The evolution of *Sinorhizobium meliloti.*

by

Kim Wong, B.Sc. (Hons.)

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2002) McMaster University
(Biology) Hamilton, Ontario

TITLE: Title

AUTHOR: Kim Wong, B.Sc. Hons. (McMaster University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: [x], 112

PREFACE

Each chapter of this thesis has been written as a separate manuscript. Chapter 1 is a multi-author work that is the result of a collaboration between researchers at the University of Bielefeld in Germany and McMaster University, as part of an international initiative to sequence the *Sinorhizobium meliloti* genome. My contribution to this research includes cloning, sequence assembly, editing, analysis and annotation of part of the pSymB replicon, and contribution to the preparation of the manuscript. Data collection, analysis, and manuscript preparation for Chapter 2 and Chapter 3 was primarily an individual effort, with contributions in editing and writing from the co-authors T. M. Finan and G. B. Golding.

ABSTRACT

The genome of the α-proteobacterium *Sinorhizobium meliloti* has been completely sequenced and annotated, providing a wealth of information about this endosymbiotic N_2-fixing organism. Although the structure of the genome, consisting of a circular chromosome and two smaller pSymA and pSymB replicons, has long been known, only a small portion of ORFs have previously been characterized. Sequence analysis of pSymB has revealed that a large portion of the 1570 ORFs code for solute uptake systems and polysaccharide biosynthesis. The pSymB replicon been referred to as a "megaplasmid," implying that pSymB is non-essential for viability of the organism. However, coded on pSymB are several essential genes, including a $\mathrm{tRNA}^{\mathrm{Arg}}_{\mathrm{CCG}}$ gene and the *minCDE* genes, which are not found elsewhere in the genome. Replication of pSymB is controlled by *repABC* genes, a typical property of plasmids among Rhizobiaceae. Therefore, the genome signature, a compositional analysis that allows comparison of whole replicons rather than focusing on particular genes, was used to provide support for designation of pSymB as a second chromosome in *S. meliloti.* It was found that among α-proteobacteria, plasmids and chromosomes have distinctive patterns of dinucleotide biases, and in this respect, pSymB is chromosome-like while pSymA is plasmid-like. This brings into question how the pSymB replicon came to acquire chromosome-like properties while appearing to be maintained as a plasmid in the genome. Whole-genome nearest neighbor analysis shows that the linear chromosome of *Agrobacterium tumefaciens* and pSymB may have a common origin. Despite conservation of gene order within small groups of genes, it is evident that rearrangements, duplications, and horizontal transfer of genes since the divergence of these species have contributed to the mosaic nature of pSymB. Since synteny between the *S. meliloti* chromosome and *A. tumefaciens* circular chromosome is highly conserved, it appears that the instability of pSymB has played a key role in the adaptation and evolution of *S. meliloti.*

ACKNOWLEGEMENTS

I would like to thank Brian Golding for his guidance and support throughout this project. I would also like to thank Turlough Finan for his helpful ideas and critiques. Most of all, my thanks go out to my fellow graduate students and friends that I have been lucky enough to meet along the way, especially Liisa Koski, Amanda Hawley, Mike Mitsch, Patrick Chain, Melanie Huntley, and members of the Golding lab and Finan lab, as well as my friends and family.

Contents

List of Figures

List of Tables

Chapter 1

The complete sequence of the 1,683 kilobase pSymB megaplasmid from the N_2-fixing endosymbiont *Sinorhizobium meliloti*

This is a multi-author work previously published in: *Proc Natl Acad Sci USA* (2001) **98**:9889-9894.

1.1 Introduction

Among the bacteria, the α-proteobacteria appear unusual because of the presence of multiple replicons within the same bacterial strain (Jumas-Bilak *et al.* 1998). In the case of *Agrobacterium tumefaciens*, the causative agent of crown gall disease, the genome contains both a linear and a circular chromosome (Allardet-Servent *et al.* 1993). Many (but not all) of the bacteria which form N_2-fixing root nodules on leguminous plants are characterized by the presence of multiple plasmids greater than 400 kilobases (kb) in size. In the N_2-fixing symbiont, *Sinorhizobium meliloti*, there are three replicons, a 3,654 kb circular chromosome (Meade and Signer 1977; Honeycutt, McClelland and Sobral 1993), and two megaplasmids 1,354 kb and 1,683 kb in size (Galibert *et al.* 2001; Capela *et al.* 2001; Barnett *et al.* 2001). The smaller of the megaplasmids, pSymA, is known to carry many of the genes involved in root nodule formation (*nod*) and nitrogen fixation (*nif*) (Banfalvi *et al.* 1981; Rosenberg *et al.* 1981). The 1,683 kb megaplasmid, referred to as pSymB, is known to carry various gene clusters involved in exopolysaccharide synthesis, C_4-dicarboxylate transport, and lactose metabolism (Finan *et al.* 1986; Hynes *et al.* 1986; Müller *et al.* 1993). Early studies focussed on mutations that abolished synthesis of the succinoglycan exopolysaccharide (EPS I) since these mutations resulted in a loss of the ability to form normal N_2-fixing root nodules. This symbiotic defect was rescued by second-site mutations that increased the synthesis of a second galactoglucan exopolysaccharide (EPS II) whose biosynthetic genes were also located on the pSymB megaplasmid (Glazebrook and Walker 1989; Becker *et al.* 1997). Other genes located on pSymB that are required for the formation of N_2-fixing root nodules include the C_4-dicarboxylate (*dctA*) and phosphate transport (*phoCDET*) genes and the *bacA* gene (Bardin *et al.* 1996; Finan, Oresnik and Bottacin 1988; Glazebrook, Ichige and Walker 1993; Watson *et al.* 1988). The presence of large plasmids in bacteria that form associations with plants was described over twenty years ago (Rosenberg *et al.* 1982). However, with the exception of the symbiotic genes in relatively small regions of these plasmids, the broader biological role of the plasmids in the biology of the organism has remained obscure. We constructed a genetic map of the pSymB megaplasmid and then generated strains in which over 1,200 kb of pSymB was removed (Charles and Finan 1990; Charles and Finan 1991). However, phenotypic analysis of these deletion derivatives revealed a small number of phenotypes and the functions of over 90% of the plasmid remains unknown. Here we describe the complete nucleotide sequence and annotation of the pSymB megaplasmid. We find that the plasmid has a high gene density, similar to that of previously sequenced bacterial chromosomes. Annotation of the sequence revealed a previously unsuspected richness with respect to the number of solute transport systems, polysaccharide synthesis gene clusters, transcriptional regulators, cell protection, and other genes that appear to have catabolic roles. In addition, the megaplasmid contains genes which we assume are essential for viability of the bacterium. The question as to whether the pSymB replicon should be called a megaplasmid or chromosome is one of the issues that arises from the nucleotide sequence analysis described in this report. For the sake of clarity, and to be consistent with the previous literature, we refer to the 3.6 Mb replicon as the chromosome and the pSymB and pSymA replicons as megaplasmids.

1.2 Methods

Bacterial strains. *Sinorhizobium meliloti* was isolated from New South Wales, Australia in 1939 and was originally referred to as strain SU47 (Vincent 1941). DNA from the streptomycin resistant SU47 derivative Rm1021 was employed for sequencing.

Library construction and sequencing of shotgun clones. The large DNA inserts from pSymB region Ω5056 to Ω5102 were cloned into BAC vectors using an *in vivo* cloning procedure (Chain *et al.* 2000) and the BAC DNA then was purified, nebulized, size fractionated to 1-2 kb fragments and cloned into phosphatase-treated, *SmaI*-restricted pUC118. A minimal set of 24 overlapping BAC clones (Barloy-Hubler *et al.* 2000) covering the complete pSymB megaplasmid was also used to construct shotgun libraries with different insert sizes (1.3-1.7 kb and 2.8-3.2 kb) by nebulization of BAC clone DNA and cloning into pTZ18R

(Pharmacia).

Sequencing of shotgun clones for all of pSymB was performed using M13 forward and reverse standard sequencing primers to achieve approximately 7.5-fold coverage. Contigs were closed by PCR amplification from specific primers. Sequencing reactions were performed with dye-terminator and dye-primer chemistry on ABI Prism 377-96, ABI 373S (Applied Biosystems) and LI-COR LongRead IR 4200 (LI-COR, Lincoln, NE) DNA sequencers.

Contig assembly and sequence editing. The replicon was partly sequenced and annotated by a team based at McMaster University and partly by a team based at the University of Bielefeld. The final sequence and annotation was checked and corrected by both groups. Sequence data was processed and assembled using the ACeDB assembly package (http://www.acedb.org) at McMaster University or the Phred/Phrap and Staden (Gap4) packages of base calling, sequence-assembly, and finishing/editing software at the University of Bielefeld (Ewing and Green 1998; Ewing *et al.* 1998; Staden 1996). Vector sequence was identified and removed from individual clones prior to assembly. Regions of base-call uncertainty, low coverage, or potential frameshifts were re-sequenced from shotgun clones, BAC clones, or from sequencing of the PCR product from amplification of genomic DNA. Custom made primers designed by PRIDE (Haas *et al.* 1998) were used for these purposes.

Sequence analysis and annotation. Annotation of the completed sequence was performed with the aid of various software packages. Individual open reading frames (ORFs) greater than 150 bp with BLAST (Altschul *et al.* 1997) hits with Expect values less than 10^{-10} were examined for the presence of ribosomal binding site(s), potential start sites were decided by considering ATG, GTG, and TTG start sites. FrameD, a hidden

Markov chain method trained for *Rhizobium* (Schiex, Thébault and Kahn 2000; see also http://www.toulouse.inra.fr/FrameD.html) was also used in conjunction with `BLAST` results to determine potential ORFs and start sites. Final editing of the annotation included a search for homology with genes within the three replicons, and annotations were confirmed with `ProDom` information (Sonnhammer and Kahn 1994). Annotation at University of Bielefeld employed the `GenBD` annotation database environment developed in Bielefeld. ORFs were predicted using `GLIMMER 2.0` (Delcher *et al.* 1999). Each ORF was subject to comparison with sequences in public databases using `BLAST`. In addition, `PFSCAN` (www.isrec.isb-sib.ch) and `SignalP` (www.cbs.dtu.dk/services/SignalP) were used to search protein sequences profile databases and to identify signal peptides. Predicted ORFs were reviewed individually by annotators for mis-prediction, start-codon assignment based on contextual information, and assignment of categories based on the functional gene classification of *E. coli* (Riley and Labedan 1996). *Rhizobium*-specific intergenic mosaic elements (RIMEs), motif A, B, C palindromic elements (Østeras, Stanley and Finan 1995; Østeras *et al.* 1998) and insertion sequence (IS) elements were included in the annotation of the complete sequence. The complete pSymB sequence can be viewed on the consortium website at http://sequence.toulouse.inra.fr/meliloti.html.

1.3 Results and Discussion

1.3.1 General features of the nucleotide sequence and replicon organization.

The total length of the *S. meliloti* pSymB megaplasmid is 1,683,333 base pairs, with an overall G+C content of 62.4%. Variation of G+C content throughout the replicon is evident, ranging from as low as ~56% in the *rkp* gene region (surface polysaccharide-associated export proteins) to as high as ~66% (within windows of 10,000 bp and 5,000 bp step). Ninety percent of the nucleotide sequence is predicted as protein-coding, constituting 1570 ORFs with a mean length of 959 base pairs. The distribution of ORFs on the forward and reverse strands is symmetrical, with 52% on the forward strand and 48% on the reverse strand (see Figure 1.1). A limited number of IS elements (12), RIME elements (27), and motif A, B, C elements (5) were also identified throughout the megaplasmid.

1.3.2 Origin of replication and *rep* genes.

The origin of replication of pSymB is predicted to lie within the *repA1B1C1* gene cluster and subclones of this region are sufficient to allow its autonomous replication in *Agrobacterium* (Chain *et al.* 2000). There is a second *rep* gene set which lacks the *repC* gene, *repA3B3*, that lies ~131 kb from the *repA1B1C1* cluster. RepC is believed to be directly involved in the control of the initiation of plasmid replication, whereas RepA and RepB appear to play a role in plasmid segregation; whether the *repA3B3* genes play redundant roles in plasmid segregation in *S. meliloti* remains to be established. We note that RepA1 has higher homology to RepA from other organisms than to the pSymB RepA3 protein, or the RepA2 protein encoded by pSymA.

1.3.3 Cell division, chaperonins and tRNA$^{\mathrm{Arg}}_{\mathrm{CCG}}$.

The MinCDE proteins play important roles in the placement of the cell division site in *E. coli* (de Boer, Crossley and Rothfield 1989). The only proteins in *S. meliloti* that are homologous to the *N. meningitidis*, *E. coli*, and *X. fastidiosa* MinCDE proteins are located downstream of a two-component sensor histidine kinase and response regulator (coded on the opposite strand) and a glycine-rich hypothetical protein. Whether the *S. meliloti minCDE* genes are essential has yet to be established, however, we note that these genes lie in a region for which deletion derivatives have not previously been identified (Charles and Finan 1991). The FtsK protein plays a role in septa formation and recombinational resolution of dimeric chromosomes in *E. coli*. Two FtsK homologs in *S. meliloti* share 74% amino acid identity; *ftsK1* is located on the chromosome, and *ftsK2* (a homolog of SpoIIIE) is located on pSymB close to the *repA3B3* genes. We believe the chromosomal *ftsK1* is functional for cell division since deletion derivatives which remove the *ftsK2* region of pSymB showed no obvious growth defects. It is interesting that FtsK2 is more similar in sequence and length to the *Rickettsia prowazekii* (AJ235273) FtsK homolog than is FtsK1. There are four chaperonins on pSymB, *groEL5*, *htpG*, which is a member of heat-shock-protein Hsp90 family (Bardwell and Craig 1987), and two members of the Hsp20 family (SMb21294 and SMb21295).

The sole copy of the gene coding for the tRNA specific for CCG, the second-most frequently used arginine codon, is located on pSymB next to a putative transposase gene (SMb20905) followed by a 3 kb-region containing seven small, possibly degenerating ORFs of unknown function. In view of this gene context, it is possible that the location of this tRNA gene on pSymB resulted from a transposition event.

1.3.4 Exopolysaccharide and lipopolysaccharide biosynthetic Genes.

Gram-negative bacteria exhibit complex sets of surface polysaccharides including lipopolysaccharides (LPS) (Schnaitman and Klena 1993), capsular polysaccharides (CPS) (Whitfield and Roberts 1999), exopolysaccharides (EPS) (Coplin and Cook 1990; Stevenson *et al.* 1996) and periplasmic glucans (Kennedy 1996). Genes involved in the biosynthesis and export of cell surface carbohydrates are often clustered and the *exo/exs* and *exp* gene clusters directing the synthesis of EPS I (Becker *et al.* 1993; Glucksmann, Reuber and Walker 1993) and EPS II (Glazebrook and Walker 1989; Becker *et al.* 1997) were previously mapped on pSymB of *S. meliloti.* The production of surface polysaccharides is essential for successful nodule invasion by rhizobia.

Analysis of the DNA sequence of pSymB revealed many genes whose products are typically involved in the synthesis of cell surface carbohydrates. Most of these genes are organized in 11 clusters (Table 1.1) but surprisingly, the existence of nine of these clusters ranging in size from 5 to 42 kb was previously unknown. In addition to these gene clusters, there are several isolated genes whose products also appear to be involved in the synthesis of cell surface carbohydrates. The 11 gene clusters contain 188 predicted genes and have a total size of 223 kb. Hence over 12% of the genes on pSymB are involved in the synthesis of cell surface carbohydrates.

The pSymB gene clusters 3, 4, 8 (*exo/exs*) and 9 (Table 1.1) comprise genes encoding proteins of the Wzy-dependent polymerization mechanism, although the specific target polysaccharide cannot be specified. In cluster 3 and 4 some genes encoding key functions of the Wzy-dependent export mechanism are missing. Cluster 2 contains (ABC)-2 export proteins and several of these find their closest homologs in the export machinery of capsular polysaccharides implying that this cluster may be involved in the export of CPS (Whitfield and Roberts 1999). Cluster 2 is homologous to *rkp* loci in *S. meliloti* strain 41 (Kereszt *et al.* 1998) and comprises genes *rkpRSTZ* and probable homologs of *rkpLM*. Other *rpk* genes (*rkpAGHIS* and *rkpK-lpsL*) are located on the chromosome of strain Rm1021. The *rkp* genes code for K antigens, capsular polysaccharides that have been implicated in root nodule invasion and can functionally replace EPS I or EPS II during the invasion process

(Reuhs *et al.* 1995).

Besides the ABC-transporter and Wzy-dependent export mechanisms, a third mechanism involving the ABC-transporter protein MsbA has been proposed to mediate the export of nascent LPS molecules consisting of the lipid A and core components (Zhou *et al.* 1998). A MsbA-like protein encoded by the *S. meliloti ndvA* gene is involved in the export of β-glucan (Stanfield *et al.* 1988). Other MsbA-homologs have been identified in pSymB gene clusters 1, 5 and 8 (*exo/exs*). While the cluster 1 MsbA1 protein may be involved in the synthesis and export of nascent LPS molecules, the functions of the other MsbA-like proteins are unknown.

1.3.5 Genes involved in metabolic pathways.

Biosynthesis. There are few pSymB genes predicted to play roles in amino acid or vitamin biosynthesis. Two loci are involved in thiamin biosynthesis, *thiCOGE* and *thiD*, and mutations at either loci result in thiamin auxotrophy (Finan *et al.* 1986). Despite the presence of these genes, it has been determined that availability of thiamin is growth-limiting in the rhizosphere and is a key factor promoting root colonization (Streit, Joseph and Phillips 1996). The *thiCOGE* cluster lies next to the *dct* genes encoding a C_4-dicarboxylate transport system. Since *thi* and *dct* genes are found in the symbiotic island of *Mesorhizobium loti* and *Bradyrhizobium japonicum*, it will be interesting to compare the organization of these regions across these three organisms (Göttfert *et al.* 2001; Sullivan and Ronson 1998) to determine the extent of gene shuffling and/or horizontal gene transfer that has occurred since the divergence of these species.

A putative ornithine cyclodeaminase (SMb21494), involved in L-proline biosynthesis, has 74% identity to the *Rhodobacter capsulatus* ornithine cyclodeaminase. Shikimate

5-dehydrogenase (*aroE2*) for biosynthesis of aromatic amino acids, and glutatmine synthetase II (*glnII*) are found on pSymB, and both have chromosomal homologs. Asparagine is synthesized from aspartic acid and ammonia (or glutamine) by the enzyme asparagine synthetase. Two *S. meliloti* genes, *asnB* and *asnO* encode proteins similar in size and sequence (27-33% identity) to asparagine synthetases of *E. coli*, and both are found on pSymB. While the activity of these proteins remains to be established, in the absence of an alternate route to synthesize asparagine, it would appear that at least one of these genes is essential for growth of *S. meliloti* in minimal medium.

Catabolism. A number of putative enzymes involved in amino acid degradative pathways were identified; these include components of the enzymes 2-oxoisovalerate dehydrogenase (SMb20019), propionyl-CoA carboxylase (*pccAB*), and methylmalonyl-CoA mutase (*bhbA*), which are involved in valine, leucine, isoleucine degradation. The histidine utilization genes (*hutUGHIL*) whose products are responsible for the conversion of L-histidine to glutamate and formamide are present as a single operon. Genes involved in the utilization of poly-3-hydroxybutyrate, the α-galactosides melibose and raffinose (*apaA, agpL, agpT*), and lactose (*lacEFGZ1K* and *lacZ2*), were previously identified (Charles and Finan 1991; Aneja and Charles 1999; Gage and Long 1998; Galbraith *et al.* 1998; Férrandez *et al.* 1998). A 20-kb region of pSymB (SMb21277 to SMb21293) encodes 17 genes involved in purine/pyrimidine nucleotide salvage/catabolic pathways. This cluster includes a putative uracil/xanthine permease, a purine/pyrimidine phosphoribosyltransferase, adenine deaminase, xanthine dehydrogenase, uricase and allantoicase.

Other catabolic gene clusters include pathways for the utilization of aromatic compounds associated with degradation of plant material and the mineralization of lignin. These include a 15 kb gene cluster encoding proteins involved in a pathway for conversion of hydroxyphenylpyruvate and 4-hydroxybenzoate via protocatechuate, and β-ketoadipate to tricarboxylic acid cycle intermediates (*pcaB-pcaF*). A gene cluster (*paaGZEDBAX*) similar in sequence and gene order to what is believed to be a multicomponent oxygenase involved in phenylacetic acid catabolism in *E. coli* (Férrandez *et al.* 1998), was also identified. There are four genes encoding proteins with homology to inositol monophosphatases (SMb20150, SMb20159, SMb20362, SMb21225). At least one of these is believed to be involved in inositol utilization while others may be involved in utilization of rhizopines, or possible degradation of the plant-derived phosphate storage compound, phytic acid (myo-inositol hexaphosphate).

Members of the *Rhizobiacea* family, including *S. meliloti* strain Rm1021, have been reported to be able to utilize phosphonates as a source of phosphate (Liu *et al.* 1991). Several gene clusters involved in phosphonate degradation are located on pSymB (*phnA, phnGHIJKL, phnM* and *phoCDET*). In a different strain of *S. meliloti* the *pta* and *ackA* genes encoding phosphotransacetylase and acetate kinase are part of a single operon whose expression is induced upon phosphate starvation (Summers, Denton and McDermott 1999). Interestingly, in strain Rm1021 we find that the putative *pta* and *ackA* genes are located over 50-kb apart on pSymB. The significance of the latter observation is unclear and it remains to be determined whether there is considerable gene shuffling within *S. meliloti* strains.

Among the many genes with high similarity to alcohol dehydrogenases, one gene cluster (SMb20170 to SMb20175) includes the methanol dehydrogenase structural gene. This gene is likely involved in methanol utilization, since methanol dehydrogenase requires the redox coenzyme pyrrolo-quinoline quinone (PQQ), and a complete PQQ biosynthesis gene cluster (*pqqABCDE*) is also located on pSymB.

A CO_2-fixation gene cluster, *cbbR-cbbFPTALSX*, encoding the enzymes of the Calvin-Benson-Bassham (CBB) cycle is located directly upstream of the *pqq* genes. The enzymes ribulose-1,5-bisphosphate carboxylase (small and large subunits) and phosphoribulokinase are unique to the CBB cycle and are encoded by *cbbS, cbbL,* and *cbbP*, respectively. The *cbbR* gene encodes a LysR family transcriptional regulator and, together with the other *cbb* genes, is found in both photo- and chemoautrotrophic bacteria. To date, *Bradyrhizobium japonicum* is the only rhizobium that has been shown to grow chemolithoautotrophically, utilizing CO_2 as a carbon source and H_2 as the electron donor (Lepo, Hanus and Evans 1980). It remains to be established experimentally whether *S. meliloti* can grow as an autotroph and under what conditions the *cbb* genes are expressed. However, because of the close vicinity of the methanol utilization gene cluster to the *cbb* genes, it may be speculated that growth on methanol is a carbon-limiting condition which could lead to expression of the CO_2-fixation genes (Shively, van Keulen and Meijer 1998).

ABC and other transport proteins. A prominent feature of the *S. meliloti* genome is the number genes encoding for ATP-binding cassette (ABC) transport systems. These systems contain an ATP-binding protein, one or two integral membrane proteins, and, in the case of uptake systems, a periplasmic solute binding protein with a N-terminal export-signal sequence. These genes are generally arranged as an operon. Of the 430 ABC transport system genes predicted in the whole genome, over half (235) are found on pSymB. This constitutes 17% of the total pSymB coding capacity. Almost half of the 64 pSymB ABC-transporter systems (some with missing components) are predicted to transport sugars (29), including previously identified lactose and trehalose/maltose transporters (*thuRE-FGK*). Other predicted solutes include iron (4), amino acids (6), peptides and oligopeptides (6), spermidine/putricine (2), sulfate (1), aliphatic sulfonate (1), phosphate (1), choline (1), glycerol-3-phosphate (1), rhizopine (1) and taurine (1).

In addition to the ABC-family of transporters, there are transporter proteins belonging to the Major Facilator Superfamily including the C_4-dicarboxylate permease DctA, and possible nitrate (SMb20436), sulfate (SMb20070) and xanthine/uracil (SMb21281) permeases. Although a number of genes show similarity to C_4-dicarboxylate transporters, they are unlikely to transport these substances, since mutation of *dctA* has been shown to abolish C_4-dicarboxylate transport in *S. meliloti* (Watson *et al.* 1988). Transmembrane efflux proteins were also identified, several of which appear to be involved in exporting toxic compounds from the bacterium (SMb20071, SMb20338, SMb20345, SMb21575).

The requirement for specific transport systems in the production of normal N_2-fixing nodules has already been demonstrated for the *dct*, *bacA*, and *pho* loci, highlighting the important role of pSymB in the infection process and in endosymbiosis. The wealth of transport systems uncovered by sequence analysis suggests that pSymB may play a broader role in adaptation of the bacterium to the local environment, both as a free-living saprophyte and as an endosymbiont with host plants.

Transcriptional regulators. We have identified 134 ORFs on pSymB as transcriptional regulators. Included are the response regulators from two component systems, for which there are an additional 15 sensor histidine kinase-like proteins. Nineteen regulators belong to the LysR family which activate transcription in response to co-inducers, 21 belong to the GntR family, 16 belong to the LacI/GalR family, and the remainder belong to the TetR, AraC, ArsR, AsnC, DeoR, MerR, SorC families. Based on sequence homology, transcription direction and relative location, we have assigned gene names to a number of the LysR regulators (GstR, CbbR, PcaQ, GbpR), since these activators are generally transcribed divergently from the genes they regulate (Schell 1993). GntR proteins bind to promoter regions and negatively regulate transcription (Haydon and Guest 1991). In addition to the above regulators, we have identified four genes coding for RpoE-like proteins (SMb21484, SMb20592, SMb20531, and SMb20030), alternative sigma factors related to the extra-cytoplasmic function (ECF) subfamily of bacterial RNA polymerase sigma factors (Missiakas and Raina 1998).

Nodulation and nitrogen metabolism. It was known that genes involved in nodulation and in N_2-fixation are mainly localized on the pSymA replicon; only a few genes involved in these processes were found on pSymB. Previously identified and functionally interchangeable copies of *nodP* and *nodQ* are found on pSymB and pSymA (Schwedock and Long 1992). The *nfeD* gene is involved in nodulation competitiveness (Garcia-Rodriguez and Toro 2000). The SMb20472 protein is 65% identical to NolO, a protein involve in production of Nod factors, however, this is likely a protein with carbamoyl transferase activity rather than a protein with a direct role in nodulation.

Several genes involved in nitrate/nitrite reduction were identified, including a potential nitrate/nitrite response regulator (SMb20077-SMb20078), a periplasmic nitrate reductase (SMb20997) similar to NnuR, and a nitrate transporter (SMb21114). A nitrate/nitrite reductase and a siroheme synthase for nitrate assimilation (Lin, Goldman and Stewart 1993) is encoded by a four-gene cluster (SMb20987-SMb20984). One of three glutatmine synthetase structural genes in *S. meliloti* is located on SymB. Transcription of *glnII* requires the RpoN sigma factor and the NtrC transcriptional activator, both of which are coded by chromosomal genes (Shatters, Somerville and Kahn 1989).

Protective response and antibiotic resistance. There are a number of pSymB genes that may be involved in detoxifying reactions. There are two non-heme haloperoxidases (SMb20054 and SMb20860) which are possibly involved in dehalogenation reactions (van Pée 1996), and two glutathione *S*-transferase genes (SMb20005, SMb21149) of which there are a total of sixteen in the genome. Several genes may be involved in antibiotic resistance. Three genes (SMb20345, SMb20346, and SMb20698) putatively encode multidrug efflux permeases. The genes *acrE* and *acrF* encode putative acriflavin resistance proteins, which also have homology to other transmembrane multidrug efflux proteins (Okusu, Ma and Nikaido 1996). A putative *ampC* coding for β-lactamase, and *aacC4*, an aminoglycoside 6'-N-acetyltransferase gene rendering resistance to amikacin, were also identified. The hypothetical gene SMb21154 codes for a protein that belongs to the bleomycin resistance protein family.

Genes that appear to play a role in responding to osmotic stress protection include two trehalose synthases (SMb20099 and SMb20574; there is a third on pSymA) and a previously reported trehalose/maltose transport gene cluster. In *S. meliloti*, trehalose acts as an osmolyte. A metallo-regulatory gene coding for HmrR2 of the MerR family, is located next to the previously identified *atcU2* gene, whose product appears to be a copper export ATPase rendering resistance to copper. A similar gene pair is also located on the pSymA replicon.

There are only eight genes assigned functions in DNA modification or degradation (compared to 57 on the chromosome, 10 on pSymA), four DNA ligases (SMb20008, SMb206868, SMb20912, SMb21044), a DNA topoisomerase I (SMb21445), a methylated-DNA-protein-cysteine methyltransferase (SMb20708), a 3-methyladenine DNA glycosylase (SMb20709), and an exodeoxyribonuclease III (*xthA4*).

Dehydrogenases, oxidoreductases, and sugar kinases. A large number of genes encoding proteins potentially involved in oxidative metabolism were located on pSymB. The numbers of pSymB genes predicted to encode dehydrogenases (68), oxidoreductases (42), and dehydratases (19) are similar to the proportions of these genes predicted for the chromosome and pSymA. However, 10 predicted sugar kinase genes were located on pSymB; this contrasts the single putative sugar kinase on pSymA and eleven on the chromosome. Several of the pSymB sugar kinase genes (SMb20852, SMb21217, SMb21373) appear to be part of sugar catabolic gene clusters which include ABC transport genes.

1.4 Conclusion

The 1,683,333 base pair size of the pSymB replicon is similar to that predicted from previous genetic and restriction analyses (Honeycutt, McClelland and Sobral 1993; Charles and Finan 1990). The size of this replicon is comparable to the entire genomes of *Haemophilus influenzae* (1.8 Mb; Fleischmann *et al.* 1995) and *Methanococcus jannaschii* (1.66 Mb; Bult *et al.* 1996). Our annotation revealed that the gene density of pSymB is similar to the *S. meliloti* chromosome and to the density of other bacteria genomes (1 ORF per 1.1 kb). Moreover, with few exceptions, we did not find evidence of single genes or gene clusters carrying nonsense or other mutations suggestive of regions on their way to being eliminated from the genome. This is interesting, as our previous deletion studies revealed that much of the pSymB replicon is dispensable for growth in minimal medium in the laboratory. We identified two loci, the $\text{tRNA}^{\text{Arg}}_{\text{CCG}}$ gene and the *minCDE* genes, which are likely to be essential for growth of the bacterium. The $\text{tRNA}^{\text{Arg}}_{\text{CCG}}$ gene lies within a region of pSymB that could not be deleted in previous experiments (Charles and Finan 1991).

Two major observations to emerge from an analysis of the annotated pSymB sequence are the large number of solute transport systems and genes involved in polysaccharide synthesis. In addition to these, we have observed many genes that have potential catabolic activities, such as alcohol dehydrogenases. Thus, it appears that pSymB endows the bacteria with the ability to take up and presumably oxidize many different compounds from the soil environment. The presence of pathways for the oxidation of methanol and plant degradation products, such as protocatechuate, is consistent with this hypothesis. While it is noticeable that there are few if any pSymB genes which play a direct role in nodulation and symbiotic N_2-fixation, we note that pSymB does play a role in adaptation to the endosymbiotic lifestyle as emphasized by the fact that mutations in the *exo, dct, pho*, and *bacA* genes abolish symbiotic nitrogen fixation. Additionally, pSymB codes for several detoxification and antibiotic resistance functions. Hence, we envisage pSymB as playing an important role in the survival of the bacterium under the presumably diverse nutritional living conditions encountered in the soil and rhizosphere. The increased accessibility to carbon sources, and increased surface variability may point to the particular importance of pSymB in enhancement of competitive abilities of *S. meliloti* in the natural habitat.

The sequence of the pSymB replicon revealed that *S. meliloti* carries many more putative polysaccharide synthesis genes than previously envisioned. This surprising wealth of genes encoding cell surface polysaccharides may reflect the very different conditions (such as desiccation and starvation) and environments *S. meliloti* has had to adapt to, *e.g.* soil, rhizosphere and legume nodule. Additionally, these polysaccharides may be important for root nodule invasion, as shown for the previously characterized exopolysaccharide synthesis gene clusters. The actual polysaccharides synthesized by most of the newly identified gene clusters are unknown. It will be of interest to determine the conditions under which these gene clusters are transcribed and how they cooperate functionally. It has been shown that some key components of the surface carbohydrate export machinery can be involved in the export of more than one polysaccharide structure (Whitfield, Amor and Köplin 1997; Feldman *et al.* 1999). The absence of key genes in some of the new clusters encoding Wzy-dependent export machineries may indicate that this phenomenon also occurs in *S. meliloti.*

Almost all of the *S. meliloti* genes required for cell growth and viability are located on the 3.6 Mb chromosome (Galibert *et al.* 2001). However, the presence on pSymB of the single essential genomic copy of the $\mathrm{tRNA}^{\mathrm{Arg}}_{\mathrm{CCG}}$ gene, and other likely essential genes such as *minCDE* and the *asn* genes, clearly suggest that pSymB is indispensable to the cell and hence can justifiably be viewed as a second chromosome. We interpret the biased distribution of RIME and A, B, C palindromic elements (Østeras, Stanley and Finan 1995; Østeras *et al.* 1998) on the three replicons together with the G+C content of the replicons as evidence that pSymB was acquired by an ancestral *S. meliloti* prior to pSymA (Galibert *et al.* 2001). This is reminiscent of the observations concerning the two chromosomes of *Vibrio* species where the 2.9 Mb Chromosome 1 carries most genes required for cell growth and viability and the 1 Mb Chromosome 2 also carries a few essential genes (Heidelberg *et al.* 2000; Yamaichi *et al.* 1999).

While the pSymB nucleotide sequence has revealed a wealth of new genes, for the most part, the precise biological functions of these genes remain to be determined. The identification of these functions will lead to a clearer understanding of the interaction of *S. meliloti* with plants and, more generally, how this bacteria lives and survives in the soil environment.

1.5 Acknowledgements

This work was a joint collaboration between Turlough M. Finan[a], Stefan Weidner[b], Kim Wong[a], Jens Buhrmester[b], Patrick Chain[a], Frank J. Vorhölter[b], Ismael Hernandez-Lucas[a], Anke Becker[b], Alison Cowie[a], Jérôme Gouzy[c], Brian Golding[a], and Alfred Pühler[b].

Work at McMaster University was supported by NSERC Research, Strategic and Genomics grants to T.M.F. and B.G. Work in Bielefeld was supported by grants from the Bundesministerium für Forschung und Technologie (0311752) to A.P. We thank F. Barloy-Hubler, D. Capela and F. Galibert for providing us the pSymB BAC clones, H. Voss and B. Remmel from Lion Bioscience AG Heidelberg (Germany) and W. Arnold, W. Engelhardt and F. Hecht from IIT Biotech GmbH Bielefeld.

[a] Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 4K1 Canada.

[b] Universität Bielefeld, Fakulätt für Biologie, Lehrstuhl für Genetik, Universitätsstr. 25, D-33615 Bielefeld, Germany.

[c] Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes, UMR215, Chemin de Borde Rouge, BP27, F-31326 Castanet Tolosan, France.

Figure 1.1: Map of the *Sinorhizobium meliloti* strain Rm1021 pSymB megaplasmid. The inner circle displays ORFs on the leading (*red*) and lagging (*green*) strands of pSymB. The outer circle shows predicted gene regions encoding transcriptional regulators (*pink*), ABC transport systems (*yellow*), and genes involved in polysaccharide biosynthesis (*orange*). The positions of specific genes or sets of genes (*e.g. exp*, *exs*, and *exo*) are also shown on the outer edge of the map.

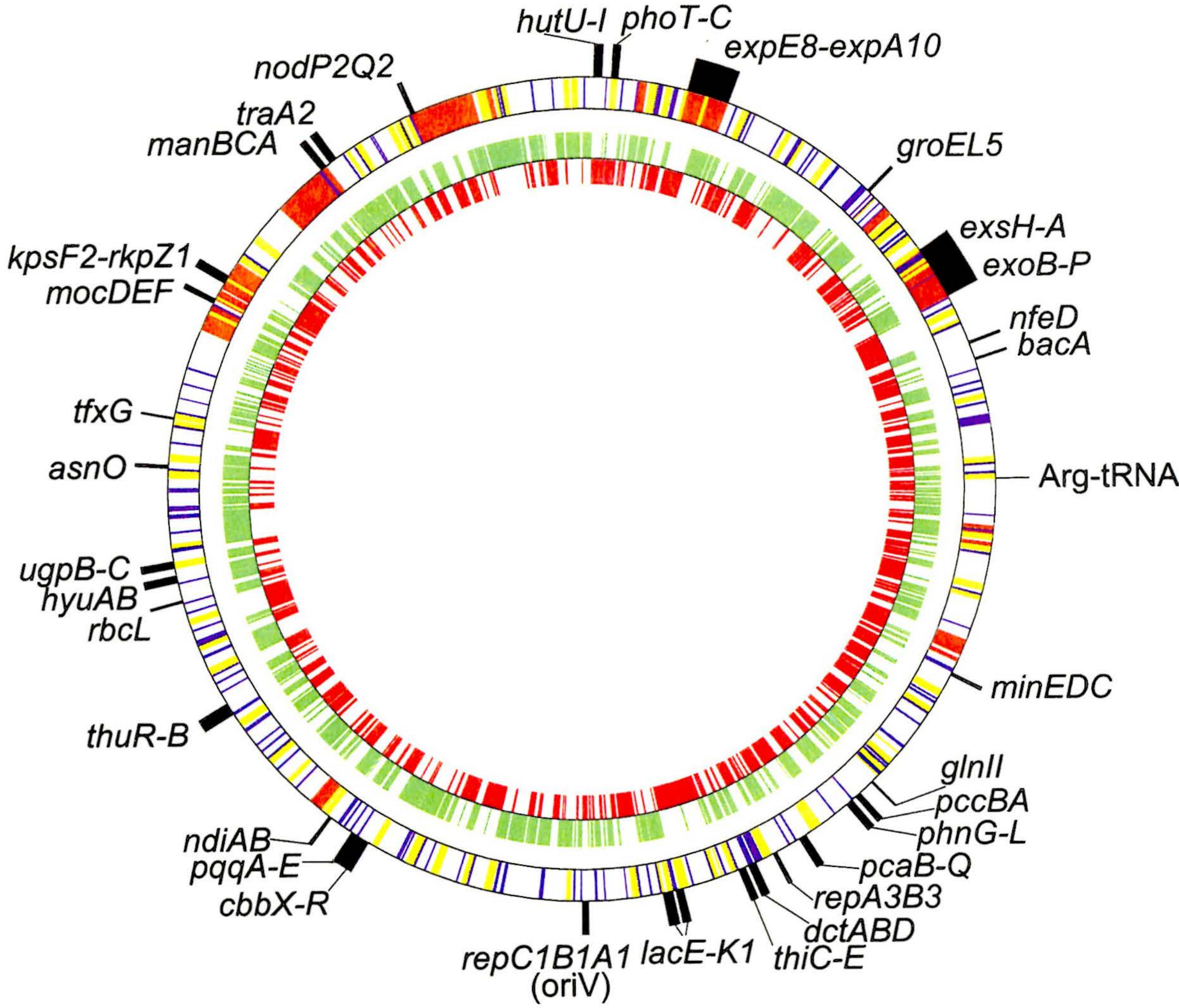

hutU-I
phoT-C
expE8-expA10
nodP2Q2
traA2
manBCA
groEL5
exsH-A
exoB-P
kpsF2-rkpZ1
mocDEF
nfeD
bacA
tfxG
asnO
Arg-tRNA
ugpB-C
hyuAB
rbcL
minEDC
thuR-B
glnII
pccBA
phnG-L
ndiAB
pqqA-E
cbbX-R
pcaB-Q
repA3B3
dctABD
repC1B1A1
(oriV)
lacE-K1
thiC-E

Table 1.1: Overview on cell surface carbohydrate synthesis gene clusters of pSymB. The 11 gene clusters contain 188 predicted genes and have a total size of 223 kb. Hence over 12% of the genes on pSymB are involved in the synthesis of cell surface carbohydrates.

Cluster	Size (kb)	Gene Number	ORFs	Surface Carbohydrate
1	14	11	SMb2080 to SMb20816	Possibly LPS core/lipid A
2	26	27	SMb20821 to SMb21013	Possibly CPS
3	40	33	SMb21050 to SMb21082	Unidentified
4	42	33	SMb21223 to SMb21256	Unidentified
5	6	4	SMb21188 to SMb21191	Unidentified
6	28	21	*expE8* to *expA10*	EPS II
7	5	5	SMb21581 to SMb21585	Unidentified
8	25	23	*exsH* to *exoP*	EPS I
9	13	10	SMb21499 to SMb2150, SMb21512 to SMb21513	Unidentified
10	14	14	SMb2150, SMb21512 to SMb21513	Unidentified
11	9	8	SMb20238 to SMb20245	Unidentified

Chapter 2

Dinucleotide compositional analysis of *Sinorhizobium meliloti* using the genome signature: distinguishing chromosomes and plasmids

Wong, K., T.M. Finan, and G B. Golding (2002) *Funct Integr Genomics*, in press.

2.1 Introduction

It is well recognized that the nucleotide composition of a genome is non-random; elements contributing to this heterogeneity include distinct regions high in G+C or A+T (e.g. isochores, Bernardi *et al* 1985), dispersed and tandem repeated sequences, and transposable elements. Coding regions also have different compositions than non-coding regions (Aota and Ikemura 1986; Muto and Osawa 1987), and even strand composition may be biased (Wu and Maeda 1987). Thus, local compositional variations make it difficult to generalize about a whole genome based on analyses of small regions of DNA.

Genomes of organisms representing all domains of life have been sequenced, allowing in-depth compositional and comparative analysis of these genomes. Examination of the frequencies of short oligonucleotides (di-, tri- and tetranucleotides) in prokaryotic, eukaryotic, and mitochondrial DNA sequences has revealed consistent patterns of oligonucleotide biases, some of which are common to all groups. However, when the relative frequencies of all dinucleotides are considered together, the pattern of dinucleotide bias is unique to each species (Nussinov 1984b; Burge, Campbell and Karlin 1992; Karlin and Ladunga 1994; Karlin, Ladunga and Blaisdell 1994; Karlin and Mrázek 1997; Karlin, Mrázek and Campbell 1997; Campbell, Mrázek and Karlin 1999). Unlike G+C content, dinucleotide biases tend to be consistent throughout a genome, in both coding and non-coding DNA (Karlin and Mrazek 1996), giving a genome-wide perspective of the patterns of nucleotide composition within a genome.

The preference or avoidance of specific dinucleotides was first quantified by Bird (1980) who observed a CG dinucleotide (or CpG) suppression in vertebrate sequences. The frequency of the CG dinucleotide is up to five-fold lower than the expected frequency based on C and G mononucleotide frequencies. It has been suggested that the high mutability of methylated cytosine to thymine due to deamination is responsible for the CG under-representation in these organisms (Bird 1980). Evidence for this is given by the existence of "CpG islands," G+C-rich DNA sequences of variable length which are abundant in unmethylated CG dinucleotides (Bird 1986). More recent studies establish that the CG suppression observed in vertebrates is also prevalent in fungi, plants, protists, and some bacteria (Cardon *et al.* 1994). Considering that mitochondria and bacteria lack the appropriate DNA methylases, the CG suppression observed in these organisms cannot be explained by the methylation/deamination/mutation hypothesis. Additionally, this hypothesis cannot account for other dinucleotide biases, such as the TA dinucleotide suppression. Rather, it has been suggested that dinucleotide biases reflect the avoidance of unfavorable base-step conformations and stacking energies (Nussinov 1984a; Nussinov 1984b). Since dinucleotide biases are a genome-wide property, it has also been suggested that the mutational biases of the modification, replication and repair machinery play a role in the generation and maintenance of species-specific dinucleotide biases (Karlin, Mrázek and Campbell 1997; Karlin and Ladunga 1994).

Karlin and his colleagues used dinucleotide relative abundance, the observed frequency of a given dinucleotide relative to the expected frequency based on base composition, to compare genomes (Burge, Campbell and Karlin 1992; Karlin, Ladunga and Blaisdell 1994; Karlin and Burge 1995; Karlin, Mrázek and Campbell 1997). The set of 16 dinucleotide relative abundance values is referred to as the *dinucleotide relative abundance profile* of an organism. The difference between two profiles is the dinucleotide relative abundance distance, or δ^*-distance. Because dinucleotide relative abundance profiles are unique to each organism, and within-species δ^*-distances between disjoint 50-kb regions of a replicon are more similar than between-species δ^*-distances, the dinucleotide relative abundance profile has been termed the *genome signature*. In addition, organisms that are closely related, as determined by 16S rDNA analysis, generally have more similar genome signatures than more distantly related organisms. Analysis of plasmids and chromosomes has shown that δ^*-distances tend to be small (but not necessarily the smallest) between plasmids and natural host chromosomes (Campbell, Mrázek and Karlin 1999). Thus, plasmids generally tend to track host chromosomal signatures.

In this paper, dinucleotide relative abundance profiles and δ^*-distances have been used to characterize the genome of *Sinorhizobium meliloti*. This α-proteobacterium is able to carry out endosymbiotic nitrogen fixation in nodules of host leguminous plants. The genome consists of three replicons: megaplasmids pSymA and pSymB, and one chromosome. The pSymA megaplasmid (1.4 Mb) has long been known to carry genes essential for symbiotic nitrogen fixation and root nodulation (Banfalvi *et al.* 1981; Rosenberg *et al.* 1981). The pSymB megaplasmid (1.7 Mb) also carries genes essential for the establishment of a successful endosymbiotic relationship with host legumes (Finan *et al.* 1986; Hynes *et al.* 1986).

The complete genome of *S. meliloti* has recently been sequenced and annotated (Galibert *et al.* 2001; Barnett *et al.* 2001; Finan *et al.* 2001; Capela *et al.* 2001). Although pSymB and pSymA both play roles in symbiosis, the difference in the G+C contents of the replicons imply that their evolutionary histories are different; the G+C content of pSymA (60.4%) is lower than the chromosome (62.7%) and pSymB (62.4%). Because of this difference, it has been suggested that pSymA had been acquired much later in evolution than pSymB (Galibert *et al.* 2001). However, the similarity in G+C content alone does not necessarily reflect the degree of relatedness between sequences. Because dinucleotide frequencies reflect restrictions in DNA conformation and mutational biases of DNA modification, replication, and repair enzymes, the application of genome signatures and δ^*-distances in this paper give a more precise picture of the compositional differences and similarities between the replicons than base composition alone. The results presented here demonstrate that the pSymB genome signature is chromosome-like and, in this respect, that the pSymB replicon is atypical of other α-proteobacterial plasmids. Taken together with previously observed chromosome-like features, genome signatures support the argument that pSymB should be considered a chromosome rather than a plasmid.

2.2 Methods

Sequence data. Sequences were downloaded from the National Center for Biotechnology Information (NCBI) website at http://www.ncbi.nlm.nih.gov as of August, 2001. Organisms were selected based on their close relationship to *S. meliloti* or similar soil habitat. Complete genome sequences from the following organisms were used in the analysis: *Mesorhizobium loti* (chromosome, plasmids pMLa and pMLb), *Agrobacterium tumefaciens* C58 (circular chromosome, linear chromosome, plasmids pAT and pTi; Cereon Genomics), *Bacillus subtillis, Pseudomonas aeruginosa, Escherichia coli* K12, *Caulobacter crescentus, Haemophilus influenzae, Deinococcus radiodurans* (chromosomes I and II, and plasmids MP1 and CP1), *Helicobacter pylori, Mycobacterium tuberculosis, Synechocystis* sp. PCC 6803, *Rickettsia prowazekii, Vibrio cholerae* (chromosomes I and II), *Thermotoga maritima*, and *Halobacterium* sp. NRC-1 (chromosome and plasmids pNRC100 and pNRC200). The *Brucella melitensis* genome sequence became available in December 2001. Complete plasmid sequences from the following organisms were used in analysis: *Rhizobium sp.* NGR234 (pNGR234a), *Agrobacterium rhizogenes* (pRi1724), *Lactococcus lactis* (pMRC01), *Yersinia pestis* (pMT1), and *Sphingomonas aromaticivorans* (pNL1).

In addition, preliminary sequence data of *Rhodobacter sphaeroides* was downloaded from the DOE Joint Genome Institute at http://www.jgi.doe.gov. Six contigs greater than 100-kb from the *R. sphaeroides* genome and were available (February 2001), totaling 807,975 bp. The *R. sphaeroides* genome is composed of two chromosomes. However, because sequencing and assembly of contigs are incomplete, the sequences were joined to form one continuous sequence for dinucleotide analysis. Complete *S. meliloti* sequences for the chromosome, pSymA, and pSymB were downloaded from the *S. meliloti* sequencing consortium website at:

http://sequence.toulouse.inra.fr/rhime/Complete/doc/Complete.html.

Dinucleotide relative abundance values and δ^*-distance. Dinucleotide extremes were determined by calculating the observed frequency divided by the expected frequency of each dinucleotide. The set of 16 dinucleotide relative frequencies, $\{\rho^*_{XY}\}$, for a particular sequence has been termed the *dinucleotide relative abundance profile*, where

$$\rho^*_{XY} = f^*_{XY}/f^*_X f^*_Y$$

for all dinucleotides XY where f^*_X is the frequency of nucleotide X and f^*_{XY} is the frequency of dinucleotide XY (Burge, Campbell and Karlin 1992). To control for strand differences, the frequencies are computed from the sequence concatenated with its inverted complementary sequence. Statistical analysis from previous studies (Karlin, Mrázek and Campbell 1997) has determined the values of dinucleotide relative abundance values that represent statistically significant extremes for double-stranded 50-kb sequences, and they are applied in this work. Over-representation of a dinucleotide is indicated according to the following scheme: $1.23 \leq \rho^* < 1.30$ (marginally high), $1.30 \leq \rho^* < 1.50$ (very high), and $\rho^* \geq 1.50$ (extremely high); under-representation of a dinucleotide is indicated by: $0.70 < \rho^* \leq 0.78$ (marginally low), $0.50 < \rho^* \leq 0.70$ (very low), and $\rho^* \leq 0.50$ (extremely low). These values represent the extremes which would occur in a random, double-stranded 50-kb nucleotide sequence with the probabilities $P \leq 10^{-3}, P \leq 10^{-6}$ and $P \leq 10^{-9}$ for the marginally high/low, very high/low, and extremely high/low categories, respectively. Values $0.78 \leq \rho^* < 1.23$ are considered within the "normal" range.

The difference between two dinucleotide relative abundance profiles was calculated by the following formula:

$$\delta^*(f,g) = (1/16) \sum \mid \rho^*_{XY}(f) - \rho^*_{XY}(g) \mid *1000$$

where δ^* is the dinucleotide relative abundance distance, X and Y are nucleotides A, T, C and G, f and g denote the two sequences, and the sum extends over all nucleotides. This is the average absolute dinucleotide relative abundance difference, referred to as the δ^*-distance. Empirical, qualitative rankings of the δ^*-distances for 50-kb sequences were modified from Karlin *et al.* (1999), and an example of the relatedness reflected by the δ^*-distance is given in parentheses: closely similar, $\delta^* \leq 55$ (*E. coli* vs. *S. typhimurium*); moderately similar, $55 < \delta^* \leq 85$ (*E. coli* vs. *H. influenzae*); weakly similar, $85 < \delta^* \leq 115$ (*Sulfolobus* sp. vs. *M. jannaschii*); distantly similar, $115 < \delta^* \leq 145$ (human vs. *S. cerevisiae*); distant, $145 < \delta^* \leq 185$ (*E. coli* vs. *H. pylori*); very distant, $\delta^* > 185$ (human vs. *E. coli*).

Unless otherwise indicated, whole genome signatures were calculated from complete genome sequences and δ^*-distances for within- and between-species comparisons were calculated from non-overlapping 50-kb regions spanning the sequence of the replicon. Two-sample t-tests were performed to compare mean δ^*-distances of 50-kb regions.

2.3 Results

2.3.1 The *S. meliloti* pSymB replicon is atypical of other α-proteobacterial plasmids.

The dinucleotide relative abundance profile, or signature, was determined for each replicon in *S. meliloti* and completely sequenced chromosomes and plasmids from other α-proteobacteria (Table 2.1). Dinucleotide abundances of a sequence are calculated relative to expected values based on the actual nucleotide content of that sequence; therefore, the dinucleotide abundances of two sequences with different G+C contents can be compared. The plasmids have a distinctive pattern of dinucleotide extremes, when compared to those in the chromosomes. The chromosomal sequences, including the *S. meliloti* chromosome, tend to have an over-representation of CG, a common feature of halobacterial and proteobacterial chromosomes (Karlin, Mrázek and Campbell 1997). The dinucleotide biases of TA and AT are more extreme in the chromosomes than the plasmids, and biases are also observed for AC/GT, GA/TC, TT/AA, and GC in the chromosomes. The pSymB replicon clearly tracks the signature of the *S. meliloti* chromosome, as the relative abundance of each dinucleotide is not significantly different from the chromosome. Unlike pSymB, the pSymA replicon has a typical plasmid-like signature, marked by "very low" relative abundance of TA, "marginally high" relative abundance of AT, and the absence of a bias in AC/GT or GA/TC dinucleotides.

To quantify the differences in genome signatures among the three *S. meliloti* replicons, mean δ^*-distances were calculated from pairwise comparisons of 50-kb regions with other plasmids and chromosomes found in closely related species or species sharing a similar habitat (Table 2.2). In accordance with the previous observation that plasmids tend to track host chromosome signatures (Campbell, Mrázek and Karlin 1999), both pSymB and pSymA have δ^*-distances which are "closely similar" to the *S. meliloti* chromosome (29.7 and 49.4, respectively). The smallest δ^*-distance to the *S. meliloti* chromosome is pSymB, and *vice versa*. However, the δ^*-distances for pSymA contrast those of pSymB; the δ^*-distances between pSymA and the chromosome (49.4) and pSymA and pSymB (42.7) are "closely similar," however, all seven of the plasmids tested have smaller δ^*-distances to pSymA than does pSymB or the chromosome. Thus, the pSymB dinucleotide relative abundance profile is atypical of other α-proteobacterial plasmids, and more similar to that of the *S. meliloti* chromosome.

2.3.2 Comparison of δ^*-distances between chromosomes within a species suggest the pSymB replicon is a chromosome rather than a plasmid.

Additional evidence for the chromosome-like nature of the pSymB signature is provided by a δ^*-distance comparison of genome signatures among and between plasmids and chromosomes. In previous studies it was shown that (i) plasmid signatures tend to track host chromosomal sequences (Campbell, Mrázek and Karlin 1999), and (ii) sequences within the *same* replicon have smaller δ^*-distances than sequences from two different species (Karlin, Ladunga and Blaisdell 1994; Karlin and Burge 1995).

In organisms harboring more than one plasmid, comparisons can be made between plasmid-chromosome δ^*-distances and plasmid-plasmid δ^*-distances (Table 2.3). The standard errors for all mean δ^*-distances listed are less than 4.5, with the exception of the pTi within-plasmid δ^*-distance which has a standard error of 8.4. The mean δ^*-distance between the *M. loti* plasmids (21.4) is significantly smaller ($P < 0.01$) than the mean δ^*-distances between either plasmid and the *M. loti* chromosome (40.8 and 43.3 for pMLa and pMLb, respectively). The same relationship is also observed for the chromosome and plasmids of the archaebacterium *Halobacterium* sp. NRC-1 (Table 2.3). In contrast, the mean δ^*-distance between the *S. meliloti* chromosome and pSymB (29.7) is significantly lower ($P < 0.01, t = -43.4$) than that between pSymA and the *S. meliloti* chromosome (49.4) and between pSymA and pSymB ($P < 0.01, t = -23.0$). Consistent with previous observations (Campbell, Mrázek and Karlin 1999; Karlin, Ladunga and Blaisdell 1994; Karlin and Burge 1995), the within-chromosome δ^*-distances are significantly smaller than plasmid-chromosome δ^*-distances for all three species.

A similar relationship is observed with organisms comprised of two chromosomes as well as plasmids. In these cases, comparisons can be made between chromosome-chromosome δ^*-distances, plasmid-chromosome δ^*-distances, and plasmid-plasmid δ^*-distances. *Agrobacterium tumefaciens* has one circular chromosome, one linear chromosome, and two plasmids. The δ^*-distance of 37.8 between the two plasmids is significantly lower ($P < 0.01$) than the δ^*-distances between either plasmid to either chromosome (Table 2.3). The *Deinococcus radiodurans* genome is comprised of two chromosomes (2,649 kb and 412 kb), one megaplasmid, and one small plasmid. The 412 kb replicon was designated a second chromosome due to the presence of amino acid utilization and cell envelope formation genes, which were deemed "likely essential" (White *et al.* 1999). Chromosomes I and II comprise 80.5% and 12.6% of the entire genome, respectively. The signatures of the two chromosomes are very similar, as reflected by the small (30.2) δ^*-distance (Table 2.3). The megaplasmid (MP1) is "closely similar" to both of the chromosomes and the δ^*-distances between the megaplasmid and each chromosome (30.1 and 33.2) are not significantly different from the between-chromosome δ^*-distance at the 5% level. The small plasmid (CP1) is only "moderately similar" to each of the other replicons. This is analogous to the relationship amongst pSymB, pSymA, and the *S. meliloti* chromosome; however, MP1 does not carry any essential genes and only comprises 5.4% of the whole genome and thus was not designated a chromosome.

The mean δ^*-distance between pSymB and the *S. meliloti* chromosome (29.7) is not significantly different ($P > 0.05$) from the between-chromosome δ^*-distance of *D. radiodurans* (30.2) or the α-proteobacterium *B. melitensis* (30.1), and significantly lower than the *V. cholerae* between-chromosome δ^*-distance ($P = 0.03, t = -2.2$). Thus, the difference in relative dinucleotide frequencies between pSymB and the *S. meliloti* chromosome is generally at the same level or lower than that found between two chromosomes in the same organism. Only the *A. tumefaciens* between-chromosome δ^*-distance of 27.0 was smaller

($P < 0.01, t = 7.1$). Within-chromosome δ^*-distance values are comparable to between-chromosome values, ranging from 22.0 to 37.3, while within-plasmid values range from 15.0 to 44.7 (Table 2.3).

These data indicate that the application of genome signatures is a good indicator of the degree of relatedness of replicons not only between species, but within-species as well. Among replicons within the same organism, the similarity in genome signatures probably reflects long-term replication by the same polymerase and repair enzymes.

2.4 Discussion

Plasmids are considered "facultative" genetic elements, not essential for cell viability but they often carry genes that allow for adaptation to different environments, lifestyles, or stress conditions (Joset and Guespin-Michel 1994). The pSymB replicon, which carries many genes involved in small molecule transport in addition to those genes involved in the endosymbiotic lifestyle, appears to be specialized for adaptation to different environments. However, most of these genes are non-essential to cell viability, as a large portion of the pSymB replicon can be deleted without loss of viability (Charles and Finan 1991), nor does pSymB carry any ribosomal RNA (*rrn*) genes (Finan *et al.* 2001). Control of the replication initiation and inheritance by plasmid-encoded RepABC proteins is characteristic of α-proteobacterial plasmids, especially within the *Agrobacterium* and *Rhizobium* genera (Tabata, Hooykaas and Oka 1989; Palmer, Turner and Young 2000). However, there are exceptions; *repABC* is also found on the linear chromosome of *A. tumefaciens* C58, leading to the speculation that this chromosome was derived from a plasmid (Goodner *et al.* 2001). A *repABC* operon and a replication origin (*oriV*) on pSymB have been shown to allow autonomous plasmid replication and stable inheritance (Chain *et al.* 2000). There is, however, also evidence that the pSymB replicon may have another replication origin that is chromosome-like in nature. This region is AT-rich, and contains potential DnaA binding sites and a 13-mer sequence similar to the *C. crescentus oriC* (Margolin and Long 1993).

Considering the presence of *repABC* and that the overall role of pSymB appears to be adaptation rather than cell viability, pSymB appears plasmid-like. However, when other characteristics of pSymB are taken into consideration, the task of designating this replicon a chromosome or a plasmid becomes more difficult.

The suggestion that a bacterial genome contains more than one chromosome is not unprecedented. Traditionally, within a genome of multiple replicons, the designation of a replicon as a chromosome has been based on the presence of genes essential to cell viability, such as 16S rRNA genes, or essential housekeeping genes such as *dnaK*. The presence of multiple chromosomes in prokaryotic genomes was first reported for *Rhodobacter sphaeroides* 2.4.1 based on the finding that two rRNA cistrons and the gene coding for glyceraldehyde-3-phosphate, two "chromosomal" loci, are found on a second replicon (Suwanto and Kaplan 1989). Since then, several other bacteria, including *Deinococcus radiodurans* (White *et al.* 1999), *Vibrio* (Heidelberg *et al.* 2000; Yamaichi *et al.* 1999), *Brucella* (Michaux *et al.* 1993; Cheng and Lessie 1994), and *Burkholderia* (Rodley, Romling and Tummler 1995) species, have been reported to harbor multiple circular chromosomes. Aside from the presence of essential "chromosomal" genes, criteria that are used to differentiate chromosomes from plasmids are significant replicon size, non-self-transmissibility, and presence in all strains (Trucksis *et al.* 1998). It has also been suggested that replication machinery and evolutionary history should also be taken into account (Ng *et al.* 1998).

The *S. meliloti* pSymB replicon meets each of these criteria, while pSymA has characteristics typical of plasmids. The pSymB replicon comprises 25% of the whole genome, and is present in all *S. meliloti* strains. Unlike pSymA, sequence analysis of pSymB did not reveal characteristics of a self-transmissible plasmid (*e.g.* it lacks an *oriT* sequence and conjugative transfer genes). The only exception is a single copy of the *traA* gene, also found on pSymA. The codon usage of pSymB is very similar to that of the chromosome, while pSymA codon usage is notably different from the chromosome or pSymB (Galibert *et al.* 2001). Sequencing and annotation of pSymB has revealed the presence of a single, essential $\text{tRNA}^{\text{Arg}}_{\text{CCG}}$ gene and two loci involved in cell division: one of two genomic copies of *ftsK*, and the single-copy *minCDE* genes (Finan *et al.* 2001). Two *ftsK* genes and the *minCDE* genes are also found on the *M. loti* chromosome, outside of the transmissible "symbiotic island" (Kaneko *et al.* 2000). The presence of these genes on pSymB suggests that this replicon plays a central role in control of cell division and chromosome partitioning. No essential genes were identified on pSymA. Additionally, while pSymB has resisted all attempts, the pSymA megaplasmid has been successfully cured from *S. meliloti* strain 2011 (Oresnik *et al.* 2000).

Here, the application of the dinucleotide relative abundance analysis has shown that, beyond the presence of a few essential genes and similar G+C content, pSymB shares genome-wide characteristics with the chromosome of *S. meliloti* that are atypical of other α-proteobacterial plasmids, including pSymA. When compared to other δ^*-distances observed between chromosomes within the same organism, the difference between pSymB and the *S. meliloti* chromosome falls within the same level of similarity; this is indicative of a high degree of relatedness and/or long-term residence of pSymB in the *S. meliloti* genome.

The genome of the γ-proteobacterium *V. cholerae* is comprised of two circular chromosomes (Trucksis *et al.* 1998). Like pSymB, the smaller chromosome of *V. cholerae* has plasmid-like characteristics, such as an integron region. However, at 1.6 Mb, it represents

40% of the genome and also carries essential genes and thus was designated a chromosome (Heidelberg *et al.* 2000; Trucksis *et al.* 1998). Further support is given here by the similar dinucleotide relative abundance profiles of the two replicons. It has been suggested that the *V. cholerae* chromosome II was derived from a megaplasmid captured in an ancestral *Vibrio* which has acquired essential genes (Heidelberg *et al.* 2000). The pSymB replicon may also have a similar history, since it contains both plasmid-like features (an *oriV*, *repABC* genes, absence of *rrn* genes), chromosomal-like features (a tRNA gene, *min* and *ftsK* genes and large size), and a signature similar to the *S. meliloti* chromosome rather than other α-proteobacterial plasmids. Over evolutionary time, pSymB may have acquired genes essential to the organism's viability as well as a similar signature as the chromosome due to long-term residence in *S. meliloti* and exposure to mutational biases of its replication and repair machinery. Taken together, these functional and compositional characteristics indicate that pSymB is not merely an accessory genetic element, but a chromosome-like replicon.

2.5 Conclusion

It has already been recognized that pSymB has chromosome-like features; pSymB comprises a large proportion of the genome, carries essential genes, is non-self-transmissible, and strains cured of pSymB cannot be produced. In terms of nucleotide composition, not only is the resemblance of pSymB to the *S. meliloti* chromosome reflected in the similarity of their G+C contents, but more precisely by the dinucleotide relative abundances. It was shown here that pSymB dinucleotide extremes parallel those of α-proteobacterial chromosomes rather than those of other α-proteobacterial plasmids, as pSymA does. In addition, the level of variability in dinucleotide frequencies between chromosomes in the same organism corresponds to the amount of variability between pSymB and the *S. meliloti* chromosome. Collectively, these characteristics of pSymB justify the designation of this replicon as a second chromosome in *Sinorhizobium meliloti*.

2.6 Acknowledgements

This work was supported by Natural Sciences and Engineering Research Council (Canada) Research, Strategic, and Genomics grants to TMF and GBG.

Table 2.1: Dinucleotide relative abundance profiles of α-proteobacterial replicons. Italics indicate chromosomal sequences. *A. tumefaciens I* and *A. tumefaciens II* refer to the cirular and linear chromosomes, respectively. The plasmid sequences of α-proteobacteria are as follows: *Rhizobium* sp. NGR234 (pNGR234a), *M. loti* (pMLa, pMLb), *A. tumefaciens* C58 (pTi, pAT), *A. rhizogenes* (pRi1724), and *S. aromaticivorans* (pNL1). Significant over-representation of a dinucleotide is indicated by the following scheme: $1.23 \leq \rho^* < 1.30$ (marginally high, *green box*), $1.30 \leq \rho^* < 1.50$ (very high, *blue box*), and $\rho^* \geq 1.50$ (extrememly high, *black box*); under-representation of a dinucleotide is indicated by: $0.70 < \rho^* \leq 0.78$ (marginally low, *yellow box*), $0.50 < \rho^* \leq 0.70$ (very low, *magenta box*), and $\rho^* \leq 0.50$ (extremely low, *red box*).

Replicon	CG	GC	TA	AT	CC/ GG	TT/ AA	TG/ CA	AG/ CT	AC/ GT	GA/ TC	G+C (%)
S. meliloti	1.29	1.15	0.47	1.39	0.82	1.18	0.93	0.89	0.77	1.28	**62.7**
pSymB	1.27	1.15	0.48	1.38	0.82	1.16	0.94	0.90	0.78	1.27	**62.4**
pSymA	1.23	1.15	0.52	1.28	0.84	1.15	1.00	0.91	0.81	1.22	**60.4**
pNGR234a	1.20	1.18	0.55	1.22	0.84	1.17	1.03	0.91	0.81	1.17	**58.5**
pMLa	1.20	1.17	0.52	1.27	0.85	1.17	1.04	0.89	0.81	1.17	**59.3**
pMLb	1.20	1.18	0.50	1.25	0.84	1.20	1.03	0.90	0.80	1.17	**59.9**
pTi	1.22	1.14	0.56	1.23	0.86	1.17	1.00	0.89	0.80	1.20	**56.7**
pAT	1.21	1.15	0.53	1.24	0.86	1.18	1.04	0.87	0.81	1.17	**57.3**
pRi1724	1.23	1.15	0.52	1.22	0.85	1.19	1.00	0.90	0.79	1.21	**57.3**
pNL1	1.19	1.16	0.47	1.36	0.85	1.16	1.05	0.88	0.81	1.18	**62.2**
M .loti	1.23	1.21	0.44	1.41	0.81	1.17	1.05	0.87	0.79	1.18	**62.8**
A. tumefaciens I	1.24	1.21	0.47	1.37	0.86	1.26	1.04	0.82	0.75	1.14	**59.4**
A. tumefaciens II	1.22	1.20	0.48	1.39	0.87	1.24	1.05	0.82	0.76	1.14	**59.3**
B. melitensis I	1.20	1.30	0.52	1.36	0.88	1.30	1.08	0.81	0.70	1.06	**57.2**
B. melitensis II	1.18	1.31	0.53	1.39	0.89	1.29	1.11	0.80	0.70	1.03	**57.3**
C. crescentus	1.16	1.11	0.45	1.29	0.85	1.09	1.01	0.96	0.86	1.22	**67.2**
R. sphaeroides	1.17	1.12	0.38	1.57	0.84	0.97	0.99	0.99	0.76	1.31	**68.6**
R. prowazekii	0.77	1.53	0.98	0.98	1.03	1.05	1.02	1.06	0.86	0.91	**29.0**

Table 2.2: Comparison of mean δ^*-distances by *S. meliloti* replicon. Chromosomal sequences are indicated in italics. *A. tumefaciens I* and *A. tumefaciens II* refer to the cirular and linear chromosomes, respectively. Plasmid sequences are from the following organisms: *A. rhizogenes* (pRi1724), *S. aromaticovorans* (pNL1), *M. loti* (pMLa, pMLb), *Rhizobium* sp. NGR234 (pNGR234a), *A. tumefaciens* (pTi, pAT), and *Y. pestis* (pMT1). *S. meliloti* replicons are indicated in *bold.* The seven α-proteobacterial plasmids have smaller δ^*-distances to pSymA than pSymB does to pSymA.

Chromosome		pSymB		pSymA	
Sequence	δ^*-distance	Sequence	δ^*-distance	Sequence	δ^*-distance
pSymB	29.7	***S. meliloti***	29.7	pMLa	28.2
pSymA	49.4	**pSymA**	42.7	pNGR234a	30.5
pNL1	53.8	pNL1	47.3	pNL1	31.1
pTi	56.7	pTi	51.0	pAT	33.8
M. loti	57.9	pMLa	53.3	pTi	34.4
pRi1724	58.5	pRi1724	54.7	pMLb	34.6
pMLb	59.8	*M. loti*	55.4	pRi1724	39.8
pMLa	62.7	pNGR234a	57.1	*C. crescentus*	42.7
pNGR234a	63.6	pMLb	58.2	**pSymB**	42.7
pAT	65.6	pAT	59.6	*M. loti*	48.9
C. crescentus	72.4	*C. crescentus*	64.5	***S. meliloti***	49.4
A. tumefaciens I	72.4	*A. tumefaciens I*	72.2	*A. tumefaciens II*	65.4
A. tumefaciens II	73.1	*A. tumefaciens II*	72.3	*A. tumefaciens I*	66.3
R. sphaeroides	87.3	*R. sphaeroides*	83.8	*P. aeruginosa*	79.0
B. melitensis I	111.6	*P. aeruginosa*	110.8	*R. sphaeroides*	91.1
B. melitensis II	119.6	*B. melitensis I*	111.0	pMT1	97.0
P. aeruginosa	120.4	*B. melitensis II*	119.1	*B. melitensis I*	100.9
pMT1	139.4	pMT1	129.7	*B. melitensis II*	110.2

Table 2.3: Mean δ^*-distances between plasmids and chromosome. *Bold* numbers indicate mean δ^*-distances between 50-kb regions in plasmid and natural host chromosome(s). "Within-chromosome" and "Within-plasmid" refer to comparison of 50-kb regions within the same chromosome and plasmid, respecitively. "Between-chromosome" and "Between-plasmid" refers to comparisons between two different chromosomes or plasmids, respectively. All standard errors of mean δ^*-distances are less than 4.5 with the exception of the pTi "within-plasmid" δ^*-distance (S.E.=8.4). A "Within-plasmid" δ^*-distance was not determined for CP1, since the plasmid is too short. For "Between-chromosome" δ^*-distances, *asterisk* (*) indicates values significantly different (at the level of 5%) from the *S. meliloti* chromosome-pSymB mean δ^*-distance of 29.7. See text for *t*-test results for comparison of means.

Chromosome	Plasmid										Within-	Between-
	pSymA	pSymB	pMLa	pMLb	pAT	pTi	MP1	CP1	pNCR100	pNCR200	chromosome	chromosome
S. meliloti	**49.4**	**29.7**	59.8	62.7	65.6	56.7	162.4	183.1	146.8	147.5	26.6	
M.loti	48.9	55.4	**40.8**	**43.3**	43.9	55.8	137.9	157.0	182.7	179.0	31.1	
A. tumefaciens I	66.3	72.2	52.5	48.6	**51.9**	**68.2**	138.4	167.0	208.3	202.2	23.7	27.0*
A. tumefaciens II	65.4	72.3	50.4	47.4	**49.1**	**66.9**	133.8	160.8	208.7	202.8	27.2	
D. radiodurans I	122.1	151.2	107.7	100.8	106.4	120.9	**30.1**	**81.3**	190.1	193.9	22.0	30.2
D. radiodurans II	122.7	150.9	108.9	102.9	108.5	121.3	**33.2**	**81.6**	190.0	194.4	37.3	
V. cholerae I	127.7	158.0	110.9	108.3	108.9	121.4	65.9	107.8	199.5	204.7	28.9	30.8*
V. cholerae II	135.1	165.1	117.4	114.3	114.1	127.3	70.1	112.2	210.4	214.6	30.4	
B. melitensis I	101.0	111.0	83.1	79.6	80.8	99.5	136.5	174.6	243.0	237.1	27.2	30.1
B. melitensis II	110.2	119.1	91.8	88.2	89.4	108.1	130.9	174.4	253.3	247.3	27.2	
Halobacterium sp.	176.0	171.6	190.1	198.0	187.6	179.3	231.9	214.9	**75.9**	**66.9**	32.9	
Within-plasmid	25.1	30.3	17.3	21.9	22.6	44.7	17.9	n/a	15.0	35.3		
Between-plasmid	42.7		21.4		37.8		66.5		25.2			

Chapter 3

A phylogenetic analysis of the pSymB replicon from the *Sinorhizobium meliloti* genome reveals a complex evolutionary history

3.1 Introduction

Five rhizobacterial genomes have been completely sequenced, including *Sinorhizobium meliloti*, *Agrobacterium tumefaciens*, *Mesorhizobium loti*, *Methylobacterium extorquens*, and *Brucella melitensis*. Several large plasmids belonging to other Rhizobiaceae have also been sequenced. The availability of sequence data for these closely related species allows a thorough examination of genome organization and evolution of these species.

The organization of α-proteobacterial genomes has been described as "unconventional" because members of this subgroup often have multiple chromosomes (occasionally linear) as well as plasmids and mega-sized plasmids (Jumas-Bilak *et al.* 1998). *Sinorhizobium meliloti* is an endosymbiotic N_2-fixing bacterium, whose genome is composed of a chromosome and two large replicons, pSymA (1.4 Mb) and pSymB (1.7 Mb), with GC contents of 62.7%, 62.4%, and 60.4%, respectively (Galibert *et al.* 2001). The completed genome sequence of *S. meliloti* has shown that, in addition to having a similar G+C content to the chromosome, the pSymB replicon carries essential genes (Finan *et al.* 2001), and comparative dinucleotide analysis has also revealed chromosome-like compositional properties of this replicon (Wong, Finan and Golding 2002). These findings have led to the suggestion that the pSymB replicon should be designated a second chromosome in *S. meliloti.*

Agrobacterium tumefaciens is plant pathogen causing crown gall disease, and the 5.67 Mb genome (strain C58) consists of one circular chromosome, one linear chromosome, and plasmids pAt and pTi (Allardet-Servent *et al.* 1993). Comparative analysis using NCBI's COG analysis and BLASTP identified 67% of the *A. tumefaciens* circular chromosomal genes as likely orthologs of *S. meliloti* chromosomal genes (Wood *et al.* 2001), and showed that global gene order is highly conserved between these two replicons (Goodner *et al.* 2001). These observations support the idea that these chromosomes were derived from a recent ancestral chromosome (Wood *et al.* 2001). A strong syntenic relationship between pSymB and any of the *A. tumefaciens* replicons, however, was not observed, so the origin and evolutionary history of pSymB remain obscure.

The amount of horizontal transfer, extent of rearrangement and duplication of pSymB genes is unknown. There is some evidence for horizontal transfer of insertion sequecne (IS) elements from Rhizobia to Agrobacterium (Deng, Gordon and Nester 1995), horizontal acquisition of glutamine synthetase in some species of Rhizobia (Turner and Young 2000) and the *nod* genes, key genes involved in symbiosis, are thought to have spread among rhizobia by horizontal gene transfer (Suominen *et al.* 2001; Young and Johnston 1989). IS elements are found on pSymB, in addition to Rhizobium-specific intergenic mosaic elements (RIMEs) and A, B, C palindromic elements (Galibert *et al.* 2001; Finan *et al.* 2001), which may contribute to recombinations, rearrangements and/or horizontal gene transfer of pSymB genes. However, the G+C content and signature dinucleotide frequencies have been maintained at relative chromosome-like levels, suggesting that the level of recent horizontal transfer is low.

Although COG analysis can identify likely orthologs (Tatusov *et al.* 2000) of pSymB gene products, whole-genome nearest phylogenetic neighbor analysis allows for a investigation of evolutionary relationships to the closest relatives, and possible identification of horizontally transferred genes. This approach to whole-genome analysis, termed *phylogenomics* (Eisen 1998), has been applied to predict protein function. Sicheritz-Ponten and Andersson (2001) have used phylogenomics to analyze the evolution of microbial proteomes, focusing on biochemical pathways and horizontal gene transfer. Whole-genome phylogenetic analysis may help to decipher the origin of pSymB and how it has evolved to its present-day form.

A potential limitation to a phylogenetic approach to whole-genome analysis is the lack of similar sequences to a particular protein, in which case the nearest phylogenetic neighbor cannot be determined conventionally. Contextual information, such as conservation of gene order across taxa, can be used to predict operons (Ermolaeva, White and Salzberg 2001), and gene function and functional interactions between genes (Huynen *et al.* 2000). Here, we explored the application of contextual information to predict the nearest neighbor in cases where phylogenetic trees cannot be constructed, or when the assignment of a nearest neighbor is not statistically supported by bootstrap values.

In addition to using phylogenetic analysis, native and foreign genes may be identified by estimating the amount of time a gene has been residing in a genome. Native genes generally have a base composition and codon usage that is characteristic of the entire genome (Médigue *et al.* 1991). Through the process of amelioration, horizontally transferred genes acquire the base composition and codon usage of the recipient genome, and therefore, the time required for amelioration can used as an estimate of the time of introgression (Lawrence and Ochman 1997). Here, the origin and evolution of the *S. meliloti* pSymB replicon was investigated using whole-genome nearest neighbor analysis and amelioration times. These methods reveal a complicated evolutionary history for this replicon.

3.2 Methods

Sequence data and determination of nearest phylogenetic neighbors. The complete *Sinorhizobium meliloti* genome, including all protein and nucleotide sequences for protein-coding regions were downloaded from the sequencing consortium website at http://sequence.toulouse.inra.fr/rhime/Complete/doc/Complete.html. Each of the 1570 pSymB gene products were analysed. Initially, similarity searches were conducted using NCBI's BLASTP alogrithm (August 23, 2001). A maximum of 50 hits with an Expect value less than 10^{-20} were used in the phylogenetic analysis, excluding any hits to other *S. meliloti* sequences. Those pSymB proteins with less than three significant hits were excluded from further analysis. Sequence similarity searches to *S. meliloti* sequences were performed using the stand-alone BLAST search against a local database containing only *S. meliloti* protein sequences. Hits with Expect values less than 10^{-20} were combined with the hits from the non-redundant NCBI protein database. The protein sequences were aligned using CLUSTALW (Thompson, Higgins and Gibson 1994). The data was bootstrapped 100 times using SEQBOOT (PHYLIP version 3.5c; Felsenstein 1993), and distance matrices were generated with PUZZLE (version 5.0; Strimmer and von Haeseler 1996) using the JTT model of substitution and a Gamma model of rate heterogeneity (with eight rate categories).

Phylogenetic relationships were determined using the neighbor-joining method (NEIGHBOR program from PHYLIP version 3.5c). That protein separated from the pSymB protein by the fewest number of nodes was defined as the nearest phylogenetic neighbor. If there was more than one possibility, then the nearest neighbor was chosen by determining the minimum total branch length from each of the possible nearest neighbors. The nearest neighbor was assigned if the corresponding bootstrap value was 95% or greater. If this was not the case, contextual information was used to help predict the nearest neighbor. Initially, the neighbor with the highest bootstrap value was chosen as the potential nearest neighbor. One of the following two conditions were satisfied in order for a nearest neighbor assignment to be made:

1. The gene is flanked on both sides by genes that have a nearest neighbor in the same species, with bootstrap values $\geq$ 95%;
2. On one side only, the gene is next to at least two successive genes that have a nearest neighbor in the same species, with bootstrap values $\geq$ 95%.

Nearest neighbor assignments were also made for proteins for which a tree could not be constructed due to the lack of sequences with BLAST E-values less than 10^{-20} (see above). In this case, a prediction was made only if condition (1) was satisfied.

Identification of potential operons. An operon can be described as a group of genes that are transcribed into a single mRNA molecule. We followed the method of Ermolaeva, White and Salzberg (2001) on prediction of operons, and their observation that operon gene order and orientation tend to be conserved across genomes. Because we have not identified operons experimentally, potential operons were identified based on the following criteria: a potential operon is a group of two or more successive genes with a nearest neighbor from the same species, coded on the same strand with intergenic regions less than or equal to 200 bp.

Dinucleotide analysis. The overall dinucleotide signature of pSymB was determined using methods previously described (Karlin and Burge 1995). For analysis of regions within pSymB, dinucleotide relative abundances were determined for 50-kb windows, with a 10-kb overlap, and δ^*-distances were determined between each 50-kb region and the overall dinucleotide signature.

Calculation of substitution rates for *S. meliloti*. Glutamine synthetase I (GSI) and glutamine synthetase II (GSII) have been shown to behave as good molecular clocks (Pesole *et al.* 1991). Using GSI and GSII sequences, Turner and Young (2000) have estimated divergence times among Rhizobia. They gave four estimates for the *S. meliloti-M. loti* split, and the average value of 265 million years was used to calculate substitution rates.

Substitution rates were determined by comparing 30 pairs of highly conserved nucleotide sequences from *S. meliloti* and *M. loti*: 25 ribosomal protein genes and 5 elongation factors (*tufA, tufB, sigA, fusA*, and *infA*). Synonymous and nonsynonynmous substitution rates (d_S and d_N) were calculated using `codeml` in the PAML program package (version 3.1; Yang 1997) which applies the codon substitution model of Goldman and Yang (1994). The transition to transversion ratio was estimated by the algorithm, and a global clock was assumed. Each of the 30 pairs have d_S less than 1.5. The mean d_S was 1.16 and the mean d_N was 0.08. Based on a divergence time of 265 million years, the following substitution rates were calculated for the first, second, and third codon positions, respectively: 0.0262%, 0.0160%, and 0.1626% substitutions per million years per lineage (Table 3.1). The mean transition to transversion ratio was estimated to be 1.6, and this value was used for the amelioration simulation.

Amelioration simulation. Muto and Osawa (1987) observed a linear correlation between the G+C content of a genome and the G+C content of each codon position in coding regions. Lawrence and Ochman (1997) describe these relationships as:

$$GC_{1st} = 0.615 \times GC_{Genome} + 26.9$$

$$GC_{2st} = 0.270 \times GC_{Genome} + 26.7$$

$$GC_{3st} = 1.692 \times GC_{Genome} - 32.3$$

where GC_{Genome} is the G+C content of the total genome. Lawrence and Ochman validated this model for $20\% \leq GC_{Genome} \leq 80\%$. They also determined that the use of under 1500 codons gave unreliable estimates of codon-specific G+C content and results in inaccurate estimates of the amelioration time. Therefore, pSymB genes were pooled according to similarity of G+C contents in the first, second, and third codon positions, with no fewer than 1500 codons (4.5 kb) per group. From these pooled genes, the G+C content of each codon position was determined, and the amelioration time was estimated for the entire pool.

We conducted simulations starting from random sequences generated with codon-position-specific G+C contents for each GC_{Genome} ranging from 20% to 80%. To simulate the mutational biases of the *S. meliloti* genome, Tamura's (1992) model of substitution was applied to each codon position, and each sequence was ameliorated in 1 Myr intervals for 1000 Myr toward the mean codon-position-specific G+C contents of the pSymB genes: 64.8%, 46.3%, and 76.8%. This was iterated 1000 times for each GC_{Genome}. Tamura's model takes into account the G+C content of the genome, and the transition to transversion ratio.

To estimate the amelioration time for each pool of genes, the G+C contents of the generated sequences were compared to the actual sequences. For each 1 Myr interval, the base composition of the generated sequence was determined, and the weighted least-squares difference was calculated between the G+C contents of each codon position in the generated sequence and the pooled pSymB genes. Because the third codon position ameliorates more quickly, while the first and second ameliorate more slowly, the weights 3.7, 1.3, and 5.0 were applied to the first, second, and third codon positions, respectively (Lawrence and Ochman 1998; Lawrence 1995). The mean least-squares difference was calculated for the 1000 iterates at each 1 Myr interval. The overall minimum least-squares difference was determined, and the time at which this minimum occurred was taken as the best estimate of the time of introgression of the pooled pSymB genes. The initial GC_{Genome} which gave the minimum least-squares difference was taken as the G+C content of the genes at the time of introgression.

3.3 Results

3.3.1 Nearest neighbor analysis.

Although the BLAST algorithm if often used to determine the similarity of a gene or protein to those in databases, it has been shown that the best BLAST hit is often not the nearest phylogenetic neighbor (Koski and Golding 2001). Therefore, the closest relative of *S. meliloti* pSymB protein sequences were determined using a phylogenetic method. Table 3.3 summarizes the nearest neighbor analysis.

In total, 510 nearest neighbor assignments were made. This represents approximately one-third of pSymB ORFs. Of the 33 species represented, 31 are bacterial, and two are archaeal belonging to the euryarchaeota phylum. Contextual information allowed us to assign nearest neighbors to 31 genes with nearest neighbor bootstrap values under 95%. For the 19 nearest neighbor predications based on contextual information and no phylogenetic information (due to limited sequence data), the BLAST output for those genes were inspected to confirm that a homologous protein in that species exists. Only one gene did not have a homolog in that predicted species, indicating that when combined with phylogenetic information, contextual information is a reliable method to predict the nearest neighbor of a protein.

Approximately 13% of pSymB genes have been involved in horizontal gene transfer. Of the pSymB proteins with assigned nearest neighbors, nearly half (45.1%) are nearest neighbors to *A. tumefaciens* proteins (Table 3.4). The majority of these *A. tumefaciens* proteins are encoded on the linear chromosome (140/230) while the remainder are encoded on the circular chromosome (66/230) and pAt (24/230). None are found on the pTi plasmid, despite the COG analysis indicating that many likely orthologs of pSymB proteins are present (Wood *et al.* 2001). *Mesorhizobium loti* proteins are the nearest neighbors for approximately 28% of the pSymB gene products, and nearly 11% are related to other *S. meliloti* proteins, which likely are results of gene duplications. The location of duplicated pSymB genes is shown in Figure 3.2. Another 3% have nearest neighbors from other members of the Rhizobiaceae family. These data suggest that *A. tumefaciens*, *M. loti*, and *S. meliloti* form a clade of closely related species, despite their divergence into pathogenic and symbiotic lifestyles. In total, 87% of pSymB genes are a result of a duplication or have a nearest neighbor in other Rhizobiaceae.

Approximately 10% of pSymB genes have non-rhizobial nearest neighbors, in the α-,

β-, and γ-proteobacterial subgroups (Table 3.4). In many instances a similar protein is absent in *A. tumefaciens* and/or *M. loti*, or other α-proteobacteria for which sequences are available. Alternatively, the non-rhizobial nearest neighbor is clustered within a clade which includes *S. meliloti*, *A. tumefaciens* and *M. loti* proteins. The remaining 3% of the pSymB proteins have nearest neighbors from the distantly related *Deinococcus* group, cyanobacteria (blue-green algae), six species of Gram positive bacteria, and two archaea (Table 3.4). The 13% of pSymB genes with these unusual phylogenetic relationships have likely been involved in horizontal transfer. Although the list of nearest neighbors represents a wide spectrum of bacterial species, all are pathogens and/or inhabitants of soils, sediments or water; their shared environment with *S. meliloti* may provide opportunities for genetic exchanges. There does not appear to be a correlation between the location of these genes and the location of IS elements, RIMEs and A, B, C palindromic elements (Figure 3.1).

Almost half of the pSymB genes involved in horizontal transfer are involved in the metabolism of small molecules, such as amino acid metabolism and sugar-nucleotide synthesis, or involved in the uptake of compounds such as iron, amino acids and carbohydrates (Figure 3.3). However, several of these genes are hypothetical and code for proteins of unknown function.

The pSymB replicon is known to play an important role in early stages of nodule formation and symbiotic nitrogen fixation. Our analysis shows that some of these genes have been involved in horizontal transfer. Only four proteins coded on pSymB have been identified as nodulation proteins, NodPQ, NfeD, and SMb20472, a putative nodulation protein belonging to the NodU family. Nearest neighbor analysis agrees with previous suggestions that the *nodPQ* genes have been duplicated on pSymA (Galibert *et al.* 2001), and a protein from *Rhizbium etli* is the nearest neighbor to NfeD, a protein involved in nodulation competetiveness. SMb20472 has likely been involved in horzontal transfer, since the nearest neighbor, with which it shares 62% identity, is a transferase from the cyanobacterium *Synechocystis* sp. PCC 6803. Genes involved in surface polysaccharide biosynthesis are important for successful nodule invasion (Hynes *et al.* 1986). In our analysis, a nearest neighbor was not assigned to many of these genes, however, some were identified as being involved in horizontal transfer. Nearest neighbors to these genes are from *C. vibrioides* and *P. aeruginosa*.

One of two β-galactosidase genes and a gene coding for a putative membrane protein (SMb20185) have nearest neighbors found in an archaeal species. Analysis of these phylogenies suggest that both of these genes were horizontally transferred into the *S. meliloti* genome from an archaea. The only two beta-galactosidase genes in *S. meliloti* are coded on pSymB. Phylogenetic analysis supports that the direction of transfer for each of these genes is from a distantly-related donor into the *S. meliloti* genome. One copy (*lacZ1*) appears to have been acquired from *Thermoanaerobacterium thermosulfurigenes* and the other (*lacZ2*) from the halophillic archaea *Haloferax alicantei.* Unlike the *lacZ1* gene, the *lacZ2* gene is not clustered on pSymB with lactose transporter genes, and the LacZ2 protein has significant similarity only to archaeal β-galactosidases.

A putative membrane protein (GenBank accession AAB85581.1) from the archaea *Methanothermobacter thermoautotrophicus* has been determined as the nearest neighbor of the putative solute-binding membrane protein encoded by SMb20185. This protein is part of a putative ABC-type transporter on pSymB. The SMb20185 protein has significant sequence similarity to other archaeal membrane proteins from *Pyrococcus* species, *Archaeoglobus fulgidus, Thermoplasma volcanium, Sulfolobus solfataricus*, and *Aeropyrum pernix*, as well as the bacterium *Thermotoga maritima.*

Local gene order is conserved in "patches" which likely contain at least one operon. Here we define a "patch" as two or more successive genes with a nearest neighbor from the same species. A patch may or may not be comprised of one or more potential operons. Two thirds of the genes with assigned nearest neighbors are arranged in patches. The distribution of these patches among different nearest neighbors is shown in Figure 3.4. Most (187/230 or 81%) of the pSymB genes encoding proteins with a nearest neighbor from *A. tumefaciens* are grouped in patches which range from two to nine genes. Comparison of gene order between *A. tumefaciens* and *S. meliloti* patches shows that there have been some local rearrangements since their divergence, however, gene order is conserved in most of these patches. Nearly two-thirds of these genes have nearest neighbors located on the *A. tumefaciens* linear chromosome, suggesting a common origin between the two replicons.

Nearest neighbors to *M. loti* genes are also arranged in patches, although to a lesser extent. Patches comprise 69% (99/144) of the genes with *M. loti* nearest neighbors (Figure 3.4) and they consist of 2 to 11 genes. Gene order is not well conserved. The exception is one patch of 11 genes (SMb20422 to SMb20432) located adjacent to a the IS element ISRm21, in which gene order is completely conserved. These genes code for an alcohol dehydrogenase (SMb20422), an aminotransferase (SMb20423), succinic semialdehyde dehydrogenase (SMb20424), transcriptional regulators (SMb20425, SMb20426), an amino acid ABC transporter (SMb20427 to SMb20430), arylmalonate decarboxylase (SMb20431), and threonine dehydratase (SMb20432).

Six patches comprise 35% of the genes that were determined to have *S. meliloti* nearest neighbors. Gene order is conserved in most of these patches, which implies that the duplication of these genes occurred as a single event rather than several independent events.

Six other species are represented in the remaining patches, three α-proteobacteria

(*S. fredii, R. etli, R .sphaeroides*), and three γ-proteobacteria (*P. aeruginosa, V .cholerae, P. multocida*). On average these patches consist of fewer genes than the *A. tumefaciens* patches, ranging from 2 to 5 genes per patch. Comparison of gene order between *S. meliloti* and these species showed that a rearrangement or insertion/deletion of a gene has occurred in only two of these patches.

Potential operons were identified within patches, and occasionally these operons were flanked by one or more genes that were in the same patch, but not included in an operon. Such cases are only observed for patches with nearest neighbors to *A. tumefaciens*, *M. loti*, and *S. meliloti* genes (Figure 3.4). Overall, few patches (21/86) consist of two or more potential operons, or no potential operons.

As stated above, pSymB genes with nearest neighbors from species outside of the Rhizobiaceae family are likely to have been involved in horizontal transfer. The patches with nearest neighbors from *R .sphaeroides*, *P. aeruginosa*, *V .cholerae*, and *P. multocida* are all comprised of one (or two) potential operons with no adjacent "single" genes. This, and conservation of gene order, indicate that genes that were horizontally transferred together likely comprise an operon. However, two-thirds of the genes involved in horizontal transfer are not part of patches or operons, and in many cases this is because only one nearest neighbor from a particular species has been identified (Table 3.4). These observations show that although horizontal transfer of operons may occur, the long-term result of horizontal transfer tends to predominantly result in horizontal transfer of single genes.

3.3.2 Simulation of amelioration.

For all 1570 protein-coding genes on pSymB, a computer simulation of amelioration was applied to determine the initial G+C content and the time required for a sequence to reach its current codon-position-specific G+C contents. The estimation was considered valid only if a minimum least-squares difference in the G+C contents (of all three codon positions) between the simulated data and the actual data was reached at a time between 0 to 1000 Myr. These genes are shown in light grey or red (if the time was less than 10 Myr) in the outer circle of Figure 3.1. The 725 genes for which a minimum between 0 and 1000 Myr was not reached were excluded from further analysis.

The average least-squares difference of 1000 iterates at each 1 Myr interval was used to produce amelioration curves. Figure 3.5a shows the amelioration curves for 4.5-kb sequences with initial G+C contents of 49% to 58%, with the least-squares differences calculated between the generated sequences and a pool of genes with the present-day G+C contents of 58.8%, 47.1%, and 64.5% at the first, second, and third codon positions. The simulation gave an estimate of 172 million years as the amelioration time, with the sequence having a G+C content of 53% at the time of introgression. According to the Muto and Osawa relationships, the G+C contents at each of the three codon positions were 59.4%, 41.0%, and 57.3%. In Figure 3.5a, the largest and smallest minima differ by less than 0.002, yet the time required to reach these minima range from 1 to 287 million years. At 172 Myr, the minimum least-squares difference was 0.00322, with a standard deviation of 0.00130. Two-sample t-tests showed that this minimum is not significantly different than those reached by sequences with initial G+C contents of 52% and 54% (at the 5% level of confidence), giving a range of 135 to 204 million years for the amelioration time. Figure 3.5b shows 10 iterates of the simulation for sequences with an initial G+C content of 53%. The simlulations demonstrate a large variance in the estimates of amelioration time. Each individual realization of evolutionary history shows a completely different level of G+C content over time.

Figure 3.5c shows the amelioration curves for sequences with initial G+C contents of

54% to 65%. The pool of genes to which these sequences were being compared had G+C contents of 62.4%, 50.6%, and 74.7% at the first, second, and third codon positions. The minimum least-squares difference, averaged over 1000 iterates, occurred after 586 million years with a sequence with an initial G+C content of 59% (63.1%, 42.6%, and 67.5% at the first, second, and third codon positions, respectively). However, the curves do not reach a minimum value. The apparent amelioration occurred by chance due to the high variation of the least-squares squares difference. This variation is demonstrated in Figure 3.5d (as well as Figure 3.5b), which shows 10 iterates of the simulation involving sequences with an initial G+C content of 59%. Since only one of these realizations might have occurred, the estimation of amelioration times may be highly variable and unreliable.

However, amelioration times can provide gross estimates of potential times of introgression for some genes. Figure 3.6 shows amelioration times for 845 pSymB genes. Genes that are predicted to have been acquired by pSymB less than 10 million years ago are distributed throughout the replicon in no apparent pattern (Figure 3.1). This group of genes comprises 148 of the 845 genes that were assigned an amelioration time, and the species to which the nearest neighbor belongs is shown in Table 3.5. The majority of these genes (61%) do not have a nearest neighbor assigned and proteins found in Rhizobiaceae make up 30% of this group. The remaining genes have distantly-related nearest neighbors and the predicted amelioration times of less than 10 Myr suggest that these genes were horizontally transferred recently into the *S. meliloti* genome. The *lacZ2* gene acquired from *Haloferax alicantei* is predicted to have been ameliorating for approximately 100 million years.

3.3.3 Dinucleotide analysis.

It has been shown that dinucleotide biases are consistant throughout a genome, and distinguishable between species (Burge, Campbell and Karlin 1992; Karlin, Ladunga and Blaisdell 1994; Karlin, Mrázek and Campbell 1997). A measure of dinucleotide bias for all 16 dinucleotides is the *dinucleotide relative abundance profile*, the frequency of a dinucleotide relative to expected values based on G+C content (Burge, Campbell and Karlin 1992). Since genomic profiles are unique to a species, this has been termed the *genome signature*. Dinucleotide analysis of overlapping 50-kb regions revealed regions with large differences in dinucleotide abundances, relative to the overall abundances and the average abundances in the 50-kb regions (Figure 3.2). The two regions with the largest differences from the overall signature contain groups of dispersed, small hypothetical ORFs, transposons, and genes involved in sugar-nucleotide biosynthesis and exopolysaccharide synthesis. In addition, the G+C content of this region is slightly below 60%. The unusual dinucleotide and base composition of this large region suggests that it may have a foreign origin. However, many of the duplicated pSymB genes are found in this region (Figure 3.2), suggesting that the small, hypothetical ORFs represent degenerating genes that may have been duplicated but not maintained on pSymB. Regions with the smallest differences include those with the *repA1B1C1* genes, *nodP2Q2*, and a large cluster of *exp* genes.

3.4 Discussion

Sequencing of the complete *S. meliloti* genome has revealed the mosaic nature of this genome, as demonstrated in the functions of the genes distibuted among the three replicons, G+C content and codon usage, and numbers and distribution of IS elements, RIMEs and A, B, C palindromic elements (Galibert *et al.* 2001). It has been argued that the pSymB replicon is a second chromosome in *S. meliloti* (Finan *et al.* 2001). Unlike the main chromosome, the pSymB replicon does not show a high degree of syntenty with an *A. tumefaciens* replicon, and orthologs are found on all four *A. tumefaciens* replicons (Goodner *et al.* 2001; Wood *et al.* 2001). Therefore, the origin and evolutionary history of pSymB are less obvious.

Several aspects were examined in an attempt to determine how pSymB might have evolved. Dinucleotide analysis shows regions with biases that deviate from the average abundances on pSymB, but since this analysis is limited to regions 50-kb or larger, methods which allow a finer gene-by-gene examination are required. A phylogenomic approach, whole-genome nearest neighbor analysis, provides evolutionary information, and the origins of genes and groups of genes can be examined. In addition to determining the most closely related genes, phylogenetic relationships indicate which genes may have been involved in horizontal gene transfer. The length of time a gene has spent in a genome can also point to horizontal transfer, and it has been argued that amelioration time corresponds to the time of introgression (Lawrence and Ochman 1997).

A nearest neighbor was assigned for one-third of the pSymB protein-coding genes. In addition to assignments of the nearest neighbor based on statistically significant data, we made additional assignments based on contextual information (10% of the nearest neighbors). Contextual information has also been used to predict protein function and functional interaction (Huynen *et al.* 2000); these predictions are also based on the observation that gene order is conserved in operons and across genomes. Here, it has been demonstrated that contextual information also can supplement phylogenetic information.

These data show that pSymB is truly mosaic; duplicated genes and genes with nearest neighbors to 32 other species are dispersed seemingly randomly throughout pSymB, and gene order is conserved only locally in small groups of genes. Because of this random distribution, and because the proportion of genes involved in horizontal transfer is small (13%), the dinucleotide signature remains similar to that of the main *S. meliloti* chromosome. The complexity of pSymB also is contrasted with the scarce number of IS elements and other repeat elements on pSymB. The proportion of pSymB genes acquired by horizontal transfer or horizontally transferred to other species is estimated here to be 13%. This value is comparable to other estimates of horizontally acquired genes in other species (Garcia-Vallve, Romeu and Palau 2000), which range from 1.5% to 14.5%. These estimates were determined using a statistical procedure which identified genes with unusual G+C content, codon usage and amino acid usage; gene position was also taken into consideration. This is a more stringent procedure than that employed by Lawrence and Ochman (1998), who used atypical G+C content and codon bias to identify horizontally acquired genes in *E. coli*. This method may, however, falsely identify native genes as horizontally acquired genes (Koski, Morton and Golding 2001; Wang 2001). Lawrence and Ochman (1998) estimate that approximately 18% of *E. coli* genes have been acquired since its divergence from *Salmonella* 100 million years ago, while Garcia-Vallve, Romeu and Palau (2000) estimate this value to be 9.6%. This concurs with the estimate of 10-15%, based on phylogenetic analysis (Koski, Morton and Golding 2001).

Although genes with nearest neighbors found in members of the Rhizobiaceae family were excluded here as potentially involved in horizontal transfer, there is evidence for horizontal transfer among rhizobia. It has been shown that an IS element has been horizontally transferred from *S. meliloti* to *A. tumefaciens* (Deng, Gordon and Nester 1995), and the incongruence of phylogenies based on *nod* genes with those based on 16*S* rDNA show that *nod* genes have been horizontally transferred among rhizobia (Suominen *et al.* 2001;

Young and Johnston 1989). It has also been suggested that glutamine synthetase II may have been horizontally acquired by *Bradyrhizobium japonicum*, *Mesorhizobium huakuii*, and *Rhizobium galegae* (Turner and Young 2000). Here, it was observed that the largest patch has *M. loti* genes as the nearest neighbors, and its location adjacent to a transposon indicates it was likely acquired intact from *M. loti*. In addition, the *A. tumefaciens* genes most closely related to pSymB genes are dispersed among three of the four *A. tumefaciens* replicons, and this is likely a result of horizontal transfer between these species as well as vertical transmission from a common ancestor. Our estimate of 13% of pSymB genes involved in horizontal transfer, therefore, may be a conservative estimate.

Genes involved in horizontal transfer that are found in patches comprise exactly one or two operons (Figure 3.4) in which gene order is conserved. This shows that the genes within an operon were transferred as a single unit, rather than in separate transfer events. Comparison of operon structures in complete microbial genomes has shown that, generally, gene order within operons was found to be less conserved with longer-term evolution and higher numbers of IS elements (Itoh *et al.* 1999). Therefore, given the conservation of gene order within the potentially horizontally transferred operons on pSymB, they have been involved in recent horizontal transfers. However, in *S. meliloti*, horizontal transfer of single genes has occurred more often than transfer of operons, since two-thirds of the predicted horizontally transferred genes are not found in potential operons.

Nearest neighbor analysis can indicate which genes may have a foreign origin, but does not provide an estimate of the time of introgression. The amelioration simulation applied here is a simple and flexible method that gives the time required for a sequence with a specific G+C content to reach its present-day codon-position-specific G+C contents, given the mutational biases of the genome. The resulting estimates have high variation, but give estimates of introgression times for some genes.

Not all genes predicted by nearest neighbor analysis to be involved in horizontal transfer were found to have amelioration times that would indicate a possible foreign origin, nor was there a strong agreement of amelioration times among genes within potential operons that have been involved in horizontal transfer (data not shown). In addition, the amelioration simulation was not successful for many genes, in that the least-squares difference increased with time, resulting in a minumum least-squares difference at 0 Myr, or, the least-squares difference decreased with time but did not reach a minimum even after 1000 million years of amelioration. Some problems with this method are that it may give innacurate amelioration times for (1) genes which have been horizontally acquired from a donor with a similar genomic G+C contents at each codon position, and (2) genes with unusual codon usage or G+C contents due to functional constraints (Jukes and Kimura 1984; Miyata and Yasunaga 1980). Therefore, the amelioration time does not appear to correlate with the actual time of a gene's residence in a genome, and estimates given by this method should not be used alone as an indicator of horizontal transfer.

There are conflicting views concerning the origins of the *S. meliloti* pSymA and pSymB replicons. Galibert *et al.* (2001) suggest that these replicons were acquired by an ancestral rhizobium, and that pSymA was acquired more recently than pSymB because overall G+C content and codon usage are distinctive from pSymB and the chromosome. These authors argue against a chromosomal origin for pSymB, citing the small number of insertion sequence elements and large proportion of unique genes. With the completion of sequencing of the *A. tumefaciens* genome, BLAST and COG analysis revealed a substantial number of orthologs between all three *S. meliloti* replicons and all four *A. tumefaciens* replicons (Goodner *et al.* 2001; Wood *et al.* 2001). Wood *et al.*suggest that ancestral pSymA and pSymB replicons were acquired prior to divergence of *A. tumefaciens* and *S. meliloti.*

In addition to having a large portion of the nearest neighbors to pSymB genes, most of which are organized in operons where local gene order is conserved, key genes on the *A. tumefaciens* linear chromosome involved in plasmid replication and cell division share a common origin with those on pSymB. Together, these observations suggest that pSymB and the *A. tumefaciens* linear chromosome have a common origin, which was a plasmid present in the last common ancestor of *S. meliloti* and *A. tumefaciens*.

Both the linear chromosome and pSymB have a *repABC* system of replication, which is common among α-proteobacterial plasmids (Tabata, Hooykaas and Oka 1989; Palmer, Turner and Young 2000). The *S. meliloti repA1B1* are involved in pSymB segregation, and *repC1* is involved in pSymB replication (Chain *et al.* 2000). The ability to self-replicate is an essential property of plasmids, and genes that confer this property must be present in order for a new plasmid to come into existence (Thomas 2000). Therefore, the *rep* genes found on a replicon are likely also to have been present in its ancestral plasmid. Our analysis showed that, indeed, the pSymB RepA1B1 proteins are nearest neighbors to the *repAB* gene products from the *A. tumefaciens* linear chromomsome (with bootstrap values of 98% and 78% for RepA1 and RepB1, respectively. The nearest neighbor analysis for RepC1 was inconclusive due to low bootstrap values.) The *min* genes are essential genes involved in cell division (de Boer, Crossley and Rothfield 1989). We found that the pSymB *min* genes also share a common origin with the *min* genes found on the linear chromosome, which are also likely to have been present in the ancestral plasmid due to their essential function.

Goodner *et al.* (2001) suggest that portions of the linear chromosome may have originated from an excision from the ancestral main chromosome, with subsequent insertion events and linearization. Given that pSymB and the *A. tumefaciens* linear chromosome have a common origin, it is likely that the excision occurred after the divergence of *S. meliloti* and *A. tumefaciens*, since this region is still intact in *S. meliloti*. COG analysis showed that

36% of genes on linear chromosome are orthologous to *S. meliloti* chromosomal genes, 25% are orthologous to pSymB genes, while only 12% are orthologous to pSymA genes (Wood *et al.* 2001). The remaining 27% were not orthologous to any *S. meliloti* genes, which suggests that significant gene loss and acquisition played a key role in the differentiation of these replicons. Since the divergence of these species, while the *S. meliloti* chromosome has remained fairly stable, the pSymB replicon has been substantially rearranged and horizontal acquisition of genes led to expansion of this replicon and therefore the *S. meliloti* genome. The reason for the instability of pSymB, relative to the main chromsome of *S. meliloti,* is unclear. It cannot be explained by the presence of IS or RIME elements, or A, B, C palindromic elements, since these make up a smaller proportion of pSymB than the chromosome.

3.5 Conclusion

It is evident that the pSymB replicon is mosaic in nature. Our analysis shows that there are pSymB genes with nearest neighbors found in archaea, *Deinococcus*, *Synechocystis*, Gram positive bacteria, and proteobacteria. It is probable that pSymB and the *A. tumefaciens* linear chromsome have a common origin, made less obvious by a complex series of rearrangements, gene duplications and horizontal acquisition of genes since the divergence of *S. meliloti* and *A. tumefaciens*. The pSymB replicon appears to be a patchwork of parts from several other genomes, some of which have likely been acquired recently through horizontal transfer. A minimum of 13% of pSymB genes have been involved in horizontal transfer. The majority of these genes are involved in small molecule metabolism or ABC-type transport, suggesting that the acquisition of genes on pSymB enabled *S. meliloti* to adapt to various niches or soil environments by transporting and utilizing various nutritional sources. In contrast, synteny between the main chromosomes of *A. tumefaciens* and *S. meliloti* is highly preserved throughout (Goodner *et al.* 2001; Wood *et al.* 2001), an indication of the recent divergence and stability of these chromosomes. Given this, it is evident that horizontal transfer, rearrangements, and duplications of pSymB genes has no doubt played a major role the evolution and adaptation of *S. meliloti*.

Figure 3.1: Map of pSymB showing genes with amelioration times less than 10 Myr and genes potentially involved in horizontal transfer. Shown on the outer circle are the genes for which amelioration was successful (*light grey*). Genes that are estimated to have been ameliorating for less than 10 million years are shown in *red*. The second circle from the outside shows the distribution of genes with non-rhizobial nearest neighbors (see figure legend) and rhizobial nearest neighbors (*light grey*). The third circle shows the location of IS elements (*green*), RIMEs (*purple*), A, B, and C palindromic elements (*pink*), and short partial IS elements and group II introns (*black*). The solid line shows the δ^*-distance of overlapping 50-kb regions from overall (outer dashed line) and the average (inner dashed line).

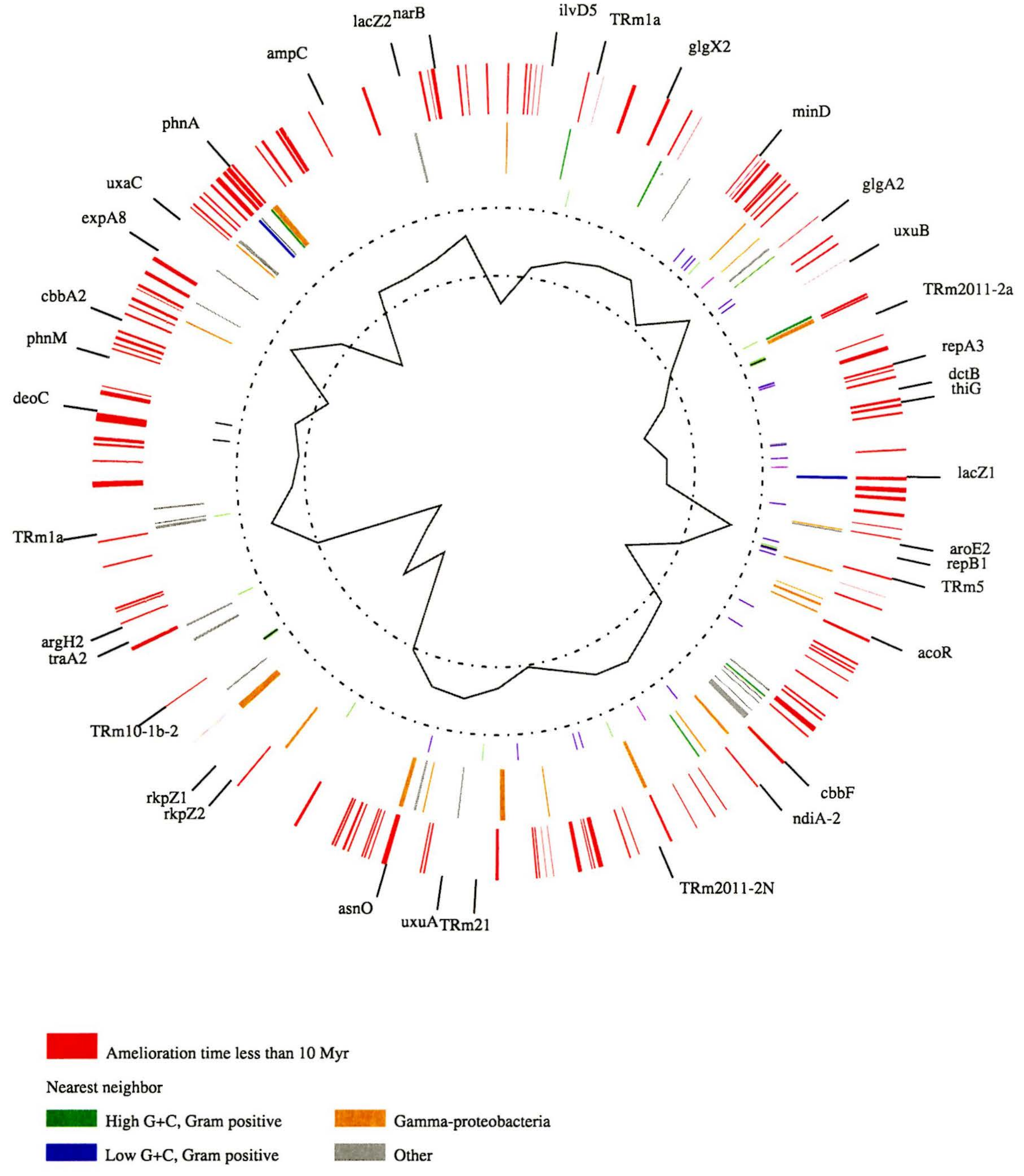
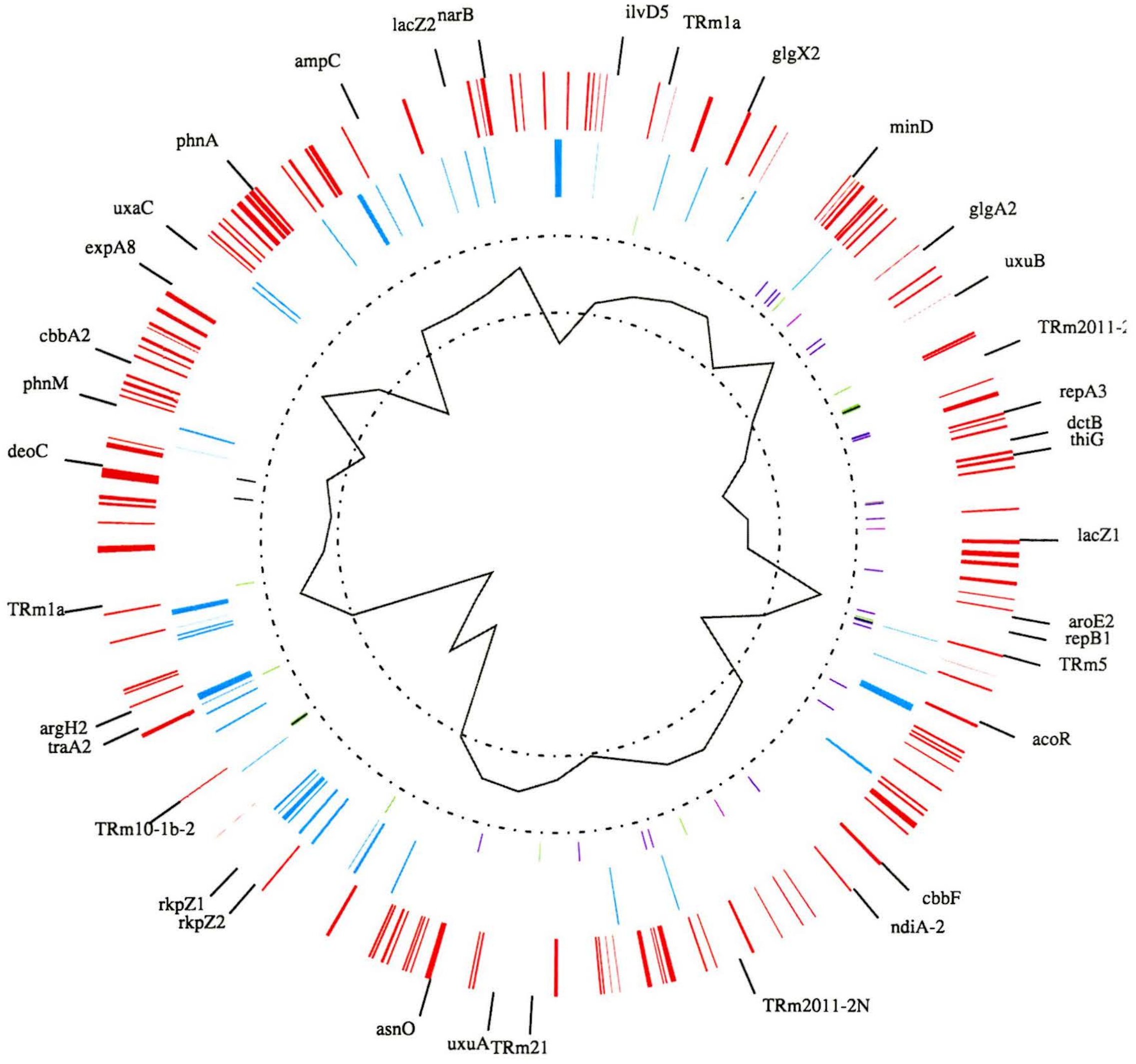

lacZ2
narB
ilvD5
TRm1a
glgX2
minD
glgA2
uxuB
TRm2011-2a
repA3
dctB
thiG
lacZ1
aroE2
repB1
TRm5
acoR
cbbF
ndiA-2
TRm2011-2N
TRm21
uxuA
asnO
rkpZ1
rkpZ2
TRm10-1b-2
argH2
traA2
TRm1a
deoC
phnM
cbbA2
expA8
uxaC
phnA
ampC
🟥 Amelioration time less than 10 Myr
Nearest neighbor
🟩 High G+C, Gram positive
🟦 Low G+C, Gram positive
🟧 Gamma-proteobacteria
▨ Other

Figure 3.2: Duplicated pSymB genes. Shown are genes that have duplicate copies on one of the three *S. meliloti* replicons (second circle from the outside). See Figure 3.1 for an explanation of the outer and inner circles.



🟥 Amelioration time less than 10 Myr

🟦 pSymB genes duplicated in the S. meliloti genome

Figure 3.3: Functions of pSymB genes involved in horizontal transfer. Included are genes involved in metabolism of amino acids and various carbon compounds, sugar-nucleotide synthesis, uptake of iron, amino acids and carbohydrates, and a putative nodulation protein.

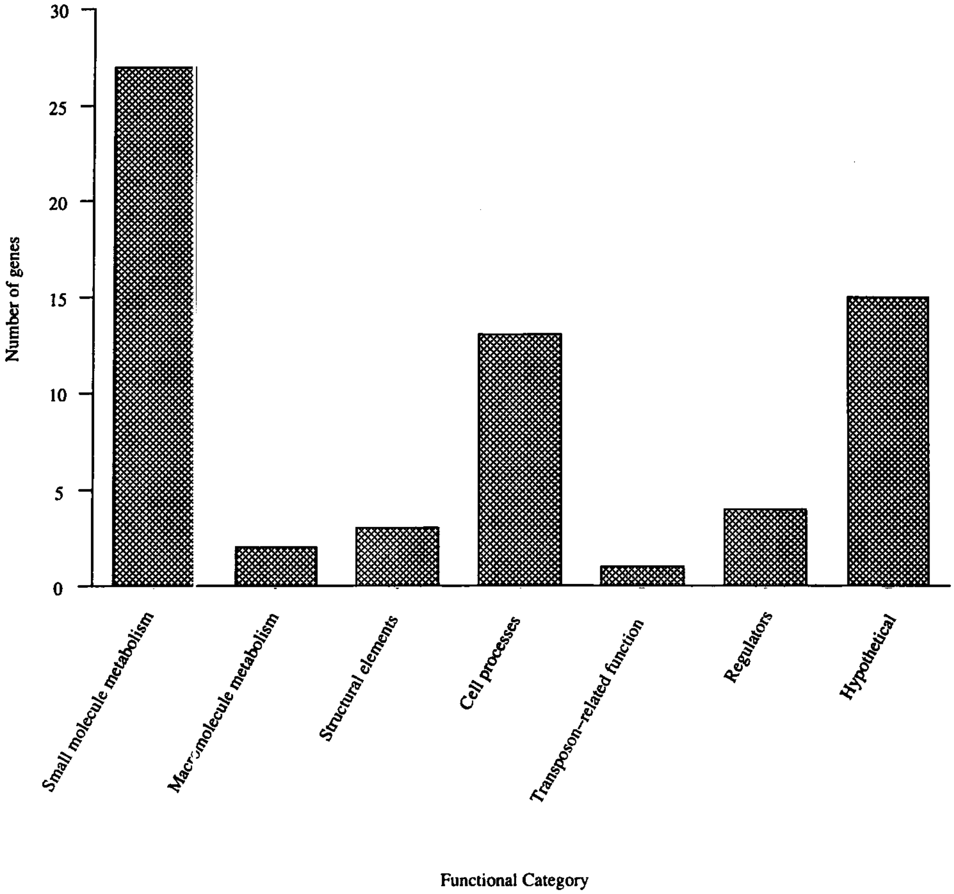

30
25
20
15
10
5
0
Number of genes
Small molecule metabolism
Macromolecule metabolism
Structural elements
Cell processes
Transposon-related function
Regulators
Hypothetical
Functional Category

Figure 3.4: Distribution of patches and potential operons among species with nearest neighbors to pSymB genes. *Dark grey* is the total number of genes with the specified nearest neighbor; *grey* is the number of genes found in patches; *light grey* is the genes found in potential operons. *C*, *L*, and *pAt* refer to the *A. tumefaciens* circular and linear chromosomes and plasmid pAt, respectively.

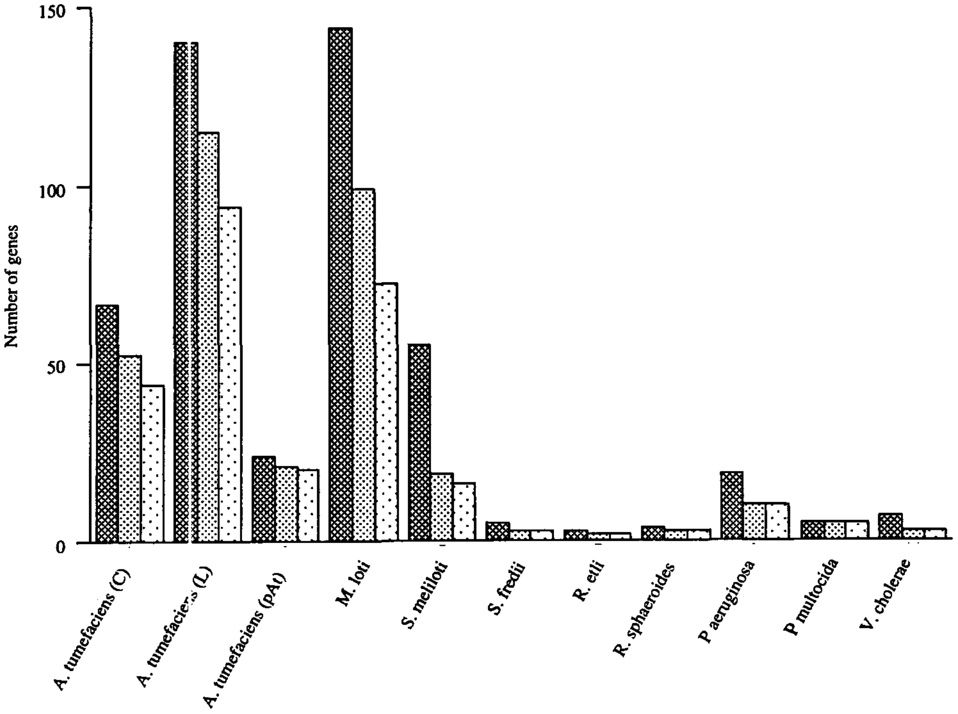

150
100
50
0
Number of genes
A. tumefaciens (C)
A. tumefaciens (L)
A. tumefaciens (pAt)
M. loti
S. meliloti
S. fredii
R. etli
R. sphaeroides
P aeruginosa
P multocida
V. cholerae

Figure 3.5: Amelioration simulation. (a) Amelioration curves (mean of 1000 iterates) for sequences with initial G+C contents of 49% to 58%. The minumum least-squares difference was reached after 172 Myr when the initial G+C content was 53% (*blue line*). (b) Ten iterates of amelioration of a 4.5-kb sequence with an initial G+C content of 53%. The G+C contents of sequences in (a) and (b) were compared to a pool of genes with 58.8%, 47.1%, and 64.5% G+C at each codon position. (c) Amelioration curves for sequences with initial G+C contents of 54% to 65%. The minumum least-squares difference was reached after 586 Myr when the initial G+C content was 59% (*yellow line*). (d) Ten iterates of amelioration of a 4.5-kb sequence with an initial G+C of 59%. The G+C contents of sequences in (c) and (d) were compared to a pool of genes with 62.4%, 50.6%, and 74.7% G+C at each codon position.

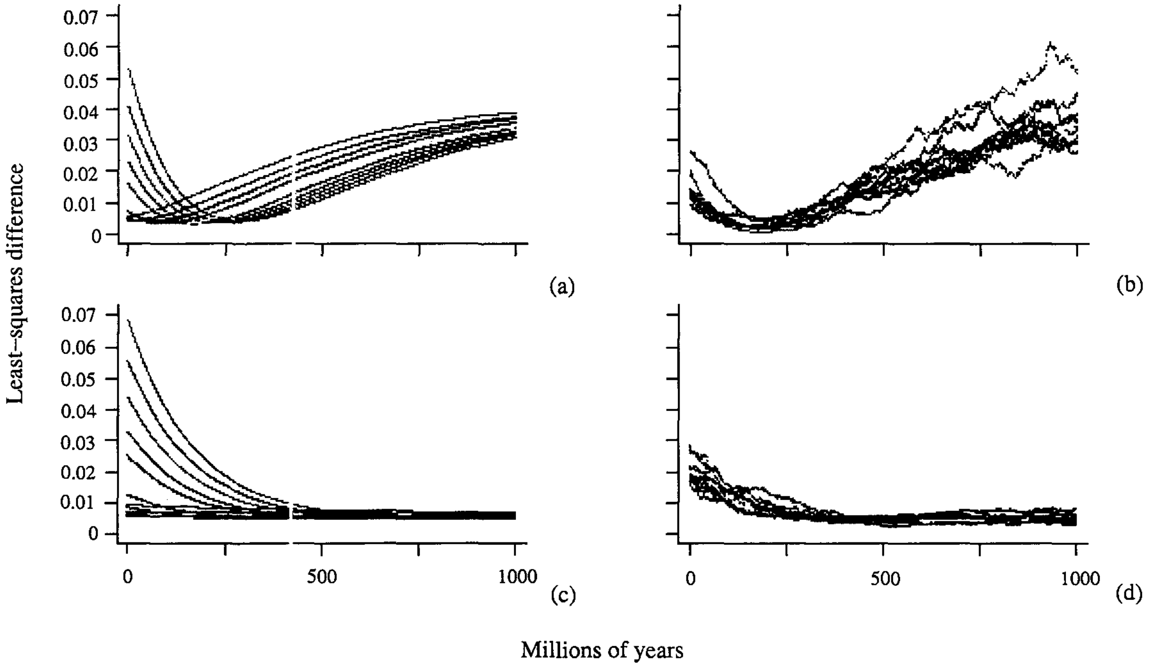

0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Least-squares difference
(a)
(b)
(c)
(d)
0
500
1000
0
500
1000
Millions of years

Figure 3.6: Estimated amelioration times of pSymB genes. Simulations that gave an estimate of 0 or 1000 million years are excluded. Proportions are calculated from a total of 845 genes. The initial G+C content of these genes ranges from 49-79%, with a mean of 64%.

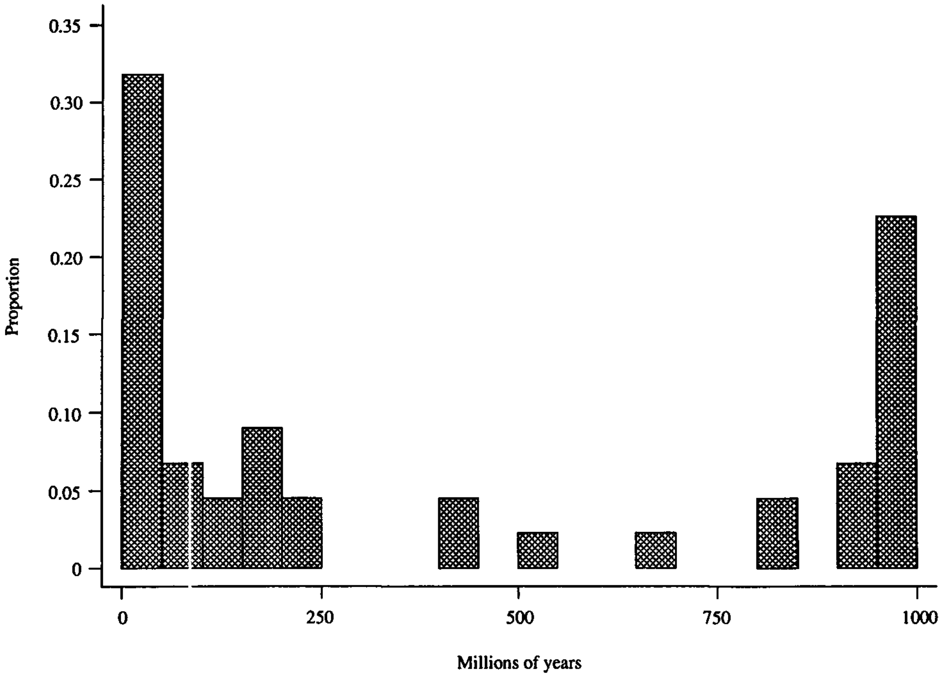

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
Proportion
0
250
500
750
1000
Millions of years

Table 3.1: *S. meliloti* substitution rates. The calculation of substitution rates at each codon position per lineage was based on the proportion of synonymous and non-synonymous sites at each position and a divergence time of 265 million years between *S. meliloti* and *M. loti*. The divergence was calculated based on the maximum likelihood codon substitution model of Goldman and Yang (Goldman and Yang 1994).

Table 3.2: The Tamura nucleotide substitution model. The Tamura model (1992) was used to simulate amelioration of the pSymB genes. Rates of substitution are shown for the original nucleotide (first column) changing to any of the four nucleotides. The rates are dependent on the G+C content (θ) and the transition (α) and transversion (β) rates.

	Proportion of sites		Substitutions	Substitution rate per lineage
	Synonymous sites	Non-synonymous sites	per site	(% substitutions/Myr.)
Synonymous rate	1	0	1.16	0.220
Non-synonymous rate	0	1	0.08	0.016
1st codon position	0.05	0.95	0.14	0.026
2nd codon position	0	1	0.08	0.016
3rd codon position	0.72	0.28	0.86	0.163

Nucleotide	A	T	C	G
A	$1-(\theta\alpha+\beta)$	$(1-\theta)\beta$	$\theta\beta$	$\theta\alpha$
T	$(1-\theta)\beta$	$1-(\theta\alpha+\beta)$	$\theta\alpha$	$\theta\beta$
C	$(1-\theta)\beta$	$(1-\theta)\alpha$	$1-[(1-\theta)\alpha+\beta]$	$\theta\beta$
G	$(1-\theta)\alpha$	$(1-\theta)\beta$	$\theta\beta$	$1-[(1-\theta)\alpha+\beta]$

Table 3.3: Summary of nearest neighbor analysis for pSymB ORFs. BLAST hits refer to hits to species other than *S. meliloti.* Phylogenies were constructed for pSymB ORFs with 3 or more hits, and if the same nearest neighbor occured in at least 95 out of 100 trees, it was assigned as the nearest neighbor of that particular ORF. For ORFs without significant nearest neighbor bootstraps, or no phylogeny constructed, assignments were made using contextual information where possible (see Methods).

	No. of genes	% of total
Total protein-coding genes	1570	100
One or more BLAST hits with $E \leq 10^{-20}$	1211	77
Phylogenies constructed	984	63
Nearest neighbors with significant bootstraps	**460**	**29**
Predictions for ORFs with non-significant bootstraps	31	2.0
Predictions for ORFs with no phylogeny constructed	20	1.3
Confirmed false predictions	1	0.06
Total nearest neighbors predicted using contextual information	**50**	**3**
Total nearest neighbor assignments	**510**	**33**

Table 3.4: pSymB nearest neighbors. *Pred.* is the number of predicted nearest neighbors based on contextual information; *Prop.* is the proportion of total nearest neighbors assigned. Asterisk (*) indicates completely sequenced genomes at the time the BLAST search was performed. The total number of pSymB ORFs is 1570.

Species		Nearest	Pred.	Prop.
		Neighbors		(%)
α-proteobacteria	_Sinorhizobium meliloti_*	50	4	10.6
	_Agrobacterium tumefaciens_*	207	23	45.1
	_Mesorhizobium loti_*	127	17	28.2
	Rhizobium leguminosarum	4		0.8
	Sinorhizobium fredii	4	1	1.0
	Rhizobium etli	3		0.6
	Rhizobium sp. NGR234	2		0.4
	Brucella melitensis	1		0.2
	_Methylobacterium extorquens_*	1		0.2
non-Rhizobiaceae	_Caulobacter vibrioides (crescentus)_	6		1.2
	Paracoccus denitrificans	4		0.8
	Paracoccus pantotrophus	1		0.2
	Rhodobacter sphaeroides	3	1	0.6
	Rhodobacter capsulatus	1		0.2
β-proteobacteria	_Bordetella bronchiseptica_	1		0.2
	Ralstonia sp.	1		0.2
γ-proteobacteria	_Pseudomonas aeruginosa_*	17	2	3.7
	_Vibrio cholerae_*	5	1	1.2
	_Pasteurella multocida_*	4	1	1.0
	Pseudomonas syringae	1		0.2
	Acidithiobacillus ferrooxidans	1		0.2
	_Escherichia coli_*	1		0.2
	_Yersinia pseudotuberculosis (pestis)_*	1		0.2
High G+C, Gram$^+$	_Streptomyces coelicolor_	3		0.6
	_Mycobacterium tuberculosis_*	2		0.4
	Mycobacterium avium	1		0.2
	Streptomyces hygroscopicus	1		0.2
Low G+C, Gram$^+$	_Desulfonispora thiosulfatigenes_	1		0.2
	Thermoanaerobacterium sp.	1		0.2
Thermus/Deinococcus	_Deinococcus radiodurans_*	2		0.4
Cyanobacteria	_Synechocystis_ sp.*	1		0.2
Euryarchaeota	_Haloferax alicantei_	1		0.2
	_Methanothermobacter thermautotrophicus_*	1		0.2
Total		460	50	100

Table 3.5: Genes with amelioration times less than 10 million years. Listed are species with nearest neighbors to pSymB genes that are estimated to have been ameliorating for less than 10 Myr. A total of 148 genes on pSymB are predicted to have an amelioration time of less than 10 million years. *Total N.N.* is the total number of nearest neighbors assigned from a species.

Nearest neighbor	Number of genes	Total N.N.
S. meliloti	1	54
M. loti	19	147
A. tumefaciens	23	230
R. leguminosarum	1	4
B. melitensis	1	1
P. aeruginosa	6	19
V. cholerae	2	6
P. denitrificans	1	4
Thermoanaerobacterium sp.	1	1
D. thiosulfatigenes	1	1
S. hygroscopicus	1	1
None assigned	91	1060

Appendix A

Table A.1: Nearest neighbors of potentially horizontally transferred pSymB genes. The identity of the nearest neighbor is given by the NCBI accession number, and pSymB genes are identified by the number and name, if applicable. *Hits* refers to BLAST hits (E-value $\leq 10^{-20}$) to the pSymB gene to *A. tumefaciens* (A), *M. loti* (M), and *S. meliloti* (S).

Species	Nearest Neighbor	pSymB gene	Name	Hits
C. vibrioides (*crescentus*)	AAK23604	SMb21041		-
	AAK22151	SMb21070	exoP2	A M S
	AAK24235	SMb21081	manB	-
	AAK25172	SMb21232		M
	AAK23150	SMb21244		-
	AAK23876	SMb21312	expE3	M
P. denitrificans	CAA11377	SMb20163		-
	AAC44554	SMb20171		A M S
	AAC44557	SMb20175		-
	PTA_PARDE	SMb21532	pta	M S
P. pantotrophus	AF295359_4	SMb20436		-
R. sphaeroides	RBL1_RHOSH	SMb20198	cbbL	S
	AAA26114	SMb20199	cbbA	M S
	TKT_RHOSH	SMb20200	cbbT	A M S
	AAD20227	SMb21370		M S
R. capsulatus	T03485	SMb21494	ocd	A M
B. bronchiseptica	CAA07640	SMb21235		-
Ralstonia sp.	AAC34291	SMb20035		A S
P. aeruginosa	E83615	SMb20033		A S
	F83328	SMb20069		S

Continued on next page

Continued from previous page

Species	Nearest Neighbor	pSymB gene	Name	Hits
	F83372	SMb20091		-
	C83382	SMb20095		S
	G83133	SMb20218		A M S
	F83133	SMb20219		A M S
	H83444	SMb20402		A M S
	B83445	SMb20403		A M S
	A83445	SMb20404		M S
	D83214	SMb20466		-
	B83214	SMb20481	asnO	M S
	C83214	SMb20482		-
	E82991	SMb20810		A M S
	C83364	SMb21367	cycA	-
	C83309	SMb20861		-
	C83502	SMb20684		A M
	F83369	SMb20769		A S
	G83369	SMb20770		-
	H83369	SMb20771		A S
V. cholerae	H82157	SMb20295		A S
	G82157	SMb20296		-

Continued on next page

Continued from previous p ıge

Species	Nearest Neighbor	pSymB gene	Name	Hits
	F82157	SMb20297		A S
	B82357	SMb20368		A S
	A82441	SMb21540		A M S
	G82440	SMb21542		A M S
P. multocida	AAK03360	SMb21017		A M S
	AAK03359	SMb21018		A M S
	AAK03358	SMb21019		A M S
	AAK03357	SMb21021		A M S
	AAK03356	SMb21022		A S
P. syringae	NP_114214	SMb20101		A M
A. ferrooxidans	AF173880_9	SMb21207		-
E. coli	YGBK_ECOLI	SMb20670		-
Y. pseudotuberculosis (pestis)	AAB48319	SMb20244	tyv	M S
S. coelicolor	T36586	SMb20249		-
	T35169	SMb21463		M S
	CAB61274	SMb20767	dak	A M S
M. tuberculosis	C70585	SMb20168		-
	YL40_MYCTU	SMb20703		-

Continued on next page

Continued from previous pαge

Species	Nearest Neighbor	pSymB gene	Name	Hits
M. avium	AAG44885	SMb20905		S
S. hygroscopicus	BAA07116	SMb21539		A M S
D. thiosulfatigenes	AF305552_1	SMb21530	ilvB2	M S
Thermoanaerobacterium sp.	BGAL_THETU	SMb21655	lacZ1	-
D. radiodurans	H75575	SMb21340		A M S
	F75497	SMb20697		-
Synechocystis sp.	YB78_SYNY3	SMb20472		M
H. alicantei	T44793	SMb20966	lacZ2	-
M. thermautotrophicus	B69012	SMb20185		-

Bibliography

Allardet-Servent, A., S. Michaux-Charachon, E. Jumas-Bilak, L. Karayan, and M. Ramuz (1993). Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J Bacteriol 175*, 7869–7874.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res 25*, 3389–3402.

Aneja, P. and T. C. Charles (1999). Poly-3-hydroxybutyrate degradation in *Rhizobium* (*Sinorhizobium*) *meliloti*: isolation and characterization of a gene encoding 3- hydroxybutyrate dehydrogenase. *J Bacteriol 181*, 849–857.

Aota, S. and T. Ikemura (1986). Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res 14*, 6345–6355.

Banfalvi, Z., V. Sakanyan, C. Koncz, A. Kiss, I. Dusha, and A. Kondorosi (1981). Location of nodulation and nitrogen fixation genes on a high molecular weight plasmid of *R. meliloti*. *Mol Gen Genet 184*, 318–325.

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

Bardwell, J. C. and E. A. Craig (1987). Eukaryotic Mr 83,000 heat shock protein has a homologue in *Escherichia coli*. *Proc Natl Acad Sci USA 84*, 5177–5181.

Barloy-Hubler, F., D. Capela, J. Batut, and F. Galibert (2000). High-resolution physical map of the pSymb megaplasmid and comparison of the three replicons of *Sinorhizobium meliloti* strain 1021. *Curr Microbiol 41*, 109–113.

Barnett, M. J., R. F. Fisher, T. Jones, C. Komp, A. P. Abola, F. Barloy-Hubler, L. Bowser, D. Capela, F. Galibert, J. Gouzy, M. Gurjal, A. Hong, L. Huizar, R. W. Hyman, D. Kahn, M. L. Kahn, S. Kalman, D. H. Keating, C. Palm, M. C. Peck, R. Surzycki, D. H. Wells, K. C. Yeh, R. W. Davis, N. A. Federspiel, and S. R. Long (2001). Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid. *Proc Natl Acad Sci USA 98*, 9883–9888.

Becker, A., A. Kleickmann, M. Keller, W. Arnold, and A. Pühler (1993). Identification and analysis of the *Rhizobium meliloti exoAMONP* genes involved in exopolysaccharide biosynthesis and mapping of promoters located on the *exoHKLAMONP* fragment. *Mol Gen Genet 241*, 367–379.

Becker, A., S. Rüberg, H. Küster, A. A. Roxlau, M. Keller, T. Ivashina, H. P. Cheng, G. C. Walker, and A. Pühler (1997). The 32-kilobase *exp* gene cluster of *Rhizobium meliloti* directing the biosynthesis of galactoglucan: genetic organization and properties of the encoded gene products. *J Bacteriol 179*, 1375–1384.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier (1985). The mosaic genome of warm-blooded vertebrates. *Science 228*, 953–958.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res 8*, 1499–1504.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature 321*, 209–13.

Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. Geoghagen, and J. C. Venter (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science 273*, 1058–1073.

Burge, C., A. M. Campbell, and S. Karlin (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA 89*, 1358–1362.

Campbell, A., J. Mrázek, and S. Karlin (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA 96*, 9184–9189.

Capela, D., F. Barloy-Hubler, J. Gouzy, G. Bothe, F. Ampe, J. Batut, P. Boistard, A. Becker, M. Boutry, E. Cadieu, S. Dréano, S. Gloux, T. Godrie, A. Goffeau, D. Kahn, E. Kiss, V. Lelaure, D. Masuy, T. Pohl, D. Portetelle, A. Pühler, B. Purnelle, U. Ramsperger, C. Renard, P. Thébault, M. Vandenbol, S. Weidner, and F. Galibert (2001). Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti strain 1021*. *Proc Natl Acad Sci USA 98*, 9877–9882.

Cardon, L. R., C. Burge, D. A. Clayton, and S. Karlin (1994). Pervasive CpG suppression in animal mitochondrial genomes. *Proc Natl Acad Sci USA 91*, 3799–3803.

Chain, P. S., I. Hernández-Lucas, B. Golding, and T. M. Finan (2000). *oriT*-directed cloning of defined large regions from bacterial genomes: identification of the *Sinorhizobium meliloti* pExo megaplasmid replicator region. *J Bacteriol 182*, 5486–

5494.

Charles, T. C. and T. M. Finan (1990). Genetic map of *Rhizobium meliloti* megaplasmid pRmeSU47b. *J Bacteriol 172*, 2469–2476.

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

Cheng, H. P. and T. G. Lessie (1994). Multiple replicons constituting the genome of *Pseudomonas cepacia* 17616. *J Bacteriol 176*, 4034–4042.

Coplin, D. L. and D. Cook (1990). Molecular genetics of extracellular polysaccharide biosynthesis in vascular phytopathogenic bacteria. *Mol Plant Microbe Interact 3*, 271–279.

de Boer, P. A., R. E. Crossley, and L. I. Rothfield (1989). A division inhibitor and a topological specificity factor coded for by the minicell locus determine proper placement of the division septum in *E. coli*. *Cell 56*, 641–649.

Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res 27*, 4636–4641.

Deng, W., M. P. Gordon, and E. W. Nester (1995). Sequence and distribution of *IS*1312: evidence for horizontal DNA transfer from *Rhizobium meliloti* to *Agrobacterium tumefaciens*. *J Bacteriol 177*, 2554–2559.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res 8*, 163–167.

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

Ewing, B. and P. Green (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res 8*, 186–194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res 8*, 175–185.

Feldman, M. F., C. L. Marolda, M. A. Monteiro, M. B. Perry, A. J. Parodi, and M. A. Valvano (1999). The activity of a putative polyisoprenol-linked sugar translocase (Wzx) involved in *Escherichia coli* O antigen assembly is independent of the chemical structure of the O repeat. *J Biol Chem 274*, 35129–35138.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package), version 3.5c.* University of Washington, Seattle, Washington.

Férrandez, A., B. Minambres, B. Garcia, E. R. Olivera, J. M. Luengo, J. L. Garcia, and E. Diaz (1998). Catabolism of phenylacetic acid in *Escherichia coli.* Characterization of a new aerobic hybrid pathway. *J Biol Chem 273*, 25974–25986.

Finan, T. M., B. Kunkel, G. F. De Vos, and E. R. Signer (1986). Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. *J Bacteriol 167*, 66–72.

Finan, T. M., I. Oresnik, and A. Bottacin (1988). Mutants of *Rhizobium meliloti* defective in succinate metabolism. *J Bacteriol 170*, 3396–3403.

Finan, T. M., S. Weidner, K. Wong, J. Buhrmester, P. Chain, F. J. Vorhölter, I. Hernandez-Lucas, A. Becker, A. Cowie, J. Gouzy, B. Golding, and A. Pühler (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.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick et al (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.

Science 269, 496–512.

Gage, D. J. and S. R. Long (1998). alpha-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.

Galbraith, M. P., S. F. Feng, J. Borneman, E. W. Triplett, F. J. de Bruijn, and S. Rossbach (1998). A functional myo-inositol catabolism pathway is essential for rhizopine utilization by *Sinorhizobium meliloti*. *Microbiology 144*, 2915–2924.

Galibert, F., T. M. Finan, S. R. Long, A. Pühler, P. Abola, F. Ampe, F. Barloy-Hubler, M. J. Barnett, A. Becker, P. Boistard, G. Bothe, M. Boutry, L. Bowser, J. Buhrmester, E. Cadieu, D. Capela, P. Chain, A. Cowie, R. W. Davis, S. Dreano, N. A. Federspiel, R. F. Fisher, S. Gloux, T. Godrie, A. Goffeau, B. Golding, J. Gouzy, M. Gurjal, I. Hernandez-Lucas, A. Hong, L. Huizar, R. W. Hyman, T. Jones, D. Kahn, M. L. Kahn, S. Kalman, D. H. Keating, E. Kiss, C. Komp, V. Lelaure, D. Masuy, C. Palm, M. C. Peck, T. M. Pohl, D. Portetelle, B. Purnelle, U. Ramsperger, R. Surzycki, P. Thebault, M. Vandenbol, F. J. Vorholter, S. Weidner, D. H. Wells, K. Wong, K. C. Yeh, and J. Batut (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science 293*, 668–672.

Garcia-Rodriguez, F. M. and N. Toro (2000). *Sinorhizobium meliloti nfe* (nodulation formation efficiency) genes exhibit temporal and spatial expression patterns similar to those of genes involved in symbiotic nitrogen fixation. *Mol Plant Microbe Interact 13*, 583–591.

Garcia-Vallve, S., A. Romeu, and J. Palau (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res 10*, 1719–1725.

Glazebrook, J., A. Ichige, and G. C. Walker (1993). A *Rhizobium meliloti* homolog of the *Escherichia coli* peptide-antibiotic transport protein SbmA is essential for bacteroid development. *Genes Dev 7*, 1485–1497.

Glazebrook, J. and G. C. Walker (1989). A novel exopolysaccharide can function in place of the calcofluor- binding exopolysaccharide in nodulation of alfalfa by *Rhizobium meliloti*. *Cell 56*, 661–672.

Glucksmann, M. A., T. L. Reuber, and G. C. Walker (1993). Genes needed for the modification, polymerization, export, and processing of succinoglycan by *Rhizobium meliloti*: a model for succinoglycan biosynthesis. *J Bacteriol 175*, 7045–7055.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol 11*, 725–736.

Goodner, B., G. Hinkle, S. Gattung, N. Miller, M. Blanchard, B. Qurollo, B. S. Goldman, Y. Cao, M. Askenazi, C. Halling, L. Mullin, K. Houmiel, J. Gordon, M. Vaudin, O. Iartchouk, A. Epp, F. Liu, C. Wollam, M. Allinger, D. Doughty, C. Scott, C. Lappas, B. Markelz, C. Flanagan, C. Crowell, J. Gurson, C. Lomo, C. Sear, G. Strub, C. Cielo, and S. Slater (2001). Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science 294*, 2323–2328.

Göttfert, M., S. Rothlisberger, C. Kundig, C. Beck, R. Marty, and H. Hennecke (2001). Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the *Bradyrhizobium japonicum* chromosome. *J Bacteriol 183*, 1405–1412.

Haas, S., M. Vingron, A. Poustka, and S. Wiemann (1998). Primer design for large scale sequencing. *Nucleic Acids Res 26*, 3006–3012.

Haydon, D. J. and J. R. Guest (1991). A new family of bacterial regulatory proteins. *FEMS Microbiol Lett 63*, 291–295.

Heidelberg, J. F., J. A. Eisen, W. C. Nelson, R. A. Clayton, M. L. Gwinn, R. J. Dodson,

D. H. Haft, E. K. Hickey, J. D. Peterson, L. Umayam, S. R. Gill, K. E. Nelson, T. D. Read, H. Tettelin, D. Richardson, M. D. Ermolaeva, J. Vamathevan, S. Bass, H. Qin, I. Dragoi, P. Sellers, L. McDonald, T. Utterback, R. D. Fleishmann, W. C. Nierman, and O. White (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature 406*, 477–483.

Honeycutt, R. J., M. McClelland, and B. W. Sobral (1993). Physical map of the genome of *Rhizobium meliloti* 1021. *J Bacteriol 175*, 6945–6952.

Huynen, M., B. Snel, W. Lathe III, and P. Bork (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res 10*, 1204–1210.

Hynes, M., R. Simon, P. Müller, K. Niehaus, M. Labes, and A. Pühler (1986). The two megaplasmids of *Rhizobium meliloti* are involved in the effective nodulation of alfalfa. *Mol Gen Genet 202*, 356–362.

Itoh, T., K. Takemoto, H. Mori, and T. Gojobori (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol 16*, 332–346.

Joset, F. and J. Guespin-Michel (1994). *Prokaryotic Genetics: Genome organization, transfer, and plasticity.* Oxford: Blackwell Science, Ltd.

Jukes, T. H. and M. Kimura (1984). Evolutionary constraints and the neutral theory. *J Mol Evol 21*, 90–92.

Jumas-Bilak, E., S. Michaux-Charachon, G. Bourg, M. Ramuz, and A. Allardet-Servent (1998). Unconventional genomic organization in the alpha subgroup of the Proteobacteria. *J Bacteriol 180*, 2749–2755.

Kaneko, T., Y. Nakamura, S. Sato, E. Asamizu, T. Kato, S. Sasamoto, A. Watanabe,

K. Idesawa, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, Y. Mochizuki, S. Nakayama, N. Nakazaki, S. Shimpo, M. Sugimoto, C. Takeuchi, M. Yamada, and S. Tabata (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. *DNA Res 7*, 331–338.

Karlin, S., L. Brocchieri, J. Mrazek, A. M. Campbell, and A. M. Spormann (1999). A chimeric prokaryotic ancestry of mitochondria and primitive eukaryotes. *Proc Natl Acad Sci USA 96*, 9190–9195.

Karlin, S. and C. Burge (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet 11*, 283–290.

Karlin, S. and I. Ladunga (1994). Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA 91*, 12832–12836.

Karlin, S., I. Ladunga, and B. E. Blaisdell (1994). Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci USA 91*, 12837–12841.

Karlin, S. and J. Mrazek (1996). What drives codon choices in human genes? *J Mol Biol 262*, 459–472.

Karlin, S. and J. Mrázek (1997). Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci USA 94*, 10227–10232.

Karlin, S., J. Mrázek, and A. M. Campbell (1997). Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol 179*, 3899–3913.

Kennedy, E. P. (1996). Membrane-Derived Oligosaccharides (Periplasmic Beta-d-Glucans) of *Escherichia coli*. In F. C. Neidhardt (Ed.), *Escherichia coli and Salmonella: cellular and molecular biology, 2nd Ed.*, pp. 1064–1070. Washington, D.C.: Am. Soc. Microbiol. Press.

Kereszt, A., E. Kiss, B. L. Reuhs, R. W. Carlson, A. Kondorosi, and P. Putnoky (1998). Novel *rkp* gene clusters of *Sinorhizobium meliloti* involved in capsular polysaccharide production and invasion of the symbiotic nodule: the *rkpK* gene encodes a UDP-glucose dehydrogenase. *J Bacteriol 180*, 5426–5431.

Koski, L. B. and G. B. Golding (2001). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol 52*, 540–542.

Koski, L. B., R. A. Morton, and G. B. Golding (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol 18*, 404–412.

Lawrence, J. G. (1995). Ameliorator, Version 1.0.

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

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

Lepo, J. E., F. J. Hanus, and H. J. Evans (1980). Chemoautotrophic growth of hydrogen-uptake-positive strains of *Rhizobium japonicum*. *J Bacteriol 141*, 664–670.

Lin, J. T., B. S. Goldman, and V. Stewart (1993). Structures of genes *nasA* and *nasB*, encoding assimilatory nitrate and nitrite reductases in *Klebsiella pneumoniae* M5al. *J Bacteriol 175*, 2370–2378.

Liu, C.-M., P. A. McLean, C. C. Sookdeo, and F. C. Cannon (1991). Degradation of the herbicide glyphosate by members of the family *Rhizobiaceae*. *Appl Environ Microbiol 59*, 1799–1804.

Margolin, W. and S. R. Long (1993). Isolation and characterization of a DNA replication origin from the 1,700-kilobase-pair symbiotic megaplasmid pSym-b of *Rhizobium meliloti*. *J Bacteriol 175*, 6553–6561.

Meade, H. M. and E. R. Signer (1977). Genetic mapping of *Rhizobium meliloti*. *Proc Natl Acad Sci USA 74*, 2076–2078.

Médigue, C., T. Rouxel, P. Vigier, A. Hénaut, and A. Danchin (1991). Evidence for horizontal gene transfer in Escherichia coli speciation. *J Mol Biol 222*, 851–856.

Michaux, S., J. Paillisson, M. J. Carles-Nurit, G. Bourg, A. Allardet-Servent, and M. Ramuz (1993). Presence of two independent chromosomes in the *Brucella melitensis* 16M genome. *J Bacteriol 175*, 701–705.

Missiakas, D. and S. Raina (1998). The extracytoplasmic function sigma factors: role and regulation. *Mol Microbiol 28*, 1059–1066.

Miyata, T. and T. Yasunaga (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol 16*, 23–36.

Müller, P., M. Keller, W. M. Weng, J. Quandt, W. Arnold, and A. Pühler (1993). Genetic analysis of the *Rhizobium meliloti exoYFQ* operon: ExoY is homologous to sugar transferases and ExoQ represents a transmembrane protein. *Mol Plant Microbe Interact 6*, 55–65.

Muto, A. and S. Osawa (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci USA 84*, 166–169.

Ng, W. V., S. A. Ciufo, T. M. Smith, R. E. Bumgarner, D. Baskin, J. Faust, B. Hall, C. Loretz, J. Seto, J. Slagel, L. Hood, and S. DasSarma (1998). Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? *Genome Res 8*, 1131–1141.

Nussinov, R. (1984a). Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res 12*, 1749–1763.

Nussinov, R. (1984b). Strong doublet preferences in nucleotide sequences and DNA geometry. *J Mol Evol 20*, 111–119.

Okusu, H., D. Ma, and H. Nikaido (1996). AcrAB efflux pump plays a major role in the antibiotic resistance phenotype of *Escherichia coli* multiple-antibiotic-resistance (Mar) mutants. *J Bacteriol 178*, 306–308.

Oresnik, I. J., S. L. Liu, C. K. Yost, and M. F. Hynes (2000). Megaplasmid pRme2011a of *Sinorhizobium meliloti* is not required for viability. *J Bacteriol 182*, 3582–3586.

Østeras, M., E. Boncompagni, N. Vincent, M. C. Poggi, and D. Le Rudulier (1998). Presence of a gene encoding choline sulfatase in *Sinorhizobium meliloti bet* operon: choline-O-sulfate is metabolized into glycine betaine. *Proc Natl Acad Sci USA 95*, 11394–11399.

Østeras, M., J. Stanley, and T. M. Finan (1995). Identification of *Rhizobium*-specific intergenic mosaic elements within an essential two-component regulatory system of *Rhizobium* species. *J Bacteriol 177*, 5485–5494.

Palmer, K. M., S. L. Turner, and J. P. Young (2000). Sequence diversity of the plasmid replication gene *repC* in the *Rhizobiaceae*. *Plasmid 44*, 209–219.

Pesole, G., M. P. Bozzetti, C. Lanave, G. Preparata, and C. Saccone (1991). Glutamine synthetase gene evolution: a good molecular clock. *Proc Natl Acad Sci U S A 88*, 522–526.

Reuhs, B. L., M. N. Williams, J. S. Kim, R. W. Carlson, and F. Cote (1995). Suppression of the Fix- phenotype of *Rhizobium meliloti exoB* mutants by *lpsZ* is correlated to a modified expression of the K polysaccharide. *J Bacteriol 177*, 4289–4296.

Riley, M. and B. Labedan (1996). *E. coli* gene products: Physiological functions and common ancestries. In F. C. Neidhardt (Ed.), *Escherichia coli and Salmonella: cel-*

lular and molecular biology, 2nd Ed., pp. 2118–2202. Washington, D.C.: Am. Soc. Microbiol. Press.

Rodley, P. D., U. Romling, and B. Tummler (1995). A physical genome map of the *Burkholderia cepacia* type strain. *Mol Microbiol 17*, 57–67.

Rosenberg, C., P. Boistard, J. Dénarié, and F. Casse-Delbart (1981). Genes controlling early and late functions in symbiosis are located on a megaplasmid in *Rhizobium meliloti. Mol Gen Genet 184*, 326–333.

Rosenberg, C., F. Casse-Delbart, I. Dusha, M. David, and C. Boucher (1982). Megaplasmids in the plant-associated bacteria *Rhizobium meliloti* and *Pseudomonas solanacearum. J Bacteriol 150*, 402–406.

Schell, M. A. (1993). Molecular biology of the LysR family of transcriptional regulators. *Annu Rev Microbiol 47*, 597–626.

Schiex, T., P. Thébault, and D. Kahn (2000). Recherche des gènes et des erreurs de séquençage dans les génomes bactériens GC-riches (et autres...). In *Proceedings of the JOBIM'2000 Conference, Montpellier, France*, pp. 321–328.

Schnaitman, C. A. and J. D. Klena (1993). Genetics of lipopolysaccharide biosynthesis in enteric bacteria. *Microbiol Rev 57*, 655–682.

Schwedock, J. S. and S. R. Long (1992). *Rhizobium meliloti* genes involved in sulfate activation: the two copies of *nodPQ* and a new locus, *saa. Genetics 132*, 899–909.

Shatters, R. G., J. E. Somerville, and M. L. Kahn (1989). Regulation of glutamine synthetase II activity in *Rhizobium meliloti* 104A14. *J Bacteriol 171*, 5087–5094.

Shively, J. M., G. van Keulen, and W. G. Meijer (1998). Something from almost nothing: carbon dioxide fixation in chemoautotrophs. *Annu Rev Microbiol 52*, 191–230.

Sicheritz-Pontén, T. and S. G. Andersson (2001). A phylogenomic approach to microbial evolution. *Nucleic Acids Res 29*, 545–552.

Sonnhammer, E. L. and D. Kahn (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci 3*, 482–492.

Staden, R. (1996). The Staden sequence analysis package. *Mol Biotechnol 5*, 233–241.

Stanfield, S. W., L. Ielpi, D. O'Brochta, D. R. Helinski, and G. S. Ditta (1988). The *ndvA* gene product of *Rhizobium meliloti* is required for beta-(1— 2)glucan production and has homology to the ATP-binding export protein HlyB. *J Bacteriol 170*, 3523–3530.

Stevenson, G., K. Andrianopoulos, M. Hobbs, and P. R. Reeves (1996). Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. *J Bacteriol 178*, 4885–4893.

Streit, W. R., C. M. Joseph, and D. A. Phillips (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.

Strimmer, K. and A. von Haeseler (1996). Quartet puzzling: A quartet maximum-likelihood method for recontructing tree topoplogies. *Mol. Biol. Evol. 13*, 964–969.

Sullivan, J. T. and C. W. Ronson (1998). Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci USA 95*, 5145–5149.

Summers, M. L., M. C. Denton, and T. R. McDermott (1999). Genes coding for phosphotransacetylase and acetate kinase in *Sinorhizobium meliloti* are in an operon that is inducible by phosphate stress and controlled by *phoB*. *J Bacteriol 181*, 2217–2224.

Suominen, L., C. Roos, G. Lortet, L. Paulin, and K. Lindstrom (2001). Identification and structure of the *Rhizobium galegae* common nodulation genes: evidence for horizontal gene transfer. *Mol Biol Evol 18*, 907–916.

Suwanto, A. and S. Kaplan (1989). Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. *J Bacteriol 171*, 5850–5859.

Tabata, S., P. J. Hooykaas, and A. Oka (1989). Sequence determination and characterization of the replicator region in the tumor-inducing plasmid pTiB6S3. *J Bacteriol 171*, 1665–1672.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol 9*, 678–687.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res 28*, 33–36.

Thomas, C. M. (2000). Paradigms of plasmid organization. *Mol Microbiol 37*, 485–491.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res 22*, 4673–4680.

Trucksis, M., J. Michalski, Y. K. Deng, and J. B. Kaper (1998). The *Vibrio cholerae* genome contains two unique circular chromosomes. *Proc Natl Acad Sci USA 95*, 14464–14469.

Turner, S. L. and J. P. Young (2000). The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. *Mol Biol Evol 17*, 309–319.

van Pée, K. H. (1996). Biosynthesis of halogenated metabolites by bacteria. *Annu Rev Microbiol 50*, 375–399.

Vincent, J. (1941). Serological studies of the root-nodule bacteria. I. Strains of Rhizobium meliloti. *Proc Linn Soc. N.S.W 66*, 145–154.

Wang, B. (2001). Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol 53*, 244–250.

Watson, R. J., Y. K. Chan, R. Wheatcroft, A. F. Yang, and S. H. Han (1988). *Rhizobium meliloti* genes required for C_4-dicarboxylate transport and symbiotic nitrogen fixation are located on a megaplasmid. *J Bacteriol 170*, 927–934.

White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft, M. L. Gwinn, W. C. Nelson, D. L. Richardson, K. S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J. J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K. S. Makarova, L. Aravind, M. J. Daly, K. W. Minton, R. D. Fleischmann, K. A. Ketchum, K. E. Nelson, S. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser (1999). Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science 286*, 1571–1577.

Whitfield, C., P. A. Amor, and R. Köplin (1997). Modulation of the surface architecture of gram-negative bacteria by the action of surface polymer:lipid A-core ligase and by determinants of polymer chain length. *Mol Microbiol 23*, 629–638.

Whitfield, C. and I. S. Roberts (1999). Structure, assembly and regulation of expression of capsules in *Escherichia coli* . *Mol Microbiol 31*, 1307–1319.

Wong, K., T. M. Finan, and G. B. Golding (2002). Dinucleotide compositional analysis of *Sinorhizobium meliloti* using the genome signature: distinguishing chromosomes and plasmids. *Funct. Integr. Genomics*, in press.

Wood, D. W., J. C. Setubal, R. Kaul, D. E. Monks, J. P. Kitajima, V. K. Okura, Y. Zhou, L. Chen, G. E. Wood, J. r. Almeida NF, L. Woo, Y. Chen, I. T. Paulsen, J. A. Eisen, P. D. Karp, S. r. Bovee D, P. Chapman, J. Clendenning, G. Deatherage, W. Gillet, C. Grant, T. Kutyavin, R. Levy, M. J. Li, E. McClelland, A. Palmieri, C. Raymond, G. Rouse, C. Saenphimmachak, Z. Wu, P. Romero, D. Gordon, S. Zhang, H. Yoo, Y. Tao, P. Biddle, M. Jung, W. Krespan, M. Perry, B. Gordon-Kamm, L. Liao, S. Kim, C. Hendrick, Z. Y. Zhao, M. Dolan, F. Chumley, S. V. Tingey, J.-F. Tomb, M. P. Gordon, M. V. Olson, and E. W. Nester (2001). The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science 294*, 2317–2323.

Wu, C. I. and N. Maeda (1987). Inequality in mutation rates of the two strands of DNA. *Nature 327*, 169–170.

Yamaichi, Y., T. Iida, K. S. Park, K. Yamamoto, and T. Honda (1999). Physical and genetic map of the genome of *Vibrio parahaemolyticus*: presence of two chromosomes in *Vibrio* species. *Mol Microbiol 31*, 1513–1521.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci 13*, 555–556.

Young, J. P. W. and A. W. B. Johnston (1989). The evolution of specificity in the legume-rhizobium symbiosis. *TREE 4*, 341–349.

Zhou, Z., K. A. White, A. Polissi, C. Georgopoulos, and C. R. Raetz (1998). Function of *Escherichia coli* MsbA, an essential ABC family transporter, in lipid A and phospholipid biosynthesis. *J Biol Chem 273*, 12466–12475.