UNIVERSITY OF CALIFORNIA
RIVERSIDE

Drivers of Genotypic Abundance and Spatial Spread in Wild *Bradyrhizobium*

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Amanda C. Hollowell

March 2015

Dissertation Committee:
Dr. Joel L. Sachs, Chairperson
Dr. Cheryl Y. Hayashi
Dr. Mark S. Springer

The Dissertation of Amanda C. Hollowell is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the superb mentorship of Professor Joel Sachs. He has taught me to strive for excellence in all things from the construction of our presentations to the architecture of this very dissertation. Moreover, he has fundamentally altered the way I approach scientific questions by compelling me to construct frameworks with which to relate my work to the larger scientific community.

I also must thank my dissertation committee members: Professors Cheryl Hayashi and Mark Springer, for their advice and guidance. They have been unfailingly optimistic even when delivering constructive criticism.

Lastly, I thank the following current and former Sachs Laboratory members for their contributions to this research: Dr. John Regus; PhD candidate Kelsey Gano; Research Assistants Roxanne Bantay, Nathaniel Gleason, and Yifan Lii; Undergraduate Researchers Andrew Bernardo, Donna Centeno, Gabriel Lopez, Jonathan Lyu, Devora Moore, Victor Pahua, Seema Patel, Akshay Patil, Jonathan Pham, and Kareem Shahin.

The following sources of funding supported this research: NSF grants DEB – 0816663 and Career grant DEB – 1150278, UCR Graduate Research Mentorship fellowships (2012, 2014), a UCR Genetics, Genomics and Bioinformatics graduate division fellowship (2009-2010), a UCR Newell travel award (2011), the UCR College of Natural and Agricultural Sciences James and Margaret Lesley Annual Prize (2013), and UCR College of Natural and Agricultural Sciences Herbert Kraft scholarships (2010, 2012).

DEDICATION

Dedicated to my parents, without whose love and support this dissertation would not have been possible.

ABSTRACT OF THE DISSERTATION

Drivers Of Genotypic Abundance And Spatial Spread In Wild *Bradyrhizobium*

by

Amanda C. Hollowell

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, March 2015
Dr. Joel L. Sachs, Chairperson

Understanding patterns of genotypic abundance and spatial spread is a fundamental objective in studying critically important bacterial strains. Many clinically and agriculturally relevant strains are host-associated, forming either pathogenic or mutualistic symbioses with crops, livestock and humans. Many of these relationships are facilitated by the acquisition of horizontally transferred genomic islands that encompass genes necessary (or beneficial) for association with a host. Genomic islands have been implicated in the clinical epidemic spread of pathogenic strains. However, the impact of genomic islands on natural populations of bacteria has not been well studied.

The legume-rhizobia mutualism is a particularly well studied symbiosis mediated by the acquisition of plasmids or genomic islands and this dissertation research focused on *Bradyrhizobium*, the most cosmopolitan rhizobial lineage. Like all rhizobia, *Bradyrhizobium*, exhibit a bipartite lifestyle and the genome reflects this with upregulation in the genomic (symbiosis) island when within a host and upregulation in the rest of the genome (chromosome) when free-living in the soil. Thus, the evolutionary

drivers of each lifestyle can be studied through the analysis of their respective genome region.

To examine the effect of symbiosis island acquisition on *Bradyrhizobium* abundance and spatial spread, the distributions of symbiotic and non-symbiotic *Bradyrhizobium* were compared. No support was found for an evolutionary association between symbiosis island gain and greater abundance or spatial spread. The frequencies of particular symbiosis island and chromosome genotypes were analyzed in order to examine the role of the host plant versus the soil in structuring *Bradyrhizobium* populations. Chromosome genotypes exhibited high abundance and spatial spread while symbiosis island genotypes did not. These results taken together suggest that competition within the soil, as opposed to selection by the host plant, is the major driver of population structure.

Antibiotic resistance is a pervasive problem in host-associated pathogens. However, these traits are ancient in environmental populations, where antibiotics are thought to be relevant to intermicrobial communication and conflict. In order to test the degree to which natural populations select for antibiotic resistance, *Bradyrhizobium* were tested for antibiotic resistance to 17 antibiotics. Resistance traits to all but one antibiotic were recovered and multidrug resistance was ubiquitous.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GENERAL INTRODUCTION

A fundamental objective in the study of bacteria is to uncover and understand genotypic patterns of abundance and spatial spread. Among host-associated bacteria, the relationship between pathogens and symbionts with their host is often facilitated by the acquisition of genomic islands. These islands contain accessory genes that can enhance bacterial pathogenicity or symbiosis and can have dramatic effects on bacterial fitness (Hacker and Carniel, 2001; Dobrindt et al., 2004). For instance, the acquisition of genomic islands in bacteria has been associated with clinical epidemics (Bach, 2000; Diep et al., 2006; Qiu et al., 2006). But other than clinical settings, there is little understanding of the role of genomic island acquisition in structuring bacterial populations.

Rhizobia are bacteria characterized by their capability to nodulate legume host plants and fix atmospheric nitrogen in exchange for plant derived photosynthates (Sprent, 2001). The legume-rhizobia symbiosis is a particularly well-studied mutualism mediated by the acquisition of genomic islands or plasmids. Through their association with legumes, symbiotic rhizobia fix ~150 teragrams of nitrogen annually (Sugawara et al., 2013). Because legumes constitute ~25% of global crop production (Ferguson et al., 2010), humans have attempted to maximize crop yield through the inoculation of desired rhizobial strains. However, inoculation programs have yielded only modest results due to a competition problem in which inoculation strains are purged from soil populations by indigenous rhizobia (Triplett and Sadowsky, 1992; Vlassak et al., 1996; Hungria et al.,

2003). Thus, understanding the evolutionary drivers of rhizobial abundance and spatial spread in natural populations is critical for the improvement of agriculture.

This dissertation examines drivers of abundance and spatial spread in the most cosmopolitan rhizobial lineage, *Bradyrhizobium,* which is found associated with plants and animals, including humans (Chaintreuil et al., 2000; Silva et al., 2005, 2005; Vinuesa et al., 2008; Sachs et al., 2009; Hunt et al., 2011; Bhatt et al., 2013; Costello et al., 2013; Parker, 2014). In *Bradyrhizobium*, symbiotic capacity upon legumes can be predicted by the presence or absence of a symbiosis island (Kaneko et al., 2002, 2011; Sachs et al., 2010, 2011). The bipartite *Bradyrhizobium* genome mirrors the bipartite lifestyle of rhizobia in which portions of time are spent both within a host plant and free-living in the soil (Pessi et al., 2007). Thus the evolutionary drivers of each lifestyle can be investigated through analysis of their respective genome region.

In the first chapter of my dissertation, I investigated the different lifestyles of *Bradyrhizobium* by genotyping bacteria isolated from either root nodules or root surfaces. Isolates were assigned the symbiotic or non-symbiotic lifestyle based on their collection location (for nodule isolates) or a combination of greenhouse inoculations and PCR amplification assays of symbiosis island loci (for root surface isolates). I tested the paradigm that acquisition of a genomic island confers a fitness benefit, via increased abundance and spread in soil populations.

In the second chapter I examined the drivers of *Bradyrhizobium* epidemics. I genotyped symbiotic *Bradyrhizobium* isolates across the genome using loci from both the symbiosis island and the chromosome. The bipartite lifestyle and genome of

Bradyrhizobium suggest that selection on *Bradyrhizobium* that occurs when free-living will affect chromosomal loci, while selection that occurs while in symbiosis will alter the symbiosis island. Host plants have been shown to act as a driver of the prevalence of particular symbiosis island genotypes across different legume species (Parker, 2012), but little work has examined symbionts across an individual host species. I tested the relative roles of the plant versus the soil in structuring *Bradyrhizobium* populations by examining population genetic structure across the genome.

In the final chapter I assessed the role of antibiotic resistance traits in driving abundance and spread in wild strains of *Bradyrhizobium.* Among host-associated bacteria, especially ones that affect the well being of humans, our livestock and our crops, one of the most important traits is antibiotic resistance. Resistance in bacteria includes many traits that are ancient and predate the usage of clinical antibiotics (D'Costa et al., 2011) and it has been suggested that antibiotic resistance traits in bacteria might be relevant in intermicrobial communication or conflict (Martinez, 2008). Recent work has uncovered environmental resistomes (D'Costa, 2006; Nesme et al., 2014), but there is little understanding of the benefit that antibiotic resistance confers outside of clinical environments. I assessed *Bradyrhizobium* isolates for antibiotic resistance traits to 17 antibiotics and tested the degree to which resistance is favored by natural selection in natural settings.

References

Bach, S. (2000) The *Yersinia* high-pathogenicity island is present in different members of the family Enterobacteriaceae. *FEMS Microbiol. Lett.* **183**: 289–294.

Bhatt, A.S., Freeman, S.S., Herrera, A.F., Pedamallu, C.S., Gevers, D., Duke, F., et al. (2013) Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N. Engl. J. Med.* **369**: 517–528.

Chaintreuil, C., Giraud, E., Prin, Y., Lorquin, J., Ba, A., Gillis, M., et al. (2000) Photosynthetic bradyrhizobia are natural endophytes of the African wild rice *Oryza breviligulata*. *Appl. Environ. Microbiol.* **66**: 5437–5447.

Costello, E.K., Carlisle, E.M., Bik, E.M., Morowitz, M.J., and Relman, D.A. (2013) Microbiome assembly across multiple body sites in low-birthweight infants. *mBio* **4**: e00782–13–e00782–13.

D'Costa, V.M. (2006) Sampling the antibiotic resistome. *Science* **311**: 374–377.

D'Costa, V.M., King, C.E., Kalan, L., Morar, M., Sung, W.W.L., Schwarz, C., et al. (2011) Antibiotic resistance is ancient. *Nature* **477**: 457–461.

Diep, B.A., Gill, S.R., Chang, R.F., Phan, T.H., Chen, J.H., Davidson, M.G., et al. (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet* **367**: 731–739.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* **2**: 414–424.

Ferguson, B.J., Indrasumunar, A., Hayashi, S., Lin, M.-H., Lin, Y.-H., Reid, D.E., and Gresshoff, P.M. (2010) Molecular analysis of legume nodule development and autoregulation. *J. Integr. Plant Biol.* **52**: 61–76.

Hacker, J. and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO Rep.* **2**: 376–381.

Hungria, M., Campo, R.J., and Mendes, I.C. (2003) Benefits of inoculation of the common bean (*Phaseolus vulgaris*) crop with efficient and competitive *Rhizobium tropici* strains. *Biol. Fertil. Soils* **39**: 88–93.

Hunt, K.M., Foster, J.A., Forney, L.J., Schütte, U.M.E., Beck, D.L., Abdo, Z., et al. (2011) Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS ONE* **6**: e21313.

Kaneko, T., Maita, H., Hirakawa, H., Uchiike, N., Minamisawa, K., Watanabe, A., and Sato, S. (2011) Complete genome sequence of the soybean symbiont *Bradyrhizobium japonicum* strain USDA6T. *Genes* **2**: 763–787.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., et al. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* **9**: 189 –197.

Martinez, J.L. (2008) Antibiotics and antibiotic resistance genes in natural environments. *Science* **321**: 365–367.

Nesme, J., Cécillon, S., Delmont, T.O., Monier, J.-M., Vogel, T.M., and Simonet, P. (2014) Large-scale metagenomic-based study of antibiotic resistance in the environment. *Curr. Biol.* **24**: 1096–1100.

Parker, M.A. (2012) Legumes select symbiosis island sequence variants in *Bradyrhizobium*. *Mol. Ecol.* **21**: 1769–1778.

Parker, M.A. (2014) The spread of *Bradyrhizobium* lineages across host legume clades: from *Abarema* to *Zygia*. *Microb. Ecol.*

Pessi, G., Ahrens, C.H., Rehrauer, H., Lindemann, A., Hauser, F., Fischer, H.-M., and Hennecke, H. (2007) Genome-wide transcript analysis of *Bradyrhizobium japonicum* bacteroids in soybean root nodules. *Mol. Plant. Microbe Interact.* **20**: 1353–1363.

Qiu, X., Gurkar, A.U., and Lory, S. (2006) Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci.* **103**: 19830–19835.

Sachs, J.L., Ehinger, M.O., and Simms, E.L. (2010) Origins of cheating and loss of symbiosis in wild *Bradyrhizobium*. *J. Evol. Biol.* **23**: 1075–1089.

Sachs, J.L., Kembel, S.W., Lau, A.H., and Simms, E.L. (2009) In Situ phylogenetic structure and diversity of wild *Bradyrhizobium* communities. *Appl. Environ. Microbiol.* **75**: 4727–4735.

Sachs, J.L., Russell, J.E., and Hollowell, A.C. (2011) Evolutionary instability of symbiotic function in *Bradyrhizobium japonicum*. *PLoS ONE* **6**: e26370.

Silva, C., Vinuesa, P., Eguiarte, L.E., Souza, V., and Martínez-Romero, E. (2005) Evolutionary genetics and biogeographic structure of *Rhizobium gallicum sensu lato*, a widely distributed bacterial symbiont of diverse legumes. *Mol. Ecol.* **14**: 4033–4050.

Sprent, J.I. (2001) Nodulation in legumes Royal Botanic Gardens, Kew.

Sugawara, M., Epstein, B., Badgley, B.D., Unno, T., Xu, L., Reese, J., et al. (2013) Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol.* **14**: R17.

Triplett, E.W. and Sadowsky, M.J. (1992) Genetics of competition for nodulation of legumes. *Annu. Rev. Microbiol.* **46**: 399–422.

Vinuesa, P., Rojas-Jimenez, K., Contreras-Moreira, B., Mahna, S.K., Prasad, B.N., Moe, H., et al. (2008) Multilocus sequence analysis for assessment of the biogeography and evolutionary genetics of four *Bradyrhizobium* species that nodulate soybeans on the Asiatic continent. *Appl. Environ. Microbiol.* **74**: 6987–6996.

Vlassak, K., Vanderleyden, J., and Franco, A. (1996) Competition and persistence of *Rhizobium tropici* and *Rhizobium etli* in tropical soil during successive bean (*Phaseolus vulgaris* L.) cultures. *Biol. Fertil. Soils* **21**: 61–68.

CHAPTER 1

Epidemic spread of symbiotic and non-symbiotic *Bradyrhizobium* across California

Abstract

Understanding how bacterial strains increase in abundance and spread spatially is a critical problem in epidemiology. Host-associated bacteria, including pathogens and symbionts, are of particular agronomic and clinical importance, and many of these relationships are mediated by horizontal acquisition of genomic islands. The acquisition of genomic islands has been implicated in many clinical epidemics. However, less is know about the role of genomic islands in structuring natural populations of bacteria including rhizobia, where 'symbiosis' islands and plasmids confer symbiotic capacity. Here we studied the most cosmopolitan rhizobial genus, *Bradyrhizobium*, in order to study drivers of abundance and spread in natural populations. We cultured 1,292 *Bradyrhizobium* from both root nodules and root surfaces across a >840 km transect of California. Isolates were genotyped and we inferred the presence or absence of the symbiosis island through a combination of greenhouse nodulation assays and PCR. Our goals were to investigate *Bradyrhizobium* strain dominance and spread, examine the role of symbiosis island acquisition in driving strain dominance, and test for community structure of rhizobial isolates due to other factors. We found that *Bradyrhizobium* populations were extremely diverse, but dominated by few haplotypes with a single epidemic haplotype constituting nearly 30% of collected isolates and spreading nearly

statewide. We found that *Bradyrhizobium* populations are structured by sampling locale, portion of legume inhabited, symbiotic capacity, and collection year. Symbiotic strains were more abundant, and more likely to be dominant, but we did not find evidence that the symbiosis island is an evolutionary driver of strain dominance.

Introduction

A critical goal in bacteriology is to understand patterns of genotypic abundance and epidemic spread. Of particular interest are host-associated bacteria, including pathogens and symbionts. These diverse bacterial lineages colonize host surfaces, can inhabit specific tissues or cells and can often persist free in soils and or aquatic habitats (Bright and Bulgheresi, 2010; Sachs et al., 2013). The capacity of bacteria to thrive in host tissues is often modulated by the presence of plasmids and genomic islands, cassettes of loci specific to host association that can get transmitted among genomes. Horizontal transfer of these accessory loci often engenders bacteria with suites of fitness-enhancing traits including host infection capacity, multidrug resistance, pathogenicity, and metabolic flexibility (Groisman and Ochman, 1996; Jain et al., 2003; Gal-Mor and Finlay, 2006). The acquisition of plasmids and genomic islands has been implicated for epidemics in human clinical infections, including *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Yersinia* (Bach, 2000; Diep et al., 2006; Qiu et al., 2006). But in natural settings, we understand little about how bacterial strains vary in their capacity to dominate local sites and host populations or to spread among sites across ecological barriers. In particular, almost nothing is known about patterns of dominance and epidemic spread in symbiotic bacteria, which are important for human health, the success of leguminous crops and other ecosystem services.

Rhizobia are proteobacteria characterized by their capacity to infect leguminous plants and fix atmospheric nitrogen for their hosts (Sprent, 2001). Globally, rhizobia are responsible for the fixation of ~150 teragrams of nitrogen per year (Sugawara et al.,

2013) and their symbiosis with legumes represents the largest input source of nitrogen into terrestrial ecosystems (Cleveland et al., 1999). In agriculture, legumes account for ~27% of global crop production (Graham and Vance, 2003), and are valued for their capacity to grow in nitrogen depauperate soils. Similar to bacterial pathogens, rhizobia can acquire accessory DNA that confers the capability to colonize and infect hosts. Rhizobial genomes are thus subdivided into portions specific for their life stages, with chromosomal loci expressed during free-living phases in the soil, and symbiosis loci expressed inside of host cells (Uchiumi et al., 2004; Pessi et al., 2007). Genes required for host nodulation and nitrogen fixation are clustered onto large plasmids or genomic islands (Kaneko, 2000; Galibert, 2001; Kaneko et al., 2002; Young et al., 2006; Lee et al., 2008), and these 'symbiosis loci' can be transferred among lineages, presumably via conjugation (Young, 1996; Perret et al., 2000; Moulin et al., 2004). Non-nodulating rhizobia are also common (Segovia et al., 1991; Sachs et al., 2010), and these strains often lack some or all of the characterized symbiosis loci (Segovia et al., 1991; Sullivan et al., 1996; Saito et al., 1998; Pongsilp et al., 2002; Sachs et al., 2010; Okubo et al., 2012).

Bradyrhizobium is a cosmopolitan rhizobial lineage found free-living in soils and in aquatic environments, as well as in symbiotic association with plant and animal hosts, including humans (Chaintreuil et al., 2000; Vinuesa et al., 2005, 2008; Sachs et al., 2009; Hunt et al., 2011; Bhatt et al., 2013; Costello et al., 2013). *Bradyrhizobium* nodulates diverse wild legumes as well as essential crops such as soybeans (*Glycine*), peanuts (*Arachis*), and cowpea (*Vigna*) (Rivas et al., 2009; Parker, 2014). In the model genome,

Bradyrhizobium diazoefficiens strain USDA110 (previously *B. japonicum*), symbiosis-specific genes are clustered within a 410-kb region in which the G + C content differs from the rest of the genome (Göttfert et al., 2001). However some symbiotic *Bradyrhizobium sp.* BTAi1 and ORS278 lack a symbiosis island, and use a different mechanisms to nodulate hosts (Giraud et al., 2007). Previous studies have revealed that *Bradyrhizobium* populations can exhibit epidemic distributions, in which a few genotypes exist at high frequency at a single site (Sachs et al., 2009) and or spread among multiple locations (Vinuesa et al., 2005, 2008). But such epidemics remain poorly understood and it is unclear what role the symbiosis island might play as a driver of increased abundance and epidemic spread.

Here, we investigated the population genetic structure of *Bradyrhizobium spp.* cultured from *Lotus strigosus,* a native annual legume common across the Pacific Southwest of the United States. We cultured 850 *L. strigosus* nodules from 14 natural sites across California encompassing 185 plants collected over a >840 km transect. In parallel we isolated 442 root surface *Bradyrhizobium* from three focal host populations within this range, which includes strains that lack symbiosis islands and cannot infect *Lotus* hosts. All 1,292 isolates were sequenced at two chromosomal loci and we used a combination of PCR and inoculation assays to test for presence of the symbiosis island in all the root surface isolates. We assigned haplotypes and symbiotic capacity information to all isolates, and examined the frequency and spatial spread of epidemic rhizobial genotypes within and among host populations. Our goals were to i) investigate strain dominance and epidemic spread of *Bradyrhizobium* genotypes in native *L. strigosus* hosts,

ii) infer the presence or absence of the symbiosis island in *Bradyrhizobium*, iii) test for the role of symbiosis-island presence as a driver of *Bradyrhizobium* strain dominance and iv) test for community structure of rhizobial isolates due to other abiotic or biotic factors.

Materials and Methods

Collection of Bradyrhizobium isolates: Bradyrhizobium was isolated from the nodules and root surfaces of *L. strigosus,* and clonal cultures were grown and archived for genotyping following published protocols (Sachs et al., 2009). Briefly, whole plants were transported in sealed plastic bags to the laboratory where they were washed with tap water and sterilized tools were used to remove root nodules. Nodules were surface sterilized with bleach and rinsed with water before being crushed with glass rods and the contents plated on a modified arabinose glucanate medium (MAG; Sachs et al. 2009). For root surface isolates, the roots were dissected into sections ~1cm in length before being vortexed in a sterile solution of 0.01% Tween 20 (Fisher Scientific Fair Lawn, NJ). The wash solution was then serially diluted and plated on glucose-based rhizobium defined medium (GRDM) with cyclohexamide as an antifungal and bromothymol blue as a pH indicator (Sachs et al., 2009). Among the resultant colonies, we selected for *Bradyrhizobium* based on growth rate, color, and ability to grow on MAG and GRDM, but not on Luria-Bertani medium (LB; (Sachs et al., 2009)). Plant hosts for culturing were collected from 14 sites across California covering a ~840 km transect. Collection sites included University of California Natural Reserves (Bodega Marine Reserve, Burns Piñon Ridge Reserve, and Motte Rimrock Reserve), an undeveloped site in the hills above University of California - Riverside, a biological field station in Claremont, CA (Robert J. Bernard Biological Field Station), natural preserves (Madrona Marsh Preserve, Pismo Dunes Natural Preserve, and Whitewater Preserve), a wildlife refuge (Guadalupe-Nipomo Dunes National Wildlife Refuge), two separate sites within a large state park

(Anza Borrego Desert State Park), a municipal park (Griffith Park), a site adjacent to the San Dimas Reservoir, and an undeveloped site adjacent to human development (San Dimas Canyon) (Supplementary Table S1.1). Nodule isolates were collected from plants at all sites, but root surface isolates were only collected from plants at the Bodega Marine Reserve, Motte Rimrock Reserve, and the undeveloped site in the hills above the University of California – Riverside.

Sequencing and haplotype analysis:

Partial sequences from two chromosomal loci: *glnII* and *recA* (totaling 974 bp) were PCR amplified using published protocols, and sequenced at the Institute for Integrative Genome Biology of UC Riverside (Vinuesa et al., 2005). Only sequences with unambiguous bases were utilized leading to a total of 1,292 sequenced isolates. Sequences from each locus were aligned separately using Clustal Omega, and isolates with identical sequences for each locus were determined using the "find redundant" command within the MacClade program (Maddison and Maddison, 2005). Each unique sequence, or haplotype, was defined for each locus separately and for the concatenated dataset. Abundance was calculated for the concatenated dataset as the number of times each haplotype was isolated.

Symbiotic capacity assessment

Isolates from nodules were automatically classified as symbiotic (i.e., presence of a functional symbiosis island; Sachs et al., 2010). We conducted a combination of assays

on root surface isolates to test for symbiotic capacity. A subset of isolates (75) were previously assessed using greenhouse nodulation assays on *L. strigosus*, which has already been shown to be a permissive host on diverse *Bradyrhizobium* lineages (Sachs et al., 2010, 2011; Ehinger et al., 2014). Here, we conducted greenhouse nodulation experiments on an additional 55 isolates, using identical procedures. Briefly, at least 5 hosts per tested *Bradyrhizobium* isolate were grown in sterile conditions and were inoculated clonally with $5x10^8$ cells, and parallel control hosts were inoculated with sterile water. At 8 weeks post inoculation all hosts were unpotted, roots and shoots were weighed, and roots were checked for nodules. In all cases controls lacked nodules. Hosts given the same inoculated strains either all became nodulated, or were all lacking nodules. The remaining 342 root surface isolates were classified as symbiotic or non-symbiotic based on success or failure of PCR amplification of at least one symbiosis island locus (*nifD*, *nodD-A*, *nodZ*, and *nolL*) (Parker, 2000; Moulin et al., 2004; Sachs et al., 2010). Earlier analyses showed that successful amplification of these loci, giving a band of the correct size, is a reliable indicated of presence of the symbiosis island (Sachs et al., 2010). Many isolates were tested at two or more loci (160/342).

Phylogenetic tree reconstruction and species designation:

A phylogenetic tree was reconstructed using the concatenated *glnII* and *recA* sequences as well as homologous sequences from the following reference strains: *Bradyrhizobium arachidis* (CCBAU33067), *B. betae* (PL7HG1), *B. canariense* (SEMIA928), *B. cytisi* (LGM25866), *B. diazoefficiens* (SEMIA5080), *B. elkanii* (USDA46), *B. iriomotense*

(EK05), *B. japonicum* (USDA110), *B. lablabi* (CCBAU61434), *B. liaoningense* (SEMIA5025), *B. retamae* (Ro19), and *B. yuanmingense* (R2m). Reference strains were chosen to utilize all known species of *Bradyrhizobium* that aligned fully with our sequenced *glnII* and *recA* regions (NCBI as of 11-18-2014) and *Mesorhizobium loti* (MAFF303099) was used as an outgroup. The GTR model of evolution was selected from the Akaike information criterion in jModelTest2 (Darriba et al., 2012) and the phylogenetic tree was reconstructed in PhyML 3.0 (Guindon et al., 2010) utilizing a BioNJ starting tree and subtree pruning and regrafting (SPR). Branch support was estimated with the fast approximate likelihood ratio test (aLRT) with the Shimodaira-Hasegawa-like (SH-like) procedure (Anisimova and Gascuel, 2006). *Bradyrhizobium* species were defined as the monophyletic clades including no more than one reference species with branch support ≥ 0.90 (Shimodaira and Hasegawa, 1999) and attempting to adhere to past species demarcations that utilized some of the same loci (Vinuesa et al., 2008). We analyzed inter-species variation using the ratio of fixed to shared polymorphisms using DNASP (Librado and Rozas, 2009).

Sequence Statistics:

Using the concatenated dataset we calculated strain richness (number of unique haplotypes/number of isolates) and strain dominance (abundance of each haplotype/number of isolates) analogues of species richness and evenness (Mcinnes, 2004). For each host population sampling site, haplotypes were defined as dominant if they were collected at least five times and represented at least 10% of the total isolates at

that site. Spatial spread was defined as the maximum distance between any individual collections sites with the same haplotype. GPS coordinates for distances used the midpoint of each collection site, because distances within sites were small compared to between site differences. We also calculated Hd (haplotype diversity – probability that two haplotypes drawn uniformly at random from the population are not the same), π (nucleotide diversity – average number of nucleotide differences per site between two sequences), k (average # nucleotide differences), linkage disequilibrium (average absolute D'), recombination (R), and the minimum number of recombination events (Lewontin, 1964; Tajima, 1983; Hudson and Kaplan, 1985; Hudson, 1987; Nei, 1987) using DnaSP (Librado and Rozas, 2009).

Trait analysis

We tested for phylogenetic signal, or the faithful transmission of traits from one generation to the next, on the traits of symbiotic capacity, abundance, and spatial spread. Significant phylogenetic signal is a prerequisite for testing the evolution of traits using phylogenetic data. We used Pagel's lambda, estimated with the "fit discrete" function in the "Geiger" package (Harmon et al., 2007), and for symbiotic capacity we also used Fritz Purvis' D, which was estimated using the "phylo.d" function in the "Caper" package (Orme, 2012). The Mk1 model of maximum likelihood as well as parsimony were used for ancestral state reconstruction of symbiotic capacity with a modified phylogenetic tree in Mesquite (Maddison, W. P. and Maddison, D. R., 2011). Because multiple states are not possible for a single taxon (i.e., haplotype), duplicate taxa were

added to the phylogenetic tree whenever a single haplotype encompassed both symbiotic and non-symbiotic isolates. We tested for correlated evolution between symbiotic capacity and haplotype abundance with the phy.anova command in the Geiger package in R (Harmon et al., 2007) utilizing the subset of isolates collected from plants where both root surface and nodule collections had been made (to avoid sampling bias). The resultant dataset included 442 root surface isolates and 116 nodule isolates from three field locales. We also used a standard ANOVA in JMP (SAS Institute Inc, 1989) to examine variation between symbiotic and non-symbiotic isolates in terms of abundance and spatial spread. This latter analysis does not take phylogenetic relationships into account and thus assumes that data are independent of strain relatedness.

Community structure

We analyzed isolation by distance with a Mantel test correlating F_{st} and physical distance matrices within PASSaGE (Rosenberg and Anderson, 2011). We used Fast UniFrac (Hamady et al., 2010) to test for significant differentiation among *Bradyrhizobium* communities at different collection sites. The 'Cluster Samples' tool was used to cluster the collection sites based on the phylogenetic lineages they contained, and the 'Jackknife Sample Clusters' tool was used to assess confidence in the collection site clusters. We utilized the 'Sample Distance Matrix' to numerically compare distances between all collection sites. Abundance was incorporated into Fast UniFrac analyses whenever possible. The jackknife analysis was performed with the number of sequences kept equal to the smallest sample size with 100 permutations. We used the "exact test of population

differentiation" in Arlequin (Excoffier and Lischer, 2010) to assess differentiation among collection sites and to investigate other drivers of *Bradyrhizobium* community structure including root isolate type (nodule, root-tip surface, old root surface), symbiotic status (symbiotic, non-symbiotic), and collection year.

Results

Haplotype designation, abundance, and spatial spread:

The 1,292 concatenated *glnII- recA* sequences resulted in 290 haplotypes, most of which were unique (isolated a single time, 184/290; Supplementary Table S1.1). Among the remaining haplotypes, 13 were defined as dominant in at least one site and these 13 haplotypes constituted the majority of collected isolates (706/1,292). We found dominant haplotypes at all but one collection site (Anza Borrego Desert State Park – Palm Canyon), which had the lowest sampling (Table 1.1). Most haplotypes (257/290) were only found at a single collection site; however, among the dominant haplotypes, most were also found to be epidemic (7/13; collected at a site ≥ 10 km away). Spatial spread for epidemic haplotypes varied from ~ 100 – 750 km and we collected epidemic haplotypes at all but the two least sample sites (Anza Borrego Desert State Park – Palm Canyon and Pismo Dunes Natural Preserve) (Table 1.1). One epidemic haplotype (G03_R01) encompassed 27% of all isolates collected (355 isolates) and was found at all but four collection sites (Pismo Dunes Natural Preserve, Griffith Park, and Anza Borrego State Park (Road/Palm Canyon) sites) (Figure 1.1).

Symbiotic capacity assessment:

We inferred 886 isolates to be symbiotic and 406 to be non-symbiotic (Supplementary Table S1.1). Sources of conflicting information occurred in < 6% of isolates, including eighteen nodule isolates that failed to amplify symbiosis island loci, conflicting information between nodulation assays and PCR amplification in 3/100 nodulation assays,

and conflicting results between PCR amplification assays in 9/160 isolates. We found 13 dominant haplotypes at collection sites. Most dominant haplotypes only had symbiotic isolates and a single dominant haplotype (G64_R29) only had non-symbiotic isolates. Of the dominant haplotypes that included both symbiotic and non-symbiotic isolates, mean abundance was higher for symbiotic (36) versus non-symbiotic isolates (15.6), but the difference was not significant (t=1.48, *df*=4, p=0.214) Among the epidemic haplotypes, most encompassed symbiotic and non-symbiotic isolates (5/7), with symbiotic isolates being more frequent on average than non-symbiotic ones, but without a significant difference (90.6 versus 18.4; t=2.05, *df*=4, p=0.110).

Phylogenetic reconstruction and species designation:

We reconstructed relationships among 20 species including six that were previously identified (*B. betae, B. canariense*, *B. cytisi*, *B. liaoningense* and *B. retamae*) and fourteen that are unnamed (Figure 1.2 and Supplementary Figure S1.1). Most (161) between species comparisons uncovered more fixed than shared polymorphisms, two comparisons have the same number of fixed and shared polymorphisms and 27 have more shared than fixed polymorphisms (Supplementary Table S1.2). Almost half of the species (8/20) were only collected at a single site. However, nearly all collection sites (13/14) were inhabited by multiple species of *Bradyrhizobium* (Supplementary Table S1.3). *Bradyrhizobium canariense* was particularly widespread and was collected at 11 sites.

Sequence Statistics:

Population genetic statistics were analyzed within each individual species that was sampled multiple times (Table 1.2). Linkage was high, between all SNPs in the concatenated sequence, for all species (> 0.9). Strain richness, Hd, π, and recombination varied between species. Some of this variation is probably due to the numbers of representative isolates. As the number of isolates increased, strain richness tended to decrease. Suggesting that we had not found all the potential variation in some species. When only species with over 40 isolates were assessed, Hd, π, and recombination were comparable.

Ancestral state reconstruction: Ancestral state reconstructions were similar for parsimony and likelihood models. (Figure 1.2 and Supplementary Figure S1.2). Gains and losses of symbiotic capacity occur across *Bradyrhizobium* lineages with more reconstructed losses (~29 parsimony/23 maximum likelihood) than gains (~21/10) (Supplementary Table S1.4). Estimated numbers of losses and gains were likely lower for maximum likelihood due to the high number of ambiguous states found on the tree (Figure 1.2). In particular, most of the deepest tree nodes are ambiguous in the maximum likelihood model.

Community analyses:

No evidence was found for isolation by distance using the Mantel test (correlation = 0.00190; p-value = 0.99271). This is supported by the Fast UniFrac analyses in which

clustering does not appear to be occurring due to geographical location (Supplementary Figure S1.3 and S1.4) and instead appears to closely follow their species make-up. Collection sites were statistically significantly different from one another ($p < 0.0000$) with the exception of the following two pairs: San Dimas Reservoir and UC Riverside hills ($p = 0.25930 \pm 0.0845$); UC Riverside Hills and Burns Piñon Ridge Reserve sites ($p = 0.75532 \pm 0.0852$) (Supplementary Table S1.5). We also found support for differentiation between root isolate type communities (nodule, root-tip surface, old root surface), symbiotic status (symbiotic, non-symbiotic), and collection year in structuring populations. Populations from nodules, old root, and new root were significantly differentiated ($p < 0.0012$) except for populations from old root as compared to tip of the root at the Bodega Marine Reserve ($p = 0.46524 \pm 0.0221$) (Supplementary Table S1.6). Non-symbiotic populations were significantly differentiated from isolates that were symbiotic within the same site ($p < 0.00000$) at the Bodega Marine Reserve and Motte Rimrock reserve. However, UC Riverside did not have statistically different non-symbiotic and symbiotic populations ($p = 0.09395 \pm 0.0221$) (Supplementary Table S1.7). Two sites exhibited population differentiation from year to year: San Dimas Canyon ($p = 0.004209 \pm 0.0023$) and Burns Piñon Ridge Reserve (0.00442 ± 0.0006). However, Motte Rimrock Reserve was statistically significant if the error range is not factored in (0.04209 ± 0.0083). The last two sites did not significantly differ from the first collection year to the second: Robert J. Bernard Biological Field Station ($p = 0.11420 \pm 0.0053$) and the UC Riverside hills (0.51000 ± 0.0947), suggesting that haplotype composition can change from year to year but varies by local environment (Supplementary Table S1.8).

Phylogenetic Signal and Correlated Evolution:

Symbiotic capacity exhibited significant phylogenetic signal, but abundance and spatial spread did not (Supplementary Table S1.9). When we assessed symbiotic capacity and abundance for correlated evolution using Chi-squared tests in JMP, we found a statistically significant positive correlation, hence that symbiotic clades on average exhibit higher abundance than non-symbiotic clades. When we tested for correlated evolution of these traits we did not find a statistically significant relationship (Supplementary Table S1.10 and S1.11), hence that evolutionary gains of symbiotic capacity are not statistically associated with gains in abundance. We did not find evidence for correlated evolution of symbiotic capacity with spatial spread using either the Chi-squared tests in JMP or the phy.anova in Geiger.

Discussion

Investigations of bacterial epidemics have largely focused on pathogens. Yet researchers have scant understanding of the drivers of strain dominance and spread in ecologically and economically important taxa such as rhizobia. Population genetic analysis from rhizobia suggest that it is common for a small subset of rhizobial strains to dominate populations (Mcinnes, 2004), but most datasets have focused on agricultural sites and do not explore evolutionary drivers. Our study uncovered an epidemic distribution of *Bradyrhizobium* haplotypes across the state of California with a dramatic divide between rare and dominant haplotypes. Although we recovered 290 haplotypes, the majority of isolates were from the 13 dominant haplotypes (707/1292), which supports the pattern that skewed abundance distributions are a common feature of rhizobial lineages (Silva et al., 1999; Mcinnes, 2004; Sachs et al., 2009). We analyzed regional spread and we found six dominant haplotypes that were endemic. However we also recovered a single epidemic haplotype (G03_R01) that was dominant at most sites (10/14), exhibited a spatial spread of 750 km, and constituted nearly 30% of the total isolates (355/1292) (Table 1.1). This suggests that patterns of dominance occur not only on a local scale, but can be regional as well. It is striking that this dominant haplotype was found in sites that vary a great deal in their patterns of rainfall, temperature, plant community, and soil types.

We uncovered a surprisingly broad diversity of *Bradyrhizobium* species nodulating and inhabiting the root surfaces of *L. strigosus*. We recovered 14 novel species (Supplementary Figure S1.1) and these species varied widely in strain richness. Six out of the seven epidemic haplotypes are found within the *B. canariense* and *B. sp.*

novel species, which had the fewest number of haplotypes and the lowest strain richness (Table 1.2). Differences are known to exist between rhizobial lineages including in the diversity of host plant infected, metabolic utilization, and antibiotic resistance (Cole and Elkan, 1979; Dupuy et al., 1994; Geurts, 1996) and these differences may be driving this trend for differences in epidemic distributions between species of *Bradyrhizobium*.

Collection sites varied in species diversity (Supplementary Table S1.3). However, *Lotus strigosus* exhibits a likely contiguous range (www.Calflora.com), so divergence among hosts is unlikely to drive this differentiation. We also failed to find significant isolation by distance using the Mantel test. Thus other local differences between sites are likely driving local selection for particular *Bradyrhizobium* species. We found the greatest diversity at the three sites where we collected the most isolates, which is consistent with species rarefaction curves in which the greater the sample numbers, the more species that are recovered (Heck et al., 1975). However, these three sites were also the only sites where we collected root surface samples, and greater diversity has been shown to exist within root surface isolates as opposed to nodule populations (Sachs et al., 2009). A sampling size effect is unlikely, as the same pattern of more species with greater sampling did not occur in sites with only nodule collections.

The acquisition of genomic islands can facilitate new suites of traits in bacteria and confer fitness benefits in particular environments (Dobrindt et al., 2004). Previous studies have shown that symbiosis island acquisition can occur in *Bradyrhizobium* and *Mesorhizobium* (Sullivan and Ronson, 1998; Parker et al., 2002; Vinuesa et al., 2005). Our dataset shows that acquisition and loss of the symbiosis island occurs more

frequently than has been shown in previous datasets based on incongruence between SI and CHR (Parker et al., 2002; Sachs et al., 2010). Ancestral state reconstruction gains were estimated at ~20/10 (parsimony/likelihood), while the losses were more common at ~29/23 (parsimony/likelihood) consistent with loss-of-symbiosis islands being more common than gain. Many types of mutations can result in the conversion of symbiotic strain to a non-symbiotic. However, only the whole-scale horizontal transmission of the symbiosis island has been associated with gain of nodulation and nitrogen fixation in *Bradyrhizobium*. Given the more frequent loss events, expansion in symbiotic isolates must occur in order to find greater abundance of symbiotic haplotypes (Table 1.1). Although we found symbiotic isolates were more common than non-symbiotic isolates of the same haplotype, we did not find statistical support for the hypothesis that symbiosis island acquisition is an evolutionary driver of strain abundance. Importantly, we uncovered epidemics of a few extremely abundant symbiotic haplotypes within a single symbiotic clade, so the evolutionary test is weak (Supplementary Table S1.11).

We found evidence for community structuring driven by the host plant. We found support for host plants as drivers of differentiation in 2/3 comparisons of symbiotic and non-symbiotic communities (Supplementary Table S1.7) and we also found 8/9 comparisons of rhizobial communities collected from different plant parts exhibited differentiation (Supplementary Table S1.6). Thus the host plant remains a strong evolutionary driver of rhizobial population structure, though the plant-rhizobial association might not be the main driver of epidemic distributions. In addition to the evidence for host plant driving population differentiation, we also found evidence for

local differences in collection sites driving rhizobial population differentiation (Supplementary Table S1.5). Moreover these environments are dynamic as we found evidence for temporal differentiation in 3/5 comparisons (Supplementary Table S1.8). Taken together this is consistent with previous evidence showing plant and soil factors affect population structure in the rhizosphere (Berg and Smalla, 2009).

In conclusion, our analysis found that native *Bradyrhizobium* populations across California are dominated by a handful of haplotypes with a single epidemic haplotype constituting nearly 30% of collected isolates and spreading nearly statewide. Although symbiotic strains are significantly more common, more likely to be dominant, and more likely to be epidemic, we found evidence of both symbiotic and non-symbiotic strains spreading across great distances. We did not find support for the hypothesis that acquisition of symbiosis islands serve as a driver of strain dominance or spread. We also uncovered other potential drivers of population differentiation in rhizobial communities including locale, portion of the legume inhabited, symbiotic capacity, and collection year.

References

Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**: 539–552.

Bach, S. (2000) The *Yersinia* high-pathogenicity island is present in different members of the family Enterobacteriaceae. *FEMS Microbiol. Lett.* **183**: 289–294.

Berg, G. and Smalla, K. (2009) Plant species and soil type cooperatively shape the structure and function of microbial communities in the rhizosphere: Plant species, soil type and rhizosphere communities. *FEMS Microbiol. Ecol.* **68**: 1–13.

Bhatt, A.S., Freeman, S.S., Herrera, A.F., Pedamallu, C.S., Gevers, D., Duke, F., et al. (2013) Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N. Engl. J. Med.* **369**: 517–528.

Bright, M. and Bulgheresi, S. (2010) A complex journey: Transmission of microbial symbionts. *Nat. Rev. Microbiol.* **8**: 218–230.

Chaintreuil, C., Giraud, E., Prin, Y., Lorquin, J., Ba, A., Gillis, M., et al. (2000) Photosynthetic bradyrhizobia are natural endophytes of the african wild rice *Oryza breviligulata*. *Appl. Environ. Microbiol.* **66**: 5437–5447.

Cleveland, C.C., Townsend, A.R., Schimel, D.S., Fisher, H., Howarth, R.W., Hedin, L.O., et al. (1999) Global patterns of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. *Glob. Biogeochem. Cycles* **13**: 623–645.

Cole, M.A. and Elkan, G.H. (1979) Multiple antibiotic resistance in *Rhizobium japonicum*. *Appl. Environ. Microbiol.* **37**: 867–870.

Costello, E.K., Carlisle, E.M., Bik, E.M., Morowitz, M.J., and Relman, D.A. (2013) Microbiome assembly across multiple body sites in low-birthweight infants. *mBio* **4**: e00782–13–e00782–13.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012) jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**: 772.

Diep, B.A., Gill, S.R., Chang, R.F., Phan, T.H., Chen, J.H., Davidson, M.G., et al. (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet* **367**: 731–739.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* **2**: 414–424.

Dupuy, N., Willems, A., Pot, B., Dewettinck, D., Vandenbruaene, I., Maestrojuan, G., et al. (1994) Phenotypic and genotypic characterization of bradyrhizobia nodulating the leguminous tree *Acacia albida*. *Int. J. Syst. Bacteriol.* **44**: 461–473.

Ehinger, M., Mohr, T.J., Starcevich, J.B., Sachs, J.L., Porter, S.S., and Simms, E.L. (2014) Specialization-generalization trade-off in a *Bradyrhizobium* symbiosis with wild legume hosts. *BMC Ecol.* **14**: 8.

Excoffier, L. and Lischer, H.E.L. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**: 564–567.

Galibert, F. (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**: 668–672.

Gal-Mor, O. and Finlay, B.B. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell. Microbiol.* **8**: 1707–1719.

Geurts, R. (1996) Signal transduction in *Rhizobium*-induced nodule formation. *Plant Physiol.* **112**: 447–453.

Giraud, E., Moulin, L., Vallenet, D., Barbe, V., Cytryn, E., Avarre, J.-C., et al. (2007) Legumes symbioses: Absence of Nod genes in photosynthetic bradyrhizobia. *Science* **316**: 1307–1312.

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

Graham, P.H. and Vance, C.P. (2003) Legumes: Importance and constraints to greater use. *Plant Physiol.* **131**: 872–877.

Groisman, E.A. and Ochman, H. (1996) Pathogenicity islands: Bacterial evolution in quantum leaps. *Cell* **87**: 791–794.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**: 307–321.

Hamady, M., Lozupone, C., and Knight, R. (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**: 17–27.

Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E., and Challenger, W. (2007) GEIGER: Investigating evolutionary radiations. *Bioinformatics* **24**: 129–131.

Heck, K.L., van Belle, G., and Simberloff, D. (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* **56**: 1459.

Hudson, R.R. (1987) Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**: 245–250.

Hudson, R.R. and Kaplan, N.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

Hunt, K.M., Foster, J.A., Forney, L.J., Schütte, U.M.E., Beck, D.L., Abdo, Z., et al. (2011) Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS ONE* **6**: e21313.

Jain, R., Rivera, M.C., Moore, J.E., and Lake, J.A. (2003) Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**: 1598–1602.

Kaneko, T. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res.* **7**: 331–338.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., et al. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* **9**: 189 –197.

Lee, K.-B., De Backer, P., Aono, T., Liu, C.-T., Suzuki, S., Suzuki, T., et al. (2008) The genome of the versatile nitrogen fixer *Azorhizobium caulinodans* ORS571. *BMC Genomics* **9**: 271.

Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49–67.

Librado, P. and Rozas, J. (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Maddison, W.P. and Maddison, D.R. (2005) MacClade: Analysis of phylogeny and character evolution.

Maddison, W. P. and Maddison, D. R. (2011) Mesquite.

Mcinnes, A. (2004) Structure and diversity among rhizobial strains, populations and communities - A review. *Soil Biol. Biochem.* **36**: 1295–1308.

Moulin, L., Béna, G., Boivin-Masson, C., and Stępkowski, T. (2004) Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the *Bradyrhizobium* genus. *Mol. Phylogenet. Evol.* **30**: 720–732.

Nei, M. (1987) Molecular evolutionary genetics Columbia University Press, New York.

Okubo, T., Tsukui, T., Maita, H., Okamoto, S., Oshima, K., Fujisawa, T., et al. (2012) Complete genome sequence of *Bradyrhizobium* sp. S23321: insights into symbiosis evolution in soil oligotrophs. *Microbes Environ. JSME* **27**: 306–315.

Orme, D. (2012) The caper package: Comparative analysis of phylogenetics and evolution in R.

Parker, M. (2000) Divergent *Bradyrhizobium* symbionts on *Tachigali versicolor* from Barro Colorado Island, Panama. *Syst Appl Microbiol* **23**: 585–90.

Parker, M.A. (2014) The spread of *Bradyrhizobium* lineages across host legume clades: from *Abarema* to *Zygia*. *Microb. Ecol.*

Parker, M.A., Lafay, B., Burdon, J.J., and van Berkum, P. (2002) Conflicting phylogeographic patterns in rRNA and *nifD* indicate regionally restricted gene transfer in *Bradyrhizobium*. *Microbiol. Read. Engl.* **148**: 2557–2565.

Perret, X., Staehelin, C., and Broughton, W.J. (2000) Molecular basis of symbiotic promiscuity. *Microbiol. Mol. Biol. Rev.* **64**: 180 –201.

Pessi, G., Ahrens, C.H., Rehrauer, H., Lindemann, A., Hauser, F., Fischer, H.-M., and Hennecke, H. (2007) Genome-wide transcript analysis of *Bradyrhizobium japonicum* bacteroids in soybean root nodules. *Mol. Plant. Microbe Interact.* **20**: 1353–1363.

Pongsilp, N., Teaumroong, N., Nuntagij, A., Boonkerd, N., and Sadowsky, M. (2002) Genetic structure of indigenous non-nodulating and nodulating populations of *Bradyrhizobium* in soils from Thailand. *Symbiosis* **33**: 39–58.

Qiu, X., Gurkar, A.U., and Lory, S. (2006) Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci.* **103**: 19830–19835.

Rivas, R., Martens, M., de Lajudie, P., and Willems, A. (2009) Multilocus sequence analysis of the genus *Bradyrhizobium*. *Syst. Appl. Microbiol.* **32**: 101–110.

Rosenberg, M.S. and Anderson, C.D. (2011) PASSaGE: Pattern analysis, spatial statistics and geographic exegesis. Version 2: PASSaGE. *Methods Ecol. Evol.* **2**: 229–232.

Sachs, J.L., Ehinger, M.O., and Simms, E.L. (2010) Origins of cheating and loss of symbiosis in wild *Bradyrhizobium*. *J. Evol. Biol.* **23**: 1075–1089.

Sachs, J.L., Kembel, S.W., Lau, A.H., and Simms, E.L. (2009) In Situ phylogenetic structure and diversity of wild *Bradyrhizobium* communities. *Appl. Environ. Microbiol.* **75**: 4727–4735.

Sachs, J.L., Russell, J.E., and Hollowell, A.C. (2011) Evolutionary instability of symbiotic function in *Bradyrhizobium japonicum*. *PLoS ONE* **6**: e26370.

Sachs, J.L., Skophammer, R.G., Bansal, N., and Stajich, J.E. (2013) Evolutionary origins and diversification of proteobacterial mutualists. *Proc. R. Soc. B Biol. Sci.* **281**: 20132146–20132146.

Saito, A., Mitsui, H., Hattori, R., Minamisawa, K., and Hattori, T. (1998) Slow-growing and oligotrophic soil bacteria phylogenetically close to *Bradyrhizobium japonicum*in. *FEMS Microbiol. Ecol.* **25**: 277–286.

SAS Institute Inc (1989) JMP SAS Institute Inc., Cary, NC.

Segovia, L., Piñero, D., Palacios, R., and Martínez-Romero, E. (1991) Genetic structure of a soil population of *Rhizobium leguminosarum*. *Appl. Environ. Microbiol.* **57**: 426–433.

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

Silva, C., Eguiarte, L.E., and Souza, V. (1999) Reticulated and epidemic population genetic structure of *Rhizobium etli* biovar *phaseoli* in a traditionally managed locality in Mexico. *Mol. Ecol.* **8**: 277–287.

Sprent, J.I. (2001) Nodulation in legumes Royal Botanic Gardens, Kew.

Sugawara, M., Epstein, B., Badgley, B.D., Unno, T., Xu, L., Reese, J., et al. (2013) Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol.* **14**: R17.

Sullivan, J.T., Eardly, B.D., van Berkum, P., and Ronson, C.W. (1996) Four unnamed species of nonsymbiotic rhizobia isolated from the rhizosphere of *Lotus corniculatus*. *Appl. Environ. Microbiol.* **62**: 2818–2825.

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

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.

Uchiumi, T., Ohwada, T., Itakura, M., Mitsui, H., Nukui, N., Dawadi, P., et al. (2004) Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. *J. Bacteriol.* **186**: 2439–2448.

Vinuesa, P., Rojas-Jimenez, K., Contreras-Moreira, B., Mahna, S.K., Prasad, B.N., Moe, H., et al. (2008) Multilocus sequence analysis for assessment of the biogeography and evolutionary genetics of four *Bradyrhizobium* species that nodulate soybeans on the Asiatic continent. *Appl. Environ. Microbiol.* **74**: 6987–6996.

Vinuesa, P., Silva, C., Werner, D., and Martínez-Romero, E. (2005) Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation. *Mol. Phylogenet. Evol.* **34**: 29–54.

Young, J.P.W. (1996) Diversity and phylogeny of rhizobia. *New Phytol.* **133**: 87–94.

Young, J.P.W., Crossman, L.C., Johnston, A.W.B., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., et al. (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **7**: R34.

Table 1. 1 Dominant and Epidemic haplotypes. All dominant haplotypes (collected 5 times and encompassing at least 10% of the isolates) are listed for each collection site followed by epidemic haplotypes (dominant haplotypes that have spread at least 10km away). Abundance indicates the number of times the haplotype was collected at a given site. Whereas the spatial spread is the greatest distance between collection sites the haplotype was found at. The information is then broken down by symbiotic capacity.

		# isolates from collection site	Total			Symbiotic		Non-symbiotic	
			Haplotype	Abundance	Spatial Spread	Abundance	Spatial Spread	Abundance	Spatial Spread
Dominant Haplotypes	Anza Borrego State Park - Palm Canyon	9	none			none		none	
	Anza Borrego State Park - Roadside	36	G58_R27	23	135.7	23	135.7	none	
			G58_R34	9	135.7	9	135.7	none	
	Bodega Marine Reserve	211	G03_R01	72	750.6	71	750.6	1	723.8
			G05_R02	24	0	22	0	2	0
	Burns Piñon Ridge Reserve	39	G03_R01	24	750.6	24	750.6	none	
			G11_R07	9	175.9	9	175.9	none	
			G40_R38	5	168.8	5	168.8	none	
	Gaudalupe-Nipomo Dunes National Wildlife Refuge	100	G03_R01	18	750.6	18	750.6	none	
			G05_R65	35	0	35	0	none	
	Griffith Park	69	G11_R07	7	175.9	7	175.9	none	
			G40_R38	50	168.8	50	168.8	none	
	Madrona Marsh Preserve	86	G03_R01	24	750.6	24	750.6	none	
			G11_R01	25	335.3	25	335.3	none	
			G36_R35	10	100.2	10	100.2	none	
	Motte Rimrock Reserve	120	G03_R01	35	750.6	23	750.6	12	723.8
			G11_R01	22	335.3	19	335.3	3	18.86
			G64_R29	13	0	none		13	0
	Pismo Dunes Natural Preserve	21	G99_R62	5	0	5	0	none	
	Robert J. Bernard Biological Field Station	68	G03_R01	26	750.6	26	750.6	none	
			G11_R01	11	335.3	11	335.3	none	
			G71_R32	11	7.04	11	7.04	none	
	San Dimas Canyon	67	G03_R01	11	750.6	11	750.6	none	
			G36_R35	9	100.2	9	100.2	none	
			G71_R39	8	0	8	0	none	
	San Dimas Reservoir	63	G03_R01	18	750.6	18	750.6	none	
			G11_R07	16	175.9	16	175.9	none	
	UC Riverside Hills	356	G03_R01	105	750.6	45	750.6	60	723.8
	Whitewater Preserve	47	G03_R01	22	750.6	22	750.6	none	
Epidemic Haplotypes			G03_R01	355	750.6	282	750.6	73	723.8
			G11_R01	81	335.3	70	335.3	11	18.86
			G11_R07	62	175.9	58	175.9	4	0
			G36_R35	20	100.2	19	100.2	1	0
			G40_R38	55	168.8	55	168.8	none	
			G58_R27	27	135.7	24	135.7	3	0
			G58_R34	10	135.7	9	135.7	none	

Table 1. 2 Species statistics for 19 *Bradyrhizobium* species calculated in DnaSP. *Bradyrhizobium retamae* was excluded from analyses as it included only a single isolate. Analyses included strain richness (number of unique haplotypes/number of isolates), Hd (haplotype diversity – probability that two haplotypes drawn uniformly at random from the population are not the same), π (nucleotide diversity – average number of nucleotide differences per site between two sequences), k (average # nucleotide differences), linkage disequilibrium (average absolute D'), recombination (R), and the minimum number of recombination events.

	# isolates	# haplotypes	# variable sites	Strain Richness (# haplotypes/# isolates)	Hd (haplotype diversity)	π (nucleotide diversity)	k (average # of nucleotide differences)	Linkage disequilibrium	Recombination		
								Average |D'|	Estimate of R/gene	Estimate of R between adjacent sites	Minimum # recombination events
B. betae	15	7	77	0.47	0.84	0.03	28	1.00	0.001	0.000	0
B. canariense	798	105	215	0.13	0.78	0.01	8	0.97	0.001	0.000	41
B. cytisi	40	14	92	0.35	0.85	0.03	30	0.98	0.300	0.000	9
B. lablabi	9	7	101	0.78	0.94	0.03	32	0.99	1.200	0.001	8
B. liaoningense	9	6	125	0.67	0.89	0.04	39	1.00	0.001	0.000	4
B. novel I	134	17	54	0.13	0.78	0.01	8	0.91	0.001	0.000	5
B. novel II	73	22	111	0.30	0.87	0.02	16	0.95	0.001	0.000	17
B. novel III	12	6	189	0.50	0.82	0.09	84	0.95	5.800	0.006	16
B. novel IV	3	2	4	0.67	0.67	0.00	3	1.00	7.300	0.008	0
B. novel V	6	2	33	0.33	0.60	0.02	20	1.00	0.700	0.001	0
B. novel VI	6	5	180	0.83	0.93	0.10	93	0.95	5.500	0.006	10
B. novel VII	3	2	67	0.67	0.67	0.05	45	1.00	1.500	0.002	0
B. novel VIII	15	4	48	0.27	0.67	0.01	11	1.00	0.001	0.000	3
B. novel IX	8	4	4	0.50	0.64	0.00	1	1.00	54.600	0.056	0
B. novel X	70	18	40	0.26	0.74	0.01	6	0.96	1.700	0.002	5
B. novel XI	60	45	80	0.75	0.99	0.01	12	0.93	19.800	0.020	12
B. novel XII	4	4	51	1.00	1.00	0.03	32	1.00	> 10,000	n/a	0
B. novel XIII	2	2	4	1.00	1.00	0.00	4	0.96	20.700	0.021	6
B. novel XIV	2	2	7	1.00	1.00	0.01	7	1.00	> 10,000	n/a	0

Figure 1. 1 Map of California indicates collection sites with black dots. Pie charts connected to black dots illustrate the proportion of isolates haplotypes encompass. The five haplotypes chosen incorporate the four haplotypes with the highest abundance and the four haplotypes with the greatest spatial spread. The distribution of epidemic haplotype G03_R01 is estimated based on spatial spread among collection sites

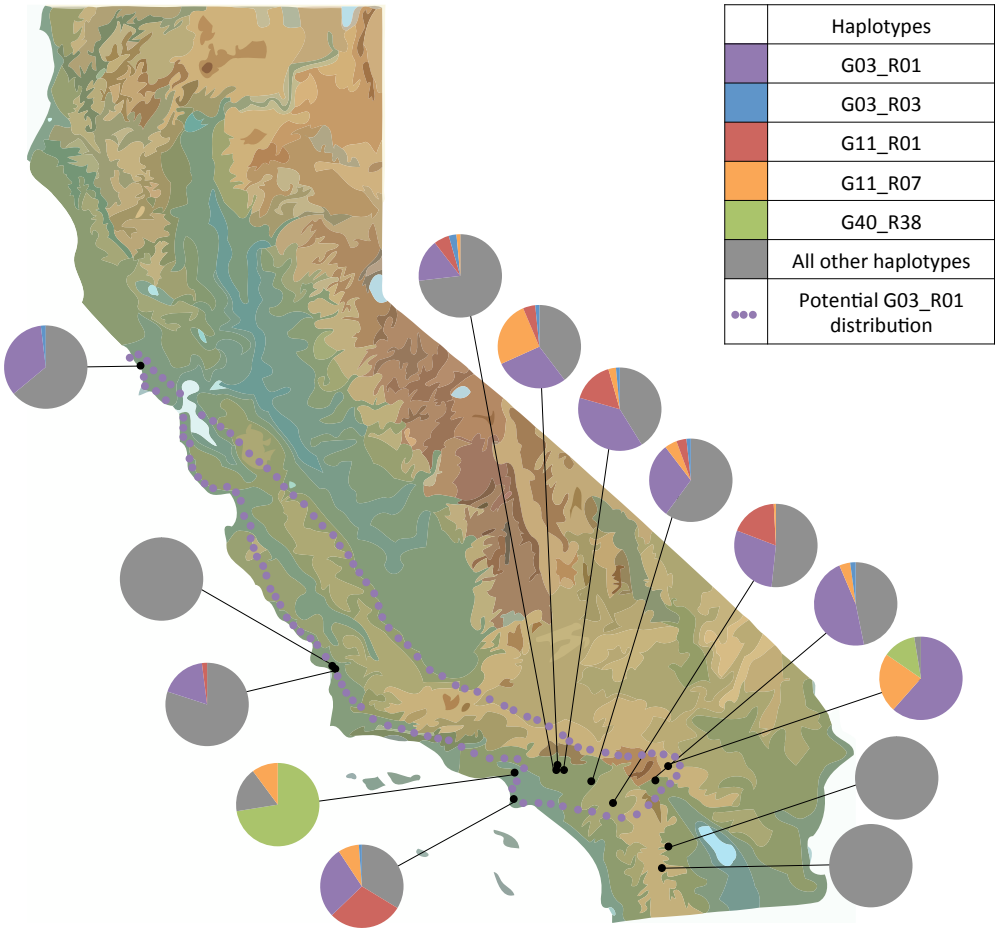
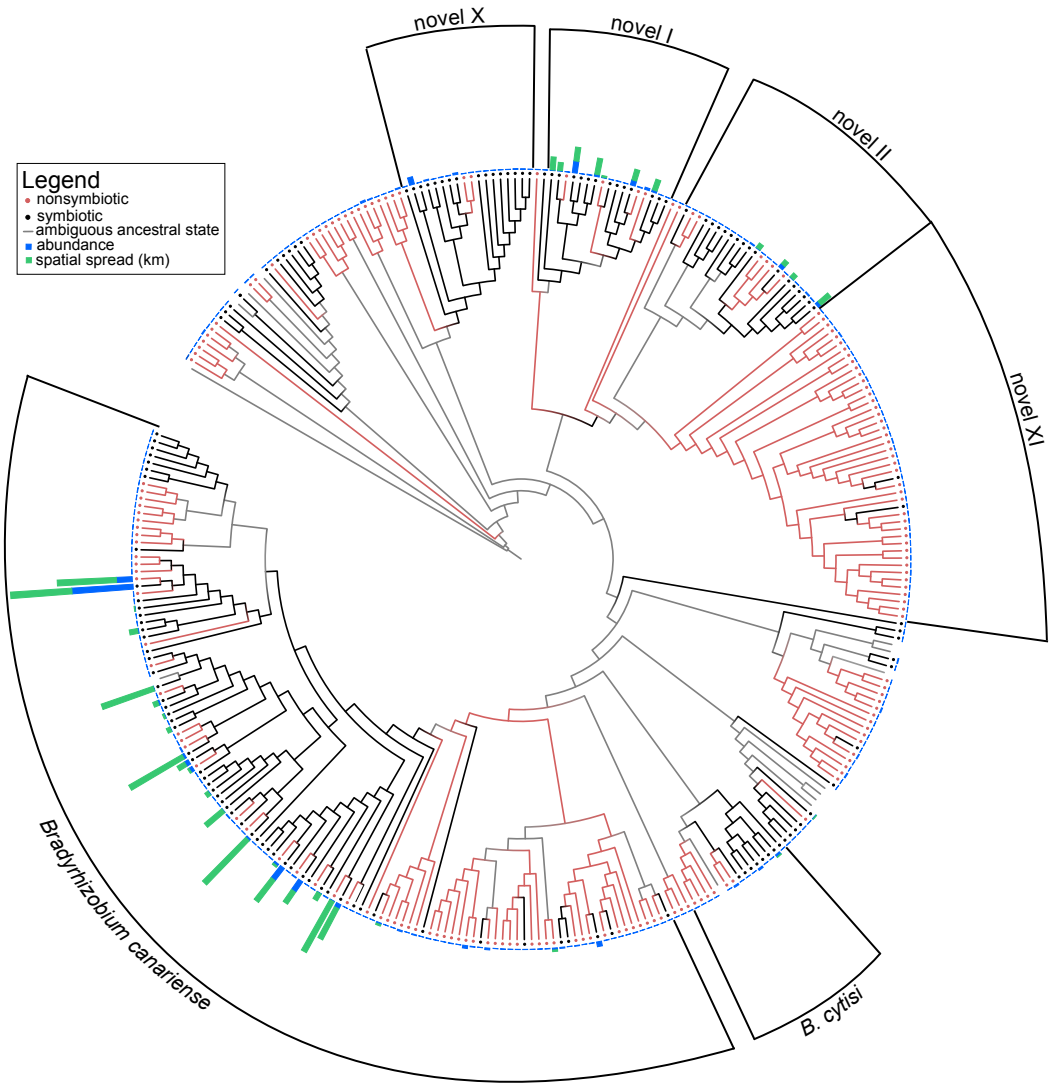


Figure 1. 2 PhyML 3.0 phylogenetic tree reconstructed from concatenated *glnII* and *recA* loci haplotypes. Non-symbiotic tips indicated with red dots and symbiotic tips with black. Ancestral state reconstruction conducted in Mesquite. Non-symbiotic and symbiotic branches are colored red and black respectively if their estimated likelihood proportions were > 66.66. Ambiguous lineages are colored grey. The relative abundance of a haplotype and the spatial spread are indicated by stacked blue and green bars. Major species clades are indicated with brackets. Reference strains can be identified by the lack of symbiotic capacity, abundance, and spatial spread data. The strains include (clockwise from the tree root): *Mesorhizobium loti* MAFF303099, *Bradyrhizobium retamae* Ro19, *Bradyrhizobium elkanii* USDA46, *Bradyrhizobium lablabi* CCBAU61434, *Bradyrhizobium diazoefficiens* SEMIA5080, *Bradyrhizobium diazoefficiens* (*japonicum*) USDA110, *Bradyrhizobium betae* PL7HG1, *Bradyrhizobium iriomotense* EK05, *Bradyrhizobium arachidis* CCBAU33067, *Bradyrhizobium yuanmingense* R2m, *Bradyrhizobium liaoningense* SEMIA5025, *Bradyrhizobium cytisi* LGM25866, and *Bradyrhizobium canariense* SEMIA928


Legend
nonsymbiotic
symbiotic
ambiguous ancestral state
abundance
spatial spread (km)
novel X
novel I
novel II
novel XI
Bradyrhizobium canariense
B. cytisi

CHAPTER 2

Massive clonal expansion of *Bradyrhizobium* across a 700km *Lotus* metapopulation

Abstract

Rhizobial bacteria capable of nodulating legume roots and fixing atmospheric nitrogen have profound impacts on host plant fitness and the ecosystem. Previous genetic analyses of rhizobia have focused on agricultural isolates, and have uncovered genetically diverse populations dominated by few strains. But researchers have little understanding of why certain genotypes dominate local sites or spread among sites or host populations. Here we genotyped >350 natural isolates of *Bradyrhizobium* from *Lotus strigosus* across an 840 km transect of native host populations. Like other rhizobia, *Bradyrhizobium* replicates both in soil and within host root tissues, and this dual lifestyle is reflected in the *Bradyrhizobium* genome with chromosomal loci being expressed under free-living conditions and symbiosis loci expressed *in planta*. We analyzed population genetic parameters of chromosomal and symbiosis island loci and examined haplotype diversity and abundance separately in each genome region. Our goal was to make predictions about the relative importance of the soil versus the plant host in structuring rhizobial populations. We uncovered a massive clonal expansion restricted to the *Bradyrhizobium* chromosome, largely driven by a single epidemic haplotype representing 17% of all isolates collected and found at the majority of the sites. The epidemic chromosomal haplotype was found associated with 42 symbiosis island haplotypes, none of which were

epidemic. This is contrary to the paradigm of a genomic island conferring a fitness advantage and increasing in frequency within the population. Our analyses reveal that selection in soils on chromosomal loci can be an important driver of rhizobial population structure, and suggests that selection by plants is more important at the scale of local host populations.

Introduction

Understanding bacterial epidemiology, including patterns of incidence, spatial spread, and human management, remains a critical challenge for ecologically and economically important taxa. Rhizobia are proteobacteria that form root nodules on legumes, convert dinitrogen gas into ammonia, and improve plant growth and productivity (Sprent, 2001). Rhizobial symbioses with legumes generate ~150 million tons of reactive nitrogen annually for the global nutrient cycle (Sugawara et al., 2013) and are a major factor in agriculture, since legumes contribute ~25% of global crop output (Ferguson et al., 2010). But the positive economic impact from manipulation of rhizobia remains unrealized for crop legumes. Agronomists have repeatedly failed to increase legume production with rhizobial inoculation, often because indigenous rhizobia dominate nodule occupancy without any yield increase -- a dilemma known as the "rhizobial competition problem" (Triplett and Sadowsky, 1992). In most cases, inoculant genotypes are rapidly purged by locally dominant rhizobial strains, even when inoculation occurs at high densities (Tang et al., 2012). Some inoculation programs have modestly improved crop yields by focusing on locally adapted strains, but even in these cases the inoculant strains make little impact on the rhizobial soil population (Hungria and Vargas, 2000; Mostasso et al., 2002; Hungria et al., 2003). Thus, a critical goal for basic and applied researchers is to understand the genetic and ecological drivers of strain dominance that structure rhizobial populations.

Population structure in bacteria is shaped by the degree to which strains propagate clonally, spread epidemically among sites and or hosts, and recombine at different spatial

scales (Maynard Smith et al., 1993). Rhizobia exhibit a bipartite lifecycle, replicating both in the soil and within host root tissues (Sachs et al., 2009). Rhizobial lifecycles are reflected in their bipartite genomes, with chromosomal loci being expressed under free-living conditions, and symbiosis loci being expressed *in planta* (Uchiumi et al., 2004; Pessi et al., 2007). Rhizobial symbiosis loci are grouped on transmissible plasmids or genomic islands that can be transmitted horizontally among chromosomal backgrounds (Kaneko et al., 2000, 2002; Young et al., 2006; Martinez-Abarca et al., 2013). In both agricultural and natural settings, rhizobial populations are often dominated by one or few rhizobial genotypes (Mcinnes, 2004), in some cases revealing that a subset of genotypes have rapidly increased in frequency (i.e., selective sweeps (Epstein et al., 2012)). Some work has suggested that legume hosts are the dominant force in shaping rhizobial populations. For instance, legume hosts can favor the *in planta* fitness of beneficial rhizobia over less-effective strains (Kiers et al., 2003; Regus et al., 2014; Sachs et al., 2010b; Simms et al., 2006) resulting in subsets of rhizobial symbiosis genotypes being structured among plant species (Parker, 2012). In contrast, selection on rhizobia *ex planta* remains poorly understood, which is problematic since competition is thought to be intense in soils (Denison and Kiers, 2004; Sachs et al., 2009), and the majority of the rhizobial genome encodes traits that are expressed outside of the host interaction (Uchiumi et al., 2004).

Bradyrhizobium is the most cosmopolitan rhizobial lineage (Parker, 2014), thriving in soil and aquatic environments and colonizing both plant and animal hosts, including humans (Chaintreuil et al., 2000; Vinuesa et al., 2005, 2008; Sachs et al., 2009;

Hunt et al., 2011; Bhatt et al., 2013; Costello et al., 2013). Diverse wild legumes are nodulated by *Bradyrhizobium* as are global staple crops such as soybeans (*Glycine*), peanuts (*Arachis*), and cowpea (*Vigna*). *Bradyrhizobium* populations have been reported to exhibit epidemic characteristics, defined as a small subset of individual genotypes dominating a local site (Sachs et al., 2009) or spreading among multiple locales (Vinuesa et al., 2005, 2008), but no study has analyzed *Bradyrhizobium* epidemics at the regional level. Here, we investigated the population genetic structure of *Bradyrhizobium* isolated from a metapopulation of *Lotus strigosus,* a common herb native to the Pacific Southwest of the United States. We cultured >350 *L. strigosus* nodules from natural sites across California encompassing 72 plants collected over a 840 km transect of the *L. strigosus* range. We sequenced isolates at 8 loci (~5.5kb) distributed across the ~9Mbp *Bradyrhizobium* genome including four loci on the chromosome and four within the integrated symbiosis island (~0.8Mb) (Kaneko et al., 2002, 2011). We examined the frequency and spatial distribution of epidemic rhizobial genotypes to address whether *Bradyrhizobium* can spread among host populations and across significant ecological barriers. To make predictions about the relative importance of the soil versus the plant host in structuring rhizobial populations, we compared population genetic parameters in the chromosomal loci versus the symbiosis island, and examined rates of recombination across these two genome regions. We investigated the fit of our dataset to four hypothetical scenarios of genome evolution, based on the relative roles of selective sweeps versus recombination across the *Bradyrhizobium* genome.

Materials and Methods

Collection of nodule isolates

Root nodules were collected from *L. strigosus* host plants at nine sampling locales across California including UC Reserves (Bodega Marine Reserve, Motte Rimrock Reserve, Burns Piñon Ridge Reserve), a natural site at UC Riverside, a biological field station in Claremont, CA (Robert J. Bernard Biological Field Station), two separate sites within Anza Borrego Desert State Park (Palm Canyon, Roadside), and two separate sites within an undeveloped natural area in San Dimas, CA (San Dimas Canyon, Reservoir) (Supplementary Table S2.1). Sampling took place over a span of eight years and some sites were sampled multiple times, allowing for temporal analysis. From each sampled nodule we isolated a single clone of *Bradyrhizobium* following published protocols (Sachs et al., 2009).

DNA amplification and sequencing

Genomic DNA extracts were purified using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) and were PCR amplified and sequenced at four loci located on the *Bradyrhizobium* chromosome, including *dnaK*, *glnII*, ITS, and *recA*, and four symbiosis island loci, including *nifD, nodD-A, nodZ* and *nolL.* The chromosome and symbiosis island are hereafter referred to as genomic regions ‘CHR’ and ‘SI’ respectively. PCR amplification followed previously published protocols (Vinuesa et al., 1998; Parker, 2000; van Berkum and Fuhrmann, 2000; Stępkowski et al., 2003; Moulin et al., 2004; Silva et al., 2005; Stepkowski et al., 2005; Sachs et al., 2009; Sachs, Ehinger, et al., 2010).

Amplicons were sequenced at the Institute for Integrative Genome Biology at UC Riverside. Sequences were analyzed using Lasergene 8 (DNASTAR, Madison, WI) and only reads with unambiguous peaks at all nucleotides were included in the analyses.

Phylogenetic Analyses

Sequences for each gene were aligned separately and were concatenated per genome region using Clustal Omega (Sievers et al., 2014) and using reference strains from *Mesorhizobium loti* (MAFF303099, NZP2037) and *Bradyrhizobium arachidis* (CCBAU051107, CCBAU23155, CCBAU33067, CCBAU45332), *B. betae* (PL7HG1), *B. canariense* (BTA-1, SEMIA928), *B. cytisi* (CTAW11, LGM25866), *B. daqingense* (CCBAU15774), *B. denitrificans* (LMG8443), *B. diazoefficiens* (SEMIA5080/USDA110), *B. elkanii* (USDA46, USDA76), *B. huanghuaihaiense* (CCBAU233), *B. iriomotense* (EK05, LMG24129), *B. japonicum* (USDA6), *B. jicamae* (PAC68), *B. lablabi* (CCBAU23086, CCBAU61434), *B. liaoningense* (SEMIA5025, SEMIA5062, USDA3622), *B. oligotrophica* (S58), *B. pachyrhizi* (PAC48), *B. retamae* (Ro19), *B. rifense* (CTAW71), *B. yuanmingense* (CCBAU33079, CCBAU33109, CCBAU53119, LMG21827, R2m, R3), and *Bradyrhizobium* sp. (BTAi1, ORS278, WM9). We used Akaike information criterion results from jModelTest 2 (Guindon and Gascuel, 2003; Darriba et al., 2012) to select the GTR model of nucleotide substitution for both genome regions. Phylogenetic trees of each genome region were reconstructed in PhyML 3.0 (Guindon et al., 2010) utilizing BioNJ as the starting tree with subtree pruning and regrafting (SPR). Branch support was estimated using the fast approximate

likelihood ratio test (aLRT) with the Shimodaira-Hasegawa-like (SH-like) procedure (Anisimova and Gascuel, 2006). Based on the phylogeny, different species of *Bradyrhizobium* were defined as highly supported, non-nested, monophyletic clades (Shimodaira-Hasegawa support $\geq$0.90) including no more than one reference species, attempting to follow past species demarcations that used some of the same loci (Shimodaira and Hasegawa, 1999; Vinuesa et al., 2008). A tanglegram connecting phylogenetic trees of each genome region was reconstructed in Treemap 3.0 (Charleston, 1998). Statistical significance of congruence between chromosome and symbiosis island genome region phylogenies was tested using AxParafit and AxPcoords (Stamatakis et al., 2007) within CopyCat (Meier-Kolthoff et al., 2007) using default parameters.

Statistical analyses

We estimated π (nucleotide diversity; (Nei, 1987), Hd (haplotype diversity; (Nei, 1987), k (average # nucleotide differences; (Tajima, 1983), linkage disequilibrium (average absolute D') (Lewontin, 1964; Parker, 2012), recombination (R) (Hudson, 1987), the minimum number of recombination events (Hudson and Kaplan, 1985), and Tajima's D. Population differentiation was calculated using F_{ST}, conducted on a base pair basis using a Perl script and the Weir-Cockerham method (Weir and Cockerham, 1984). Average F_{ST}s were calculated between collection sites for each genome region. Isolation by distance was tested with a Mantel test (Mantel, 1967) correlating F_{ST} and distance matrices between collection sites in PASSaGE (Rosenberg and Anderson, 2011). Loci were analyzed separately or were grouped into genome regions, recognizing the potential

for horizontal transfer of the SI (Sullivan and Ronson, 1998; Sachs, Ehinger, et al., 2010), and we analyzed collection sites and or *Bradyrhizobium* species separately when appropriate. We analyzed inter-species variation using the ratio of fixed to shared polymorphisms using DNASP (Librado and Rozas, 2009).

We identified isolates with identical haplotypes within one or both genome regions using the "find redundant" command in the MacClade program (Maddison, W. P. and Maddison, D. R., 2005). For each genome region haplotype, we calculated raw abundance (number of times a haplotype was isolated), and an adjusted abundance (only counting identical haplotypes from unique GPS locations), which discounts repeated isolation of the same haplotype from an individual plant, or closely neighboring plants. Distances within collection sites were small (<.3km), thus we used the geographic midpoint at each collection site (Clark et al., 2014) to calculate distance among collection sites. Strain richness and dominance were estimated, which are akin to species richness and dominance (Hurlbert, 1971; Peet, 1975). Strain richness was calculated for each locus and genome region by dividing the number of haplotypes by the number of isolates collected (Mcinnes, 2004). Strain dominance was calculated for each genome region haplotype as the percentage of isolates each haplotype contains (Mcinnes, 2004). Haplotypes were defined as 'dominant' if they were isolated at least 5 times and constituted at least 10% of the isolates within one of the nine collection sites. Haplotypes were defined as being 'epidemic' if they were dominant and were also found to have spread among multiple sites separated by at least 10km.

Testing of evolutionary-genomic scenarios

To investigate the relative roles of selective sweeps versus recombination across the two *Bradyrhizobium* genome regions, we examined the fit of our data to four potential evolutionary-genomic scenarios, including **i)** selective sweeps restricted to the SI, predicted if host plants select on symbiosis loci that recombine among diverse CHR backgrounds (SI Sweep; e.g., (Sullivan et al., 1995)), **ii)** selective sweeps restricted to the CHR, predicted if soil competition selects on CHR loci which recombine into multiple SI backgrounds (CHR Sweep), **iii)** whole genome selective sweeps, predicted if selection affects the whole genome without recombination (CHR-SI Sweep; e.g., (Diep et al., 2006)), and **iv)** a scenario of stable whole genome linkage, predicted in the absence of selective sweeps or recombination (CHR-SI Linkage; e.g., (Juhas et al., 2007)). To discriminate among these models we compared GC%, π, haplotype #, Hd, strain richness, linkage, and Tajima's D within and among genome regions (see Supplementary Information Appendix for predictions).

Results

Phylogenetic reconstruction

A total of 358 isolates were included in addition to reference strain sequences. (Supplementary Ins S2.1). We reconstructed a phylogeny using the CHR loci and recovered seven lineages that fit our operational definition of species (monophyletic; Sh-like branch support > 0.9; ≤1 reference species) including four previously defined species, *B. japonicum, B. canariense, B. retamae,* and *B. yuanmingense*, and two new ones, *B. sp. nov. I* and *B. sp. nov. II* (Figure 2.1 and Supplementary Figure S2.1). *B. sp. nov. I* was recovered multiple times at most Southern California sampling sites (6/8), and was most abundant at the Anza Borrego sites. *B. sp. nov. II* haplotypes were recovered only four times, within the two San Dimas sites (Supplementary Figure S2.2). The species designations were supported by a higher ratio of fixed differences to shared polymorphisms among species, with only two exceptions (*B. canariense* x *B japonicum*, *B. canariense* x *B. yuanmingense;* Supplementary Table S2.2). Two isolates, 12LoS3_5 and 12LoS6_1 did not fit in any of the recovered species.

We reconstructed a phylogeny using the SI loci and compared topologies of the CHR and SI trees (Supplementary Figure S2.3). A tanglegram analysis revealed congruence among major CHR and SI clades, especially within the *B. canariense, B. sp. nov. I*, and *B. retamae* CHR lineages that were each consistently associated with SI clades (#'s 4, #3, #1, respectively), and in the paraphyletic taxa *B. japonicum*, and *B. sp. nov. II* that consistently associated with SI clade #2 (Supplementary Figure S2.4 and S2.5). At least seven independent HGT events are also evident among the defined species

lineages (Supplementary Table S2.3 and Figure S2.5). We did not find significant support for congruence of the CHR and SI trees when we used the programs AxParafit and AxPcoords within CopyCat (ParaFitGlobal = 0.11097; p=0.53), but this test is sensitive to poor phylogenetic resolution near branch tips (Meier-Kolthoff et al., 2007; Stamatakis et al., 2007).

Genome region evolution

We examined the same number of loci and similar numbers of nucleotide sites and variable sites within the CHR and the SI (Supplementary Figure S2.6 and Table S2.4). The SI loci have reduced GC content compared with the CHR loci (except for ITS which encodes rDNA), consistent with sequenced *Bradyrhizobium* genomes USDA6 and USDA110 (Kaneko et al., 2002, 2011) and indicative of horizontal transfer of the symbiosis island (CHR, ~59% GC; SI, 55% GC; Supplementary Table S2.4 and Figure S2.6). The SI loci exhibit many more haplotypes on average (SI, 225; CHR, 138), greater haplotype diversity (SI, 0.993; CHR, 0.947), and greater strain richness than the CHR loci (SI, 0.63; CHR, 0.39), but the SI haplotypes have lower nucleotide diversity (SI, 0.02; CHR, 0.03) and differ by fewer nucleotides on average (SI, 52; CHR, 66; Supplementary Table S2.4 and Figure S2.6). We found relatively high linkage among all loci (average $|D'| > 0.9$), with the SI exhibiting greater linkage on average than the CHR (average $|D'| = 0.968$ versus 0.925). High linkage values between the genome regions (average $|D'| = 0.937$), and low estimates of recombination ($R = 0.001$ per gene, $R = 0.0000$ between adjacent sites (Hudson, 1987)) suggests that horizontal gene transfer

(HGT) of the SI occurs very infrequently (Supplementary Table S2.4 and Figure S2.6). Most parameters were similar among the different *Bradyrhizobium* species (Supplementary Table S2.5). However, we did not find consistently greater haplotype diversity in the SI (*B. yuanmingense* exhibited the opposite pattern). Moreover, we found the opposite pattern of higher nucleotide diversity in the SI than CHR for all species except *B. canariense* in which the diversity values were equivalent. Finally, we found that the average number of nucleotide differences (k) within the SI and CHR varied widely among the species (Supplementary Table S2.5).

Spatial Analyses

Bradyrhizobium species exhibited large differences in sampling range, with *B. canariense* having the largest range since it was the only species to be found in northern California (*B. canariense* ~750 km; *B. retamae*, *B. sp. nov. I* ~150 km; *B. sp. nov. II*, *B. japonicum* <10 km; Supplementary Figure S2.2). Species diversity of *Bradyrhizobium* also varied among sites, and was greatest at San Dimas Canyon, which contained all six species and was the only site with *B. yuanmingense*. Three locales only contained a single species each (Bodega Marine Reserve, Motte Rimrock reserve, and Anza Borrego Desert State Park - Roadside).

Nucleotide and haplotype diversity roughly paralleled species diversity among sampling locations and results were largely congruent between CHR and SI datasets (Supplementary Table S2.6). Differentiation among populations was lower for the SI than the CHR (mean F_{ST} 0.08, 0.20, respectively; Supplementary Table S2.6). Mean F_{ST} values

for the SI varied little among populations (0.06 - 0.13). In contrast, two populations exhibited F_{STs} for the CHR loci that were well beyond this range (Palm Canyon, mean F_{ST} = 0.25; Burns Piñon Ridge, F_{ST} = 0.83). Using a Mantel test, we did not find support for isolation by distance within the CHR or SI datasets (CHR; R = -0.00145; p = 0.99; SI, R = 0.09132; p= 0.72).

Whole genome haplotypes were assigned from concatenated sequences of all eight loci. Most whole-genome haplotypes were unique (86%) and included only a single representative isolate. Only five whole-genome haplotypes were found to be dominant in any single site, ranging from ~10-14% dominance of local isolates. No whole-genome haplotypes were found at multiple sites. Ten of the 138 CHR haplotypes were categorized as dominant and only two of these were found at multiple sites ≥10km distant, and thus were categorized as epidemic. Seven of the 225 SI haplotypes were categorized as dominant within a site, and none were found at multiple sites (Table 2.1). Among the 10 dominant CHR haplotypes, different ones dominated at each site, except in one case (Figure 2.2). The haplotype K01_G03_I01_R01, was found at all six non-desert collection sites and was dominant at five of them.

CHR Haplotype K01_G03_I01_R01 exhibited a spatial spread of 728.3 km, and encompassed >17% of all isolates assayed, consistent with a massive clonal expansion of the CHR (Maynard Smith et al., 1993; Mcinnes, 2004; Silva et al., 2005) (Figure 2.1 and Supplementary Figure S2.1). A broad diversity of SI haplotypes are associated with K01_G03_I01_R01 (42 SI haplotypes, encompassing 61 nucleotide changes), revealing that the epidemic CHR haplotype has recombined with divergent SI haplotypes as it has

spread (Figure 2.3). Two of the central and most genetically diverse locales were each sampled two times over different years (Claremont, San Dimas), and data from both sites suggest that K01_G03_I01_R01 has contracted over that period.

Hypothesis testing of evolutionary-genomic scenarios

The four evolutionary-genomic scenarios (SI Sweep, CHR Sweep, CHR-SI Sweep, and CHR-SI Linkage) were tested separately on each well-sampled *Bradyrhizobium* species (>20 isolates) because these lineages were diverged and exhibited varied evolutionary genetic parameters (Supplementary Table S2.7). For *B. canariense,* for which we have the largest sample, the data support the CHR Sweep model, as all the tested data are consistent with this hypothesis (Supplementary Table S2.7). For *B. nov. I,* n=54, with the next largest dataset, the data also support the CHR Sweep model, except that there is no evidence of a negative Tajima's D, which is a test for selective sweeps. The data for *B. japonicum,* and *B. yuanmingense,* do not strongly support any of the models, but both have relatively small sample sizes.

Discussion

Nitrogen-fixing rhizobial symbionts are ubiquitous and have profound impacts on host plant fitness and the ecosystem. But we have little understanding of why certain genotypes dominate local sites or spread among host populations. Previous studies of rhizobia have mostly focused on agricultural samples, and have showed that rhizobial populations are often genetically diverse, but dominated by a small subset of strains that reach high local abundance (Mcinnes, 2004). We investigated native legume hosts in natural soils to avoid the potentially confounding problems of tilling, introduced or genetically altered plants, and biological soil amendments.

We found a surpising amount of *Bradyrhizobium* genetic diversity given that we only sampled from one host species. *L. strigosus* was only previously found to nodulate *B. japonicum, B. diazoefficiens* and *B. canariense* (Sachs et al., 2009; Ehinger et al., 2014). In addition to these species, we also recovered *B retamae*, and *B. yuanmingense*, as well as two novel species, temporarily named *B. sp. novel I* and *B. sp. novel II* (Figure 2.1 and Supplementary Figure S2.1). We found wide variation among collection sites in terms of the *Bradyrhizobium* species that were recovered. Some locations were species diverse (San Dimas Canyon) while three included only a single species (Supplementary Figure S2.2). It does not appear that this pattern is a result of sampling bias, as the site we sampled most densely (Bodega Marine Reserve; n = 108) included only a single species and one of the least well sampled (San Dimas Reservoir; n = 13) was one of the most diverse (Supplementary Figure S2.2). Among the four most abundant species (*B. canariense*, *B. japonicum*, *B. sp. Novel I*, and *B. yuanmingense*), we observed differences

in nucleotide diversity, haplotype diversity, strain richness, and abundance, especially within the SI, suggesting that the evolution of this genome region differs among lineages. Most of the recovered isolates (244/358) were classified as *B. canariense,* resulting in overrepresentation in analyses utilizing the total dataset. Despite greater sampling, strain richness for *B. canariense* was lower than the other species for all loci. This is consistent with our finding that most abundant haplotypes were *B. canariense* (6/10 dominant haplotypes and 2/2 epidemic haplotypes) (Table 2.1).

We uncovered strikingly different population genetic parameters between the CHR and SI genome regions despite their similar numbers of nucleotide sites and variable sites. We found significantly fewer haplotypes, lower haplotype diversity, and lower strain richness within the CHR loci, all consistent with additional pressure of natural selection on the CHR. Patterns of haplotype abundance and spatial spread were also consistent with the role of natural selection shaping the CHR. Focusing only upon the CHR, we found that most field sites had one or a handful of dominant strains that were unique to that site. Among the two CHR haplotypes that were categorized as epidemic, one was found to be dominant at the majority of sampled sites and spread over a 700km span to represent >17% of all isolates sampled (CHR haplotype, K01_G03_I01_R01, Figure 2.3 and Table 2.1). This striking evidence of recent CHR clonal expansions was not mirrored in the SI dataset. Although we uncovered locally dominant SI haplotypes within 5/9 collection sites, none had spread >10km and thus were not defined as epidemic (Table 2.1). Moreover, all the dominant SI haplotypes were

nested subsets of CHR haplotypes, suggesting that SI dominance is driven by hitchhiking with the CHR.

A key challenge in bacterial population genetics is to understand the drivers of bacterial epidemics, especially in pathogens (Maynard Smith et al., 1993, 2000; Karaolis et al., 1998; Diep et al., 2006). One known mechanism is the horizontal transfer of accessory DNA, wherein acquisition of antibiotic resistance traits or vaccine-escape loci can result in epidemic spread of pathogens (Croucher et al., 2009, 2011). Parallel processes can occur in bacterial symbionts, except with the host promoting rather than countering bacterial spread. For instance, in rhizobial populations host plants can favor beneficial over ineffective rhizobial genotypes (Kiers et al., 2003; Simms et al., 2006; Sachs, Russell, et al., 2010) and thus select certain symbiosis locus variants over others (Koppell and Parker, 2012; Parker, 2012). If plant selection is intense, it can promote selective sweeps of certain symbiosis genotypes (Sullivan and Ronson, 1998) thus degrading molecular variation in symbiosis encoding genome regions (Vinuesa et al., 2005). In a striking example from an agricultural site, a single SI genotype from an introduced *Mesorhizobium loti* strain was acquired and spread through a diverse population of non-symbiotic *Mesorhizobium spp* (Sullivan and Ronson, 1998). Converse to this example, we found evidence of a CHR sweep. We considered four potential models of genome evolution during epidemic spread and found the best support for a model of a selective sweep restricted to the chromosome, with the bulk of the evidence within the *B. canariense* lineage (other lineages were not as well sampled; Supplementary Table S2.7). Thus our data are not consistent with selection by plant hosts

as driving rhizobial dominance, but instead suggest that selection among *Bradyrhizobium* in the soil is the driving force structuring these populations.

Genomic data of low nucleotide diversity in the chromosome, indicative of extensive hitchhiking following a selective sweep, was also found in *Sinorhizobium melliloti* (Epstein et al., 2012). But in the case of our dataset, the epidemic distribution in *B. canariense* is strictly limited to the CHR. For instance, of the 42 SI haplotypes associated with the epidemic CHR haplotype, only one is dominant (Figure 2.3 and Table 2.1). Thus, the most likely explanation is that the epidemic CHR haplotype has spread rapidly via HGT among divergent SI haplotypes, with little evidence of the SI hitchhiking. Importantly, our data allow us to reject the hypothesis of a genome-wide sweep followed by variation in the SI because we found significantly faster molecular evolution in the CHR versus the SI (328 versus 236 mutations, respectively within the best sampled *B. canariense/*Clade 4 clade). Instead, this epidemic distribution supports a sweep of the CHR and the acquisition of different symbiosis islands via horizontal gene transfer. Previous work shows that *Bradyrhizobium* and *Mesorhizobium* genomes exhibit 'expression islands', wherein the symbiosis loci are primarily expressed in nodules and chromosomal loci are primarily expressed *ex planta* (Uchiumi et al., 2004; Pessi et al., 2007). Thus, our dataset strongly suggests that selection in the soil is the key driver structuring rhizobial populations, whereas *in planta* selection is relatively less important.

References

Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**: 539–552.

Van Berkum, P. and Fuhrmann, J.J. (2000) Evolutionary relationships among the soybean bradyrhizobia reconstructed from 16S rRNA gene and internally transcribed spacer region sequence divergence. *Int. J. Syst. Evol. Microbiol.* **50**: 2165–2172.

Bhatt, A.S., Freeman, S.S., Herrera, A.F., Pedamallu, C.S., Gevers, D., Duke, F., et al. (2013) Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N. Engl. J. Med.* **369**: 517–528.

Chaintreuil, C., Giraud, E., Prin, Y., Lorquin, J., Ba, A., Gillis, M., et al. (2000) Photosynthetic bradyrhizobia are natural endophytes of the African wild rice *Oryza breviligulata*. *Appl. Environ. Microbiol.* **66**: 5437–5447.

Charleston, M.A. (1998) Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.* **149**: 191–223.

Clark, R.W., Schuett, G.W., Repp, R.A., Amarello, M., Smith, C.F., and Herrmann, H.-W. (2014) Mating systems, reproductive success, and sexual selection in secretive species: A case study of the western diamond-backed rattlesnake, *Crotalus atrox*. *PLoS ONE* **9**: e90616.

Costello, E.K., Carlisle, E.M., Bik, E.M., Morowitz, M.J., and Relman, D.A. (2013) Microbiome assembly across multiple body sites in low-birthweight infants. *mBio* **4**: e00782–13–e00782–13.

Croucher, N.J., Harris, S.R., Fraser, C., Quail, M.A., Burton, J., van der Linden, M., et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**: 430–434.

Croucher, N.J., Walker, D., Romero, P., Lennard, N., Paterson, G.K., Bason, N.C., et al. (2009) Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone *Streptococcus pneumoniae* Spain23F ST81. *J. Bacteriol.* **191**: 1480–1489.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012) jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**: 772.

Denison, R.F. (2000) Legume sanctions and the evolution of symbiotic cooperation by rhizobia. *Am. Nat.* **156**: 567–576.

Denison, R. and Kiers, E. (2004) Lifestyle alternatives for rhizobia: mutualism, parasitism, and forgoing symbiosis. *FEMS Microbiol. Lett.* **237**: 187–193.

Diep, B.A., Gill, S.R., Chang, R.F., Phan, T.H., Chen, J.H., Davidson, M.G., et al. (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet* **367**: 731–739.

Ehinger, M., Mohr, T.J., Starcevich, J.B., Sachs, J.L., Porter, S.S., and Simms, E.L. (2014) Specialization-generalization trade-off in a *Bradyrhizobium* symbiosis with wild legume hosts. *BMC Ecol.* **14**: 8.

Epstein, B., Branca, A., Mudge, J., Bharti, A.K., Briskine, R., Farmer, A.D., et al. (2012) Population genomics of the facultatively mutualistic bacteria *Sinorhizobium meliloti* and *S. medicae*. *PLoS Genet.* **8**: e1002868.

Ferguson, B.J., Indrasumunar, A., Hayashi, S., Lin, M.-H., Lin, Y.-H., Reid, D.E., and Gresshoff, P.M. (2010) Molecular analysis of legume nodule development and autoregulation. *J. Integr. Plant Biol.* **52**: 61–76.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**: 307–321.

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**: 696–704.

Hudson, R.R. (1987) Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**: 245–250.

Hudson, R.R. and Kaplan, N.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

Hungria, M., Campo, R.J., and Mendes, I.C. (2003) Benefits of inoculation of the common bean (*Phaseolus vulgaris*) crop with efficient and competitive *Rhizobium tropici* strains. *Biol. Fertil. Soils* **39**: 88–93.

Hungria, M. and Vargas, M.A.T. (2000) Environmental factors affecting N2 fixation in grain legumes in the tropics, with an emphasis on Brazil. *Field Crops Res.* **65**: 151–164.

Hunt, K.M., Foster, J.A., Forney, L.J., Schütte, U.M.E., Beck, D.L., Abdo, Z., et al. (2011) Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS ONE* **6**: e21313.

Hurlbert, S.H. (1971) The nonconcept of species diversity: A critique and alternative parameters. *Ecology* **52**: 577.

Juhas, M., Power, P.M., Harding, R.M., Ferguson, D.J., Dimopoulou, I.D., Elamin, A.R., et al. (2007) Sequence and functional analyses of *Haemophilus* spp. genomic islands. *Genome Biol.* **8**: R237.

Kaneko, T., Maita, H., Hirakawa, H., Uchiike, N., Minamisawa, K., Watanabe, A., and Sato, S. (2011) Complete genome sequence of the soybean symbiont *Bradyrhizobium japonicum* strain USDA6T. *Genes* **2**: 763–787.

Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., et al. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **7**: 331–338.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., et al. (2002) Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res.* **9**: 189 –197.

Karaolis, D.K.R., Johnson, J.A., Bailey, C.C., Boedeker, E.C., Kaper, J.B., and Reeves, P.R. (1998) A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains. *Proc. Natl. Acad. Sci.* **95**: 3134–3139.

Kiers, E.T., Rousseau, R.A., West, S.A., and Denison, R.F. (2003) Host sanctions and the legume–rhizobium mutualism. *Nature* **425**: 78–81.

Koppell, J.H. and Parker, M.A. (2012) Phylogenetic clustering of *Bradyrhizobium* symbionts on legumes indigenous to North America. *Microbiology* **158**: 2050–2059.

Lewontin, R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49–67.

Librado, P. and Rozas, J. (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**: 1451–1452.

Maddison, W. P. and Maddison, D. R. (2005) MacClade.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27**: 209–220.

Martinez-Abarca, F., Martinez-Rodriguez, L., Lopez-Contreras, J.A., Jimenez-Zurdo, J.I., and Toro, N. (2013) Complete Genome Sequence of the Alfalfa Symbiont *Sinorhizobium/Ensifer meliloti* Strain GR4. *Genome Announc.* **1**: e00174–12–e00174–12.

Maynard Smith, J., Feil, E.J., and Smith, N.H. (2000) Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssays* **22**: 1115–1122.

Maynard Smith, J., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993) How clonal are bacteria? *Proc. Natl. Acad. Sci.* **90**: 4384 –4388.

Mcinnes, A. (2004) Structure and diversity among rhizobial strains, populations and communities - A review. *Soil Biol. Biochem.* **36**: 1295–1308.

Meier-Kolthoff, J.P., Auch, A.F., Huson, D.H., and Goker, M. (2007) COPYCAT : cophylogenetic analysis tool. *Bioinformatics* **23**: 898–900.

Mostasso, L., Mostasso, F.L., Dias, B.G., Vargas, M.A.., and Hungria, M. (2002) Selection of bean (*Phaseolus vulgaris* L.) rhizobial strains for the Brazilian Cerrados. *Field Crops Res.* **73**: 121–132.

Moulin, L., Béna, G., Boivin-Masson, C., and Stępkowski, T. (2004) Phylogenetic analyses of symbiotic nodulation genes support vertical and lateral gene co-transfer within the *Bradyrhizobium* genus. *Mol. Phylogenet. Evol.* **30**: 720–732.

Nei, M. (1987) Molecular evolutionary genetics Columbia University Press, New York.

Parker, M. (2000) Divergent *Bradyrhizobium* symbionts on *Tachigali versicolor* from Barro Colorado Island, Panama. *Syst Appl Microbiol* **23**: 585–90.

Parker, M.A. (2012) Legumes select symbiosis island sequence variants in *Bradyrhizobium*. *Mol. Ecol.* **21**: 1769–1778.

Parker, M.A. (2014) The spread of *Bradyrhizobium* lineages across host legume clades: from *Abarema* to *Zygia*. *Microb. Ecol.*

Peet, R.K. (1975) Relative diversity indices. *Ecology* **56**: 496.

Pessi, G., Ahrens, C.H., Rehrauer, H., Lindemann, A., Hauser, F., Fischer, H.-M., and Hennecke, H. (2007) Genome-wide transcript analysis of *Bradyrhizobium japonicum* bacteroids in soybean root nodules. *Mol. Plant. Microbe Interact.* **20**: 1353–1363.

Regus, J.U., Gano, K.A., Hollowell, A.C., and Sachs, J.L. (2014) Efficiency of partner choice and sanctions in *Lotus* is not altered by nitrogen fertilization. *Proc. R. Soc. B Biol. Sci.* **281**: 20132587–20132587.

Rosenberg, M.S. and Anderson, C.D. (2011) PASSaGE: Pattern analysis, spatial statistics and geographic exegesis. Version 2: PASSaGE. *Methods Ecol. Evol.* **2**: 229–232.

Sachs, J.L., Ehinger, M.O., and Simms, E.L. (2010) Origins of cheating and loss of symbiosis in wild *Bradyrhizobium*. *J. Evol. Biol.* **23**: 1075–1089.

Sachs, J.L., Kembel, S.W., Lau, A.H., and Simms, E.L. (2009) In Situ phylogenetic structure and diversity of wild *Bradyrhizobium* communities. *Appl. Environ. Microbiol.* **75**: 4727–4735.

Sachs, J.L., Russell, J.E., Lii, Y.E., Black, K.C., Lopez, G., and Patil, A.S. (2010) Host control over infection and proliferation of a cheater symbiont. *J. Evol. Biol.* **23**: 1919–1927.

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

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., et al. (2014) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**: 539–539.

Silva, C., Vinuesa, P., Eguiarte, L.E., Souza, V., and Martínez-Romero, E. (2005) Evolutionary genetics and biogeographic structure of *Rhizobium gallicum sensu lato*, a widely distributed bacterial symbiont of diverse legumes. *Mol. Ecol.* **14**: 4033–4050.

Simms, E.L., Taylor, D.L., Povich, J., Shefferson, R.P., Sachs, J.., Urbina, M., and Tausczik, Y. (2006) An empirical test of partner choice mechanisms in a wild legume-rhizobium interaction. *Proc. R. Soc. B Biol. Sci.* **273**: 77–81.

Sprent, J.I. (2001) Nodulation in legumes Royal Botanic Gardens, Kew.

Stamatakis, A., Auch, A.F., Meier-Kolthoff, J., and Göker, M. (2007) AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa. *BMC Bioinformatics* **8**: 405.

Stępkowski, T., Czaplińska, M., Miedzinska, K., and Moulin, L. (2003) The variable part of the dnaK gene as an alternative marker for phylogenetic studies of rhizobia and related alpha proteobacteria. *Syst. Appl. Microbiol.* **26**: 483–494.

Stepkowski, T., Moulin, L., Krzyzanska, A., McInnes, A., Law, I.J., and Howieson, J. (2005) European origin of *Bradyrhizobium* populations infecting lupins and serradella in soils of Western Australia and South Africa. *Appl. Environ. Microbiol.* **71**: 7041–7052.

Sugawara, M., Epstein, B., Badgley, B.D., Unno, T., Xu, L., Reese, J., et al. (2013) Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. *Genome Biol.* **14**: R17.

Sullivan, J.T., Patrick, H.N., Lowther, W.L., Scott, D.B., and Ronson, C.W. (1995) Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. *Proc. Natl. Acad. Sci.* **92**: 8985–8989.

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

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.

Tang, J., Bromfield, E.S.P., Rodrigue, N., Cloutier, S., and Tambong, J.T. (2012) Microevolution of symbiotic *Bradyrhizobium* populations associated with soybeans in east North America. *Ecol. Evol.* **2**: 2943–2961.

Triplett, E.W. and Sadowsky, M.J. (1992) Genetics of competition for nodulation of legumes. *Annu. Rev. Microbiol.* **46**: 399–422.

Uchiumi, T., Ohwada, T., Itakura, M., Mitsui, H., Nukui, N., Dawadi, P., et al. (2004) Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. *J. Bacteriol.* **186**: 2439–2448.

Vinuesa, P., Rademaker, J.L.W., de Bruijn, F.J., and Werner, D. (1998) Genotypic characterization of *Bradyrhizobium* strains nodulating endemic woody legumes of the Canary Islands by PCR-restriction fragment length polymorphism analysis of genes encoding 16S rRNA (16S rDNA) and 16S-23S rDNA intergenic spacers, repetitive extragenic palindromic PCR genomic fingerprinting, and partial 16S rDNA sequencing. *Appl. Environ. Microbiol.* **64**: 2096–2104.

Vinuesa, P., Rojas-Jimenez, K., Contreras-Moreira, B., Mahna, S.K., Prasad, B.N., Moe, H., et al. (2008) Multilocus sequence analysis for assessment of the biogeography and evolutionary genetics of four *Bradyrhizobium* species that nodulate soybeans on the Asiatic continent. *Appl. Environ. Microbiol.* **74**: 6987–6996.

Vinuesa, P., Silva, C., Werner, D., and Martínez-Romero, E. (2005) Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation. *Mol. Phylogenet. Evol.* **34**: 29–54.

Weir, B.S. and Cockerham, C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**: 1358.

Young, J.P.W., Crossman, L.C., Johnston, A.W.B., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., et al. (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **7**: R34.

Table 2.1 Table of dominant and epidemic haplotypes for chromosome, symbiosis island and the whole genome. Dominant haplotypes are assigned within sites, must be collected 5 times, and constitute at least 10% of the isolates. Epidemic haplotypes must be dominant at a site and collected at another site that is >10km away as indicated by spatial spread.

		# isolates	Chromosome			Symbiosis island			Whole Genome		
			Haplotype	Abundance	Spatial Spread	Haplotype	Abundance	Spatial Spread	Haplotype	Abundance	Spatial Spread
Dominant Haplotypes	**Anza Borrego State Park - Palm Canyon**	**9**									
	Anza Borrego State Park - Roadside	**36**	K14_G58_X_R27	17	0	D46_N26_Z15_L24	7	0	K14_G58_X_R34_D46_N26_Z15_L24	5	0
			K14_G58_X_R34	7	0						
	Bodega Marine Reserve	**108**	K04_G05_I07_R02	19	0	D09_N01_Z01_L07	13	0	K04_G05_I07_R02_D09_N01_Z01_L07	13	0
			K02_G03_I01_R01	18	0	D07_N14_Z01_L10	12	0	K02_G03_I01_R01_D07_N14_Z01_L10	12	0
			K01_G03_I01_R01	14	723.8						
			K01_G01_I01_R01	13	0						
			K01_G03_I05_R01	12	0						
	Burns Piñon Ridge Reserve	**15**	K01_G03_I20_R01	6	120	D40_N01_Z02_L07	5	0			
	Motte Rimrock Reserve	**25**	K01_G03_I01_R01	9	723.8						
	Robert J. Bernard Biological Field Station	**68**	K01_G03_I01_R01	18	723.8	D29_N30_Z01_L04	9	0	K01_G03_I01_R01_D29_N30_Z01_L04	7	0
	San Dimas Canyon	**67**	K08_G36_I27_R35	9	0	D34_N33_Z12_L28	9	0	K11_G71_I21_R39_D37_N20_Z24_L16	7	0
			K01_G03_I01_R01	7	723.8	D37_N20_Z24_L16	7	0			
			K11_G71_I21_R39	7	0						
	San Dimas Reservoir	**13**									
	University of California - Riverside Hills	**17**	K01_G03_I01_R01	11	723.8						
Epidemic Haplotypes			K01_G03_I01_R01	61	723.8						
			K01_G03_I20_R01	10	120						

Figure 2. 1 Reconstructed phylogram of CHR haplotypes. Tree reconstructed in PhyML using concatenated *dnak*, *glnII*, ITS, and *recA* loci. Zig zags indicate long branches shortened for visibility. Red circles and blue circles on perimeter indicate the relative abundance and spatial spread of each CHR haplotype respectively.

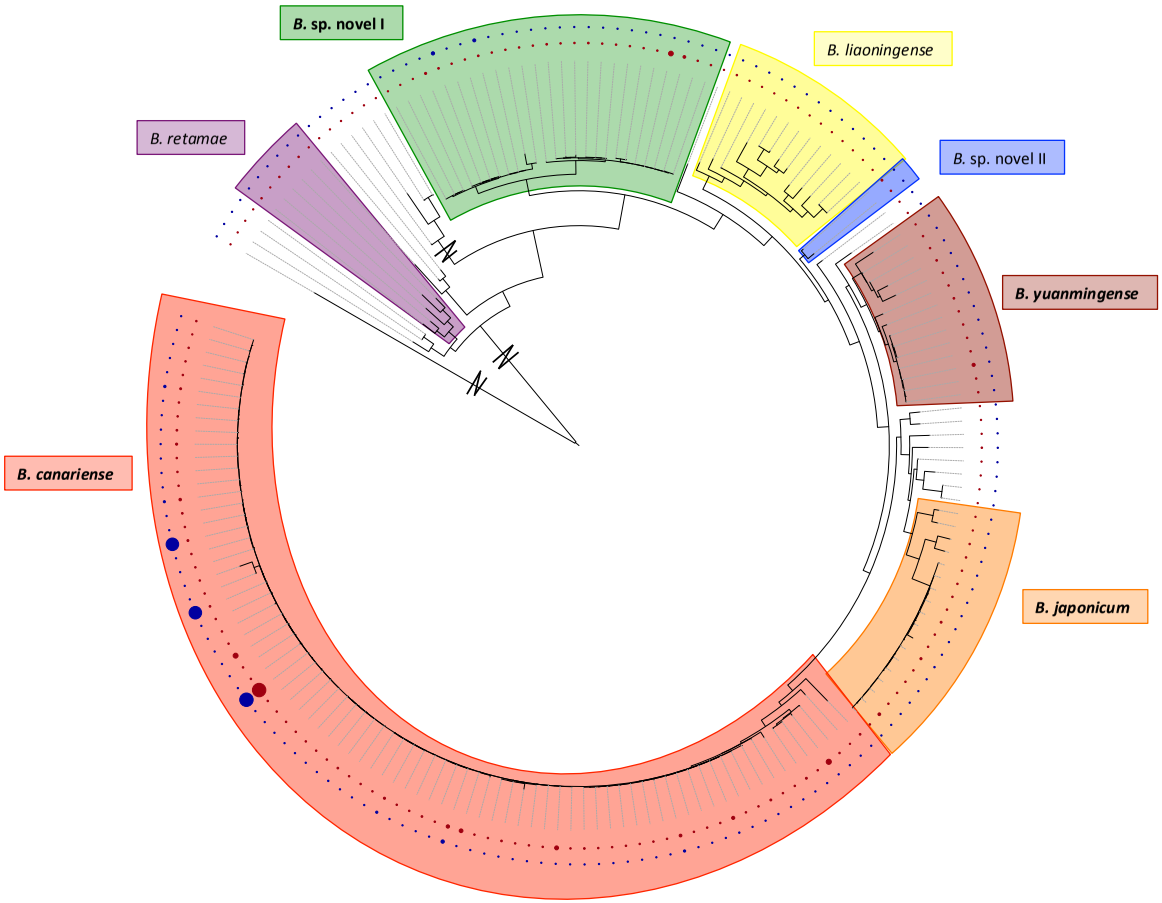


Figure 2. 2 Map of dominant haplotypes. Location of nine collection sites indicated by black dots. Proportional breakdown of dominant haplotypes at each location indicated by piecharts. Potential distribution of epidemic haplotypes indicated by colored dotted outlines.

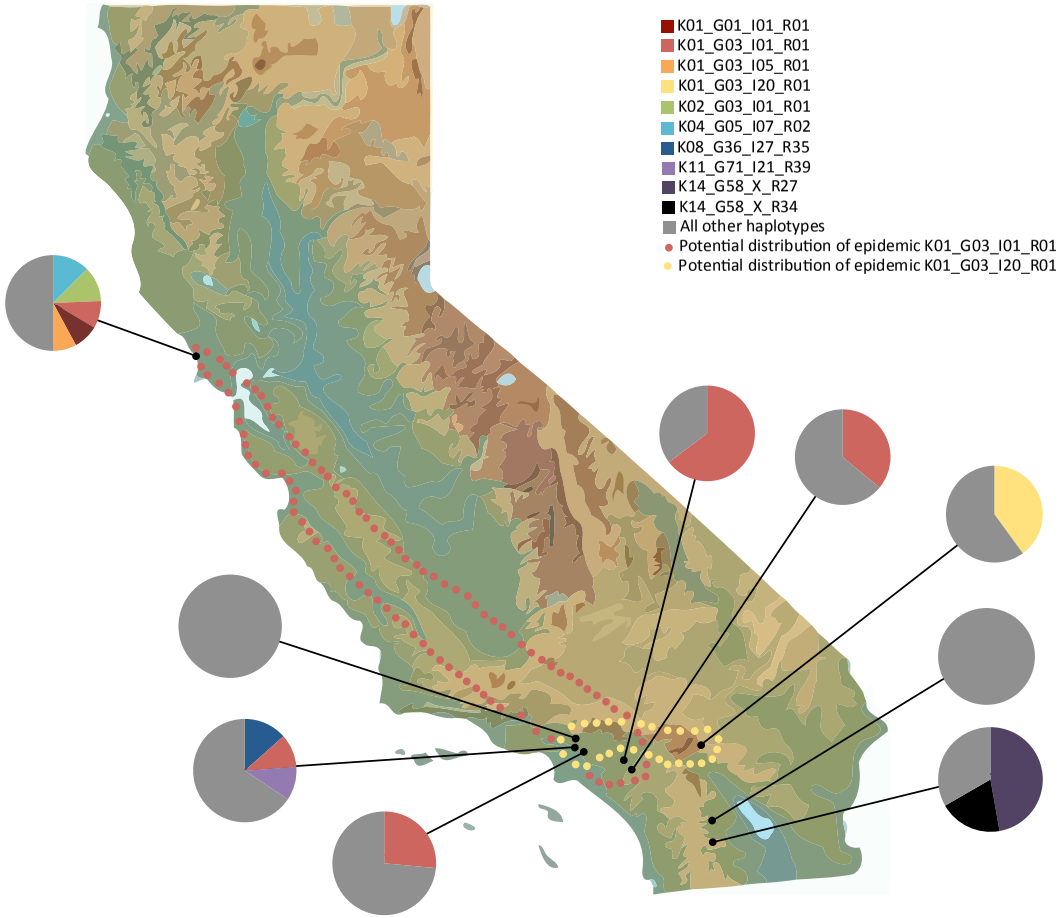
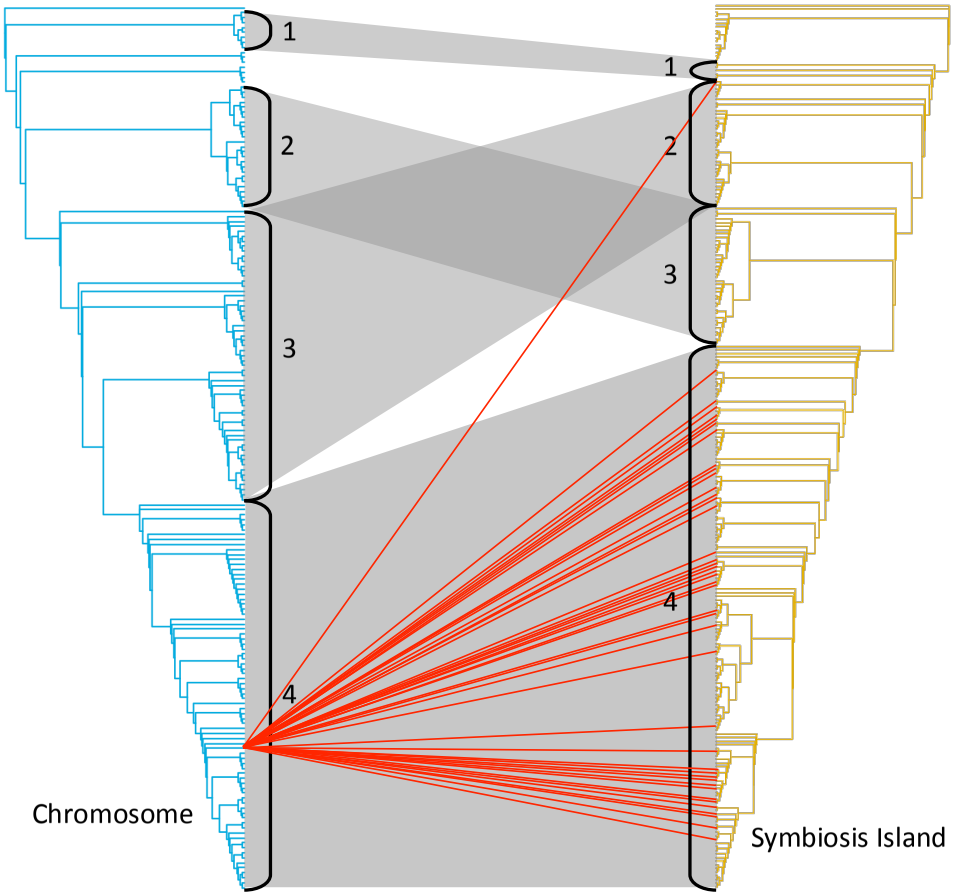


Figure 2. 3 Tanglegram with epidemic haplotype SI associations. TreeMap cophylogeny tanglegram of CHR and SI haplotypes. Blue cladogram reconstructed with PhyML from CHR loci (*dnak*, *glnII*, ITS, and *recA*). Yellow cladogram reconstructed with PhyML from SI loci (*nifD*, *nodD-A*, *nodZ*, and *nolL*). Overall patterns in associations between CHR and SI haplotypes indicated by grey bars (Individual associations between haplotypes viewable in Supplementary Figure S4). Clades of associated CHR and SI haplotypes bracketed and numbered. Red lines connect the 42 symbiosis island haplotypes associated with the epidemic haplotype K01_G03_I01_R01.



CHAPTER 3

Native California soils are selective reservoirs for multidrug resistant bacteria

Abstract

Soil bacteria can exhibit extensive antibiotic resistomes and act as reservoirs of important antibiotic resistance traits. However, the geographic sources and evolutionary drivers of resistance traits are poorly understood in these natural settings. We investigated the prevalence, spatial structure, and evolutionary drivers of multidrug resistance in natural populations of *Bradyrhizobium*, a cosmopolitan bacterial lineage that thrives in soil and aquatic systems as well as in plant and human hosts. We genotyped > 400 isolates from plant roots and soils across California and assayed 98 of them for resistance traits against 17 clinically relevant antibiotics. We investigated the geographic and phylogenetic structure of resistance traits, and analyzed correlations of resistance with strain abundance, host infection capacity, and *in vitro* fitness. We found: (i) multidrug resistance at all sites, (ii) subsets of resistance traits that are spatially structured, and (iii) significant associations between resistance traits and increased strain abundance or host infection capacity. Our results highlight multiple selective factors that can result in the spread of resistance traits in native *Bradyrhizobium* populations.

Introduction

Antibiotic resistance is ancient in bacterial populations, and the origins of most resistance traits predate human selective pressures (D'Costa et al., 2011). In natural settings, resistance traits are thought to be important in modulating inter-microbial communication and conflict (Martinez, 2008) irrespective of any anthropogenic effects (Singer et al., 2006). Native bacterial communities can exhibit extensive 'natural resistomes' (D'Costa, 2006; Nesme et al., 2014), but little is known about the abundance, spatial distribution, or selective drivers of multidrug resistance traits. Here, we investigated the population genetics and evolution of antibiotic resistance in *Bradyrhizobium* (alphaproteobacteria). *Bradyrhizobium* are cosmopolitan bacteria that thrive in soils and aquatic environments (Chaintreuil et al., 2000), and are found associated with plants (Vinuesa et al., 2005, 2008; Sachs et al., 2009) and animals, including humans (Hunt et al., 2011; Bhatt et al., 2013; Costello et al., 2013). *Bradyrhizobium* vary in antibiotic resistance (Mueller et al., 1988) and can be multidrug resistant (Cole and Elkan, 1979). Multiple putative resistance proteins are annotated within *Bradyrhizobium* genomes, but without predicted specificity to any antibiotic in most cases (http://genome.microbedb.jp/rhizobase/Bradyrhizobium). Moreover, *Bradyrhizobium* genomes contain low proportions of "core" proteins (i.e., proteins are that are universally shared among strains 23–33%), so strains can vary a great deal in gene content and are often sparsely annotated (Tian et al., 2012). We genotyped *Bradyrhizobium* isolates cultured from root nodules and the soil-root interface of native California *Lotus* species (herbaceous legumes) and assessed resistance to seventeen antibiotics. We investigated the phylogenetic, population genetic, and

geographic structure of resistance traits, as well as the correlation of resistance with strain abundance, nodulation capacity, in vitro fitness, and geographic distribution.

Materials and Methods

Isolation of *Bradyrhizobium*

Bradyrhizobium isolates were cultured from the root nodules and the soil-root interface of *Lotus strigosus* following published protocols (Sachs et al., 2009) (Supplementary Appendix S3). We included previously cultured isolates from *L. strigosus*, *L. micranthus*, *L. heermannii*, and *L. wrangelianus* (Sachs et al., 2009) (Supplementary Table S3.1). Most isolates (301/417) were collected from preserves with negligible human impact [i.e., gated, limited public access, lack of organic wastewater contaminant sources, etc. (Kolpin et al., 2002)], including University of California Reserves (Bodega Marine Reserve, Motte Rimrock Reserve, and Burns Piñon Ridge Reserve) and a biological field station (Robert J. Bernard Biological Field Station). We sampled 37 isolates from three sites with minimal human impact, including an undeveloped natural site in the hills at the University of California Riverside, and two state parks (Anza Borrego Desert State Park and Sonoma Coast State Park). Finally, we sampled 79 isolates from a canyon with limited human access, but close to human development (San Dimas Canyon).

PCR and Sequencing

Genomic DNA was isolated using the Qiagen DNeasy Blood and Tissue kit (Qiagen, Valencia, CA). For each isolate we amplified three loci encoded on the bacterial chromosome, including the ITS (internal transcribed spacer between the 16S and 23S rDNA) (van Berkum and Fuhrmann, 2000; Sachs et al., 2009), and partial sequences of *glnII* (glutamine synthetase) (Vinuesa et al., 2005, 2005; Sachs et al., 2011) and *recA*

(recombinase-A) (Vinuesa et al., 2005; Sachs et al., 2011). We also performed PCR for *nifD* (nitrogen fixation gene D) (Parker, 2000) and *nodD-A* (noncoding region, hypothetical protein, and partial sequence of nodulation gene A) (Sachs *et al*., 2010), both encoded on an integrated genomic island 'symbiosis island' that is present in strains with the capacity to nodulate hosts (Sullivan and Ronson, 1998; Sachs et al., 2010). PCR amplicons were sequenced at the Institute for Integrative Genome Biology at UC Riverside. Only sequences with unambiguous peaks were used for analysis. Sequences were aligned using ClustalW (Thompson et al., 1994) and indels were coded as a fifth character. Sequences were separately concatenated by genome region (chromosome: *glnII*, ITS, *recA;* symbiosis island: *nifD* and *nodD-A*) and clonal groups (haplotypes) were designated for each using the 'find redundant' function in MacClade (Maddison and Maddison, 2005). Concatenated chromosome haplotypes were assigned abundance scores based on the number of isolates recovered with identical sequences (Supplementary Table S3.1).

Antibiotic resistance assay

We chose a representative subset of 98/417 focal isolates to test for antibiotic resistance and other traits. We initially studied isolates from the Bodega Marine Reserve (Sachs et al., 2009), then expanded this study to multiple sites attempting to sample diverse soils and to maximize genetic diversity among collection sites (Supplementary Table S3.1). Seventeen different antibiotics were tested, including ampicillin (Acros Organics, Morris Plains, NJ); carbenicillin, ciprofloxacin, erythromycin, gentamicin, hygromycin B,

kanamycin, neomycin, penicillin, rifampicin (Fisher Scientific, Pittsburgh, PA); chloramphenicol, spectinomycin, streptomycin (MP Biomedical, Santa Ana, CA); doxycycline, novobiocin, paromycin, and tetracycline (Sigma-Aldrich, St. Louis, MO). These antibiotics can be grouped into 7 defined 'classes', including aminocourmains, aminoglycosides, amphenicols, fluoroquinolones, macrolides, penicillins, rifamycins, and tetracyclines. Two researchers independently assayed growth of the focal isolates on plates with modified arabinose-gluconate (MAG) (Sachs et al., 2009) containing one of seventeen antibiotics at four different concentrations (25, 50, 100 and 200 μg/ mL). Different concentrations of a given antibiotic were made with the same batch of media to minimize variation. Sterile toothpicks were used to create streaks on antibiotic plates, followed by control plates, in randomly assigned locations (Haahr, 1998).

Antibiotic and control plates were incubated at 29ºC, photographed and scored on the fifth day following plating to optimize growth of the isolates while minimizing risk of antibiotic degradation. Streaks on antibiotics were compared to the respective control streaks and binary assignments of resistance were recorded (resistance = growth on the plate; sensitivity = no growth). In cases of score discrepancies between independent researchers, the lower inhibitory antibiotic threshold was chosen. Five levels of antibiotic resistance were possible ranging from sensitivity to all concentrations to resistance to all concentrations.

Nodulation capacity assays

Nodulation capacity was assessed using two different methods. Some isolates were previously inoculated on *Lotus strigosus* (a permissive host with a broad symbiont range; (Sachs et al., 2009)) to confirm nodulation capacity. For the remainder, nodulation capacity was assessed via PCR amplification of the symbiosis loci *nifD* and *nodD-A*. Amplification of both of these loci is consistent with a root-nodulating strain, whereas soil isolates that lack amplifiable DNA in both these symbiosis-island loci consistently fail to form nodules on *L. strigosus* (Sachs et al., 2010, 2011).

Phylogenetic tree reconstruction

For each isolate we amplified five loci. Evolutionary relationships were reconstructed among the focal strains using three chromosomal loci (glnII, ITS, *recA* (van Berkum and Fuhrmann, 2000; Vinuesa et al., 2005; Sachs et al., 2011); and separately using two loci on a genomic island that encodes symbiosis traits (*nifD* and *nodD-A*) (Sachs et al., 2010), because this 'symbiosis island' can be horizontally transferred among *Bradyrhizobium* lineages (Sullivan and Ronson, 1998). Phylogenetic trees were reconstructed using maximum likelihood with RAxML (GUI1.3) (Silvestro and Michalak, 2012) with the following parameters: maximum likelihood, 1000 rapid bootstrap replicates, per-partition branchlengths, and a GTR + Γ nucleotide substitution model using partitioned datasets (Supplementary Appendix S3).

Data independence tests

Multilocus sequence typing of the 98 focal isolates uncovered several isolates with identical sequences (haplotypes). To test whether these isolates were independent for resistance traits, we compared three potential sources of variation in antibiotic resistance score (experimenter gathering the data, different isolates with the same haplotype, and different isolates with different haplotypes). Resistance to antibiotics of the same class can also be non-independent due to structural similarity. To assess independence, we compared two sources of variation (pairs of antibiotics of the same class versus different classes). Variation was calculated as the sum of individual absolute differences in the five level antibiotic resistance scores. All possible combinations, without replacement, were chosen using a random number generator (Haahr, 1998). Data were analyzed using the 'Tests' function of 'fit y by x' in JMP (SAS Institute Inc, 1989), analogous to an ANOVA (but using ordinal data).

Population genetic analyses

We analyzed molecular diversity indices and interpopulation variation (FST) using Arlequin (Excoffier and Lischer, 2010) for concatenated *glnII*, ITS, and *recA* sequences (2,686 bp).

Structure in antibiotic resistance patterns

Phylogenetic signal— We tested antibiotic resistance traits, colony growth rates, and nodulation capacity for phylogenetic signal by quantifying Pagel's lambda (λ; (Pagel,

1999a) and Fritz and Purvis' D (Fritz and Purvis, 2010). These tests are key prerequisites for the evolutionary analyses presented below. Among closely related taxa, traits might be correlated due to shared ancestry and thus would not be independent in statistical analyses.

Geographic sites - Resistance traits might be selectively favored in soils depending on local soil microbiota or other factors that shape resistance traits (Singer et al., 2006). Associations between the five level antibiotic resistance score and seven field collection sites were tested using chi-squared tests (Likelihood ratio and Pearson's) in JMP (SAS Institute Inc, 1989).

Nodulation capacity and haplotype abundance - Resistance traits might be advantageous for bacterial genotypes that compete in the rhizosphere to infect plant hosts. Root-nodule forming genotypes of *Bradyrhizobium* can form intra-cellular infections on legumes, whereas many *Bradyrhizobium* genotypes do not have this capacity (and persist in the soil or the soil-root interface). We issued binary abundance assignments (haplotypes with ≥ five isolates and constituting > 1% of the sampled population were considered abundant and the remainder of haplotypes were considered 'rare'). For antibiotic resistance traits exhibiting no phylogenetic signal, associations between host association/abundance and antibiotic resistance traits (100 μg/mL cutoff) were tested with chi-squared tests (likelihood ratio and Pearson's) in JMP (SAS Institute Inc, 1989).

For antibiotics exhibiting phylogenetic signal we used the likelihood ratio test with the ML option of the discrete function in BayesTraits (Pagel, 1999b).

Colony growth rate - Resistance traits might also be advantageous for competition and spread in the environment. Soil bacteria often exhibit a highly skewed distribution with many rare haplotypes and a handful of highly abundant ones (Mcinnes, 2004). To estimate bacterial fitness in multiple types of media, a phylogenetically diverse subset of 13 strains was chosen for colony growth rate assays on three media types including a glucose-based rhizobium-defined medium (GRDM) (Sullivan et al., 1996), yeast mannitol medium (YM) (Somasegaran et al., 1994), and MAG (Sachs et al., 2009). Isolates were cultured on six replica plates (~50-150 CFU/plate), were incubated at 29ºC, and colony size was recorded at 360 hours (MAG, YM) or 480 hours (GRDM) of incubation using photographs on gridded paper in ImageJ (Schneider et al., 2012). Plates were analyzed if there were at least 10 distinct colonies, at least 2 mm from the edge of the plate or other colonies.

We examined correlations between resistance scores at the 100 μg/mL cutoff and colony growth rate as measured by average colony size on each of three different media (see above) using ANOVA in JMP (SAS Institute Inc, 1989). Antibiotic resistance traits that exhibited significant phylogenetic signal were tested with phylogenetic independent contrasts using the crunch command in the Caper package (Orme, 2012) of R (R Core Team, 2013).

Results

Antibiotic resistance assay and phylogenetic tree reconstruction

Two antibiotics did not exhibit variation in resistance. All isolates were sensitive to novobiocin and were resistant to erythromycin (Supplementary Table S3.2). At the conservative cutoff of 100 μg/mL to assign resistance, we failed to find resistance traits to only four antibiotics (ciprofloxacin, novobiocin, spectinomycin, and tetracycline) (Figure 3.1). Moreover, we found multidrug resistant *Bradyrhizobium* to be ubiquitous in native California soils, even in pristine settings (Supplementary Table S3.2). At the conservative 100 μg/mL cutoff for resistance, we found isolates to be resistant to an average of 4 antibiotics and uncovered isolates that were resistant to up to 10 antibiotics (Figure 3.2). The most common 4x resistance trait combination (chloramphenicol, erythromycin, neomycin, and paromycin) was uncovered in 21 isolates from all collection sites (Supplementary Table S3.2). Most isolates (76/82) were resistant to at least two unrelated classes of antibiotics.

Population genetic analyses

Gene diversity (Nei, 1987) was high within all eight populations (all > 0.8333). (Supplementary Table S3.3). The Anza Borrego State Park site was greatly differentiated from all other populations, and the other between population FST values were often not significant (17/28; Supplementary Tables S3.4, S3.5).

Data independence tests

Isolates with identical haplotypes were non-independent for antibiotic resistance, exhibiting similar variation in resistance scores to identical isolates tested by different experimenters (Pearson's $X^2 = 16.109$, p = 0.1371, $df = 11$), and significantly less variation than isolates with different haplotypes (Pearson's $X^2 = 38.842$, p =0.0019, $df = 17$) (Supplementary Table S3.6). Thus, a single isolate per each haplotype was selected for further analyses using a random number generator (Haahr, 1998), resulting in a total of 82 unique haplotypes.

Variation in resistance scores within antibiotic classes was not significantly different from variation in resistance among antibiotic classes (Pearson's $X^2 = 292.389$, p = 0.5156, $df = 294$) (Supplementary Table S3.6). Thus, all antibiotics were treated as independent.

Structure in antibiotic resistance patterns

Phylogenetic signal - Resistance to hygromycin B, neomycin, and chloramphenicol exhibited statistically significant phylogenetic signal for both Pagel's lambda and Fritz Purvis' D. Kanamycin, paromycin, and streptomycin resistance exhibited statistically significant phylogenetic signal for Pagel's lambda only and gentamicin exhibited phylogenetic signal for Fritz Purvis' D only. Symbiotic capacity exhibited statistically significant phylogenetic signal for both Pagel's lambda and Fritz Purvis' D. Colony growth rate on MAG and YM media exhibited statistically significant phylogenetic signal for Pagel's lambda (Supplementary Table S3.7).

Geographic sites – Quantitative antibiotic resistance traits exhibited non-random spatial structure among geographic sites for six resistance traits, using both likelihood ratio and Pearson's chi-squared tests (Pearson's; ciprofloxacin, $X^2 = 59.274$, $p < 0.0001$, $df = 12$; kanamycin, $X^2 = 37.177$, $p = 0.0002$, $df = 12$; neomycin, $X^2 = 61.148$, $p < 0.0001$, $df = 18$; penicillin, $X^2 = 37.762$, $p = 0.0367$, $df = 24$; spectinomycin, $X^2 = 36.449$, $p = 0.0003$, $df = 12$; tetracycline, $X^2 = 30.479$, $p = 0.0024$, $df = 12$). Resistance to carbenicillin was statistically significant for the likelihood ratio chi-squared test only ($X^2 = 40.556$, $p = 0.0186$, $df = 24$) (Supplementary Table S3.8). Among resistance traits with spatial structure, ciprofloxacin, kanamycin, neomycin, and spectinomycin resistance were mostly localized at inland desert populations (Anza Borrego Desert State Park, Burns Piñon Ridge Reserve) but also at Bodega Marine Reserve and San Dimas Canyon (Supplementary Figure S3.1A-D). Resistance to higher concentrations of tetracycline was localized to inland Southern California (Supplementary Figure S3.1E) and a pattern of higher resistance in the western populations was observed for carbenicillin (Supplementary Figure S3.1F).

Nodulation capacity – We uncovered support for the dependent evolution of hygromycin B and neomycin resistance with root-nodulation capacity (hygromycin, $D = 13.26533$, $p = 0.01$, $df = 4$; neomycin, $D = 20.317124$, $p = 0.0004$, $df = 4$) (Supplementary Table S3.9). Additionally, evolutionary transitions rates among these states (nodulation

capacity, hygromycin B resistance) support a correlation between nodulation capacity and antibiotic resistance (Supplementary Figure S3.2A,B).

Haplotype abundance – Ampicillin resistance was associated with increased abundance (ampicillin, $X^2 = 6.005$, p-value = 0.0143, $df = 1$) (Supplementary Table S3.10) and likelihood ratio tests significantly support the dependent evolution of abundance and neomycin resistance traits ($D = 10.88488$, p = 0.0279, $df = 4$). Similarly, most transitions rates among these states (high vs. low abundance, neomycin resistance) support a correlation between increased abundance and neomycin resistance (Supplementary Figure S3.2C).

Colony growth rate – Colony growth rates on three media types uncovered no consistent evidence for costs to antibiotic resistance. On GRDM, resistance to hygromycin B was correlated with smaller colony size (F = 54.120, p = 0.000) while paromycin resistance was correlated with larger colony size (F = 34.700, p = 0.000). The results were converse on MAG (hygromycin B: F = 16.570, p = 0.001; paromycin: F = 52.420, p = 0.000) and YM (hygromycin B: F = 17.880, p = 0.008; paromycin: F = 7.474, p = 0.017). A further two antibiotics exhibited statistically significant correlations between antibiotic resistance and colony size on YM only (gentamicin: F = 9.669, p = 0.008; streptomycin: F = 6.969, p = 0.020) (Supplementary Table S3.11). Although some resistance traits exhibited significant associations with colony growth rate, no overall pattern was apparent.

Discussion

We found multidrug resistant *Bradyrhizobium* to be ubiquitous in native California soils, even in pristine settings with limited human contact. Using a conservative 100 μg/mL cutoff to assign resistance, we found isolates to be resistant to an average of 4 antibiotics and uncovered isolates that were resistant to up to 10 antibiotics (Figure 3.1). The most common 4x resistance trait combination (chloramphenicol, erythromycin, neomycin, and paromycin) was uncovered in 21 isolates from all collection sites (Supplementary Table S3.2). Most isolates (76/82) were resistant to at least two unrelated classes of antibiotics. Thus patterns of multidrug resistance were not driven by cross-resistance against related antibiotics. Our data support previous research that showed intrinsic antibiotic resistance in *Bradyrhizobium*, and demonstrates the ubiquity of multidrug resistance in wild *Bradyrhizobium* populations.

We found significant geographical structure of resistance traits, which is consistent with localized or regional selective pressures favoring the spread of resistance traits (e.g., carbenicillin, ciprofloxacin, kanamycin, neomycin, penicillin, spectinomycin, and tetracycline) (Supplementary Table S3.8, Supplementary Figure S3.1A-G). Although each sampled population was genetically diverse (Supplementary Table S3.3), some collection sites have small samples and it is possible that these patterns are driven by relatively few isolates. However, the high gene diversity found across populations (>0.8333) indicates that our sampling captured variation in the populations.

Selection upon antibiotic resistance traits has largely been studied in anthropogenic contexts. It is now clear that bacterial resistance traits can be found in native, protected

sites (D'Costa, 2006), but little is known about the selective drivers of resistance traits in natural populations. To understand the evolution of resistance, we examined correlated evolution of resistance with host association (nodulation capacity), abundance, and colony growth rate. We found statistically significant associations between antibiotic resistance and nodulation capacity (hygromycin B, neomycin) (Supplementary Table S3.9 and Supplementary Figure S3.2A and B) as well as strain abundance (ampicillin, neomycin) (Supplementary Table S3.10 and Supplementary Figure 3.2C). In particular, resistance to neomycin is correlated with the recent spread of an abundant haplotype found across California, suggesting that this resistance trait might be a key fitness driver in natural bacterial populations. Lastly, although some resistance traits exhibited significant associations with colony growth rate (Supplementary Table S3.11), no overall pattern was apparent.

We inferred that antibiotic resistance traits vary substantially in phylogenetic signal, which suggests that these resistance traits evolve via multiple processes. Even low levels of horizontal gene transfer can eliminate statistical dependence upon the phylogenetic tree, as occurs in cophylogenetic analyses of horizontally transferred symbionts and their hosts (Schilthuizen and Stouthamer, 1997). Thus, traits with significant phylogenetic signal (resistance to chloramphenicol, gentamicin, hygromycin B, kanamycin, neomycin, paromycin, streptomycin; largely aminoglycosides) (Supplementary Table S3.7) are most likely faithfully passed from parent to offspring. In contrast, traits without significant phylogenetic signal (resistance to ampicillin, carbenicillin, doxycycline, penicillin, and rifampicin) might be rapidly evolving, swiftly being lost due to local selection, or

evolving via horizontal gene transfer. However, multiple evolutionary processes can be correlated with variation in phylogenetic signal, so other sources of data are needed to corroborate these inferences (Revell et al., 2008).

References

Van Berkum, P. and Fuhrmann, J.J. (2000) Evolutionary relationships among the soybean bradyrhizobia reconstructed from 16S rRNA gene and internally transcribed spacer region sequence divergence. *Int. J. Syst. Evol. Microbiol.* **50**: 2165–2172.

Bhatt, A.S., Freeman, S.S., Herrera, A.F., Pedamallu, C.S., Gevers, D., Duke, F., et al. (2013) Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N. Engl. J. Med.* **369**: 517–528.

Chaintreuil, C., Giraud, E., Prin, Y., Lorquin, J., Ba, A., Gillis, M., et al. (2000) Photosynthetic bradyrhizobia are natural endophytes of the African wild rice *Oryza breviligulata*. *Appl. Environ. Microbiol.* **66**: 5437–5447.

Cole, M.A. and Elkan, G.H. (1979) Multiple antibiotic resistance in *Rhizobium japonicum*. *Appl. Environ. Microbiol.* **37**: 867–870.

Costello, E.K., Carlisle, E.M., Bik, E.M., Morowitz, M.J., and Relman, D.A. (2013) Microbiome assembly across multiple body sites in low-birthweight infants. *mBio* **4**: e00782–13–e00782–13.

D'Costa, V.M. (2006) Sampling the antibiotic resistome. *Science* **311**: 374–377.

D'Costa, V.M., King, C.E., Kalan, L., Morar, M., Sung, W.W.L., Schwarz, C., et al. (2011) Antibiotic resistance is ancient. *Nature* **477**: 457–461.

Excoffier, L. and Lischer, H.E.L. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**: 564–567.

Fritz, S.A. and Purvis, A. (2010) Selectivity in mammalian extinction risk and threat Types: A new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* **24**: 1042–1051.

Haahr, M. (1998) RANDOM.ORG - True Random Number Service. *random.org*.

Hunt, K.M., Foster, J.A., Forney, L.J., Schütte, U.M.E., Beck, D.L., Abdo, Z., et al. (2011) Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS ONE* **6**: e21313.

Kolpin, D.W., Furlong, E.T., Meyer, M.T., Thurman, E.M., Zaugg, S.D., Barber, L.B., and Buxton, H.T. (2002) Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. Streams, 1999–2000: A national reconnaissance. *Environ. Sci. Technol.* **36**: 1202–1211.

Maddison, W.P. and Maddison, D.R. (2005) MacClade: Analysis of phylogeny and character evolution.

Martinez, J.L. (2008) Antibiotics and antibiotic resistance genes in natural environments. *Science* **321**: 365–367.

Mcinnes, A. (2004) Structure and diversity among rhizobial strains, populations and communities - A review. *Soil Biol. Biochem.* **36**: 1295–1308.

Mueller, J.G., Skipper, H.D., Shipe, E.R., Grimes, L.W., and Wagner, S.C. (1988) Intrinsic antibiotic resistance in *Bradyrhizobium japonicum*. *Soil Biol. Biochem.* **20**: 879–882.

Nei, M. (1987) Molecular evolutionary genetics Columbia University Press, New York.

Nesme, J., Cécillon, S., Delmont, T.O., Monier, J.-M., Vogel, T.M., and Simonet, P. (2014) Large-scale metagenomic-based study of antibiotic resistance in the environment. *Curr. Biol.* **24**: 1096–1100.

Orme, D. (2012) The caper package: Comparative analysis of phylogenetics and evolution in R.

Pagel, M. (1999a) Inferring the historical patterns of biological evolution. *Nature* **401**: 877–884.

Pagel, M. (1999b) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48**: 612–622.

Parker, M. (2000) Divergent *Bradyrhizobium* symbionts on *Tachigali versicolor* from Barro Colorado Island, Panama. *Syst Appl Microbiol* **23**: 585–90.

R Core Team (2013) R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria.

Revell, L., Harmon, L., and Collar, D. (2008) Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* **57**: 591–601.

Sachs, J.L., Ehinger, M.O., and Simms, E.L. (2010) Origins of cheating and loss of symbiosis in wild *Bradyrhizobium*. *J. Evol. Biol.* **23**: 1075–1089.

Sachs, J.L., Kembel, S.W., Lau, A.H., and Simms, E.L. (2009) In Situ phylogenetic structure and diversity of wild *Bradyrhizobium* communities. *Appl. Environ. Microbiol.* **75**: 4727–4735.

Sachs, J.L., Russell, J.E., and Hollowell, A.C. (2011) Evolutionary instability of symbiotic function in *Bradyrhizobium japonicum*. *PLoS ONE* **6**: e26370.

SAS Institute Inc (1989) JMP SAS Institute Inc., Cary, NC.

Schilthuizen, M.O. and Stouthamer, R. (1997) Horizontal transmission of parthenogenesis-inducing microbes in *Trichogramma* wasps. *Proc. R. Soc. B Biol. Sci.* **264**: 361–366.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Meth* **9**: 671–675.

Silvestro, D. and Michalak, I. (2012) raxmlGUI: A graphical front-end for RAxML. *Org. Divers. Evol.* **12**: 335–337.

Singer, R.S., Ward, M.P., and Maldonado, G. (2006) Can landscape ecology untangle the complexity of antibiotic resistance? *Nat Rev Micro* **4**: 943–952.

Somasegaran, P., Hoben, H.J., and Somasegaran, P. (1994) Handbook for rhizobia Springer-Verlag, New York; London.

Sullivan, J.T., Eardly, B.D., van Berkum, P., and Ronson, C.W. (1996) Four unnamed species of nonsymbiotic rhizobia isolated from the rhizosphere of *Lotus corniculatus*. *Appl. Environ. Microbiol.* **62**: 2818–2825.

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

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (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.

Tian, C.F., Zhou, Y.J., Zhang, Y.M., Li, Q.Q., Zhang, Y.Z., Li, D.F., et al. (2012) Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. *Proc. Natl. Acad. Sci.* **109**: 8629–8634.

Vinuesa, P., Rojas-Jimenez, K., Contreras-Moreira, B., Mahna, S.K., Prasad, B.N., Moe, H., et al. (2008) Multilocus sequence analysis for assessment of the biogeography and evolutionary genetics of four *Bradyrhizobium* species that nodulate soybeans on the Asiatic continent. *Appl. Environ. Microbiol.* **74**: 6987–6996.

Vinuesa, P., Silva, C., Werner, D., and Martínez-Romero, E. (2005) Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation. *Mol. Phylogenet. Evol.* **34**: 29–54.

Figure 3. 1. Population frequency of resistance traits among antibiotics. Resistant *Bradyrhizobium* isolates were uncovered for all but four antibiotics at 100ug/ml (spectinomycin, novobiocin, ciprofloxacin, tetracycline).

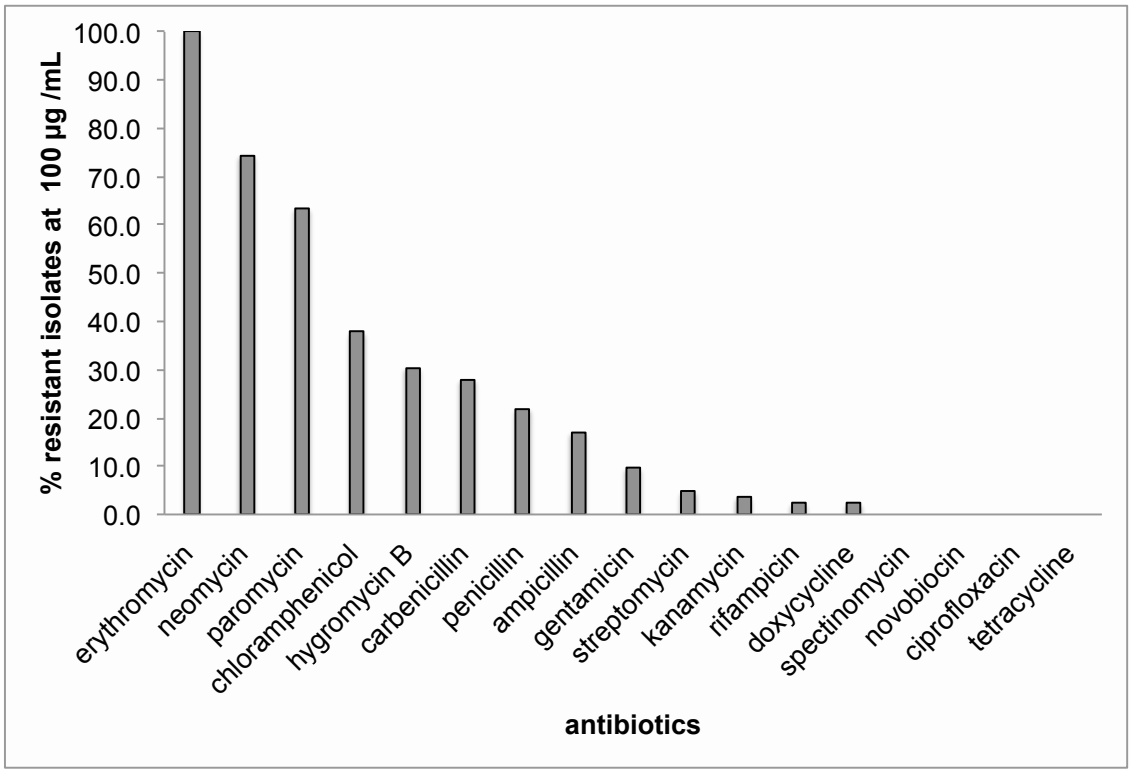


Figure 3. 2. Population frequency of multidrug antibiotic resistance: Gray bars indicate the number of isolates that are resistant to given number of antibiotics at 100 μg/mL.

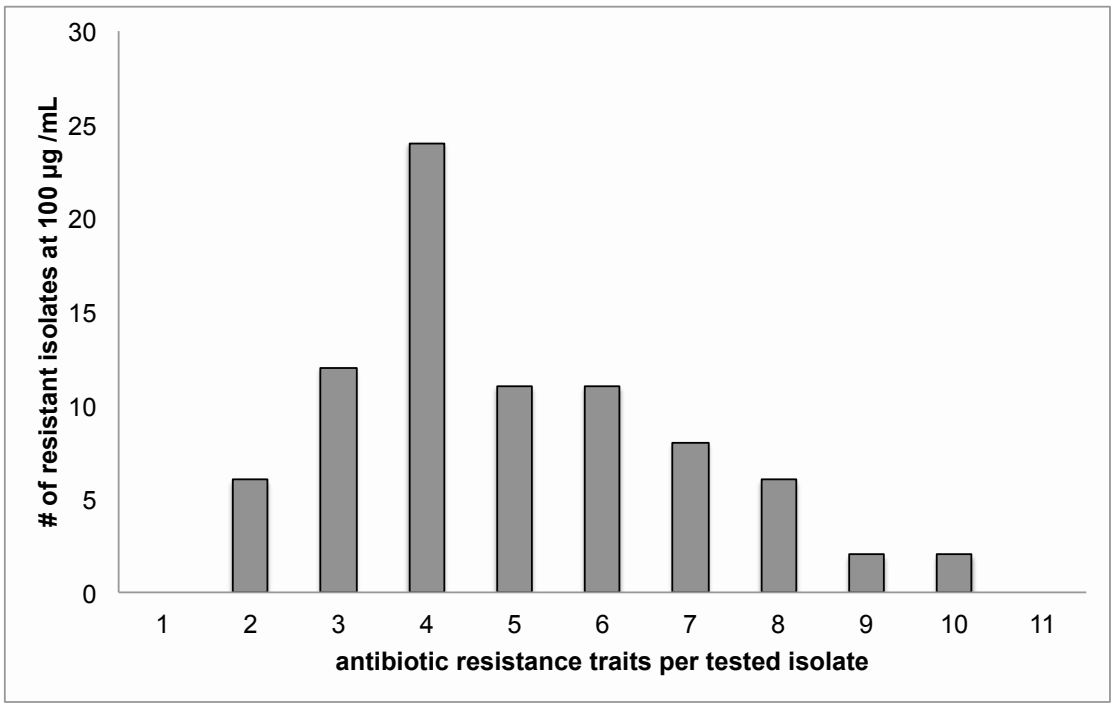
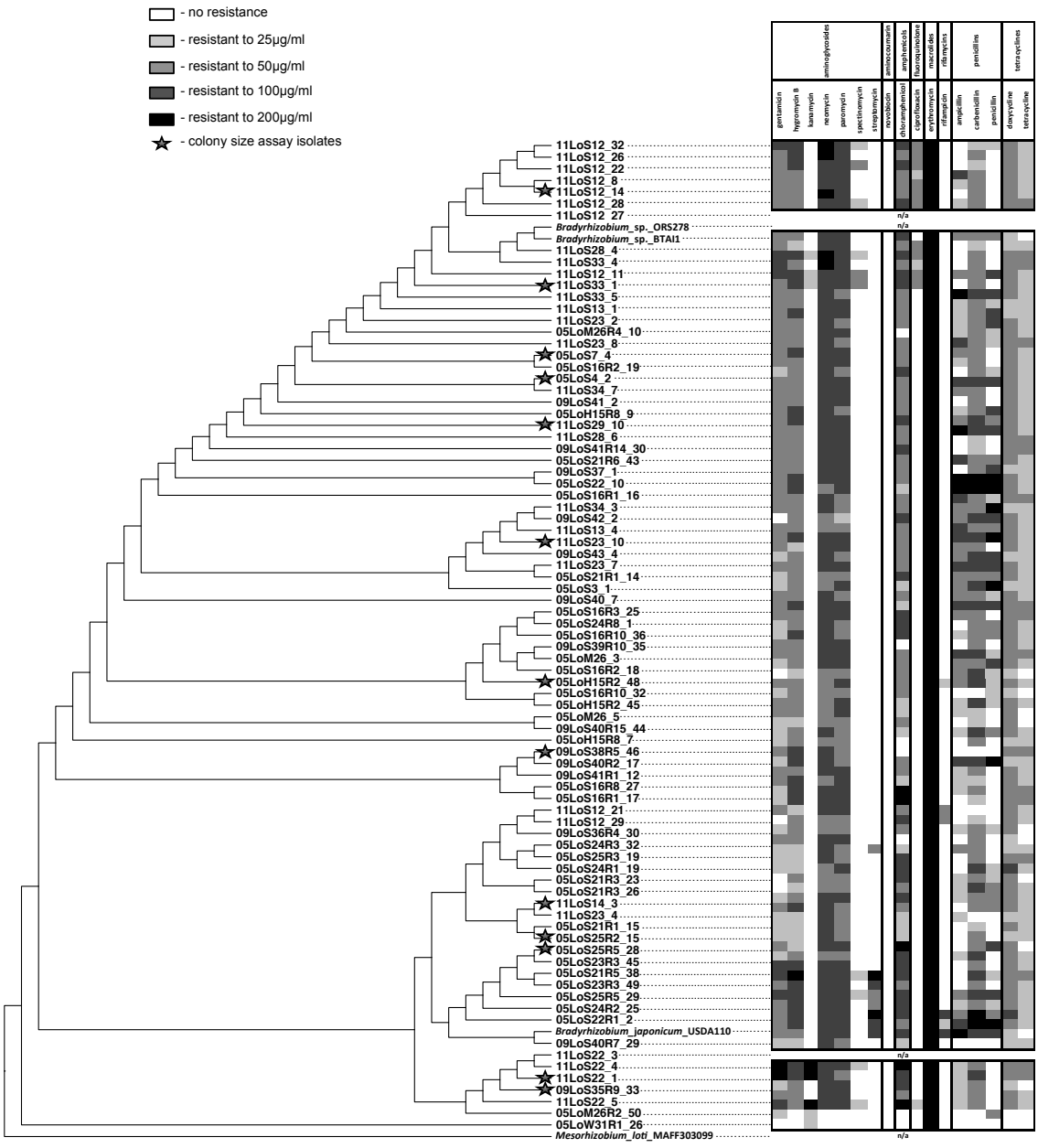


Figure 3. 3. Maximum likelihood cladogram reconstructed from concatenated *glnII*, ITS, and *recA* loci from 82 isolates with corresponding heat map of antibiotic resistance. White indicates sensitivity at all tested concentrations while black indicates resistance at all tested concentrations. Isolates chosen for colony size assessment on multiple solid media types are phylogenetically diverse as indicated by stars.



GENERAL CONCLUSION

Acquisition of genomic islands can confer bacteria with massive fitness benefits including the capacity to spread epidemically among sensitive hosts (Bach, 2000; Hacker and Carniel, 2001; Diep et al., 2006; Qiu et al., 2006). In the legume-rhizobia symbiosis, inoculation experiments suggest intense selection by the host plant, given that hosts can efficiently favor beneficial over ineffective rhizobial partners. These data would predict symbiotic capacity will be correlated with increased abundance and host plants may select for particular symbiosis island genotypes. However, the data herein shows the supreme importance of selection in the soil among rhizobial genotypes. In the first chapter of my dissertation I tested the paradigm that acquisition of a genomic island confers a fitness benefit. However, I found that non-symbiotic rhizobia -- that cannot infect plant hosts -- can spread among sites and achieve high abundance that is similar to symbiotic strains. In the second chapter of my dissertation I examined role of the plant and soil in structuring *Bradyrhizobium* populations by analyzing population genetic structure across the genome. I found that genome regions that encode soil traits are most important for structuring rhizobial populations, whereas symbiosis encoding loci show no such evidence. In the final chapter of my dissertation I tested the selection for antibiotic resistance in natural settings. I found that antibiotic resistance traits, that are thought to be key factors mediating strain competition, are favored differentially dependent on specific soil habitats. Thus the paradigm of host plants acting as a major selective force upon *Bradyrhizobium* populations is not supported by the results presented in this dissertation.

References

Bach, S. (2000) The *Yersinia* high-pathogenicity island is present in different members of the family Enterobacteriaceae. *FEMS Microbiol. Lett.* **183**: 289–294.

Diep, B.A., Gill, S.R., Chang, R.F., Phan, T.H., Chen, J.H., Davidson, M.G., et al. (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet* **367**: 731–739.

Hacker, J. and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO Rep.* **2**: 376–381.

Martinez, J.L. (2008) Antibiotics and antibiotic resistance genes in natural environments. *Science* **321**: 365–367.

Qiu, X., Gurkar, A.U., and Lory, S. (2006) Interstrain transfer of the large pathogenicity island (PAPI-1) of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci.* **103**: 19830–19835.