MULTIPLE GENE GENEALOGICAL ANALYSIS OF

SINORHIZOBIUM MELILOTI

MULTIPLE GENE GENEALOGICAL ANALYSIS OF *SINORHIZOBIUM MELILOTI*

By

SHENG SUN

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

For the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2005) McMASTER UNIVERSITY

(Biology) Hamilton, Ontario

TITLE: Multiple Gene Genealogical Analysis of *Sinorhizobium meliloti*

AUTHOR: SUN, SHENG B.Sc (Xiamen University, China)

SUPERVISOR: Dr. XU, JIANPING (J-P)

NUMBER OF PAGES: viii, 58

ABSTRACT

Sinorhizobium meliloti is an economically important bacterium as it forms nodules on and fixes nitrogen for alfalfa, an important agricultural crop. The complete genome of a laboratory strain, Rm1021, was published in 2001 and this strain was found to have three replicons: a chromosome with 3.65 million base pairs (MB) and two megaplasmids called pSymA (1.35 MB) and pSymB (1.68 MB). In this study, I sequenced 3 genes from each replicon (9 genes total) for each of 33 natural *S. meliloti* strains and analyzed the DNA sequence variation. The mean sequence divergence between strains varied significantly among the nine genes, ranging from 0.11% to 5.02%. Overall, the three genes located on the chromosome showed a lower level polymorphism than those on pSymA and pSymB. My population genetic analyses revealed that: (i) within each of the nine genes, polymorphic nucleotide sites were in significant linkage disequilibrium (LD); (ii) between genes within a replicon, those on the chromosome were in significant LD while those on the two megaplasmids were in linkage equilibrium (LE); and (iii) between genes on different replicons, a variable proportion showed LD. Gene genealogical analysis indicated a lack of host or geographic pattern for the observed molecular variation. My results suggest a dynamic pattern of molecular evolution in the genomes of natural strains of *S. meliloti.*

ACKNOWEDGMENTS

I would like to thank Dr. Jianping Xu for giving me the opportunity to work in his lab for the past two years and for his support, advice and encouragement throughout the course of this study. I would like to express my gratitude to the members of the entire Biology Department, especially the members of Dr. Xu's lab: Zhun, Hong, Rovena, Susan, Xiaogang, and those undergraduate students worked in our lab.

I would like to thank my parents for their continuous support. I also would like to thank my son for enriching our lives and for the joy he has giving me over the last year. Lastly, I would like to thank my wife, Li Xu, for her love, support and understanding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. GENERAL INTRODUCTION

Symbiotic Nitrogen-Fixing Bacteria

Symbiotic bacteria are those that can form mutually beneficial relationships with other organisms. They can be obligate symbionts. For example, the bacterium *Buchnera aphidicola* can grow only within aphids. They can also be non-obligate symbionts, growing in both free-living and symbiotic conditions. All legume symbiotic nitrogen-fixing bacteria are non-obligate symbionts. However, they fix nitrogen only when in a symbiotic condition.

Symbiotic nitrogen-fixing bacteria fall into two main types: those interact with legumes and those with non-legumes. The non-leguminous nitrogen-fixing symbionts include species in the actinomycete *Frankia* spp. and certain species of Cyanobacteria. Bacteria that form nitrogen-fixing symbiotic associations with legumes have been confirmed in more than forty species in 12 genera (Sawada *et al.*, 2003). Comparative analysis of the 16S rRNA gene sequences showed that these taxa were not clustered in one lineage but distributed in the classes α- and β-proteobacteria. These species were dispersed in 9 monophyletic groups, with each group containing both nitrogen-fixing symbionts and free-living species incapable of fixing nitrogen (Sawada *et al.*, 2003). Many of these nitrogen-fixing species are of significant agricultural, environmental, and economical importance. As a result, there are significant research efforts from around the world to understand the processes of biological nitrogen fixation and to enhance the efficiencies of nitrogen fixation. Most studies have focused on species in genera *Azorhizobium*, *Bradyrhizobium*, *Mesorhizobium*, *Rhizobium* and *Sinorhizobium*.

Sinorhizobium meliloti

Sinorhizobium meliloti is one of 12 known species in the genus *Sinorhizobium*. It is a gram-negative α-proteobacterium that can form symbiotic relationships with alfalfa (*Medicago sativa*) and occasionally several other plant species (including those in genera *Medicago*, *Medica* and *Tetrinella*). The genome of a laboratory strain of this species, Rm1021, has been completely sequenced (Galibert *et al.*, 2001). This strain was found to contain three replicons: a 3.65 mega-base (MB) chromosome and two megaplasmids called pSymA (1.35 MB) and pSymB (1.7 MB) (Barnett *et al.*, 2001; Capela *et al.*, 2001; Finan *et al.*, 2001; Galibert *et al.*, 2001). Most genes involved in essential metabolic processes are located on the chromosome and genes involved in symbiotic relationship with plants and substrate utilization are mainly located on pSymA and pSymB.

Population Studies of *S. meliloti*

Most of the methods used in bacterial population studies have been applied to *Sinorhizobium*. These include Multi-locus Enzyme Electrophoresis (MLEE), Randomly Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP) and Restriction Fragment Length Polymorphism (RFLP). Below are brief descriptions of these studies.

Eardly *et al.* (1990) analyzed a collection of 232 strains of *Rhizobium meliloti* (syn. *Sinorhizobium meliloti*) using MLEE with 15 metabolic enzyme genes. A total of 50 distinct electrophoresis types (ETs) were found based on the analysis of 14 polymorphic isoenzymes. Cluster analysis of these ETs revealed two primary divisions. RFLP analysis of *Bam*HI-digested rRNA sequences of 12 representative isolates revealed no significant variation within each of the two divisions but showed substantial differences between these two divisions. Genetic diversities within both divisions were relatively low and no strong linkage disequilibrium among ETs within each division was observed. One division (subdivision B in Eardly et al.'s study) was subsequently named a new species, *Sinorhizobium medicae* by Rome *et al.* in 1996.

Paffetti *et al.* (1996) investigated the genetic diversity of 96 *Rhizobium meliloti* strains isolated from nodules of four *Medicago sativa* varieties. These plants were grown in two different soils in several distinct geographic areas in northern Italy. Strains were analyzed using RAPD, RFLP of the intergenic spacer region (IGS) of the ribosomal operon, and RFLP of a 25-kb region of pSymA megaplasmid that contains the *nod* genes. RAPD analysis revealed considerable variation within each population. They found strains from different soils were more divergent than strains from the same soil. Also, strains from different soil showed a different level of within-population variation. A dendrogram showing the relationships among the 96 strains they analyzed was constructed based on the RAPD result. Their analyses indicated that recombination might have been frequent in Italian *S. meliloti* populations. Interestingly, they observed several differences among the markers. For example, RFLP analysis of the IGS from the 96 strains using *Hae*III digests revealed only three restriction digest patterns, with pattern 1 dominant in one soil type and pattern 2 more prevalent in the other. Pattern 3 was found in only one strain. In contrast, RFLP analysis of the *nod* gene region revealed 7 different restriction digest patterns. Results of the two RFLP analysis indicated that plasmid sequences might be more variable than chromosomal sequences.

Carelli *et al.* (2000) analyzed the genetic diversity of 531 *S. meliloti* strains isolated from nodules of *Medicago sativa* in two different Italian soils during a 4-year period. Using RAPD markers, they found a high level of genetic polymorphism. Analysis of molecular variance with regard to factors such as soil type, alfalfa cultivar, individual plants within a cultivar, and time showed that the population structure changed during the 4-year period. At the beginning, soil and cultivar contributed significantly to the distribution of genetic variation. However, after 3 years, the genetic structure was influenced mainly by individual plants and soil while host cultivar types contributed little to the overall variation. They found that the influence of soil and cultivar on the patterns of genetic variation in *S. meliloti* population may not act independently, but rather the growth of a particular cultivar in a particular soil may be responsible for the genetic differences among the symbiotic populations. This study demonstrated the importance of the plant partner in determining the genetic structure of a symbiotic microbial population and the importance of monitoring for longer periods of time in the analyses of symbiotic bacterial populations.

Jebara *et al* (2001) analyzed the megaplasmid genotype profiles of a collection of 134 *Sinorhizobium* isolates obtained from different ecological areas of Tunisia. 89

isolates were further analyzed by MLEE and Southern hybridization using probes from the 16S rRNA, IGS and *nif*KD genes. Their results identified 18 different plasmid types. MLEE analysis gave similar results to those of Eardly *et al.* (1990) and identified two divisions that correspond to *S. meliloti* and *S. medicae*, respectively. RFLP analyses identified 2, 62 and 33 digestion patterns respectively using probes from the 16S rRNA, IGS and *nif*KD probes. There was little correlation between diversities revealed by *nif*KD and IGS probes, indicating horizontal transfer and recombination might have occurred in the population.

Roumiantseva *et al.* (2002) investigated 27 isolates collected from soils and nodules in western Tajikistan, a center of diversity for host plants of *S. meliloti.* The genetic diversity was assessed by plasmid profiling, *Rsa*I digestion of amplified 16S rRNA gene, and RFLP of 10 single-copy loci and 4 insertion sequence (IS) elements. Isolates trapped from soil differed significantly from those from field nodules in their genotype frequencies. Their analyses identified 7 chromosomal types, 15 pSymA types and 6 pSymB types. In their RFLP analysis of the 10 single-copy loci, the most common allele in this sample corresponded to that of the model laboratory strain Sm1021. Their analysis suggested a low level recombination and significant linkage disequilibrium was found in 37 of the 78 pair wise loci comparisons.

In a study conducted by Biondi *et al.* (2003), 30 *Sinorhizobium* isolates (including both *S. meliloti* and *S. medicae*) were collected from soil of centers of legume diversity located in the Caucasus, Tajikistan and Siberia. They used RFLP of *nodD* genes, IGS fingerprinting and AFLP to investigate the evolutionary relationship between *S. meliloti* and *S. medicae.* RFLP analysis of both the *nodD* genes and the IGS revealed relatively high levels of genetic diversity in this population. The diversity was found higher in *S. meliloti* than in *S. medicae.*

In summary, using a variety of molecular markers, studies have shown that there are relatively high levels of genetic diversity within natural populations of *S. meliloti.* Analysis of the collected genetic information revealed a predominantly clonal population structure in *S. meliloti* but with a low level genetic recombination. However, because of the timing of these studies and/or the nature of the analyzed markers, little is known about the potential differences in population structure inferred from markers located on different replicons.

Multi-locus Sequence Typing

Multilocus sequence typing (MLST) is a powerful method for strain typing and for analyzing the structure of microbial populations (Maiden *et al.* 1998; Cooper *et al.* 2004). Many species have been analyzed using this method, including human pathogens *Neisserria meningitidis*, *Staphylococcus aureus*, *Escherichia coli*, *Cryptococcus neoformans*, and environmental microbes such as *Campylobacter jejuni* (Xu *et al.* 2000; Colles *et al.* 2003; for a recent review, see Cooper *et al.* 2004). Compared to other molecular markers, data generated by MLST are unambiguous, can be easily stored in public databases, and are readily shared among researchers. In a study conducted by Vinuesa et al (2005), four loci (*atpD*, *glnII*, *recA*, and *nifH*) were sequenced for a collection of *Bradyrhizobium* species. Phylogenetic inference revealed that the four loci yield significantly incongruent topologies, indicating the occurrence of recombination in *Bradyrhizobium* species (Vinuesa et al., 2005). To my knowledge, there is no published MLST study of *S. meliloti*.

This work: MLST Study of *S. meliloti*

In this study, I used MLST to examine the patterns of DNA sequence variation in a collection of natural strains of the nitrogen-fixing bacterium *S. meliloti*. Three genes from each of the three replicons were analyzed for each of 33 strains. These strains were previously analyzed by Eardly *et al.* (1990) using the technique of multilocus enzyme electrophoresis (MLEE). In this study, I attempt to address the following questions. First, how much divergence is there among strains in *S. meliloti*? Do genes on all three replicons show similar levels of divergence? Second, do genes on the same replicon show similar relationships among strains? And, do genes on different replicons show different relationships among strains? Third, is there any evidence of a phylogenetic pattern based on geographic origin or host species within this collection of *S. meliloti* strains?

CHAPTER 2. MATERIALS AND METHODS

Strains

The 33 isolates of *S. meliloti* analyzed in this study were part of the collection used for the MLEE study reported by Eardly *et al.* (1990). The MLEE type, geographic origin and host plant species of these strains are presented in Table 1. Each strain represents a different MLEE type. Dr. B. D. Eardly of Pennsylvania State University, Berks-Lehigh Valley College, Reading, Pennsylvania (USA) kindly provided us these strains.

DNA Isolation

For each isolate, storage culture from a -70°C freezer was first streaked onto TY (Trypton-Yeast extract) agar plate and incubated at 30°C. For each strain, a single colony was picked to inoculate a liquid LBmc broth (per liter: 10 grams of pancreatic digest of casein, 5 grams of NaCl, 5 grams of yeast extract, 2.5mM $MgSO_4$ and 2.5mM $CaCl_2$, pH7). Cells were incubated at 30°C with constant agitation at 120rpm and harvested through centrifugation when population density reached an OD600 reading between 0.8-1.0. Genomic DNA was extracted using a method modified from the previously described for *S. meliloti* (Eardly *et al.* 1990). The quantity and quality of DNA were assessed using the UltraSpec 2000 *pro* spectrophotometer.

Table 1. Strains used in this study[1]

Strain	Multilocus enzyme electrophoretic type[2] (ET)	Original host species (genus *Medicago*)	Geographic Origin[3]
9930	1	*M. sativa*	USA
M56	2	*M. rotato*	SYR
A145	3	*M.sativa*	SYR
M98	4	*M. rotato*	SYR
M275	5	*M. rigidula*	JOR
M270	6	*M. truncatula*	JOR
102F85	7	*M.sativa*	CAN
74B3	8	*M. sativa*	PAK
N6B1	9	*M. falcate*	NPL
M95	10	*M. rotato*	SYR
128A7	11	*M. sativa*	PAK
56A14	12	*M. sativa*	PAK
M286	15	*M. rotato*	JOR
M289	16	*M. truncatula*	JOR
15B4	17	*M.sativa*	PAK
128A10	18	*M. sativa*	PAK
N6B5	19	*M. falcate*	NPL
N6B11	20	*M. falcate*	NPL
17B6	21	*M. sativa*	PAK
N6B9	22	*M. falcate*	NPL
1322	23	*M. sativa*	NZL
CC2003	24	*M. sativa*	AUS
N6B4	25	*M. sativa*	NPL
M248	26	*M.polymorpha*	JOR
15A5	27	*M.sativa*	PAK
M294	28	*M.polymorpha*	JOR
M119	29	*Unspecified*	SYR
S33	30	*M.sativa*	USA
102F51	31	*M.sativa*	USA
74B4	32	*M.sativa*	PAK
74B12	33	*M.sativa*	PAK
74B15	34	*M.sativa*	PAK
CC2013	35	*M.sativa*	AUS

[1], Modified from Eardly et al. (1990)

[2], Electrophoretic types refer those defined by Eardly et al. (1990)

[3], Country names: AUS, Australia; CAN, Canada; JOR, Jordan; NPL, Nepal; NZL, New Zealand; PAK, Pakistan; SYR, Syria; USA, United States;

PCR

Ten pairs of primers were used in this study to amplify nine DNA fragments from all 33 isolates (Table 2). The primers were designed based on the genome sequence of strain Rm1021 (http://bioinfo.genopole-toulouse.prd.fr/annotation/iANT/bacteria/rhime/). Three genes were chosen from each replicon and these genes occupy distinct locations in the genome (Figure 1, Table 2). The exoF3-2 primer pair was used for isolates M56, M275, N6B9, CC2003, and M294 because the initial exoF3-1 primer pair did not work for these five strains. The exoF3 gene fragments in the remaining 28 strains were amplified and sequenced using the exoF3-1 primer pair.

A typical PCR reaction contained 6 μl of diluted genomic DNA template (~ 20 ng), 0.5unit Taq DNA polymerase, 1 μM of each primer and 200 μM of each of the four deoxyribonucleotide triphosphates in a total volume of 30 μl. The following PCR conditions were used: 4 min at 95°C, followed by 30 cycles of 30s at 95°C, 30s at 56°C (at 50°C for exoF3-2), 45s at 72°C, and finally 7min at 72°C.

Table 2. Primers and sequences used in this study

Gene			Primer Sequence (5'—>3')	Position in Database
OxyR	Hydrogen Peroxide-inducible Gene - Activator	F	AGGCGGATATGGCGTTTGCA	839182-840117[a] (839259-840017, 759bp)[b]
		R	TGGAAGAACATCTGGGCGTGA	
Aqpz1	Aquaporin Z (Bacterial Nodulin-Like Intrinsic Protein)	F	GGCACTCGAGTATGCGTCGAGCCAAGAATGATGAG	2339739-2340425 (2339959-2340428, 470bp)
		R	TTCAAGATCTGGAAGCTCTCTGTGGAATTTC	
mdh	Malate Dehydrogenase	F	GCACGCGCTTCTTGTCCTTGA	3318498-3319406 (3318701-3319431, 731bp)
		R	TTCGGGGATGATTGGTGGCA	
CbbR	Transcriptional Regulator	F	AAGGATGGCGCAAAAGGGGA	212468-213409 (212494-213258, 765bp)
		R	TGATCGTCTCGTTCGAAGCGA	
ExoF3	Putative OMA Family Outer Membrane Protein Precursor	1-F	TTCCTTGACGATGCCGAGCTG	813621-814871 (813783-814607, 822bp)
		1-R	TGCAAGCTTTGCGAGCTGCA	
		2-F	ACTTCCTTGACGATGCCGAG	813621-814871 (813781-814694, 914bp)
		2-R	TTCGGCGGAGTGTTTTCCAG	
MinC	Putative Cell Division Inhibitor	F	CCCTCTAGAAGCGTCCCGTAGATATG	1447281-1448060 (1447462-1448050, 589bp)
		R	CCCCGGATCCGCTAGCAATAATTAACGAAGATG	
FdhE	Probable FdhE Formate Dehydrogenase Formation	F	AAGCCGAATTTGGCACGCCT	6149-7069 (6218-7064, 847bp)
		R	AGCCCATCAGGAACGGGTCAA	
NifH	Nitrogenase Fe Protein	F	CCGAACAACCGAAATAGCTTAAAC	453556-454449 (453517-454404, 888bp)
		R	AAGCATCTGCTCGTCGCTCTTCATG	
sma1440	5-dehydro-4-deoxyglucarate Dehydratase	F	CGCCAGTTCCGGCACGAAATT	794368-795373 (794608-795362, 755bp)
		R	CGAGCGAAAAAACCGATGCG	

[a]: Start and stop positions of the gene. Data from GenBank

[b]: Start and stop positions of the sequenced region and their exact length are presented in parenthesis.

Cleaning and Sequencing of PCR products

The PCR products were cleaned using the PCR cleanup kit (DiaMed) according to the manufacturer's manual. The purified PCR products were sequenced using an Applied Biosystems Prism 3100 automated sequencer with dRhodamine-labeled terminators (PE Applied Biosystems), following the manufacturer's instructions.

Data Analyses

Sequence Variation

The analyses of pair-wise DNA sequence variation between strains were performed using the computer program PAUP 4.0 (Swofford, 2004).

Construction of Phylogenetic Trees

Phylogenetic trees were constructed using the maximum parsimony method implemented in PAUP 4.0 (Swofford, 2004) and maximum likelihood method implemented in the PHYLIP software (Felsenstein, 2005, http://evolution.gs.washington.edu/phylip/software.html).

Linkage Disequilibrium Analysis

Two computer programs, LIAN (LInkage Analysis, Haubold and Hudson, 2000) and MultiLocus (Agapow *et al.* 2001), were used to perform the linkage disequilibrium analysis. The bases of these two programs are similar and both compute the standardized index of association (I_A^s in LIAN and r_d in MultiLocus), a measure of multilocus linkage disequilibrium.

The traditional I_A was defined by Maynard Smith *et al.* (1993) as indicates such a population has a structure not significantly different from random recombination.

Three levels of the index of association analysis were performed: (i) among nucleotide sites within a gene, (ii) among genes within a replicon, and (iii) between genes on different replicons.

For level (i), two kinds of nucleotide sites (all nucleotide sites and all polymorphic nucleotide sites), were analyzed using the LIAN program. Because of the limitation of the number of loci (nucleotides) that can be analyzed by the MultiLocus software, only polymorphic nucleotide sites were analyzed by the MultiLocus program. These two programs and the two different kinds of nucleotide sites in LIAN analysis gave consistent results (see below).

For subsequent levels (ii) and (iii), it is necessary to define all nucleotides within each gene as one linkage group to eliminate the influence of the clonal population structure inferred within each gene on the analysis of between genes and replicons. Because the function for defining linkage groups is not available in LIAN, I only used the MultiLocus software and polymorphic nucleotide sites for these two levels of analyses.

Incompatibility Ratio Analysis

Another method called incompatibility ratio analysis (Maynard Smith, 1999) was also used to infer the population structure. In the simplest case in a haploid species, assuming two loci (A and B) with two alleles each (A1 and A2; B1 and B2), if all four possible genotypes (A1B1, A1B2, A2B1, and A2B2) are found in the population, these two loci are considered incompatible and are indicators of recombination at the population level. Incompatibility ratio (IR), where IR=(number of incompatible pairs of sites in the test data set)/(number of incompatible pairs of sites in a shuffled data set), can be used as a statistic of the population structure (Maynard Smith, 1999). Similar to the linkage disequilibrium analysis, this analysis also tests against the null hypothesis of a randomly recombining population structure. For each test, 1000 randomizations were performed and 95% confidence interval was generated. A *P* value greater than 0.05 indicates the analyzed population in linkage equilibrium.

Shimodaira and Hasegawa Test (SHTest)

The above two methods compare the observed data against the null hypothesis of random mating. In small populations with highly skewed allele frequencies, these tests have a significant type II error – the error to accept a null but false hypothesis. Because the sample size here is relatively small and singleton alleles are common, to minimize this error, I also used a different but complementary test, the Shimodaira and Hasegawa Test (SHTest, Shimodaira and Hasegawa, 1999), to examine the association among haplotypes in different genes. The null hypothesis of SHTest is strict clonality, different from the linkage disequilibrium analysis and incompatibility ratio test. This analysis was done using SHTest software (http://evolve.zoo.ox.ac.uk/software.html?id=shtest).

To perform this test, we first examined the models of sequence variation for individual genes using the program ModelTest (Posada and Crandall, 1998). The inferred models were then used to construct the maximum likelihood tree for each gene using PHYLIP. These ML trees from different genes were then compared against each other for phylogenetic congruency through the SHTest. Congruent phylogenies indicate lack of recombination and a clonal population structure.

Specifically, in the SHTest, when two topologies are compared, one of them is the most likelihood tree for the sequence. The program will calculate the likelihood (L) of the test tree as well as δ of that topology ($\delta=L_{ML}-L$). Statistical significance of this test is derived using the nonparametric bootstrap replicate data sets. If the test topology's δ falls within the 95% confidence interval of the δ distribution obtained from the bootstrap data sets, this topology is considered to be no worse than the maximum likelihood tree in explaining the sequence and the two topologies are considered congruent.

Isolation by Distance and Isolation by Host Plant Species

In addition to the analysis of clonality and recombination, I also examined the potential patterns of molecular variation based on geographic distances and host plant species. These potential patterns were tested using simple Mantel Test whose website interface is available at http://phage.sdsu.edu/%7Ejensen/ (Jensen *et al.* 2005). In these tests, pair-wise genetic distances between strains of *S. meliloti* were compared with the corresponding geographic distances between the sampling sites and the corresponding phylogenetic genetic distances between the host plants.

The evolutionary genetic distance between plant species was calculated using DNA sequences of three regions, IT1, IT2 and ETS. These sequences were retrieved from GenBank.

CHAPTER 3. RESULTS

Overall Sequence Variation among Genes and Strains

My sequence analysis identified that strain CC2013 is significantly different from the other 32 strains (Figure 2), similar to what was found in the study by Eardly *et al* using MLEE data (Eardly *et al.* 1990). Therefore, to avoid skewed data that may bias statistical analyses, I excluded this strain in the analysis of sequence divergence among genes and strains and in the analysis of linkage disequilibria. However, this strain will be used as an outgroup in my phylogenetic analysis.

In all, I obtained a total of 6309 nucleotides from the nine genes for each strain. Based on the combined sequence of genes from each replicon, I found 14 unique sequence types on the Chromosome, 29 types on pSymA and 24 types on pSymB. The combined sequence data from all nine genes identified that each of the 33 strains had a unique multilocus sequence type (Table 3).

The mean pair-wise distance between strains was calculated for each gene (Table 4). The smallest was found in the *mdh* gene located on the Chromosome (mean = 0.0034 mismatch per nucleotide site with a standard deviation of 0.0053). The largest was found in the *exoF3* gene located on pSymB (mean= 0.0522 mismatch per nucleotide site with standard deviation of 0.064).

To investigate whether different genes have different levels of among-strain divergence, I performed one-tailed t-tests for each of the 36 gene pairs (9x8/2). The t-values are presented in Table 5. Of the 36 pair-wise comparisons, all except four (*aqpz1– mdh*, *aqpz1 – oxyR*, *cbbR – sma1440* and *fdhE – nifH*) showed statistically significant difference in the degrees of divergence between strains (Table 5).

When genes on different replicons were considered, those on the two megaplasmids showed, on average, a greater divergence than genes located on the Chromosome (Table 5). However, in the comparisons between genes on pSymA and pSymB, a mixed pattern was found.

The *minC* gene on pSymB showed a lower mean divergence than all three examined genes on pSymA. In contrast, the *exoF3* gene on pSymB showed the greatest divergence among the nine genes, including all three examined genes on pSymA. The third gene on pSymB, *cbbR,* showed an intermediate divergence between *minC* and *exoF3* (Table 5).

Fig 2A: ML tree (left) and one most parsimonious tree (right) for gene *oxyR*

(Based on sequence of 711 bp and bootstrap values higher than 50 are listed)

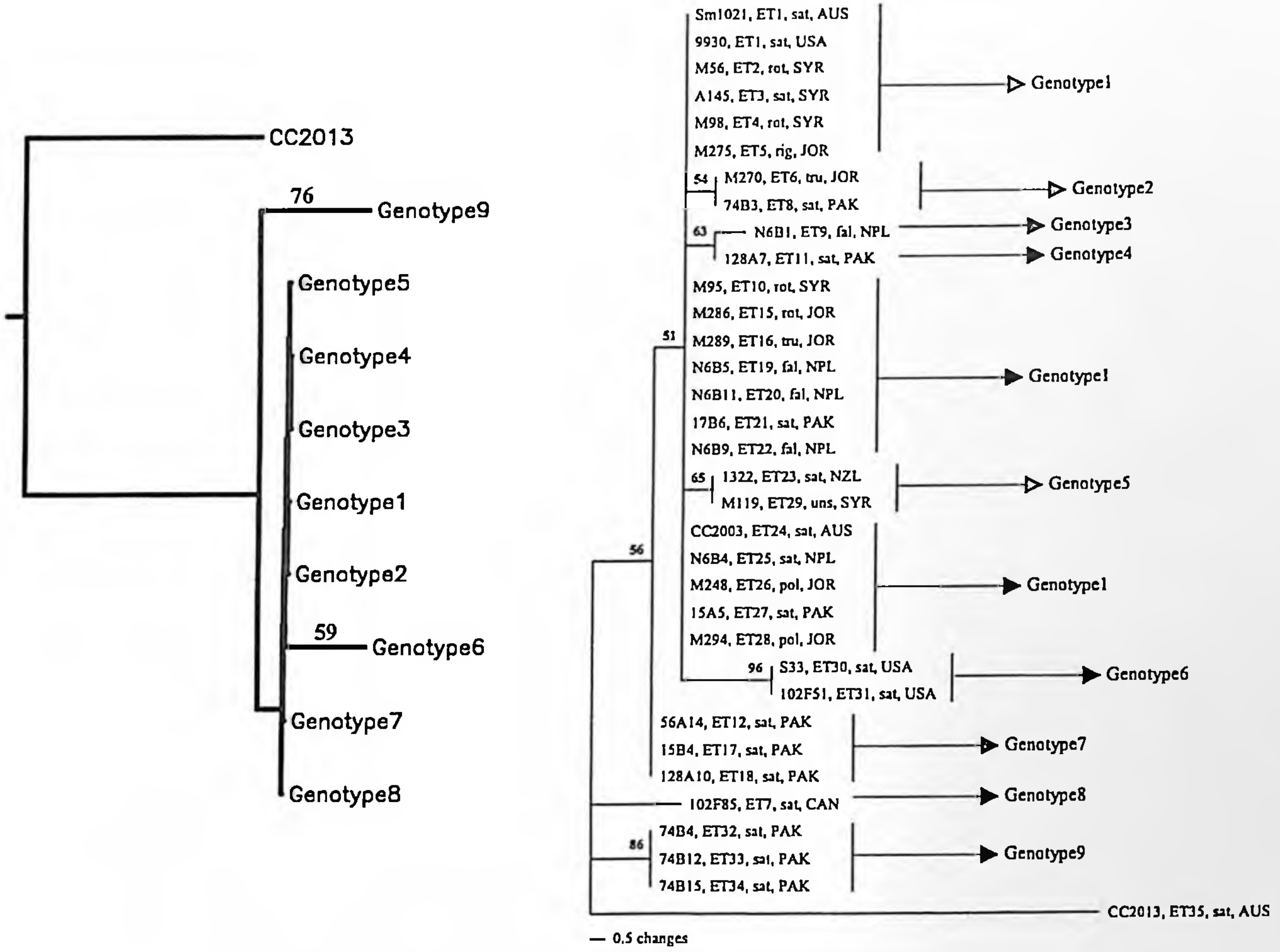

CC2013
76
Genotype9
Genotype5
Genotype4
Genotype3
Genotype1
Genotype2
59
Genotype6
Genotype7
Genotype8
Sm1021, ET1, sat, AUS
9930, ET1, sat, USA
M56, ET2, rot, SYR
A145, ET3, sat, SYR
M98, ET4, rot, SYR
M275, ET5, rig, JOR
Genotype1
54
M270, ET6, tru, JOR
74B3, ET8, sat, PAK
Genotype2
63
N6B1, ET9, fal, NPL
Genotype3
128A7, ET11, sat, PAK
Genotype4
M95, ET10, rot, SYR
M286, ET15, rot, JOR
51
M289, ET16, tru, JOR
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
17B6, ET21, sat, PAK
N6B9, ET22, fal, NPL
Genotype1
65
1322, ET23, sat, NZL
M119, ET29, uns, SYR
Genotype5
CC2003, ET24, sat, AUS
56
N6B4, ET25, sat, NPL
M248, ET26, pol, JOR
15A5, ET27, sat, PAK
M294, ET28, pol, JOR
Genotype1
96
S33, ET30, sat, USA
102F51, ET31, sat, USA
Genotype6
56A14, ET12, sat, PAK
15B4, ET17, sat, PAK
128A10, ET18, sat, PAK
Genotype7
102F85, ET7, sat, CAN
Genotype8
86
74B4, ET32, sat, PAK
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
Genotype9
CC2013, ET35, sat, AUS
0.5 changes

Fig.2A oxyR

Fig 2B: ML tree (left) and one most parsimonious tree (right) for each of gene *aqpz1*

(Based on sequence of 435 bp and bootstrap values higher than 50 are listed)

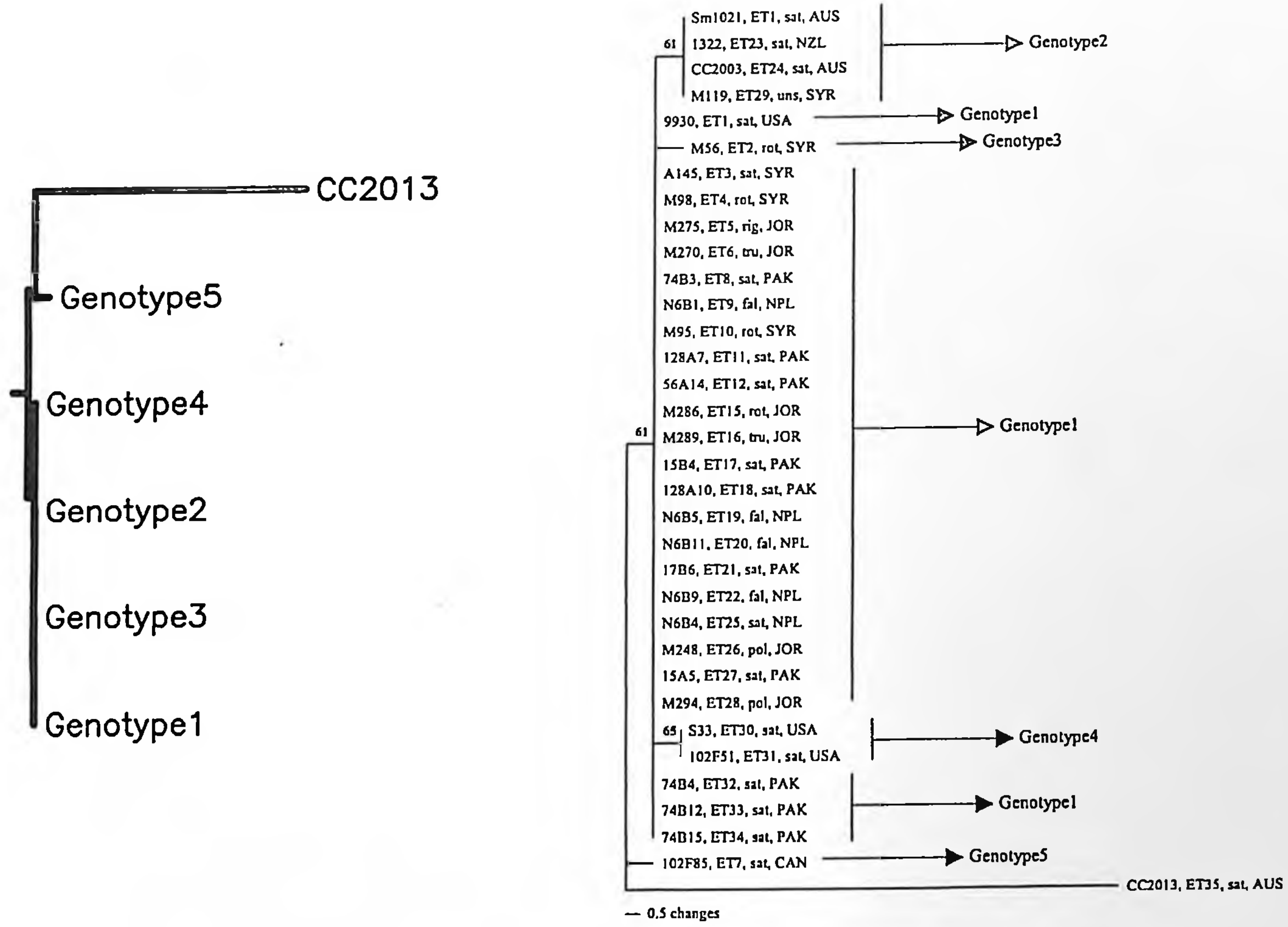

CC2013
Genotype5
Genotype4
Genotype2
Genotype3
Genotype1
61
Sm1021, ET1, sat, AUS
1322, ET23, sat, NZL
CC2003, ET24, sat, AUS
M119, ET29, uns, SYR
Genotype2
9930, ET1, sat, USA
Genotype1
M56, ET2, rot, SYR
Genotype3
A145, ET3, sat, SYR
M98, ET4, rot, SYR
M275, ET5, rig, JOR
M270, ET6, tru, JOR
74B3, ET8, sat, PAK
N6B1, ET9, fal, NPL
M95, ET10, rot, SYR
128A7, ET11, sat, PAK
56A14, ET12, sat, PAK
M286, ET15, rot, JOR
61
M289, ET16, tru, JOR
Genotype1
15B4, ET17, sat, PAK
128A10, ET18, sat, PAK
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
17B6, ET21, sat, PAK
N6B9, ET22, fal, NPL
N6B4, ET25, sat, NPL
M248, ET26, pol, JOR
15A5, ET27, sat, PAK
M294, ET28, pol, JOR
65
S33, ET30, sat, USA
102F51, ET31, sat, USA
Genotype4
74B4, ET32, sat, PAK
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
Genotype1
102F85, ET7, sat, CAN
Genotype5
CC2013, ET35, sat, AUS
0.5 changes

Fig.2B aqpzl

Fig 2C: ML tree (left) and one most parsimonious tree (right) for each of gene *mdh*

(Based on sequence of 693 bp and bootstrap values higher than 50 are listed)

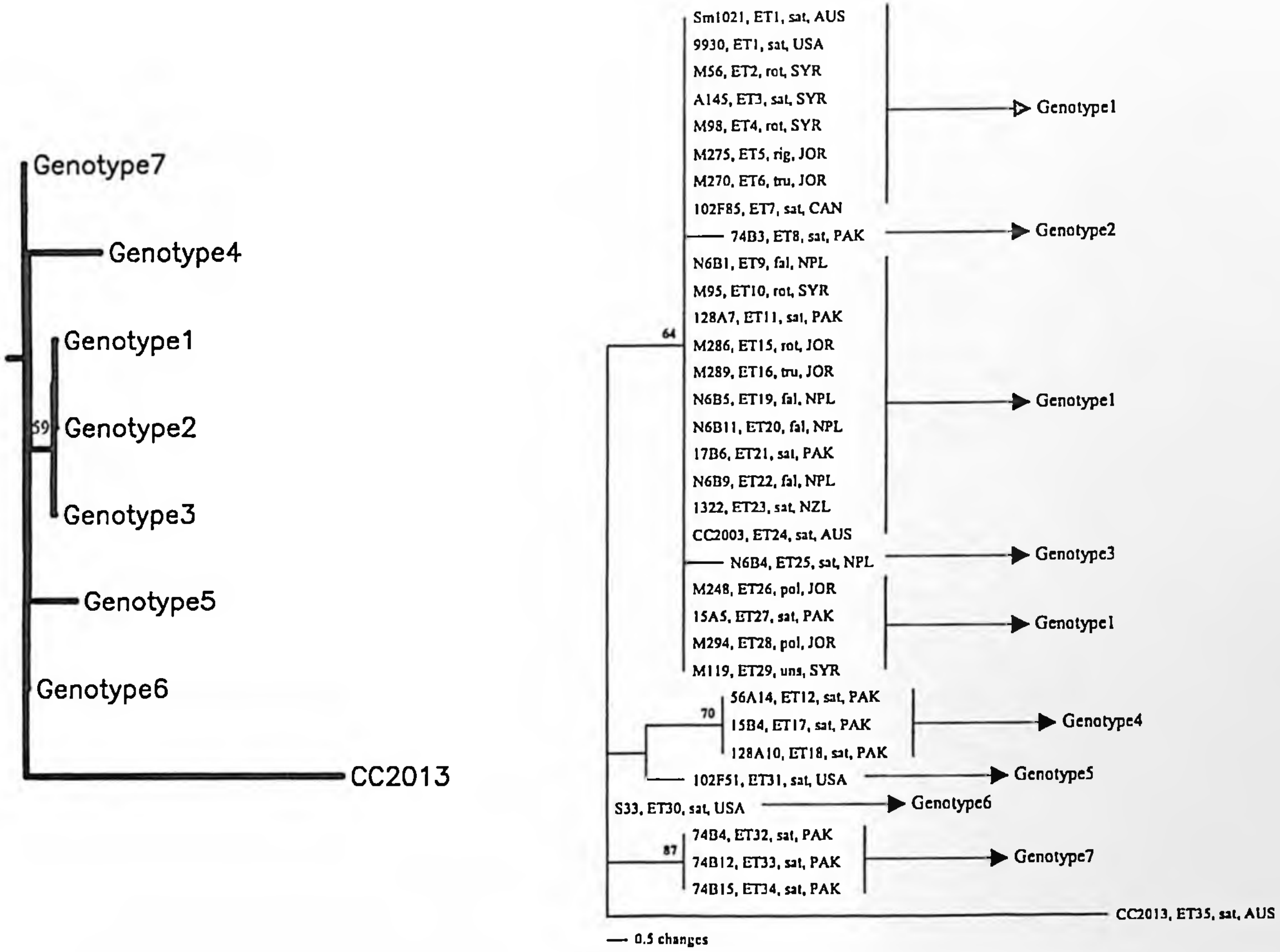

Genotype7
Genotype4
Genotype1
59
Genotype2
Genotype3
Genotype5
Genotype6
CC2013
Sm1021, ET1, sat, AUS
9930, ET1, sat, USA
M56, ET2, rot, SYR
A145, ET3, sat, SYR
M98, ET4, rot, SYR
M275, ET5, rig, JOR
M270, ET6, tru, JOR
102F85, ET7, sat, CAN
74B3, ET8, sat, PAK
N6B1, ET9, fal, NPL
M95, ET10, rot, SYR
128A7, ET11, sat, PAK
64
M286, ET15, rot, JOR
M289, ET16, tru, JOR
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
17B6, ET21, sat, PAK
N6B9, ET22, fal, NPL
1322, ET23, sat, NZL
CC2003, ET24, sat, AUS
N6B4, ET25, sat, NPL
M248, ET26, pol, JOR
15A5, ET27, sat, PAK
M294, ET28, pol, JOR
M119, ET29, uns, SYR
56A14, ET12, sat, PAK
70
15B4, ET17, sat, PAK
128A10, ET18, sat, PAK
102F51, ET31, sat, USA
S33, ET30, sat, USA
74B4, ET32, sat, PAK
87
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
CC2013, ET35, sat, AUS
0.5 changes
Genotype1
Genotype2
Genotype1
Genotype3
Genotype1
Genotype4
Genotype5
Genotype6
Genotype7

Fig.2C mdh

Fig 2D: ML tree (left) and one most parsimonious tree (right) for each of gene *fdhE*

(Based on sequence of 810 bp and bootstrap values higher than 50 are listed)

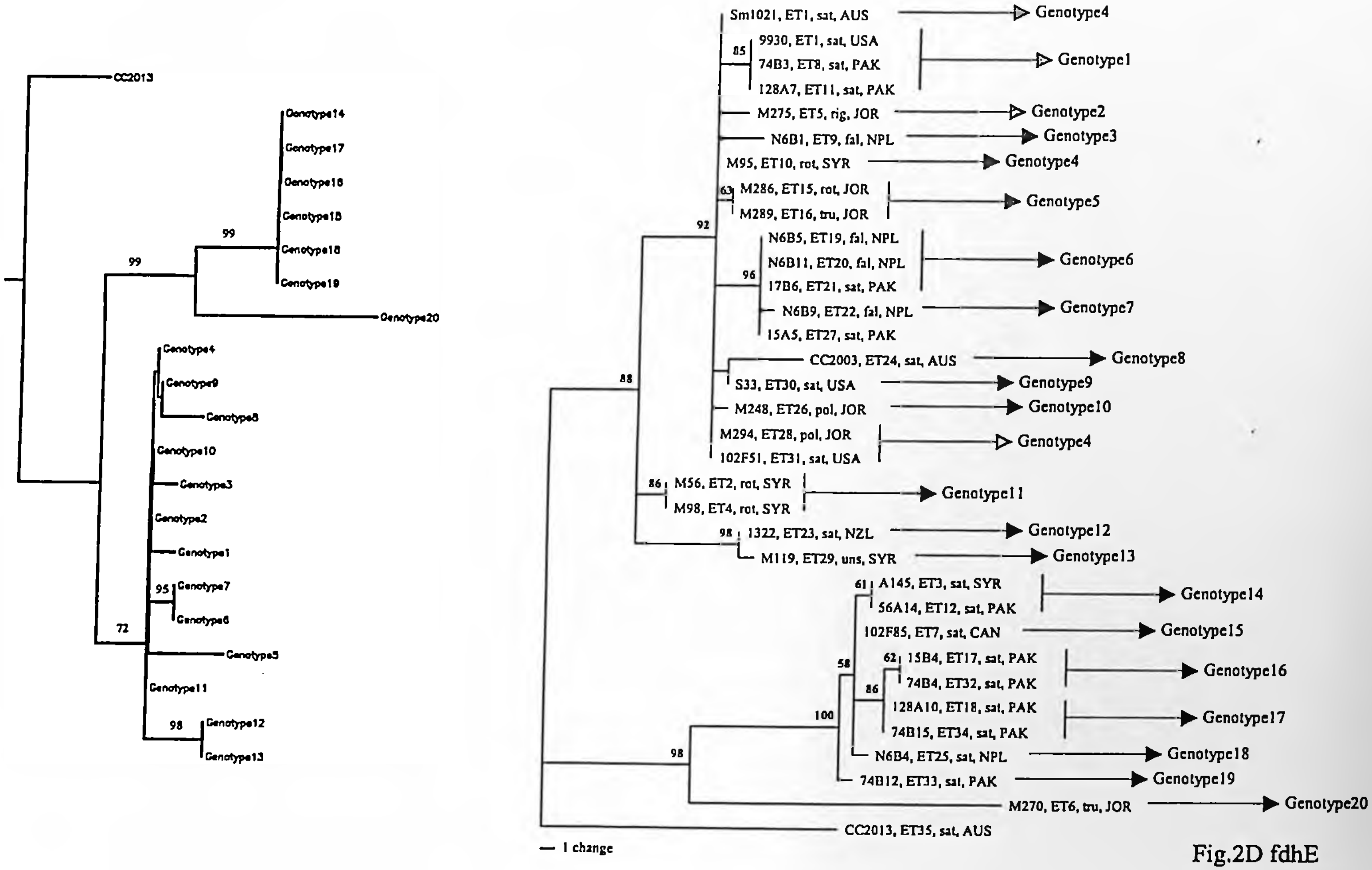
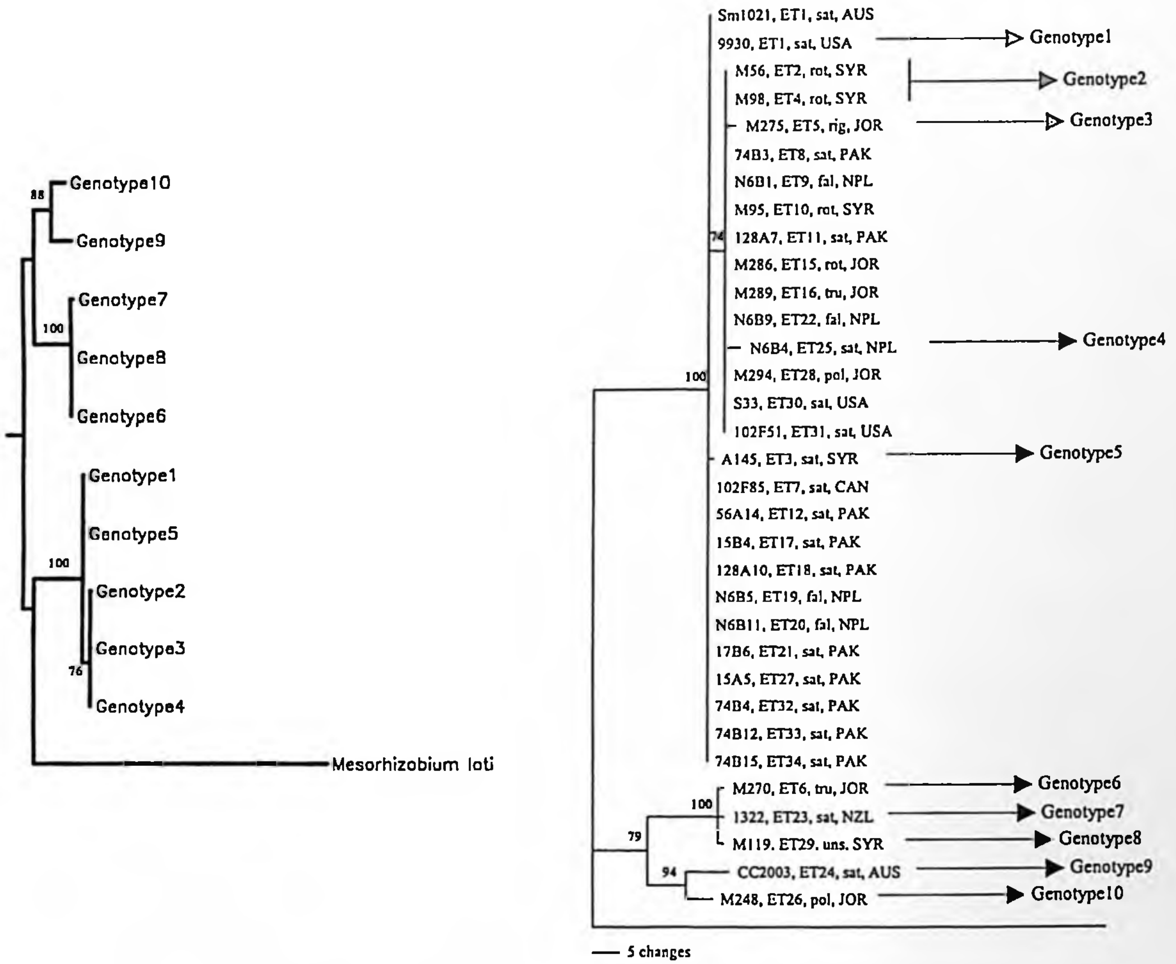

CC2013
Genotype14
Genotype17
Genotype16
Genotype15
Genotype18
Genotype19
Genotype20
99
99
Genotype4
Genotype9
Genotype8
Genotype10
Genotype3
Genotype2
Genotype1
95
Genotype7
Genotype6
Genotype5
72
Genotype11
98
Genotype12
Genotype13
Sm1021, ET1, sat, AUS
Genotype4
85
9930, ET1, sat, USA
74B3, ET8, sat, PAK
128A7, ET11, sat, PAK
Genotype1
M275, ET5, rig, JOR
Genotype2
N6B1, ET9, fal, NPL
Genotype3
M95, ET10, rot, SYR
Genotype4
63
M286, ET15, rot, JOR
M289, ET16, tru, JOR
Genotype5
92
96
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
17B6, ET21, sat, PAK
Genotype6
N6B9, ET22, fal, NPL
Genotype7
15A5, ET27, sat, PAK
CC2003, ET24, sat, AUS
Genotype8
88
S33, ET30, sat, USA
Genotype9
M248, ET26, pol, JOR
Genotype10
M294, ET28, pol, JOR
102F51, ET31, sat, USA
Genotype4
86
M56, ET2, rot, SYR
M98, ET4, rot, SYR
Genotype11
98
1322, ET23, sat, NZL
Genotype12
M119, ET29, uns, SYR
Genotype13
61
A145, ET3, sat, SYR
56A14, ET12, sat, PAK
Genotype14
102F85, ET7, sat, CAN
Genotype15
58
62
15B4, ET17, sat, PAK
74B4, ET32, sat, PAK
Genotype16
86
128A10, ET18, sat, PAK
74B15, ET34, sat, PAK
Genotype17
100
N6B4, ET25, sat, NPL
Genotype18
98
74B12, ET33, sat, PAK
Genotype19
M270, ET6, tru, JOR
Genotype20
CC2013, ET35, sat, AUS
1 change

Fig.2D fdhE

Fig 2E: ML tree (left) and one most parsimonious tree (right) for each of gene *nifH*

(Based on sequence of 843 bp and bootstrap values higher than 50 are listed)



Genotype10
Genotype9
Genotype7
Genotype8
Genotype6
Genotype1
Genotype5
Genotype2
Genotype3
Genotype4
Mesorhizobium loti
88
100
100
76
Sm1021, ET1, sat, AUS
9930, ET1, sat, USA
M56, ET2, rot, SYR
M98, ET4, rot, SYR
M275, ET5, rig, JOR
74B3, ET8, sat, PAK
N6B1, ET9, fal, NPL
M95, ET10, rot, SYR
128A7, ET11, sat, PAK
M286, ET15, rot, JOR
M289, ET16, tru, JOR
N6B9, ET22, fal, NPL
N6B4, ET25, sat, NPL
M294, ET28, pol, JOR
S33, ET30, sat, USA
102F51, ET31, sat, USA
A145, ET3, sat, SYR
102F85, ET7, sat, CAN
56A14, ET12, sat, PAK
15B4, ET17, sat, PAK
128A10, ET18, sat, PAK
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
17B6, ET21, sat, PAK
15A5, ET27, sat, PAK
74B4, ET32, sat, PAK
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
M270, ET6, tru, JOR
1322, ET23, sat, NZL
M119. ET29. uns. SYR
CC2003, ET24, sat, AUS
M248, ET26, pol, JOR
Genotype1
Genotype2
Genotype3
Genotype4
Genotype5
Genotype6
Genotype7
Genotype8
Genotype9
Genotype10
74
100
100
79
94
5 changes

Fig.2E nifH

Fig 2F: ML tree (left) and one most parsimonious tree (right) for each of gene

sma1440

(Based on sequence of 721 bp and bootstrap values higher than 50 are listed)

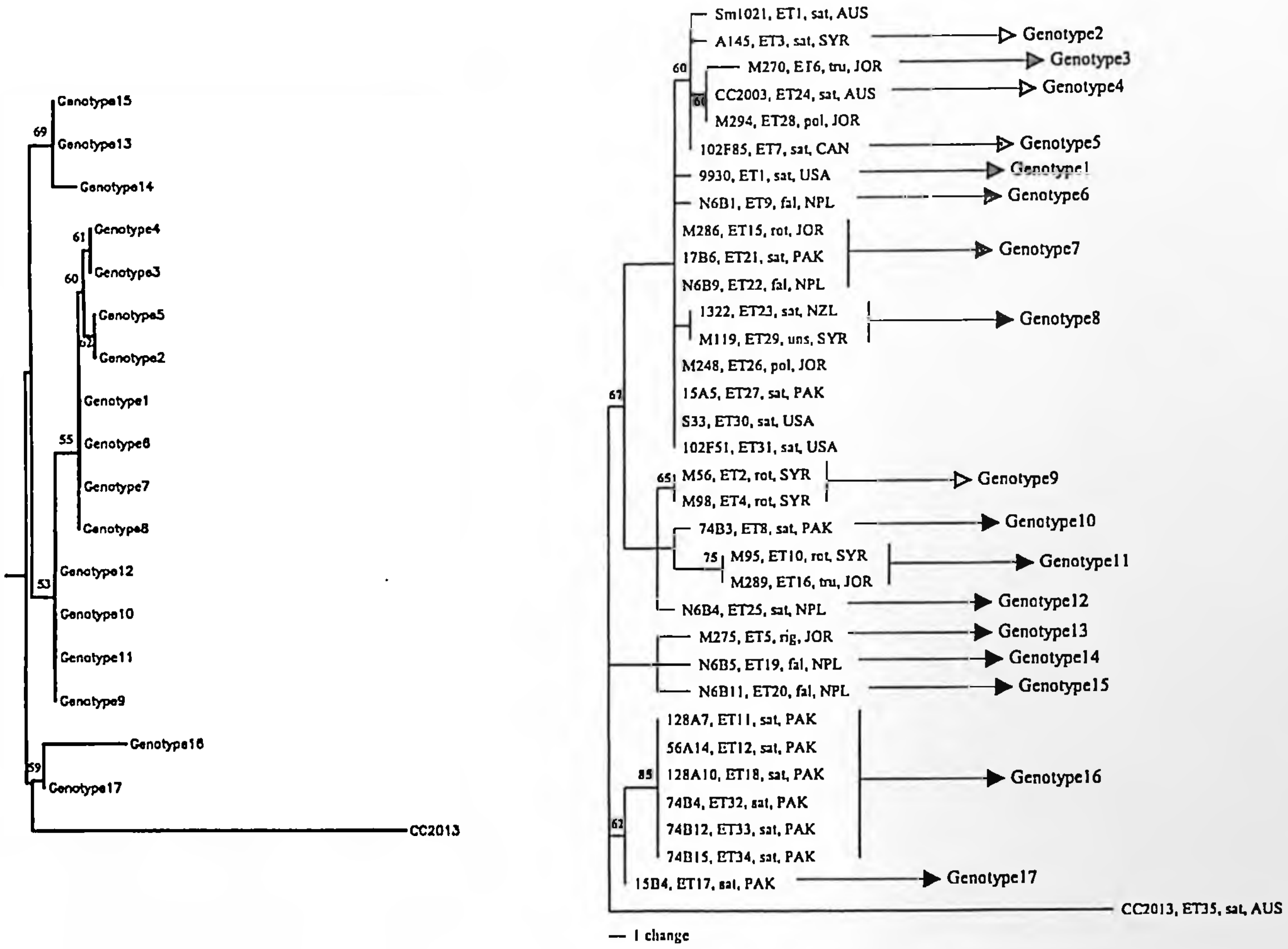


Fig.2F sma1440

Fig 2G: ML tree (left) and one most parsimonious tree (right) for each of gene *cbbR*

(Based on sequence of 729 bp and bootstrap values higher than 50 are listed)

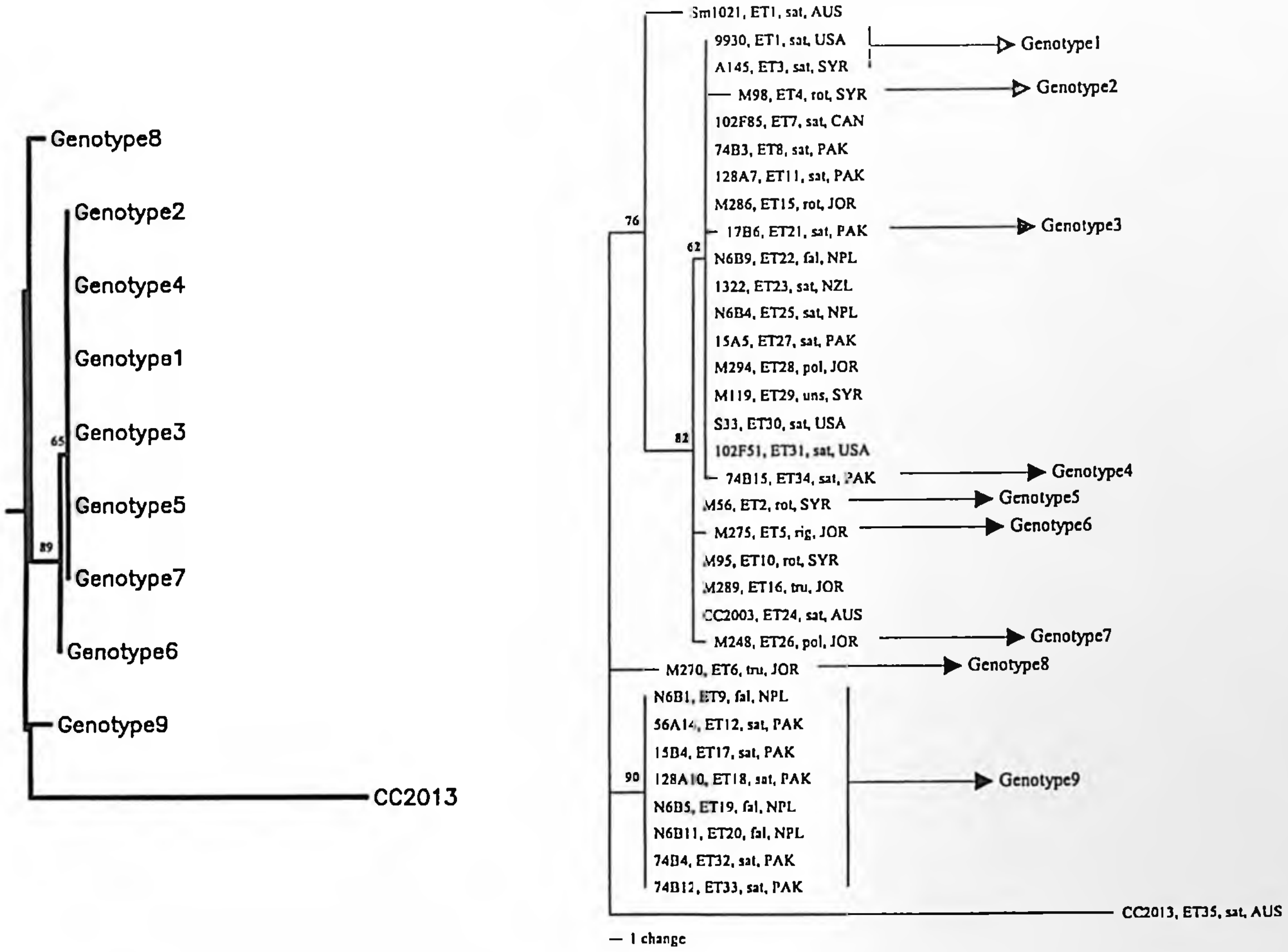

Genotype8
Genotype2
Genotype4
Genotype1
65
Genotype3
Genotype5
89
Genotype7
Genotype6
Genotype9
CC2013
Sm1021, ET1, sat, AUS
9930, ET1, sat, USA
A145, ET3, sat, SYR
Genotype1
M98, ET4, rot, SYR
Genotype2
102F85, ET7, sat, CAN
74B3, ET8, sat, PAK
128A7, ET11, sat, PAK
M286, ET15, rot, JOR
76
17B6, ET21, sat, PAK
Genotype3
62
N6B9, ET22, fal, NPL
1322, ET23, sat, NZL
N6B4, ET25, sat, NPL
15A5, ET27, sat, PAK
M294, ET28, pol, JOR
M119, ET29, uns, SYR
S33, ET30, sat, USA
82
102F51, ET31, sat, USA
74B15, ET34, sat, PAK
Genotype4
M56, ET2, rot, SYR
Genotype5
M275, ET5, rig, JOR
Genotype6
M95, ET10, rot, SYR
M289, ET16, tru, JOR
CC2003, ET24, sat, AUS
M248, ET26, pol, JOR
Genotype7
M270, ET6, tru, JOR
Genotype8
N6B1, ET9, fal, NPL
56A14, ET12, sat, PAK
15B4, ET17, sat, PAK
90
128A10, ET18, sat, PAK
Genotype9
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
74B4, ET32, sat, PAK
74B12, ET33, sat, PAK
CC2013, ET35, sat, AUS
— 1 change

Fig.2G cbbR

Fig 2H: ML tree (left) and one most parsimonious tree (right) for each of gene *exoF3*

(Based on sequence of 793 bp and bootstrap values higher than 50 are listed)

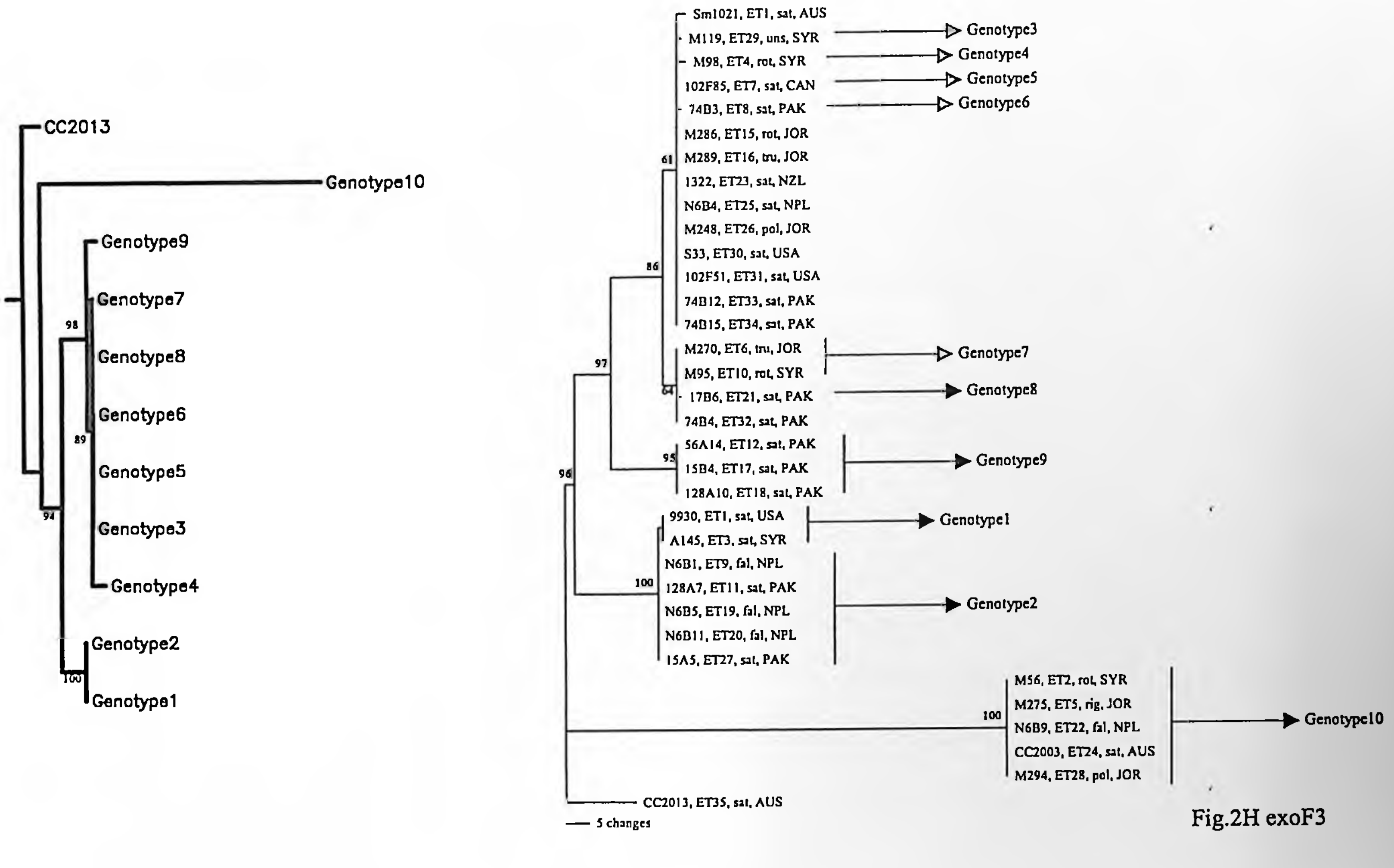

CC2013
Genotype10
Genotype9
Genotype7
98
Genotype8
Genotype6
89
Genotype5
94
Genotype3
Genotype4
Genotype2
100
Genotype1
Sm1021, ET1, sat, AUS
M119, ET29, uns, SYR
M98, ET4, rot, SYR
102F85, ET7, sat, CAN
74B3, ET8, sat, PAK
M286, ET15, rot, JOR
61
M289, ET16, tru, JOR
1322, ET23, sat, NZL
N6B4, ET25, sat, NPL
M248, ET26, pol, JOR
S33, ET30, sat, USA
86
102F51, ET31, sat, USA
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
M270, ET6, tru, JOR
97
M95, ET10, rot, SYR
64
17B6, ET21, sat, PAK
74B4, ET32, sat, PAK
56A14, ET12, sat, PAK
95
15B4, ET17, sat, PAK
96
128A10, ET18, sat, PAK
9930, ET1, sat, USA
A145, ET3, sat, SYR
N6B1, ET9, fal, NPL
100
128A7, ET11, sat, PAK
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
15A5, ET27, sat, PAK
M56, ET2, rot, SYR
M275, ET5, rig, JOR
100
N6B9, ET22, fal, NPL
CC2003, ET24, sat, AUS
M294, ET28, pol, JOR
CC2013, ET35, sat, AUS
5 changes
Genotype3
Genotype4
Genotype5
Genotype6
Genotype7
Genotype8
Genotype9
Genotype1
Genotype2
Genotype10

Fig.2H exoF3

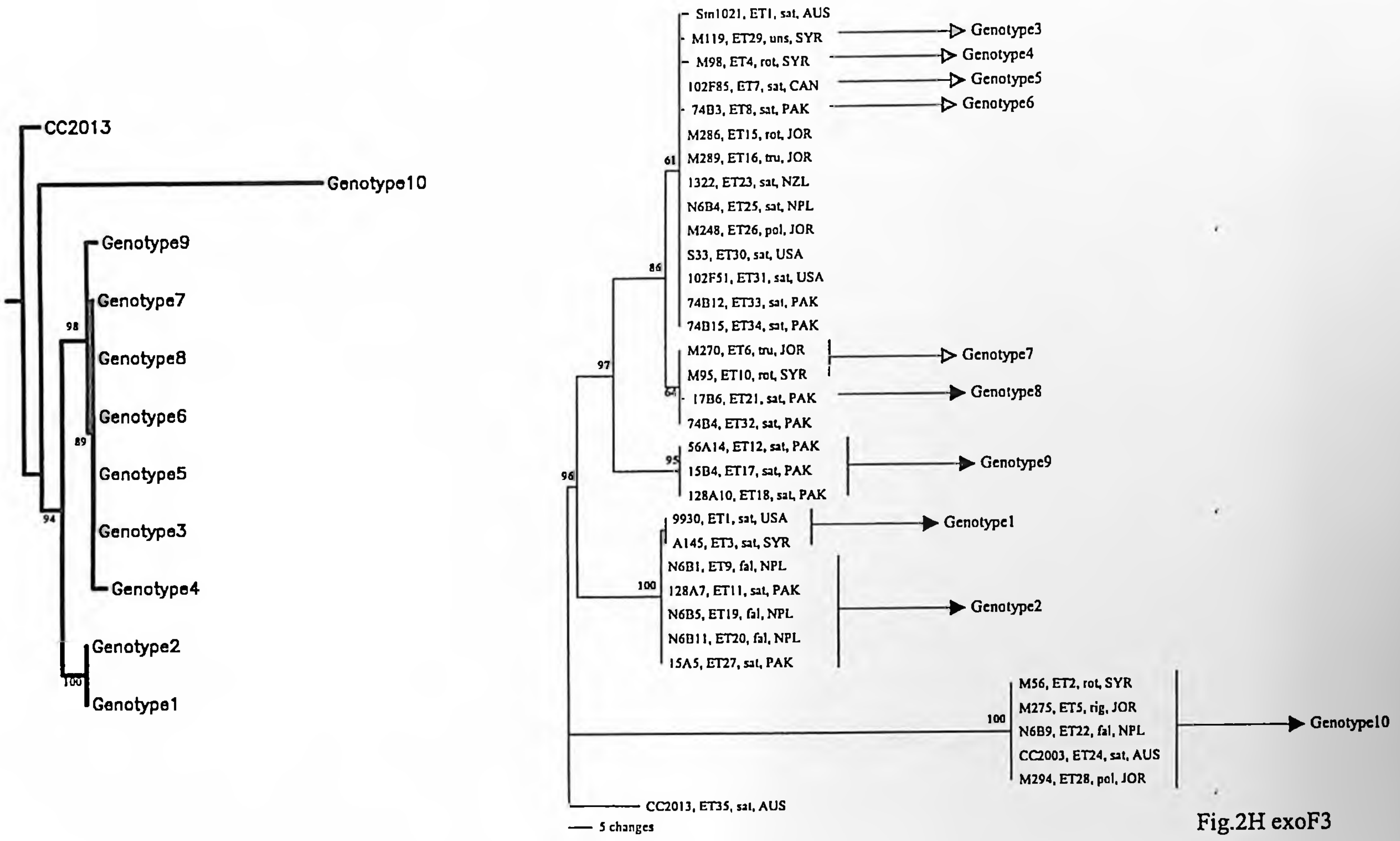

CC2013
Genotype10
Genotype9
Genotype7
98
Genotype8
Genotype6
89
Genotype5
94
Genotype3
Genotype4
Genotype2
100
Genotype1
Sm1021, ET1, sat, AUS
M119, ET29, uns, SYR
M98, ET4, rot, SYR
102F85, ET7, sat, CAN
74B3, ET8, sat, PAK
M286, ET15, rot, JOR
61
M289, ET16, tru, JOR
1322, ET23, sat, NZL
N6B4, ET25, sat, NPL
M248, ET26, pol, JOR
S33, ET30, sat, USA
86
102F51, ET31, sat, USA
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
M270, ET6, tru, JOR
97
M95, ET10, rot, SYR
64
17B6, ET21, sat, PAK
74B4, ET32, sat, PAK
56A14, ET12, sat, PAK
95
15B4, ET17, sat, PAK
96
128A10, ET18, sat, PAK
9930, ET1, sat, USA
A145, ET3, sat, SYR
N6B1, ET9, fal, NPL
100
128A7, ET11, sat, PAK
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
15A5, ET27, sat, PAK
M56, ET2, rot, SYR
M275, ET5, rig, JOR
100
N6B9, ET22, fal, NPL
CC2003, ET24, sat, AUS
M294, ET28, pol, JOR
CC2013, ET35, sat, AUS
5 changes
Genotype3
Genotype4
Genotype5
Genotype6
Genotype7
Genotype8
Genotype9
Genotype1
Genotype2
Genotype10

Fig.2H exoF3

Fig 2I: ML tree (left) and one most parsimonious tree (right) for each of gene *minC*

(Based on sequence of 574 bp and bootstrap values higher than 50 are listed)

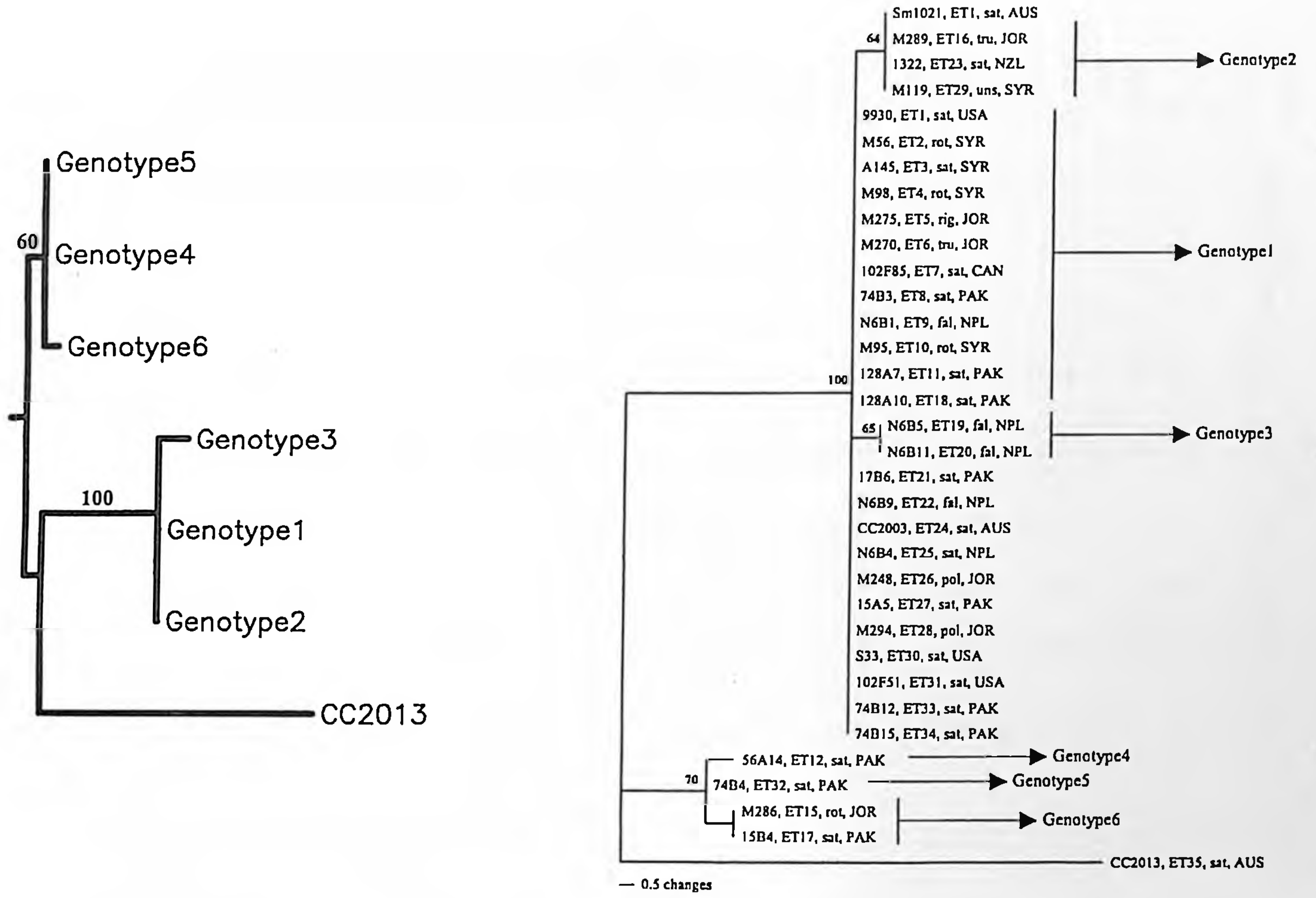

Genotype5
60
Genotype4
Genotype6
Genotype3
100
Genotype1
Genotype2
CC2013
Sm1021, ET1, sat, AUS
64
M289, ET16, tru, JOR
1322, ET23, sat, NZL
M119, ET29, uns, SYR
Genotype2
9930, ET1, sat, USA
M56, ET2, rot, SYR
A145, ET3, sat, SYR
M98, ET4, rot, SYR
M275, ET5, rig, JOR
M270, ET6, tru, JOR
102F85, ET7, sat, CAN
74B3, ET8, sat, PAK
N6B1, ET9, fal, NPL
M95, ET10, rot, SYR
128A7, ET11, sat, PAK
128A10, ET18, sat, PAK
Genotype1
100
65
N6B5, ET19, fal, NPL
N6B11, ET20, fal, NPL
Genotype3
17B6, ET21, sat, PAK
N6B9, ET22, fal, NPL
CC2003, ET24, sat, AUS
N6B4, ET25, sat, NPL
M248, ET26, pol, JOR
15A5, ET27, sat, PAK
M294, ET28, pol, JOR
S33, ET30, sat, USA
102F51, ET31, sat, USA
74B12, ET33, sat, PAK
74B15, ET34, sat, PAK
56A14, ET12, sat, PAK
Genotype4
70
74B4, ET32, sat, PAK
Genotype5
M286, ET15, rot, JOR
15B4, ET17, sat, PAK
Genotype6
CC2013, ET35, sat, AUS
0.5 changes

Fig.2I minC

Table 3. Haplotypes within each of the nine genes in the 32 strains analyzed here

Strain	Genotype									Genotype
	oxyR	*aqpz1*	*mdh*	*cbbR*	*exoF3*	*minC*	*nifH*	*fdhE*	*sma1440*	Overall
9930	1	1	1	1	1	1	1	1	1	1
M56	1	3	1	5	10	1	2	11	9	2
A145	1	1	1	1	1	1	5	14	2	3
M98	1	1	1	2	4	1	2	11	9	4
M275	1	1	1	6	10	1	3	2	13	5
M270	2	1	1	8	7	1	6	20	3	6
102F85	8	5	1	1	5	1	1	15	5	7
74B3	2	1	2	1	6	1	2	1	10	8
N6B1	3	1	1	9	2	1	2	3	6	9
M95	1	1	1	5	7	1	2	4	11	10
128A7	4	1	1	1	2	1	2	1	16	11
56A14	7	1	4	9	9	4	1	14	16	12
M286	1	1	1	1	5	6	2	5	7	13
M289	1	1	1	5	5	2	2	5	11	14
15B4	7	1	4	9	9	6	1	16	17	15
128A10	7	1	4	9	9	1	1	17	16	16
N6B5	1	1	1	9	2	3	1	6	14	17
N6B11	1	1	1	9	2	3	1	6	15	18
17B6	1	1	1	3	8	1	1	6	7	19
N6B9	1	1	1	1	10	1	2	7	7	20
1322	5	2	1	1	5	2	7	12	8	21
CC2003	1	2	1	5	10	1	9	8	4	22
N6B4	1	1	3	1	5	1	4	18	12	23
M248	1	1	1	7	5	1	10	10	7	24
15A5	1	1	1	1	2	1	1	6	7	25
M294	1	1	1	1	10	1	2	4	4	26
M119	5	2	1	1	3	2	8	13	8	27
S33	6	4	6	1	5	1	2	9	7	28
102F51	6	4	5	1	5	1	2	4	7	29
74B4	9	1	7	9	7	5	1	16	16	30
74B12	9	1	7	9	5	1	1	19	16	31
74B15	9	1	7	4	5	1	1	17	16	32

Table 4. Mean pairwise nucleotide difference per site among strains for each of the nine genes. The means and standard deviations were calculated from all 561 pairwise comparisons of the 32 strains.

Replicon	Gene	Mean Pairwise Distance	SD
Chromosome	*oxyR*	0.004307771	0.00675
Chromosome	*aqpz1*	0.003807787	0.01057
Chromosome	*mdh*	0.003427944	0.005302
pSymA	*fdhE*	0.025156082	0.019574
pSymA	*nifH*	0.022909196	0.037753
pSymA	*sma1440*	0.009397311	0.010837
pSymB	*cbbR*	0.009955083	0.015867
pSymB	*exoF3*	0.052175896	0.064024
pSymB	*minC*	0.007014616	0.011694

Gene Genealogy Analysis

The maximum parsimony method implemented in PAUP and maximum likelihood method implemented in PHYLIP program were used to infer the relationships among strains for each of the nine genes. For maximum likelihood analysis, only one represent of each genotype was used and for the convenience of comparison, the corresponding genotypes were showed in the maximum parsimony trees. These trees are shown in Fig 2.

For maximum parsimony analysis, I also included sequences from the model laboratory strain Rm1021. These sequences were retrieved from the GenBank. In eight of the nine genealogies, strain Rm1021 was found in the main cluster. However, the alleles in Rm1021 were in the most frequent category in only three of the nine genes. The three genes were *mdh*, *oxyR* and *nifH*. The most abundant alleles in these three genes were present in 22, 17 and 12 strains respectively (out of 32).

The genealogies of the three Chromosomal genes showed highly similar, though not identical, relationships among the strains. However, none of the three genes showed a clustering of strains based on either geographic origin or host species (Figure 2A-C).

For the three genes on pSymA, their phylogenies differed from each other (Figure 2D-F). Both *fdhE* and *nifH* genealogies revealed two divergent clades supported by strong bootstrap support. Similar to genealogies inferred from the three Chromosomal genes, phylogenies inferred from those on pSymA showed no pattern based on geographic origin or host species. Interestingly, phylogenies inferred here were very different from that inferred from MLEE data.

Table 5. T-test of pairwise strain divergence between genes. A positive t value indicates that the gene in the top row has a lower among-strain divergence than the gene in the furthest left column.

	aqpz1	*mdh*	*oxyR*	*cbbR*	*exoF3*	*minC*	*sma1440*	*fdhE*	*nifH*
aqpz1	-								
mdh	-0.76	-							
oxyR	0.94	2.43*	-						
cbbR	7.64**[a]	9.24**	7.76**	-					
exoF3	17.65**	17.97**	17.61**	15.16**	-				
minC	4.82**	6.62**	4.75**	-3.53**	-16.44**	-			
sma1440	8.75**	11.72**	9.44**	-0.69	-15.60**	3.54**	-		
fdhE	22.73**	25.38**	23.85**	14.29**	-9.56**	18.85**	16.68**	-	
nifH	11.54**	12.10**	11.49**	7.49**	-9.33**	9.53**	8.15**	-1.25	-

([a], P>0.05, *P<0.05, **P<0.001)

The three genes on pSymB also showed divergent phylogenetic patterns. Each of the three phylogenies revealed two clades supported by strong bootstrap values (87% to 100%). Again, in all three phylogenies, I found no evidence of clustering based on geographic origin or host species. Similar to the genes on pSymA, all three genealogies were very different from that inferred from MLEE data.

Overall, maximum likelihood method gave similar topology for each of the nine genes (Fig 2). Maximum likelihood trees were further analyzed using SHTest in testing the congruence and incongruence between topologies (see below).

Comparisons between the MultiLocus and LIAN Softwares

Both the LIAN software and the Multi-Locus software were used to infer the population structures based on the nucleotides within each single gene.

For MultiLocus software, two kinds of data, phylogenetically informative sites and polymorphic sites were used to infer the population structures for each gene. For LIAN software, in addition to the two kinds of data used in MultiLocus analysis, all nucleotides of each gene were also used to infer the population structure. The results are summarized in Table 6.

Comparing results from different datasets using the LIAN software, although standardized IAs differed slightly, I found all three datasets gave the same Pmc value for the same gene (except aqpz1, Table 6). A similar but not identical pattern was observed when the datasets were analyzed using the MultiLocus program. Overall, MultiLocus analysis generated P values slightly lower than those from LIAN. The only exception is aqpz1 (Table 6).

Overall, these two softwares gave consistent results that the population structures are clonal based on nucleotides within each gene (except aqpz1, however, see below). Although the kind of sites used for the analysis had influenced the values of rds and IAs, they did not influence the overall P values.

Based on these results, I proceed to use only MultiLocus software and polymorphic sites in the following analysis between different genes. This is because the MultiLocus software allows you to define linkage groups, a feature not available in the LIAN software. In addition, as seen above, the population structures inferred from all polymorphic sites were consistent with the results obtained using all nucleotide sites. One practical problem of using all nucleotide sites is that multilocus has a limit on the number of loci (2000) that can be analyzed.

Table 6. Results of Multilocus and LIAN analyses of single genes

Gene tested	Multilocus-- Informative Sites		Multilocus-- All Polymorphic Sites		LIAN-- Informative Sites		LIAN-- All Polymorphic Sites		LIAN-- All Sites	
	rd	P	rd	P	IA	Pmc[a]	IA	Pmc	IA	Pmc
oxyR	0.1572	< 0.002	0.1443	< 0.005	0.153	0.01	0.1391	0.01	N/A	
aqpz1	-0.0736	1	0.0251	0.17	-0.0708	1	0.0244	0.17	N/A	
mdh	0.3555	< 0.002	0.2215	< 0.005	0.3526	0.01	0.2049	0.01	N/A	
fdhE	0.386	< 0.01	0.2172	< 0.005	0.3786	0.01	0.1971	0.01	0.02	0.01
nifH	0.6065	< 0.01	0.4468	< 0.005	0.5991	0.01	0.4225	0.01	0.0342	0.01
sma1440	0.1016	< 0.002	0.0651	< 0.005	0.1015	0.01	0.0601	0.01	0.0021	0.01
cbbR	0.7889	< 0.002	0.4043	< 0.005	0.7907	0.01	0.3663	0.01	0.0081	0.01
exoF3	0.6418	< 0.01	0.5385	< 0.005	0.6405	0.01	0.5214	0.01	0.0967	0.01
minC	0.6602	< 0.01	0.4352	< 0.005	0.65	0.01	0.4131	0.01	0.0122	0.01

[a]: P value based on Monte Carlo simulation.

Multilocus and Incompatibility Ratio Test

Using both the Multilocus program and the Incompatibility Ratio test, I calculated the standardized overall index of association (r_d) among alleles at different sites using polymorphic nucleotide sites and obtained the percentage of pair-wise sites that are phylogenetically incompatible by using phylogenetic informative sites. Below I summarize the results of the analyses conducted at three different levels: (i) among polymorphic nucleotide sites within each of the nine genes, (ii) among genes within each replicon, and (iii) among genes between replicons.

Among nucleotide sites within a gene

Within eight of the nine genes, alleles among polymorphic nucleotide sites showed overall significant linkage disequilibrium (Table 7). The only gene showed linkage equilibrium was *aqpz*1 on the chromosome. This result is similar to what I obtained from the incompatibility ratio test (Table 7).

I examined further the nucleotide polymorphisms within this gene. Among the 32 strains, only two polymorphic nucleotide sites were found and each of these two sites had two alleles (site one, position 242, with alleles A and G; site two, position 383, with alleles A and G). Among the four possible genotypes (AA, AG, GA, and GG), only three were found: 3 individuals with the AG genotype combination, 2 with the GA genotype, 27 with the GG genotype, and none for the AA genotype. Therefore, there was no convincing evidence for phylogenetic incompatibility and random recombination within the *aqpz1* gene. The false linkage equilibrium result given by the two analyses was due to a type II error likely resulted from the small number of polymorphic sites (five polymorphic nucleotide sites in which two are phylogenetic informative) and the highly skewed allele frequencies (one allele accounts for more than 84% of all), as discussed in the Materials and Methods section.

In summary, both analyses indicated very little, if any, evidence of recombination within individual genes.

Among genes within a replicon

In this analysis, I structured all polymorphic nucleotide sites within each gene as one haplotype (i.e. one linkage group in the Multilocus program option). This structuring thus eliminated the effect of within-gene linkage disequilibrium on that between genes. The among gene linkage equilibrium analysis was then conducted and the results are summarized in Table 7. My analyses indicated that the three genes on the Chromosome were in significant linkage disequilibrium while those on either pSymA or pSymB were in linkage equilibrium.

Incompatibility ratio test also showed that the rate of recombination is very low within the Chromosome but high within each of the two replicons (Table 7). These results are consistent with those from linkage disequilibrium analysis.

Among genes from different replicons

My third level analysis examined the associations of alleles among genes located on different replicons. Two different analyses were performed. In the first, I defined each replicon as one linkage group and analyzed all three replicons together. The results showed that the three replicons were all in linkage equilibrium (Table 7).

In the second analysis, I examined further the relationships between individual pairs of genes located on different replicons. To perform this analysis, I again defined each gene as one linkage group and analyze their individual pair-wise relationships. The results of this analysis are summarized in Table 7. Of the nine pairwise comparisons between genes located on the Chromosome and pSymA, seven were in linkage equilibrium and two in linkage disequilibrium. Of the nine pairwise tests between genes located on the chromosome and pSymB, seven were in linkage equilibrium and two in linkage disequilibrium. Among the nine pairwise tests between genes located on the megaplasmids pSymA and pSymB, six were in linkage equilibrium and three in linkage disequilibrium.

Incompatibility ratio analyses were also performed between individual pairs of genes located on different replicons and the results are summarized in Table 7. Of the nine pairwise comparisons between genes located on the chromosome and pSymA, two showed incompatibility ratios not significantly different from randomized data sets. Of the nine pairwise comparisons between genes located on the Chromosome and pSymB, seven showed incompatibility ratios not significantly different from randomized data set. Of the nine pairwise comparisons between genes located on the pSymA and pSymB, all showed incompatibility ratios not significantly different from randomized data set (Table 7).

Shimodaira and Hasegawa Test (SHTest) and Partition Homogeneity Test (PHT)

SHTest tests whether two topologies are significantly different from each other in a given dataset. One limitation of this method is that it requires the topology must be bifurcating. For example, if X and Y (different ETs) have different sequence in gene A but have same sequence in gene B, then they can not be used in SHTest, just because in gene B they are not bifurcating but staying together at the end of the same branch. After checking all possible gene pairs, only two of them are good for the SHTest and they are the nifH-fdhE pair and the fdhE-sma1440 pair. All of these three genes are located on pSymA. Results of SHTest are summarized in Table 8. If we use 90% (*P*=0.1) as our confidence interval (which is suggested for SHTest), both of the comparisons showed that the topologies of the two genes involved are not congruent.

PHTests were also performed using the same two gene pairs. Results showed that for both gene pairs the tree lengths in the randomized dataset are significantly longer than the shortest possible trees, indicating these two gene pairs were incongruent (Table 8). These results are consistent with that from SHTest.

To further confirm the results, I used the MultiLocus software to infer the population structure by calculating rd and incompatibility ratio, using all polymorphic nucleotide sites from these two gene pairs. Results of both these two analyses were consistent with that from SHTest and PHTest, i.e. the pairs of loci were in linkage equilibrium and consistent with recombination.

Table 7. Results of linkage disequilibrium analysis and incompatibility ratio test

Gene(s) tested	Linkage Disequilibrium Analysis[a]				Incompatibility Ratio Test			
	No. of loci	r_d	*P*	LE or LD[b]	No. of loci	IR[c]	*P*	Significant recombination
Single Gene								
oxyR	14	0.1443	< 0.005	LD	13	0.34	< 0.002	No
aqpz1	5	0.0251	0.17	LE	2	0	0.802	Yes[g]
mdh	19	0.2215	< 0.005	LD	6	0.15	< 0.002	No
fdhE	87	0.2172	< 0.005	LD	50	0.14	< 0.01	No
nifH	65	0.4468	< 0.005	LD	47	0.18	< 0.01	No
sma1440	27	0.0651	< 0.005	LD	17	0.39	< 0.002	No
cbbR	17	0.4043	< 0.005	LD	10	0.09	< 0.002	No
exoF3	147	0.5385	< 0.005	LD	126	0.05	< 0.01	No
minC	18	0.4352	< 0.005	LD	13	0	< 0.01	No
Single Replicon								
Chromosome	38	0.1104	0.025	LD	21	0.52	0.04	No
pSymA	179	0.1286	0.08	LE	114	0.96	0.52	Yes
pSymB	182	0.3576	0.985	LE	149	0.95	0.42	Yes
Chromosome-pSymA-pSymB [e]	284	18.6235	0.19	LE				
Chromosome-pSymA								
oxyR-fdhE	101	0.1807	0.005	LD	63	0.53	0.005	No
oxyR-nifH	79	0.3026	0.75	LE	60	0.58	0.04	No[d]
oxyR-sma1440	41	0.054	0.105	LE	30	0.66	0.025	No[d]
aqpz1-fdhE	92	0.1981	0.53	LE	52	0.80	0.075	Yes
aqpz1-nifH	70	0.4064	0.07	LE	49	0.93	0.45	Yes
aqpz1-sma1440	32	0.0411	0.975	LE	19	0.76	0.03	No[d]
mdh-fdhE	106	0.1778	0.01	LD	56	0.77	0.03	No
mdh-nifH	84	0.2801	0.92	LE	53	0.47	0.02	No[d]
mdh-sma1440	46	0.0643	0.16	LE	23	0.65	0.01	No[d]
Chromosome-pSymB								
oxyR-cbbR	31	0.1811	0.1	LE	23	1.15	0.66	Yes

oxyR-exoF3	161	0.4547	1	LE	139	0.32	< 0.01	No[d]
oxyR-minC	32	0.1781	0.26	LE	26	1.35	0.72	Yes
aqpz1-cbbR	22	0.2493	0.995	LE	12	0.34	0.115	Yes
aqpz1-exoF3	152	0.5151	0.295	LE	128	1.01	0.38	Yes
aqpz1-minC	23	0.2692	0.875	LE	15	0.16	0.48	Yes
mdh-cbbR	36	0.2112	0.01	LD	16	1.13	0.62	Yes[d]
mdh-exoF3	166	0.4447	0.88	LE	132	0.4	0.01	No[d]
mdh-minC	37	0.2479	0.01	LD	19	2.09	0.94	Yes[d]
pSymA-pSymB								
fdhE-cbbR	104	0.1797	< 0.005	LD	60	0.97	0.18	Yes[d]
fdhE-exoF3	234			LE	176	0.54	0.07	Yes
fdhE-minC	105	0.179	0.035	LD	63	1.01	0.25	Yes[d]
nifH-cbbR	82	0.2805	0.76	LE	57	0.67	0.11	Yes
nifH-exoF3	212	0.318	0.39	LE	173	1.16	0.46	Yes
nifH-minC	83	0.2785	0.635	LE	60	0.51	0.28	Yes
sma1440-cbbR	44	0.1242	0.005	LD	27	0.9	0.17	Yes[d]
sma1440-exoF3	174	0.4048	0.4	LE	143	1.03	0.51	Yes
sma1440-minC	45	0.0972	0.27	LE	30	0.84	0.2	Yes
MLEE data [f]	9	0.0144	0.77	LE	9	0.67	<0.01	No[d]

[a]: Results from MultiLocus program, using all polymorphic sites;

[b]: LE-Linkage Equilibrium; LD-Linkage Disequilibrium;

[c]: IR, Incompatibility Ratio;

[d]: Results from Linkage Disequilibrium analysis and Incompatibility Ratio test are incongruent;

[c]: Only phylogenetic informative sites were used;

[f]: The loci locating on the chromosome were used.

[g]: No evidence of recombination was found in the data by eye inspection and the false recombination result from the program was due to type II error because of the small sample size.

Table 8. Results of SHTest, PHTest, MultiLocus analysis, and IR analysis on gene pairs of nifH-fdhE and fdhE-sma1440

Gene Paris	SHTest		PHTest			Multilocus --All Polymorphic Sties		Incompatibility Ratio Test --All Polymorphic Sites	
	No. of ETs	*P*	Shortest	Randomized	*P*	rd	*P*	I.R.	*P*
nifH-fdhE	7	0/0	278	301-311	0.001	0.1713	0.18	0.098	0.155
fdhE-sma1440	7	0/0.073	171	187-196	0.001	0.1443	0.105	0.072	0.15

Isolation by Distance and Isolation by Host Plant

The Mantel Test was used to infer whether there is genetic isolation by geographic distance or isolation by host plant species in this collection of strains. The results of these analyses are summarized in Table 9.

Overall, when pair-wise genetic distance of *S.meliloti* strains was compared with the corresponding geographic distance, no obvious correlation was found. The only one exception is aqpz1, which showed a positive relationship between genetic and geographic distance with a *P* value of 0.021 (Table 9).

Similarly, when pair-wise genetic distance of *S. meliloti* strains was compared with the corresponding phylogenetic distance of the host plants, no obvious correlation was found (Table 9).

Table 9. Results of the Analyses for Isolation by Distance and Isolation by Host Plants

Gene	Isolation by Distance		Isolation by Host Plant	
	r	P	r	P
oxyR	0.0593	0.371	-0.4417	0.984
aqpz1	0.31	0.021	-0.1788	0.589
mdh	-0.0314	0.5160	-0.3415	0.886
fdhE	-0.151	0.811	-0.3075	0.746
nifH	0.3806	0.045	-0.2514	0.632
sma1440	-0.1763	0.869	-0.4238	0.833
cbbR	-0.05	0.549	-0.3709	0.837
exoF3	0.0562	0.378	-0.1474	0.634
minC	-0.1828	0.847	0.3233	0.305

CHAPTER 4. DISCUSSION

In this study, I sequenced portions of nine genes from each of 33 strains of *S. meliloti*. My sequence comparisons revealed a diverse pattern of variation among strains and significantly different rates of nucleotide substitutions among the nine sequenced genes. My results suggest divergent and dynamic patterns of molecular evolution among the three replicons within the *S. meliloti* genomes. Below I discuss the potential mechanisms for the observed patterns of gene sequence variation and their relevance to the evolution of *S. meliloti*.

Divergent Rate of Molecular Evolution among Genes

Overall, I found significant difference among the nine sequenced genes in the mean divergence among strains, from a low of 0.34% for the *mdh* gene to a high of 5.22% for the *exo*F3 gene, a difference of over 15 folds. This difference is larger than those in most species examined so far using MLST. For example, the sequenced loci in the human pathogen *Neisserria meningitides* showed a 6-fold difference in the mean sequence divergence among strains (Cooper *et al.* 2004). The difference in the magnitude of among-gene variation in bacteria was likely influenced by the examined genes. In most MLST studies, house-keeping, conserved genes were used to ensure that sequences from all strains can be obtained for analysis. As a result, there may be greater functional and nucleotide sequence constraints on those genes.

In this study, my criterion for genes to be sequenced was primarily based on their genomic positions, with little consideration of function. In fact, the idea of house-keeping essential genes in *S. meliloti* may not be applicable to most genes on pSymA and pSymB. Laboratory studies have shown that most DNA on the pSymA and pSymB megaplasmids can be deleted with little or no fitness consequences. Interestingly, the gene with the highest mean divergence *exoF3* is located on pSymB and it encodes a putative outer membrane protein. Its high rate of divergence might have significant functional implications, e.g. in host recognition and niche specialization.

Compared to genes on the two megaplasmids, genes on the Chromosome showed a lower level of divergence among strains. The mean pairwise nucleotide dissimilarity for each replicon was 0.38%, 1.92% and 2.3% for the Chromosome, pSymA and pSymB, respectively. The examined genes on pSymA and pSymB showed an average of 5 and 6-fold higher rates of substitution than those on the Chromosome. While the exact mechanism(s) for these differences are unknown at present, there are two possibilities. The first is that genes on pSymA and pSymB are under significantly less functional constraints than those on the Chromosome. Therefore, genes on pSymA and pSymB might be more prone to mutation accumulation. The second hypothesis is that genes on pSymA and pSymB might be under positive selection. Signatures of positive selection have been detected for genes involved in sexual reproduction in several animals and the rapid diversification of some of these genes has been proposed as a key factor driving niche specialization and speciation. It is possible that many genes on pSymA and pSymB are highly niche-specific and their divergence might be associated with niche specialization or switching. Additional sequences from closely related species such as *S. medicae* and *S. fredii* could help distinguishing these two hypotheses.

Relationship between Genotype and Geographic Origin or Host Species

In this study, my analyses revealed little geographic or host species - based patterns of molecular variation. Instead, the results suggested significant gene flow between geographic regions and host species. Migration between different geographic areas could have been brought about by human activities such as the widespread cultivation of the host plant alfalfa in many parts of the world.

Extensive gene flow between geographic populations has been found in other *Rhizobium* species (e.g. Oyaizu *et al.* 1993; Moreiar *et al.* 1998). For example, strains of *Rhizobium etli*, another common nitrogen fixating species that can form symbiotic relationship with legumes, have been found capable of dispersal along with the seeds of its host plant, *Phaseolus vulgaris* (Perez-Ramirez *et al.* 1998). The lack of strictly host-specific clades in *S. meliloti* is also consistent with its life style in nature. *S. meliloti* is not an obligate symbiont but exists mostly as a free-living bacterium in the soil. As a result, each strain/genotype might have been exposed to many different host species and an obligate host-specialization might be detrimental to the strain's long-term survival in nature. At present, the host range for each of the 33 strains analyzed here are unknown. It would be interesting to determine if host range itself are phylogenetically constrained.

I would like to point out that the lack of a global geographic or host pattern of molecular variation in this set of strains does not imply that host or geography plays no role in the distributions of molecular variation in *S. meliloti*. Indeed, experimental evidence has shown that isolates of *S. meliloti* trapped by certain host plant species may show similar genotype profiles (Jebara *et al.* 2001). In addition, several studies have shown that chemical and physical differences, such as pH and clay and organic matter contents, in the soil can influence the genetic diversity of *Rhizobium* populations (Harrison *et al.* 1989; Richaume *et al.* 1989; Strain *et al.* 1994). The 32 strains analyzed here were selected based on their MLEE types and therefore, they may not represent any natural population from a specific geographic area.

Comparisons between MLEE and MLST Data

While the analysis of the nine chromosomal MLEE loci using the same set of strains suggested significant linkage equilibrium (Eardly et al. 1990; Table 7), my analysis of the nucleotide polymorphisms among the three genes on the chromosome found significant linkage disequilibrium (Table 7). In addition, the IR analysis of the same 9 chromosomal MLEE loci showed that the population is in linkage disequilibrium. This result clearly demonstrated that the selections of markers and analytical methods could have a significant influence on the inference of population structure.

Three tests were used here to infer clonality and recombination: the index of association (linkage equilibrium), incompatibility ratio test, and the SHTest. In this study, while the three methods showed overall consistent results, minor inconsistencies were also found. Several factors could have contributed to these inconsistencies.

First, the three methods have different null models. The tests for index of association and incompatibility ratio used the null models of random mating while the SHtest used strict clonality as the null model. Second and perhaps more importantly, all three analytical methods are highly sensitive to sample size, including both the number of strains as well as the number of polymorphic nucleotide sites and the frequencies of individual alleles at these sites. The number of strains used in this study was relatively small and many polymorphic sites had highly skewed allele frequencies, thus potentially contributing to the observed inconsistencies.

Divergent Patterns of Evolution among Genes on Different Replicons

My results identified that the three chromosomal genes were in linkage disequilibrium while those on pSymA and pSymB were in linkage equilibrium. In addition, genes on different replicons showed widespread linkage equilibrium. While these results suggested that genetic exchange was common in natural populations of *S. meliloti*, the differences between the Chromosome and the two megaplasmids required additional explanations. The following four, non-mutually exclusive hypotheses could have contributed to these dynamic patterns.

First, the differences found here between the Chromosome and the two megaplasmids could be the artifact of the specific genes analyzed and may not reflect the general patterns of the three replicons. For example, the nine genes analyzed here were selected to represent the wide genomic regions based on data from the model laboratory strain Rm1021. However, the locations of these genes might be different in other natural strains. The second hypothesis invokes mating and fusion of natural strains followed by homologous recombination. In this hypothesis, the rate of recombination among chromosomal genes was lower than those on pSymA and pSymB. The third hypothesis is that the loss and gain of megaplasmids might be common for wild strains in natural populations and that linkage equilibria observed here for genes on megaplasmids were the results of recombination among homologous megaplasmids in an unknown reservoir. The fourth hypothesis is that genes located on megaplasmids could be easily lost and gained while those on the chromosome were more stable. Therefore, the recombining population structure observed here for the two megaplasmids could be the result of many horizontal gene transfer events.

Indeed, horizontal gene transfer has been found in many natural bacterial populations, often involve genetic materials from divergent origins (Nelson *et al.* 1999). In addition, horizontal transfers of large segments of DNA have been demonstrated in both pathogenic as well as symbiotic bacteria. For example, Sullivan *et al.* (2002) found that a 502-kb symbiosis island in *Mesorhizobium loti* strain R7A was transferred to a non-symbiotic *Mesorhizobium* strain in the soil and converting the recipient cell to a symbiont. Although no evidence for direct genetic exchange between marked strains of *S. meliloti* has been found in nature, a cluster of genes encoding a type IV pilus was found on pSymA of Rm1021 and in *Rhizobium* sp. NGR234 (Streit *et al.* 2004). Type IV pili are unique structures on the bacterial surface that are found in many gram-negative bacteria. They play important roles in adhesions to host cells, in infections by bacteriophages, and in conjugative DNA transfers (Door *et al.* 1998; Ashelford *et al.* 2003).

REFERENCE

Agapow, P. M., and A. Burt. 2001. Indices of multilocus linkage disequilibrium. Molecular Ecology Notes, 1:101-102.

Ashelford, K. E., M. J. Day and J. C. Fry. 2003. Elevated abundance of bacteriophage infecting bacteria in soil. Appli. Environ. Microbiol. 69: 285-289.

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. 98:9883-9888.

Biondi, E. G., E. Pilli, E. Giuntini, M. L. Roumiantseva, E. E. Andronov, O. P. Onichtchouk, O. N. Kurchak, B. V. Simarov, N. I. Dzyubenko, A. Mengoni, and M. Bazzicalupo. 2003. Genetic relationship of *Sinorhizobium meliloti* and *Sinorhizobium medicae* strains isolated from Caucasian region. FEMS Microbiology Letters 220:207-213.

Brown, A. H. D., M. W. Feldman, and E. Nevo. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics 96:523-536.

Capela, D., F. Barloy-Hubler, J. Gouzy, G. Bothe, F. Ampe, J. Batut, P. Boistard, A. Becker, M. Boutry, E. Cadieu, S. Dreano, S. Gloux, T. Godrie, A. Goffeau, D.

Kahn, E. Kiss, V. Lelaure, D. Masuy, T. Pohl, D. Portetelle, A. Puhler, B. Purnelle, U. Ramsperger, C. Renard, P. Thebault, 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.

Carelli, M., S. Gnocchi, S. Fancelli, A. Mengoni, D. Paffetti, C. Scotti, and M. Bazzicalupo. 2000. Genetic diversity and dynamics of *Sinorhizobium meliloti* populations nodulating different alfalfa cultivars in Italian soils. Appli. Environ. Microbiol. 66: 4785-4789.

Cooper, J. E., and E. J. Feil. 2004. Multilocus sequence typing – what is resolved? Trends in Microbiology 12:373-377.

Colles, F. M., K. Jones, R. M. Harding and M. C. Maiden. 2003. Genetic diversity of *Campylobacter jejuni* isolates from farm animals and the farm environment. Appli. Environ. Microbiol. 69: 7409-7413.

Door, J., T. Hurek, and B. Reinhold-Hurek. 1998. Type IV pili are involved in plant-microbe and fungus-microbe interactions. Mol. Microbiol. 30:7-17.

Eardly, B. D., L. A. Materon, N. H. Smith, D. A. Johnson, M. D. Rumbaugh, and Selander R. K. 1990. Genetic structure of natural populations of the nitrogen-fixing bacterium *Rhizobium meliloti*. Appl. Environ. Microbiol. 56:187-194.

Felsenstein, J. 2005. PHYLIP, Version 3.63. Seattle: University of Washington.

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

2001. The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont *Sinorhizobium meliloti*. Proc. Natl. Acad. Sci. USA 98:9889-9894.

Galibert, F., T. M. Finan, S. R. Long, A. Puhler, 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.

Harrison, S. P., D. G. Jones, and J. P. Young. 1989. *Rhizobium* population genetics: genetic variation within and between populations from diverse locations, J. Gen. Microbiol. 135:1061-1069.

Haubold, B., and R. R. Hudson. 2000. LIAN 3.0: detecting linkage disequilibrium in lultilocus data. Bioinformatics Applications Note. 16:847-848.

Jebara, M., R. Mhamdi, M. E. Aouani, R. Ghrir, and M. Mars. 2001. Genetic diversity of *Sinorhizobium* populations recovered from different *Medicago* varieties cultivated in Tunisian soils. Can. J. Microbiol. 47:139-147.

Jensen, J. L., A. J. Bohonak, and S. T. Kelley. 2005. Isolation by distance, web service. BMC Genetics 6: 13. http://phage.sdsu.edu/~jensen/

Maiden, M.C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. 95:3140-3145.

Maynard Smith, J., N. H. Smith, C. G. Dowson and B. G. Spratt. 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. U. S. A. 90:4384-4388.

Maynard Smith, J. 1999. The detection and measurement of recombination from sequence data. Genetics 153:1021-1027.

Moreiar, F. M. S., K. Haukka, and J. P. Young. 1998. Biodiversity of rhizobia isolated from a wide range of forest legumes in Brazil. Molecular Ecology 7:889-895.

Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen and C. M. Fraser. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323-329.

Oyaizu, H., S. Matsumoto, K. Minamisawa, and T. Gamou. 1993. Distribution of rhizobia in leguminous plants surveyed by phylogenetic identification. Journal of General and Applied Microbiology 39:339-354.

Paffetti, D, C. Scotti, S. Gnocchi, S. Fancelli, and M. Bazzicalupo. 1996. Genetic diversity of an Italian *Rhizobium meliloti* population from different *Medicago sativa* varieties. Appl. Environ. Microbiol. 62 (7): 2279—2285.

Perez-Ramirez, N. O., M. A. Rogel, E. Wang, J. Z. Castellanos, and E. Martinez-Romero. 1998. Seeds of *Phaseolus vulgaris* bean carry *Rhizobium etli.* FEMS Microbiology Ecology 26:289-296.

Posada, D. and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14(9): 817-818.

Richaume, A., J. S. Angle, and M. J. Sadowsky. 1989. Influence of soil variables on in situ plasmid transfer from *Escherichia coli* to *Rhizobium fredii.* Appl. Environ. Microbiol. 55:1730-1734.

Rome, S., M.P. Fernandez, B. Brunel, P. Normand, and J.C. Cleyet-Marel. 1996. Sinorhizobium medicae sp. Nov., isolated from annual Medicago spp. Int. J. Syst. Bacteriol. 46:972-980.

Roumiantseva, M. L., E. E. Andronov, L. A. Sharypova, T. Dammann-Kalinowski, M. Keller, and J. P. Young. 2002. Diversity of *Sinorhizobium meliloti* from the central Asian alfalfa gene center. Appl. Environ. Microbiol. 68:4694-4697.

Sawada, H., L.D. Kuykendall and J.M. Young. 2003. Changing concepts in the systematics of bacterial nitrogen-fixing legume symbionts. *J. Gen. Appl. Microbiol.* 49:155-79.

Shimodaira, H., and M. hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16:**1114-1116.

Strain, S. R., K. Leung, T. S. Whittam, F. J. De Brujin, and P. J. Bottomley. 1994. Genetic structure of *Rhizobium leguminosarum* biovar *trifolii* and *viciae* populations found in two Oregon soils under different plant communities. Appl. Environ. Microbiol. **60:**2772-2778.

Streit, W. R., R. A. Schmitz, X. Perret, C. Staehelin, W. J. Deakin, C. Raasch, H. Liesegang, and W. J. Broughton. 2004. An evolutionary hot spot: the pNGR234b replicon of *Rhizobium* sp. strain NGR234. Journal of Bacteriology **186:**535-542.

Sullivan, J. T., J. R. Trzebiatowski, R. W. Cruickshank, J. Gouzy, S. D. Brown, R. M. Elliot, D. J. Fleetwood, N. G. McCallum, U. Rossbach, G. S. Stuart, J. E. Weaver, R. J. Webby, F. J. de Brujin, and C. W. Ronson. 2002. Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. J. Bacteriol. **184:**3086-3095.

Swofford, D. L. 2004. PAUP*: Phylogenetic Analysis Using Parsimony (and other methods). Sinaur Associates, MA, USA.

Vinuesa, P., C. Silva, D. Werner, and E. Martinez-Romero. 2005. Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in Bradyrhizobium species cohesion and delineation.

Molecular Phylogenetics and Evolution **34**:29-54.

Xu, J., R. Vilgalys, and T. G. Mitchell. 2000. Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus Cryptococcus neoformans. Molecular Ecology. 9:1471-1482.