-galactosamine¥
ΔB181	L-lysine* Dulcitol* *N*-acetyl-D-galactosamine* D-glucosamine	L-histidine D-tagatose	D-galactosamine¥
ΔB108	L-ornithine L-serine L-asparagine L-alanine	None	None
ΔB109	None	None	None
ΔB179	Glycerol L-lactic acid	L-arabinose M-inositol	None
ΔB118	None	L-ornithine M-inositol L-arginine	None
ΔB182	Acetic acid Asparagine** α-methyl-D-galactoside** Melibionic acid** Bromosuccinate**	D-melibiose L-malic acid D-raffinose Succinic acid D,L-malate Fumaric acid L-aspartate	Protocatechuate†
ΔB161	Dulcitol* α-methyl-D-galactoside** Melibionic acid**	D-melibiose D-raffinose α-D-lactose Lactulose Methyl-β-D-galactoside	None

The 'Biolog' column indicates the phenotypes observed experimentally that were not seen *in silico* with the iGD1575 model, and *vice versa* for the 'Model' column. The 'Both' column lists the phenotypes observed both experimentally and *in silico*. * The model did not produce biomass with this compound. ** The model did not include this compound. † This compound was not present in the PM1 or PM2A plates, but the phenotype was confirmed in the literature. ¥ This compound was not present in the PM1 or PM2A plates, and the phenotype was not been reported in the literature. Ø Included all tested L-amino acids, gly-glu, ala-gly, gly-asp, L-lactic acid, acetic acid, pyruvic acid, methylpyruvic acid, melibionic acid, and GABA.

Figure 3.1. Agreement between experimental and *in silico* metabolic capabilities of *S. meliloti*.

True positives, growth was observed experimentally and *in silico*. True negatives, growth was not observed experimentally or *in silico*. False negatives, compounds that support growth experimentally but not *in silico*. False positives, compounds that support growth *in silico* but not experimentally. The complete set of compounds and growth predictions can be found in Supplementary Data 3.S1.

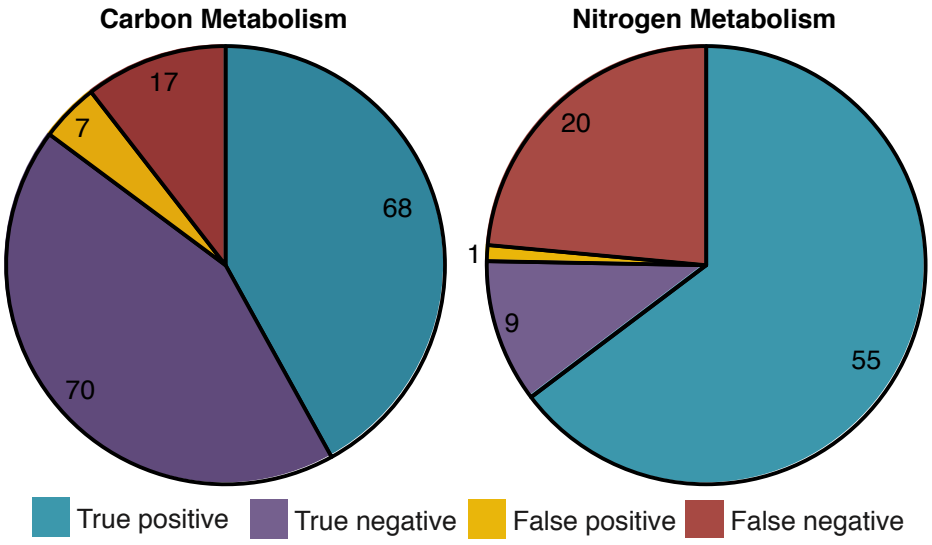

Carbon Metabolism
17
7
68
70
Nitrogen Metabolism
20
1
9
55
True positive
True negative
False positive
False negative

Figure 3.2. The effect of niche conditions on the reconstructed metabolic network.

Networks were visualized following optimization in (**A**) bulk soil, (**C**) rhizosphere, and (**E**) nodule environments. Lines are colour coded based on fitness effect of deleting each reaction: blue indicates a fitness decrease < 1%, dark purple indicates a fitness decrease < 50%, bright purple indicates a fitness decrease > 50%, and red indicates a fitness decrease > 99%. Thin grey lines indicate inactive reactions. Line thickness shows the flux through each reaction on a log scale. The graphs summarize the metabolic changes detected during the (**B**) bulk soil to rhizosphere and (**D**) rhizosphere to nodule transitions. Summary of Changes graphs: on and off – reactions carrying flux only in the second and first environment, respectively; up and down – reactions carrying increased flux (≥ 50%) in the second and first environment, respectively; reverse – reactions whose directionality is switched; greater and lesser – reactions whose removal have a greater (≥ 10%) fitness impact in the second and first environment, respectively; essential and non-essential – reactions essential only in the second and first environment, respectively. The nine classifications are not mutually exclusive. The reactions present in each category are described in Supplementary Data 3.S3 and 3.S4. The COG Analysis graphs summarize the functional annotation of the genes associated with the reactions in the summary of changes graphs. The blue and red bars include the genes associated with the blue and red bars, respectively, in the summary of changes graphs. Asterisks indicate statistically significant changes (P-value < 0.01) as determined by Pearson's Chi-squared tests. In the Reaction Flux figures, each point represents the amount of flux through individual reactions in the two environments. Blue and purple symbols indicate reactions with the same or reverse directionality, respectively. The angled line indicates the position of a perfect correlation. In the Reaction Deletion Fitness figures, each point represents the fitness of individual reaction deletion mutants in the two environments. The angled line indicates the position of a perfect correlation.

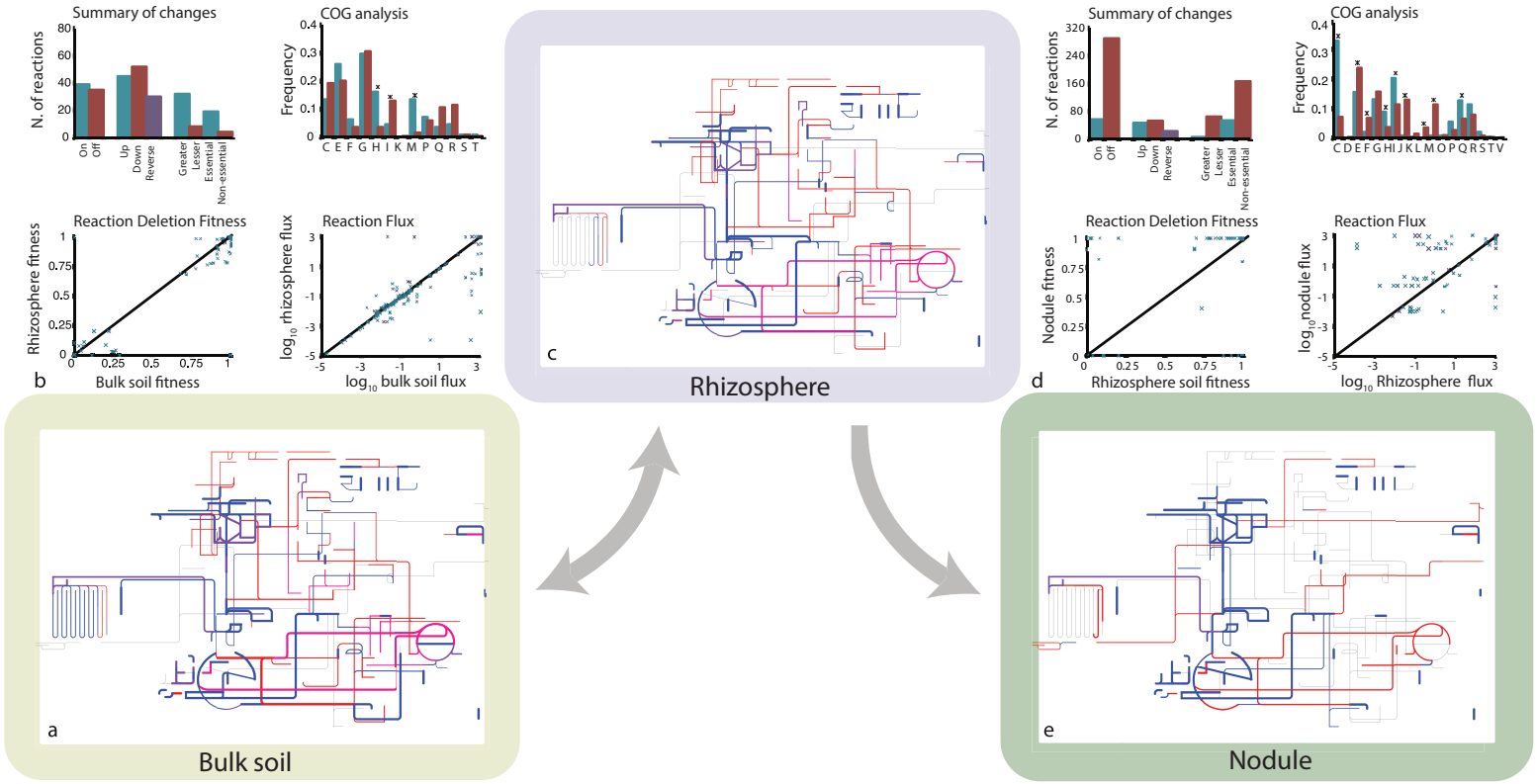
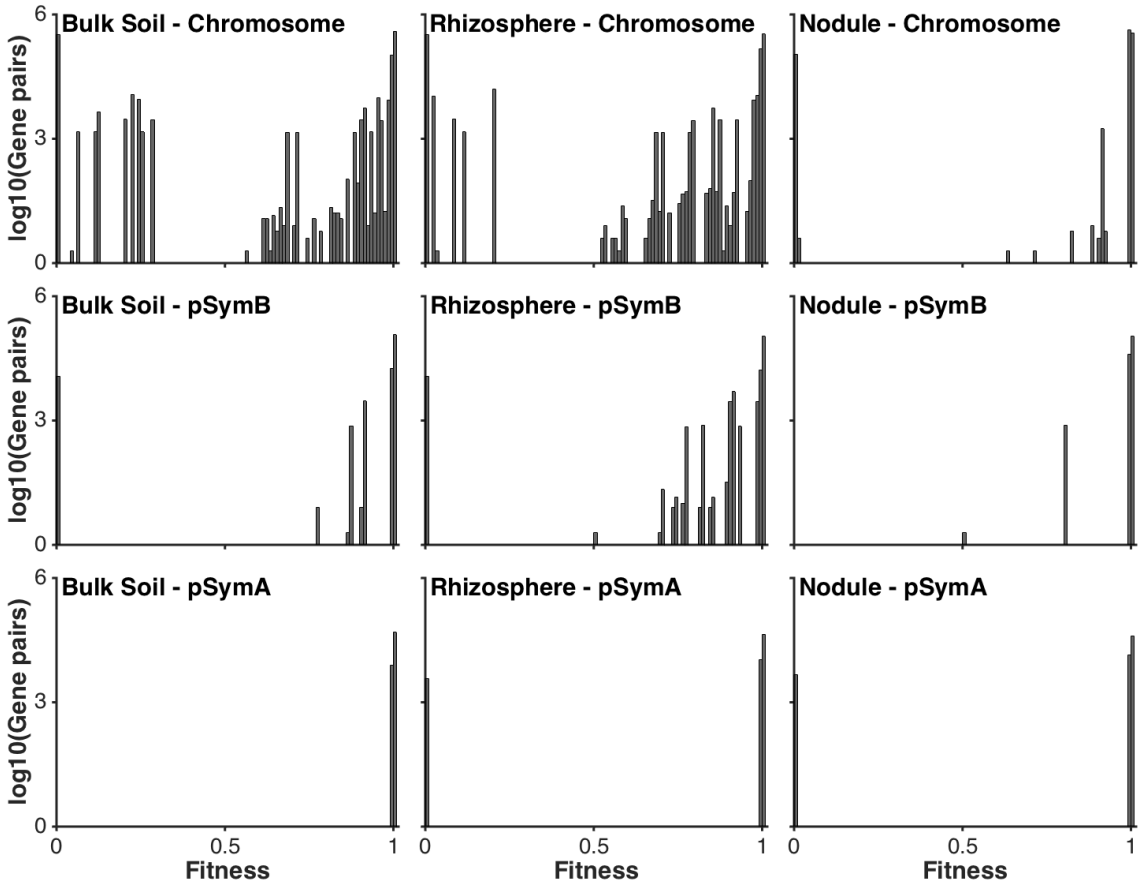

Summary of changes
N. of reactions
On
Off
Up
Down
Reverse
Greater
Lesser
Essential
Non-essential
COG analysis
Frequency
C E F G H I K M P Q R S T
Reaction Deletion Fitness
Rhizosphere fitness
Bulk soil fitness
Reaction Flux
log10 rhizosphere flux
log10 bulk soil flux
b
c
Rhizosphere
Summary of changes
N. of reactions
COG analysis
Frequency
C D E F G H I J K L M O P Q R S T V
Reaction Deletion Fitness
Nodule fitness
Rhizosphere soil fitness
Reaction Flux
log10 nodule flux
log10 Rhizosphere flux
d
a
Bulk soil
e
Nodule

Figure 3.3. Fitness costs associated with double gene deletions in bulk soil, the rhizosphere, and the nodule.

All possible pairs of genes present on the same replicon were individually removed from the model, the ability of the resulting mutant to produce flux through the objective function was examined with FBA, and the fitness (solution value of the mutant / solution value of the wild type) of each mutant was calculated. The histograms summarize the calculated fitness values for each mutant in each of the three environments separately for each replicon. The fitness is displayed on the X-axis, with the number of mutants displaying that fitness level on the Y-axis. The metabolic relevance of a replicon in a particular environment is represented by the number of mutants showing phenotypes between the two extremes (1 and 0); the greater the metabolic relevance, the greater the number of non-extreme phenotypes.


Bulk Soil - Chromosome
Rhizosphere - Chromosome
Nodule - Chromosome
Bulk Soil - pSymB
Rhizosphere - pSymB
Nodule - pSymB
Bulk Soil - pSymA
Rhizosphere - pSymA
Nodule - pSymA
log10(Gene pairs)
0
3
6
0
0.5
1
Fitness

3.9 Supplementary materials

3.9.1 Supplementary text

3.9.1.1 Prediction of the genetic basis of D-galacactosamine metabolism

The genetic basis for several of the observed carbon utilization phenotypes had not been previously examined in *S. meliloti*. By combining our mutant phenotype data with iGD1575, the output of the DuctApe software (Galardini *et al.*, 2014), the *S. meliloti* genome annotation (Galibert *et al.*, 2001), and a previously published ABC transporter induction study (Mauchline *et al.*, 2006), we predicted novel carbon catabolic loci. One example compound is D-galactosamine that was not present in the Phenotype MicroArray^{TM}, but *N*-acetyl-D-galactosamine, whose catabolism proceeds via D-galactosamine, was present. Both the ΔB180 and ΔB181 mutants failed to utilize *N*-acetyl-D-galactosamine as a carbon source. Two ABC transporters within ΔB180, one encoded by *smb21135-smb21138* and the second by *smb21216, smb21219-smb21221*, are induced by D-galactosamine (Mauchline *et al.*, 2006). The *smb21216, smb21219-smb21221* operon also includes a putative sugar isomerase (*smb21218*) and sugar amine kinase (*smb21217*). Thus, we hypothesize that the Smb21216, Smb21219-Smb21221 transporter, as well as potentially the Smb21135-Smb21138 transporter, transports D-glucosamine, which is phosphorylated by Smb21217 to D-glucosamine-6-phosphate, and converted to D-tagatose-6-phosphate by Smb21218. D-tagatose-6-phosphate can then be further metabolized via two steps to glyceraldehyde-3-phosphate and dihydroxyacetone phosphate, both which can enter glycolysis. At least one of these steps is located within

ΔB181 as this mutant cannot grow with *N*-acetyl-D-glucosamine or D-tagatose. In fact, the *smb21373-smb21377* operon, removed in ΔB181, encodes an ABC transporter that we predict transports D-tagatose and that indeed is induced by tagatose (Mauchline *et al.*, 2006), as well as a putative carbohydrate kinase and a putative D-tagatose-1,6-bisphosphate aldolase that could complete the D-galactosamine catabolic pathway.

3.9.1.2 Validation of the metabolic switching detection during niche adaptation

Flux variability analysis (FVA) was employed to ensure that the observed flux changes between environments were true changes. During FBA, one flux distribution solution is returned although multiple flux distribution solutions may provide the same maximal flux through the objective function. FVA examines the range of flux that can be carried by each reaction while maintaining maximal flux through the objective function.

A total of 199 reactions were considered to change in importance between simulated growth in bulk soil and the rhizosphere based on changes in flux and/or the fitness contribution. Of these, 63 (31.7%) were supported by fitness data while 136 (68.7%) were classified as changing solely on the basis of flux. Of the 136 reactions, the predicted reaction flux during growth in the rhizosphere for 61 of them was outside the FVA range for the reaction when grown in bulk soil. This supported that optimal biomass production in the rhizosphere required that flux through these 61 reactions changed. To further examine the remaining 75 reactions, where the reaction flux in the rhizosphere was within the FVA range of the reaction when grown in bulk soil, the flux through each reaction was individually set to the rhizosphere value, and the flux distribution during

growth in bulk soil was monitored. In all but two cases, doing so caused the flux through at least one other reaction (with an average of 9 reactions and standard deviation of 4) to move outside of the rhizosphere FVA range of that reaction. Or in other words, while 61 of the predicted changing reactions between bulk soil and rhizosphere could be artificially set to the rhizosphere flux rate without impairing growth in bulk soil, in all but two cases, doing so would require an average of nine new changes following transition from growth in bulk soil to the rhizosphere. Additionally, when flux through all 61 of the reactions were simultaneously set to the corresponding flux value from the rhizosphere, and growth simulated in bulk soil, the model became unsolvable. Together, these data support that 197 of the 199 reactions (99.0%) predicted to change in importance during transition between growth in bulk soil and the rhizosphere are true changes.

The same procedure was used to confirm the flux changes observed during transition between the rhizosphere and the nodule environments. Of the 451 reactions that were considered changing, 292 (64.7%) were supported by fitness data while 159 (35.3%) were based solely on flux data. Of these 159 reactions, the predicted flux through 57 in the nodule was outside the FVA range for the reaction during simulated growth in the rhizosphere. For all the remaining 102 reactions, individually setting the flux during growth in the rhizosphere to the flux value predicted in the nodule resulted in the flux through at least one (average of 4.5, standard deviation 2) other reaction to move outside the nodule FVA range of that reaction. Additionally, the model became unsolvable in the rhizosphere when flux through all 102 reactions were simultaneously set

to the corresponding nodule flux value. When considered together, these data support that all of the reactions predicted to change in importance during transition between growth in the rhizosphere and nitrogen fixation in the nodule were true changes.

3.9.1.3 Robustness with respect to changes in the nutrients composition and uptake rates

We explored the effect of random variations in the composition and uptake rates of the nutrients present in each of the simulated ecological niches (bulk soil, rhizosphere, and nodule) on the main outcomes of the model, i.e. predicted growth rates, number of essential and fitness contributing genes and gene pairs. Results obtained (shown in Supplementary Figures 3.S7 and 3.S8) revealed that, overall, conclusions drawn concerning the role of pSymA and pSymB throughout the *S. meliloti* lifecycle and the growth phenotypes in different conditions were not influenced by such variations. In all environments, the number of essential genes showed very little variation across all the iterations, and the number of essential plus fitness contributing genes on each replicon was predominately unaffected by the environmental variations. This trend was also observed in the case of essential and fitness contributing gene-pairs (Supplementary Figure 3.S8). A notable exception was represented by the effectiveness of the symbiosis in the nodule (Supplementary Figure 3.S7), which seemed to be more dependent on the composition and the utilization rate of the input compounds. However, this higher variation was not surprising as a limited number of nutrients were available to the bacteroid in the nodule; as such, any change in one nutrient was likely to influence flux through the objective function much more. Overall, the stability of the number of

essential/fitness contributing genes across all permutations confirmed that the presented results were robust and supported the conclusions drawn in this work.

3.9.1.4 Few biases were detected in the genomic localization of the transcriptional regulators of metabolic genes

The genes present in iGD1575 were compared with a recent *in silico S. meliloti* regulon analysis (Galardini *et al.*, 2015) to look for biases in the location of predicted transcriptional regulators. The low overlap between the genes of iGD1575 and the regulon analysis meant that no conclusive correlations were observed. Nevertheless, there appeared to be a bias for genes associated with reactions classified as less important during the bulk soil to rhizosphere or the rhizosphere to nodule transitions to be regulated by a pSymB encoded transcription factor (Supplementary Table 3.S4 and Supplementary Data 3.S5). However, while this may be suggestive of a bias for the regulatory machinery associated with niche adaptation to be encoded by pSymB, it may also simply reflect a bias in the dataset as most of the regulated genes were located on pSymB and transcription factors tend to regulate genes on their own replicon (Galardini *et al.*, 2015).

3.9.1.5 The core S. meliloti *genome is over-represented among fitness contributing genes*

An advantage of using *S. meliloti* as a model organism is that many strains have been fully sequenced, facilitating the study of the *S. meliloti* pangenome. We classified all genes associated with the environmentally variable or fitness promoting reactions as belonging to the 'core', 'accessory', or 'unique' genome (Supplementary Data 3.S5) based on previous results (Galardini *et al.*, 2015). Perhaps not surprisingly given that

iGD1575 is enriched in the core metabolic processes relative to the entire *S. meliloti* genome, the core genome was over-represented in the model (Supplementary Table 3.S5). Remarkably, however, nearly all genes contributing to fitness in either bulk soil, the rhizosphere, or the nodule belonged to the core genome, a clear enrichment relative to the percentage of core genes in iGD1575 overall. This observation highlights that metabolic genes contributing to the fitness of the cell are highly likely to be or become part of the core genome, emphasizing the functional role of core genome as common tool set of genes for a given bacterial species.

3.9.1.6 Identification of essential model genes

Currently, no high throughput systematic knock-out studies of *S. meliloti* exist in the literature, and so the complete set of essential genes in this organism is not known. However, previous work has shown that only two essential genes exist on the pSymB chromid and none are present on pSymA (diCenzo *et al.*, 2013; 2014; Oresnik *et al.*, 2000). In this work we examined a library of pSymB deletion mutants that collectively remove 98% of pSymB (Milunovic *et al.*, 2014) for carbon utilization phenotypes using Phenotype MicroArrays™. We also replicated this experiment *in silico* using iGD1575. In order to do so, it was necessary to ensure that none of the *in silico* deletion mutants were lethal. To test this, we used an *in silico* single gene deletion analysis to identify single copy essential genes in iGD1575 when grown in a minimal medium with sucrose as the carbon source and with thiamine supplementation (diCenzo *et al.*, 2014; Oresnik *et al.*, 2000). In this context, an essential gene refers to a gene whose deletion prevents the

formation of at least one biomass precursor when grown. This analysis identified a total of 231 single copy essential metabolic genes (Supplementary Table 3.S8), none of which were on pSymA. However, of the 231 essential genes, 216 were on the chromosome while 15 were on pSymB.

Of the 15 genes on pSymB, 12 were involved in the biosynthesis of succinoglycan. While these genes were not truly essential, the inclusion of succinoglycan in the biomass reaction resulted in them being considered essential in the model. Two of the essential pSymB genes, *wgaG* and *wgaJ*, were predicted to be required for lipopolysaccharide (LPS) biosynthesis. Further examination, revealed another two reactions associated with multiple redundant pSymB genes that were essential for LPS synthesis. All four reactions were required for the synthesis of dTDP-rhamnose, necessary for the production of the O-antigen. Little is known about LPS biosynthesis in *S. meliloti*, and it is possible that these genes are truly required for complete LPS synthesis. But whereas *Sinorhizobium* LPS mutants that cannot incorporate rhamnose into their LPS produce a truncated LPS yet survive (Ardissone *et al.*, 2011), the rigidity of the model and incorporation of complete LPS in the biomass reaction meant that such a mutant would fail to produce biomass *in silico*. Finally, *ansB* was not surprisingly determined to be essential as it is the only *S. meliloti* gene predicted to be involved in asparagine biosynthesis, although experimental evidence indicates it is not essential for asparagine biosynthesis and the asparagine biosynthetic pathway in *S. meliloti* remains unidentified (diCenzo *et al.*, 2014). The two genes on pSymB that are truly essential are

a tRNA and a protein involved in ribosomal biogenesis (*engA*) (diCenzo *et al.*, 2013), and are therefore not present in the model.

When succinoglycan was removed from the biomass reaction, and unknown GPRs were added to the AsnB and the four dTDP-rhamnose synthesis reactions, the deletion of all pSymA and pSymB genes from the model did not prevent biomass formation. Thus, for the *in silico* experiments testing the metabolic capacity of iGD1575, the model was modified as described in the previous sentence so that all deletion mutants were viable.

3.9.2 Supplementary materials and methods

3.9.2.1 Metabolic network reconstruction

A draft metabolic model was constructed using the KBase Narrative Interface (www.kbase.us). The *S. meliloti* 1021 annotated genome was imported from the public KBase database, and a draft metabolic model was reconstructed using the 'Build metabolic model' method. This draft model was gap-filled using the 'Gapfill metabolic model' method to allow biomass formation when grown in a minimal medium containing metal ions, succinate, ammonium, sulphate, phosphate, and biotin. The model was then downloaded, the KBase annotations in the 'Gene_association' field replaced with the *S. meliloti* 1021 locus tags, and the KBase annotations in the 'Protein_association' field replaced with the *S. meliloti* 1021 gene names (Galibert *et al.*, 2001). The KBase biomass objective was removed, new objective functions were formulated (representing biomass formation and symbiosis, as described below) and the model manually gap filled to produce flux through each of the objective functions.

Manual curation and further expansion of the model through the inclusion of additional reactions and gene-protein-reaction associations (GPRs) was performed in several main stages. Where possible, 'Unknown' GPRs were replaced with genes from the *S. meliloti* 1021 genome, and when supported by published experimental data, incorrect GPRs and reactions were removed from the draft model. We next identified and included additional metabolic and transport genes in the *S. meliloti* 1021 genome annotation. Following this, the genes present in the draft model were compared to the list of genes included in the existing *S. meliloti* metabolic model iHZ565 (Zhao *et al.*, 2012), and the majority of the additional genes and associated reactions from iHZ565 were added to our draft model. However, 31 genes were not transferred from iHZ565 (Supplementary Table 3.S1) as the annotation was extremely general and we lacked confidence in the true substrates/products of the reaction, experimental data was inconsistent with their inclusion, or the reaction produced a dead-end metabolite (a metabolite produced or consumed by only a single reaction, meaning that the reaction will never be active during flux balance analysis). An iterative gap-filling procedure was then employed to reconcile the predictions of 'growth' or 'no growth' with various carbon and nitrogen substrates with the known ability or inability of *S. meliloti* 1021 to grow on these substrates. Most of the experimental growth data came from a previous Phenotype MicroArray^{TM} (Biolog) experiment (Biondi *et al.*, 2009), with a few additional substrates taken from additional literature sources (Boivin *et al.*, 1991; Chen *et al.*, 2010; Kohler *et al.*, 2011; MacLean *et al.*, 2009). Finally, a library of large *S. meliloti* deletion mutants

was screened for carbon utilization defects using the OmniLog Phenotype MicroArray™ system (Biolog), as described below. *In silico* predictions were compared with the experimental results, and where necessary and possible, discrepancies were fixed by further refinement and expansion of the metabolic model. Where possible, GPRs for transport reactions were added based on published mutation or induction studies; otherwise, transport reactions were added as a diffusion reaction with no associated GPR. Support for manually added transporters and metabolic reactions came from experimental evidence, review articles, and the KEGG and BioCyc databases (Ampomah *et al.*, 2013; Biondi *et al.*, 2009; Boncompagni *et al.*, 2000; Borisova *et al.*, 2011; Caspi *et al.*, 2014; Charles *et al.*, 1997; Chen *et al.*, 2010; Cheng *et al.*, 2011; de Rudder *et al.*, 1999; diCenzo *et al.*, 2015; Dominguez-Ferreras *et al.*, 2009a, b; Dunn, 2015; Dupont *et al.*, 2004; Gage & Long, 1998; Gao *et al.*, 2004; Geddes & Oresnik, 2012a, b; 2014; Geddes *et al.*, 2010; Geiger & López-Lara, 2002; Goldmann *et al.*, 1994; Gu *et al.*, 2010; Harrison *et al.*, 2005; Jebbar *et al.*, 2005; Jensen *et al.*, 2002; Jones *et al.*, 2007; Kanehisa *et al.*, 2014; Kohler *et al.*, 2010; 2011; Lambert *et al.*, 2001; las Nieves Peltzer *et al.*, 2008; Lerouge *et al.*, 1990; Lynch *et al.*, 2001; MacLean *et al.*, 2006; 2009; 2011; Mauchline *et al.*, 2006; Phillips *et al.*, 1998; Poysti *et al.*, 2007; Raimunda & Elso-Berberián, 2014; Richardson & Oresnik, 2007; Richardson *et al.*, 2004; Sagot *et al.*, 2010; Weissenmayer *et al.*, 2002; White *et al.*, 2012; Willis & Walker, 1999; Wilson & Kappler, 2009; Yurgel & Kahn, 2005; Yurgel *et al.*, 2013).

The final *S. meliloti* model was termed iGD1575 in accordance with the

nomenclature standard (Reed *et al.*, 2003), and includes 1575 genes, 1825 reactions, and 1579 metabolites. The SBML file of the model was validated by the online SBML validator (sbml.org/Facilities/Validator/), and is available as Supplementary Data 3.S6.

3.9.2.2 Metabolic modeling

The ability of the model to produce flux through the objective functions was examined using flux balance analysis (FBA) in Matlab R2015a (Mathworks), using scripts from the Cobra Toolbox (Schellenberger *et al.*, 2011) and the Gurobi 6.0.1 solver (www.gurobi.com). Single gene deletion and double gene deletion analyses were performed using methods in the Cobra Toolbox. Iteratively removing each reaction from the model and then examining the effect with FBA determined the essentiality of each individual reaction. Metabolic capacity was determined *in silico* by iteratively providing the model with a unique carbon or nitrogen source and then observing the ability of the model to produce biomass with FBA. The predicted effect of the large-scale genome deletions on the growth phenotype (either growth or no growth) was addressed by simultaneously removing all reactions dependent on the deleted genes from the model and then running the *in silico* phenotype microarray experiment. For this analysis, exopolysaccharide was removed from the objective function and an unknown GPR added to the other four essential reactions dependent on pSymB genes (see Section 3.9.1.6 for additional details) in order to allow all deletion mutants to grow in the minimal media.

3.9.2.3 Evaluation of robustness to changes in the nutrients composition and uptake rates

Simulation of growth/symbiosis in 1000 randomly generated media for each

environment was performed as follows. For each of the 1000 iterations, a random variation in the allowable uptake rate for each of the nutrients in the medium was introduced, with the allowable variation set to a value 50% greater or lower than the original one (e.g. if the original uptake rate was 1 mmol gm^{-1} hr^{-1}, in each of the 1000 conditions, the uptake rate was randomly set between 0.5 to 1.5 mmol gm^{-1} hr^{-1}). Additionally, further noise was introduced by randomly removing, at each iteration, two nutrients from the niche simulated nutrients set and restored for the following iteration. FBA was then used, at each iteration, to evaluate i) the predicted growth rate in each of the iterations for each of the three environments, ii) the variations in the number of essential genes, and iii) the variations in replicon specific essential plus fitness contributing genes. Due to time constraints, when determining the variation in the number of replicon specific essential/fitness-contributing gene pairs related to random changes in the nutrients composition, the number of iterations was reduced to 100.

3.9.2.4 Flux visualization

Metabolic networks were visualized with the online tool iPath 2.0 (Yamada *et al.*, 2011). Where possible, KEGG IDs were associated with each reaction based on comparison with the Seed Reference list, and these reactions mapped to the corresponding reaction, if one existed, in the 'Metabolic Pathway' map of iPath.

3.9.2.5 In silico *environmental representations*

In silico representations of the nutritional composition of the rhizosphere and bulk soil were constructed from data available in the literature. For both the rhizosphere and

bulk soil *in silico* representations, ammonium and nitrate were included at a one to one ratio, and sufficient ammonium, nitrate, phosphate, sulphate, metal ions, and gases were included so that these compounds were not growth rate limiting.

The sugar content of the *in silico* rhizosphere was primarily set according to the average monosaccharide composition of pea (*Pisum sativum*) and cowpea (*Vigna unguiculata*) root mucilage (Knee *et al.*, 2001; Moody *et al.*, 1988). Sucrose, raffinose, and stachyose were added to the list of sugars, with the ratio of these three sugars set according to their approximate ratio in alfalfa (*Medicago sativa*) root exudate (Bringhurst *et al.*, 2001). The total amount of sucrose, raffinose, and stachyose was arbitrarily set so that 60% of the total glucose was within these sugars, and the amounts of free glucose and galactose were reduced accordingly. The organic acids included in the rhizosphere, and their relative ratios, were based on their prevalence in unstressed alfalfa root exudate (Lipton *et al.*, 1987). The total ratio of organic acids to sugars in legume root exudate was not found; however, organic acids were ~ 10, 4, and 2 fold more prevalent than sugars in the root exudates of tomato (*Lycopersicon esculentum*), cucumber (*Cucumis sativus*), and sweet pepper (*Capsicum annum*), respectively (Kamilova *et al.*, 2006). Therefore, the total amount of organic acid in the rhizosphere was set at a molar ratio two fold greater than the total carbohydrates. The amino acid content of the rhizosphere was primarily based on the amount of each amino acid present in pea root exudate when grown in quartz sand (Boulter *et al.*, 1966). To this, hydroxyproline was added according to the serine to hydroxyproline ratio in pea root mucilage (Knee *et al.*, 2001). The

carbohydrate to amino acid ratio in legume root exudate was not found; however, an approximate carbohydrate to protein ratio of four to one, or greater, was observed in three rice varieties (*Oryza sativa*) (Naher *et al.*, 2008). Therefore, the total amount of amino acids in the rhizosphere was set at a molar ratio four fold less than the total carbohydrates.

The molar ratio of sugars in the *in silico* bulk soil representation was set as determined previously (Murayama, 1981). Unlike in the rhizosphere condition, sucrose, raffinose, and stachyose were not added to the bulk soil as they appear to be largely absent (Bringhurst *et al.*, 2001; Jaeger *et al.*, 1999). The concentration of organic acids appears to be quite low in bulk soil, with the organic acid content of the rhizosphere likely at least 50-fold higher than that of bulk soil (Jones, 1998). Therefore, the three organic acids included in the rhizosphere formulation were also included in bulk soil, but at 2% the concentration. The dominant amino acid content of bulk soil was set as the average of two previously analyzed soil samples (Hertenberger *et al.*, 2002). These amino acids accounted for 66.9% of the amino acids, with the remaining 33.1% split evenly between the non-modified amino acids not displayed in the reference (Hertenberger *et al.*, 2002) as they never exceeded 5% of the total amino acid population. The total carbohydrate content of the bulk soil was set ten fold higher than the total amino acid content, as determined previously (Hertenberger *et al.*, 2002).

The nutritional environment within the root nodule was set as used previously for constraint based modelling of iHZ565 (Zhao *et al.*, 2012). The upper and lower bounds for exchange reactions in all three environments are listed in Supplementary Table 3.S6.

3.9.2.6 Phenotype MicroArray™ analysis

S. meliloti RmP110 (Yuan *et al.*, 2006), a derivative of *S. meliloti* 1021 in which a frameshift mutation within *pstC* was fixed, was used as the wild type reference strain. All deletion mutant strains were described previously (diCenzo *et al.*, 2013; Milunovic *et al.*, 2014), and consist of large deletions of the pSymB replicon that each span ~ 40 to 370 kilobase pairs. Phenotype MicroArray™ experiments were performed largely as described previously (Biondi *et al.*, 2009), using Biolog plates PM1 and PM2A. Strains were initially grown at 30°C on LBmc agar supplemented with $CoCl_2$ (diCenzo *et al.*, 2014). To begin the Phenotype MicroArray™ analysis, colonies were picked up with sterile cotton swabs and resuspended in 0.8% NaCl to a cell density of 81% turbidity (OD_{600} ~ 0.1) as measured with a Biolog turbidimeter. Two mL of each suspension was diluted in 22 mL of carbon free M9 minimal medium (diCenzo *et al.*, 2014) containing 240 µL of the redox dye MixA (Biolog), and 100 µL of the final mixture was added to each well of the Biolog plates. The one exception was *S. meliloti* RmP2754 (ΔB180), which grew very poorly when inoculated directly from the agar plate. Therefore, this strain and a second replicate of *S. meliloti* RmP110 were pregrown in liquid M9-glucose, washed twice with 0.8% saline, and diluted in 0.8% saline to a turbidity of 81%. This cellular suspension was then treated as described above and used to inoculate the PM plates. All PM plates were incubated at 30°C in an OmniLog plate reader, which measured reduction of the dye every 15 minutes for 120 hours.

3.9.3 Supplementary Tables and figures

Table 3.S1. iHZ565 model genes excluded from iGD1575.

Gene	iHZ565 reaction	Reason for exclusion from iGD1575
smc04455	acetolactate synthase	Does not complement auxotrophy of a *smc01431* mutation (las Nieves Peltzer *et al.*, 2008)
smc03785	ADPribose diphosphatase	Required a dead end metabolite
smc03763	DNA (cytosine5)methyltransferase	Would produce a dead end metabolite
smc03243	Dihydroneopterin dephosphorylase	Too general of an annotation in the *S. meliloti* 1021 genome annotation
smc03236	GMP synthase	Lacked confidence in the annotation
smc00356	LysyltRNA synthetase	Lacked confidence in the annotation
smc02689	Aminobutyraldehyde dehydrogenase	Too general of an annotation in the *S. meliloti* 1021 genome annotation
smc02377	putrescine oxidase	Lacked confidence in the annotation
smc02263	acetolactate synthase	Does not complement auxotrophy of a *smc01431* mutation (las Nieves Peltzer *et al.*, 2008)
smc01404	Aspartate racemase	Would produce a dead end metabolite
smc01153	3hydroxyacylCoA dehydratase (3hydroxytetradecanoylCoA)	Too general of an annotation in the *S. meliloti* 1021 genome annotation
smc01147	coproporphyrinogen oxidase (O2 required)	Lacked confidence in the annotation
smc01146	thiamin pyrophosphatase	Lacked confidence in the annotation
smc00810	Guanosine 3'-diphosphate 5'-triphosphate	Lacked confidence in the annotation
smc00808	protoporphyrinogen oxidase (aerobic)	Lacked confidence in the annotation
smc00410	NADH dehydrogenase	Lacked confidence in the annotation
smb21586	glutathione synthetase	Does not complement mutation of *gshB1* (Harrison *et al.*, 2005)
smb21301	Aminobutyraldehyde dehydrogenase	Too general of an annotation in the *S. meliloti* 1021 genome annotation
smb20857	glucose6phosphate isomerase	Excluded based on (Geddes & Oresnik, 2012b)
smb20433	ornithine cyclodeaminase	Does not complement mutation of *ocd* (diCenzo *et al.*, 2015)
smb20115	Dihydroxyacid dehydratase (2,3dihydroxy3methylpentanoate)	Does not complement auxotrophy of a *smc04045* mutation (las Nieves Peltzer *et al.*, 2008)
smb20072	Cytosolic source for myoinositol	Lacked confidence in the annotation
sma2213	Aminobutyraldehyde dehydrogenase	Too general of an annotation in the *S. meliloti* 1021 genome annotation
sma2211	acetolactate synthase	Does not complement auxotrophy of a *smc01431* mutation (las Nieves Peltzer *et al.*, 2008)
sma1871	ornithine cyclodeaminase	Does not complement mutation of *ocd* (diCenzo *et al.*, 2015)
sma1844	Aminobutyraldehyde dehydrogenase	Too general of an annotation in the *S. meliloti* 1021 genome annotation
sma1155	magnesium transport via ABC system	Too general of an annotation in the *S. meliloti* 1021 genome annotation
sma0959	3oxoacylacylcarrierprotein reductase (nC10:0)	Too general of an annotation in the *S. meliloti* 1021 genome annotation
sma0958	acetolactate synthase	Does not complement auxotrophy of a *smc01431* mutation (las Nieves Peltzer *et al.*, 2008)
sma0956	glutamate1semialdehyde aminotransferase	Too general of an annotation in the *S. meliloti* 1021 genome annotation
sma0486	ornithine cyclodeaminase	Does not complement mutation of *ocd* (diCenzo *et al.*, 2015)

Table 3.S2. Biomass composition used in this study.

Component	Percent dry mass	Composition - % [a]
DNA [b]	2.8	Guanine - 31.05 Cytosine - 31.05 Adenine - 18.95 Thymine - 18.95
RNA [c]	7.1	Guanine - 28.09 Cytosine - 28.09 Adenine - 21.91 Uracil - 21.91
Protein [d]	49.3	Lysine - 3.20 Alanine - 12.01 Leucine 10.19 Phenylalanine - 3.94 Arginine 7.33 Glutamine - 2.90 Glycine - 8.46 Methionine - 2.44 Valine - 7.58 Proline - 5.03 Tyrosine - 2.29 Aspartate - 5.31 Glutamate - 5.84 Histidine - 2.11 Threonine - 5.15 Cysteine - 0.93 Isoleucine - 5.48 Tryptophan - 1.38 Asparagine - 2.64 Serine - 5.80
Lipid [e]	12.8	PG(36:2) - 7.82 CL(36:2) - 3.11 PE(36:2) - 25.35 PC(36:2) - 59.92 SL(36:2) - 2.00 OL(36:1) - 1.80
PHB	17.6	N/A
Glycogen	0.4	N/A
LPS	3	N/A
Cell wall	2	N/A
LMW Succinoglycan [f]	4 [g]	N/A
HMW Succinoglycan [f]	1 [g]	N/A
Putrescine	Trace	N/A
Spermidine	Trace	N/A
Nod factor [h]	1 mg per gm cell dry weight	N/A

[a] Where applicable, the subunit composition of each macromolecule is given, as is what

percentage of the macromolecule that each subunit accounts for.

[b] Composition based on the overall GC content of *S. meliloti* (Galibert *et al.*, 2001).

[c] The GC content for mRNA was estimated from the overall GC content of *S. meliloti* (Galibert *et al.*, 2001). The GC content of tRNA was estimated based on the GC content of the 10 most common codons in *S. meliloti* (Galibert *et al.*, 2001; Nakamura *et al.*, 1999). The GC content of rRNA was determined based on the *rrn* loci of *S. meliloti* (Galibert *et al.*, 2001). The overall composition of cellular RNA was determined assuming 80% rRNA, 15% tRNA, and 5% mRNA.

[d] The amino acid composition was estimated based on the codon usage of *S. meliloti* (Nakamura *et al.*, 1999).

[e] The lipid composition used was as previously determined for *S. meliloti*, with each lipid class represented by a single lipid of the most common lipid size (Basconcillo *et al.*, 2009; Gao *et al.*, 2004; Weissenmayer *et al.*, 2002; Zavaleta-Pastor *et al.*, 2010).

[f] A 4:1 ratio of low molecular weight (LMW) to high molecular weight (HMW) succinoglycan was set as previously determined (Glenn *et al.*, 2007).

[g] These numbers were for when growth was modelled in bulk soil. When growth was simulated in the rhizosphere, the amount of both LMW and HMW succinoglycan was doubled.

[h] Nod factor was included in the biomass composition when growth was modelled in the rhizosphere, but not when growth was modelled in bulk soil.

Table 3.S3. Summary of the fitness effect of individual gene deletions.*

Replicon	Niche	Non-essential†	Essential†	Fitness contributing†
Chromosome	Bulk soil	704 (74.6)	191 (20.2)	49 (5.2)
	Rhizosphere	702 (74.4)	193 (20.4)	48 (5.1)
	Nodule	882 (93.6)	59 (6.3)	1 (0.1)
pSymB	Bulk soil	369 (94.9)	15 (3.9)	5 (1.3)
	Rhizosphere	356 (91.6)	15 (3.9)	18 (4.6)
	Nodule	388 (99.7)	0 (0.0)	1 (0.3)
pSymA	Bulk soil	240 (100)	0 (0.0)	0 (0.0)
	Rhizosphere	232 (96.7)	8 (3.3)	0 (0.0)
	Nodule	230 (95.8)	10 (4.2)	0 (0.0)

* Values indicate the number of genes (% of genes from the given replicon). † Non-essential – fitness of mutant > 99% of the wild type; Essential – fitness of mutant < 1% of the wild type; Fitness contributing – fitness of mutant ≥ 1% and ≤ 99% of the wild type.

Table 3.S4. The location of transcription factors regulating genes associated with environmental variable and fitness determining reactions.

Replicon	Entire Genome	Bulk Fit	Rhizo Fit	Nodule Fit	B2R More	B2R Less*	R2N More	R2N Less*
Chromosome	0.488 (312)	0.500 (1)	0.333 (1)	0.000 (0)	0.667 (6)	0.259 (7)	0.619 (13)	0.184 (7)
pSymB	0.353 (226)	0.500 (1)	0.667 (2)	1.000 (1)	0.333 (3)	0.741 (20)	0.143 (3)	0.632 (24)
pSymA	0.159 (102)	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)	0.238 (5)	0.184 (7)

Values are ratios of all genes in the given group (column) with the gene count given in brackets. Replicon – the replicon on which the transcriptional regulator is located. Entire Genome – the distribution in the entire *S. meliloti* 1021 genome. Bulk Fit – genes whose deletion results in a fitness decrease in bulk soil. Rhizo Fit – genes whose deletion results in a fitness decrease in the rhizosphere. Nodule Fit – genes whose deletion results in a fitness decrease in the nodule. B2R More and B2R Less– genes associated with more important and less important reactions, respectively, following the bulk soil to rhizosphere transition. R2N More and R2N Less– genes associated with more important and less important reactions, respectively, following the rhizosphere to nodule transition. Statistically significant biases within each group, compared to the entire genome, were determined using Pearson's Chi Squared tests: * p-value < 0.001.

Table 3.S5. Summary of the pangenome classification of genes present in iGD1575.

Compartment	Genome	iGD1575*	Bulk Act*	Rhizo Act*	Nodule Act*	Bulk Fit*†	Rhizo Fit*†	Nodule Fit
Core	0.697 (4337)	0.827 (1296)	0.834 (637)	0.848 (6.38)	0.851 (344)	0.962 (51)	0.970 (64)	1.000 (2)
Accessory	0.285 (1775)	0.172 (269)	0.165 (126)	0.150 (113)	0.146 (59)	0.038 (2)	0.030 (2)	0.000 (0)
Unique	0.017 (106)	0.001 (2)	0.001 (1)	0.001 (1)	0.001 (1)	0.000 (0)	0.000 (0)	0.000 (0)

Values are ratios of all genes in the given group (column) with the gene count given in brackets. Compartment – the pangenome classifications. The distribution is shown for: the entire *S. meliloti* 1021 genome (entire genome), the model (iGD1575), genes associated with reactions active in bulk soil (Bulk Act), the rhizosphere (Rhizo Act), or the nodule (Nodule Act), and genes contributing to fitness in bulk soil (Bulk Fit), the rhizosphere (Rhizo Fit), or the nodule (Nodule Fit). * Distribution was significantly different (p-value < 0.001) from that of the *S. meliloti* 1021 genome as determined using Pearson's Chi Squared tests. † Distribution was significantly different (p-value < 0.05) from that of iGD1575 as determined using Pearson's Chi Squared tests.

Table 3.S6. Exchange reaction bounds for setting the environmental conditions.

		Bulk Soil			Rhizosphere			Nodule		
Compound	Exchange reaction*	Lower Bound	Upper Bound	Flux rate	Lower Bound	Upper Bound	Flux rate	Lower Bound	Upper Bound	Flux rate
L-arabinose	EX_cpd00224_e0	-0.28889705	1000	-0.28889705	-0.26974143	1000	-0.26974143	0	1000	0
D-galactose	EX_cpd00108_e0	-0.25290619	1000	-0.25290619	-0.18317325	1000	-0.18317325	0	1000	0
D-glucose	EX_cpd00027_e0	-0.61972444	1000	-0.61972444	-0.04098397	1000	-0.04098397	0	1000	0
D-mannose	EX_cpd00138_e0	-0.2540567	1000	-0.2540567	-0.05436649	1000	-0.05436649	0	1000	0
L-rhamnose	EX_cpd00396_e0	-0.13913636	1000	-0.13913636	-0.02927426	1000	-0.02927426	0	1000	0
Xylose	EX_cpd00154_e0	-0.21280795	1000	-0.21280795	-0.04600242	1000	-0.04600242	0	1000	0
D-ribose	EX_cpd00105_e0	-0.01862144	1000	-0.01862144	0	1000	0	0	1000	0
Sucrose	EX_cpd00076_e0	0	1000	0	-0.02049199	1000	-0.02049199	0	1000	0
D-raffinose	EX_cpd00382_e0	0	1000	0	-0.01024599	1000	-0.01024599	0	1000	0
Stachyose	EX_cpd01133_e0	0	1000	0	-0.03073798	1000	-0.03073798	0	1000	0
Succinate	EX_cpd00036_e0	-0.0056245	1000	-0.0056245	-0.12240603	1000	-0.12240603	-1.326	0	-1.326
L-malate	EX_cpd00130_e0	-0.03649016	1000	-0.03649016	-0.79413596	1000	-0.79413596	-1.1122	1000	-1.1122
L-aspartate	EX_cpd00041_e0	0	1000	0	-0.03205557	1000	-0.03205557	0	1000	0
L-threonine	EX_cpd00161_e0	-0.0068175	1000	-0.0068175	-0.00961667	1000	-0.00961667	0	1000	0
L-serine	EX_cpd00054_e0	0	1000	0	-0.01004986	1000	-0.01004986	0	1000	0
L-glutamate	EX_cpd00023_e0	0	1000	0	-0.0283302	1000	-0.0283302	-2	10	-2
L-proline	EX_cpd00129_e0	-0.01102953	1000	-0.01102953	0	1000	0	0	1000	0
Glycine	EX_cpd00033_e0	-0.0068175	1000	-0.0068175	-0.00693094	1000	-0.00693094	0	1000	0
L-alanine	EX_cpd00035_e0	-0.02780553	1000	-0.02780553	-0.00823049	1000	-0.00823049	0	1000	0
L-valine	EX_cpd00156_e0	-0.02928849	1000	-0.02928849	-0.00337883	1000	-0.00337883	0	1000	0
L-cysteine	EX_cpd00084_e0	-0.0068175	1000	-0.0068175	0	1000	0	0	1000	0
L-isoleucine	EX_cpd00322_e0	-0.0199273	1000	-0.0199273	-0.00186269	1000	-0.00186269	0	1000	0
L-leucine	EX_cpd00107_e0	-0.03596182	1000	-0.03596182	-0.00303228	1000	-0.00303228	0	1000	0
L-tyrosine	EX_cpd00069_e0	-0.0068175	1000	-0.0068175	-0.00233919	1000	-0.00233919	0	1000	0
GABA	EX_cpd00281_e0	0	1000	0	0	1000	0	0	1000	0
L-ornithine	EX_cpd00064_e0	-0.0068175	1000	-0.0068175	-0.00831712	1000	-0.00831712	0	1000	0
L-histidine	EX_cpd00119_e0	-0.0068175	1000	-0.0068175	-0.00506825	1000	-0.00506825	0	1000	0
L-arginine	EX_cpd00051_e0	-0.0068175	1000	-0.0068175	-0.00948672	1000	-0.00948672	0	1000	0
Trans-4-hydroxy-l-proline	EX_cpd00851_e0	0	1000	0	-0.26974144	1000	-0.26974144	0	1000	0
Ammonia	EX_cpd00013_e0	-7.5	1000	-7.5	-7.5	1000	-7.5	0	1000	0
Nitrate	EX_cpd00209_e0	-7.5	1000	0	-7.5	1000	0	0	1000	0
Sulfate	EX_cpd00048_e0	-15	1000	-0.05219940	-15	1000	-0.05604370	-1000	1000	-0.01900238
Phosphate	EX_cpd00009_e0	-15	1000	-0.17306777	-15	1000	-0.14804998	-1000	0	-0.01425178
H_2O	EX_cpd00001_e0	-15	1000	19.38238966	-15	1000	16.74436329	-1000	1000	2.53455439
O_2	EX_cpd00007_e0	-15	1000	-0.06956818	-15	1000	-0.12160182	-1.26	1000	-0.01187648
CO_2	EX_cpd00011_e0	-15	1000	-5.60835703	-15	1000	-4.83015412	-1000	1000	6.05032969
Mn_2	EX_cpd00030_e0	-15	1000	0	-15	1000	0	0	1000	0

Zn_2	EX_cpd00034_e0	-15	1000	0	-15	1000	0	0	1000	0
Cu_2	EX_cpd00058_e0	-15	1000	0	-15	1000	0	0	1000	0
Ca_2	EX_cpd00063_e0	-15	1000	0	-15	1000	0	0	1000	0
H^+	EX_cpd00067_e0	-15	1000	-15	-15	1000	-15	-1000	1000	-12.54547743
Cl^-	EX_cpd00099_e0	-15	1000	0	-15	1000	0	0	1000	0
Biotin	EX_cpd00104_e0	-15	1000	0	-15	1000	0	0	1000	0
Co_2	EX_cpd00149_e0	-15	1000	0	-15	1000	0	0	1000	0
K^+	EX_cpd00205_e0	-15	1000	0	-15	1000	0	0	1000	0
Mg^{2+}	EX_cpd00254_e0	-15	1000	0	-15	1000	0	-1000	1000	-0.00475059
Na^+	EX_cpd00971_e0	-15	1000	0	-15	1000	0	0	1000	0
Fe^{2+}	EX_cpd10515_e0	-15	1000	0	-15	1000	0	-1000	1000	-0.02375297
Fe^{3+}	EX_cpd10516_e0	-15	1000	0	-15	1000	0	0	1000	0
Cob(I)alamin	EX_cpd00635_e0	0	1000	0	0	1000	0	-1000	1000	-0.00475059
Homocitrate	EX_cpd00919_e0	0	1000	0	0	1000	0	-1000	1000	-0.00475059
Molybdate	EX_cpd11574_e0	0	1000	0	0	1000	0	-1000	1000	-0.00475059
Thiamine	EX_cpd00305_e0	0	1000	0	0	1000	0	-1000	1000	-0.00475059
M-inositol	EX_cpd00121_e0	0	1000	0	0	1000	0	-0.01	0	-0.01
N_2	EX_cpd00528_e0	0	1000	0	0	1000	0	-1000	0	-0.47505938
Urea	EX_cpd00073_e0	0	1000	2.5546521	0	1000	2.89929159	0	1000	0
Biomass	EX_cpd11416_c0	0	1000	0.32084806	0	1000	0.25948388	0	1000	0
5-Methylthioadenosine	EX_cpd00147_e0	0	1000	0.00001692	0	1000	0.00001369	0	1000	0
H_2	EX_cpd11640_e0	0	1000	0	0	1000	0	0	1000	0.47505938
Fixed NH_3	EX_cpdFixed_e0	0	1000	0	0	1000	0	0	1000	0.47505938

*If an exchange reaction is not listed, the lower boundary was set to ‘0’, the upper boundary set to ‘1000’, and the reaction did not carry flux in any of the three conditions.

Table 3.S7. Single copy iGD1575 genes essential for growth in M9-sucrose minimal medium.

pSymA
None
pSymB
smb21690, smb21327, smb21324, smb20961, smb20959, smb20958, smb20957, smb20956, smb20954, smb20949, smb20948, smb20946, smb20944, smb20943, smb20652
Chromosome
smc04434, smc04410, smc04406, smc04405, smc04346, smc04320, smc04268, smc04213, smc04088, smc04045, smc04003, smc04002, smc04001, smc03990, smc03934, smc03925, smc03885, smc03881, smc03863, smc03861, smc03859, smc03858, smc03856, smc03826, smc03823, smc03807, smc03797, smc03795, smc03772, smc03770, smc03173, smc03172, smc03112, smc02912, smc02905, smc02899, smc02898, smc02837, smc02804, smc02802, smc02790, smc02782, smc02767, smc02766, smc02765, smc02755, smc02725, smc02717, smc02700, smc02692, smc02686, smc02652, smc02574, smc02572, smc02570, smc02569, smc02568, smc02567, smc02551, smc02496, smc02450, smc02438, smc02305, smc02245, smc02165, smc02124, smc02113, smc02101, smc02099, smc02093, smc02091, smc02089, smc02080, smc02076, smc02073, smc02070, smc02069, smc02064, smc01934, smc01880, smc01878, smc01875, smc01871, smc01868, smc01867, smc01866, smc01864, smc01863, smc01862, smc01861, smc01804, smc01803, smc01801, smc01761, smc01756, smc01732, smc01726, smc01720, smc01567, smc01563, smc01461, smc01444, smc01431, smc01430, smc01369, smc01364, smc01360, smc01353, smc01352, smc01350, smc01343, smc01321, smc01320, smc01319, smc01318, smc01317, smc01316, smc01314, smc01313, smc01310, smc01309, smc01308, smc01307, smc01306, smc01305, smc01304, smc01303, smc01302, smc01301, smc01300, smc01299, smc01298, smc01297, smc01296, smc01295, smc01294, smc01293, smc01292, smc01291, smc01290, smc01287, smc01286, smc01285, smc01283, smc01233, smc01231, smc01209, smc01192, smc01189, smc01161, smc01152, smc01127, smc01121, smc01116, smc01109, smc01100, smc01096, smc01027, smc01025, smc01018, smc01010, smc01004, smc00993, smc00919, smc00918, smc00917, smc00908, smc00892, smc00855, smc00851, smc00723, smc00711, smc00704, smc00696, smc00695, smc00659, smc00643, smc00615, smc00614, smc00595, smc00568, smc00567, smc00565, smc00554, smc00552, smc00551, smc00526, smc00508, smc00495, smc00494, smc00493, smc00488, smc00485, smc00475, smc00421, smc00415, smc00414, smc00412, smc00408, smc00394, smc00385, smc00366, smc00365, smc00364, smc00363, smc00335, smc00334, smc00333, smc00323, smc00296, smc00293, smc00236, smc00235, smc00232, smc00155, smc00007

Table 3.S8. *Sinorhizobium meliloti* strains used in this study.

Strain	Relevant characteristics	Reference
RmP110	Wild type Rm1021 (SU47 *str-21*) with fixed *pstC*; Sm^R	(Yuan *et al.*, 2006)
RmP790	RmP110, ΔB108 (1,131,168-1,169,073); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP799	RmP110, ΔB109 (1,180,466-1,204,770); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP811	RmP110, ΔB118 (1,323,078-1,528,150); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP876	RmP110, ΔB141 (101,396-466,499); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP1055	RmP110, ΔB161 (1,679,723-49,523); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP1059	RmP110, ΔB154 (62,137-100,636); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2712	RmP110, ΔB179 (1,207,052–1,322,226); Sm^R Sp^R Nm^R Gm^R	(diCenzo *et al.*, 2013)
RmP2715	RmP110, ΔB182 (1,529,711-1,677,882); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2716	RmP110, ΔB181 (870,505-1,129,758); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2717	RmP110, ΔB163 (451,557-651,863); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2754	RmP110, ΔB180 (635,940-869,642); Sm^R Nm^R	(Milunovic *et al.*, 2014)

Sm – streptomycin; Nm – neomycin; Gm – gentamicin; Tc – tetracycline; Sp – spectinomycin

Figure 3.S1. COG analysis of the iGD1575 model genes.

The total number of genes annotated with each COG class is represented by the crosses, which are plotted on the right axis. For each COG class represented in the gene set, the relative enrichment of each class for each replicon is shown and plotted on the left axis: blue – chromosome, red – pSymB, gold – pSymA. Statistically significant biases in distribution, as determined by Pearson's Chi Squared tests, are indicated by the asterisks: * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001. The observed biases were consistent with the previously reported COG biases for the entire *S. meliloti* chromosome, pSymA, and pSymB replicons (Galardini *et al.*, 2015; Galibert *et al.*, 2001).

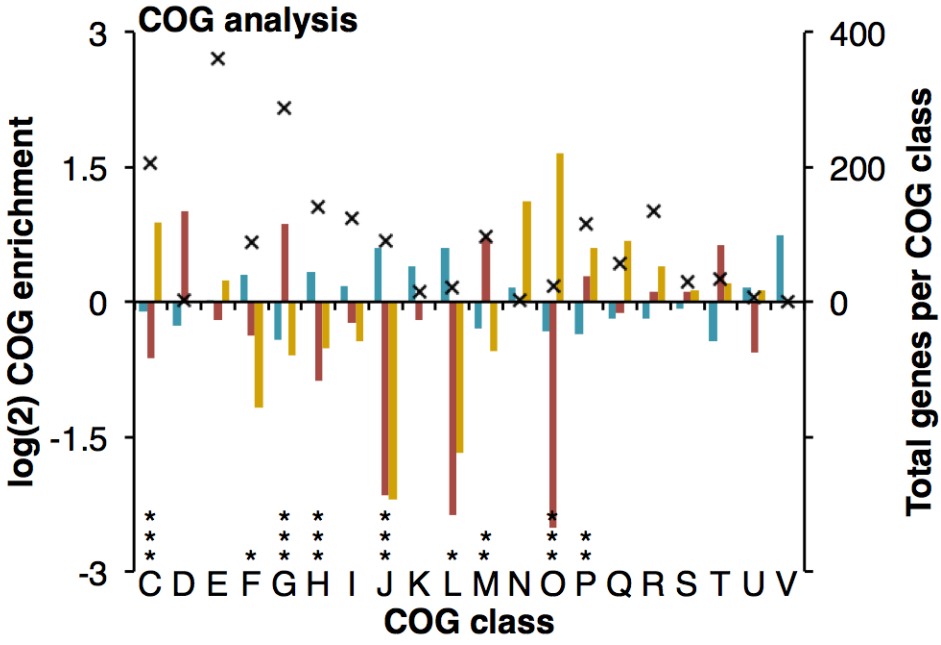
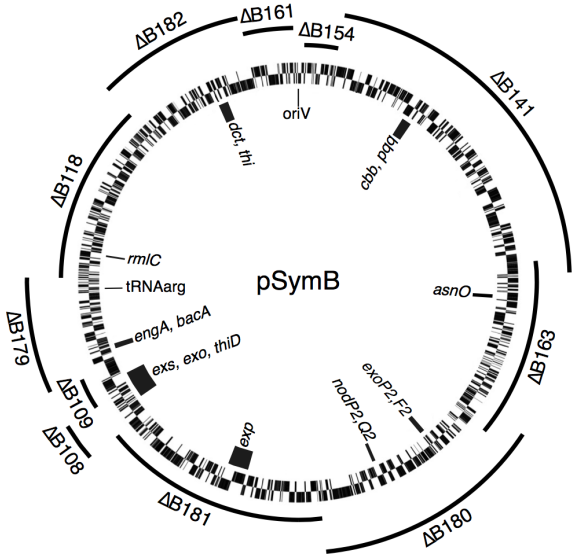

COG analysis
log(2) COG enrichment
Total genes per COG class
COG class
3
1.5
0
-1.5
-3
400
200
0
C D E F G H I J K L M N O P Q R S T U V

Figure 3.S2. Deletion library mutants screened for carbon metabolic phenotypes.

A schematic representation of pSymB and the location of the deletions in the studied mutants is shown. The inner circle represents pSymB, with the individual lines showing the position of annotated genes. The outer lines indicate the region of pSymB that has been removed in the corresponding deletion mutant. Several notable loci are indicated along the inner circle for reference. *dct*: *dctA,B,D.* *thi*: *thiC,O,G,E.* *exs*: *exsA-I.* *exo*: *exoA,B,F1,H,I,K-Q,T-Z.* *exp*: *wgeA-H, wgdA,B, wggR, wgcA, wgaA,B,D-J.* *cbb*: *cbbA,F,L,P,R,S,T,X.* *pqq*: *pqqA-E.*



Strain	Deleted region (nucleotide position)
ΔB154	62,137-100,636
ΔB141	101,396-466,499
ΔB163	451,557-651,863
ΔB180	635,940-869,642
ΔB181	870,505-1,129,758
ΔB108	1,131,168-1,169,073

Strain	Deleted region (nucleotide position)
ΔB109	1,180,466-1,204,770
ΔB179	1,207,052-1,322,226
ΔB118	1,33,078-1,528,150
ΔB182	1,529,711-1,677,882
ΔB161	1,679,723-49,523

Figure 3.S3. Metabolic activity in the PM1 plates.

Pairwise comparisons of the growth of *S. meliloti* RmP110 (blue) and the indicated deletion mutant (red) in each well of the PM1 plates. Wells were not blanked with the carbon free well (A1).

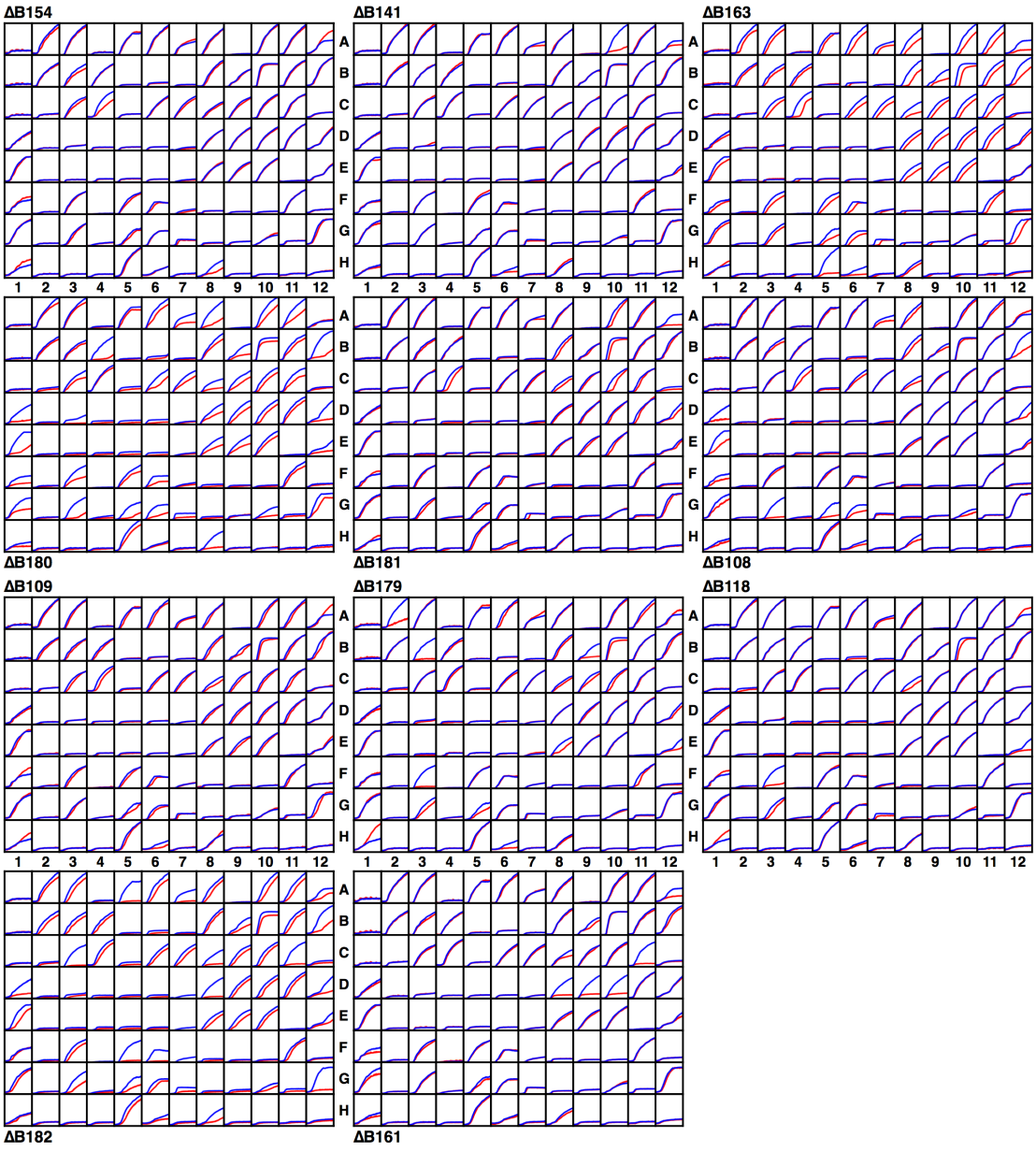

ΔB154
ΔB141
ΔB163
ΔB180
ΔB181
ΔB108
ΔB109
ΔB179
ΔB118
ΔB182
ΔB161
A
B
C
D
E
F
G
H
1 2 3 4 5 6 7 8 9 10 11 12

Figure 3.S4. Metabolic activity in the PM2A plates.

Pairwise comparisons of the growth of *S. meliloti* RmP110 (blue) and the indicated deletion mutant (red) in each well of the PM2A plates. Wells were not blanked with the carbon free well (A1).

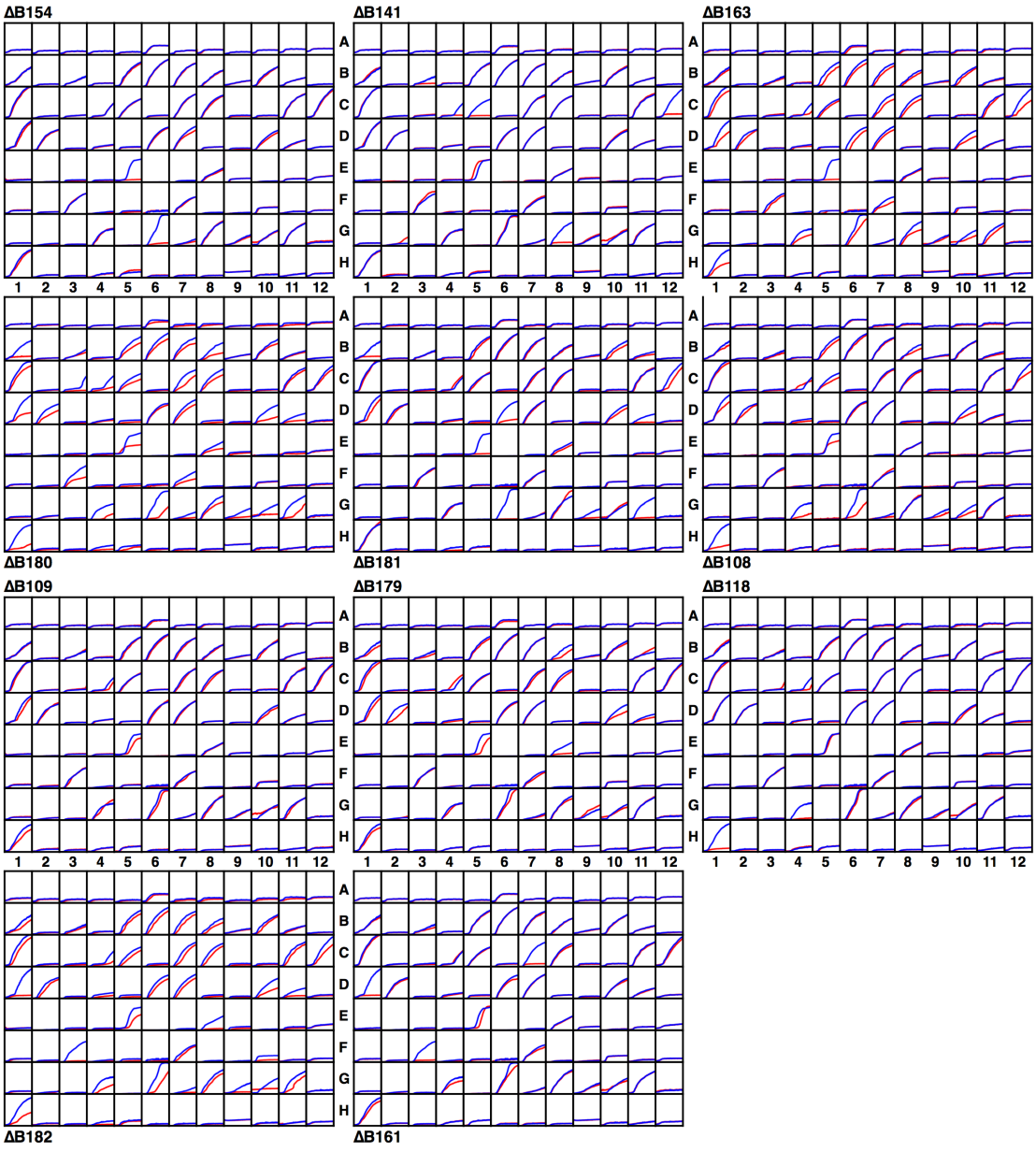

ΔB154
ΔB141
ΔB163
ΔB180
ΔB181
ΔB108
ΔB109
ΔB179
ΔB118
ΔB182
ΔB161
A
B
C
D
E
F
G
H
1 2 3 4 5 6 7 8 9 10 11 12

Figure 3.S5. Changes in reaction flux/essentiality during environmental transitions. Scatterplots, linear regression lines, equations, and R^2 values (all determined in Microsoft Excel) are shown for (**A**) individual reaction flux in bulk soil and the rhizosphere (**B**) fitness of individual reaction deletion mutants in bulk soil and the rhizosphere (**C**) individual reaction flux in the rhizosphere and the nodule (**D**) fitness of individual reaction deletion mutants in the rhizosphere and the nodule.

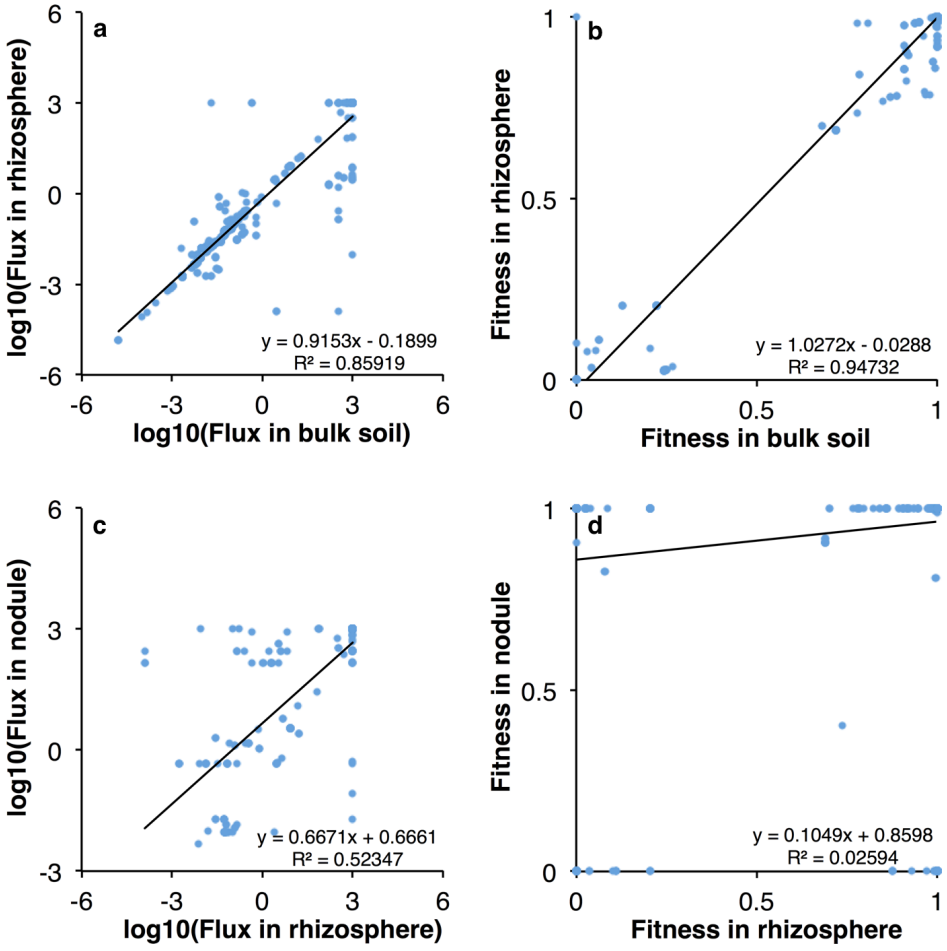

a
log10(Flux in rhizosphere)
log10(Flux in bulk soil)
y = 0.9153x - 0.1899
R² = 0.85919
b
Fitness in rhizosphere
Fitness in bulk soil
y = 1.0272x - 0.0288
R² = 0.94732
c
log10(Flux in nodule)
log10(Flux in rhizosphere)
y = 0.6671x + 0.6661
R² = 0.52347
d
Fitness in nodule
Fitness in rhizosphere
y = 0.1049x + 0.8598
R² = 0.02594

Figure 3.S6. Fitness costs associated with single gene deletions during growth in bulk soil, the rhizosphere, and the nodule.

All genes present within iGD1575 were individually removed from the model, the ability of the resulting mutants to produce flux through the objective function was examined with FBA, and the fitness (flux through objective function in the mutant / flux through the objective function in the wild type) of each mutant was calculated. The histograms summarize the calculated fitness values for each mutant in each of the three environments, plotted separately for each replicon.

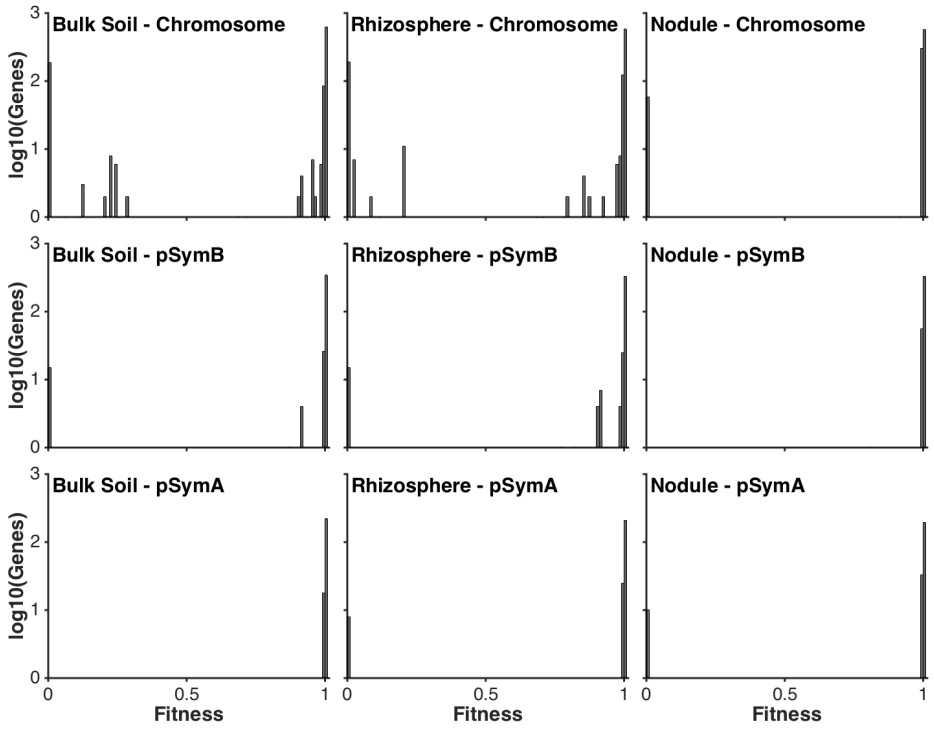

Bulk Soil - Chromosome
Rhizosphere - Chromosome
Nodule - Chromosome
Bulk Soil - pSymB
Rhizosphere - pSymB
Nodule - pSymB
Bulk Soil - pSymA
Rhizosphere - pSymA
Nodule - pSymA
log10(Genes)
0
1
2
3
0.5
Fitness

Figure 3.S7. Robustness to nutrients composition variation of fitness costs associated with single gene deletions and predicted growth rates.

To account for the influence of the nutrients composition of the three niches analysed (bulk soil, rhizosphere, and nodule), 1000 iterations were performed and, for each iteration, new random uptake rates were generated. For each of these iterations, all genes present within iGD1575 were individually removed from the model, the ability of the resulting mutants to produce flux through the objective function was examined with FBA, and the fitness (flux through objective function in the mutant / flux through the objective function in the wild type) of each mutant was calculated. This, in turn, was repeated for each ecological niche (bulk soil, rhizosphere, and nodule). Essential and fitness promoting genes include all genes whose deletion decreases flux through the objective function by a value greater than or equal to 1%.

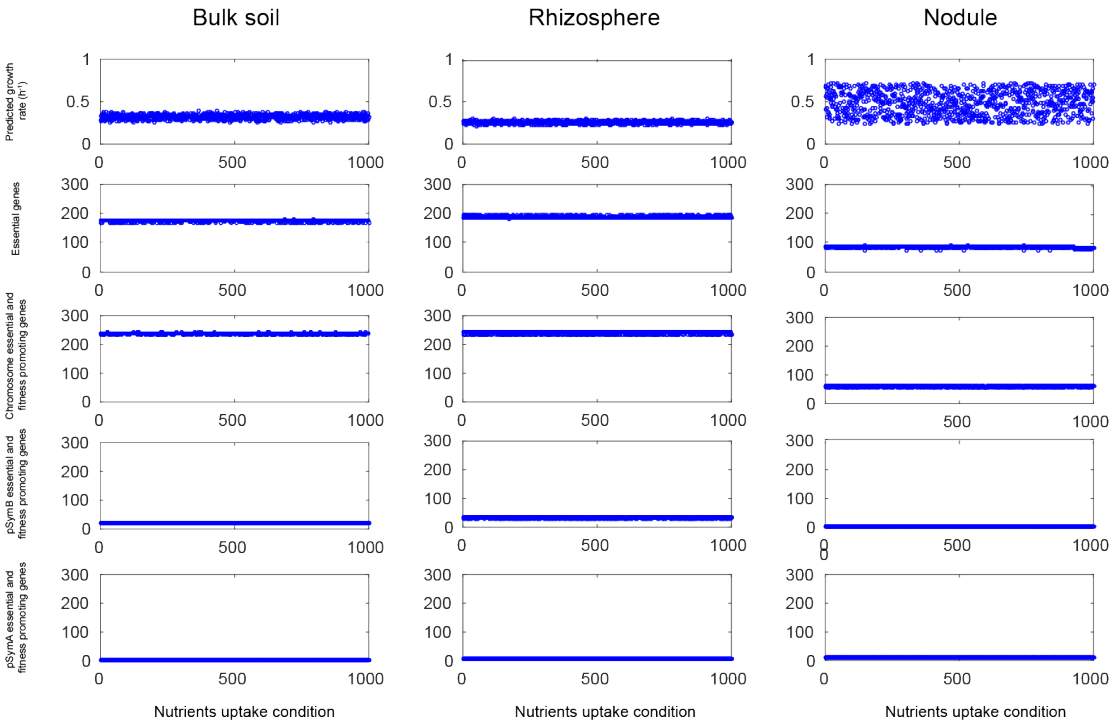
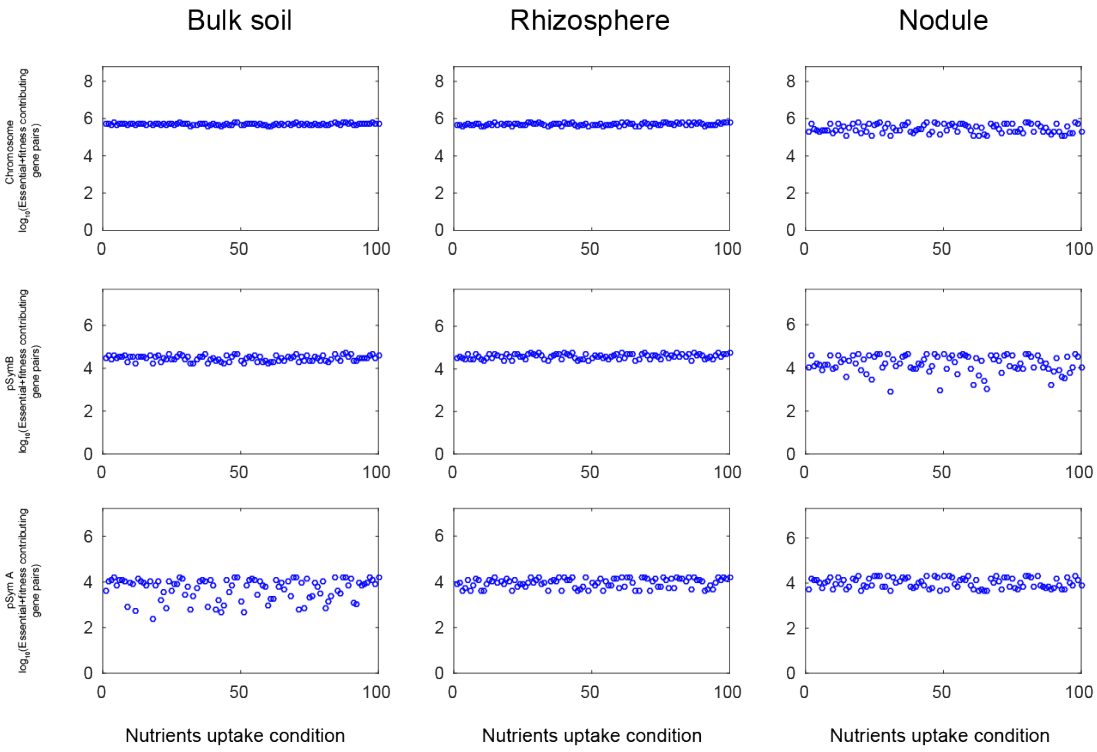

Bulk soil
Rhizosphere
Nodule
Predicted growth rate (h-1)
Essential genes
Chromosome essential and fitness promoting genes
pSymB essential and fitness promoting genes
pSymA essential and fitness promoting genes
Nutrients uptake condition
Nutrients uptake condition
Nutrients uptake condition

Figure 3.S8. Robustness to nutrient composition variation of fitness costs associated with replicon specific double gene deletions.

To account for the influence of the nutrient composition of the three niches analysed (bulk soil, rhizosphere, and nodule) 100 iterations were performed and, for each iteration, new random uptake rates were generated. For each of these iterations, all the replicon specific gene pairs were removed from the model and the fitness (flux through objective function in the mutant / flux through the objective function in the wild type) of each double mutant was calculated. This, in turn, was repeated for each ecological niche (bulk soil, rhizosphere, and nodule). Essential and fitness promoting gene pairs include all gene pairs whose deletion decreases flux through the objective function by a value greater than or equal to 1%.


Bulk soil
Rhizosphere
Nodule
Chromosome
log10(Essential+fitness contributing gene pairs)
pSymB
log10(Essential+fitness contributing gene pairs)
pSym A
log10(Essential+fitness contributing gene pairs)
Nutrients uptake condition
Nutrients uptake condition
Nutrients uptake condition

3.9.4 Supplementary data sets

All supplementary data sets can be found online at:

www.nature.com/articles/ncomms12219

Data set 3.S1. Contains the analyzed data of all Phenotype MicroArray^{TM} data (experimental and in silico) used or generated in this study.

Data set 3.S2. Contains all the raw Phenotype MicroArray^{TM} data that was generated in this study in the form of .csv files. Related descriptive files and summary files are also included.

Data set 3.S3. Contains tables that list all the reactions showing different characteristics during growth in bulk soil versus the rhizosphere.

Data set 3.S4. Contains tables that list all the reactions showing different characteristics during growth in the rhizosphere versus the nodule.

Data set 3.S5. Contains previously published pangenome information and regulon data for all genes that are included in iGD1575.

Data set 3.S6. The sbml file of iGD1575.

Data set 3.S7. Contains the COG annotations for all genes included within iGD1575, as generated by WebMGA.

CHAPTER 4. GENOMIC RESOURCES FOR IDENTIFICATION OF THE MINIMAL N_2-FIXING SYMBIOTIC GENOME

Citation: diCenzo GC, Zamani M, Milunovic B, Finan TM. 2016. Genomic resources for identification of the minimal N_2-fixing symbiotic genome. Environ Microbiol. 18(8):2534-2547.

4.1 Preface

In this chapter, a detailed phylogenetic analysis of the essential gene region of pSymB is presented. Previously, we experimentally identified two essential genes on pSymB, *engA* and the sole copy of a tRNA with the CCG anticodon (diCenzo *et al.*, 2013). I had previously shown that these two genes are part of a gene region we referred to as the *engA*-tRNA-*rmlC* region, or ETR region for short, that appeared to have translocated from the chromosome to pSymB in an ancestral species. A more detailed analysis of the phylogenetic history of this region is presented in this chapter. The ETR region from 15 *Sinorhizobium* or *Ensifer* strains was delineated, and alignments of the entire region were produced. The analysis confirmed that the ETR region did indeed translocate from the chromosome to pSymB, and not *vice versa*, and that the translocation appeared to co-incide with the split of *Sinorhizobium* into two clades. This analysis provides perhaps the best example of how secondary replicons can become essential for cell viability through the transfer of genetic content between a secondary replicon and a co-evolving chromosome. In this way, these data supports the tenant of the model described in Chapter 2 that states that secondary replicons gain essential functions and become more integrated into core cellular metabolism over time through inter-replicon gene flow.

This chapter also contains an analysis of the symbiotic phenotype of a set of *S. meliloti* large-scale deletion mutants that cumulatively remove greater than 95% of pSymA and pSymB. Results of the analysis suggest that outside of the known symbiotic

genes, none of the pSymA genes contributed to symbiosis with alfalfa in the given conditions, whereas many of the pSymB deletions had moderate effects. These results are consistent with the multipartite evolution model described in Chapter 2. Outside of the genes necessary to inhabit the nodule, none of the genes on pSymA, whose genetic content is highly variable, are globally required for nitrogen fixation but may instead benefit specific symbioses. In contrast, the older and more genetically stable pSymB replicon has more of an universal role, likely as it is more integrated into the core *S. meliloti* metabolism that has been co-opted to produce a successful symbiosis. These results can be further interpreted to suggest that megaplasmids provide a fitness advantage in a very particular environment whereas chromids provide an advantage in a more general niche.

In addition, this chapter contributes to the secondary theme of this thesis: forward genetic analyses of symbiosis and engineering synthetic symbioses. This chapter outlines the construction of genomic tools necessary to undertake such studies. The work with the ETR region allowed the identification of the most ancestral version of it in the extant *Sinorhizobium*, the introduction of this region into the *S. meliloti* chromosome, and the construction of a ΔpSymAB derivative carrying the ETR region in the chromosome. This strain will facilitate the identification of the horizontally acquired necessary and sufficient symbiotic gene set. Furthermore, the screen of the deletion library strains for symbiotic phenotypes with alfalfa delineated an initial target region for this necessary and sufficient gene set. Together, these represent two of the important steps in working towards the

synthetic engineering of symbiosis.

4.2 Abstract

The lack of an appropriate genomic platform has precluded the use of gain-of-function approaches to study the rhizobium–legume symbiosis, preventing the establishment of the genes necessary and sufficient for symbiotic nitrogen fixation (SNF) and potentially hindering synthetic biology approaches aimed at engineering this process. Here, we describe the development of an appropriate system by reverse engineering *Sinorhizobium meliloti*. Using a novel *in vivo* cloning procedure, the *engA*-tRNA-*rmlC* (ETR) region, essential for cell viability and symbiosis, was transferred from *Sinorhizobium fredii* to the ancestral location on the *S. meliloti* chromosome, rendering the ETR region on pSymB redundant. A derivative of this strain lacking both the large symbiotic replicons (pSymA and pSymB) was constructed. Transfer of pSymA and pSymB back into this strain restored symbiotic capabilities with alfalfa. To delineate the location of the single-copy genes essential for SNF on these replicons, we screened a *S. meliloti* deletion library, representing > 95% of the 2900 genes of the symbiotic replicons, for their phenotypes with alfalfa. Only four loci, accounting for < 12% of pSymA and pSymB, were essential for SNF. These regions will serve as our preliminary target of the minimal set of horizontally acquired genes necessary and sufficient for SNF.

4.3 Introduction

The rhizobium – legume symbiosis is an agriculturally and ecologically important biological process. This complex relationship begins in the soil, where the rhizobia must

successfully compete for occupancy in the rhizosphere. Following an exchange of signals between the rhizobial and the legume partners, the rhizobia infect the plant root and differentiate into bacteroids within the plant cells of a specialized organ known as the nodule (Garg & Geetanjali, 2007; Oldroyd *et al.*, 2011). Here, the rhizobia convert atmospheric N_2 gas to ammonia (NH_3) for assimilation by the legume host. In 2008, biological nitrogen fixation in crop legumes was estimated to fix ~ 21 teragrams of nitrogen globally per year (Herridge *et al.*, 2008), equivalent to > 10 billion US dollar worth of nitrogen fertilizer.

For decades, the study of the steps and underlying genetics of this interaction has been of interest to many researchers, with a long-term goal of manipulating this process for agricultural gains. This includes developing improved biological inoculants as well as engineering symbiotic nitrogen fixation (SNF) in non-legumes such as cereals (Archana, 2010; Oldroyd & Dixon, 2014; Rogers & Oldroyd, 2014). Doing so will require intimate knowledge of the genetic basis of the interaction through-out all stages of the relationship. Although great strides have been made in this area, the lack of an appropriate genomic platform for forward genetics studies has limited certain analyses, for example, establishing the sufficiency of the identified symbiotic gene complement. Previously, researchers have attempted to use non-rhizobia that are closely related to the rhizobia as the genetic background, and have transferred entire large symbiotic plasmids into these species (for example, see Marchetti *et al.*, 2010; Martínez *et al.*, 1987; Rogel *et al.*, 2001). However, the resulting strains either show incomplete or poor symbiotic phenotypes

relative to the donor rhizobia. It is difficult to draw conclusions as to why the resulting strain is a poor symbiont in such experiments; e.g. were relevant genes not transferred or was there a lack of recruitment of housekeeping genes into symbiotic regulons? As a result, non-rhizobia do not show much promise as the background genome in which to perform forward genetic analysis of SNF.

We believe a more promising alternative is to reverse engineer a rhizobium into a non-nitrogen fixer. Rhizobia acquired their symbiotic genes via horizontal gene transfer (HGT) and these genes are generally located either on large plasmids or as symbiotic gene islands (MacLean *et al.*, 2007). Thus, the complete removal of these plasmids or islands should remove most, if not all, the horizontally acquired genes related to symbiosis. Such a strain would provide an appropriate genomic background for gain-of-function (forward) genetic analyses of SNF as the chromosomal background is known to be permissive to SNF, and the plasmids or islands removed represent a genetic pool that contains all genes necessary for SNF.

The genome of a model rhizobium, *Sinorhizobium meliloti* Rm1021, is divided into three replicons: a 3.7 Mb chromosome, a 1.7 Mb chromid (pSymB) and a 1.4 Mb megaplasmid (pSymA) (Galibert *et al.*, 2001). Whereas the chromosome was primarily inherited via vertical transmission, the majority of pSymA and pSymB were acquired via (ancient) HGT (Galibert *et al.*, 2001; Wong & Golding, 2003), and both are essential for symbiosis (Barnett *et al.*, 2001; Finan *et al.*, 2001). We recently reported the construction of a *S. meliloti* strain lacking both pSymA and pSymB (termed ΔpSymAB; diCenzo *et al.*,

2014), and thus the majority, if not all, of the horizontally acquired genes required for symbiosis. However, pSymB contains a 129 kb region that includes several symbiotically important genes, including the housekeeping gene *bacA* (diCenzo *et al.*, 2013). Phylogenetic analysis showed that 69 kb of this region translocated from the chromosome (diCenzo *et al.*, 2013), and the absence of this region in the ΔpSymAB strain limits its potential for use in forward genetic analysis of SNF.

Here, we describe two fundamental steps towards defining the rhizobial minimal symbiotic genome. First, we detail the construction of a *S. meliloti* ΔpSymAB derivative in which the entire 69 kb region has been re-introduced into the chromosome, and confirm this mutant as a viable platform for gain-of-function analysis of the rhizobium–legume symbiosis. Second, we delineated the pSymA and pSymB regions that carry single-copy genes essential for symbiosis with *Medicago sativa* (alfalfa). These data will serve as the basis for an ongoing project to use the ΔpSymAB strain reported here to identify the minimal N_2-fixing symbiotic gene complement.

4.4 Results

4.4.1 Reconstruction of the ancestral engA-*tRNA*-rmlC *(ETR) region*

The ETR region was originally identified as a pSymB locus carrying essential genes (diCenzo *et al.*, 2013). Initial comparative genomics revealed that this region was chromosomally situated between the *kdgK* and the *dppF2* genes in a *S. meliloti* ancestor and translocated to pSymB as part of a 69 kb translocation (diCenzo *et al.*, 2013). To identify which sequenced *Sinorhizobium* isolate carries an ETR region most similar to the

predicted ancestral ETR region, a detailed phylogenetic analysis was performed.

The analysis included all the *Sinorhizobium* and *Ensifer* species with a fully assembled genome, as well as some species still in contigs if the entire ETR region was present on a single contig (Figure 4.1). This phylogeny suggested the ancestor common to the nitrogen fixing *Sinorhizobia* was present at the base of the *Sinorhizobium* following the divergence from *Ensifer adhaerens*. Alignments of the ETR region revealed that a core region was conserved across all strains, although some minor differences were present between clades 1 and 2 (Figure 4.1). However, all the clade 1 strains had at least one large strain- or species-specific insertion, as did *Sinorhizobium fredii* USDA 257, which was not desired. Of the remaining four organisms, the ETR region was strongly conserved between *S. fredii* NGR234, *S. fredii* HH103 and *Sinorhizobium americanum* CCGM7, with the ETR region of *Sinorhizobium terangae* WSM1721 showing slight deviation from the others. As the ETR region of *S. terangae* seemed more divergent from the ETR region of clade 1 strains, it was reasoned that the other three (NGR234, HH103 and CCGM7) were likely the best representation of the ETR region in the ancestral *Sinorhizobium*. As each appeared equally good, we chose to clone the ETR region of *S. fredii* NGR234 as it is a model organism.

4.4.2 Development of an in vivo *cloning and genome manipulation technique*

To transfer the 69 kb ETR region from the chromosome of *S. fredii* NGR234 to the chromosome of *S. meliloti*, we developed an *in vivo* cloning method for *Sinorhizobium*. We previously exploited the Flp/FRT recombination system to make

large, defined deletions in the *S. meliloti* genome (Milunovic *et al.*, 2014), and had adapted the φC31 integrase system for the stable integration of plasmids into the *S. meliloti* chromosome (diCenzo *et al.*, 2013). We therefore combined these two systems into a streamlined protocol for the *in vivo* cloning of large DNA fragments followed by the stable integration of this DNA into the desired recipient.

As summarized in the experimental procedures and Figures 4.2, 4.S1 and 4.S2, the region to be cloned are flanked by FRT (Flippase recognition target) sites through single cross-over plasmid integration. The resulting strain carries FRT sites flanking not only the region to be cloned, but also p15A and ColE1 *oriV* and *rep* genes, two RK2 *oriT* sequences, neomycin (Nm) and gentamicin (Gm) resistance genes (*nptII* and $aacC_4$) and an *attP* sequence. Following Flp-mediated recombination between the two FRT sites, a circular DNA fragment containing the sequence of interest is excised from the genome. Although this fragment cannot replicate in *Sinorhizobium* species, it can be mobilized to *Escherichia coli* through the expression of the RK2 *tra* genes *in trans*, where it is effectively captured as a replicating plasmid. The resulting plasmid is then conjugated into an appropriate recipient that carries a landing pad that includes an *attB* site and the φC31 recombinase. The plasmid will integrate into the genome via the *attB* and *attP* sites, following which the scar regions at either ends of the insertion can be removed using *sacB*-mediated double cross-over recombination.

The effectiveness of the Flp/FRT cloning technique was initially assessed by isolating a ~ 10 kb fragment from the *S. meliloti* chromosome that included the *phoR* and

pstSCABphoUphoB operons. Using a quadriparental mating method (Figure 4.2A), we were able to recover *E. coli* transconjugants carrying the desired plasmid at a frequency on the order of 10^{-9} per *S. meliloti* donor and per *E. coli* recipient. However, when attempting to clone the 69 kb ETR region from the chromosome of *S. fredii* NGR234, we failed to isolate the desired plasmid using the quadriparental technique. We therefore switched to using a two-step procedure that involved sequential triparental matings (Figure 4.2B). As expanded on below, we successfully obtained *E. coli* transconjugants at a frequency on the order of 10^{-7} per donor *S. meliloti* and *E. coli* recipient.

4.4.3 Re-introduction of the ETR region into the S. meliloti *chromosome*

As single cross-over plasmid integrants could recombine out of the genome, it was possible that the *E. coli* transconjugants did not carry the *S. fredii* ETR region as a plasmid (pETR or pTH2938) and instead carried the two plasmids used to integrate the FRT sites. Successful polymerase chain reaction (PCR) amplification of *engA* and *rmlC* from the plasmid DNA purified from the *E. coli* transconjugants and, more importantly, 100% linkage of the Gm and kanamycin (Km) resistance genes following transformation of *E. coli* DH5α with the purified pETR provided evidence that the plasmid in the *E. coli* transconjugants was indeed pETR (data not shown).

Following transfer of pETR to *S. meliloti* RmP3331 (*attB*, φC31 integrase), transconjugants were recovered at a frequency of 10^{-4}, whereas transconjugants were recovered at a reduced frequency of just 10^{-7} into wild type RmP110. This suggested that ~ 99.9% of transconjugants had the plasmid integrated into the *attB* site and not

elsewhere in the genome (e.g. the ETR region on pSymB). As there was 13% divergence in the sequence identity of the conserved portion of the ETR region between *S. meliloti* Rm2011 and *S. fredii* NGR234, it was not surprising that little homologous recombination between pETR and pSymB was detected. The ability to transduce the ΔB179 deletion, which removed the pSymB-encoded copies of the essential *engA* and $tRNA^{arg}$ gene, into one of the transconjugants confirmed that pETR integrated at the *attB* site and that the *S. fredii engA* and $tRNA^{arg}$ genes properly complemented the deletion of the *S. meliloti engA* and $tRNA^{arg}$ genes (data not shown). Although the resulting transconjugant (RmP3340) carried the ETR region in the desired location in the chromosome, unwanted scar regions that included the φC31 integrase, the *att*, p15A and ColE1 sequences were present at both ends of the integration, one marked with Sp^R and one with Nm^R Gm^R (Figure 3.S2). Therefore, a seamless integrant (designated as RmP3380 or $Rm2011_{\Omega NGR69}$) was constructed by removing the scar regions via recombination using *sacB* negative selection (see the experimental procedures; Figures 4.S1 and 4.S2). The co-transduction frequency (0.153, 46/300) of the spectinomycin (Sp) and Nm/Gm resistances from RmP3340 into $Rm2011_{\Omega NGR69}$ was not statistically different from the expected frequency (0.161) calculated with the modified Wu formula (Sanderson & Roth, 1988). This suggested that no large deletions were present in the cloned ETR region present in the *S. meliloti* chromosome. Subsequent whole-genome sequencing confirmed the presence of the complete *S. fredii* NGR234 ETR region, and unpublished RNA-seq data showed that the genes of the *S. fredii* ETR region were expressed in the *S. meliloti* host.

Sinorhizobium meliloti Rm2011 derivatives that contain the ΩNGR69 insertion but lack pSymA (RmP3409 or $Rm2011_{\Omega NGR69}$ ΔpSymA), pSymB (RmP3413 or $Rm2011_{\Omega NGR69}$ ΔpSymB), or pSymA + pSymB (RmP3414 or $Rm2011_{\Omega NGR69}$ ΔpSymAB) were made as described in the experimental procedures. The removal of pSymB was confirmed based on the inability to grow with protocatechuate (PCA) (MacLean *et al.*, 2006), maltitol (Ampomah *et al.*, 2013), succinate (Finan *et al.*, 1988), glycerol (diCenzo *et al.*, 2014), melibiose (Gage & Long, 1998), hydroxyproline (MacLean *et al.*, 2009), or taurine (Mostafavi *et al.*, 2014) as carbon sources, thiamine auxotrophy (Finan *et al.*, 1986), and a requirement for cobalt supplementation of a yeast extract – tryptone based medium (LB) (Cheng *et al.*, 2011; data not shown). Further confirming the proper expression of the *S. fredii* ETR region, the $Rm2011_{\Omega NGR69}$ ΔpSymB and $Rm2011_{\Omega NGR69}$ ΔpSymAB strains were able to grow with L-arabinose as the sole carbon source, unlike our previous ΔpSymB (RmP3009) and ΔpSymAB (RmP2917) strains that lack the ETR region, as the transport and catabolism of L-arabinose is located within the ETR region (data not shown) (diCenzo *et al.*, 2013; Poysti *et al.*, 2007). Additionally, integration of the entire ETR region into the chromosome alleviated part of the growth and density defects observed in LBmc when pSymB is removed from *S. meliloti* (Table 4.1 and data not shown).

4.4.4 Re-introduction of pSymA and pSymB into $Rm2011_{\Omega NGR69}$ ΔpSymAB

To use $Rm2011_{\Omega NGR69}$ ΔpSymAB as a platform for forward genetic analyses of SNF, it was necessary to establish that SNF could be restored to this strain. Therefore,

the protocol of Nogales and colleagues was used to transfer the entire pSymA and/or pSymB replicons back into this strain (Nogales *et al.*, 2013). This involved mobilizing these replicons from *S. meliloti* Rm5000 [rifampicin (Rif^R), Sm^S; Finan *et al.*, 1984] through overexpression of *rctB*, whose protein product negatively modulates the activity of the conjugal repressor protein RctA (Blanca-Ordóñez *et al.*, 2010; Nogales *et al.*, 2013; Pérez-Mendoza *et al.*, 2005). By selecting transconjugants based on carbon and vitamin phenotypes (see the experimental procedures), we were able to select for the gain of pSymA and/or pSymB replicons that were not genetically modified. The pSymA and/or pSymB replicons were transferred to both $Rm2011_{\Omega NGR69}$ ΔpSymAB (RmP3414) and our original Rm2011 ΔpSymAB strain (RmP2917; diCenzo *et al.*, 2014). Transconjugants were obtained at a frequency of 10^{-2} to 10^{-3} per recipient, and 10^{-3} to 10^{-5} per donor. The genotypes of all transconjugants expected to carry either pSymA, pSymB, or both were confirmed with whole-genome sequencing. In all cases, mapping the reads to the reference *S. meliloti* Rm2011 genome (Sallet *et al.*, 2013) indicated that no deletions were present on the transferred pSymA and pSymB replicons. Inoculating alfalfa plants with the $Rm2011_{\Omega NGR69}$ ΔpSymAB transconjugants revealed that introduction of pSymA was sufficient to restore the Nod^+ phenotype, although the strain remained Fix^-, whereas re-introduction of both pSymA and pSymB fully restored it's symbiotic capabilities (Table 4.2). Thus, $Rm2011_{\Omega NGR69}$ ΔpSymAB can serve as an appropriate genomic background for forward genetic analyses of SNF.

A second objective of re-introducing pSymA and pSymB into both $Rm2011_{\Omega NGR69}$

ΔpSymAB (RmP3414) and Rm2011 ΔpSymAB (RmP2917) was to construct strain sets with highly isogenic chromosomal backgrounds. Analysis of the genome sequences for the original Rm2011, ΔpSymA (SmA818), ΔpSymB (RmP3009) and ΔpSymAB (RmP2917) strain set revealed numerous polymorphisms between the strains (Table 4.S1), many of which separate Rm2011 and ΔpSymB from ΔpSymA and ΔpSymAB. In contrast, only four polymorphisms were found between ΔpSymAB and the ΔpSymAB derivatives with pSymA, pSymB, or both re-introduced (Table 4.S2). Strains resulting from the reintroduction of the replicons into Rm2011 ΔpSymAB or $Rm2011_{\Omega NGR69}$ ΔpSymAB showed growth profiles distinct from the equivalent strains made through replicon removal from Rm2011 (strains RmP3380, RmP3409, and RmP3413 versus strains RmP3501-RmP3503 in Table 4.1).

4.4.5 Symbiotic phenotypes of the pSymA and pSymB deletion library mutants

To delineate the pSymA and pSymB regions that carry single-copy genes essential for symbiosis, we screened strains from our previously constructed deletion mutant library (Milunovic *et al.*, 2014) for their symbiotic phenotypes with alfalfa (Table 4.2 and Figure 4.3). Collectively, the deletions employed in this study removed 97.5% of pSymA and 95.7% of pSymB. The majority of the DNA that was not deleted is in close proximity to either the *oriV*s of pSymA or pSymB or one of the essential pSymB genes. The rest is present as 0.6–2 kb gaps between the deletions.

For the pSymA megaplasmid, 21 deletion mutants were screened and only four

deletions were found to be either Nod$^+$ Fix$^-$ or Nod$^-$ Fix$^-$ (ΔA116, ΔA117, ΔA118 and ΔA121 in Table 4.2 and Figure 4.3A). These four deletions covered two contiguous regions spanning a total of ~ 156 kb, or < 12% of pSymA. The phenotypes of these four deletion mutants were not surprising as they lost *nod*, *nif*, or *fix* genes known to be essential for SNF (Barnett *et al.*, 2001; Batut *et al.*, 1985; Rosenberg *et al.*, 1981). The other 17 deletions were statistically indistinguishable from the wild type RmP110. These results were used to guide construction of two new larger deletions (ΔA301 and ΔA303), which flank the essential symbiotic regions, by combining some of the Fix$^+$ deletions. These deletions were ~ 116 kb and ~ 354 kb, or ~ 8.6% and ~ 26.2% of pSymA, and neither showed a symbiotic defect with alfalfa based on shoot dry weight (Table 4.2 and Figure 4.3A).

In our analysis of pSymB, 15 deletion mutants were examined on alfalfa. Three of these (ΔB108, ΔB109 and ΔB123 in Table 4.2 and Figure 4.3B), forming two contiguous regions spanning a total of ~ 196 kb, or < 12% of pSymB, were not surprisingly found to be Nod$^+$ Fix$^-$ as they encompass *exo* or *dct* genes known to be required for an effective symbiosis (Finan *et al.*, 1986; 2001; Watson *et al.*, 1988). The remaining 12 pSymB deletion strains were all Nod$^+$ Fix$^+$, although five deletion mutants (ΔB106, ΔB107, ΔB116, ΔB118 and ΔB180) repeatedly showed moderate symbiotic phenotypes and one (ΔB122) showed a severe phenotype. Based on the results of this screen, a novel deletion strain (ΔB201) was constructed that lacks ~ 748 kb or ~ 44.4% of pSymB. Despite lacking nearly half of pSymB, ΔB201 was Nod$^+$ Fix$^+$ with only a slight

shoot dry weight phenotype on alfalfa (Table 4.2 and Figure 4.3).

4.5 Discussion

In this study, we constructed a *S. meliloti* derivative with the 69 kb ETR region from *S. fredii* NGR234 integrated into its ancestral location in the *S. meliloti* chromosome. The decision to clone the ETR region from *S. fredii* NGR234 was guided by a phylogenetic analysis of this gene region (Figure 4.1). While performing the phylogenetic analysis, two interesting observations were noted that held true regardless of whether the phylogeny was constructed with the EngA protein sequence, the *engA* DNA sequence, or 16S rRNA sequences (Figs 4.1 and 4.S3). First, in all phylogenies, the translocation of the ETR region to the pSymB ancestor corresponded with the divergence of the *Sinorhizobium* into two unique clades. Whether this is representative of the translocation having an evolutionary effect or is simply a coincidence remains unclear. Nevertheless, the unique genomic location of the ETR region in these two lineages means that it may be an ideal locus for further examination on the distinct evolutionary forces acting on each replicon in a multipartite genome (Cooper *et al.*, 2010; Galardini *et al.*, 2013). Second, *E. adhaerens* diverged from the nitrogen fixing *Sinorhizobium* at the base of all constructed phylogenies, consistent with *Ensifer* and *Sinorhizobium* referring to closely related, but distinct, sister taxa. Both of these findings are well supported by previously constructed phylogenies using multi-locus sequence analyses (Degefu *et al.*, 2012; Martens *et al.*, 2007; 2008; Mousavi *et al.*, 2015; Rudder *et al.*, 2014), as well as by phylogenomic analyses (Figure 4.S4; Ormeño-Orrillo *et al.*, 2015).

Throughout this study, we worked with strains that were genotypically wild type, ΔpSymA, ΔpSymB or ΔpSymAB, made either through the removal of the appropriate replicon(s) from the wild type or through re-introduction of the desired replicons back into the ΔpSymAB strain. We were surprised to find that the growth phenotype of genotypically equivalent strains displayed distinct growth rates (strains RmP3380, RmP3409 and RmP3413 versus strains RmP3501–RmP3503 in Table 4.1). This was particularly evident in comparing wild type versus ΔpSymA and ΔpSymB versus ΔpSymAB. Whereas the strains produced through replicon re-introduction are highly isogenic (Table 4.S2), the strains produced through replicon removal carry numerous polymorphisms (Table 4.S1). Thus, it is likely that the growth differences among the equivalent strains produced through replicon re-introduction should solely reflect the effect of the presence/absence of pSymA and/or pSymB. In contrast, a portion of the growth rate differences among the wild-type, ΔpSymA, ΔpSymB and ΔpSymAB strains produced through replicon removal can likely be explained by the chromosomal polymorphisms, highlighting the importance of ensuring chromosomal homogeneity when comparing phenotypes between strains. Moreover, we observed that strains lacking pSymB grew faster (Table 4.1) and to a higher density in LBmc medium, but not M9 sucrose medium, if they carried the ETR region integrated into the chromosome. Given that this phenotype was specific to the complex medium (Table 4.1), and that the growth of pSymB-cured strains is carbon limited in LBmc (Fei *et al.*, 2016), we speculate this phenotype is due to the ETR region allowing the catabolism of additional, preferred

carbon sources in the LBmc medium, such as L-arabinose (Poysti *et al.*, 2007).

We effectively generated quantitative symbiotic shoot dry weight data for more than 2500 pSymA and pSymB genes through screening the deletion library mutants for their symbiotic effectiveness with alfalfa. All single-copy genes essential for N_2-fixing symbiosis with alfalfa on pSymA and pSymB were present in a total of four regions accounting for less than 12% of these replicons (Table 4.2 and Figure 4.3), and all of these phenotypes were expected based on the presence of known SNF genes (Batut *et al.*, 1985; Finan *et al.*, 1986; Rosenberg *et al.*, 1981; Watson *et al.*, 1988). The lack of additional symbiotic genes is largely consistent with previous work (Charles & Finan, 1991; Yurgel *et al.*, 2013), although these other studies did not provide quantitative data for all the deletions thus preventing comparison of intermediate phenotypes. The one exception was *phoCDET* on pSymB that was initially identified as being essential for SNF (Bardin *et al.*, 1996; Charles & Finan, 1991), which follow-up work showed to be dependent on a *pstC* mutation in Rm1021 that is corrected in RmP110 (our wild type) (Yuan *et al.*, 2006). None of the other pSymA deletion derivatives, even those removing up to ~ 25% of pSymA, showed a symbiotic phenotype statistically distinguishable from the wild type (Figure 4.3 and Table 4.2). In contrast, many of the Fix^+ pSymB deletion mutants displayed moderate, but noticeable, symbiotic defects on alfalfa (Figure 4.3 and Table 4.2). Only one deletion (ΔB122) showed a severe shoot dry weight phenotype, and we are currently following up on the genetic/biochemical basis for this phenotype.

The phenotypes of the 32 Fix^+ deletion mutants have some interesting implications

in relation to *S. meliloti* genome evolution and the genomics of SNF. The lack of phenotypes suggest that aside from the classical symbiotic genes, pSymA-encoded genes are generally not directly involved in nitrogen fixation. However, many are likely to contribute to the nodule occupancy competiveness of *S. meliloti* (Pobigaylo *et al.*, 2006) or provide host-specific benefits not detected on alfalfa (Ardourel *et al.*, 1995; diCenzo *et al.*, 2015). This could potentially explain, in part, the high variability in the gene content of pSymA in various nodule isolates (Epstein *et al.*, 2012; Galardini *et al.*, 2013); most pSymA genes provide little global symbiotic advantage and so are gained or lost as necessary to better compete for nodule occupancy and to improve N_2 fixation in each individual environment. In contrast, the gene content of pSymB is more strongly conserved than that of pSymA (Epstein *et al.*, 2012; Galardini *et al.*, 2013), and encodes a number of 'core' enzymes and pathways that are expected to contribute to the fitness of both free-living and symbiotic *S. meliloti* (Finan *et al.*, 2001). Thus, we expect that most of the moderate symbiotic phenotypes (shoot dry weights between 50% and 70% of the wild type) associated with the pSymB deletion mutants (ΔB106, ΔB107, ΔB116, ΔB118 and ΔB180) can be attributed to the loss of such 'core' genes that have been recruited to fulfill a need during symbiosis. Overall, we interpret the results of these symbiotic assays as suggesting that gain of the classical symbiotic genes is sufficient to confer SNF capabilities. Universal improvements in this symbiosis then occur primarily through adaptive evolution of the chromosomal backbone (Marchetti *et al.*, 2010), better integration of core genes into symbiotic regulons, and evolution of the symbiotic genes,

whereas further gene gain largely provides environment- or symbiosis-specific enhancements.

There were two primary objectives of this work. The first was to construct a strain (Rm2011$_{\Omega NGR69}$ ΔpSymAB) that could effectively be used in forward genetic analysis of SNF, particularly for the identification of the minimal, horizontally acquired, N_2-fixing symbiotic gene set. This constructed strain overcomes the primary complication of attempting to perform such work in a species that does not naturally participate in SNF. That is, there are many possible explanations of a negative result, i.e. lack of SNF following introduction of a suite of SNF genes into a SNF-naïve species, making interpretation and problem shooting difficult. These interpretations include that not all the genes are expressed, that all essential SNF genes were not introduced, or that the full set of dedicated symbiotic genes is present and expressed but housekeeping chromosomal genes are not properly integrated into the symbiotic regulons. These issues are avoided by using the Rm2011$_{\Omega NGR69}$ ΔpSymAB strain reported here, which was engineered to contain all chromosomal *S. meliloti* genes (the chromosome and the ETR region) while lacking the rest of the horizontally acquired pSymA and pSymB replicons. We showed that re-introduction of all ~ 2900 genes of pSymA and pSymB into Rm2011$_{\Omega NGR69}$ ΔpSymAB can transform this strain into a fully effective symbiont of alfalfa (Table 4.2). In essence, this serves as a positive control and means that the only explanation for the lack of an effective symbiosis following the re-introduction of a subset of these 2900 genes is the absence of at least one necessary symbiotic gene.

An alternative approach would be to isolate a natural non-symbiotic strain of a rhizobial species from soil, as such strains can be prevalent and transfer of the symbiotic plasmid/island from strains of the same species can confer SNF (Segovia *et al.*, 1991; Sullivan *et al.*, 1995). However, using the strain reported here has advantages. These non-symbiotic rhizobia usually only lack the most recently acquired symbiotic plasmid or island while retaining more ancestral plasmids required for SNF (like pSymB), thus lacking fewer genes than Rm2011$_{\Omega NGR69}$ΔpSymAB. This could hinder some subsequent analyses; as an example, it would result in a smaller number of genes identified as required for the minimal SNF gene set. Additionally, we are unaware of *S. meliloti* strains lacking the symbiotic plasmid that have been naturally isolated from the environment or any method to select for such strains, although such strains may exist (Trabelsi *et al.*, 2009). Importantly, the large number of genetic/genomic manipulation techniques and the large number of available resources, such as the deletion library, facilitate studies using the model *S. meliloti* strain Rm2011/1021 rather than another soil isolate for which genetic/genomic techniques are not developed.

The second objective was to delineate the location of the single-copy pSymA and pSymB genes essential for symbiosis, which will serve as our preliminary target for the minimal symbiotic genome. Identification of the minimal symbiotic gene complement would ensure that we are aware of all genes necessary for the establishment of a N_2-fixing symbiosis. Once identified, the construction of a broad host range plasmid carrying these genes could contribute to the development of 'personalized bioinoculants'. A primary

cause of the failure of rhizobial inoculants to significantly improve legume crop growth is the failure of these inoculants to compete well in the soil/rhizosphere with the indigenous microbiome (Archana, 2010). This often results in the inoculants being unable to establish a good foothold in the soil microbiome, and they often fail to persist in the population over long periods. This problem could be overcome by isolating the major members of an individual field, transferring the symbiotic genes to this organism and re-introducing this strain to the field (Geddes *et al.*, 2015). Additionally, a *S. meliloti* strain with just the minimal SNF gene set would represent a stripped-down version of a rhizobium, and would serve as a valuable platform for screening environmental metagenomic libraries for the identification of SNF-promoting genes, including genes contributing to improved N_2 fixation, nodule competitiveness, etc. The data reported here suggest that all single-copy genes essential or highly important for SNF on pSymA and pSymB can be localized to just 350 kb. However, it remains possible that additional genes lie within the < 5% of pSymA and pSymB that was not represented in our screen. Moreover, given the extent of redundancy within the *S. meliloti* genome (diCenzo & Finan, 2015), it is highly likely that essential symbiotic functions are encoded by duplicate genes and therefore remained undetected in the current study. Such loci will be revealed as we continue to reduce the size of pSymA and pSymB to just the essential 350 kb identified here.

4.6 Materials and Methods

4.6.1 Media, growth condition, and bacterial strains

All media (LB, LBmc, TY, M9) were prepared as previously described (diCenzo

et al., 2014). The concentration of the carbon sources in the M9 minimal media was 5 or 10 mM, as described previously (Mauchline *et al.*, 2006). Most antibiotic concentrations for *S. meliloti* and *E. coli* [streptomycin (Sm), Sp, Nm, Km, tetracycline (Tc), Gm and chloramphenicol (Cm)] and growth conditions were as described elsewhere (diCenzo & Finan, 2015; diCenzo *et al.*, 2014). Antibiotic concentrations for *S. fredii* NGR234 were as follows: 50 μg mL^{-1} Gm, 50 μg mL^{-1} Km, and 5 μg mL^{-1} Tc. Rif was added to 50 μg mL^{-1} for *Sinorhizobium* and 20 μg mL^{-1} for *E. coli*. Growth curves were performed and analysed as described before (diCenzo *et al.*, 2014). Bacterial strains and plasmids are listed in Table 4.S3.

4.6.2 Genetic manipulations

General DNA manipulations and recombinant techniques, bacterial conjugation, isolation of genomic *S. meliloti* DNA, and ΦM12 transductions were performed as previously described (Cowie *et al.*, 2006; Finan *et al.*, 1984; Milunovic *et al.*, 2014; Sambrook *et al.*, 1989). Mating spots for transfer of plasmids to/from *E. coli* that involved *S. meliloti* were performed on LB or LBmc, whereas mating spots involving *S. fredii* were performed on TY. DNA cloning, oligonucleotide sequences (Table 4.S4) and plasmid construction are described in Section 4.10.

4.6.3 Symbiotic assays

Plant growth experiments were performed largely as described previously (diCenzo *et al.*, 2015; Yarosh *et al.*, 1989). Briefly, ~ 5×10^8 colony-forming units of *S. meliloti* were added to each of the sterile Magenta jars containing six to eight alfalfa (*M.*

sativa cv. Iroquois) seedlings, quartz sand, vermiculite and Jensen's medium (Jensen, 1942). Plants were grown for 26 - 28 days in a Conviron growth chamber with a day (18 h, 21°C) and night (6 h, 17°C) cycle, and the shoot dry weight was determined after drying at 55°C for 10 days.

4.6.4 Genome sequencing and analysis

S. meliloti genomic DNA was isolated as previously described (Cowie *et al.*, 2006). Genomic DNA samples were sequenced at the Farncombe Family Digestive Health Research Institute at McMaster University using an Illumina HiSeq. Sequencing reads were mapped to the Rm2011 reference genome using Geneious R8 (mapper – Geneious; sensitivity – medium-low sensitivity/fast; fine tuning – up to five iterations; reads not trimmed), and the presence/absence of reads mapping to pSymA/pSymB was examined. Polymorphisms were identified using the Geneious 'Find Variations/SNPs' function following the mapping of reads to the Rm2011 reference genome (Sallet *et al.*, 2013).

4.6.5 Sequence analysis of the ETR region

All genome/contig sequences used in this study were downloaded from the National Center for Biotechnology Information (NCBI) Genome database. The borders of the ETR region in each genome/contig were identified following whole-genome alignment with MAUVE using the progressiveMAUVE algorithm and default settings (Darling *et al.*, 2010). The sequence corresponding to the ETR regions in each genome/contig was extracted, pairwise alignments produced with DoubleACT (http://

www.hpa-bioinfotools.org.uk/pise/double_act.html), and alignments viewed with the Artemis Comparison Tool (Carver *et al.*, 2005). Phylogenies of the species of interest were constructed with the MEGA 5.2.2 package (Tamura *et al.*, 2011), multiple sequence alignments were constructed with MUSCLE (Edgar, 2004) and phylogenetic trees were produced using both the maximum likelihood or minimum evolution algorithms in MEGA 5.2.2. All accession numbers for the genomes/contigs used in this study are listed in Table 4.S5 (Galardini *et al.*, 2011; Galibert *et al.*, 2001; González *et al.*, 2006; Martínez-Abarca *et al.*, 2013; Mora *et al.*, 2014; Reeve *et al.*, 2006; 2010; Rudder *et al.*, 2014; Sallet *et al.*, 2013; Schmeisser *et al.*, 2009; Schneiker-Bekel *et al.*, 2011; Schuldes *et al.*, 2012; Toro *et al.*, 2014; Weidner *et al.*, 2012; 2013).

4.6.6. Flp-mediated in vivo *cloning of FRT-flanked regions*

Two methods were employed for *in vivo* cloning of large DNA fragments from species of the *Sinorhizobium* genus. In the first procedure (Figure 4.1A), a single quadriparental mating was performed. This mating consisted of a helper strain (*E. coli* MT616), an *E. coli* strain carrying a plasmid (pTH1944) that constitutively expressed *flp* (Milunovic *et al.*, 2014), a *Sinorhizobium* strain with two FRT sites in direct orientation, and Rif^R *E. coli* DH5α. In the second method (Figure 4.2B), two consecutive triparental matings were performed. In the first mating, the plasmid pTH2505 (Zhang *et al.*, 2012), carrying *flp* under control of a PCA-inducible promoter, was transferred to the desired *Sinorhizobium* strain. The second conjugation included the *Sinorhizobium* transconjugant from the first mating, *E. coli* MT616 and Rif^R *E. coli* DH5α, and was performed on media

containing 2.5 mM PCA to induce expression of *flp*. Regardless of the method used, the cloned region was isolated as a Gm^R Km^R circular plasmid in the Rif^R *E. coli* DH5α.

4.6.7. Construction of a S. meliloti *strain with a landing pad for the integration of the ETR region from* S. fredii *NGR234*

Plasmid pTH2937 contained two fragments homologous to the chromosome of *S. meliloti* Rm2011 (nt. 2,601,218 – 2,602,232 and 2,603,745 – 2,604,747), and these fragments flanked an *attB* site, the ϕC31 integrase and a Sm^R/Sp^R cassette. This plasmid was conjugated into *S. meliloti* Rm2011, Sp^R single cross-over recombinants isolated on M9-hydroxyproline with Sp, and double cross-over recombinants identified by screening for a Nm^S phenotype. A double cross-over recombinant was purified and termed *S. meliloti* RmP3331.

4.6.8. Construction of S. meliloti *strains carrying the* S. fredii *NGR234 ETR region*

Construction of $Rm2011_{\Omega NGR69}$, including the cloning of the ETR region and construction of *S. meliloti* RmP3331, is represented visually in Figures 4.S1 and 4.S2. Plasmids pTH2916, containing a FRT site, and pTH2917, containing a FRT site and an *attB* site, were sequentially conjugated into *S. fredii* NGR234. The resulting strain, *S. fredii* P3285, contained two FRT sites flanking the ETR region. The plasmid pTH2505 (Zhang *et al.*, 2012) carrying a PCA-inducible *flp* gene was transferred to P3285. A resulting transconjugant was used in a triparental mating, performed on TY supplemented

with 2.5 mM PCA, with a Rif-resistant *E. coli* DH5α and *E. coli* MT616 (pRK600). This resulted in the excision of the ETR region from the *S. fredii* chromosome as a plasmid (pETR or pTH2938), the mobilization of pETR by pRK600, and the capture of pETR in the Rif-resistant DH5α. Plasmid pETR was conjugated into *S. meliloti* RmP3331, and transconjugants were isolated and purified on LBmc Sm Nm. The resulting strain (*S. meliloti* RmP3340) contained the *S. fredii* NGR234 ETR region integrated into the *attB* site in the chromosome, and the integration was marked by Sm^R/Sp^R on one side and Gm^R/Nm^R on the other. The Gm^R/Nm^R scar region was removed through *sacB* mediated double cross-over recombination of plasmid pTH2929 as described previously (diCenzo & Finan, 2015; Quandt & Hynes, 1993), and the double recombinants (*S. meliloti* RmP3362) were confirmed by a Tc^S Gm^S Nm^S phenotype. The Sp^R scar region was removed through *sacB*-mediated double cross-over recombination of plasmid pTH2930, and double recombinants [*S. meliloti* RmP3380 ($Rm2011_{\Omega NGR69}$)] were confirmed by a Tc^S Sp^S phenotype.

To produce *S. meliloti* strains carrying the ΩNGR69 integration but lacking pSymA and or pSymB, the Sp^R chromosomal insertion of the ETR region in *S. meliloti* RmP3362 was transduced into *S. meliloti* SmA818 (cured of pSymA) (Oresnik *et al.*, 2000). The Sp^R scar region was then removed via pTH2930 as described above, creating *S. meliloti* RmP3409 ($Rm2011_{\Omega NGR69}$ ΔpSymA). pSymB was removed from *S. meliloti* RmP3380 and RmP3409 as described previously (diCenzo *et al.*, 2014), creating *S. meliloti* RmP3413 ($Rm2011_{\Omega NGR69}$ ΔpSymB) and RmP3414 ($Rm2011_{\Omega NGR69}$ ΔpSymAB).

4.6.9. Re-introduction of pSymA and/or pSymB into S. meliloti *ΔpSymAB*$_{\Omega NGR69}$

The protocol for the conjugation of pSymA and pSymB was based on a previously described method involving RctB overexpression (Nogales *et al.*, 2013). Plasmid pTE3Yp028 (also called pTEYp028) (Pérez-Mendoza *et al.*, 2004) was transferred to *S. meliloti* Rm5000 (Rif^R Sm^S) (Finan *et al.*, 1984), creating *S. meliloti* RmP3491. An overnight culture of *S. meliloti* RmP3491 was diluted to an OD_{600} ~ 0.06 and grown to OD_{600} ~ 0.3. Next, 0.3 OD_{600} units (i.e. the equivalent of 1 ml at an OD_{600} of 0.3) of *S. meliloti* RmP3491 and of *S. meliloti* $Rm2011_{\Omega NGR69}$ ΔpSymAB were centrifuged, both pellets re-suspended in the same 1 ml of LBmc, pelleted and re-suspended in 50 µL LBmc. The entire 50 µL mixture was spotted onto a LBmc agar plate, with cobalt chloride, and incubated overnight. Following incubation, the mating spot was re-suspended in saline, serial diluted and plated on M9-trigonelline + thiamine Sm, M9-hydroxyproline Sm and M9-trigonelline Sm to select for $Rm2011_{\Omega NGR69}$ ΔpSymAB cells that had gained pSymA, pSymB, or pSymA + pSymB, respectively. Following their isolation, transconjugants were screened on all three media to confirm they only contained the desired replicons, and eventually confirmed through whole-genome sequencing.

4.6.10. Construction of deletion mutants

The majority of the deletion mutant strains used in this study were constructed

previously (Milunovic *et al.*, 2014), with newly constructed deletions making use of previously reported deletions and plasmids (Cowie *et al.*, 2006; diCenzo *et al.*, 2013; Milunovic *et al.*, 2014; Zhang *et al.*, 2012). ΔB165 was made by transducing ΔB148 into RmP2720 (ΔB142 with the *S. meliloti engA* and $tRNA^{arg}$ genes in the chromosome) and deleting the region between ΔB142 and ΔB148 with Flp via pTH2505. ΔB201 was constructed by transducing ΔB180 into RmP2720 and deleting the intervening region between ΔB142 and ΔB180 with Flp via pTH2505. The deletion ΔA301 was made by transducing pFL1439 from RmFL1439 into RmP3542 (ΔA119) and deleting the intervening region with Flp via pTH2505. To prepare the deletion ΔA302, the pTH2038 plasmid integrant from RmP1052 was transduced into RmP936 (ΔA106), following which the region between the pTH1937 plasmid integrant and ΔA106 was deleted with Flp, via pTH2505. Transducing the pFL1142 plasmid integrant from RmFL1142 into RmP3551 (ΔA302) and deleting the region between pFL1142 and ΔA302 with Flp via pTH2505 produced ΔA303. All deletions were confirmed with a mix of PCR, antibiotic resistances and carbon utilization phenotypes.

4.9 Tables and Figures

Table 4.1. Growth rates (h) of the replicon cured and replicon-re-introduced strains.

Strain name	Genotype	LBmc	M9-sucrose
Rm2011	Rm2011 wild type	1.9 (0.1)	3.6 (0.1)
SmA818	Rm2011 ΔpSymA	1.7 (0.1)	3.3 (0.1)
RmP3009	Rm2011 ΔpSymB	4.4 (0.2)	6.7 (0.2)
RmP2917	Rm2011 ΔpSymAB	3.2 (0.1)	5.3 (0.1)
RmP3380	$Rm2011_{\Omega NGR69}$	2.0 (0.1)	3.5 (0.2)
RmP3409	$Rm2011_{\Omega NGR69}$ ΔpSymA	1.6 (0.1)	3.3 (0.2)
RmP3413	$Rm2011_{\Omega NGR69}$ ΔpSymB	3.7 (0.1)	6.3 (0.3)
RmP3414	$Rm2011_{\Omega NGR69}$ ΔpSymAB	2.5 (0.1)	5.0 (0.2)
RmP3501	$Rm2011_{\Omega NGR69}$ ΔpSymAB + pSymA	2.9 (0.1)	5.4 (0.4)
RmP3502	$Rm2011_{\Omega NGR69}$ ΔpSymAB + pSymB	1.7 (0.1)	3.4 (0.1)
RmP3503	$Rm2011_{\Omega NGR69}$ ΔpSymAB + pSymAB	1.7 (0.1)	3.4 (0.1)

Values represent the average of triplicate samples with the standard deviation presented in parentheses. Growth rates were calculated between OD_{600} values of 0.1 and 0.3 for LBmc, and between 0.1 and 0.5 for M9-sucrose.

Table 4.2. Alfalfa shoot dry weights (SDW) of plants inoculated with *S. meliloti* strains.

pSymA deletion library			pSymB deletion library			Replicon re-introduction	
Strain	Deleted region (nt)	Average SDW (% RmP110)	Strain	Deleted region (nt)	Average SDW (% RmP110)	Strain	Average SDW (% RmP110)
RmP110-1	N/A	100 (22)	RmP110-2	N/A	100 (14)	RmP110-1	100 (22)
dme tme*	N/A	18 (9)¥	dme tme*	N/A	15 (4)¥	dme tme*	18 (9)¥
ΔA160	10,988-184,519	95 (16)	ΔB154†	62,137-100,636	74 (11)	Rm2011	125 (11)
ΔA106	186,200-250,917	103 (36)	ΔB141	101,396-466,499	88 (3)	$Rm2011_{\Omega NGR69}$ ΔpSymAB**	16 (2)¥
ΔA109	251,809-311,877	106 (7)	ΔB165	122,108-762,942	94 (31)	$Rm2011_{\Omega NGR69}$ ΔpSymAB + pSymA*	24 (11)¥
ΔA112	308,586-347,789	96 (10)	ΔB180	635,940-869,642	57 (15)¥	$Rm2011_{\Omega NGR69}$ ΔpSymAB + pSymB**	19 (4)¥
ΔA116*	313,654-458,916	27 (11)¥	ΔB106	870,505-1,091,104	69 (18)¥	$Rm2011_{\Omega NGR69}$ ΔpSymAB + pSymAB	157 (35)¥
ΔA117*	402,136-458,916	20 (4)¥	ΔB107	1,092,289-1,129,758	63 (12)¥		
ΔA118**	459,668-505,335	18 (4)¥	ΔB108*	1,131,168-1,169,073	15 (4)¥		
ΔA119	507,338-575,671	89 (17)	ΔB109*	1,170,466-1,204,770	13 (2)¥		
ΔA120	577,241-623,673	92 (13)	ΔB112	1,207,052-1,224,621	87 (38)		
ΔA121*	624,863-677,157	18 (5)¥	ΔB116	1,256,503-1,307,752	63 (25)¥		
ΔA122	678,150-726,673	103 (28)	ΔB118	1,323,078-1,528,150	64 (7)¥		
ΔA123	727,921-774,293	121(19)	ΔB122	1,529,711-1,572,422	32 (3)¥		
ΔA124	775,476-828,417	124 (36)	ΔB123*	1,529,711-1,652,558	17 (3)¥		
ΔA125	830,143-879,960	94 (25)	ΔB124	1,654,191-1,677,882	104 (3)		
ΔA127	881,169-930,000	104 (29)	ΔB161†	1,679,723-49,523	75 (14)		
ΔA128	927,009-1,063,642	101 (11)	ΔB201	122,108-869,642	79 (17)		
ΔA129	1,064,644-1,122,176	97 (15)					
ΔA130	1,123,504-1,173,115	104 (5)					
ΔA131	1,173,730-1,231,998	90 (8)					
ΔA132	1,232,916-1,283,082	97 (14)					
ΔA133	1,284,751-1,348,238	100 (25)					
ΔA301	507,338-623,673	94 (36)					
ΔA303	47,717-402,136	115 (21)					

Values represent the average of triplicate samples with the standard deviation presented in parentheses. All deletion library mutant strains were screened on alfalfa at least two independent times, and values from one trial per strain are shown. Plasmid re-introduction strains were screened on plants once. Absolute weights (mg $shoot^{-1}$) of wild type controls: RmP110-1 – 20.9 (4.7); RmP110-2 – 36.1 (8.2). * Nod+Fix-. ** Nod-Fix-. † In the second trial, average weight of these plants was > 80% of the RmP110 control. ¥ Statistically different from the appropriate RmP110 control as determined via an ANOVA followed by a Duncan's post hoc test with an alpha level of 0.05. See Tables 4.S5, 4.S6, and 4.S7 for statistical groupings.

Figure 4.1. Evolution of the ETR region in the *Sinorhizobium*.

The phylogenetic tree is the minimum evolution bootstrap consensus tree made from the EngA protein sequence, and is rooted with EngA from *Rhizobium etli* CFN42. Values represent bootstrap values from 1000 replicates. The asterisk represents the lowest common ancestor of the nitrogen-fixing *Sinorhizobium*. On the right, pairwise alignments of the ETR region, visualized with the Artemis Comparison Tool, are shown. The green indicates the *engA*-tRNA region, whereas the cyan indicates the tRNA-*rmlC* region. Strain names are listed following the alignments. Rm41, SM11, AK83, BL225C, Rm1021, RMO17 and GR4 are *S. meliloti* strains; LMG 14919 is a *Sinorhizobium arboris* strain; WSM419 is a *Sinorhizobium medicae* strain; CCGM7 is a *S. americanum* strain; USDA 257, HH103 and NGR234 are *S. fredii* strains; WSM1721 is a *S. terangae* strain; OV14 is an *E. adhaerens* strain.

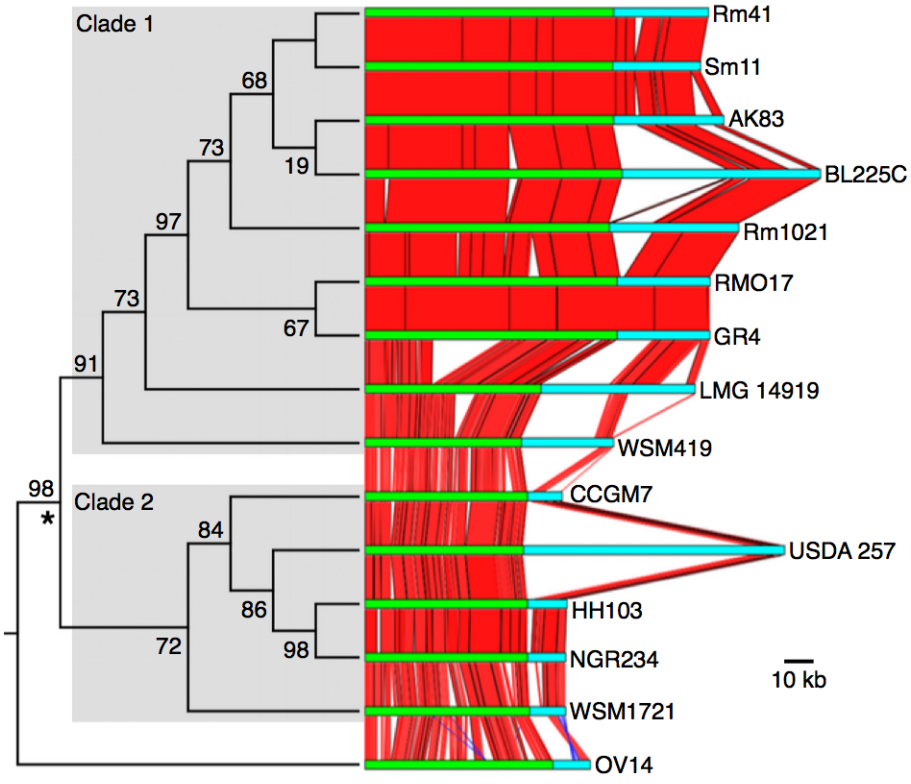

Clade 1
Clade 2
68
73
19
97
73
67
91
98
*
84
86
98
72
Rm41
Sm11
AK83
BL225C
Rm1021
RMO17
GR4
LMG 14919
WSM419
CCGM7
USDA 257
HH103
NGR234
WSM1721
OV14
10 kb

Figure 4.2. Flp/FRT mediated *in vivo* cloning.

A. In procedure A, a single quadriparental mating is performed. (**1**) In no particular order, the plasmids pTH1944 (red; mobilized by pRK600) and pRK600 (blue) are transferred to a *Sinorhizobium* strain with two FRT sites integrated into its genome. (**2**) Constitutive expression of the Flp recombinase catalyzes recombination between the two FRT sites (arrowheads), resulting in the excision of a RK2 mobilizable and narrow host range plasmid. Excision of the plasmid from the *Sinorhizobium* genome may occur prior to the gain of pRK600. (**3**) The excised plasmid is mobilized by pRK600 and transferred to DH5α, (**4**) which is selected for on LB Rif Gm. **B**. In procedure B, two consecutive triparental matings are performed. (**5**) The plasmid pTH2505 is transferred to a *Sinorhizobium* strain with two FRT sites integrated into its genome. (**6**) The resulting strain is used in a second mating where, in no particular order, pRK600 is transferred to the *Sinorhizobium* strain and expression of the *flp* gene is induced by the protocatechuate in the medium. The newly synthesized Flp recombinase mediates recombination between the two FRT sites in the genome, excising a RK2 mobilizable and narrow host range plasmid. (**7**) The excised plasmid is mobilized by pRK600 and transferred to DH5α, (**4**) which is selected for on LB Rif Gm. Black arrows indicate new experimental steps, whereas grey arrows indicate steps occurring in a single mating spot.

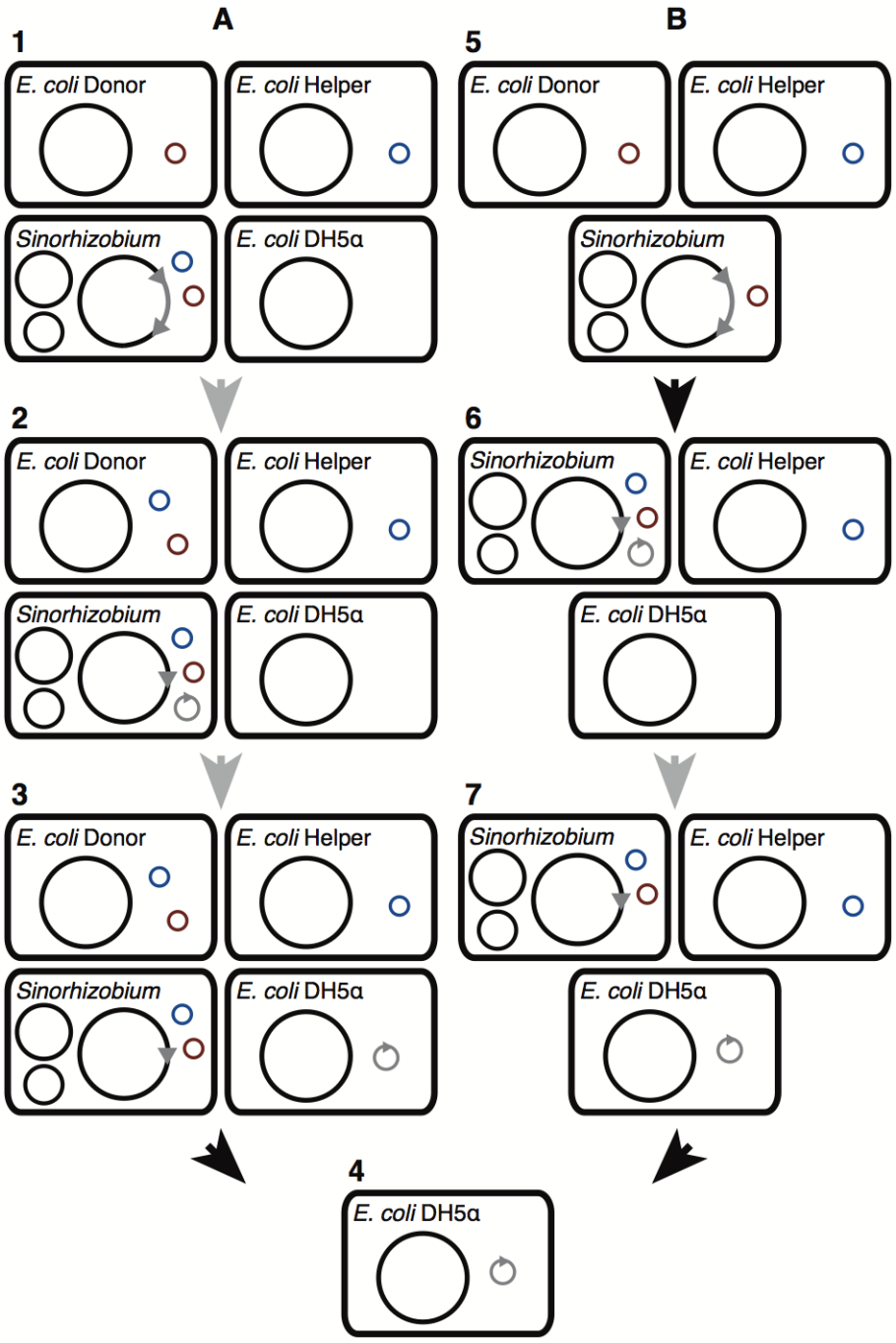
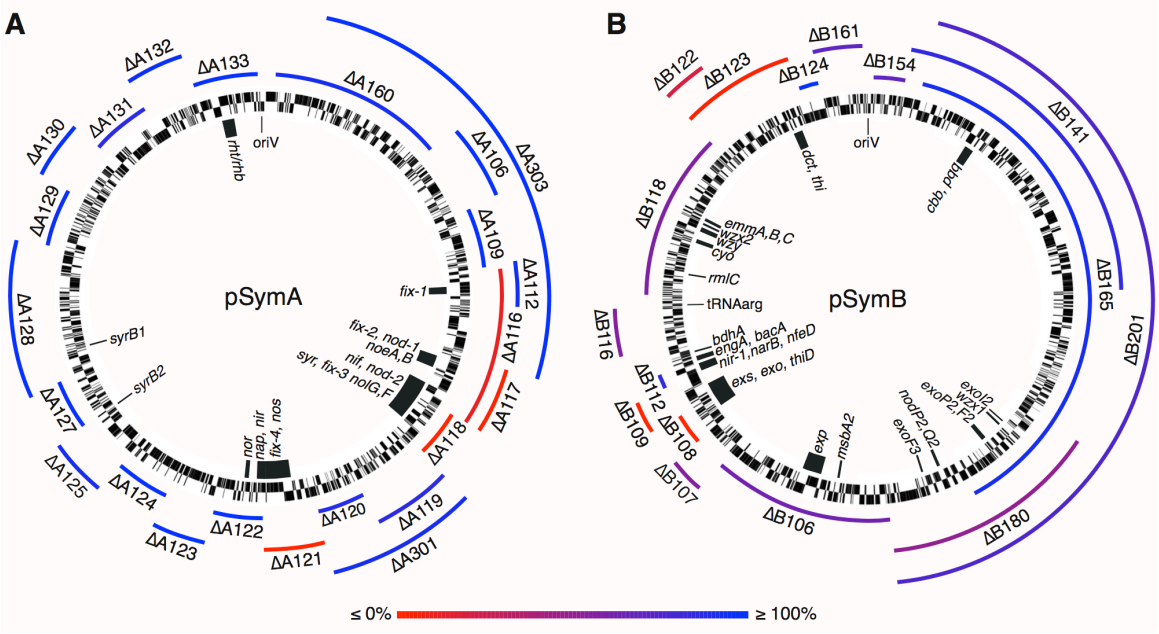

A
B
1
E. coli Donor
E. coli Helper
Sinorhizobium
E. coli DH5α
2
E. coli Donor
E. coli Helper
Sinorhizobium
E. coli DH5α
3
E. coli Donor
E. coli Helper
Sinorhizobium
E. coli DH5α
4
E. coli DH5α
5
E. coli Donor
E. coli Helper
Sinorhizobium
6
Sinorhizobium
E. coli Helper
E. coli DH5α
7
Sinorhizobium
E. coli Helper
E. coli DH5α

Figure 4.3. Schematic representation of the location of the deletions in the mutants screened on alfalfa.

Circular maps of the pSymA (**A**) and pSymB (**B**) replicons of *S. meliloti*. The inner circle represents either pSymA or pSymB, with individual lines corresponding to annotated genes. The solid outer lines highlight the position of the respective deletions. The deletions are colour coded based on their symbiotic phenotypes, as indicated at the bottom of the figure. Red indicates alfalfa plants inoculated with the deletion mutant had a shoot dry weight less than or equal to that of plants inoculated with the Fix$^-$ control (*dme tme*), whereas blue indicates the shoot dry weight of inoculated alfalfa was ≥ 100% of plants inoculated with RmP110. Several relevant loci are indicated along the inner circle, with shorthand notation referring to gene clusters as follows. **fix-1**: *fixI2,N3,O3,P3,Q3,S2.* **fix-2**: *fixK2,N2,O2,P2,Q2,T2.* **fix-3**: *fixA-C,U,X.* **fix-4**: *fixG,H,I1,J,K1,L,M,N1, fixO1,P1,Q1,S1,T.* **nif**: *nifA,B,D,E,H,K,N,X.* **nod-1**: *nodD2,L.* **nod-2**: *nodA-C,D1,D3,E-J,M,NP1,Q1.* **syr**: *syrA,M,B3.* **nos**: *nosD,F,L,R,X-Z.* **nap**: *napA-F, nnrR.* **nir-1**: *nirV,K.* **nir-2**: *nirB,D.* **nor**: *norB-E,Q.* **rht/rhb**: *rhtX,A, rhbA,C-F.* **cbb**: *cbbA,F,L,P,R,S,T,X, pqqA-E.* **exp**: *wgeA-H, wgdA,B, wggR, wgcA, wgaA,B,D-J.* **exs**: *exsA-I.* exo: *exoA,B,F1,H,I,K-Q,T-Z.* **cyo**: *cyoA-D.* **dct**: *dctA,B,D.* **thi**: *thiC,O,G,E.*


A
pSymA
oriV
rhi/rhb
fix-1
fix-2, nod-1
noeA,B
nif, nod-2
syr, fix-3 nolG,F
nor
nap, nir
fix-4, nos
syrB1
syrB2
ΔA132
ΔA133
ΔA160
ΔA303
ΔA106
ΔA109
ΔA112
ΔA116
ΔA117
ΔA118
ΔA119
ΔA301
ΔA120
ΔA121
ΔA122
ΔA123
ΔA124
ΔA125
ΔA127
ΔA128
ΔA129
ΔA130
ΔA131
B
pSymB
oriV
dct, thi
cbb, pqq
emmA,B,C
wzx2
cyo
rmlC
tRNAarg
bdhA
engA, bacA
nir-1,narB, nfeD
exs, exo, thiD
exp
msbA2
exoF3
nodP2,Q2
exoP2,T2
exoI2
wzx1
ΔB161
ΔB122
ΔB123
ΔB124
ΔB154
ΔB141
ΔB118
ΔB165
ΔB201
ΔB116
ΔB112
ΔB109
ΔB108
ΔB107
ΔB106
ΔB180
≤ 0%
≥ 100%

4.10 Supplementary Materials

4.10.1 Supplementary experimental procedures

4.10.1.1 Genetic manipulations.

General DNA manipulations, recombinant techniques, and isolation of genomic *S. meliloti* DNA were performed as previously described (Cowie *et al.*, 2006; Sambrook *et al.*, 1989). Unless stated otherwise, all DNA clonings were performed using sequence- and ligation-independent cloning (Jeong *et al.*, 2012). The sequence of oligonucleotides used in this study are listed in Table 4.S2.

4.10.1.2 Construction of plasmids for isolation of the engA-*tRNA*-rmlC *(ETR) region from the chromosome of* S. fredii *NGR234.*

For flanking the ETR region with FRT sites, plasmid pTH1937 (Milunovic *et al.*, 2014) was first modified such that the excised plasmid following Flp-mediated recombination between the two FRT sites would carry a φC31 *attP* site (Thorpe & Smith, 1998); oligos attP1 and attP2 were annealed to construct an *attP* site, which was then ligated into *Eco*RV/*Eco*RI digested pTH1937, forming plasmid pTH2884. A 963 nt PCR product (oligos DF016 and DF017) from the *S. fredii* chromosome (nt. 2,528,036 – 2,528,998) was then introduced into *Spe*I digested pTH2884, constructing plasmid pTH2917. For introduction of the second FRT site, a 940 nt fragment was PCR amplified (oligos DF014 and DF015) from the *S. fredii* chromosome (nt. 2,469,829 – 2,470,768) and cloned into *Xho*I digested pTH1522 (Cowie *et al*., 2006), constructing plasmid pTH2916.

4.10.1.3 Construction of plasmids for introduction of the attB *landing pad in the* S. meliloti *Rm2011 chromosome.*

Two ~ 1 kb fragments were independently PCR amplified (oligos DF018 and DF019, plus DF020 and DF021) from the *S. meliloti* Rm2011 chromosome (nt. 2,601,218 – 2,602,232 and 2,603,745-2,604,747) and simultaneously cloned into *Spe*I digested pTH1937, making pTH2918. Plasmid pTH2918 was digested with *Apa*I and *Apa*LI, which cut between the two *S. meliloti* fragments. Oligos DF022 and DF023 were annealed to form a φC31 *attB* site (Thorpe & Smith, 1998), which was ligated into the *Apa*I/*Apa*LI digested pTH2918, forming pTH2920. A Sm^R/Sp^R cassette was PCR amplified (oligos DF024 and DF025) from pHP45Ω (Prentki & Krisch, 1984), and the φC31 integrase (Thorpe & Smith, 1998) under the control of the P_{tac} promoter was PCR amplified (oligos DF026 and DF027) from pTH1599 (Cheng and Finan, unpublished). The Sm^R/Sp^R cassette and φC31 integrase were simultaneously introduced into *Apa*LI digested pTH2920, resulting in the plasmid pTH2937.

4.10.1.4 Construction of plasmids to remove scar regions following integration of the ETR region in S. meliloti.

Plasmids pTH2929 and pTH2930 were made to remove the scar regions following integration of the *S. fredii engA*-tRNA-*rmlC* region into *S. meliloti*. PCR amplified fragments from the chromosomes of *S. meliloti* (nt. 2,601,218 – 2,602,232; oligos DF028 and DF029) and *S. fredii* (nt. 2,469,824 – 2,470,768; oligos DF030 and DF031) were simultaneously cloned into *Xba*I digested pTH2919 (Tc^R, *sacB*) (diCenzo & Finan, 2015),

forming pTH2929. Similarly, plasmid pTH2930 was made by simultaneously introducing PCR amplified fragments from the chromosomes of *S. fredii* (nt. 2,538,046 – 2,538,998; oligos DF032 and DF033) and *S. meliloti* (nt. 2,603,745 – 2,604,747; oligos DF034 and DF035) into *Xba*I digested pTH2919. In both pTH2929 and pTH2930, the fragments amplified from *S. meliloti* and *S. fredii* are directly adjacent to each other, with no intervening DNA.

4.10.2 Supplementary Tables and Figures

Table 4.S1. Chromosomal polymorphisms in the original *S. meliloti* replicon cured derivative.

						Wild type (Rm2011)		ΔpSymA (SmA818)		ΔpSymB (RmP3009)		ΔpSymAB (RmP2917)	
Start*	Stop*	Poly-morphism Type	NT Change	AA Change	Protein Effect	Coverage	Variant freq. (%)	Coverage	Variant freq. (%)	Coverage	Variant freq. (%)	Coverage	Variant freq. (%)
Polymorphisms common to all strains													
221,094	221,094	SNP (transversion)	A -> T	I -> N	Substitution	101	95	116	97.4	168	97	173	98.3
392,553	392,553	SNP (transversion)	T -> G	-	None	92	97.8	94	97.9	140	99.3	120	100
521,050	521,056	Deletion	-GCCCGAT	-	Frame Shift	97	100	99	96	126	100	138	100
1,134,054†	1,134,169	Highly variable region	N/A	N/A	N/A	0->48	N/A	0->51	N/A	10->74	N/A	0->93	N/A
1,134,134†	1,134,620	No reads	N/A	N/A	N/A	0	N/A	0	N/A	0	N/A	0	N/A
1,134,620†	1,134,727	Highly variable region	N/A	N/A	N/A	0->57	N/A	0->57	N/A	13-67	N/A	0->106	N/A
1,266,445	1,266,445	SNP (transversion)	A -> C	M -> L	Substitution	122	97.5	130	99.2	136	97.8	178	99.4
1,521,691	1,521,692	Likely ISRm19 insertion	N/A	N/A	N/A	-	-	-	-	-	-	-	-
1,716,202	1,716,202	Insertion	+G	N/A	N/A	104	99	119	97.5	110	94.5	122	100
1,716,255	1,716,255	Insertion	+CTGCGGC	N/A	N/A	106	95.3	126 -> 127	89.7	108 -> 112	94	119 -> 121	95.8
1,930,257	1,930,257	SNP (transition)	T -> C	Y -> C	Substitution	77	100	107	99.1	80	96.3	114	100
2,290,057	2,290,057	SNP (transversion)	A -> C	-	None	73	94.5	82	96.3	75	94.7	114	100
3,113,898	3,113,898	SNP (transition)	A -> G	K -> E	Substitution	95	100	103	97.1	154	96.8	163	98.8
3,622,706	3,622,706	SNP (transition)	T -> C	N/A	N/A	95	95.8	67	94	101	93.1	114	93.9
3,622,726	3,622,727	Substitution	AA -> GG	N/A	N/A	96	95.8	71	94.4	113	89.4	110	94.5
3,622,732	3,622,732	Deletion	-G	N/A	N/A	96	100	70	100	108	100	110	98.2
3,622,746	3,622,746	SNP (transition)	A -> G	N/A	N/A	99	98	70	95.7	106	95.3	110	100
Polymorphisms common to Wild type and ΔpSymB													
87,913	87,913	SNP (transition)	A -> G	-	-	127	100	-	-	143	94.4	-	-
1,622,680	1,622,680	SNP (transition)	T -> C	H -> R	Substitution	100	99	-	-	102	100	-	-
1,742,052	1,742,052	SNP (transition)	G -> A	-	Truncation	106	93.4	-	-	112	90.2	-	-

2,934,778	2,934,778	Deletion	-C	N/A	N/A	79	100	-	-	126	100	-	-
2,935,408	2,935,408	Deletion	-A	-	Frame Shift	84	100	-	-	116	100	-	-
3,363,780¥	3,363,873	Highly variable region	N/A	N/A	N/A	0->51	N/A	-	-	17->86	N/A	-	-
3,363,874¥	3,373,292	No reads	N/A	N/A	N/A	0	N/A	-	-	0	N/A	-	-
3,373,293¥	3,373,398	Highly variable region	N/A	N/A	N/A	0->61	N/A	-	-	12->69	N/A	-	-
3,435,131	3,435,131	SNP (transition)	T -> C	I -> V	Substitution	119	95.8	-	-	170	98.2	-	-
3,638,949	3,638,949	SNP (transition)	C -> T	D -> N	Substitution	98	98	-	-	158	98.7	-	-
Polymorphisms common to ΔpSymA and ΔpSymAB													
8,443	8,443	SNP (transition)	G -> A	G -> S	Substitution	-	-	117	99.1	-	-	152	98.7
1,878,249	1,878,249	SNP (transition)	C -> T	C -> Y	Substitution	-	-	107	93.5	-	-	158	86.7
2,935,408	2,935,408	Deletion	-A	-	Frame Shift	-	-	123	100	-	-	117	100
3,604,255	3,604,255	SNP (transversion)	T -> G	N/A	N/A	-	-	139	100	-	-	158	100
3,607,007	3,607,007	SNP (transversion)	C -> G	R -> P	Substitution	-	-	97	92.8	-	-	133	99.2
Unique polymorphisms													
373,159	373,159	SNP (transversion)	G -> C	L -> V	Substitution	-	-	-	-	-	-	120	100
1,621,761	1,621,761	SNP (transition)	C -> T	N/A	N/A	140	93.6	180	80.6	173	95.4	-	-
2,283,833	2,283,833	SNP (transition)	C -> T	G -> S	Substitution	-	-	-	-	-	-	143	96.5
2,285,645	2,285,645	SNP (transversion)	G -> C	I -> M	Substitution	84	100	-	-	-	-	-	-
3,198,777	3,198,777	SNP (transition)	T -> C	N/A	N/A	-	-	114	98.2	179	97.2	27	100
3,233,568	3,233,568	SNP (transition)	T -> C	N/A	N/A	-	-	-	-	-	-	44	81.8
3,239,774	3,239,774	SNP (transition)	G -> A	A -> V	Substitution	105	99	157	96.2	-	-	-	-
3,352,044	3,352,044	Deletion	-C	-	Frame Shift	-	-	-	-	87	100	-	-
3,622,727	3,622,727	SNP (transition)	A -> G	N/A	N/A	-	-	-	-	-	-	110	98.2
3,644,163	3,644,163	SNP (transversion)	T -> G	F -> C	Substitution	-	-	133	97	-	-	-	-

This table lists the nucleotide, and resulting amino acid, variations between the indicated strains and the Rm2011 reference genome. Variations were identified if there were a minimum of 10 reads with at least 80% showing the variant. * Start and stop positions of the polymorphism in reference to the published Rm2011 chromosome sequence (accession: NC_020528.1). † These three polymorphisms were likely due to a single, continuous deletion from nt. 1,134,054->1,134,727. ¥ These three polymorphisms were likely due to a single, continuous deletion from nt. 3,363,780->3,373,293.

Table 4.S2. Chromosomal polymorphisms in *S. meliloti* RmP2917 (ΔpSymAB) derivatives following replicon re-introduction.

						ΔpSymAB (RmP3496)		ΔpSymAB + pSymA (RmP3497)		ΔpSymAB + pSymB (RmP3498)		ΔpSymAB + pSymAB (RmP3499)	
Start*	Stop*	Poly-morphism Type	NT Change	AA Change	Protein Effect	Coverage	Variant freq. (%)	Coverage	Variant freq. (%)	Coverage	Variant freq. (%)	Coverage	Variant freq. (%)
Polymorphisms common to all strains													
8,443	8,443	SNP (transition)	G -> A	G -> S	Substitution	151	98	155	100	87	98.9	132	98.5
221,094	221,094	SNP (transversion)	A -> T	I -> N	Substitution	253	99.6	137	97.8	105	100	172	97.1
373,159	373,159	SNP (transversion)	G -> C	L -> V	Substitution	184	98.4	116	98.3	70	100	129	97.7
392,553	392,553	SNP (transversion)	T -> G	-	None	225	98.2	159	98.7	115	99.1	200	98.5
521,029	521,029	SNP (transition)	C -> T	-	Truncation	193	98.4	128	99.2	87	98.9	126	96.8
1,134,054†	1,134,133	Highly variable region	N/A	N/A	N/A	0->132	N/A	0->87	N/A	0->65	N/A	10->79	N/A
1,134,134†	1,134,640	No reads	N/A	N/A	N/A	0	N/A	0	N/A	0	N/A	0	N/A
1,134,641†	1,134,727	Highly variable region	N/A	N/A	N/A	0->129	N/A	10->95	N/A	0->66	N/A	0->80	N/A
1,266,445	1,266,445	SNP (transversion)	A -> C	M -> L	Substitution	259	98.5	165	99.4	90	100	175	97.1
1,521,691	1,521,692	Likely ISRm19 insertion	N/A	N/A	N/A	-	-	-	-	-	-	-	-
1,716,202	1,716,202	Insertion (tandem repeat)	(G)2 -> (G)3	N/A	N/A	247	98.4	181	100	118	99.2	181	97.2
1,716,255	1,716,255	Insertion	+CTGC GGC	N/A	N/A	233 -> 241	95	149 -> 153	97.3	102 -> 104	97.1	152 -> 155	98.7
1,878,249	1,878,249	SNP (transition)	C -> T	C -> Y	Substitution	179	97.2	175	99.4	107	99.1	132	98.5
1,930,257	1,930,257	SNP (transition)	T -> C	Y -> C	Substitution	202	99.5	180	100	100	100	127	95.3
2,283,833	2,283,833	SNP (transition)	C -> T	G -> S	Substitution	194	99	150	99.3	93	100	147	98.6
2,290,057	2,290,057	SNP (transversion)	A -> C	-	None	133	100	105	99	55	100	92	97.8
2,935,408	2,935,408	Deletion	-A	-	Frame Shift	156	99.4	131	100	90	100	141	99.3
3,113,898	3,113,898	SNP (transition)	A -> G	K -> E	Substitution	237	100	156	100	107	99.1	145	96.6
3,198,777	3,198,777	SNP (transition)	T -> C	N/A	N/A	266	98.1	175	99.4	18	100	119	95.8
3,233,568	3,233,568	SNP (transition)	T -> C	N/A	N/A	98	95.9	62	96.8	52	100	99	94.9
3,435,131	3,435,131	SNP (transition)	T -> C	I -> V	Substitution	279	97.8	131	100	124	100	165	98.8
3,604,255	3,604,255	SNP (transversion)	T -> G	N/A	N/A	215	98.1	155	98.1	92	98.9	138	94.2

3,607,007	3,607,007	SNP (transversion)	C -> G	R -> P	Substitution	143	98.6	100	100	67	100	112	98.2
3,622,706	3,622,706	SNP (transition)	T -> C	N/A	N/A	170	91.8	149	96.6	67	91	133	94.7
3,622,726	3,622,727	Substitution	AA -> GG	N/A	N/A	156 -> 157	95.5	146	97.9	59	93.2	126	94.4
3,622,732	3,622,732	Deletion	-G	N/A	N/A	156	99.4	145	100	59	100	126	100
3,622,746	3,622,746	SNP (transition)	A -> G	N/A	N/A	161	95.7	141	99.3	54	96.3	121	96.7
Unique polymorphisms													
502,020	502,055	No reads	N/A	N/A	N/A	-	-	-	-	0	N/A	-	-
738,563	738,563	SNP (transition)	C -> T	A -> V	Substitution	-	-	-	-	-	-	110	96.4
2,059,066	2,059,095	No reads	N/A	N/A	N/A	-	-	-	-	0	N/A	0->1	N/A
2,236,798	2,236,798	SNP (transition)	G -> A	G -> D	Substitution	30	93.3	17	82.4	-	-	21	81
3,194,149	3,194,149	SNP (transversion)	T -> A	Q -> L	Substitution	-	-	146	100	-	-	-	-

This table lists the nucleotide, and resulting amino acid, variations between the indicated strains and the Rm2011 reference genome. Variations were identified if there were a minimum of 10 reads with at least 80% showing the variant. * Start and stop positions of the polymorphism in reference to the published Rm2011 chromosome sequence (accession: NC_020528.1). † These three polymorphisms were likely due to a single, continuous deletion from nt. 1,134,054->1,134,727.

Table 4.S3. Strains and plasmids used in this study.

Strain / Plasmid	Characteristics	Reference
Sinorhizobium meliloti		
Rm1021	Wild type SU47 *str-21*; Sm^R	(Meade *et al.*, 1982)
Rm2011	Wild type SU47 *str-3*; Sm^R	Lab collection
Rm5000	Wild type SU47 *rif-5*; Rif^R	(Finan *et al.*, 1984)
RmFL1142	RmP110 (pFL1142); Sm^R Gm^R	(Cowie *et al.*, 2006)
RmFL1439	RmP110 (pFL1439); Sm^R Gm^R	This study
RmG994	Rm1021, *dme-3*::Tn*5*, *tme-4*::ΩSp^R; Sm^R Sp^R Nm^R	(Driscoll & Finan, 1996)
RmP110	Rm1021 with wild type *pstC*; Sm^R	(Yuan *et al.*, 2006)
RmP790	RmP110, ΔB108 (pSymB 1,131,168-1,169,073) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP791	RmP110, ΔB107 (pSymB 1,092,289-1,129,758) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP798	RmP110, ΔB122 (pSymB 1,529,711-1,572,422) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP799	RmP110, ΔB109 (pSymB 1,180,466-1,204,770) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP801	RmP110, ΔB116 (pSymB 1,256,503-1,307,752) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP803	RmP110, ΔB124 (pSymB 1,654,191-1,677,882) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP806	RmP110, ΔB123 (pSymB 1,529,711-1,652,558) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP808	RmP110, ΔB106 (pSymB 870,505-1,091,104) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP811	RmP110, ΔB118 (pSymB 1,323,078-1,528,150) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP816	RmP110 (pFL3564 and pTH1942); Sm^R Nm^R Gm^R	(Milunovic *et al.*, 2014)
RmP823	RmP110, ΔB112 (pSymB 1,207,052-1,224,621) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP876	RmP110, ΔB141 (pSymB 101,396-466,499) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP884	RmP110, ΔB148 (pSymB 635,940-762,942) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP936	RmP110, ΔA106 (pSymA 186,200-250,917) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP939	RmP110, ΔA117 (pSymA 402,136-458,916) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP941	RmP110, ΔA118 (pSymA 459,668-505,335) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP943	RmP110, ΔA119 (pSymA 507,338-575,671) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP945	RmP110, ΔA120 (pSymA 577,241-623,673) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP947	RmP110, ΔA121 (pSymA 624,863-677,157) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP949	RmP110, ΔA122 (pSymA 678,150-726,673) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP951	RmP110, ΔA123 (pSymA 727,921-774,293) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP953	RmP110, ΔA124 (pSymA 775,476-828,417) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP955	RmP110, ΔA125 (pSymA 830,143-879,960) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP957	RmP110, ΔA128 (pSymA 927,009-1,063,642) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP959	RmP110, ΔA129 (pSymA 1,064,644-1,122,176) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP961	RmP110, ΔA130 (pSymA 1,123,504-1,173,115) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP963	RmP110, ΔA133 (pSymA 1,284,751-1,348,238) (pTH1944) Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP965	RmP110, ΔA132 (pSymA 1,232,916-1,283,082) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP967	RmP110, ΔA131 (pSymA 1,173,730-1,231,998) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP969	RmP110, ΔA127 (pSymA 881,169-930,000) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP977	RmP110, ΔA112 (pSymA 308,586-347,789) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)

RmP979	RmP110, ΔA109 (pSymA 251,809-311,877) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP991	RmP110, ΔA116 (pSymA 313,654-458,916) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP1052	RmP110 (pFL4393 and pTH2038); Sm^R Gm^R	(Milunovic *et al.*, 2014)
RmP1119	RmP936 cured of pTH1944; Sm^R	This study
RmP1054	RmP110, ΔB154 (pSymB 62,137-100,636) (pTH1944); Sm^R Nm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP1055	RmP110, ΔB161 (pSymB 1,679,723-49,523) (pTH1944); Sm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2650	RmP110, ΔA160 (pSymA 10,988-184,519) (pTH1944); Sm^R Gm^R Tc^R	(Milunovic *et al.*, 2014)
RmP2712	RmP2681, deletion ΔB179 (pSymB 1,207,052-1,323,078); Sm^R Sp^R Nm^R Gm^R	(diCenzo *et al.*, 2013)
RmP2719	RmP110 with a SpR chromosomal insertion of the *engA* and tRNA genes; Sm^R Sp^R	(diCenzo *et al.*, 2014)
RmP2720	RmP2719, ΔB142 (pSymB 122,108-466,499); Sm^R Sp^R	This study
RmP2721	RmP2720, ΔB148 via ΦRmP884; Sm^R Sp^R Nm^R Gm^R	This study
RmP2727	RmP2719, ΔB165 (pSymB 122,108-762,942); Sm^R Sp^R Gm^R	This study
RmP2754	RmP110, ΔB180 (pSymB 635,940-869,642); Sm^R Nm^R	(Milunovic *et al.*, 2014)
RmP2917	Rm2011, ΔpSymA ΔpSymB, engA/tRNAarg in chromosome; Sm^R Sp^R	(diCenzo *et al.*, 2014)
RmP3009	Rm2011, ΔpSymB, *engA*/$tRNA^{arg}$ in chromosome; Sm^R Sp^R	(diCenzo *et al.*, 2014)
RmP3277	RmP110, FRT sites flanking *phoR pstSCABphoUphoB*; Sm^R Gm^R Nm^R	Lab collection
RmP3331	Rm2011, pTH2937 double recombinant; Sm^R Sp^R	This study
RmP3340	RmP3331, pTH2938 integrant via *attB/attP*; Sm^R Sp^R Nm^R Gm^R	This study
RmP3362	RmP3340, pTH2929 double recombinant; Sm^R Sp^R	This study
RmP3380	$Rm2011_{\Omega NGR69}$ (RmP3380, pTH2930 double recombinant); Sm^R	This study
RmP3382	SmA818 with NGR234 *engA*-tRNA-*rmlC* region via ΦRmP3362; Sm^R Sp^R	This study
RmP3409	$\Delta pSymA_{\Omega NGR69}$ (RmP3382, pTH2930 double recombinant); Sm^R	This study
RmP3410	RmP3380, ΔB180 via ΦRmP2745; Sm^R Nm^R	This study
RmP3411	RmP3409, ΔB180 via ΦRmP2745; Sm^R Nm^R	This study
RmP3412	RmP3380, ΔB179 via ΦRmP2712; Sm^R Nm^R Gm^R	This study
RmP3413	$\Delta pSymB_{\Omega NGR69}$ (RmP3410, pTH2930 double recombinant); Sm^R	This study
RmP3414	$\Delta pSymAB_{\Omega NGR69}$ (RmP3411, pTH2930 double recombinant); Sm^R	This study
RmP3491	Rm5000 (pTE3Yp028); Rif^R Tc^R	This study
RmP3496	ΔpSymAB (RmP2917) used to make RmP3497-RmP3499; Sm^R Sp^R	(diCenzo *et al.*, 2014)
RmP3497	ΔpSymAB with pSymA from RmP3491; Sm^R Sp^R	This study
RmP3498	ΔpSymAB with pSymB from RmP3491; Sm^R Sp^R	This study
RmP3499	ΔpSymAB with pSymA and pSymB from RmP3491; Sm^R Sp^R	This study
RmP3500	$\Delta pSymAB_{\Omega NGR69}$ (RmP3380) used to make RmP3501-RmP3503; Sm^R	This study
RmP3501	$\Delta pSymAB_{\Omega NGR69}$ with pSymA from RmP3491; Sm^R	This study
RmP3502	$\Delta pSymAB_{\Omega NGR69}$ with pSymB from RmP3491; Sm^R	This study
RmP3503	$\Delta pSymAB_{\Omega NGR69}$ with pSymA and pSymB from RmP3491; Sm^R	This study
RmP3541	RmP943 cured of pTH1944; Sm^R	This study
RmP3542	RmP3542, pFL1439 via ΦRmFL1439; Sm^R Gm^R	This study
RmP3543	RmP110, ΔA301 (pSymA 507,388-623,673); Sm^R Gm^R	This study
RmP3550	RmP1119, pTH2038 via ΦRmP1052; Sm^R Nm^R	This study
RmP3551	RmP110, ΔA302 (pSymA 47,717-184,519); Sm^R Nm^R	This study
RmP3552	RmP3551, pFL1142 via ΦRmFL1142; Sm^R Nm^R Gm^R	This study
RmP3553	RmP110, ΔA303 (pSymA 47,717-402,136); Sm^R Nm^R Gm^R	This study
RmP3557	RmP2720, ΔB180 via ΦRmP2745; Sm^R Sp^R Nm^R	This study
RmP3358	RmP2719, ΔB201 (pSymB 122,108-869,642); Sm^R Sp^R	This study
SmA818	Rm2011, ΔpSymA; Sm^R	(Oresnik *et al.*, 2000)
Sinorhizobium fredii		
NGR234	Wild type NGR234R *rif-1*; Rif^R	(Stanley *et al.*, 1988)
P3278	NGR234 (pTH2916); Rif^R Gm^R	This study
P3285	P3278 (pTH2917); Rif^R Gm^R Km^R	This study
P3330	P3285 (pTH2505); Rif^R Gm^R Km^R Tc^R	This study

Escherichia coli		
MT616	MM294A *recA-56* (pRK600), mobilizer; Cm^R	(Finan *et al.*, 1986)
MT620	MM294A *recA-56*; Rif^R	Lab collection
Plasmids		
pFL1142	pTH1522 (pSymB 400,267-402,136); Gm^R	(Cowie *et al.*, 2006)
pFL1439	pTH1522 (pSymA 623,673-624,863); Gm^R	(Cowie *et al.*, 2006)
pFL3564	pTH1522 (pSymB 1,408,135-1,408,806); Gm^R	(Cowie *et al.*, 2006)
pFL4393	pTH1522 (pSymA 1,348,238-1,349,931); Gm^R	(Cowie *et al.*, 2006)
pTE3Yp028	pTE3 with P_{trp}::*rctB* of *Rhizobium etli*; Tc^R	(Pérez-Mendoza *et al.*, 2004)
pTH1522	Reporter vector containing a FRT site, ColE1 *oriV*; Gm^R	(Cowie *et al.*, 2006)
pTH1931	Expression vector pTrcSC, derived from pTrcStrep; Sm^R Sp^R	(diCenzo *et al.*, 2013)
pTH1937	ΔTn*903* inverted repeats, pRK2 *oriT*, *nptII* from Tn*5*, p15A *oriV*; Km^R	(Milunovic *et al.*, 2014)
pTH1942	pTH1937 (pSymB 1,528,150-1,529,711); Km^R	(Milunovic *et al.*, 2014)
pTH1944	*flp* gene in a pBBR-MCS3 derivative with RK2-*tetR*-*tetA*; Tc^R	(Milunovic *et al.*, 2014)
pTH2038	pTH1937 (pSymA 4,7717-48,842); Km^R	(Milunovic *et al.*, 2014)
pTH2505	*flp* gene controlled by protocatechuate inducible promoter in pRK7813; Tc^R	(Zhang *et al.*, 2012)
pTH2884	pTH1937 with *attP* inserted via *Eco*RV and *Eco*RI; Km^R	This study
pTH2913	pTH1522/pTH2884 with smc 564,475-575,625; Gm^R Km^R	This study
pTH2916	pTH1522 (NGR234 chr 2,469,829-2,470,768 via *Xho*I); Gm^R	This study
pTH2917	pTH2884 (NGR234 chr 2,538,036-2,528,998 via *Spe*I); Km^R	This study
pTH2918	pTH1937 (smc2,601,218-2,602,232 + 2,603,745-2,604,747 via SpeI); Km^R	This study
pTH2919	Tc^R *sacB* suicide vector, derived from pJQ200mp18; Tc^R	(diCenzo & Finan, 2015)
pTH2920	pTH2918 (*attB* site introduced via *Apa*I and *Apa*LI; Km^R	This study
pTH2929	pTH2919 (smc 2,601,218-2,602,232 + NGRc 2,469,824-2,470,768 via *Xba*I); Tc^R	This study
pTH2930	pTH2919 (smc 2,603,745-2,604,747 + NGRc 2,538,046-2,538,998 via *Xba*I); Tc^R	This study
pTH2937	pTH2920 (Sp^R/Sm^R omega cassette and P_{tac}::ϕC31 integrase via *Apa*L1); Km^R	This study
pTH2938	pTH1522/pTH2884 recombinant with NGRc 2,469,829-2,538,998; Gm^R Km^R	This study

Sm – streptomycin, Nm – neomycin, Gm – gentamicin, Km – kanamycin, Sp – spectinomycin, Rif – rifampicin, Tc – tetracycline, Cm – chloramphenicol

Table 4.S4. Oligonucleotides used in this study.

Name	Sequence
attP1	5'- /PHOS/ **ATC** CCC CAA CTG GGG TAA CCT TTG AGT TCT CTC AGT TGG GG**G**
attP2	5'- /PHOS/ **AAT TC**C CCC AAC TGA GAG AAC TCA AAG GTT ACC CCA GTT GGG G**GA T**
DF014	5'- ATC ACC GGA TCT CGA AGA GGG GAT TTC GGC GGT GC
DF015	5'- CAA TCA ATC ACT CGA AGC CTA TGA CAA GAC CAA GCC G
DF016	5'- CTG CAG ATC TAC TAG CCC GTG TAC AAG ACG CAG CC
DF017	5'- TCA TTT AAA TAC TAG CTA CAC TCT ACT GCC CCG TGG
DF018	5'- TCG ACC TGC AGA TCT ACT AGA AAG GGA GGA CTG AAT GAC G
DF019	5'- **GTG CAC** GCA TGC AT**G GGC CC**C CTT ACT TCG GGA CAA TCT GCC
DF020	5'- **GGG CCC** ATG CAT GC**G TGC AC**C GGC TAC CCA GAA TCT GTC G
DF021	5'- GGA TAT CAT TTA AAT ACT AGT GGT GAT GTA TGC GGG CAG G
DF022	5'- /PHOS/ **C**GG GTG CCA GGG CGT GCC CTT GGG CTC CCC GGG CGC GTA **CG**
DF023	5'- /PHOS/ **TGC ACG** TAC GCG CCC GGG GAG CCC AAG GGC GAC GCC TGG CAC CC**G GGC C**
DF024	5'- TCC CCG GGC GCG TAC GTG CAG CAT GTG TCA GAG GTT TTC ACC
DF025	5'- TAT GAT GTC GCG TAT CAC GAG GCC CTT TCG
DF026	5'- TCG TGA TAC GCG ACA TCA TAA CGG TTC TGG
DF027	5'- GAT TAT GGG TAG CCG GTG CAC CCG GGT GTC TCG CTA
DF028	5'- GAG CTC GGT ACC CGG GGA TCA AAG GGA GGA CTG AAT GAC C
DF029	5'- AAT CCC CTC TCC TTA CTT CGG GAC AAT CTG CC
DF030	5'- CGA AGT AAG GAG AGG GGA TTT CGG CGG TGC
DF031	5'- CAG GTC GAC TCT AGA GGA TGA GCC TAT GAC AAG ACC AAG CCG
DF032	5'- GAG CTC GGT ACC CGG GGA TCC CCG TGT AGA AGA CGC AGC C
DF033	5'- TGG GTA GCC GCT ACA CTC TAC TGC CCC GTG G
DF034	5'- TAG AGT GTA GCG GCT ACC CAG AAT CTG TCG
DF035	5'- CAG GTC GAC TCT AGA GGA TCT GGT GAT GTA TGC GGG CAG G

Nucleotides that anneal to the template for PCR or sequencing are underlined, while restriction sites of interest are in boldface font.

Table 4.S5. Nucleotide sequences accessed in this study.

Accession code	Description	Reference
NC_002078.1	*S. meliloti* Rm1021 pSymB	(Galibert *et al.*, 2001)
NC_018701.1	S. meliloti Rm41 pSYMB	(Weidner *et al.*, 2013)
NC_017326.1	*S. meliloti* SM11 pSmeSM11d	(Schneiker-Bekel *et al.*, 2011)
NC_015596.1	*S. meliloti* AK83 chromosome 2	(Galardini *et al.*, 2011)
NC_017323.1	*S. meliloti* BL225C pSINMEB02	(Galardini *et al.*, 2011)
NC_019849.1	*S. meliloti* GR4 pRmeGR4d	(Martínez-Abarca *et al.*, 2013)
NZ_CP009146.1	*S. meliloti* RMO17 pSymB	(Toro *et al.*, 2014)
NC_020528.1	*S. meliloti* Rm2011 chromosome	(Sallet *et al.*, 2013)
NC_020527.1	*S. meliloti* Rm2011 pSymA	(Sallet *et al.*, 2013)
NC_020560.1	*S. meliloti* Rm2011 pSymB	(Sallet *et al.*, 2013)
ATYB01000008.1	*S. arboris* LMG 14919 scaffold2.3_C	(Reeve *et al.*, 2006)
NC_009620.1	*S. medicae* WSM419 pSMED01	(Reeve *et al.*, 2010)
JFGO01000048.1	*S. americanum* CCGM7 scaffold48	(Mora *et al.*, 2014)
NC_018000.1	*S. fredii*; USDA 257 chromosome	(Schuldes *et al.*, 2012)
NC_016812.1	*S. fredii* HH103 chromosome	(Weidner *et al.*, 2012)
NC_012587.1	*S. fredii* NGR234 chromosome	(Schmeisser *et al.*, 2009)
AZUW01000006.1	*S. terangae* WSM1721 scaffold_2.3_C	None
NZ_CP007236.1	*E. adhaerens* OV14 chromosome 1	(Rudder *et al.*, 2014)
NC_007761.1	*R. etli* CFN42 chromosome	(González *et al.*, 2006)

Figure 4.S1. Schematic illustration of the re-introduction of the ETR region into the *S. meliloti* chromosome.

(**A**) The *engA*-tRNA-*rmlC* (ETR) region from *S. fredii* NGR234 was isolated as a narrow host range plasmid through a two-step Flp/FRT mediated *in vivo* cloning procedure. (1) Two FRT sites were sequentially introduced at the borders of the ETR region (shown in grey) through single cross-over plasmid integration (yellow and light blue plasmids with the FRT sites represented by the arrow heads). (2) The plasmid pTH2505 (maroon), carrying the *flp* gene was transferred to this *S. fredii* strain. (3) The ETR region was recombined out of the *S. fredii* NGR234 chromosome, and captured in *E. coli* as a Gm^R Km^R ColE1/p15A plasmid with RK2 *oriT* sites and a *attP* site. (**B**) A *S. meliloti* Rm2011 recipient strain was prepared. (4,5) Plasmid double cross-over recombination was used to replace the ISRm5 insertion element (shown in magenta) between *kdgK* and *dppF2* in the *S. meliloti* Rm2011 chromosome with a Sm^R/Sp^R cassette, the ϕC31 integrase, and an *attB* site (all of which is represented by the green). (**C**) The ETR region from *S. fredii* NGR234 was introduced into the *S. meliloti* Rm2011 genome. (6) The plasmid containing the *S. fredii* ETR region was transferred to the *S. meliloti* recipient strain, where it integrated into the chromosomal *attB* site via the plasmid *attP* site. (7) The scar regions at the ends of the ETR integration were removed through double cross-over plasmid (pink) recombination, (8) resulting in a borderless integration of the ETR region into the *S. meliloti* chromosome.

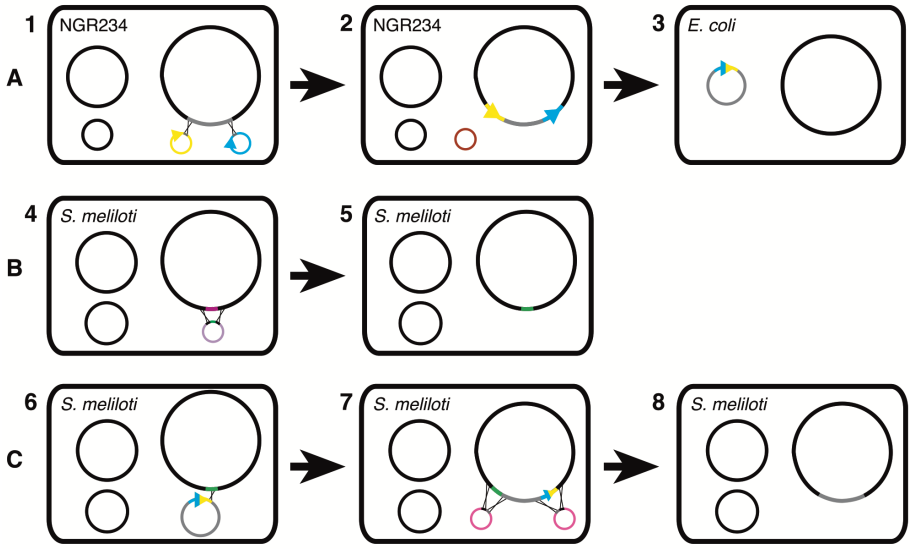

1
NGR234
2
NGR234
3
E. coli
A
4
S. meliloti
5
S. meliloti
B
6
S. meliloti
7
S. meliloti
8
S. meliloti
C

Figure 4.S2. Detailed illustrations of important constructs described in this study.

(**A**) The FRT flanked *engA*-tRNA-*rmlC* region (NGR69) in the *S. fredii* NGR234 chromosome. The figure corresponds to the chromosome shown in Figure S3-A2. (**B**) The plasmid (pTH2938 or pETR) excised from the *S. fredii* chromosome following *flp* mediated recombination between the FRT sites. The figure corresponds to the plasmid in Figure S3-A3. (**C**) The *S. meliloti* chromosome following integration of the pETR plasmid via the *attB* and *attP* sites. The double cross over plasmid recombinations to remove the scar regions are also shown. The figure corresponds to the chromosome of Figure S3-A7. *int* – ϕC31 integrase.

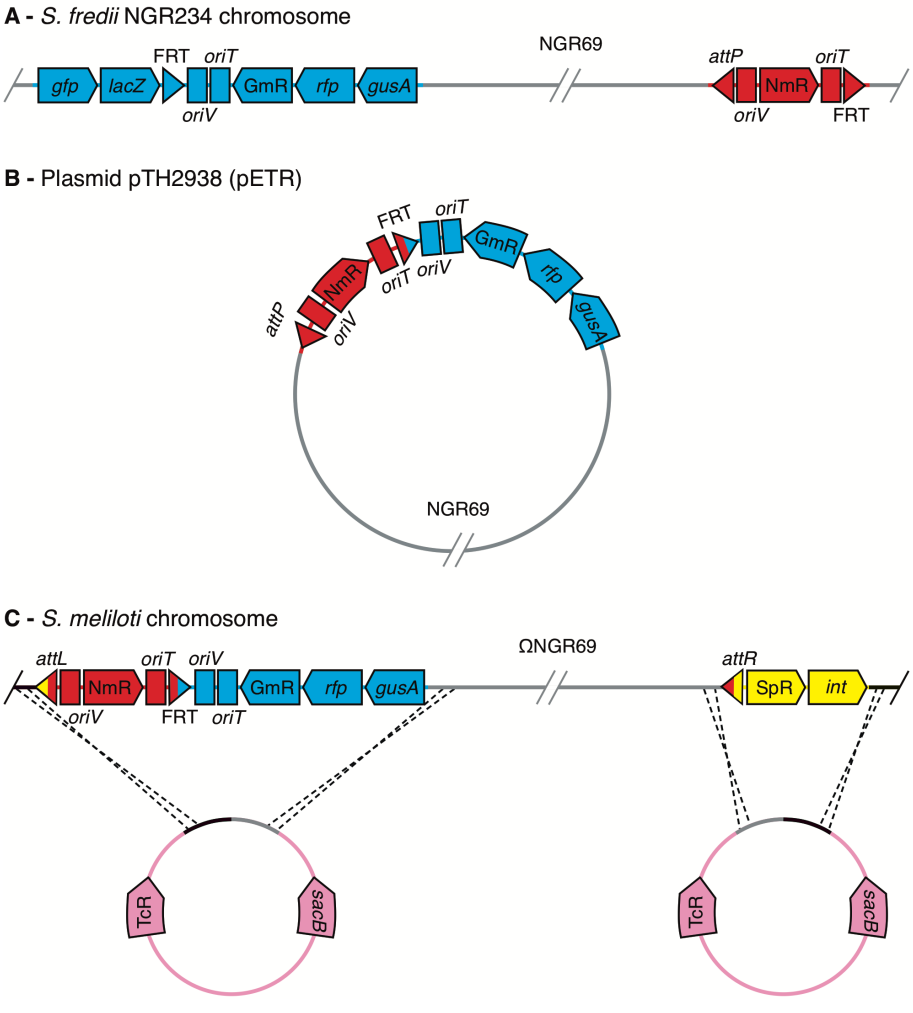

A - S. fredii NGR234 chromosome
FRT
oriT
gfp
lacZ
GmR
rfp
gusA
oriV
NGR69
attP
oriT
NmR
oriV
FRT
B - Plasmid pTH2938 (pETR)
FRT
oriT
GmR
rfp
gusA
NmR
attP
oriV
oriT
oriV
NGR69
C - S. meliloti chromosome
attL
NmR
oriV
oriT
oriV
GmR
rfp
gusA
FRT
oriT
ΩNGR69
attR
SpR
int
TcR
sacB
TcR
sacB

Figure 4.S3. Phylogenetic analysis of the *Sinorhizobium/Ensifer* group.

Phylogeny of the indicated species based on (**A**,**B**) the *engA* nucleotide sequence, (**C**,**D**) the EngA amino acid sequence, and (**E**,**F**) the rRNA nucleotide sequence using the (**A**,**C**,**E**) minimum evolution or (**B**,**D**,**F**) maximum likelihood algorithm. Numbers indicate bootstrap value from 1000 replicates, and all trees were rooted with *R. etli* CFN42. Rm41, SM11, AK83, BL225C, Rm1021, RMO17, and GR4 are *S. meliloti* strains; LMG 14919 is a *S. arboris* strain; WSM419 is a *S. medicae* strain; CCGM7 is a *S. americanum* strain; USDA 257, HH103, and NGR234 are *S. fredii* strain; WSM1721 is a *S. terangae* strain; OV14 is an *E. adhaerens* strain.

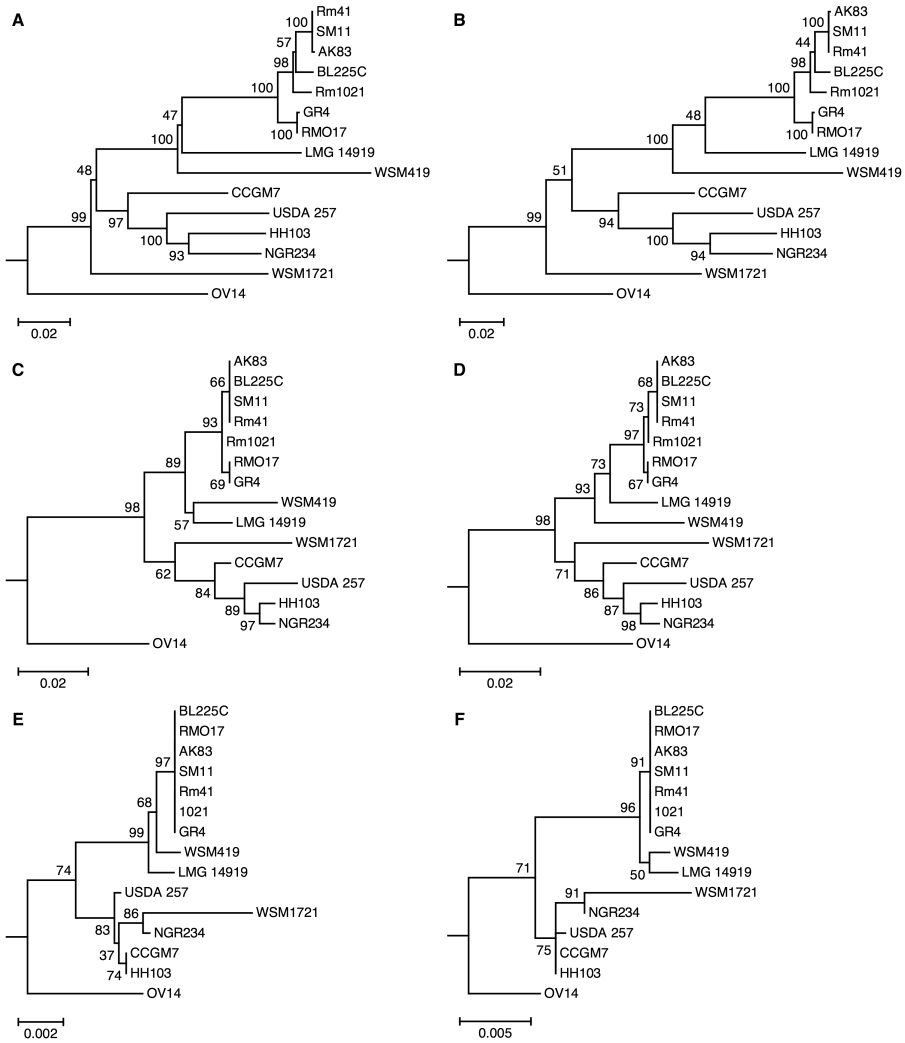
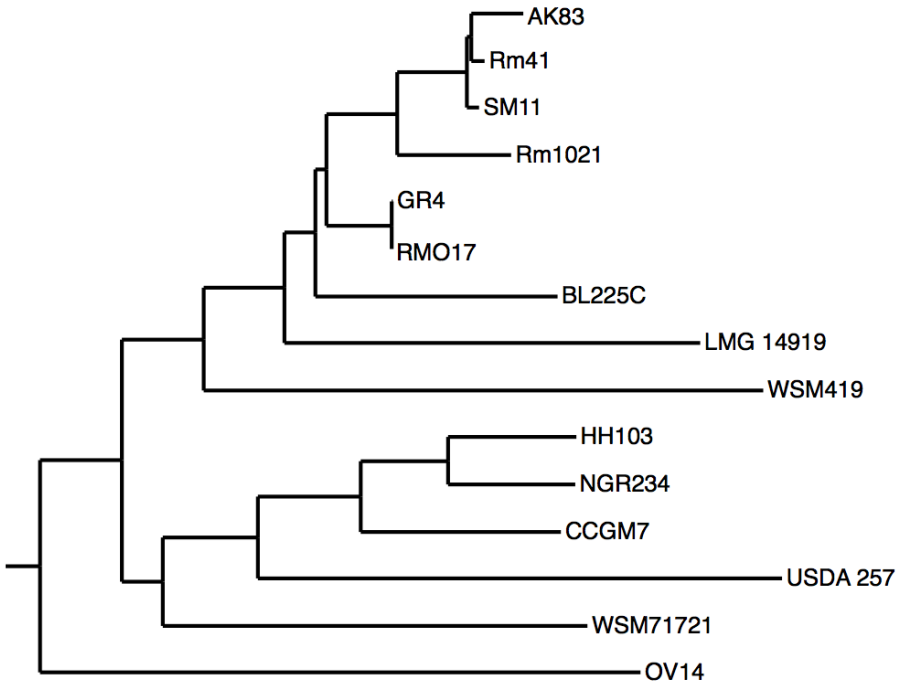

A
Rm41
SM11
AK83
BL225C
Rm1021
GR4
RMO17
LMG 14919
WSM419
CCGM7
USDA 257
HH103
NGR234
WSM1721
OV14
0.02
B
AK83
SM11
Rm41
BL225C
Rm1021
GR4
RMO17
LMG 14919
WSM419
CCGM7
USDA 257
HH103
NGR234
WSM1721
OV14
0.02
C
AK83
BL225C
SM11
Rm41
Rm1021
RMO17
GR4
WSM419
LMG 14919
WSM1721
CCGM7
USDA 257
HH103
NGR234
OV14
0.02
D
AK83
BL225C
SM11
Rm41
Rm1021
RMO17
GR4
LMG 14919
WSM419
WSM1721
CCGM7
USDA 257
HH103
NGR234
OV14
0.02
E
BL225C
RMO17
AK83
SM11
Rm41
1021
GR4
WSM419
LMG 14919
USDA 257
WSM1721
NGR234
CCGM7
HH103
OV14
0.002
F
BL225C
RMO17
AK83
SM11
Rm41
1021
GR4
WSM419
LMG 14919
WSM1721
NGR234
USDA 257
CCGM7
HH103
OV14
0.005

Figure 4.S4. Phylogenomic analysis of the *Sinorhizobium/Ensifer* species.

The progressiveMAUVE guide tree for the alignment of the entire *engA*-tRNA-*rmlC* region from the indicated species is shown. The tree was rooted with the corresponding region from *R. etli* CFN42. Rm41, SM11, AK83, BL225C, Rm1021, RMO17, and GR4 are *S. meliloti* strains; LMG 14919 is a *S. arboris* strain; WSM419 is a *S. medicae* strain; CCGM7 is a *S. americanum* strain; USDA 257, HH103, and NGR234 are *S. fredii* strain; WSM1721 is a *S. terangae* strain; OV14 is an *E. adhaerens* strain.


AK83
Rm41
SM11
Rm1021
GR4
RMO17
BL225C
LMG 14919
WSM419
HH103
NGR234
CCGM7
USDA 257
WSM71721
OV14

CHAPTER 5. GENETIC REDUNDANCY IS PREVALENT WITHIN THE 6.7 MB *SINORHIZOBIUM MELILOTI* GENOME

Citation: diCenzo GC Finan TM. 2015. Genetic redundancy is prevalent within the 6.7 Mb *Sinorhizobium meliloti* genome. Mol Genet Genomics. 290(4): 1345-1356.

5.1 Preface

This chapter details an analysis of functional redundancy between the *S. meliloti* chromosome with the pSymA megaplasmid and pSymB chromid. In short, transposon insertion mutations within the *S. meliloti* strain lacking pSymA and pSymB (ΔpSymAB) were identified that resulted in a no growth phenotype on a minimal M9-sucrose medium and that did not result in this phenotype in the wild type *S. meliloti*. Of all insertions resulting in a no growth phenotype in ΔpSymAB, greater than 10% did not have this phenotype in the wild type. Those 10% of transposon insertions were determined to be located within chromosomal genes that encode a gene product with functional overlap with a gene product encoded by a pSymA or pSymB gene. This result is the first experimental evidence for high levels of functional redundancy between a chromosome with one or more secondary replicons, which has interesting implications for multipartite genome evolution.

This particular study focused on genes essential for growth on a minimal medium and identified numerous biosynthetic genes. It is likely that essential cellular functions are also redundant between the chromosome and secondary replicons, but limitations in the experimental design precluded the identification of such genes. Nevertheless, it appears that numerous genes required for growth in a soil environment are redundant between the chromosome and secondary replicons. Hence, this redundancy provides a potential mechanism through which secondary replicons can gain core cellular functions and gain chromid status if the protein produced by the gene on the secondary replicon is

capable of fully replacing the function of the chromosomally encoded protein. In such a case, mutations inactivating the chromosomal gene would be fitness neutral as it would be complemented by the second copy. If this were to happen, the gene on the secondary replicon would become the sole gene encoding a functional protein and would therefore become essential, in essence transferring a core gene from the chromosome to a secondary replicon.

The high level of redundancy between the chromosome and secondary replicons also highlights how secondary replicons can influence the evolution of a co-evolving chromosome. Genome streamlining and genetic drift in prokaryotic organisms means that, in general, genes encoding products with identical functions are unlikely to remain in the genome. Instead, one copy is either likely to be lost or gain novel functionality, which could include novel substrate specificity or different regulatory patterns. Indeed, data presented in this chapter suggest that at least some of the functionally redundant genes identified in this study have different expression patterns or only partially overlapping substrate specificity. Thus, the presence of a secondary replicon may influence the evolution of chromosomal genes and result in the emergence of new functions.

On the other hand, significant redundancy may limit the horizontal transfer of secondary replicons. For genes that require highly regulated expression, the acquisition of a megaplasmid containing a second copy of these genes may have a fitness cost and result in the elimination of the megaplasmid from the population. Similarly, if the copy

on the megaplasmid/chromid is different enough from the chromosomal copy that they interfere with each other's function, the loss of the replicon may be favoured. In these ways, redundancy may limit the host range of a megaplasmid or chromid.

The previous few paragraphs provides scenarios in which the presence of significant functional redundancy between a chromosome and a secondary replicon can influence the evolution of the genome. Thus, by providing experimental evidence that such redundancy does exists, this chapter sheds novel light on a different aspect of the evolution of multipartite genomes.

5.2 Abstract

Biological pathways are frequently identified via a genetic loss-of-function approach. While this approach has proven to be powerful, it is imperfect as illustrated by well-studied pathways continuing to have missing steps. One potential limiting factor is the masking of phenotypes through genetic redundancy. The prevalence of genetic redundancy in bacterial species has received little attention, although isolated examples of functionally redundant gene pairs exist. Here, we made use of a strain of *Sinorhizobium meliloti* whose genome was reduced by 45% through the complete removal of a megaplasmid and a chromid (3 Mb of the 6.7 Mb genome was removed) to begin quantifying the level of genetic redundancy within a large bacterial genome. Mutagenesis of the strain with the reduced genome identified a set of transposon insertions precluding growth of this strain on minimal medium. Transfer of these mutations to the wild-type background revealed that 10 - 15% of these chromosomal mutations were located within

genes with functional overlap, as they did not prevent growth of cells with the full genome. The functionally redundant genes were involved in a variety of metabolic pathways, including central carbon metabolism, transport, and amino acid biosynthesis. These results indicate that genetic redundancy may be prevalent within large bacterial genomes. Failing to account for redundantly encoded functions in loss-of-function studies will impair our understanding of a broad range of biological processes and limit our ability to use synthetic biology in the construction of designer cell factories.

5.3 Introduction

Despite significant advances in the annotation of prokaryotic genomes, a surprisingly large percentage of genes appear to be uncharacterized. As of October 17, 2014, a search of the Entrez Gene database (Maglott *et al.*, 2011) with the query 'hypothetical' returns ~ 32% of all prokaryotic genes (~ 2,900,000 of ~ 8,900,000 entries), of which ~ 1,750,000 are returned with a query of 'conserved hypothetical'. Whereas some hypothetical genes may represent artefacts of automated genome annotation, many of the conserved hypothetical genes are likely to encode true proteins of unknown function at the time of annotation (Kolker *et al.*, 2004). This uncertainty in gene function is reflective of how genome sequencing and annotation have outpaced the functional characterization of open reading frames and highlights that much is still to be discovered about the biology of prokaryotic genes.

One commonly encountered difficulty in the functional annotation of genes is an absence of an observable phenotype following their disruption. Given that natural

selection and genetic drift lead to genome streamlining (Kuo & Ochman, 2009; Kuo *et al.*, 2009; Lynch, 2006), and pseudogenes are rapidly lost from the genome (Lerat & Ochman, 2005), it is unlikely that all of the uncharacterized genes truly lack a function. While many factors could contribute to a gene's apparent lack of function (e.g., the gene is not expressed in the tested environment), one potential source is genetic redundancy within the genome. Genetic redundancy refers to the phenomenon whereby two genes or pathways are able to functionally complement the loss of the other (Zhang, 2012). Genetic redundancy at both the pathway and single gene level is prevalent within eukaryotic genomes and has been extensively studied, particularly in yeast [see for example, *Saccharomyces cerevisiae* (Costanzo *et al.*, 2010; Gu *et al.*, 2003; Li *et al.*, 2010; Tong *et al.*, 2001), *Caenorhabditis elegans* (Tischler *et al.*, 2006), and *Homo sapiens* (Hsiao & Vitkup, 2008)]. Such redundancy can mask the fitness cost associated with the loss of a particular gene (Gu *et al.*, 2003) and may make it difficult to identify the genetic determinants involved in a particular pathway through loss-of-function studies (Cutler & McCourt, 2005).

Considering the streamlined nature of prokaryotic genomes, it may seem counter-intuitive to expect functionally redundant genes to be prevalent within a bacterial genome. Yet, there have been reports of functional redundancy within a broad range of microbial species and functions [for example, see (Belitsky *et al.*, 2001; Cheng *et al.*, 2007; Elliot *et al.*, 2003; Rabin & Stewart, 1992; Xiao & Wall, 2014)]. While these reports suggest that genetic redundancy through enzyme functional overlap is present

within prokaryotic organisms, non-targeted large-scale studies are necessary to measure the extent of its prevalence. To date, large-scale studies have predominately focused on redundancy of metabolic pathways (Nakahigashi *et al.*, 2009), predominately through *in silico* analyses (Ghim *et al.*, 2005; Suthers *et al.*, 2009; Wang & Zhang, 2009). Recently, a synthetic genetic array was developed for *E. coli* (Butland *et al.*, 2008), and genetic interaction maps have begun to be analyzed (Babu *et al.*, 2014). Despite providing significant insight into the organism, neither of these approaches provides a direct examination of the presence of functionally redundant genes. In perhaps the largest examination of prokaryotic functionally redundant genes, Thomaides *et al.* experimentally examined 120 pairs of potentially redundant *Bacillus subtilis* genes identified through a bioinformatics analysis (Thomaides *et al.*, 2007). These researchers found six pairs of redundant essential genes and a pair of redundant biosynthetic genes. However, the targeted nature of this study precludes reaching conclusions about the genome prevalence of genetic redundancy.

Sinorhizobium meliloti is a soil-dwelling, Gram-negative α-proteobacterium that enters into N_2-fixing endosymbiosis with several legumes belonging to the genera *Medicago*, *Melilotus*, and *Trigonella*. The genome of *S. meliloti* is multipartite, with all natural isolates containing at least an ~ 3.7 Mb chromosome, an ~ 1.4 Mb megaplasmid, and an ~ 1.7 Mb chromid (Epstein *et al.*, 2012; Galibert *et al.*, 2001; Guo *et al.*, 2009). The identification of only two essential genes located outside of the chromosome (diCenzo *et al.*, 2013; Milunovic *et al.*, 2014) and their subsequent integration into the

chromosome (diCenzo *et al.*, 2013) facilitated the recent construction of an *S. meliloti* strain lacking 45% of its genome through the removal of both pSymA and pSymB (diCenzo *et al.*, 2014). The presence of this significantly reduced genome, accomplished through the removal of two out of three unlinked replicons, provided an opportunity to examine the level of genetic redundancy within the *S. meliloti* genome using a large-scale, non-targeted approach.

Here, we report the identification of chromosomal *S. meliloti* insertion mutations producing a phenotype that differed dependent on the presence/absence of pSymA and pSymB (~45% of the genome). Greater than 10% of all mutations that abolished growth on minimal medium were located within redundant genes. Redundancy was found in pathways for central carbon metabolism, transport, and amino acid biosynthesis. Characterization of these redundant gene pairs provided the first experimental confirmation of the function of several genes, adding to our understanding of *S. meliloti* metabolism.

5.4 Results

5.4.1 Isolation of transposon insertions within redundant loci

To uncover redundancy between functions encoded on the pSymA or pSymB replicons and the chromosome, we sought to identify mutations whose phenotype was dependent on genome content. The experimental workflow is summarized in Figure 5.1. Approximately 9,300 mutants from multiple independent Tn*5*-B20 mutageneses of *S. meliloti* ΔpSymAB (ΔRmP2917) were screened for growth on M9-sucrose and M9-

cellobiose, and 73 (~ 0.8%) failed to grow on both media. Recombination via transduction of all 73 Tn*5*-B20 insertions back into *S. meliloti* ΔpSymAB showed linkage of the insertion with the inability to grow on M9-sucrose. At the same time, all 73 Tn*5*-B20 insertion mutations were recombined into the wild-type *S. meliloti* Rm2011 and screened for growth on M9-sucrose. Recombinants from 12 of the 73 crosses with *S. meliloti* Rm2011 grew on M9-sucrose. Recombination of these 12 Nm^R insertions back into *S. meliloti* Rm2011 and ΔpSymAB confirmed that these 12 insertions prevented growth on M9-sucrose only in the absence of pSymA and pSymB (Figure 5.2). The 12 transposon insertions were found to be located within seven unique genes (*edd*, *proC*, *argH1*, *aglE*, *glgB1*, *pgk*, *argD*) via DNA sequencing (Figure 5.3). Below, we summarize our characterization of the putatively redundant loci, and in many cases the identification of the redundant pSymA/pSymB loci. For the latter, we utilized a cosmid library carrying DNA from wild-type *S. meliloti* (Friedman *et al.*, 1982) as well as a library of strains in which defined regions from pSymA or pSymB were deleted (Milunovic *et al.*, 2014).

*5.4.2 6-phosphogluconate dehydratase (*edd*)*

Cosmids carrying a 21 kb region from nucleotide 125,586 to 147,721 of pSymA (note, all nucleotide positions for cosmid inserts and deletions are given relative to the Rm1021 reference genome) were found to complement the *edd-1*::Tn*5*-B20 allele. Confirming that the locus redundant with *edd* was located within this region, the introduction of the pSymA deletion ΔA105 (nt: 125,128–184,519) into RmP3099 (Rm2011, *edd-1*::Tn*5*-B20) resulted in extremely poor growth on M9-sucrose (data not

shown). The most likely candidate gene to be redundant with *edd* within the 21 kb region of the complementing cosmids was *sma0235* [51.3 % nt identity to *edd*, 51.5 % aa (amino acid) similarity], annotated as a dihydroxyacid dehydratase and 6-phosphogluconate dehydratase by InterPro (Hunter *et al.*, 2012). Indeed, deletion of *sma0235* in RmP3099 resulted in little growth on M9-sucrose, while growth was indistinguishable from wild type on M9-succinate (Figure 5.4A,B).

*5.4.3 Pyrroline-5-carboxylate reductase (*proC*)*

Complementation of *S. meliloti* RmP3104 (ΔpSymAB, *proC-1*::Tn*5*-B20) led to the isolation of clones carrying a 23 kb (nt: 1,669,658–9,262) region from pSymB. Introduction of the pSymB deletion ΔB181 (nt: 1,679,723–49,523) into RmP3103 (Rm2011, *proC-1*::Tn*5*-B20) resulted in no growth on M9-sucrose (data not shown), localizing the complementing locus to a 13 kb region. The *smb20003* gene (49.7 % nt identity to *proC*, 46.5 % aa similarity) within this region is annotated as a pyrroline-5-carboxylate reductase, as is *proC*. Deletion of *smb20003* in RmP3103 precluded growth on M9-sucrose unless supplemented with L-proline, L-ornithine, or L-arginine (Figure 5.4C,D).

*5.4.4 Argininosuccinate lyase (*argH1*)*

There are two annotated copies of *argH* in the *S. meliloti* genome; *argH1* is on the chromosome and *argH2* (53.7 % nt identity to *argH1*, 68.5 % aa similarity) is on pSymB. Both were isolated on separate cosmids following complementation of *S. meliloti* RmP3110 (ΔpSymAB, *argH1-1*::Tn*5*-B20). No growth in M9-sucrose was observed

when *argH2* was removed from RmP3109 (Rm2011, *argH1-1*::Tn*5*-B20) through the introduction of the pSymB deletion ΔB145 (nt: 635,940–744,320) (Figure 5.4E). Growth on M9-sucrose was recovered following L-arginine supplementation (Figure 5.4E). Furthermore, expression of the *argH2* gene *in trans* from pTH1931 complemented the arginine auxotrophy of RmP3242 (P3109, ΔB145) (Figure 5.4E).

*5.4.5 α-glucosides periplasmic substrate binding protein (*aglE*)*

RmP3108 (ΔpSymAB, *aglE-1*::Tn*5*-B20) does not grow on sucrose (Figure 5.2), but this strain grows well with cellobiose and glucose as carbon sources (data not shown). Complementation of the sucrose growth phenotype of *S. meliloti* RmP3108 identified library clones carrying a 26 kb region (nt: 327,602–354,008) from pSymB. Introduction of the pSymB deletion ΔB182 (nt: 122,108–466,499) into RmP3107 (Rm2011, *aglE*::Tn*5*- B20) confirmed the redundant locus to be within this region (data not shown). Located within this region is the *thu* ABC transporter, involved in uptake of several α-glucosides (Jensen *et al.*, 2002). As it has previously been shown that the *thu* transporter is redundant with the *agl* ABC transporter for sucrose uptake (Jensen *et al.*, 2002; Willis & Walker, 1999), this redundancy was not further studied.

*5.4.6 1,4-alpha-glucan branching enzyme (*glgB1*)*

No significant growth of RmP3106 (ΔpSymAB, *glgB1-1*::Tn*5*-B20) was observed on M9-sucrose plates (Figure 5.2). Slow growing colonies with a mucoid phenotype were observed on M9-sucrose plates for *S. meliloti* RmP3105 (Rm2011, *glgB1-1*::Tn*5*-B20) (Figure 5.2), while little growth was observed for this strain in liquid M9-sucrose

medium (Figure 5.4F). The five complementing clones that were analyzed carried the wild-type *glgB1* and we therefore failed to identify a pSymA or pSymB locus redundant with *glgB1*. However, transduction of *glgB1-1*::Tn*5*-B20 into strains lacking either pSymA or pSymB indicated that the severity of the phenotype of the *glgB1-1*::Tn*5*-B20 mutation was increased by the lack of either pSymA or pSymB (Supplementary Figure 5.S1).

*5.4.7 Phosphoglycerate kinase (*pgk*)*

All five of the isolated cosmids that complemented *S. meliloti* RmP3102 (ΔpSymAB, *pgk-1*::Tn*5*-B20) encoded *pgk*. Thus, no redundant locus was identified via complementation. Intriguingly, transduction of *pgk-1*::Tn*5*-B20 into strains lacking just pSymA or pSymB revealed the phenotype of the *pgk-1*::Tn*5*-B20 mutation was impacted by the absence of either pSymA or pSymB (Figure 5.1, Supplementary Figure 5.S1). Growth studies revealed that while RmP3101 (Rm2011, *pgk-1*::Tn*5*-B20) did grow in M9-sucrose, it was significantly impaired (Figure 5.4G). Slow growth of RmP3101 relative to the wild type was also observed in M9-glucose, while no growth was observed in M9-succinate (Figure 5.4G). However, growth was largely indistinguishable from that of the wild type when grown in M9 with both glucose and succinate (Figure 5.4H).

*5.4.8 Acetylornithine aminotransferase (*argD*)*

A Tn*5*-B20 insertion within *argD* was isolated that resulted in L-arginine auxotrophy in the ΔpSymAB strain, but not the wild-type *S. meliloti* Rm2011 (Figure 5.2 and data not shown). Transduction of *argD-1*::Tn*5*-B20 into strains lacking either

pSymA or pSymB revealed the arginine auxotrophy of *argD-1*::Tn*5*-B20 was only observed in the absence of pSymB (Supplementary Figure 5.S1). However, we did not examine the precise location of the redundant locus.

5.5 Discussion

5.5.1 Identification and characterization of functionally redundant genes

Of the transposon insertions in the *S. meliloti* chromosome that resulted in an inability of the ΔpSymAB strain to grown on M9-sucrose, 10-15% generated phenotypes dependent on the absence of the pSymA or pSymB replicons. The identified chromosomal genes with a functionally redundant gene on pSymA or pSymB were involved in amino acid biosynthesis (*proC*, *argH1*, and *argD*), central carbon metabolism (*edd*), and transport (*aglE*).

S. meliloti does not possess a complete Embden–Meyerhof–Parnas (EMP) glycolytic pathway, as it does not encode a phosphofructokinase enzyme (Arias *et al.*, 1979; Galibert *et al.*, 2001). Instead, glycolytic substrates such as sucrose and glucose are metabolized via the Entner–Doudoroff (ED) pathway to pyruvate and glyceraldehyde-3-phosphate, which is further metabolized to form a second pyruvate via the lower half of the EMP pathway (Figure 5.5) (Geddes & Oresnik, 2014; Stowers, 1985). Gluconeogenesis proceeds via the EMP pathway in *S. meliloti* and does not involve the ED pathway (Figure 5.5) (Finan *et al.*, 1988). The gene *edd* encodes a 6-phosphogluconate dehydratase that catalyzes the first step unique to the ED pathway. Despite previous study on carbon metabolism in *S. meliloti* [e.g., (Finan *et al.*, 1988)],

mutants defective in 6-phosphogluconate dehydratase activity have never been isolated. Consistent with this, disrupting *edd* decreases, but does not prevent, growth on sucrose (Figure 5.4A). However, combining this disruption with a deletion of a pSymA-encoded gene, *sma0235*, results in a synthetic phenotype and very little growth with sucrose as the sole carbon source (Figure 5.4A). As *sma0235* shows homology to dehydratases, in particular, dihydroxyacid and 6-phosphogluconate dehydratases, it is likely that the gene products of *edd* and *sma0235* share overlapping enzymatic specificity. However, while Sma0235 seemingly has activity with 6-phosphogluconate, this is unlikely to be the primary substrate of this enzyme for two reasons: one, *sma0235* in unable to completely complement the disruption of *edd*; two, *sma0235* is the third gene of a three-gene operon that also consists of an epimerase (the first gene, *sma0241*) and a dehydrogenase (the second gene, *sma0237*), suggesting that this operon is involved in a metabolic pathway distinct from that of the ED pathway.

L-proline synthesis from L-glutamate occurs via a three-enzyme pathway, with the final step the conversion of Δ^1-pyrroline-5-carboxylate to L-proline catalyzed by ProC (Figure 5.6). However, disrupting the chromosomal *proC* gene had no phenotype unless combined with a deletion of the pSymB-encoded gene *smb20003* (Figure 5.4C). As *smb20003* shares the same annotation of *proC*, a Δ^1-pyrroline-5-carboxylate reductase, it is likely that the synthetic effect observed in the double mutant is due to the elimination of enzymatic, not pathway, redundancy. Transcriptional start site mapping indicates that *smb20003* is transcribed as a monocistronic mRNA (Schlüter *et al.*, 2013), while a

reciprocal best-hit strategy suggests that both ProC and Smb20003 orthologs are present throughout the *Rhizobiaceae* family (data not shown). In addition to redundancy of *proC*, two genes are annotated as encoding the ProB enzyme on the *S. meliloti* chromosome, while *proA* appears to present in a single copy (Figure 5.6). This is similar to the situation in *B. subtilis*, which encodes two orthologs of *proB*, three of *proC*, and only one ortholog of *proA* (Belitsky *et al.*, 2001). While the two *proB* and three *proC* homologs in *B. subtilis* can complement each other, the transcriptional up-regulation of each homolog occurs in response to unique stimuli (Belitsky *et al.*, 2001).

In addition to synthesis of L-proline from L-glutamate, alternate L-proline biosynthetic pathways exist in prokaryotic species, such as the conversion of L-ornithine to L-proline through the activity of ornithine cyclodeaminases (Figure 5.6). Previous work has identified ornithine cyclodeaminase activity in *S. meliloti* cell extracts (Soto *et al.*, 1994), and four genes (*ocd*, *eutC*, *sma0486*, *sma1871*) are annotated as encoding ornithine cyclodeaminases in *S. meliloti* Rm2011. Indeed, auxotrophy of the *proC, smb20003* double mutant was eliminated by adding exogenous L-ornithine (Figure 5.4D). Furthermore, exogenous L-arginine also eliminated the proline auxotrophy (Figure 5.4D), presumably as L-arginine can be converted to L-ornithine either through an arginase (i.e., ArgI1 or ArgI2) or the ArcABC catabolic pathway (Figure 5.6). Similarly, L-arginine and L-ornithine can be converted to L-proline in *B. subtilis*; however, unlike in *S. meliloti*, this pathway proceeds through a ProC-catalyzed final step (Belitsky *et al.*, 2001). While we are unsure why such extensive enzymatic and pathway level redundancy in L-proline

biosynthesis exists, its presence in highly diverse species suggests it may provide a fitness advantage to the cell.

Two of the redundant gene pairs encode proteins involved in the eight-step pathway converting L-glutamate to L-arginine (Figure 5.6). In the final step of this pathway, L-argininosuccinate lyase (ArgH) converts L-argininosuccinate to L-arginine. Two *argH* genes are annotated in the *S. meliloti* genome, one on the chromosome (*argH1*) and one on pSymB (*argH2*). Consistent with this function being redundant, disruption/deletion of both genes renders *S. meliloti* an arginine auxotroph, whereas the presence of either gene individually is sufficient for arginine prototrophy (Figure 5.4E). The gene *argH2* is the first gene of an operon encoding an ATP transporter (*smb21095*–*smb21097*) and two hypothetical genes (*smb21098*, *smb21100*). Previous work has shown that transcription of this ATP transporter is induced in the presence of L-citrulline (Mauchline *et al.*, 2006), whose conversion to L-arginine or catabolism through the urea cycle involves an L-argininosuccinate lyase (ArgH)-catalyzed step (Kanehisa *et al.*, 2014). Thus, the presence of two copies of *argH* in the *S. meliloti* genome presumably allows differential regulation and synthesis of the L-argininosuccinate lyase isozymes: induction of *argH1* by a lack of L-arginine and the upregulation of *argH2* by the presence of L-citrulline.

The second Tn*5*-B20 insertion within a redundant arginine biosynthetic locus was situated near the 3′ end of *argD*, encoding an L-acetylornithine aminotransferase. The gene *argD* is situated less than 100 base pairs upstream of *argF1*, annotated as encoding

an L-ornithine carbamoyl-transferase that is presumably also involved in the arginine biosynthetic pathway (Figure 5.6). Two independent RNA-seq experiments suggest the presence of transcriptional start sites upstream of *argD* and upstream of *argF1* (Milunovic *et al.*, 2014; Schlüter *et al.*, 2013); however, a promoter motif was only identified upstream of *argD* (Schlüter *et al.*, 2013). Thus, it is unclear whether *argD* and *argF1* form a bicistronic operon. Furthermore, the Tn*5*-B20 transposon inserted downstream of all codons encoding conserved residues in ArgD, meaning a functional version of ArgD, may still have been translated. The redundant locus (or loci) is located on pSymB (Supplementary Figure 5.S1); however, we did not attempt to identify the precise gene(s), and therefore cannot state whether the *argD-1*::Tn*5*-B20 phenotype is associated with the loss of ArgD, ArgF1, or both. Nevertheless, we note that ArgD shows similarity to five pSymB aminotransferases (as well as five encoded by the chromosome and four encoded by pSymA), while ArgF1 shows similarity to one pSymB-encoded protein (as well as one pSymA-encoded protein).

5.5.2 Additional genome-dependant phenotypes

Phosphoglycerate kinase (Pgk) catalyzes the reversible conversion of 1,3-diphosphate-glycerate and 3-phosphoglycerate, and is thus essential for growth of *S. meliloti* with gluconeogenic substrates (Figure 5.5) (Finan *et al.*, 1988). While Pgk is also involved in the glycolytic pathway, it is not absolutely required as one of the pyruvic acid products of the ED pathway bypasses Pgk (Figure 5.5) (Finan *et al.*, 1988). Here, it was observed that *pgk-1*::Tn*5*-B20 mutants grew slowly with glycolytic substrate (sucrose and

glucose) as sole carbon sources only when pSymA and pSymB were a part of the genome (Figures 5.2, 5.4G, Supplementary Figure 5.S1). We hypothesize that disruption of *pgk* leads to a buildup of glycolytic intermediates that, unless removed through pSymA- and pSymB-dependent pathways, results in an eventual cessation of glycolysis and no formation of pyruvate (Figure 5.5). Similarly, it was previously suggested that pathways removing glycolytic intermediates explain, in part, why *pgk* mutations in *Pseudomonas aeruginosa* do not prevent growth on glycolytic substrate, unlike in *E. coli* (Banergee *et al.*, 1987).

S. meliloti predominately stores carbon as two polymers, poly-3-hydroxybutyrate and glycogen, which appear to have separate functions during N_2-fixing symbiosis with legumes (Wang *et al.*, 2007). Glycogen is synthesized via a three-enzyme pathway consisting of a glucose-1-phosphate adenylyltransferase (GlgC), glycogen synthetase (GlgA), and a glycogen branching enzyme (GlgB) (Kanehisa *et al.*, 2014). Little growth of *glgB1-1*::Tn*5*-B20 mutants was observed in liquid M9-sucrose medium (Figure 5.4F), although slow growth was observed on M9-sucrose plates unless pSymA and pSymB were removed (Figure 5.2, Supplementary Figure 5.S1). No redundant pSymA- or pSymB-encoded locus was identified through complementation with the genomic DNA cosmid library. Additionally, combining the *glgB1-1*::Tn*5*-B20 allele with pSymB deletions of candidate genes (*glgB2*, *bdhA*, *bhbA*, and the exopolysaccharide biosynthetic cluster) failed to reveal synthetic interactions (data not shown). Hence, the mechanistic basis for the pSymA or pSymB redundancy remains unknown.

5.5.3 Genetic redundancy in bacterial organisms

Redundant gene pairs encoded 10-15% of functions required for growth on minimal M9-sucrose medium. Considering the potential for intra-chromosomal redundancy, which would not have been detected in the present study, we posit that this is underestimated the total genetic redundancy encoding processes essential for growth on M9-sucrose. These results illustrate that genetic redundancy may be extensive within large bacterial genomes, an observation that raises several theoretical questions and has significant practical implications.

It may not be particularly surprising to find significant genetic redundancy in the *S. meliloti* genome given that ~ 40% of the predicted genes belong to gene families, although few of these are the result of recent duplications (Galibert *et al.*, 2001). While it is common for large bacterial genomes to have a high percentage of genes belonging to gene families (Glass *et al.*, 2005), our data show that one must be careful to extrapolate from amino acid similarity to genetic redundancy and functional complementation. As an example, Edd shows significant amino acid similarity to six proteins (IlvD4, IlvD2, IlvD3, IlvD1, AraF, Sma0235). Yet, only Sma0235 is able to partially complement *edd* mutants despite being the most dissimilar protein to Edd of the six (Edd:Sma0235, 51.5 % aa similarity; Edd:others, 54.4-54.6 % aa similarity). In many cases, two or more proteins may have similar molecular functions but serve unique biological roles and thus do not complement mutations of the others, such as the five *S. meliloti* copper ABC transporters (Patel *et al.*, 2014). Thus, while bioinformatics provides insight into the potential level of

genetic redundancy in a genome, an experimental approach is required for a full understanding.

A high level of genetic redundancy would seem to be inconsistent with the streamlined nature of prokaryotic genomes. Genetic redundancy could simply be the result of a recent gene duplication and an insufficient amount of time for purifying selection to lead to the loss of one copy. However, this is unlikely to be true in examples where the redundant genes are conserved (e.g., *proC*). Redundant genes can serve to increase the rate of production of highly expressed gene products, such as is posited for rRNA (Bremer, 1975; Condon *et al.*, 1995) and experimentally increasing gene copy number can increase gene expression (Janczarek *et al.*, 2009). Functionally redundant proteins could also be maintained if they show partial substrate overlap. For example, both AglE and ThuE bind sucrose, but only AglE binds cellobiose while ThuE uniquely binds palatinose (Ampomah *et al.*, 2013; Jensen *et al.*, 2002). Moreover, differential regulation could serve as a driving force for the maintenance of redundant genes by optimizing metabolic versatility, similar to how the redundant proline biosynthetic genes in *B. subtilis* are up-regulated in response to different stimuli (Belitsky *et al.*, 2001).

5.5.4 Practical implications/consequences of genetic redundancy

Genetic redundancy in prokaryotic genomes has practical implications/consequences, primarily the masking of phenotypes during loss-of-function studies. This would limit our ability to identify the genetic determinants of a pathway or to ascertain the biological role of a gene through a mutagenesis-based approach. For

example, the last step of gluconeogenesis in *S. meliloti* (Figure 5.5) remains uncharacterized despite intensive characterization of this pathway [e.g., (Finan *et al.*, 1988)], perhaps reflective of redundancy in this step. Phenotype masking through genetic redundancy also impacts the conclusions drawn from high-throughput transposon mutagenesis studies. The use of saturation transposon mutagenesis has been used in the global mapping of essential genes and identification of the minimal bacterial genome (e.g., Christen *et al.*, 2011; Glass *et al.*, 2005; Jacobs *et al.*, 2003), bacterial virulence (e.g., Gawronski *et al.*, 2009; van Opijnen & Camilli, 2012; Wu *et al.*, 2006), and other biological processes (e.g., Cameron *et al.*, 2008; Stewart *et al.*, 2004). While such studies provide invaluable genetic information into many biological processes, it must be recognized that greater than 10% of target genes may be missed due to genetic redundancy. A failure to consider redundantly encoded functions could lead to an incomplete understanding of many important biological processes. Not only is this an issue in the sense of fully understanding the biology of the cell, an incomplete understanding of biological pathways will hinder our ability to use synthetic biology to construct designer cell factories.

5.6 Materials and methods

5.6.1 Media, growth conditions, and bacterial strains

All media [LB, LBmc (LB plus 2.5 mM $MgSO_2$ and 2.5 mM $CaCl_2$), TY, M9] were prepared as previously described (diCenzo *et al.*, 2014). For growth of *S. meliloti*, all complex media (LB, LBmc, TY) were supplemented with 2 μM cobalt chloride

(Cheng *et al.*, 2011) and the M9 minimal medium with 5 μM thiamine-HCl (diCenzo *et al.*, 2014; Finan *et al.*, 1986). Antibiotic concentrations [streptomycin (Sm), spectinomycin (Sp), neomycin (Nm), kanamycin (Km), tetracycline (Tc), gentamicin (Gm), and chloramphenicol (Cm)] and growth conditions were as described elsewhere (diCenzo *et al.*, 2014). Bacterial strains and plasmids are listed in Supplementary Table 5.S1. When used as the sole carbon source, sucrose was added to a concentration of 10 mM, while glucose and succinate were added to 15 mM. When both were present, glucose and succinate were added to 10 mM each. Where stated, L-proline, L-arginine, and L-ornithine were added at a concentration of 1 mM.

5.6.2 Genetic manipulations

General DNA manipulations and recombinant techniques, bacterial matings, isolation of genomic *S. meliloti* DNA, and ΦM12 transductions were performed as previously described (Cowie *et al.*, 2006; Finan *et al.*, 1984; Milunovic *et al.*, 2014; Sambrook *et al.*, 1989). Where necessary, plasmids were mobilized with the helper strain *E. coli* MT616 (pRK600) (Finan *et al.*, 1986). Unless stated otherwise, cloning of DNA fragments into plasmids was performed with sequence- and ligation-independent cloning (SLIC) (Jeong *et al.*, 2012). Oligonucleotides were ordered from Integrated DNA Technologies (IDT), and sequencing was performed by the MOBIX facility at McMaster University, Hamilton, Ontario, Canada. Oligonucleotide sequences used in this study are listed in Supplementary Table 5.S2.

5.6.3 Transposon mutagenesis

The Tn*5*-B20 transposon (Simon *et al.*, 1989) was transferred to *S. meliloti* ΔpSymAB (RmP2917) via the self-transmissible suicide vector pRK600::Tn*5*-B20. Transposon insertion mutants were selected on TY Sm Nm or LBmc Sm Nm and 40 independent matings were performed. Transposon insertion mutants were screened for an inability to grow on M9-sucrose and M9-cellobiose. Following the isolation of genomic DNA from mutants of interest, the location of the insertion was determined by sequencing out from the 3′ end of the transposon using the primer ML-1160 and mapping reads to the *S. meliloti* 1021 genome sequence (Galibert *et al.*, 2001) available online at https://iant.toulouse.inra.fr/bacteria/annotation/cgi/rhime.cgi.

5.6.4 Complementation

A pLAFR1 cosmid library of *S. meliloti* 1021 genomic DNA (Friedman *et al.*, 1982) was mated *en masse* into *S. meliloti* ΔpSymAB and complemented cells were selected for on M9-sucrose Sp Tc. Transconjugants were purified and complementing clones were conjugated into *E. coli* via a triparental mating with *E. coli* MT616 and rifampicin-resistant *E. coli* DH5α. Transconjugants were selected on either LB Rif (20 μg uL^{-1}) Tc or LB Tc and recovered on LB Tc at a frequency of ~ 10^{-8} /donor and ~ 10^{-8} /recipient. Cosmids were re-introduced into the appropriate *S. meliloti* ΔpSymAB transposon insertion mutant and screened for growth on M9-sucrose to confirm the complementation. The border ends of the *S. meliloti* DNA insert in the pLAFR1 cosmids were identified through Sanger sequencing with primers P128 and P129.

5.6.5 Growth curves

Cultures grown in LBmc were washed once using carbon-free M9 medium, then resuspended and diluted into the M9 medium in which the growth curve was to be performed. All strains were grown in triplicate in 96-well microtitre plates. Growth conditions and analysis were as previously described (diCenzo *et al.*, 2014).

5.6.6 Construction of pTH2919, a Tc^R sacB vector

To construct a Tc^R *sacB* vector, the Gm^R gene was removed from pJQ200mp18 (Quandt & Hynes, 1993) through digestion with *Eco*RV and *Apa*LI. The Tc^R gene from pBBR1mcs-3 (Kovach *et al.*, 1995) was PCR amplified with the primers DF001 and DF002, and introduced into the *Eco*RV/*Apa*LI digested pJQ200mp18, creating pTH2919.

5.6.7 Construction of plasmids for allelic replacement

Plasmids pTH2982 and pTH2983 were constructed to replace *sma0235* and *smb20003*, respectively, with the gentamicin resistance gene $aacC_4$. Approximately 500 nucleotide fragments upstream and downstream of *sma0235* (primers: DF003/DF004 and DF005/DF006) and *smb20003* (primers: DF007/DF008 and DF009/DF010) were PCR amplified, and each pair was simultaneously cloned into *Xba*I digested pTH2919, resulting in plasmids pTH2978 and pTH2979, respectively. These plasmids were digested with *Swa*I and ligated with a DNA fragment encoding $aacC_4$, which was PCR amplified from pHP45Ωaac (Blondelet-Rouault *et al.*, 1997) using the primer DF011, resulting in the plasmids pTH2982 and pTH2983, respectively.

5.6.8 Construction of double deletions

To delete pSymA or pSymB genes that were putatively redundant with a chromosomal gene, Gm^R double recombinants of pTH2982 and pTH2983 into *S. meliloti* Rm2011 and the appropriate Rm2011 Tn*5*-B20 insertion mutants (pTH2982 into *edd*::Tn*5*-B20, pTH2983 into *proC*::Tn*5*- B20) were isolated. Plasmids were transferred into the appropriate recipient through a triparental mating on LB, and single recombinants were selected for on TY Sm Gm. Double recombinants were identified by screening for Tc sensitivity and for growth on LB Gm and on LB + 8 % sucrose.

5.6.9 Construction of pTH2987

pTH2987 was constructed to express *argH2 in trans* in *S. meliloti*. The coding region of *argH2* was PCR amplified from *S. meliloti* Rm2011 genomic DNA with primers DF012 and DF013, and was inserted into *Pac*I digested pTH1931 (diCenzo *et al.*, 2013), where *argH2* expression was driven from the P_{trc} promoter.

5.6.10 Sequence analysis

Nucleotide identity and amino acid similarity were determined following a global alignment of two sequences using the LALIGN algorithm provided on the ExPASy web server (Gasteiger *et al.*, 2003; Huang & Miller, 1991).

5.9 Figures

Figure 5.1. Schematic representation of the experimental workflow.

(**A**) A Tn*5*-B20 random insertion mutant library (represented by the arrowheads) of *S. meliloti* ΔpSymAB (RmP2917) was constructed. (**B**) The mutant library was replica plated (patched) on M9-sucrose, M9-cellobiose, and LB. Mutants unable to grow on M9-sucrose and M9-cellobiose were identified, as indicated by the dashed boxes. (**C**) The Tn*5*-B20 insertions from the mutants unable to grow on both minimal media were recombined (transduced) into naïve *S. meliloti* ΔpSymAB and wild-type *S. meliloti* Rm2011. (**D**) These recombinants were screened for the ability to grow on M9-sucrose. Transposon insertions resulting in no growth on M9-sucrose only in the absence of pSymA and pSymB were identified, as indicated by the dashed boxes. The location of these transposon insertions was determined via Sanger sequencing, as they were putatively located in genes with a redundant copy on pSymA/pSymB. (**E**) The loci on pSymA/ pSymB that complemented the phenotypes of the transposon insertions were identified using two methods. *S. meliloti* ΔpSymAB transposon mutants were complemented with a genomic DNA cosmid library, as represented by the small circle. Additionally, *S. meliloti* Rm2011 transposon mutants were combined with large-scale deletions on pSymA/pSymB, as represented by the dashed arch adjacent to pSymB. (**A-E**) Chr - chromosome, pA - pSymA, pB - pSymB.

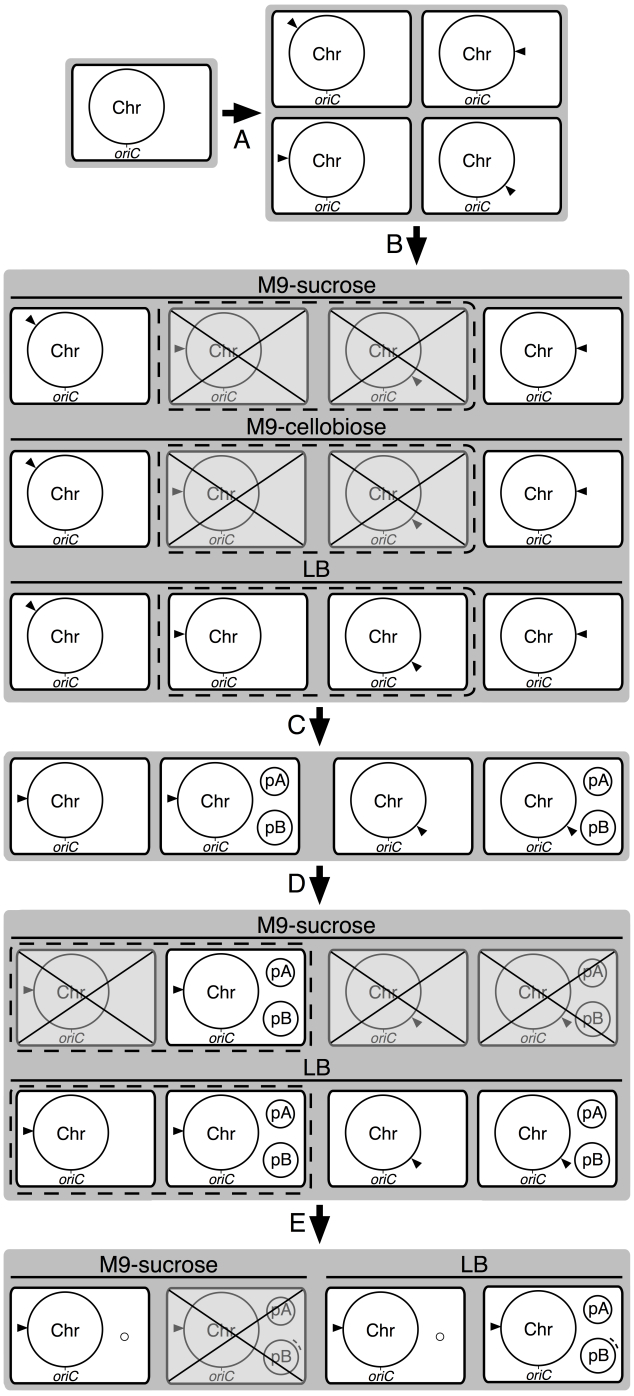

Chr
oriC
A
B
M9-sucrose
M9-cellobiose
LB
C
pA
pB
D
E

Figure 5.2. Strain specific phenotypes of Tn*5*-B20 insertions.

Twelve transposon insertions, located within seven unique genes, prevented growth of *S. meliloti* on M9-sucrose agar plates in *S. meliloti* ΔpSymAB (RmP2917), but not Rm2011 (wild type). Pictures of the growth of representative alleles are shown. *S. meliloti* Rm2011 and the corresponding Tn*5*-B20 insertion mutants were incubated for 4 days, while *S. meliloti* ΔpSymAB and the corresponding Tn*5*-B20 insertion mutants were incubated for 5 days.

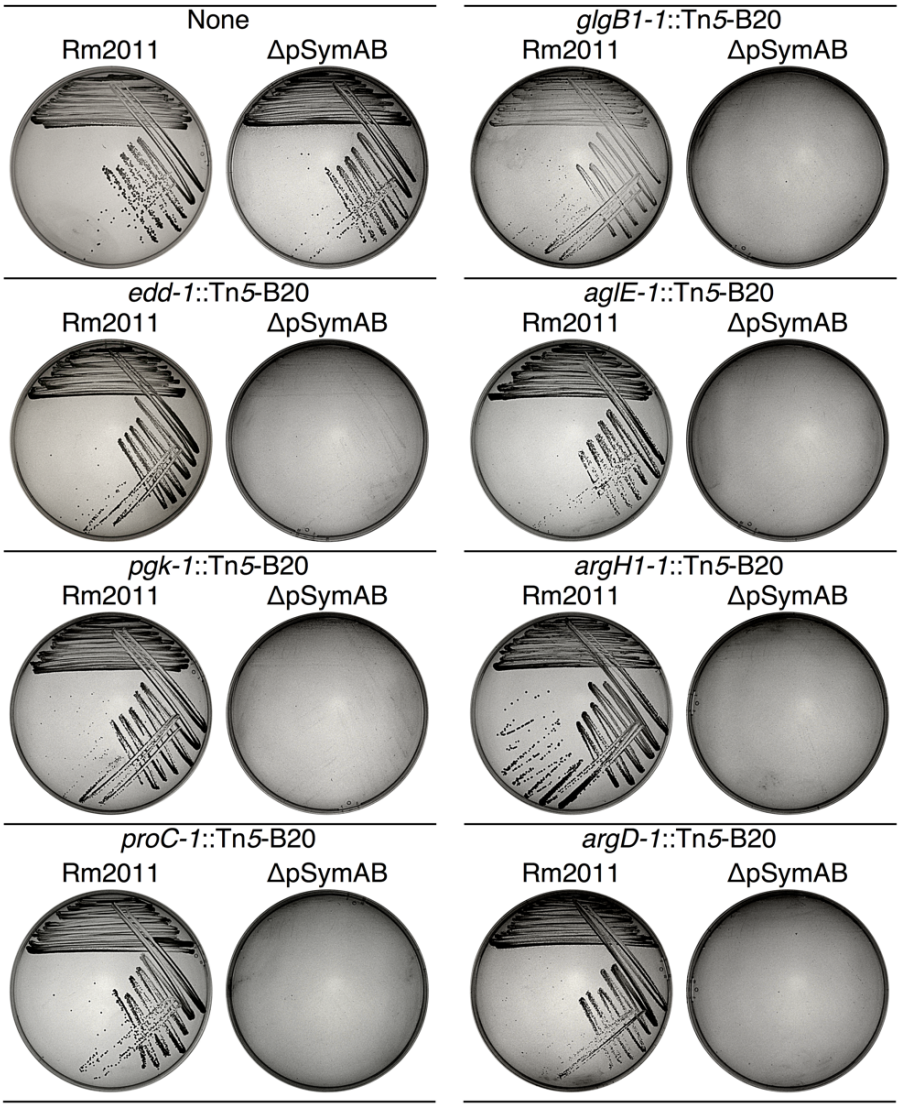

None
Rm2011
ΔpSymAB
glgB1-1::Tn5-B20
Rm2011
ΔpSymAB
edd-1::Tn5-B20
Rm2011
ΔpSymAB
aglE-1::Tn5-B20
Rm2011
ΔpSymAB
pgk-1::Tn5-B20
Rm2011
ΔpSymAB
argH1-1::Tn5-B20
Rm2011
ΔpSymAB
proC-1::Tn5-B20
Rm2011
ΔpSymAB
argD-1::Tn5-B20
Rm2011
ΔpSymAB

Figure 5.3. The genomic location of the genes and deletions of interest in this study. Black arrowheads indicate the locations of genes disrupted by a Tn*5*-B20 transposon. Gray arrowheads signify the position of pSymA of pSymB genes able to complement the disruption of chromosomal genes. The dotted curved lines indicate the regions of pSymA and pSymB removed in the deletions used in this study. The position of the *oriC* of the chromosome, *repA1* on pSymB, and *repA2* on pSymA are pointed out for reference. Chr - chromosome, pA - pSymA, pB - pSymB.

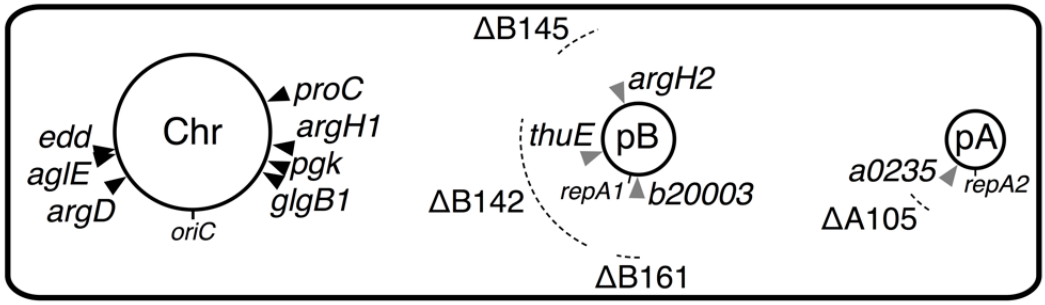

ΔB145
argH2
proC
argH1
pgk
glgB1
edd
aglE
argD
Chr
oriC
thuE
pB
repA1
b20003
ΔB142
ΔB161
pA
a0235
repA2
ΔA105

Figure 5.4. Growth profiles of Tn*5*-B20 insertion mutants and associated strains.

(**A**,**B**) Growth of wild type (square), *edd-1*::Tn*5*-B20 (diamond), Δ*sma0235*::$aacC_4$ (triangle), and *edd-1*::Tn*5*-B20 + Δ*sma0235*::$aacC_4$ (circle) in M9-sucrose (**A**) or M9-succinate (**B**). (**C**) Growth of wild type (square), *proC-1*::Tn*5*-B20 (diamond), Δ*smb20003*::$aacC_4$ (triangle), and *proC-1*::Tn*5*-B20 + Δ*smb20003*::$aacC_4$ (circle) in M9-sucrose. (**D**) Growth of wild type in M9-sucrose (square) and *proC-1*::Tn*5*-B20 + Δ*smb20003*::$aacC_4$ in M9-sucrose with no amino acid supplementation (circle), or L-proline (diamond), L-ornithine (triangle), or L-arginine (cross) supplementation. (**E**) Growth of wild type (square), *argH1-1*::Tn*5*-B20 (diamond), ΔB145 (triangle), *argH1-1*::Tn*5*-B20 + ΔB145 (closed circle), and *argH1-1*::Tn*5*-B20 + ΔB145 + *argH2 in trans* (cross) in M9-sucrose, as well as *argH1-1*::Tn*5*-B20 + ΔB145 in M9-sucrose supplemented with L-arginine (open circle). (**F**) Growth of wild type (square) and *glgB1-1*::Tn*5*-B20 (diamond) in M9-sucrose. (**G**) Growth of wild type (solid symbols) and *pgk-1*::Tn*5*-B20 (open symbols) in M9-sucrose (square), M9-glucose (triangle), and M9-succinate (circle). (**H**) Growth of wild type (closed diamond) and *pgk-1*::Tn*5*-B20 (open diamond) in M9-glucose + succinate. (**A**-**H**) Data points and error bars represent the average and standard deviation of triplicate samples. Curves were produced by plotting data points from readings every 2 hr, with the data points shown for every 8 hr.

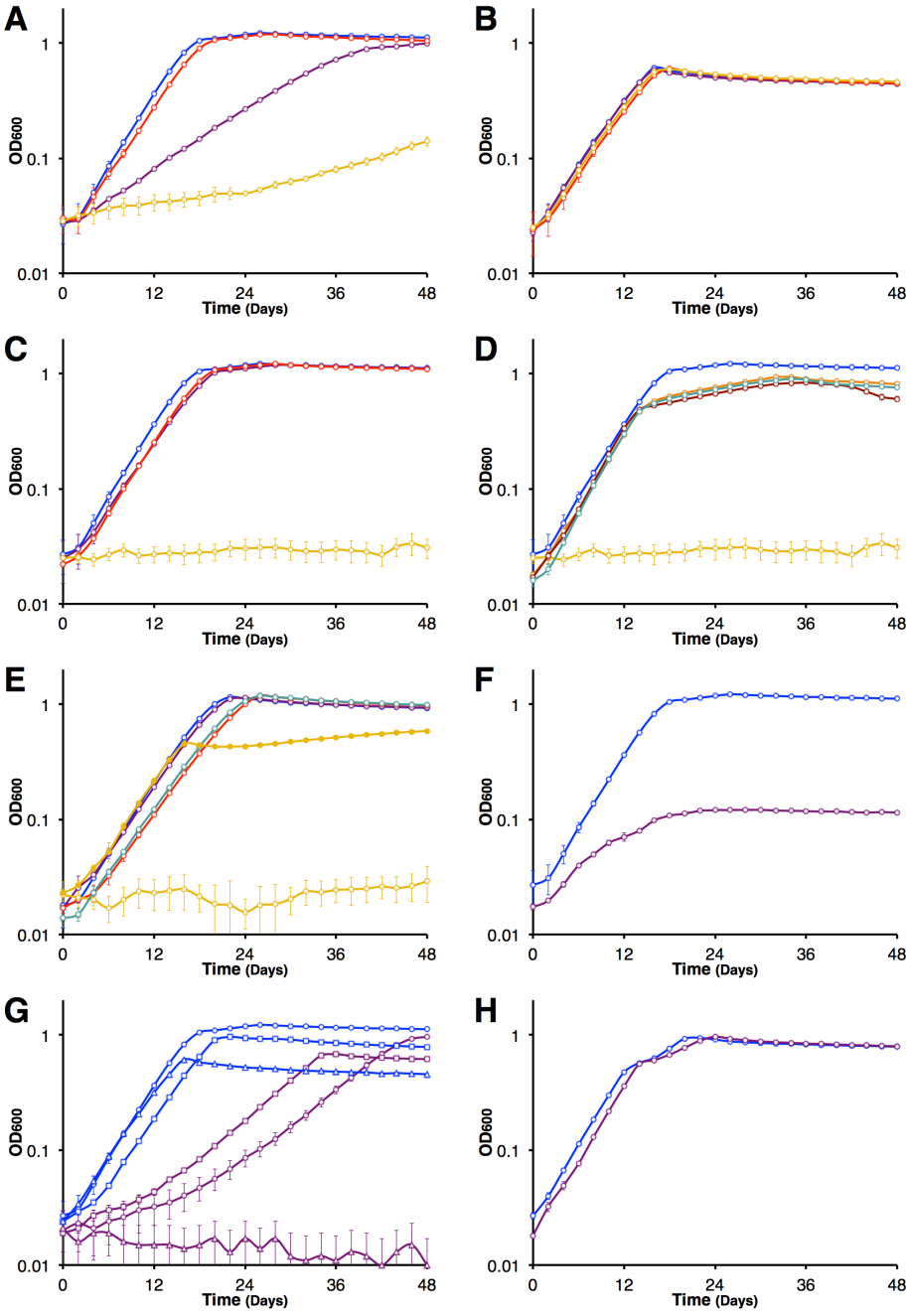

A
B
C
D
E
F
G
H
OD600
1
0.1
0.01
0
12
24
36
48
Time (Days)

Figure 5.5. Schematic of the glycolytic (GLY) and gluconeogenic (GNG) pathways of *S. meliloti*.

The short dashed gray box indicates the steps of the GLY pathway, while the long dashed black box indicates the steps of the GNG pathway. Proteins catalyzing each step are indicated. Proteins with experimental evidence supporting their designated molecular function are indicated with boldface, while the rest are classified based solely on annotation. Proteins of interest in this study are highlighted with a gray box. TCA - tricarboxylic acid cycle, PPP - pentose phosphate pathway, DHAP - dihydroxyacetone phosphate, ??? - enzyme unknown. For a review of carbon metabolism in *S. meliloti*, please refer to (Geddes & Oresnik, 2014).

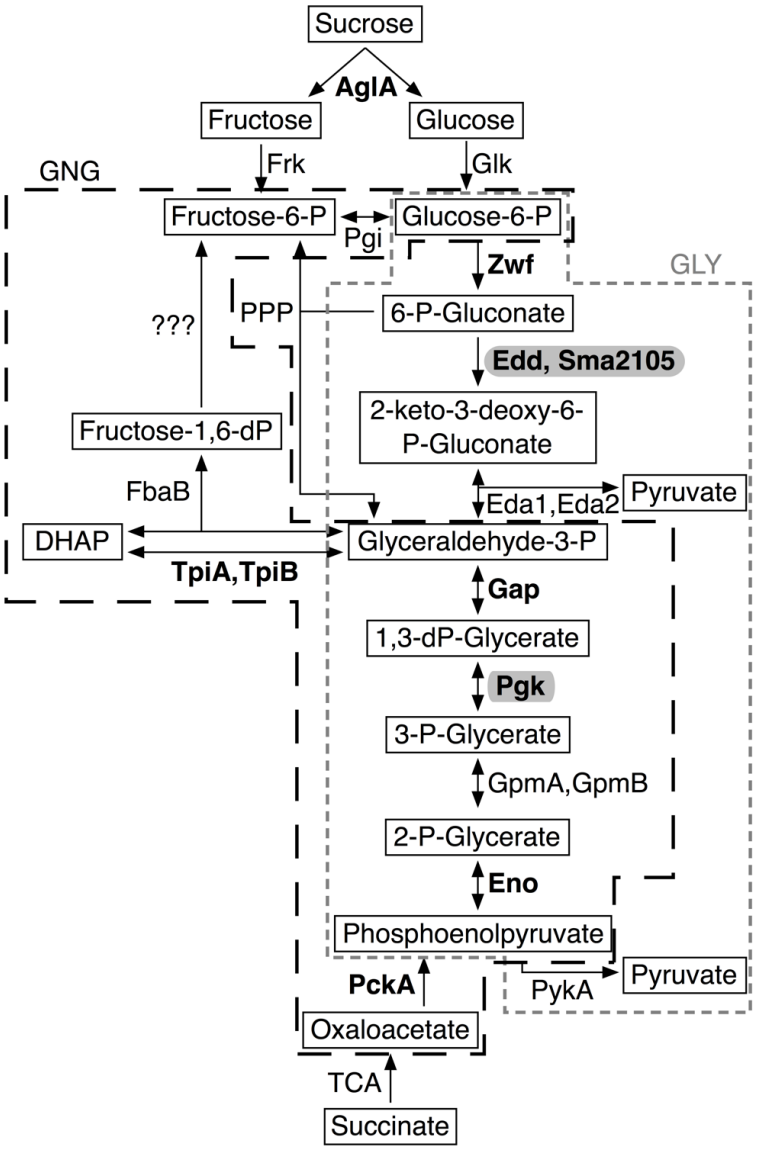
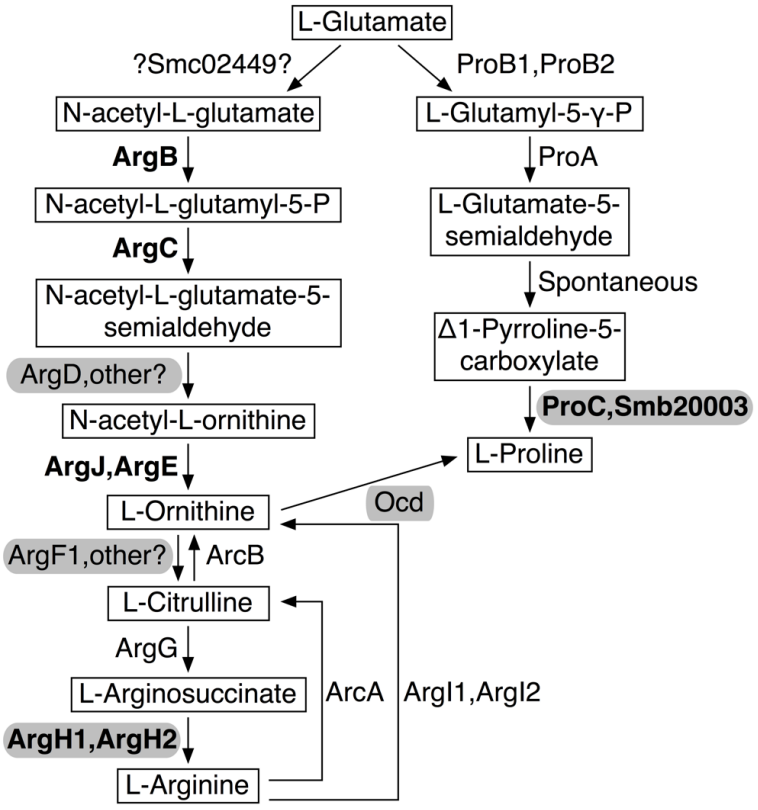

Sucrose
AglA
Fructose
Glucose
GNG
Frk
Glk
Fructose-6-P
Pgi
Glucose-6-P
Zwf
GLY
???
PPP
6-P-Gluconate
Edd, Sma2105
2-keto-3-deoxy-6-
P-Gluconate
Fructose-1,6-dP
FbaB
Eda1,Eda2
Pyruvate
DHAP
TpiA,TpiB
Glyceraldehyde-3-P
Gap
1,3-dP-Glycerate
Pgk
3-P-Glycerate
GpmA,GpmB
2-P-Glycerate
Eno
Phosphoenolpyruvate
PckA
PykA
Pyruvate
Oxaloacetate
TCA
Succinate

Figure 5.6. Schematic of the L-arginine and L-proline biosynthetic pathways of *S. meliloti*.

Proteins catalyzing each step are indicated. Proteins with experimental evidence supporting their designated molecular function are indicated with boldface, while the rest are classified based solely on annotation. The enzyme catalyzing the first reaction in the L-arginine pathway is unknown; however, it may be Smc02449 based on genome context and homology to the *Corynebacterium glutamicum* NAG synthase (data not shown). Proteins of interest in this study are highlighted with a gray box. For a review of amino acid metabolism in rhizobia, please refer to (Dunn, 2015).


L-Glutamate
?Smc02449?
ProB1,ProB2
N-acetyl-L-glutamate
L-Glutamyl-5-γ-P
ArgB
ProA
N-acetyl-L-glutamyl-5-P
L-Glutamate-5-semialdehyde
ArgC
Spontaneous
N-acetyl-L-glutamate-5-semialdehyde
Δ1-Pyrroline-5-carboxylate
ArgD,other?
ProC,Smb20003
N-acetyl-L-ornithine
L-Proline
ArgJ,ArgE
Ocd
L-Ornithine
ArgF1,other?
ArcB
L-Citrulline
ArgG
L-Arginosuccinate
ArcA
ArgI1,ArgI2
ArgH1,ArgH2
L-Arginine

5.10 Supplementary Materials

5.10.1 Supplementary Tables and Figures

Table 5.S1. Bacterial strains and plasmids.

Strain/Plasmid	Characteristics	Source/Reference
Sinorhizobium meliloti		
Rm2011	Wild type SU47 *str-3*; Sm^R	M. Hynes
RmP934	RmP110, ΔA105 (pSymA nt: 125,128-184,519); Sm^R Nm^R Gm^R	Milunovic et al. (2014)
RmP1109	RmP110, ΔB145 (pSymB nt: 635,940-744,320); Sm^R Nm^R Gm^R	Milunovic et al. (2014)
RmP2707	RmP110, ΔB181 (pSymB nt: 1,679,723-49,523); Sm^R Nm^R Gm^R	Lab collection
RmP2713	RmP110, ΔB182 (pSymB nt: 101,396-345,341); Sm^R Nm^R Gm^R	Lab collection
RmP2917	Rm2011, ΔpSymA/pSymB, *engA*/$tRNA^{arg}$ in chromosome; Sm^R Sp^R	diCenzo et al. (2014)
RmP3009	Rm2011, ΔpSymB, engA/tRNAarg in chromosome; Sm^R Sp^R	diCenzo et al. (2014)
RmP3099	Rm2011, *edd-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3100	RmP2917, *edd-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3101	Rm2011, *pgk-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3102	RmP2917, *pgk-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3103	Rm2011, *proC-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3104	RmP2917, *proC-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3105	Rm2011, *glgB1-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3106	RmP2917, *glgB1-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3107	Rm2011, *algE-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3108	RmP2917, *algE-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3109	Rm2011, *argH1-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3110	RmP2917, *argH1-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3190	SmA818 (*pgk-1*::Tn*5*-B20); Sm^R Nm^R	This study
RmP3191	RmP3009 (*pgk-1*::Tn*5*-B20); Sm^R Sp^R Nm^R	This study
RmP3195	SmA818 (*glgB1-1*::Tn*5*-B20); Sm^R Nm^R	This study
RmP3196	RmP3009 (*glgB1-1*::Tn*5*-B20); Sm^R Sp^R Nm^R	This study
RmP3224	RmP3100 (pTH2874); Sm^R Sp^R Nm^R Tc^R	This study
RmP3226	RmP3102 (pTH2876); Sm^R Sp^R Nm^R Tc^R	This study
RmP3227	RmP3104 (pTH2877); Sm^R Sp^R Nm^R Tc^R	This study
RmP3228	RmP3106 (pTH2878); Sm^R Sp^R Nm^R Tc^R	This study
RmP3229	RmP3108 (pTH2879); Sm^R Sp^R Nm^R Tc^R	This study
RmP3230	RmP3108 (pTH2880); Sm^R Sp^R Nm^R Tc^R	This study
RmP3231	RmP3110 (pTH2881); Sm^R Sp^R Nm^R Tc^R	This study
RmP3232	RmP3110 (pTH2882); Sm^R Sp^R Nm^R Tc^R	This study
RmP3235	Rm2011, ΔA105 via ΦRmP934; Sm^R Nm^R Gm^R	This study
RmP3236	RmP3099, ΔA105 via ΦRmP934; Sm^R Sp^R Nm^R Gm^R	This study
RmP3237	Rm2011, ΔB181 via ΦRmP2707; Sm^R Nm^R Gm^R	This study
RmP3238	RmP3103, ΔB181 via ΦRmP2707; Sm^R Sp^R Nm^R Gm^R	This study
RmP3239	Rm2011, ΔB182 via ΦRmP2713; Sm^R Nm^R Gm^R	This study
RmP3240	RmP3107, ΔB182 via ΦRmP2713; Sm^R Sp^R Nm^R Gm^R	This study
RmP3241	Rm2011, ΔB145 via ΦRmP1109; Sm^R Nm^R Gm^R	This study
RmP3242	RmP3109, ΔB145 via ΦRmP1109; Sm^R Sp^R Nm^R Gm^R	This study

RmP3369	Rm2011, Δ*sma0235*::*aacC4* via pTH2982; Sm^R Gm^R	This study
RmP3370	RmP3099, Δ*sma0235*::*aacC4* via pTH2982; Sm^R Sp^R Nm^R Gm^R	This study
RmP3371	Rm2011, Δ*smb20003*::*aacC4* via pTH2983; Sm^R Gm^R	This study
RmP3372	RmP3103, Δ*smb20003*::*aacC4* via pTH2983; Sm^R Sp^R Nm^R Gm^R	This study
RmP3383	Rm2011, *argD-1*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3384	RmP2917, *argD-1*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3385	Rm2011, *argH1-2*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3386	RmP2917, *argH1-2*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3387	Rm2011, *argH1-3*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3388	RmP2917, *argH1-3*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3389	Rm2011, *argH1-4*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3390	RmP2917, *argH1-4*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3391	Rm2011, *edd-2*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3392	RmP2917, *edd-2*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3393	Rm2011, *edd-3*::Tn*5*-B20; Sm^R Nm^R	This study
RmP3394	RmP2917, *edd-3*::Tn*5*-B20; Sm^R Sp^R Nm^R	This study
RmP3427	RmP3242 (pTH1931); Sm^R Sp^R Nm^R Gm^R	This study
RmP3428	RmP3242 (pTH2987); Sm^R Sp^R Nm^R Gm^R	This study
RmP3449	SmA818 (*argD-1*::Tn*5*-B20); Sm^R Nm^R	This study
RmP3450	RmP3009 (*argD-1*::Tn*5*-B20); Sm^R Sp^R Nm^R	This study
SmA818	Rm2011, ΔpSymA; Sm^R	This study
Escherichia coli		
G351	MT620 (pRK600::Tn*5*-B20), Rif^R Cm^R Km^R	Lab collection
MT616	MM294A *recA-56* (pRK600), mobilizer; Cm^R	Finan et al. (1986)
MT620	MM294A *recA-56*; Rif^R	Lab collection
Plasmids		
pTH1931	Expression vector pTrcSC, derived from pTrcStrep; Sm^R Sp^R	diCenzo et al. (2013)
pTH2874	pLAFR1 (pSymA 125,586-147,721), complements RmP3100; Tc^R	This study
pTH2876	pLAFR1 (SMc 2,966,675-2,992,226), complements RmP3102; Tc^R	This study
pTH2877	pLAFR1 (pSymB 1,669,658-9,262), complements RmP3104; Tc^R	This study
pTH2878	pLAFR1 (SMc 3,049,275-3,070,202), complements RmP3106; Tc^R	This study
pTH2879	pLAFR1 (SMc 752,710-771,842), complements RmP3108; Tc^R	This study
pTH2880	pLAFR1 (pSymB 327,602-354,008), complements RmP3108; Tc^R	This study
pTH2881	pLAFR1 (SMc 2,843,881-2,869,916), complements RmP3110; Tc^R	This study
pTH2882	pLAFR1 (pSymB 717,582-745,468), complements RmP3100; Tc^R	This study
pTH2919	TcR *sacB* suicide vector, derived from pJQ200mp18; Tc^R	This study
pTH2978	pTH2919 (pSymA 133,433-132,874 - *Swa*I - 131,122-130,577); Tc^R	This study
pTH2979	pTH2919 (pSymB nt: 5,981-6,490 - *Swa*I - 7,263-7,819); Tc^R	This study
pTH2982	pTH2978 (*aacC4* via *Swa*I); Tc^R Gm^R	This study
pTH2983	pTH2979 (*aacC4* via *Swa*I); Tc^R Gm^R	This study
pTH2987	pTH1931 (*argH2*); Sm^R Sp^R	This study

Sm – streptomycin; Sp – spectinomycin; Nm – neomycin; Gm – gentamycin; Km – kanamycin; Tc – tetracycline; Cm – chloramphenicol; Rif – rifampicin

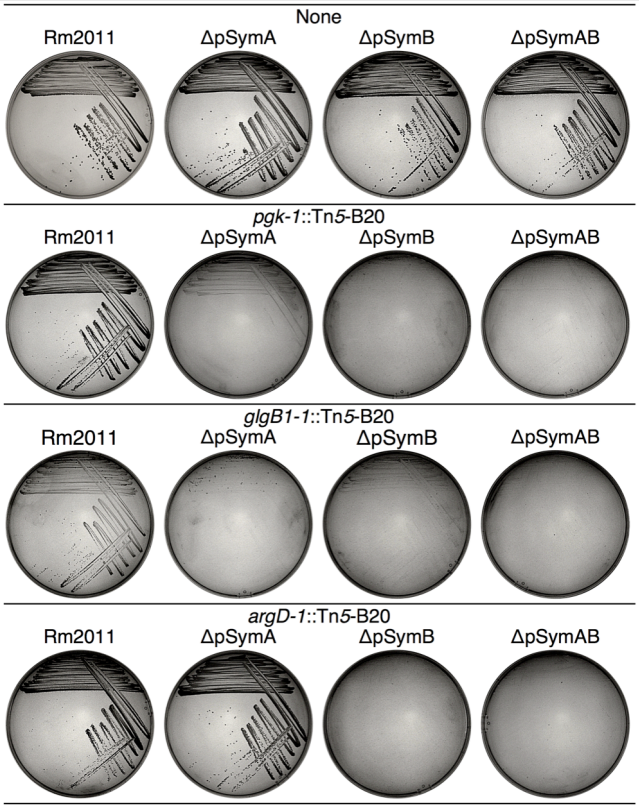
Table 5.S2. Oligonucleotides used in this study.

Name	Sequence
ML-1160	5'- CGC CAG GGT TTT CCC AGT CAC GAC GGT GTA
P128	5'- CCT CGA TCA GCT CTT GCA CTC G
P129	5'- GCA GGT GCT GGC ATC GAC AAT CAG C
DF001	5'- GGT ACT TGG GTC GAT GAG AGG CGG TTT GCG TAT TGG
DF002	5'- GGG CAG ATC CGT GCA CCG AAA AGT GCC ACC TGA CG
DF003	5'- GGT ACC CGG GGA TCC TCT AGA GGC ATC CGT CAC CTC GAC G
DF004	5'- TTG GGG **ATT TAA AT**C CCC AAC CCC ATT GTC AGT CAT GGC GGG ACC
DF005	5'- TTG GGG **ATT TAA AT**C CCC AAC CCC ACC CGC CAC AAT CAC TAG AGC
DF006	5'- TGC CTG CAG GTC GAC TCT AGT GCC GCC ATG CTC TTC AAG G
DF007	5'- GGT ACC CGG GGA TCC TCT AGT GGT CGA ACA GGT GGG CTC G
DF008	5'- TTG GGG **ATT TAA AT**C CCC AAC CCC AAC GCA TCG CTG TTC ACT CCC
DF009	5'- TTG GGG **ATT TAA AT**C CCC AAC CCC AGG TCT TCA AAG CTC GGT GCG
DF010	5'- TGC CTG CAG GTC GAC TCT AGC CGG CTA TGC AGG TAA AGA CC
DF011	5'- TGG GGT TGG GGA TTT GGT CAT TCA AAA GGT CAT CCA CCG
DF012	5'- CAG ACC ATG GCT TTA ATG ACC GAG CCC ACT CAG C
DF013	5'- CCT TCG ATT GCG TTA GAG GCT TGC CAC ATT CAT CG

Nucleotides that anneal to the template for PCR or sequencing are underlined, while restriction sites of interest are in boldface font

Figure 5.S1. Contributions of pSymA and pSymB to the phenotype of three Tn*5*-B20 insertion mutants on M9-sucrose agar plates.

The phenotypes of the *pgk-1*::Tn*5*-B20, *glgB1-1*::Tn*5*-B20, and *argD-1*::Tn*5*-B20 were examined in wild type *S. meliloti* Rm2011, ΔpSymA (SmA818), ΔpSymB (RmP3009), and ΔpSymAB (RmP2917). The phenotype of the *argD-1*::Tn*5*-B20 insertion mutant was dependent on the absence of pSymB, whereas the absence of either pSymA or pSymB exaggerated the phenotypes of the *pgk-1*::Tn*5*-B20 and *glgB1-1*::Tn*5*-B20 insertion mutations. *S. meliloti* Rm2011, ΔpSymA and the corresponding Tn*5*-B20 insertion mutants were incubated for four days, *S. meliloti* ΔpSymAB and the corresponding Tn*5*-B20 insertion mutants were incubated for five days, and *S. meliloti* ΔpSymB and the corresponding Tn*5*-B20 insertion mutants were incubated for six days


None
Rm2011
ΔpSymA
ΔpSymB
ΔpSymAB
pgk-1::Tn5-B20
Rm2011
ΔpSymA
ΔpSymB
ΔpSymAB
glgB1-1::Tn5-B20
Rm2011
ΔpSymA
ΔpSymB
ΔpSymAB
argD-1::Tn5-B20
Rm2011
ΔpSymA
ΔpSymB
ΔpSymAB

CHAPTER 6. DISCUSSION

6.1 Multipartite genome evolution

Bacterial genomes are highly structured, with their structure generally reflecting some functional purpose (Junier, 2014; Lawrence, 2003; Rocha, 2008). It is therefore reasonable to expect that the multipartite structure of some bacterial genomes plays an unique functional purpose and is not simply an evolutionary peculiarity. Based on the data presented in Chapters 1 through 5 of this thesis, I propose a generalizable model of multipartite genome evolution and function as summarized in Figure 6.1. In the following paragraphs, the rationale and support for this model is described.

The main tenet of the proposed model is that secondary replicons act as specialized replicons for adaptation to unique environments. Implicit in this argument is that the primary chromosome is sufficient for growth and survival in non-specialized soil and aquatic environments. This is supported by the experimental work presented in Chapter 2 that illustrated that the *S. meliloti* chromosome is sufficient for growth in a sterile soil environment, as well as by the *in silico* work of Chapter 3 that showed that the majority of metabolic functions required for bulk soil growth are chromosomally encoded. Moreover, computational analyses have suggested that the ancestor of the α-proteobacteria encoded ~ 3,300 genes, with lower and upper bounds of 3,000 and 5,000 proteins, respectively (Boussau *et al.*, 2004). Considering that on average one gene equals one kilobase, the median bacterial chromosome size of 3.46 Mb is likely similar to the size of the ancestral α-proteobacterial genome. Overall, these data are consistent with bacterial chromosomes being sufficient for high fitness in general soil or aquatic

environments.

If the chromosome is sufficient for growth in general soil or aquatic environments, presumably secondary replicons must provide a more specialized function. Accepting that the 'plasmid hypothesis' (Section 1.4.2) explain the initial formation of nearly all multipartite genomes, then the origin of a secondary replicon in a genome is through horizontal gene transfer (HGT). Hence, the function and evolutionary patterns of megaplasmids are expected to mimic those of individual genes acquired through HGT.

Comparative genomics and metabolic modeling studies illustrated that most genes acquired through HGT are primarily involved in adapting to a different environments (Lawrence & Ochman, 1998; Pál *et al.*, 2005), with different genomic regions responsible for the different ecologies (Niehus *et al.*, 2015). Other computational work has illustrated that genome expansion within lineages of the α-proteobacteria and the *Rhizobiales* order was linked to an association with plants and the evolution of symbiosis, respectively, and involved mainly the acquisition of transcriptional, transport, and metabolic functions that are commonly found of secondary replicons (Table 1.1; Boussau *et al.*, 2004; Pini *et al.*, 2011). The *in silico* work of Chapter 3 suggested that the metabolic functions encoded by pSymB are specialized for the rhizosphere, while the experimental work of Chapter 2 provided indirect evidence also consistent with pSymB dependent transport/metabolism being specialized. Additionally, further support is provided by the transcriptional studies that illustrated replicon biases in transcriptional responses to niche changes (see Section 1.9.4). Moreover, COG enrichment analyses have repeatedly demonstrated that

secondary replicons are enriched in functions that may assist in environmental adaptation, such as transport, metabolism, and regulation (Section 1.6.3 and Table 1.1). Together, these data support that secondary replicons generally play a role in environmental adaptation. In particular, I postulate that these environments more often than not represent new niches formed due to the emergence of eukaryotic organisms, such as the rhizosphere and the animal gut.

It has been argued that most HGT events, including those that are eventually fixed in a population, are initially deleterious and are often lost from the genome (Hao & Golding, 2006; Park & Zhang, 2012). The studies summarized in Section 1.8 and the growth data presented in Chapter 2 are consistent with megaplasmid acquisitions also at least initially being associated with fitness costs. Thus, megaplasmids must provide an immediate benefit to the host cell or risk being rapidly lost from the population. However, if no such advantage is provided by the megaplasmid, perhaps as the organism lacks access to the corresponding environment, the costs of the megaplasmid results in its loss form the population. Support for this 'do or die' nature is seen in the distribution and abundance of megaplasmids. As chromids are considered to have evolved from megaplasmids, chromids can be thought of as a sub-class of the megaplasmids and may be expected to be less abundant in the bacterial phylogeny. However, this is not the case, and there are ~ 2 times more species containing a chromid compared to a megaplasmid. Moreover, it is unusual to find large clusters of species with megaplasmids in the phylogeny, and megaplasmids are unlikely to be conserved across an entire genus, unlike

chromids (Section 1.5 and Figure 1.3). Additionally, while at least some megaplasmids retain conjugal properties, transfer to distant taxa may be limited (Romanchuk *et al.*, 2014). Taken together, it seems reasonable to argue that megaplasmids are more commonly lost than maintained following their transfer to a naïve cell, while those that are maintained rapidly transform into chromid replicons.

Genes acquired through recent HGT display elevated rates of evolution compared to those acquired in the more distant past (Hao & Golding, 2006). The data summarized in Sections 1.6.2 and 1.9.3 illustrate that megaplasmids display rapid evolution both in terms of gene gain/lost and the rate of modification of individual genes, whereas the evolution rate of chromid genes is closer to that of the chromosome. This rapid evolution of megaplasmids is expected to reflect the combined pressure of both reducing the costs associated with the replicon, as well as increasing the benefits it confers to the host cell. This includes modification of genes and regulatory elements, such as through amelioration of the codon usage (Lawrence & Ochman, 1997), and promoter modifications to better integrate the genes into existing transcriptional networks (Lercher & Pál, 2008). Moreover, megaplasmid gene content is highly variable even within the same species (Section 1.6.2). Most genes acquired through HGT are lost from the genome as the costs outweigh the benefits (Hao & Golding, 2006; 2010; Kurland *et al.*, 2003; Lawrence & Ochman, 1998; Park & Zhang, 2012). Thus, it is likely many of the genes brought into the genome as part of the megaplasmid are lost as they are deleterious, for example, due to negative genetic interactions, misfolding of the proteins due to the

differences in the cellular environment, or because they are highly expressed (Baltrus, 2013; Park & Zhang, 2012; San Millan *et al.*, 2015).

At the same time, as chromids are on average twice as large as megaplasmids, the evolution of a megaplasmid must therefore involve significant gene accumulation. Given the biases in COG distributions (Section 1.6.3) between chromids and megaplasmids, there must also be a functional bias between genes acquired by the megaplasmid once in a new cell, and the genes that were transferred as part of the replicon. Two sources of DNA for this megaplasmid enlargement are envisaged. The first is through HGT and the acquisition of novel genes that improve fitness of the cell in the novel environment that the megaplasmid allows the cell to inhabit. This source of gene gain is expected to confer improved colonization of the newly inhabited niche, and indeed HGT is often though to improve environmental adaptation (Lawrence & Ochman, 1998; Ochman *et al.*, 2000; Pál *et al.*, 2005; Wiedenbeck & Cohan, 2011). The second source of gene gain is through inter-replicon gene flow. As an extreme example, ~ 4% of the pSymB replicon of *Sinorhizobium meliloti* was derived through a single translocation event from the chromosome (Chapter 4) (diCenzo *et al.*, 2013). This transfer of genetic material from the chromosome is expected to transfer core genes to the replicon (now a chromid), stabilizing the replicon and forcing the replicon to be more integrated into core biology of the organism. Eventually, rate of evolution of the replicon decreases and the replicon stabilizes as it becomes optimized in relation to the specific host's chromosome and environment, and due to the gain of core functions.

These data that I have presented in this thesis and elsewhere (diCenzo *et al.*, 2013; 2014; 2016a; Fei *et al.*, 2016; Galardini *et al.*, 2013; 2015; Slater *et al.*, 2009; Wong & Golding, 2003) for *Sinorhizobium meliloti* provide both experimental and *in silico* support for the evolutionary history and functional roles of megaplasmids and chromids described in the previous few paragraphs. The growth data of Chapter 2, the flux balance analysis of Chapter 3, metabolomics (Fei *et al.*, 2016), and transcriptomics (diCenzo, Golding, and Finan, unpublished; Figure 6.2) illustrated that the pSymA megaplasmid is completely dispensable for, and poorly integrated into the cellular networks of, free-living *S. meliloti*. The deletion analysis of pSymA presented in Chapter 4 illustrates that the majority of pSymA is even dispensable for symbiotic nitrogen fixation, a niche inhabited by *S. meliloti* as a result of the presence of pSymA [I should note, however, that other regions of pSymA may contribute to symbiosis with other plant hosts, or play a role in competition for nodule occupancy (Pobigaylo *et al.*, 2008)]. In contrast, the data outlined above support that the pSymB chromid has a general role in supporting optimal fitness of *S. meliloti*, and also that pSymB is integrated into the metabolic and transcriptomic networks of the cell. Yet, the Phenotype MicroArray^{TM} and growth data of Chapter 2 and the flux balance analysis of Chapter 3 are consistent with the metabolic capabilities encoded by pSymB predominately being specialized for growth in a particular niche – the rhizosphere. The functions redundantly encoded by the chromosome and pSymB that were identified in Chapter 5 may also reflect a method of regulatory specialization, allowing each homolog to be differentially regulated in response to unique environmental

cues. Chapter 4 and earlier work (diCenzo *et al.*, 2013) details a clear example of how inter-replicon gene flow can transfer essential genes to a secondary replicon, forming a chromid. Additionally, the unpublished transcriptomic data illustrate that this translocation event contributed to the transcriptional integration of pSymB with the chromosome, and it is likely this genome rearrangement also contributed to their metabolic integration; for example, the elevated requirement for growth of *S. meliloti* strains lacking pSymB is possibly attributed, at least in part, to the loss of the *bacA* gene that is within the region translocated to pSymB (diCenzo, Zamani, and Finan, unpublished). Hence, together these data supports pSymA as a young megaplasmid that allows the cell to inhabit the nodule environment, but is perhaps not yet optimized for this niche, whereas pSymB is more developed and specialized for inhabiting the rhizosphere environment, but has also been integrated into the core biology of the cell through inter-replicon gene flow.

The preceding paragraphs describe a process through which niche adaptation drives the emergence of a multipartite genome. Namely, a newly acquired megaplasmid that allows the cell to inhabit a new environment, without being optimized for fitness in this niche, evolves into a large and stable replicon that is predominately specialized for growth in a specific environment, but is also generally required for cell growth. While the data supports this model, some questions remain unanswered. Below, I will highlight four unanswered questions, as well as my associated hypotheses.

The above model states that megaplasmids undergo a large size expansion in part

due to the gain of genes contributing to fitness in the cell's new niche. This implies that the majority of newly acquired DNA is incorporated into the megaplasmid instead of into the chromosome, although a reason for this bias was not addressed. I see two possible explanations, and likely a combination of the two explain the true situation. First, being newly acquired and under relaxed evolution, the megaplasmid may simply be more amenable to genome rearrangements and gene insertions. Secondly, DNA incorporated into the megaplasmid may persist in the population better than if the DNA were incorporated into the chromosome. Megaplasmids are transmissible through conjugation, meaning that a gene located can be more easily transferred through HGT than one on the chromosome. Thus, there may be selection for newly acquired genes to accumulate on the megaplasmid as it both facilitates the horizontal transfer of the gene, and may promote maintenance of the megaplasmid upon transfer to closely related cells, such as strains of the same species, by increasing the benefit to cost ratio.

On the topic of replicon conjugation, it is unclear why chromids, unlike megaplasmids, do not appear to transfer between species in nature despite retaining conjugal properties in lab settings (Banfalvi *et al*., 1985; Blanca-Ordóñez *et al.*, 2010; Harrison *et al.*, 2010). However, this may simply reflect the difference in gene content, and the costs outweighing the benefits of such a large and sudden genome expansion. Core 'information' genes, found on chromids but not megaplasmids, have previously been shown to be less likely to be successfully transferred via HGT (Jain *et al.*, 1999; Rivera *et al.*, 1998). Additionally, while a chromid is expected to contain many genes

contributing to niche adaptation, and thus theoretically has a high benefit to the cell, the genes and gene products may not be optimized for the recipient genome and thus have high costs; for example, due to genetic interactions, protein misfolding, or inappropriate expression (Baltrus, 2013; Park & Zhang, 2012; San Millan *et al.*, 2015). It has been argued that most genes acquired through HGT are maintained due to their low costs as opposed to their benefits (Park & Zhang, 2012). Thus, the higher costs associated with chromids may limit their transfer compared to smaller megaplasmids, which have lower benefits but also lower costs.

Earlier, I stated that the evolution of a chromid is associated with the transfer of essential genes from the chromosome. A clear example is the transfer of the ETR region to pSymB in a *S. meliloti* ancestor (Chapter 4; diCenzo *et al.*, 2013). However, it is unclear why such a translocation event would become fixed in the population. I hypothesize that the evolutionary driving force is the stabilization of the secondary replicon to ensure faithful inheritance of the replicon in both daughter cells; as the replicon carries essential genes, a daughter cell that fails to inherit the replicon would be non-viable. Many secondary replicons within the α-proteobacteria belong to the *repABC* family (Cevallos *et al.*, 2008). The *repAB* genes encode a partitioning system that helps ensure both daughter cells receive a copy of the replicon. Indeed, cloning of the *repAB* genes from the *S. meliloti* pSymA replicon into an unstable replicon stabilizes the plasmid, and 96% of the population contains the replicon following 38 cell divisions (MacLellan *et al.*, 2006). However, at this rate of loss, only 50% of the population would

contain the replicon following 475 generations, and only 10% would still have the replicon after 855 generations. Even at a mere 2 cell divisions per year, the replicon would be largely lost from the population within 500 years. Thus, cells in which essential genes have transferred to the secondary replicon may be selected for due to the stabilizing effect on the replicon that prevents the replicon from being lost from the population. It could also partially explain why chromids generally only contain a few essential genes as the first transfer of essential genes would stabilize the replicon, while additional essential gene transfers would provide little additional advantage.

Perhaps the biggest unanswered question relates to why the multipartite genome is maintained as opposed to integrating the genes of the secondary replicon into the chromosome. As summarized in Section 1.9, there are numerous potential benefits that may help maintain the multipartite genome structure once it has been formed. However, I hypothesize that often the multipartite genome structure is an evolutionary relic limited by what came before. As discussed before, megaplasmids are transferable. Hence, a gene on a megaplasmid is expected to have higher fitness than a gene on a chromosome due to the increased frequency of HGT mediated through megaplasmid conjugation. Therefore, early in the formation of a multipartite genome, evolutionary pressures maintain the multi-replicon structure by acting on individual genes to maximize HGT. However, chromids do not appear to be horizontally transferred, meaning that this evolutionary pressure is no longer present. However, at this point the chromid may simply be too large to integrate into the chromosome. The origin of replication and the

terminus region normally, but not always, split chromosomes into two roughly equal halves referred to as replichores (Song *et al.*, 2003). Genome rearrangements that perturb this equal distribution appear to negatively impact fitness and be selected against (Darling *et al.*, 2008; Hill & Gray, 1988; Lesterlin *et al.*, 2008; Liu *et al.*, 2006; Rocha, 2008), meaning that integration of a 1.5 Mb chromid into a bacterial chromosome is likely to be unfavourable. Hence, the maintenance of the multipartite genome architecture may reflect selection for increased HGT early in its development, and selective pressures against chromosome disruptions later during its maintenance.

6.2 Conclusions

Here, I described the development of experimental and *in silico* resources and tools for the study of both free-living and symbiotic *S. meliloti*, and how I have employed these resources to examine the biology of *S. meliloti*. These resources have the potential to advance studies within numerous areas of *S. meliloti* biology. This is particularly true in the study of symbiotic nitrogen fixation, the identification of the minimal symbiotic genome, and working towards engineering synthetic symbioses for agricultural gains.

In this thesis, I focused on how my studies of *S. meliloti* shed light on the evolution of the multipartite genome, and synthesize my data and the data of other groups to develop a comprehensive model for the evolution and function of this genome architecture. As detailed previously, this work is consistent with a model wherein the secondary replicons of a multipartite genome are specialized for colonization of a specific niche, which is often a eukaryotic host associated niche. Understanding the evolution of

this complex and unusual bacterial genome structure is more than simply interesting with respect to basic scientific curiosity; it also potentially has practical implications. The multipartite genome is present in many important plant symbionts as well as plant, animal, and human pathogens. Understanding these complex inter-kingdom relationships, and promoting or suppressing these relationships, are major goals of the scientific community. By providing evidence that the secondary replicons in these organisms are likely specialized for adaptation to the niche associated with their host, focusing efforts on these replicons may facilitate an improved understanding of these interactions and ways to manipulate them.

6.3 Figures

Figure 6.1. Model of multipartite genome evolution.

A schematic representation of the proposed model of multipartite genome evolution. Multipartite genome formation begins when a cell with a single replicon acquired a megaplasmid through horizontal gene transfer (HGT). The megaplasmid is maintained if it provides the cell a fitness benefit; this is hypothesized to be primarily by niche adaptation. The megaplasmid then undergoes rapid evolution, including sequence optimization, gene loss, and large amounts of gene gain through HGT, producing a large megaplasmid specialized for growth in a particular environment. At any point, the megaplasmid may be lost from the genome if the costs outweigh the benefits; for example, during growth in the cell's original niche. Eventually, gene movement from the chromosome to the megaplasmid results in the transfer of essential genes, and the formation of an evolutionarily stable chromid that cannot be lost from the genome.

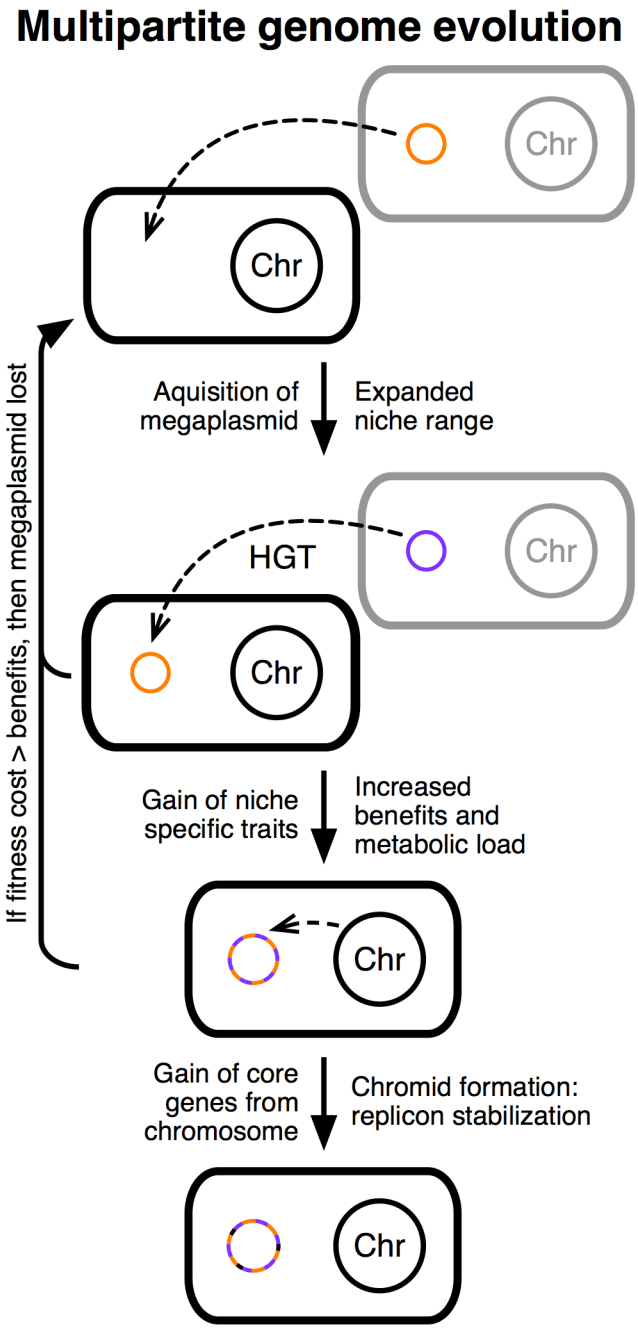
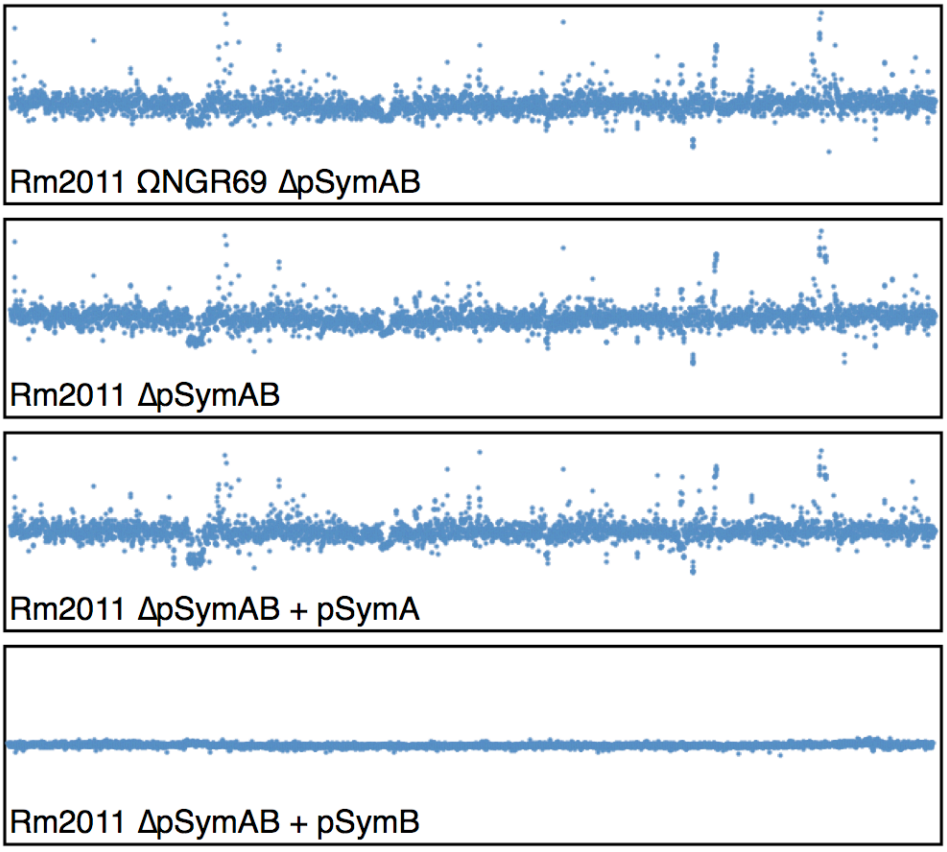

Multipartite genome evolution
Chr
Chr
Aquisition of megaplasmid
Expanded niche range
HGT
Chr
Chr
Gain of niche specific traits
Increased benefits and metabolic load
Chr
If fitness cost > benefits, then megaplasmid lost
Gain of core genes from chromosome
Chromid formation: replicon stabilization
Chr

Figure 6.2. Transcriptional consequences of removing replicons from the *S. meliloti* genome.

The following *S. meliloti* strains were grown in M9 minimal medium until mid-exponential phase: *S. meliloti* lacking both pSymA and pSymB (Rm2011 ΔpSymAB), *S. meliloti* lacking pSymA and pSymB, but with the entire *S. fredii* NGR234 ETR region integrated into the chromosome (Rm2011 ΩNGR69, ΔpSymAB), and Rm2011 ΔpSymAB derivatives in which either pSymA (Rm2011 ΔpSymAB + pSymA), pSymB (Rm2011 ΔpSymAB + pSymB), or both (Rm2011 ΔpSymAB + pSymAB) were re-introduced. Total RNA was isolated from biological duplicates of each strain and sequenced using Illumina HiSeq technology. This figures summarizes how the absence of a particular replicon influenced expression of all chromosomal genes relative to Rm2011 ΔpSymAB + pSymAB. Each dot represents an individual gene, and genes are ordered based on their position on the chromosome. Dots above the mid-line indicate genes up-regulated in the indicated strain relative to Rm2011 ΔpSymAB + pSymAB, whereas dots below the mid-line were down-regulated compared to Rm2011 ΔpSymAB + pSymAB. Expression changes are plotted on a $\log_2$ scale, from 2^{-6} at the bottom of the box to 2^6 at the top of the box. As can be seen, strains lacking pSymB showed major changes in the chromosomal transcriptome, whereas the absence of pSymA had little transcriptional effect.


Rm2011 ΩNGR69 ΔpSymAB
Rm2011 ΔpSymAB
Rm2011 ΔpSymAB + pSymA
Rm2011 ΔpSymAB + pSymB

CHAPTER 7. REFERENCES

Ampomah, O. Y., Avetisyan, A., Hansen, E., Svenson, J., Huser, T., Jensen, J. B. & Bhuvaneswari, T. V. (2013). The *thuEFGKAB* operon of rhizobia and *Agrobacterium tumefaciens* codes for transport of trehalose, maltitol, and isomers of sucrose and their assimilation through the formation of their 3-keto derivatives. *J Bacteriol* **195**, 3797–3807.

Ara, K., Ozaki, K., Nakamura, K., Yamane, K., Sekiguchi, J. & Ogasawara, N. (2007). *Bacillus* minimum genome factory: effective utilization of microbial genome information. *Biotechnol Appl Biochem* **46**, 169–178.

Archana, G. (2010). Engineering Nodulation Competitiveness of Rhizobial Bioinoculants in Soils. In *Microbes for Legume Improvement*, pp. 157–194. Vienna: Springer Vienna.

Ardissone, S., Noel, K. D., Klement, M., Broughton, W. J. & Deakin, W. J. (2011). Synthesis of the flavonoid-induced lipopolysaccharide of *Rhizobium* sp. strain NGR234 requires rhamnosyl transferases encoded by genes *rgpF* and *wbgA*. *Mol Plant Microbe Interact* **24**, 1513–1521.

Ardourel, M., Lortet, G., Maillet, F., Roche, P., Truchet, G., Promé, J. C. & Rosenberg, C. (1995). In *Rhizobium meliloti*, the operon associated with the nod box n5 comprises *nodL*, *noeA* and *noeB*, three host-range genes specifically required for the nodulation of particular *Medicago* species. *Mol Microbiol* **17**, 687–699.

Arias, A., Cerveñansky, C., Gardiol, A. & Martínez-Drets, G. (1979). Phosphoglucose isomerase mutant of *Rhizobium meliloti*. *J Bacteriol* **137**, 409–414.

Babu, M., Arnold, R., Bundalovic-Torma, C., Gagarinova, A., Wong, K. S., Kumar, A., Stewart, G., Samanfar, B., Aoki, H. & other authors. **(2014).** Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in *Escherichia coli*. *PLOS Genet* **10**, e1004120.

Baltrus, D. A. (2013). Exploring the costs of horizontal gene transfer. *Trends Ecol Evol* **28**, 489–495.

Baltrus, D. A., Nishimura, M. T., Romanchuk, A., Chang, J. H., Mukhtar, M. S., Cherkis, K., Roach, J., Grant, S. R., Jones, C. D. & Dangl, J. L. (2011). Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. *PLoS Pathog* **7**, e1002132.

Banergee, P. C., Darzins, A. & Maitra, P. K. (1987). Gluconeogenic mutations in *Pseudomonas aeruginosa*: genetic linkage between fructose-bisphosphate aldolase and phosphoglycerate kinase. *J Gen Microbiol* **133**, 1099–1107.

Banfalvi, Z., Kondorosi, E. & Kondorosi, A. (1985). *Rhizobium meliloti* has two megaplasmids. *Plasmid* **13**, 129-138.

Bardin, S., Dan, S., Osteras, M. & Finan, T. M. (1996). A phosphate transport system is required for symbiotic nitrogen fixation by *Rhizobium meliloti. J Bacteriol* **178**, 4540–4547.

Baril, C., Richaud, C., Baranton, G. & Saint Girons, I. (1989). Linear chromosome of *Borrelia burgdorferi*. *Res Microbiol* **140**, 507–516.

Barnett, M. J., Fisher, R. F., Jones, T., Komp, C., Abola, A. P., Barloy-Hubler, F.,

Bowser, L., Capela, D., Galibert, F. & other authors. **(2001).** Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid. *Proc Natl Acad Sci USA* **98**, 9883–9888.

Barnett, M. J., Toman, C. J., Fisher, R. F. & Long, S. R. (2004). A dual-genome Symbiosis Chip for coordinate study of signal exchange and development in a prokaryote–host interaction. *Proc Natl Acad Sci USA* **101**, 16636–16641.

Bartell, J. A., Yen, P., Varga, J. J., Goldberg, J. B. & Papin, J. A. (2014). Comparative metabolic systems analysis of pathogenic *Burkholderia*. *J Bacteriol* **196**, 210–226.

Basconcillo, L. S., Zaheer, R., Finan, T. M. & McCarry, B. E. (2009). A shotgun lipidomics study of a putative lysophosphatidic acid acyl transferase (PlsC) in *Sinorhizobium meliloti*. *J Chromatogr B* **877**, 2873–2882.

Batut, J., Terzaghi, B., Gherardi, M., Huguet, M., Terzaghi, E., Garnerone, A. M., Boistard, P. & Huguet, T. (1985). Localization of a symbiotic *fix* region on *Rhizobium meliloti* pSym megaplasmid more than 200 kilobases from the *nod-nif* region. *Mol Gen Genet* **199**, 232–239.

Bavishi, A., Abhishek, A., Lin, L. & Choudhary, M. (2010a). Complex prokaryotic genome structure: rapid evolution of chromosome II. *Genome* **53**, 675–687.

Bavishi, A., Lin, L., Schroeder, K., Peters, A., Cho, H. & Choudhary, M. (2010b). The prevalence of gene duplications and their ancient origin in *Rhodobacter sphaeroides* 2.4.1. *BMC Microbiol* **10**, 331.

Beatty, P. H. & Good, A. G. (2011). Future prospects for cereals that fix nitrogen. *Science* **333**, 416–417.

Becker, A., Bergès, H., Krol, E., Bruand, C., Rüberg, S., Capela, D., Lauber, E., Meilhoc, E., Ampe, F. & other authors. **(2004).** Global changes in gene expression in *Sinorhizobium meliloti* 1021 under microoxic and symbiotic conditions. *Mol Plant Microbe Interact* **17**, 292–303.

Belitsky, B. R., Brill, J., Bremer, E. & Sonenshein, A. L. (2001). Multiple genes for the last step of proline biosynthesis in *Bacillus subtilis*. *J Bacteriol* **183**, 4389–4392.

Bhattacharjee, R. B., Singh, A. & Mukhopadhyay, S. N. (2008). Use of nitrogen-fixing bacteria as biofertiliser for non-legumes: prospects and challenges. *Appl Microbiol Biotechnol* **80**, 199–209.

Biondi, E. G., Tatti, E., Comparini, D., Giuntini, E., Mocali, S., Giovannetti, L., Bazzicalupo, M., Mengoni, A. & Viti, C. (2009). Metabolic capacity of *Sinorhizobium* (*Ensifer*) *meliloti* strains as determined by Phenotype MicroArray analysis. *Appl Environ Microbiol* **75**, 5396–5404.

Blanca-Ordóñez, H., Oliva-García, J. J., Pérez-Mendoza, D., Soto, M. J., Olivares, J., Sanjuán, J. & Nogales, J. (2010). pSymA-dependent mobilization of the *Sinorhizobium meliloti* pSymB megaplasmid. *J Bacteriol* **192**, 6309–6312.

Blondelet-Rouault, M. H., Weiser, J., Lebrihi, A., Branny, P. & Pernodet, J. L. (1997). Antibiotic resistance gene cassettes derived from the omega interposon for use in *E. coli* and *Streptomyces*. *Gene* **190**, 315–317.

Bohlool, B. B., Ladha, J. K., Garrity, D. P. & George, T. (1992). Biological nitrogen fixation for sustainable agriculture: A perspective. In *Biological nitrogen fixation for sustainable agriculture*, pp. 1–11. Dordrecht: Springer Netherlands.

Boivin, C., Barran, L. R., Malpica, C. A. & Rosenberg, C. (1991). Genetic analysis of a region of the *Rhizobium meliloti* pSym plasmid specifying catabolism of trigonelline, a secondary metabolie present in legumes. *J Bacteriol* **173**, 2809–2817.

Boncompagni, E., Dupont, L., Mignot, T., Osteras, M., Lambert, A., Poggi, M. C. & Le Rudulier, D. (2000). Characterization of a *Snorhizobium meliloti* ATP-binding cassette histidine transporter also involved in betaine and proline uptake. *J Bacteriol* **182**, 3717–3725.

Bonhoeffer, F. & Messer, W. (1969). Replication of the bacterial chromosome. *Annu Rev Genet* **3**, 233–246.

Borisova, S. A., Christman, H. D., Metcalf, M. E. M., Zulkepli, N. A., Zhang, J. K., van der Donk, W. A. & Metcalf, W. W. (2011). Genetic and biochemical characterization of a pathway for the degradation of 2-aminoethylphosphonate in *Sinorhizobium meliloti* 1021. *J Biol Chem* **286**, 22283–22290.

Boulter, D., Jeremy, J. J. & Wilding, M. (1966). Amino acids liberated into the culture medium by pea seedling roots. *Plant Soil* **24**, 121–127.

Boussau, B., Karlberg, E. O., Frank, A. C., Legault, B.-A. & Andersson, S. G. E. (2004). Computational inference of scenarios for α-proteobacterial genome evolution. *Proc Natl Acad Sci USA* **101**, 9722–9727.

Bremer, H. (1975). Parameters affecting the rate of synthesis of ribosomes and RNA polynerase in bacteria. *J Theor Biol* **53**, 115–124.

Brewin, N. J. & Legocki, A. B. (1996). Biological nitrogen fixation for sustainable agriculture. *Trends Microbiol* **4**, 476–477.

Bringhurst, R. M., Cardon, Z. G. & Gage, D. J. (2001). Galactosides in the rhizosphere: Utilization by *Sinorhizobium meliloti* and development of a biosensor. *Proc Natl Acad Sci USA* **98**, 4540–4548.

Brom, S., García-de los Santos, A., Cervantes, L., Palacios, R. & Romero, D. (2000). In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons. *Plasmid* **44**, 34–43.

Brom, S., García-de los Santos, A., Stepkowsky, T., Flores, M., Davila, G., Romero, D. & Palacios, R. (1992). Different plasmids of *Rhizobium leguminosarum* bv. *phaseoli* are required for optimal symbiotic performance. *J Bacteriol* **174**, 5183–5189.

Brom, S., Girard, L., García-de-los-Santos, A., Sanjuan-Pinilla, J. M., Olivares, J. & Sanjuán, J. (2002). Conservation of plasmid-encoded traits among bean-nodulating *Rhizobium* species. *Appl Environ Microbiol* **68**, 2555–2561.

Bromfield, E. S. P., Barran, L. R. & 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*). *Mol Ecol* **4**. 183-188.

Bryant, J. A., Sellars, L. E., Busby, S. J. W. & Lee, D. J. (2015). Chromosome position

effects on gene expression in *Escherichia coli* K-12. *Nucleic Acids Res* **42**, 11383–11392.

Butland, G., Babu, M., Díaz-Mejía, J. J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A. G. & other authors. **(2008).** eSGA: *E. coli* synthetic genetic array analysis. *Nat Methods* **5**, 789–795.

Cairns, J. (1963). The bacterial chromosome and its manner of replication as seen by autoradiography. *J Mol Biol* **6**, 208–IN5. Academic Press Inc. (London) Ltd.

Cameron, D. E., Urbach, J. M. & Mekalanos, J. J. (2008). A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proc Natl Acad Sci USA* **105**, 8736–8741.

Capela, D., Barloy-Hubler, F., Gouzy, J., Bothe, G., Ampe, F., Batut, J., Boistard, P., Becker, A., Boutry, M. & other authors. **(2001).** Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc Natl Acad Sci USA* **98**, 9877–9882.

Capela, D., Filipe, C., Bobik, C., Batut, J. & Bruand, C. (2006). *Sinorhizobium meliloti* differentiation during symbiosis with alfalfa: a transcriptomic dissection. *Mol Plant Microbe Interact* **19**, 363–372.

Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.

Carbone, A., Zinovyev, A. & Képès, F. (2003). Codon adaptation index as a measure of

dominating codon bias. *Bioinformatics* **19**, 2005–2015.

Carlson, R. W., Price, N. P. & Stacey, G. (1994). The biosynthesis of rhizobial lipo-oligosaccharide nodulation signal molecules. *Mol Plant Microbe Interact* **7**, 684–695.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G. & Parkhill, J. (2005). ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422–3423.

Casjens, S. (1999). Evolution of the linear DNA replicons of the *Borrelia* spirochetes. *Curr Opin Microbiol* **2**, 529–534.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A. & other authors. **(2014).** The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* **42**, D459–71.

Castillo-Ramírez, S., Vázquez-Castellanos, J. F., González, V. & Cevallos, M. A. (2009). Horizontal gene transfer and diverse functional constrains within a common replication-partitioning system in *Alphaproteobacteria*: the *repABC* operon. *BMC Genomics* **10**, 536.

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

Chain, P. S. G., Denef, V. J., Konstantinidis, K. T., Vergez, L. M., Agulló, L., Reyes, V. L., Hauser, L., Córdova, M., Gómez, L. & other authors. **(2006).** *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility.

Proc Natl Acad Sci USA **103**, 15280–15287.

Chang, P. C., Kim, E. S. & Cohen, S. N. (1996). *Streptomyces* linear plasmids that contain a phage-like, centrally located, replication origin. *Mol Microbiol* **22**, 789–800.

Charles, T. C. & Aneja, P. (1999). Methylmalonyl-CoA mutase encoding gene of *Sinorhizobium meliloti*. *Gene* **226**, 121–127.

Charles, T. C., Cai, G. Q. & Aneja, P. (1997). Megaplasmid and chromosomal loci for the PHB degradation pathway in *Rhizobium (Sinorhizobium) meliloti*. *Genetics* **146**, 1211–1220.

Charles, T. C. & Finan, T. M. (1991). Analysis of a 1600-kilobase *Rhizobium meliloti* megaplasmid using defined deletions generated *in vivo*. *Genetics* **127**, 5–20.

Charpentier, M. & Oldroyd, G. (2010). How close are we to nitrogen-fixing cereals? *Curr Opin Plant Biol* **13**, 556–564.

Chen, A.-M., Wang, Y.-B., Jie, S., Yu, A.-Y., Luo, L., Yu, G.-Q., Zhu, J.-B. & Wang, Y.-Z. (2010). Identification of a TRAP transporter for malonate transport and its expression regulated by GtrA from *Sinorhizobium meliloti*. *Res Microbiol* **161**, 556–564.

Chen, C. W., Huang, C.-H., Lee, H.-H., Tsai, H.-H. & Kirby, R. (2002). Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet* **18**, 522–529.

Cheng, J., Poduska, B., Morton, R. A. & Finan, T. M. (2011). An ABC-type cobalt

transport system is essential for growth of *Sinorhizobium meliloti* at trace metal concentrations. *J Bacteriol* **193**, 4405–4416.

Cheng, J., Sibley, C. D., Zaheer, R. & Finan, T. M. (2007). A *Sinorhizobium meliloti minE* mutant has an altered morphology and exhibits defects in legume symbiosis. *Microbiology* **153**, 375–387.

Choudhary, M., Mackenzie, C., Nereng, K. S., Sodergren, E., Weinstock, G. M. & Kaplan, S. (1994). Multiple chromosomes in bacteria: structure and function of chromosome II of *Rhodobacter sphaeroides* 2.4.1T. *J Bacteriol* **176**, 7694–7702.

Choudhary, M., Mackenzie, C., Nereng, K., Sodergren, E., Weinstock, G. M. & Kaplan, S. (1997). Low-resolution sequencing of *Rhodobacter sphaeroides* 2.A.1T: chromosome II is a true chromosome. *Microbiology* **143**, 3085–3099.

Choudhary, M., Zanhua, X., Fu, Y. X. & Kaplan, S. (2007). Genome analyses of three strains of *Rhodobacter sphaeroides*: evidence of rapid evolution of chromosome II. *J Bacteriol* **189**, 1914–1921.

Choudhary, M., Cho, H., Bavishi, A., Trahan, C. & Myagmarjav, B.-E. (2012). Evolution of multipartite genomes in prokaryotes. In *Evolutionary Biology: Mechanisms and Trends*, pp. 301–323. Berlin, Heidelberg: Springer Berlin Heidelberg.

Choudhary, M., Fu, Y.-X., Mackenzie, C. & Kaplan, S. (2004). DNA sequence duplication in *Rhodobacter sphaeroides* 2.4.1: evidence of an ancient partnership between chromosomes I and II. *J Bacteriol* **186**, 2019–2027.

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Ben Passarelli, Coller, J. A., Fero, M. J., McAdams, H. H. & Shapiro, L. (2011). The essential genome of a bacterium. *Mol Syst Biol* **7**, 528.

Condon, C., Liveris, D., Squires, C., Schwartz, I. & Squires, C. L. (1995). rRNA operon multiplicity in *Escherichia coli* and the physiological implications of *rrn* inactivation. *J Bacteriol* **177**, 4152–4156.

Cooper, V. S., Vohr, S. H., Wrocklage, S. C. & Hatcher, P. J. (2010). Why genes evolve faster on secondary chromosomes in bacteria. *PLOS Comput Biol* **6**, e1000732.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K. & other authors. **(2010).** The genetic landscape of a cell. *Science* **327**, 425–431.

Couturier, E. & Rocha, E. P. C. (2006). Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol* **59**, 1506–1518.

Cowie, A., Cheng, J., Sibley, C. D., Fong, Y., Zaheer, R., Patten, C. L., Morton, R. M., Golding, G. B. & Finan, T. M. (2006). An integrated approach to functional genomics: construction of a novel reporter gene fusion library for *Sinorhizobium meliloti*. *Appl Environ Microbiol* **72**, 7156–7167.

Cui, T., oka, N. M., Ohsumi, K., Kodama, K., Ohshima, T., Ogasawara, N., Mori, H., Wanner, B., Niki, H. & Horiuchi, T. (2007). *Escherichia coli* with a linear

genome. *EMBO Rep* **8**, 181–187.

Curatti, L. & Rubio, L. M. (2014). Challenges to develop nitrogen-fixing cereals by direct *nif*-gene transfer. *Plant Sci* **225**, 130–137.

Cutler, S. & McCourt, P. (2005). Dude, where's my phenotype? Dealing with redundancy in signaling networks. *Plant Physiol* **138**, 558–559.

Darling, A. E., Mau, B. & Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5**, e11147.

Darling, A. E., Miklós, I. & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLOS Genet* **4**, e1000128.

de Rudder, K. E. E., Sohlenkamp, C. & Geiger, O. (1999). Plant-exuded choline is used for rhizobial membrane lipid biosynthesis by phosphatidylcholine synthase. *J Biol Chem* **274**, 20011–20016.

Degefu, T., Wolde-meskel, E. & Frostegard, A. (2012). Phylogenetic multilocus sequence analysis identifies seven novel *Ensifer* genospecies isolated from a less-well-explored biogeographical region in East Africa. *Int J Syst Evol Microbiol* **62**, 2286–2295.

diCenzo, G. C., Checcucci, A., Bazzicalupo, M., Mengoni, A., Viti, C., Dziewit, L., Finan, T. M., Galardini, M. & Fondi, M. (2016a). Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. *Nat Commun* **7**, 12219.

diCenzo, G. C. & Finan, T. M. (2015). Genetic redundancy is prevalent within the

6.7 Mb *Sinorhizobium meliloti* genome. *Mol Genet Genomics* **290**, 1345–1356.

diCenzo, G. C., MacLean, A. M., Milunovic, B., Golding, G. B. & Finan, T. M. (2014). Examination of prokaryotic multipartite genome evolution through experimental genome reduction. *PLOS Genet* **10**, e1004742.

diCenzo, G. C., Zamani, M., Cowie, A. & Finan, T. M. (2015). Proline auxotrophy in *Sinorhizobium meliloti* results in a plant-specific symbiotic phenotype. *Microbiology* **161**, 2341–2351.

diCenzo, G. C., Zamani, M., Milunovic, B. & Finan, T. M. (2016b). Genomic resources for identification of the minimal N_2-fixing symbiotic genome. *Environ Microbiol* **18**, 2534–2547.

diCenzo, G., Milunovic, B., Cheng, J. & Finan, T. M. (2013). The $tRNA^{arg}$ gene and *engA* are essential genes on the 1.7-mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. *J Bacteriol* **195**, 202–212.

Ding, H., Yip, C. B., Geddes, B. A., Oresnik, I. J. & Hynes, M. F. (2012). Glycerol utilization by *Rhizobium leguminosarum* requires an ABC transporter and affects competition for nodulation. *Microbiology* **158**, 1369–1378.

Djordjevic, M. A. (2004). *Sinorhizobium meliloti* metabolism in the root nodule: A proteomic perspective. *Proteomics* **4**, 1859–1872.

Dominguez-Ferreras, A., Muñoz, S., Olivares, J., Soto, M. J. & Sanjuán, J. (2009a). Role of potassium uptake systems in *Sinorhizobium meliloti* osmoadaptation and

symbiotic performance. *J Bacteriol* **191**, 2133–2143.

Dominguez-Ferreras, A., Soto, M. J., Pérez-Arnedo, R., Olivares, J. & Sanjuán, J. (2009b). Importance of trehalose biosynthesis for *Sinorhizobium meliloti* osmotolerance and nodulation of alfalfa roots. *J Bacteriol* **191**, 7490–7499.

Dorken, G., Ferguson, G. P., French, C. E. & Poon, W. C. K. (2012). Aggregation by depletion attraction in cultures of bacteria producing exopolysaccharide. *J R Soc Interface* **9**, 3490–3502.

Driscoll, B. T. & Finan, T. M. (1996). $NADP^+$-dependent malic enzyme of *Rhizobium meliloti*. *J Bacteriol* **178**, 2224–2231.

Dryselius, R., Izutsu, K., Honda, T. & Iida, T. (2008). Differential replication dynamics for large and small *Vibrio* chromosomes affect gene dosage, expression and location. *BMC Genomics* **9**, 559.

Dunn, M. F. (2015). Key roles of microsymbiont amino acid metabolism in rhizobia-legume interactions. *Critical Reviews in Microbiology* **41**, 411–451.

Dupont, L., Garcia, I., Poggi, M. C., Alloing, G., Mandon, K. & Le Rudulier, D. (2004). The *Sinorhizobium meliloti* ABC transporter Cho is highly specific for choline and expressed in bacteroids from *Medicago sativa* nodules. *J Bacteriol* **186**, 5988–5996.

Dziewit, L., Czarnecki, J., Wibberg, D., Radlinska, M., Mrozek, P., Szymczak, M., Schlüter, A., Pühler, A. & Bartosik, D. (2014). Architecture and functions of a multipartite genome of the methylotrophic bacterium *Paracoccus aminophilus* JCM

7686, containing primary and secondary chromids. *BMC Genomics* **15**, 1–16.

Eberhard, W. G. (1990). Evolution in bacterial plasmids and levels of selection. *Q Rev Biol* **65**, 3–22.

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

Egan, E. S., Fogel, M. A. & Waldor, M. K. (2005). Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. *Mol Microbiol* **56**, 1129–1138.

Elliot, M. A., Karoonuthaisiri, N., Huang, J., Bibb, M. J., Cohen, S. N., Kao, C. M. & Buttner, M. J. (2003). The chaplins: a family of hydrophobic cell-surface proteins involved in aerial mycelium formation in *Streptomyces coelicolor*. *Genes Dev* **17**, 1727–1740.

Epstein, B., Branca, A., Mudge, J., Bharti, A. K., Briskine, R., Farmer, A. D., Sugawara, M., Young, N. D., Sadowsky, M. J. & Tiffin, P. (2012). Population genomics of the facultatively mutualistic bacteria *Sinorhizobium meliloti* and *S. medicae*. *PLOS Genet* **8**, e1002868.

Fei, F., diCenzo, G. C., Bowdish, D. M. E., McCarry, B. E. & Finan, T. M. (2016). Effects of synthetic large-scale genome reduction on metabolism and metabolic preferences in a nutritionally complex environment. *Metabolomics* **12**, 23.

Ferdows, M. S. & Barbour, A. G. (1989). Megabase-sized linear DNA in the bacterium *Borrelia burgdorferi*, the Lyme disease agent. *Proc Indian Natn Sci Acad* **86**, 5969–5973.

Finan, T. M., Hartweig, E., LeMieux, K., Bergman, K., Walker, G. C. & Signer, E. R. (1984). General transduction in *Rhizobium meliloti*. *J Bacteriol* **159**, 120–124.

Finan, T. M., Kunkel, B., De Vos, G. F. & Signer, E. R. (1986). Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. *J Bacteriol* **167**, 66–72.

Finan, T. M., Oresnik, I. J. & Bottacin, A. (1988). Mutants of *Rhizobium meliloti* defective in succinate metabolism. *J Bacteriol* **170**, 3396–3403.

Finan, T. M., Weidner, S., Wong, K., Buhrmester, J., Chain, P., Vorhölter, F. J., Hernández-Lucas, I., Becker, A., Cowie, A. & other authors. **(2001).** The complete sequence of the 1,683-kb pSymB megaplasmid from the N_2-fixing endosymbiont *Sinorhizobium meliloti*. *Proc Natl Acad Sci USA* **98**, 9889–9894.

Fondi, M., Maida, I., Perrin, E., Mellera, A., Mocali, S., Parrilli, E., Tutino, M. L., Liò, P. & Fani, R. (2014). Genome-scale metabolic reconstruction and constraint-based modelling of the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125. *Environ Microbiol* **17**, 751–766.

Frage, B., Döhlemann, J., Robledo, M., Lucena, D., Sobetzko, P., Graumann, P. L. & Becker, A. (2016). Spatiotemporal choreography of chromosome and megaplasmids in the *Sinorhizobium meliloti* cell cycle. *Mol Microbiol* **100**, 808–823.

Frank, O., Göker, M., Pradella, S. & Petersen, J. (2015). Ocean's twelve: Flagellar and biofilm chromids in the multipartite genome of *Marinovum algicola* DG898 exemplify functional compartmentalization. *Environ Microbiol* **17**, 4019–4034.

Friedman, A. M., Long, S. R., Brown, S. E., Buikema, W. J. & Ausubel, F. M. (1982). Construction of a broad host range cosmid cloning vector and its use in the genetic analysis of *Rhizobium* mutants. *Gene* **18**, 289–296.

Fuhrer, T., Fischer, E. & Sauer, U. (2005). Experimental identification and quantification of glucose metabolism in seven bacterial species. *J Bacteriol* **187**, 1581–1590.

Gage, D. J. & Long, S. R. (1998). α-Galactoside uptake in *Rhizobium meliloti*: isolation and characterization of *agpA*, a gene encoding a periplasmic binding protein required for melibiose and raffinose utilization. *J Bacteriol* **180**, 5739–5748.

Gage, D. J. (2004). Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes. *Microbiol Mol Biol Rev* **68**, 280–300.

Galardini, M., Pini, F., Bazzicalupo, M., Biondi, E. G. & Mengoni, A. (2013). Replicon-dependent bacterial genome evolution: the case of *Sinorhizobium meliloti*. *Genome Biol Evol* **5**, 542–558.

Galardini, M., Brilli, M., Spini, G., Rossi, M., Roncaglia, B., Bani, A., Chiancianesi, M., Moretto, M., Engelen, K. & other authors. **(2015).** Evolution of intra-specific regulatory networks in a multipartite bacterial genome. *PLOS Comput Biol* **11**, e1004478.

Galardini, M., Mengoni, A., Biondi, E. G., Semeraro, R., Florio, A., Bazzicalupo, M., Benedetti, A. & Mocali, S. (2014). DuctApe: a suite for the analysis and correlation of genomic and OmniLog™ Phenotype Microarray data. *Genomics* **103**, 1–10.

Galardini, M., Mengoni, A., Brilli, M., Pini, F., Fioravanti, A., Lucas, S., Lapidus, A., Cheng, J.-F., Goodwin, L. & other authors. **(2011).** Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. *BMC Genomics* **12**, 235.

Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, A. P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A. & other authors. **(2001).** The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**, 668–672.

Galli, E., Poidevin, M., Le Bars, R., Desfontains, J-M., Muresan, L., Paly, E., Yamaichi, Y., & Barre, F-X. (2016). Cell division licensing in the multi-chromosomal *Vibrio cholerae* bacterium. *Nat Microbiol* **1**, 16094.

Galperin, M. Y. (2007). Linear chromosomes in bacteria: no straight edge advantage? *Environ Microbiol* **9**, 1357–1362.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* **43**, D261–9.

Gao, J.-L., Weissenmayer, B., Taylor, A. M., Thomas-Oates, J., López-Lara, I. M. & Geiger, O. (2004). Identification of a gene required for the formation of lyso-ornithine lipid, an intermediate in the biosynthesis of ornithine-containing lipids. *Mol Microbiol* **53**, 1757–1770.

Garcia-Fraile, P., Seaman, J. C., Karunakaran, R., Edwards, A., Poole, P. S. & Downie, J. A. (2015). Arabinose and protocatechuate catabolism genes are important

for growth of *Rhizobium leguminosarum* biovar *viciae* in the pea rhizosphere. *Plant Soil* **390**, 251–264.

García-de los Santos, A. & Brom, S. (1997). Characterization of two plasmid-borne *lps*β loci of *Rhizobium etli* required for lipopolysaccharide synthesis and for optimal interaction with plants. *Mol Plant Microbe Interact* **10**, 891–902.

Garg, N. & Geetanjali. **(2007).** Symbiotic nitrogen fixation in legume nodules: process and signaling. A review. *Agron Sustain Dev* **27**, 59–68.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784–3788.

Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V. & Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci USA* **106**, 16422–16427.

Geddes, B. A. & Oresnik, I. J. (2012a). Inability to catabolize galactose leads to increased ability to compete for nodule occupancy in *Sinorhizobium meliloti*. *J Bacteriol* **194**, 5044–5053.

Geddes, B. A. & Oresnik, I. J. (2012b). Genetic characterization of a complex locus necessary for the transport and catabolism of erythritol, adonitol and L-arabitol in *Sinorhizobium meliloti*. *Microbiology* **158**, 2180–2191.

Geddes, B. A. & Oresnik, I. J. (2014). Physiology, genetics, and biochemistry of carbon

metabolism in the alphaproteobacterium *Sinorhizobium meliloti*. *Can J Microbiol* **60**, 491–507.

Geddes, B. A., Pickering, B. S., Poysti, N. J., Collins, H., Yudistira, H. & Oresnik, I. J. (2010). A locus necessary for the transport and catabolism of erythritol in *Sinorhizobium meliloti*. *Microbiology* **156**, 2970–2981.

Geddes, B. A., Ryu, M.-H., Mus, F., Garcia Costas, A., Peters, J. W., Voigt, C. A. & Poole, P. (2015). Use of plant colonizing bacteria as chassis for transfer of N_2-fixation to cereals. *Curr Opin Biotech* **32**, 216–222.

Geiger, O. & López-Lara, I. M. (2002). Rhizobial acyl carrier proteins and their roles in the formation of bacterial cell-surface components that are required for the development of nitrogen-fixing root nodules on legume hosts. *FEMS Microbio Lett* **208**, 153–162.

Gemperline, E., Jayaraman, D., Maeda, J., Ané, J.-M. & Li, L. (2015). Multifaceted investigation of metabolites during nitrogen fixation in *Medicago* via high resolution MALDI-MS imaging and ESI-MS. *J Am Soc Mass Spectrom* **26**, 149–158.

Gerdes, K., Christensen, S. K. & Løbner-Olesen, A. (2005). Prokaryotic toxin-antitoxin stress response loci. *Nature Rev Microbiol* **3**, 371–382.

Ghim, C. M., Goh, K. I. & Kahng, B. (2005). Lethality and synthetic lethality in the genome-wide metabolic network of *Escherichia coli*. *J Theor Biol* **237**, 401–411.

Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M. R., Maruf, M., Hutchison, C. A., III, Smith, H. O. & Venter, J. C. (2005). Essential genes of a

minimal bacterium. *Proc Natl Acad Sci USA* **103**, 425–430.

Glenn, S. A., Gurich, N., Feeney, M. A. & González, J. E. (2007). The ExpR/Sin quorum-sensing system controls succinoglycan production in *Sinorhizobium meliloti*. *J Bacteriol* **189**, 7077–7088.

Goldmann, A., Lecoeur, L., Message, B., Delarue, M., Schoonejans, E. & Tepfer, D. (1994). Symbiotic plasmid genes essential to the catabolism of proline betaine, or stachydrine, are also required for efficient nodulation by *Rhizobium meliloti*. *FEMS Microbio Lett* **115**, 305–312.

González, V., Santamaría, R. I., Bustos, P., Hernández-González, I., Medrano-Soto, A., Moreno-Hagelsieb, G., Janga, S. C., Ramírez, M. A., Jiménez-Jacinto, V. & other authors. **(2006).** The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci USA* **103**, 3834–3839.

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M. & other authors. **(2001).** Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**, 2323–2328.

Gu, X., Lee, S. G. & Bar-Peled, M. (2010). Biosynthesis of UDP-xylose and UDP-arabinose in *Sinorhizobium meliloti* 1021: first characterization of a bacterial UDP-xylose synthase, and UDP-xylose 4-epimerase. *Microbiology* **157**, 260–269.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. & Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66.

Guo, H., Sun, S., Eardly, B., Finan, T. & Xu, J. (2009). Genome variation in the symbiotic nitrogen-fixing bacterium *Sinorhizobium meliloti*. *Genome* **52**, 862–875.

Guo, H. J., Wang, E. T., Zhang, X. X., Li, Q. Q., Zhang, Y. M., Tian, C. F. & Chen, W. X. (2014). Replicon-dependent differentiation of symbiosis-related genes in *Sinorhizobium* strains nodulating *Glycine max*. *Appl Environ Microbiol* **80**, 1245–1255.

Guo, X., Flores, M., Mavingui, P., Fuentes, S. I., Hernández, G., Dávila, G. & Palacios, R. (2003). Natural genomic design in *Sinorhizobium meliloti*: Novel genomic architectures. *Genome Res* **13**, 1810–1817.

Hao, W. & Golding, G. B. (2006). The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res* **16**, 636–643.

Hao, W. & Golding, G. B. (2010). Inferring bacterial genome flux while considering truncated genes. *Genetics* **186**, 411–426.

Harrison, J., Jamet, A., Muglia, C. I., Van de Sype, G., Aguilar, O. M., Puppo, A. & Frendo, P. (2005). Glutathione plays a fundamental role in growth and symbiotic capacity of *Sinorhizobium meliloti*. *J Bacteriol* **187**, 168–174.

Harrison, P. W., Lower, R. P. J., Kim, N. K. D. & Young, J. P. W. (2010). Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol* **18**, 141–148.

Hartmann, A., Girard, J. J. & Catroux, G. (1998). Genotypic diversity of *Sinorhizobium* (formerly *Rhizobium*) *meliloti* strains isolated directly from a soil and

from nodules of alfalfa (*Medicago sativa*) grown in the same soil. *FEMS Microbiol Ecol* **25**, 107-116.

Hayakawa, T., Tanaka, T., Sakaguchi, K., Ōtake, N. & Yonehara, H. (1979). A linear plasmid-like DNA in *Streptomyces* sp. producing lankacidin group antibiotics. *J Gen Appl Microbiol* **25**, 255–260.

Hayes, F. (2003). Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. *Science* **301**, 1496–1499.

He, X., Chang, W., Pierce, D. L., Seib, L. O., Wagner, J. & Fuqua, C. (2003). Quorum sensing in *Rhizobium* sp. strain NGR234 regulates conjugal transfer (*tra*) gene expression and influences growth rate. *J Bacteriol* **185**, 809–822.

Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D. & other authors. **(2000).** DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483.

Henzell, E. F. (1988). The role of biological nitrogen fixation research in solving problems in tropical agriculture. *Plant Soil* **108**, 15–21.

Herrera-Cervera, J. A., Caballero-Mellado, J., Laguerre, G., Tichy, H.-V., Requena, N., Amarger, N., Martínez-Romero, E., Olivares, J. & Sanjuán, J. (1999). At least five rhizobial species nodulate *Phaseolus vulgaris* in a Spanish soil. *FEMS Microbiol Ecol* **30**, 87–97.

Herridge, D. F., Peoples, M. B. & Boddey, R. M. (2008). Global inputs of biological

nitrogen fixation in agricultural systems. *Plant Soil* **311**, 1–18.

Hertenberger, G., Zampach, P. & Bachmann, G. (2002). Plant species affect the concentration of free sugars and free amino acids in different types of soil. *J Plant Nutr Soil Sci* **165**, 557–565.

Hessen, D. O., Jeyasingh, P. D., Neiman, M. & Weider, L. J. (2010). Genome streamlining and the elemental costs of growth. *Trends Ecol Evol* **25**, 75–80. Elsevier.

Hill, C. W. & Gray, J. A. (1988). Effects of chromosomal inversion on cell fitness in *Escherichia coli* K-12. *Genetics* **119**, 771–778.

Hinnebusch, J. & Tilly, K. (1993). Linear plasmids and chromosomes in bacteria. *Mol Microbiol* **10**, 917–922.

Hinsinger, P., Bengough, A. G., Vetterlein, D. & Young, I. M. (2009). Rhizosphere: biophysics, biogeochemistry and ecological relevance. *Plant Soil* **321**, 117–152.

Hirsch, P. R. (1979). Plasmid-determined bacteriocin production by *Rhizobium leguminosarum*. *Microbiology* **113**, 219–228.

Holden, M. T. G., Seth-Smith, H. M. B., Crossman, L. C., Sebaihia, M., Bentley, S. D., Cerdeño-Tárraga, A. M., Thomson, N. R., Bason, N., Quail, M. A. & other authors. **(2009).** The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *J Bacteriol* **191**, 261–277.

Holden, M. T. G., Titball, R. W., Peacock, S. J., Cerdeño-Tárraga, A. M., Atkins, T., Crossman, L. C., Pitt, T., Churcher, C., Mungall, K. & other authors. **(2004).** Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*.

Proc Natl Acad Sci USA **101**, 14240–14245.

Hsiao, T.-L. & Vitkup, D. (2008). Role of duplicate genes in robustness against deleterious human mutations. *PLOS Genet* **4**, e1000014.

Huang, X. & Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. *Adv Appl Math* **12**, 337–357.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P. & other authors. **(2012).** InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* **40**, D306–312.

Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J. & other authors. **(2016).** Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253–aad6253.

Hynes, M. F. & McGregor, N. F. (1990). Two plasmids other than the nodulation plasmid are necessary for formation of nitrogen-fixing nodules by *Rhizobium leguminosarum*. *Mol Microbiol* **4**, 567–574.

Hynes, M. F., Quandt, J., O'Connell, M. P. & Pühler, A. (1989). Direct selection for curing and deletion of *Rhizobium* plasmids using transposons carrying the *Bacillus subtilis sacB* gene. *Gene* **78**, 111–120.

Hynes, M. F., Simon, R., Müller, P., Niehaus, K., Labes, M. & Pühler, A. (1986). The two megaplasmids of *Rhizobium meliloti* are involved in the effective nodulation of alfalfa. *Mol Gen Genet* **202**, 356–362.

Hynes, M. F., Simon, R. & Puhler, A. (1985). The development of plasmid-free strains of *Agrobacterium tumefaciens* by using incompatibility with a *Rhizobium meliloti* plasmid to eliminate pAtC58. *Plasmid* **13**, 99–105.

Imam, S., Yilmaz, S., Sohmen, U., Gorzalski, A. S., Reed, J. L., Noguera, D. R. & Donohue, T. J. (2011). iRsp1095: A genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network. *BMC Syst Biol* **5**, 116.

Itaya, M. & Tanaka, T. (1997). Experimental surgery to create subgenomes of *Bacillus subtilis* 168. *Proc Natl Acad Sci USA* **94**, 5378–5382.

Iwadate, Y., Honda, H., Sato, H., Hashimoto, M. & Kato, J.-I. (2011). Oxidative stress sensitivity of engineered *Escherichia coli* cells with a reduced genome. *FEMS Microbio Lett* **322**, 25–33.

Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C. & other authors. **(2003).** Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA* **100**, 14339–14344.

Jaeger, C. H., III, Lindow, S. E., Miller, W., E, C. & Firestone, M. K. (1999). Mapping of sugar and amino acid availability in soil around roots with bacterial sensors of sucrose and tryptophan. *Appl Environ Microbiol* **65**, 2685–2690.

Jain, R., Rivera, M. C. & Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* **96**, 3801–3806.

Janczarek, M., Jaroszuk-Sciseł, J. & Skorupska, A. (2009). Multiple copies of *rosR*

and *pssA* genes enhance exopolysaccharide production, symbiotic competitiveness and clover nodulation in *Rhizobium leguminosarum* bv. *trifolii*. *Antonie Van Leeuwenhoek* **96**, 471–486.

Janssen, P. J., Van Houdt, R., Moors, H., Monsieurs, P., Morin, N., Michaux, A., Benotmane, M. A., Leys, N., Vallaeys, T. & other authors. **(2010).** The complete genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh and anthropogenic environments. *PLOS ONE* **5**, e10433.

Jebbar, M., Sohn-Bosser, L., Bremer, E., Bernard, T. & Blanco, C. (2005). Ectoine-induced proteins in *Sinorhizobium meliloti* include an ectoine ABC-type transporter involved in osmoprotection and ectoine catabolism. *J Bacteriol* **187**, 1293–1304.

Jensen, H. L. (1942). Nitrogen fixation in leguminous plants. I. General characters of root-nodule bacteria isolated from species of *Medicago* and *Trifolium* in Australia. *Proc Linn Soc NSW* **67**, 98–108.

Jensen, J. B., Peters, N. K. & Bhuvaneswari, T. V. (2002). Redundancy in periplasmic binding protein-dependent transport systems for trehalose, sucrose, and maltose in *Sinorhizobium meliloti*. *J Bacteriol* **184**, 2978–2986.

Jeong, J.-Y., Yim, H.-S., Ryu, J.-Y., Lee, H. S., Lee, J.-H., Seen, D.-S. & Kang, S. G. (2012). One-step sequence- and ligation-independent cloning as a rapid and versatile cloning method for functional genomics studies. *Appl Environ Microbiol* **78**, 5440–5443.

Johnson, T. J. & Nolan, L. K. (2009). Pathogenomics of the virulence plasmids of

Escherichia coli. *Microbiol Mol Biol Rev* **73**, 750–774.

Jones, D. L. (1998). Organic acids in the rhizosphere – a critical review. *Plant Soil* **205**, 25–44.

Jones, K. M., Kobayashi, H., Davies, B. W., Taga, M. E. & Walker, G. C. (2007). How rhizobial symbionts invade plants: the *Sinorhizobium–Medicago* model. *Nature Rev Microbiol* **5**, 619–633.

Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., O'Callaghan, D. & Ramuz, M. (1998). Differences in chromosome number and genome rearrangements in the genus *Brucella*. *Mol Microbiol* **27**, 99–106.

Junier, I. (2014). Conserved patterns in bacterial genomes: A conundrum physically tailored by evolutionary tinkering. *Comp Biol Chem* **53**, 125–133.

Kamilova, F., Kravchenko, L. V., Shaposchnikov, A. I., Azarova, T., Makarova, N. & Lugtenberg, B. (2006). Organic acids, sugars, and L-tryprophane in esudates of vegetables growing on stonewool and their effects on activities of rhizosphere bacteria. *Mol Plant Microbe Interact* **19**, 250–256.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**, D199–205.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., Watanabe, A., Idesawa, K., Iriguchi, M. & other authors. **(2002).** Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum*

USDA110. *DNA Res* **9**, 189–197.

Karlin, S. & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**, 283–290.

Kiss, H., Cleland, D., Lapidus, A., Lucas, S., Del Rio, T. G., Nolan, M., Tice, H., Han, C., Goodwin, L. & other authors. **(2010).** Complete genome sequence of '*Thermobaculum terrenum*' type strain (YNP1). *Stand Genomic Sci* **3**, 153–162.

Knee, E. M., Gong, F.-C., Gao, M., Teplitski, M., Jones, A. R., Goxworthy, A., Mort, A. J. & Bauer, W. D. (2001). Root mucilage from pea and its utilization by rhizosphere bacteria as a sole carbon source. *Mol Plant Microbe Interact* **14**, 775–784.

Kohler, P. R. A., Choong, E. L. & Rossbach, S. (2011). The RpiR-Like repressor IolR regulates inositol catabolism in *Sinorhizobium meliloti*. *J Bacteriol* **193**, 5155–5163.

Kohler, P. R. A., Zheng, J. Y., Schoffers, E. & Rossbach, S. (2010). Inositol catabolism, a key pathway in *Sinorhizobium meliloti* for competitive host nodulation. *Appl Environ Microbiol* **76**, 7972–7980.

Kolker, E., Makarova, K. S., Shabalina, S., Picone, A. F., Purvine, S., Holzman, T., Cherny, T., Armbruster, D., Munson, R. S. & other authors. **(2004).** Identification and functional analysis of 'hypothetical' genes expressed in *Haemophilus influenzae*. *Nucleic Acids Res* **32**, 2353–2361.

Kovach, M. E., Elzer, P. H., Hill, D. S., Robertson, G. T., Farris, M. A., Roop, R. M. & Peterson, K. M. (1995). Four new derivatives of the broad-host-range cloning

vector pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene* **166**, 175–176.

Kuo, C.-H., Moran, N. A. & Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19**, 1450–1454.

Kuo, C.-H. & Ochman, H. (2009). Deletional bias across the three domains of life. *Genome Biol Evol* **1**, 145–152.

Kurland, C. G., Canback, B. & Berg, O. G. (2003). Horizontal gene transfer: a critical view. *Proc Natl Acad Sci USA* **100**, 9658–9662.

Lambert, A., Osteras, M., Mandon, K., Poggi, M. C. & Le Rudulier, D. (2001). Fructose uptake in *Sinorhizobium meliloti* is mediated by a high-affinity ATP-binding cassette transport system. *J Bacteriol* **183**, 4709–4717.

Landeta, C., Dávalos, A., Cevallos, M. Á., Geiger, O., Brom, S. & Romero, D. (2011). Plasmids with a chromosome-like role in rhizobia. *J Bacteriol* **193**, 1317–1326.

las Nieves Peltzer, de, M., Roques, N., Poinsot, V., Aguilar, O. M., Batut, J. & Capela, D. (2008). Auxotrophy accounts for nodulation defect of most *Sinorhizobium meliloti* mutants in the branched-chain amino acid biosynthesis pathway. *Mol Plant Microbe Interact* **21**, 1232–1241.

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

Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA* **95**, 9413–9417.

Lawrence, J. G. (2003). Genome organization: Selection, selfishness, and serendipity. *Annu Rev Microbiol* **57**, 419–440.

Lerat, E. & Ochman, H. (2005). Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* **33**, 3125–3132.

Lercher, M. J. & Pál, C. (2008). Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* **25**, 559–567.

Lerivrey, J., Dubois, B., Decock, P., Micera, G., Urbanska, J. & Kozłowski, H. (1986). Formation of D-glucosamine complexes with Cu(II), Ni(II) and Co(II) ions. *Inorg Chim Acta* **125**, 187–190.

Lerouge, P., Roche, P., Faucher, C., Maillet, F., Truchet, G., Promé, J. C. & Dénarié, J. (1990). Symbiotic host-specificity of *Rhizobium meliloti* is determined by a sulphated and acylated glucosamine oligosaccharide signal. *Nature* **344**, 781–784.

Lesterlin, C., Pages, C., Dubarry, N., Dasgupta, S. & Cornet, F. (2008). Asymmetry of chromosome replichores renders the DNA translocase activity of FtsK essential for cell division and cell shape maintenance in *Escherichia coli*. *PLOS Genet* **4**, e1000288.

Li, J., Yuan, Z. & Zhang, Z. (2010). The cellular robustness by genetic redundancy in budding yeast. *PLOS Genet* **6**, e1001187.

Liang, X., Baek, C.-H. & Katzen, F. (2013). *Escherichia coli* with two linear chromosomes. *ACS Synth Biol* **2**, 734–740.

Lipton, D. S., Blanchar, R. W. & Blevins, D. G. (1987). Citrate, malate, and succinate

concentration in exudates from P-sufficient and P-stressed *Medicago sativa* L. seedlings. *Plant Physiol* **85**, 315–317.

Liu, G.-R., Liu, W.-Q., Johnston, R. N., Sanderson, K. E., Li, S.-X. & Liu, S.-L. (2006). Genome plasticity and *ori-ter* rebalancing in *Salmonella typhi*. *Mol Biol Evol* **23**, 365–371.

Long, S. R. (2001). Genes and signals in the *Rhizobium*-legume symbiosis. *Plant Physiol* **125**, 69–72.

Loper, J. E. & Henkels, M. D. (1997). Availability of iron to *Pseudomonas fluorescens* in rhizosphere and bulk soil evaluated with an ice nucleation reporter gene. *Appl Environ Microbiol* **63**, 99–105.

López-Guerrero, M. G., Ormeño-Orrillo, E., Acosta, J. L., Mendoza-Vargas, A., Rogel, M. A., Ramírez, M. A., Rosenblueth, M., Martínez-Romero, J. & Martínez-Romero, E. (2012). Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. *Plasmid* **68**, 149–158.

Lynch, D., O'Brien, J., Welch, T., Clarke, P., O Cuiv, P., Crosa, J. H. & O'Connell, M. (2001). Genetic organization of the region encoding regulation, biosynthesis, and transport of Rhizobactin 1021, a siderophore produced by *Sinorhizobium meliloti*. *J Bacteriol* **183**, 2576–2585.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* **60**, 327–349.

Mackenzie, C., Choudhary, M., Larimer, F. W., Predki, P. F., Stilwagen, S.,

Armitage, J. P., Barber, R. D., Donohue, T. J., Hosler, J. P. & other authors. **(2001).** The home stretch, a first analysis of the nearly completed genome of *Rhodobacter sphaeroides* 2.4.1. *Photosynthesis Research* **70**, 19–41.

Mackenzie, C., Eraso, J. M., Choudhary, M., Roh, J. H., Zeng, X., Bruscella, P., Puskás, Á. & Kaplan, S. (2007). Postgenomic adventures with *Rhodobacter sphaeroides*. *Annu Rev Microbiol* **61**, 283–307.

MacLean, A. M., MacPherson, G., Aneja, P. & Finan, T. M. (2006). Characterization of the β-ketoadipate pathway in *Sinorhizobium meliloti*. *Appl Environ Microbiol* **72**, 5403–5413.

MacLean, A. M., Finan, T. M. & Sadowsky, M. J. (2007). Genomes of the symbiotic nitrogen-fixing bacteria of legumes. *Plant Physiol* **144**, 615–622.

MacLean, A. M., Haerty, W., Golding, G. B. & Finan, T. M. (2011). The LysR-type PcaQ protein regulates expression of a protocatechuate-inducible ABC-type transport system in *Sinorhizobium meliloti*. *Microbiology* **157**, 2522–2533.

MacLean, A. M., White, C. E., Fowler, J. E. & Finan, T. M. (2009). Identification of a hydroxyproline transport system in the legume endosymbiont *Sinorhizobium meliloti*. *Mol Plant Microbe Interact* **22**, 1116–1127.

MacLellan, S. R., Smallbone, L. A., Sibley, C. D. & Finan, T. M. (2005). The expression of a novel antisense gene mediates incompatibility within the large *repABC* family of α-proteobacterial plasmids. *Mol Microbiol* **55**, 611–623.

MacLellan, S. R., Zaheer, R., Sartor, A. L., MacLean, A. M. & Finan, T. M. (2006).

Identification of a megaplasmid centromere reveals genetic structural diversity within the *repABC* family of basic replicons. *Mol Microbiol* **59**, 1559–1575.

Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **39**, D52–D57.

Maida, I., Fondi, M., Orlandini, V., Emiliani, G., Papaleo, M. C., Perrin, E. & Fani, R. (2014). Origin, duplication and reshuffling of plasmid genes: Insights from *Burkholderia vietnamiensis* G4 genome. *Genomics* **103**, 229–238.

Malhotra, H. C., Prakash, J. & Sharma, G. C. (1986). Kinetics of chelation of Co(II) with L-histidine. *Proc Indian Natn Sci Acad* **53**, 223–231.

Marchetti, M., Capela, D., Glew, M., Cruveiller, S., Chane-Woon-Ming, B., Gris, C., Timmers, T., Poinsot, V., Gilbert, L. B. & other authors. **(2010).** Experimental evolution of a plant pathogen into a legume symbiont. *PLOS Biol* **8**, e1000280–10.

Maróti, G. & Kondorosi, E. (2014). Nitrogen-fixing *Rhizobium*-legume symbiosis: are polyploidy and host peptide-governed symbiont differentiation general principles of endosymbiosis? *Front Microbiol* **5**, 326.

Martens, M., Dawyndt, P., Coopman, R., Gillis, M., De Vos, P. & Willems, A. (2008). Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). *Int J Syst Evol Microbiol* **58**, 200–214.

Martens, M., Delaere, M., Coopman, R., De Vos, P., Gillis, M. & Willems, A. (2007). Multilocus sequence analysis of *Ensifer* and related taxa. *Int J Syst Evol Microbiol*

57, 489–503.

Martínez, E., Palacios, R. & Sánchez, F. (1987). Nitrogen-fixing nodules induced by *Agrobacterium tumefaciens* harboring *Rhizobium phaseoli* plasmids. *J Bacteriol* **169**, 2828–2834.

Martínez-Abarca, F., Martínez-Rodríguez, L., López-Contreras, J. A., Jiménez-Zurdo, J. I. & Toro, N. (2013). Complete genome sequence of the alfalfa symbiont *Sinorhizobium/Ensifer meliloti* Strain GR4. *Genome Announc* **1**, e00174–12–e00174–12.

Mauchline, T. H., Fowler, J. E., East, A. K., Sartor, A. L., Zaheer, R., Hosie, A. H. F., Poole, P. S. & Finan, T. M. (2006). Mapping the *Sinorhizobium meliloti* 1021 solute-binding protein-dependent transportome. *Proc Natl Acad Sci USA* **103**, 17933–17938.

Meade, H. M., Long, S. R., Ruvkun, G. B., Brown, S. E. & Ausubel, F. M. (1982). Physical and genetic characterization of symbiotic and auxotrophic mutants of *Rhizobium meliloti* induced by transposon Tn*5* mutagenesis. *J Bacteriol* **149**, 114–122.

Mellata, M., Ameiss, K., Mo, H. & Curtiss, R. (2010). Characterization of the contribution to virulence of three large plasmids of avian pathogenic *Escherichia coli* χ7122 (O78:K80:H9). *Infect and Immun* **78**, 1528–1541.

Michaux, S., Paillisson, J., Carles-Nurit, M. J., Bourg, G., Allardet-Servent, A. & Ramuz, M. (1993). Presence of two independent chromosomes in the *Brucella*

melitensis 16M genome. *J Bacteriol* **175**, 701–705.

Miller, M. A., Pfeiffer, W. & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop*, pp. 1–8. Presented at the Gateway Computing Environments Workshop.

Milunovic, B., diCenzo, G. C., Morton, R. A. & Finan, T. M. (2014). Cell growth inhibition upon deletion of four toxin-antitoxin loci from the megaplasmids of *Sinorhizobium meliloti*. *J Bacteriol* **196**, 811–824.

Miura, K. (1970). An agar medium for aquatic hyphomycetes. *Trans Mycol Soc Jpn* **11**, 116–118.

Moënne-Loccoz, Y., Baldani, J. I. & Weaver, R. W. (1995). Sequential heat-curing of Tn*5*-Mob-*sac* labeled plasmids form *Rhizobium* to obtain derivatives with various combinations of plasmids and no plasmid. *Lett Appl Microbiol* **20**, 175-179.

Moody, S. F., Clarke, A. E. & Bacic, A. (1988). Structural analysis of secreted slimre from wheat and cowpea roots. *Phytochemistry* **27**, 2857–2861.

Mora, Y., Díaz, R., Vargas-Lagunas, C., Peralta, H., Guerrero, G., Aguilar, A., Encarnación, S., Girard, L. & Mora, J. (2014). Nitrogen-fixing rhizobial strains isolated from common bean seeds: phylogeny, physiology, and genome analysis. *Appl Environ Microbiol* **80**, 5644–5654.

Moreno, E. (1998). Genome evolution within the alpha Proteobacteria: why do some bacteria not possess plasmids and others exhibit more than one different

chromosome? *FEMS Microbiol Rev* **22**, 255–275.

Moretto, M., Sonego, P., Dierckxsens, N., Brilli, M., Bianco, L., Ledezma-Tejeida, D., Gama-Castro, S., Galardini, M., Romualdi, C. & other authors. **(2016).** COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses. *Nucleic Acids Res* **44**, D620–3.

Morton, E. R., Merritt, P. M., Bever, J. D. & Fuqua, C. (2013). Large deletions in the pAtC58 megaplasmid of *Agrobacterium tumefaciens* can confer reduced carriage cost and increased expression of virulence genes. *Genome Biol Evol* **5**, 1353–1364.

Mostafavi, M., Lewis, J. C., Saini, T., Bustamante, J. A., Gao, I. T., Tran, T. T., King, S. N., Huang, Z. & Chen, J. C. (2014). Analysis of a taurine-dependent promoter in *Sinorhizobium meliloti* that offers tight modulation of gene expression. *BMC Microbiol* **14**, 295.

Mousavi, S. A., Willems, A., Nesme, X., de Lajudie, P. & Lindström, K. (2015). Revised phylogeny of *Rhizobiaceae*: Proposal of the delineation of *Pararhizobium* gen. nov., and 13 new species combinations. *Syst Appl Microbiol* **38**, 84–90.

Murayama, S. (1981). Persistency and monosaccharide composition of polysaccharides of soil which received no plant materials for a certain period under field conditions. *Soil Sci Plant Nutr* **27**, 463–475.

Murphy, J. & Riley, J. P. (1962). A modified single solution method for the determination of phosphate in natural waters. *Anal Chim Acta* **27**, 31–36.

Mus, F., Crook, M. B., Garcia, K., Garcia Costas, A., Geddes, B. A., Kouri, E. D.,

Paramasivan, P., Ryu, M.-H., Oldroyd, G. E. D. & other authors. **(2016).** Symbiotic nitrogen fixation and the challenges to its extension to nonlegumes. *Appl Environ Microbiol* **82**, 3698–3710.

Naher, U. A., Radziah, O., Halimi, M. S., Shamsuddin, Z. H. & Mohd Razi, I. (2008). Effect of inoculation on root exudates carbon sugar and amino acids production of different rice varieties. *Res J Microbiol* **3**, 580–587.

Nakahigashi, K., Toya, Y., Ishii, N., Soga, T., Hasegawa, M., Watanabe, H., Takai, Y., Honma, M., Mori, H. & Tomita, M. (2009). Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol* **5**, 306.

Nakamura, Y., Gojobori, T. & Ikemura, T. (1999). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**, 292.

Ng, W. V., Ciufo, S. A., Smith, T. M., Bumgarner, R. E., Baskin, D., Faust, J., Hall, B., Loretz, C., Seto, J. & other authors. **(1998).** Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? *Genome Res* **8**, 1131–1141.

Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. (2015). Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat Commun* **6**, 8924.

Noé, L. & Kucherov, G. (2005). YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* **33**, W540–3.

Nogales, J., Blanca-Ordóñez, H., Olivares, J. & Sanjuán, J. (2013). Conjugal transfer of the *Sinorhizobium meliloti* 1021 symbiotic plasmid is governed through the concerted action of one- and two-component signal transduction regulators. *Environ Microbiol* **15**, 811–821.

Nutman, P. S. (1971). Perspectives in biological nitrogen fixation. *Sci Prog* **59**, 55–74.

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

Oldroyd, G. E. D., Murray, J. D., Poole, P. S. & Downie, J. A. (2011). The rules of engagement in the legume-rhizobial symbiosis. *Annu Rev Genet* **45**, 119–144.

Oldroyd, G. E. & Dixon, R. (2014). Biotechnological solutions to the nitrogen problem. *Curr Opin Biotech* **26**, 19–24.

Oresnik, I. J., Liu, S. L., Yost, C. K. & Hynes, M. F. (2000). Megaplasmid pRme2011a of *Sinorhizobium meliloti* is not required for viability. *J Bacteriol* **182**, 3582–3586.

Ormeño-Orrillo, E., Servín-Garcidueñas, L. E., Rogel, M. A., González, V., Peralta, H., Mora, J., Martínez-Romero, J. & Martínez-Romero, E. (2015). Taxonomy of rhizobia and agrobacteria from the *Rhizobiaceae* family in light of genomics. *Syst Appl Microbiol* **38**, 287–291.

Österman, J., Marsh, J., Laine, P. K., Zeng, Z., Alatalo, E., Sullivan, J. T., Young, J. P. W., Thomas-Oates, J., Paulin, L. & Lindström, K. (2014). Genome sequencing of two *Neorhizobium galegae* strains reveals a *noeT* gene responsible for the unusual acetylation of the nodulation factors. *BMC Genomics* **15**, 500.

Park, C. & Zhang, J. (2012). High expression hampers horizontal gene transfer. *Genome Biol Evol* **4**, 523–532.

Patel, S. J., Padilla-Benavides, T., Collins, J. M. & Argüello, J. M. (2014). Functional diversity of five homologous Cu^{+}-ATPases present in *Sinorhizobium meliloti*. *Microbiology* **160**, 1237–1251.

Paulsen, I. T., Seshadri, R., Nelson, K. E., Eisen, J. A., Heidelberg, J. F., Read, T. D., Dodson, R. J., Umayam, L., Brinkac, L. M. & other authors. **(2002).** The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proc Natl Acad Sci USA* **99**, 13148–13153.

Pál, C., Papp, B. & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genet* **37**, 1372–1375.

Perrine-Walker, F. M., Hynes, M. F., Rolfe, B. G. & Hocart, C. H. (2009). Strain competition and agar affect the interaction of rhizobia with rice. *Can J Microbiol* **55**, 1217–1223.

Peters, A. E., Bavishi, A., Cho, H. & Choudhary, M. (2012). Evolutionary constraints and expression analysis of gene duplications in *Rhodobacter sphaeroides* 2.4.1. *BMC Res Notes* **5**, 192.

Pérez-Mendoza, D., Sepúlveda, E., Pando, V., Muñoz, S., Nogales, J., Olivares, J., Soto, M. J., Herrera-Cervera, J. A., Romero, D. & other authors. **(2005).** Identification of the *rctA* gene, which is required for repression of conjugative transfer of rhizobial symbiotic megaplasmids. *J Bacteriol* **187**, 7341–7350.

Pérez-Mendoza, D., Dominguez-Ferreras, A., Muñoz, S., Soto, M. J., Olivares, J., Brom, S., Girard, L., Herrera-Cervera, J. A. & Sanjuán, J. (2004). Identification of functional mob regions in *Rhizobium etli*: evidence for self-transmissibility of the symbiotic plasmid pRetCFN42d. *J Bacteriol* **186**, 5753–5761.

Phillips, D. A., Sande, E. S., Vriezen, J. A. C., de Bruijn, F. J., Le Rudulier, D. & Joseph, C. M. (1998). A new genetic locus in *Sinorhizobium meliloti* is involved in stachydrine utilization. *Appl Environ Microbiol* **64**, 3954–2960.

Pini, F., Galardini, M., Bazzicalupo, M. & Mengoni, A. (2011). Plant-bacteria association and symbiosis: are there common genomic traits in *Alphaproteobacteria*? *Genes* **2**, 1017–1032.

Pobigaylo, N., Szymczak, S., Nattkemper, T. W. & Becker, A. (2008). Identification of genes relevant to symbiosis and competitiveness in *Sinorhizobium meliloti* using signature-tagged mutants. *Mol Plant Microbe Interact* **21**, 219–231.

Pobigaylo, N., Wetter, D., Szymczak, S., Schiller, U., Kurtz, S., Meyer, F., Nattkemper, T. W. & Becker, A. (2006). Construction of a large signature-tagged mini-Tn*5* transposon library and its application to mutagenesis of *Sinorhizobium meliloti*. *Appl Environ Microbiol* **72**, 4329–4337.

Poysti, N. J., Loewen, E. D. M., Wang, Z. & Oresnik, I. J. (2007). *Sinorhizobium meliloti* pSymB carries genes necessary for arabinose transport and catabolism. *Microbiology* **153**, 727–736.

Pósfai, G., Plunkett, G., Fehér, T., Frisch, D., Keil, G. M., Umenhoffer, K.,

Kolisnychenko, V., Stahl, B., Sharma, S. S. & other authors. **(2006).** Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046.

Prell, J., Boesten, B., Poole, P. & Priefer, U. B. (2002). The *Rhizobium leguminosarum* bv. *viciae* VF39 γ-aminobutyrate (GABA) aminotransferase gene (*gabT*) is induced by GABA and highly expressed in bacteroids. *Microbiology* **148**, 615–623.

Prentki, P. & Krisch, H. M. (1984). In vitro insertional mutagenesis with a selectable DNA fragment. *Gene* **29**, 303–313.

Prozorov, A. A. (2008). Additional chromosomes in bacteria: Properties and origin. *Mikrobiologiia* **77**, 385–394.

Quandt, J. & Hynes, M. F. (1993). Versatile suicide vectors which allow direct selection for gene replacement in Gram-negative bacteria. *Gene* **127**, 15–21.

Rabin, R. S. & Stewart, V. (1992). Either of two functionally redundant sensor proteins, *NarX* and *NarQ*, is sufficient for nitrate regulation in *Escherichia coli* K-12. *Proc Natl Acad Sci USA* **89**, 8419–8423.

Raimunda, D. & Elso-Berberián, G. (2014). Functional characterization of the CDF transporter SMc02724 (SmYiiP) in *Sinorhizobium meliloti*: Roles in manganese homeostasis and nodulation. *BBA - Biomembranes* **1838**, 3203–3211.

Ramachandran, V. K., East, A. K., Karunakaran, R., Downie, J. A. & Poole, P. S. (2011). Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol* **12**, R106.

Rasmussen, T., Jensen, R. B. & Skovgaard, O. (2007). The two chromosomes of *Vibrio*

cholerae are initiated at different time points in the cell cycle. *EMBO J* **26**, 3124–3131.

Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. Ø. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4**, R54.

Reeve, W., Chain, P., OHara, G., Ardley, J., Nandesena, K., Bräu, L., Tiwari, R., Malfatti, S., Kiss, H. & other authors. **(2010).** Complete genome sequence of the *Medicago* microsymbiont *Ensifer* (*Sinorhizobium*) *medicae* strain WSM419. *Stand Genomic Sci* **2**, 77–86.

Reeve, W., Tian, R., Lambert, B., Goodwin, L., Munk, C., Detter, C., Tapia, R., Han, C., Liolios, K. & other authors. **(2006).** Genome sequence of *Ensifer arboris* strain LMG 14919^{T}; a microsymbiont of the legume *Prosopis chilensis* growing in Kosti, Sudan. *Stand Genomic Sci* **9**, 473–483.

Remigi, P., Zhu, J., Young, J. P. W. & Masson-Boivin, C. (2016). Symbiosis within symbiosis: Evolving nitrogen-fixing legume symbionts. *Trends Microbiol* **24**, 63–75.

Resendis-Antonio, O., Hernández, M., Mora, Y. & Encarnación, S. (2012). Functional modules, structural topology, and optimal activity in metabolic networks. *PLOS Comput Biol* **8**, e1002720.

Resendis-Antonio, O., Reed, J. L., Encarnación, S., Collado-Vides, J. & Palsson, B. Ø. (2007). Metabolic reconstruction and modeling of nitrogen fixation in *Rhizobium etli*. *PLOS Comput Biol* **3**, 1887–1895.

Richardson, J. S., Hynes, M. F. & Oresnik, I. J. (2004). A genetic locus necessary for

rhamnose uptake and catabolism in *Rhizobium leguminosarum* bv. trifolii. *J Bacteriol* **186**, 8433–8442.

Richardson, J. S. & Oresnik, I. J. (2007). L-Rhamnose transport is sugar kinase (RhaK) dependent in *Rhizobium leguminosarum* bv. trifolii. *J Bacteriol* **189**, 8437–8446.

Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* **95**, 6239–6244.

Rocha, E. P. (2006). Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol* **23**, 513–522.

Rocha, E. P. C. (2004). The replication-related organization of bacterial genomes. *Microbiology* **150**, 1609–1627.

Rocha, E. P. C. (2008). The organization of the bacterial genome. *Annu Rev Genet* **42**, 211–233.

Rogel, M. A., Hernández-Lucas, I., Kuykendall, L. D., Balkwill, D. L. & Martinez-Romero, E. (2001). Nitrogen-fixing nodules with *Ensifer adhaerens* harboring *Rhizobium tropici* symbiotic plasmids. *Appl Environ Microbiol* **67**, 3264–3268.

Rogers, C. & Oldroyd, G. E. D. (2014). Synthetic biology approaches to engineering the nitrogen symbiosis in cereals. *J Exp Bot* **65**, 1939–1946.

Romanchuk, A., Jones, C. D., Karkare, K., Moore, A., Smith, B. A., Jones, C., Dougherty, K. & Baltrus, D. A. (2014). Bigger is not always better: transmission and fitness burden of ~1MB Pseudomonas syringae megaplasmid pMPPla107. *Plasmid* **73**, 16–25.

Rosenberg, C., Boistard, P., Dénarié, J. & Casse-Delbart, F. (1981). Genes controlling early and late functions in symbiosis are located on a megaplasmid in *Rhizobium meliloti*. *Mol Gen Genet* **184**, 326–333.

Roux, B., Rodde, N., Jardinaud, M.-F., Timmers, T., Sauviac, L., Cottret, L., Carrère, S., Sallet, E., Courcelle, E. & other authors. **(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**, 817–837.

Rudder, S., Doohan, F., Creevey, C. J., Wendt, T. & Mullins, E. (2014). Genome sequence of *Ensifer adhaerens* OV14 provides insights into its ability as a novel vector for the genetic transformation of plant genomes. *BMC Genomics* **15**, 1–17.

Sagot, B., Gaysinski, M., Mehiri, M., Guigonis, J.-M., Le Rudulier, D. & Alloing, G. (2010). Osmotically induced synthesis of the dipeptide *N*-acetylglutaminylglutamine amide is mediated by a new pathway conserved among bacteria. *Proc Natl Acad Sci USA* **107**, 12652–12657.

Sallet, E., Roux, B., Sauviac, L., Jardinaud, M.-F., Carrère, S., Faraut, T., de Carvalho-Niebel, F., Gouzy, J., Gamas, P. & other authors. **(2013).** Next-generation annotation of prokaryotic genomes with EuGene-P: application to *Sinorhizobium meliloti* 2011. *DNA Res* **20**, 339–354.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. New York: Cold Spring Harbor Laboratory Press.

San Millan, A., Toll-Riera, M., Qi, Q. & MacLean, R. C. (2015). Interactions between

horizontally acquired genes create a fitness cost in *Pseudomonas aeruginosa*. *Nat Commun* **6**, 6845.

Sanderson, K. E. & Roth, J. R. (1988). Linkage map of *Salmonella typhimurium*, Edition VII. *Microbiol Rev* **52**, 485–532.

Schatschneider, S., Persicke, M., Watt, S. A., Hublik, G., Pühler, A., Niehaus, K. & Vorhölter, F.-J. (2013). Establishment, *in silico* analysis, and experimental verification of a large-scale metabolic network of the xanthan producing *Xanthomonas campestris* pv. campestris strain B100. *J Biotechnol* **167**, 123–134.

Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E. & other authors. **(2011).** Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* **6**, 1290–1307.

Schlüter, J.-P., Reinkensmeier, J., Barnett, M. J., Lang, C., Krol, E., Giegerich, R., Long, S. R. & Becker, A. (2013). Global mapping of transcription start sites and promoter motifs in the symbiotic α-proteobacterium *Sinorhizobium meliloti* 1021. *BMC Genomics* 156.

Schlüter, J.-P., Reinkensmeier, J., Daschkey, S., Evguenieva-Hackenberg, E., Janssen, S., Jänicke, S., Becker, J. D., Giegerich, R. & Becker, A. (2010). A

genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium *Sinorhizobium meliloti*. *BMC Genomics* **11**, 245.

Schmeisser, C., Liesegang, H., Krysciak, D., Bakkou, N., Le Quéré, A., Wollherr, A., Heinemeyer, I., Morgenstern, B., Pommerening-Röser, A. & other authors. **(2009).** *Rhizobium* sp. strain NGR234 possesses a remarkable number of secretion systems. *Appl Environ Microbiol* **75**, 4035–4045.

Schmid, J., Sieber, V. & Rehm, B. (2015). Bacterial exopolysaccharides: biosynthesis pathways and engineering strategies. *Front Microbiol* **6**.

Schneiker-Bekel, S., Wibberg, D., Bekel, T., Blom, J., Linke, B., Neuweger, H., Stiens, M., Vorhölter, F.-J., Weidner, S. & other authors. **(2011).** The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome. *J Biotechnol* **155**, 20–33.

Schroeder, B. K., House, B. L., Mortimer, M. W., Yurgel, S. N., Maloney, S. C., Ward, K. L. & Kahn, M. L. (2005). Development of a functional genomics platform for *Sinorhizobium meliloti*: construction of an ORFeome. *Appl Environ Microbiol* **71**, 5858–5864.

Schuldes, J., Rodriguez Orbegoso, M., Schmeisser, C., Krishnan, H. B., Daniel, R. & Streit, W. R. (2012). Complete genome sequence of the broad-host-range strain *Sinorhizobium fredii* USDA257. *J Bacteriol* **194**, 4483–4483.

Segovia, L., Piñero, D., Palacios, R. & Martinez-Romero, E. (1991). Genetic structure of a soil population of nonsymbiotic *Rhizobium leguminosarum*. *Appl Environ*

Microbiol **57**, 426–433.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M. & other authors. **(2011).** Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539–539.

Silver, W. S. (1969). Biology and ecology of nitrogen fixation by symbiotic associations of non-leguminous plants. *Proc R Soc Lond, B, Biol Sci* **172**, 389–400.

Simon, R., Quandt, J. & Klipp, W. (1989). New derivatives of transposon Tn*5* suitable for mobilization of replicons, generation of operon fusions and induction of genes in Gram-negative bacteria. *Gene* **80**, 161–169.

Slater, S. C., Goldman, B. S., Goodner, B., Setubal, J. C., Farrand, S. K., Nester, E. W., Burr, T. J., Banta, L., Dickerman, A. W. & other authors. **(2009).** Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. *J Bacteriol* **191**, 2501–2511.

Sobrero, P., Schlüter, J.-P., Lanner, U., Schlosser, A., Becker, A. & Valverde, C. (2012). Quantitative proteomic analysis of the Hfq-regulon in *Sinorhizobium meliloti* 2011. *PLOS ONE* **7**, e48494.

Song, J., Ware, A. & Liu, S.-L. (2003). Wavelet to predict bacterial *ori* and *ter*: a tendency towards a physical balance. *BMC Genomics* **4**, 17.

Soto, M. J., van Dillewijn, P., Olivares, J. & Toro, N. (1994). Ornithine cyclodeaminase activity in *Rhizobium meliloti*. *FEMS Microbio Lett* **119**, 209–214.

Spini, G., Decorosi, F., Cerboneschi, M., Tegli, S., Mengoni, A., Viti, C. & Giovannetti, L. (2015). Effect of the plant flavonoid luteolin on *Ensifer meliloti* 3001 phenotypic responses. *Plant Soil* **399**, 159–178.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* **57**, 758–771.

Stanley, J., Dowling, D. N. & Broughton, W. J. (1988). Cloning of *hemA* from *Rhizobium* sp. NGR234 and symbiotic phenotype of a gene-directed mutant in diverse legume genera. *Mol Gen Genet* **215**, 32–37.

Steele, T. T., Fowler, C. W. & Griffitts, J. S. (2009). Control of gluconate utilization in *Sinorhizobium meliloti*. *J Bacteriol* **191**, 1355–1358.

Stewart, P. E., Hoff, J., Fischer, E., Krum, J. G. & Rosa, P. A. (2004). Genome-wide transposon mutagenesis of *Borrelia burgdorferi* for identification of phenotypic mutants. *Appl Environ Microbiol* **70**, 5973–5979.

Stowers, M. D. (1985). Carbon metabolism in *Rhizobium* species. *Annu Rev Microbiol* **39**, 89–108.

Streit, W. R., Joseph, C. M. & Phillips, D. A. (1996). Biotin and other water-soluble vitamins are key growth factors for alfalfa root colonization by *Rhizobium meliloti* 1021. *Mol Plant Microbe Interact* **9**, 330–338.

Sugawara, M., Epstein, B., Badgley, B. D., Unno, T., Xu, L., Reese, J., Gyaneshwar,

P., Denny, R., Mudge, J. & other authors. **(2013).** Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol* **14**, R17.

Sullivan, J. T., Patrick, H. N., Lowther, W. L., Scott, D. B. & Ronson, C. W. (1995). Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. *Proc Natl Acad Sci USA* **92**, 8985–8989.

Suthers, P. F., Zomorrodi, A. & Maranas, C. D. (2009). Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol Syst Biol* **5**, 301.

Suwanto, A. & Kaplan, S. (1989). Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: Presence of two unique circular chromosomes. *J Bacteriol* **171**, 5850–5859.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–2739.

Thomaides, H. B., Davison, E. J., Burston, L., Johnson, H., Brown, D. R., Hunt, A. C., Errington, J. & Czaplewski, L. (2007). Essential bacterial functions encoded by gene pairs. *J Bacteriol* **189**, 591–602.

Thorpe, H. M. & Smith, M. C. M. (1998). *In vitro* site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. *Proc Natl Acad Sci USA* **95**, 5505–5510.

Tischler, J., Ben Lehner, Chen, N. & Fraser, A. G. (2006). Combinatorial RNA interference in *Caenorhabditis elegans* reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol* **7**, R69.

Tong, A. H. Y., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W. V. & other authors. **(2001).** Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368.

Toro, N., Martínez-Abarca, F. & Nisa-Martínez, R. (2014). Complete genome sequence of the RmInt1 group II intronless *Sinorhizobium meliloti* Strain RMO17. *Genome Announc* **2**, e01001–14.

Trabelsi, D., Pini, F., Aouani, M. E., Bazzicalupo, M. & Mengoni, A. (2009). Development of real-time PCR assay for detection and quantification of *Sinorhizobium meliloti* in soil and plant tissue. *Lett Appl Microbiol* **48**, 355–361.

Udvardi, M. & Poole, P. S. (2013). Transport and metabolism in legume-rhizobia symbioses. *Annu Rev Plant Biol* **64**, 781–805.

Van Houdt, R. & Mergeay, M. (2012). Plasmids as secondary chromosomes. In *Molecular Life Sciences An Encyclopedic Reference*, pp. 1–4.

Van Houdt, R., Monsieurs, P., Mijnendonckx, K., Provoost, A., Janssen, A., Mergeay, M. & Leys, N. (2012). Variation in genomic islands contribute to genome plasticity in *Cupriavidus metallidurans*. *BMC Genomics* **13**, 111.

Van Melderen, L. & Saavedra De Bast, M. (2009). Bacterial toxin-antitoxin systems: more than selfish entities? *PLOS Genet* **5**, e1000437.

van Opijnen, T. & Camilli, A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res* **22**, 2541–2551.

van Passel, M. W. J., Bart, A., Luyf, A. C. M., van Kampen, A. H. C. & van der Ende, A. (2006). Compositional discordance between prokaryotic plasmids and host chromosomes. *BMC Genomics* **7**, 26.

Vieira-Silva, S., Touchon, M. & Rocha, E. P. C. (2010). No evidence for elemental-based streamlining of prokaryotic genomes. *Trends Ecol Evol* **25**, 319–20.

Villaseñor, T., Brom, S., Dávalos, A., Lozano, L., Romero, D. & Santos, A. G.-D. L. (2011). Housekeeping genes essential for pantothenate biosynthesis are plasmid-encoded in *Rhizobium etli* and *Rhizobium leguminosarum*. *BMC Microbiol* **11**, 66.

Volff, J.-N. & Altenbuchner, J. (2000). A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbio Lett* **186**, 143–150.

Wang, C., Sheng, X., Equi, R. C., Trainer, M. A., Charles, T. C. & Sobral, B. W. S. (2007). Influence of the poly-3-hydroxybutyrate (PHB) granule-associated proteins (PhaP1 and PhaP2) on PHB accumulation and symbiotic nitrogen fixation in *Sinorhizobium meliloti* Rm1021. *J Bacteriol* **189**, 9050–9056.

Wang, D., Yang, S., Tang, F. & Zhu, H. (2012). Symbiosis specificity in the legume: rhizobial mutualism. *Cell Microbiol* **14**, 334–342.

Wang, Z. & Zhang, J. (2009). Abundant indispensable redundancies in cellular

metabolic networks. *Genome Biol Evol* **1**, 23–33.

Watson, R. J., Chan, Y. K., Wheatcroft, R., Yang, A. F. & Han, S. H. (1988). *Rhizobium meliloti* genes required for C_4-dicarboxylate transport and symbiotic nitrogen fixation are located on a megaplasmid. *J Bacteriol* **170**, 927–934.

Weidner, S., Baumgarth, B., Göttfert, M., Jaenicke, S., Pühler, A., Schneiker-Bekel, S., Serrania, J., Szczepanowski, R. & Becker, A. (2013). Genome sequence of *Sinorhizobium meliloti* Rm41. *Genome Announc* **1**, e00013–12.

Weidner, S., Becker, A., Bonilla, I., Jaenicke, S., Lloret, J., Margaret, I., Pühler, A., Ruiz-Sainz, J. E., Schneiker-Bekel, S. & other authors. **(2012).** Genome sequence of the soybean symbiont *Sinorhizobium fredii* HH103. *J Bacteriol* **194**, 1617–1618.

Weissenmayer, B., Gao, J. L., López-Lara, I. M. & Geiger, O. (2002). Identification of a gene required for the biosynthesis of ornithine-derived lipids. *Mol Microbiol* **45**, 721–733.

White, C. E., Gavina, J. M. A., Morton, R., Britz-McKibbin, P. & Finan, T. M. (2012). Control of hydroxyproline catabolism in *Sinorhizobium meliloti*. *Mol Microbiol* **85**, 1133–1147.

Wiedenbeck, J. & Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* **35**, 957–976.

Willis, L. B. & Walker, G. C. (1999). A novel *Sinorhizobium meliloti* operon encodes an α-glucosidase and a periplasmic-binding-protein-dependent transport system for α-

glucosides. *J Bacteriol* **181**, 4176–4184.

Wilson, J. J. & Kappler, U. (2009). Sulfite oxidation in *Sinorhizobium meliloti*. *BBA - Bioenergetics* **1787**, 1516–1525.

Wong, K. & Golding, G. B. (2003). A phylogenetic analysis of the pSymB replicon from the *Sinorhizobium meliloti* genome reveals a complex evolutionary history. *Can J Microbiol* **49**, 269–280.

Wong, K., Finan, T. & Golding, B. (2002). Dinucleotide compositional analysis of *Sinorhizobium meliloti* using the genome signature: distinguishing chromosomes and plasmids. *Funct Integr Genomic* **2**, 274–281.

Wu, M. & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034.

Wu, Q., Pei, J., Turse, C. & Ficht, T. A. (2006). Mariner mutagenesis of *Brucella melitensis* reveals genes with previously uncharacterized roles in virulence and survival. *BMC Microbiol* **6**, 102.

Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**, 444.

Xiao, Y. & Wall, D. (2014). Genetic redundancy, proximity, and functionality of *lspA*, the target of antibiotic TA, in the *Myxococcus xanthus* producer strain. *J Bacteriol* **196**, 1174–1183.

Xu, Q., Dziejman, M. & Mekalanos, J. J. (2003). Determination of the transcriptome of *Vibrio cholerae* during intraintestinal growth and midexponential phase in vitro. *Proc*

Natl Acad Sci USA **100**, 1286–1291.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. (2011). iPath2.0: interactive pathway explorer. *Nucleic Acids Res* **39**, W412–5.

Yang, J. C., Lessard, P. A., Sengupta, N., Windsor, S. D., O'brien, X. M., Bramucci, M., Tomb, J.-F., Nagarajan, V. & Sinskey, A. J. (2007). TraA is required for megaplasmid conjugation in *Rhodococcus erythropolis* AN12. *Plasmid* **57**, 55–70.

Yarosh, O. K., Charles, T. C. & Finan, T. M. (1989). Analysis of C_4-dicarboxylate transport genes in *Rhizobium meliloti*. *Mol Microbiol* **3**, 813–823.

Yoder-Himes, D. R., Konstantinidis, K. T. & Tiedje, J. M. (2010). Identification of potential therapeutic targets for *Burkholderia cenocepacia* by comparative transcriptomics. *PLOS ONE* **5**, e8724.

Young, J. P. W. & Wexler, M. (1988). Sym plasmid and chromosomal genotypes are correlated in field populations of *Rhizobium leguminosarum*. *Microbiology* **134**, 2731–2739.

Yuan, Z.-C., Zaheer, R. & Finan, T. M. (2006). Regulation and properties of PstSCAB, a high-affinity, high-velocity phosphate transport system of *Sinorhizobium meliloti*. *J Bacteriol* **188**, 1089–1102.

Yurgel, S. N., Mortimer, M. W., Rice, J. T., Humann, J. L. & Kahn, M. L. (2013). Directed construction and analysis of a *Sinorhizobium meliloti* pSymA deletion mutant library. *Appl Environ Microbiol* **79**, 2081–2087.

Yurgel, S., Mortimer, M. W., Rogers, K. N. & Kahn, M. L. (2000). New substrates for

the dicarboxylate transport system of *Sinorhizobium meliloti*. *J Bacteriol* **182**, 4216–4221.

Yurgel, S. N. & Kahn, M. L. (2005). *Sinorhizobium meliloti dctA* mutants with partial ability to transport dicarboxylic acids. *J Bacteriol* **187**, 1161–1172.

Zavaleta-Pastor, M., Sohlenkamp, C., Gao, J. L., Guan, Z., Zaheer, R., Finan, T. M., Raetz, C. R. H., Lopez-Lara, I. M. & Geiger, O. (2010). *Sinorhizobium meliloti* phospholipase C required for lipid remodeling during phosphorus limitation. *Proc Natl Acad Sci USA* **107**, 302–307.

Zhang, J. (2012). Genetic redundancies and their evolutionary maintenance. *Adv Exp Med Biol* **751**, 279–300. New York, NY: Springer New York.

Zhang, Y., Aono, T., Poole, P. & Finan, T. M. (2012). NAD(P)+-malic enzyme mutants of *Sinorhizobium* sp. strain NGR234, but not *Azorhizobium caulinodans* ORS571, maintain symbiotic N_2 fixation capabilities. *Appl Environ Microbiol* **78**, 2803–2812.

Zhao, H., Li, M., Fang, K., Chen, W. & Wang, J. (2012). *In silico* insights into the symbiotic nitrogen fixation in *Sinorhizobium meliloti* via metabolic reconstruction. *PLOS ONE* **7**, e31287.

Zheng, J., Guan, Z., Cao, S., Peng, D., Ruan, L., Jiang, D. & Sun, M. (2015). Plasmids are vectors for redundant chromosomal genes in the *Bacillus cereus* group. *BMC Genomics* **16**, 6.