Phylogenetic clustering of Bradyrhizobium symbionts on legumes indigenous to North America

Jonathan H. Koppell and Matthew A. Parker

Department of Biological Sciences, State University of New York, Binghamton, NY 13902, USA

To analyse determinants of biogeographic structure in members of the genus Bradyrhizobium, isolates were obtained from 41 legume genera, originating from North American sites spanning 48.5° of latitude (Alaska to Panama). Sequencing of portions of six gene loci (3674 bp) in 203 isolates showed that there was only a weak trend towards higher nucleotide diversity in tropical regions. Phylogenetic relationships for nifD, in the symbiosis island region of the Bradyrhizobium chromosome, conflicted substantially with a tree inferred for five housekeeping gene loci. For both nifD and housekeeping gene trees, bacteria from each region were significantly more similar, on average, than would be expected if the source location was permuted at random on the tree. Within-region permutation tests also showed that bacteria clustered significantly on particular host plant clades at all levels in the phylogeny of legumes (from genus up to subfamily).

Nevertheless, some bacterial groups were dispersed across multiple regions and were associated with diverse legume host lineages. These results indicate that migration, horizontal gene transfer and host interactions have all influenced the geographical divergence of Bradyrhizobium populations on a continental scale.

In this study, we analysed relationships of Bradyrhizobium species nodule bacteria across regions of North America. Among the 13 genera of Proteobacteria that are legume nodule symbionts (Willems, 2006; Gyaneshwar et al., 2011; Ardley et al., 2011), Bradyrhizobium is prevalent on more legume taxa and in more habitats than any other rhizobial genus. We sampled Bradyrhizobium strains from diverse environments (Alaska to Panama), from 41 legume host genera representing 16 of the 36 recognized tribes in the Leguminosae family (Lewis et al., 2005). Bacteria were characterized by sequencing portions of five housekeeping gene loci and a sixth locus in the symbiosis island (SI) portion of the genome (nifD). In Bradyrhizobium, symbiosis-related genes are clustered in a portion of the chromosome termed the SI region, which has hallmarks of being a mobile genetic element (Kaneko et al., 2002). Markers in the SI region commonly have a different genealogical history from other genes (Parker et al., 2002; Moulin et al., 2004; Vinuesa et al., 2005; Stepkowski et al., 2007). Thus, analysis of relationships requires a separate study of both SI and non-SI loci.

Our main goal was to understand how bacteria are distributed across geographical regions and across clades of legume hosts. Phylogenetic patterns were initially assessed by a tree permutation method (Webb et al., 2008). This procedure analyses whether a set of bacterial strains sampled from a particular region or a set of strains sampled from a particular legume taxon tend to cluster as relatives on a phylogenetic tree. The mean patristic distance among pairs of strains within a group is calculated.

INTRODUCTION

It is obvious that legume biogeography must play some role in structuring the diversity of nodule symbionts, because bacteria can only thrive in regions where compatible hosts are present. However, the wide variation in host breadth among rhizobia has made this a challenging problem to analyse. Rhizobial groups show complex patterns of distribution across legume taxa (Vinuesa et al., 2005; Stepkowski et al., 2007, 2012; Nzoüe et al., 2009). For any given bacterial lineage, it is thus unclear whether the major limiting factor affecting geographical spread might involve regional differentiation of legume communities or non-symbiotic factors such as physical habitat variables and dispersal limitation.

The wide heterogeneity in bacterial host breadth has also made it difficult to identify the set of plant lineages that engage in symbiotic coevolution with particular rhizobia. It appears that few (if any) legume species harbour a symbiont assemblage that is both unique and monophyletic. Yet existing analyses have not resolved the extent to which larger host clades at any level (such as legume genera, tribes or other defined monophyletic groups of hosts) may associate with distinctive sets of symbionts.

**Abbreviations:** MPD, mean phylogenetic distance; SI, symbiosis island.

The GenBank accession numbers for the 203 North American Bradyrhizobium strains can be found in Table S3.

Two supplementary figures, six supplementary tables and supplementary material are available with the online version of this paper.
Significance of phylogenetic clustering for the group is then assessed by comparing its distance value to a null distribution generated by permuting strains at random across tips of the observed tree.

However, there is a basic problem in interpreting phylogenetic patterns detected in these analyses. Geographical origin and host legume clade are confounded in any set of samples from natural ecosystems, since few or no legume taxa are uniformly distributed across regions. In a null distribution generated by permuting across all tips on a tree, randomized data for each host-associated bacterial group will encompass a greater variety of locations than in the original sample. Thus, phylogenetic distances in the null distribution are inflated by regional differentiation arising from non-symbiotic factors such as genetic drift and limited migration. This may lead to an inference of host-related clustering that is not reliable.

An alternative permutation method was therefore developed where randomization was restricted to the set of bacterial strains within each region, yielding a more conservative test for host-related clustering of bacteria. This analysis showed that there are substantial disparities among clades of legumes in the extent of phylogenetic clustering of their associated *Bradyrhizobium* symbionts. A number of legume clades had symbiont assemblages with no detectable indication of phylogenetic clustering. Yet in other cases, there was evidence for distinctive host-associated clusters at a variety of levels ranging from legume genera to other larger monophyletic groups within the legume family.

**METHODS**

The sampling location and hosts for all 203 North American bacterial strains are provided in Tables S1 and S2 (available with the online version of this paper). In each of the eight regions (Fig. 1), 21–35 strains were analysed. A limited number of bacterial strains were also available from Alaska (three strains), California (two strains) and Illinois (one strain). These samples were not sufficient for region-based summaries of nucleotide polymorphism carried out for the eight main locations, but they were included in tree analyses to increase geographical coverage. For five of the regions [north-eastern USA, North Carolina, Chihuahua (Northern Mexico), Costa Rica, Panama], sampling details have been reported (Parker, 2002, 2003, 2004, 2008, 2012). These collections were supplemented by new samples of root nodules from areas of natural vegetation in Southern Texas, Washington State and Oaxaca, Mexico (Table S1). The primary sampling objective was to obtain symbionts from a maximum diversity of legume taxa native to each region. Due to dissimilarity among regions in the number of legume species that could be sampled within a single habitat, the spatial scale of sampling varied across regions (Table S1). Among the 87 legume species sampled, only 14 were distributed in more than one region (Table S2).

A single bacterial isolate was obtained from each root nodule. Inoculation tests using a subset of 81 isolates representing the major lineages from these populations indicated that all isolates were capable of inducing nodule development, and nearly all were symbiotically effective as shown by acetylene reduction assays or plant biomass gain (Parker, 1999, 2004, 2008 and unpublished data). Portions of three housekeeping genes (*dnaK, gyrB* and 23S rRNA) proximal to the SI chromosomal region, and two markers distal to the SI region (*rplC and rpoB*) were sequenced (see chromosomal location information in Table S4). For a marker within the SI region, the *nifD* locus was analysed. Out of a total of 1218 sequence data files (6 loci × 203 strains), 340 were available from prior studies (Parker, 2003,
2004, 2008, 2012); we obtained the other 878 sequences during the current study. Most isolates selected for sequencing (187/203) were chosen at random (without any prior molecular typing) from among the available strains for each host legume within each region. The remaining 16 isolates were from early studies (Parker, 1999, 2003, 2004) that had been chosen for sequencing based on isozyme polymorphism or rRNA length variation.

A 603 bp portion of the dnaK gene and a 5’ portion of the 23S rRNA gene (467–495 bp) were sequenced as described previously (Parker, 2004). An 822 bp portion of the nifD gene (encoding the nitrogenase \( z \) subunit) was amplified and sequenced as described previously (Parker et al., 2002). A 432 bp portion of the rplC gene (which encodes 50S ribosomal protein L3) and a 537 bp portion of the rpoB gene were amplified and sequenced as described previously (Parker, 2012). Finally, a 781 bp portion of the gyrB gene was amplified using primers 5’- GCACACATGGTCTACGAGGT-3’ and 5’-AGCCTTGCTCCGCAAC-3’ (the latter adapted from Rivas et al., 2009). gyrB amplification used a touch-down program: 59–56 °C (15 s, –1 °C per cycle for four cycles), 28 cycles of 55 °C (15 s) and 72 °C (60 s) and a final extension of 3 min at 72 °C. Nucleotide diversity (the mean pairwise number of nucleotide differences per site) was estimated using nDNA v5.10 (Librado & Rozas, 2009) for each of the six gene loci.

NeighborNet analysis using SplitsTree v4.11.3 (Huson & Bryant, 2006) indicated that concatenation of the five housekeeping gene loci resulted in a network (Fig. S1) that was only marginally more reticulated than networks for individual loci. The network for nifD sequences had a largely bifurcating structure with modest reticulation (Fig. S2). However, concatenation of nifD with the housekeeping loci resulted in a network with extensive reticulation, indicating that the phylogenetic signal in nifD sequences conflicted greatly with the other loci. To summarize bacterial relationships, we therefore present one tree for concatenated sequences of the five housekeeping loci and a separate tree for nifD sequences. Trees were inferred using MrBayes (Ronquist & Huelsenbeck, 2003) with protein-coding loci partitioned by codon position and separate estimates of substitution rates and nucleotide composition for each codon position and locus. A HKY substitution model was used. Several alternative models were tested but all yielded very similar trees. Analyses were run for 2 million generations, sampling every 250 generations, with the last 1000 sequences being saved for tree analysis. Replicate runs yielded identical consensus tree topologies. Trees were rooted using Azorhizobium caulinodans ORS571 (NC009937) and Xanthobacter autotrophicus Py2 (NC009720) as outgroups. Eight reference strains of Bradyrhizobium were also included: B. canariense BTA1\(^T\) (from Chamaecytisus proliferus in the Canary Islands; Vinuesa et al., 2005), B. elkanii USDA767\(^T\) (from soybean (Glycine max) in North America), B. japonicum USDA110 (from soybean in North America), B. lima- ligenense USDA3622\(^T\) (from soybean in China), B. jicamae PAC6\(^T\) and B. pachyrhizi PAC48\(^T\) (from Pachyrhizus erosus in Honduras and Costa Rica; Ramirez-Bahena et al., 2009), B. yuanmingsense LMG21827\(^T\) (from Lespedeza cuneata in China) and Bradyrhizobium sp. strain ORS278 (from Aeschynomone sensibila in Senegal). For phylogenetic clustering analyses, reference strains were only included if their host legume clade was represented in the primary sample of the 203 North American strains. Thus, legume genera used only by reference strains (Glycine, Pachyrhizus and Chamaecytisus) were not included in clustering analyses at the level of host genus. However, reference strains and their hosts were included in the clustering analyses focused on larger legume clades. Separate analyses showed that it made little difference whether the eight reference strains were included or excluded, because they were such a small proportion of the aggregate sample.

The Phylocom 4.0.1 software (Webb et al., 2008) was used to analyse whether bacteria were phylogenetically clustered according to the region from which they were sampled or clustered according to legume host group (Sachs et al., 2009). All clustering analyses were performed separately for the nifD tree and the housekeeping gene tree. For a given sample group (the set of bacterial strains from one region or the set of strains from one host taxon), the mean phylogenetic distance (MPD\(_{\text{sample}}\); Webb et al., 2008) was calculated as the average branch length separating pairs of strains within the group. Significance was assessed by comparing MPD\(_{\text{sample}}\) values to the null distribution inferred from 1000 random permutations of isolate names across tips of the observed tree. This test cannot be performed for legume taxa or regions represented by only one bacterial isolate. The net relatedness index, a standardized measure of the magnitude of clustering, was calculated as the difference between MPD\(_{\text{sample}}\) and the MPD value for the same sample size calculated from the null distribution (MPD\(_{\text{null}}\); divided by the standard deviation of MPD\(_{\text{null}}\) (Webb et al., 2008). As an alternative to the ‘global’ permutation implemented in Phylocom, we also performed more focused permutation tests where isolates were only randomized within specific data partitions. These tests (detailed in Results) were performed using original code written in R program (see supplementary material), which utilized the ‘distTips’ function from the Adephylo package (Jombart et al., 2010).

To analyse host-related phylogenetic clustering, isolates were first grouped according to legume genus (in our sample, too few isolates were available for individual species within a genus to permit a meaningful finer-scale analysis). We also tested whether aggregate sets of symbionts from larger legume clades clustered as relatives. A cladogram for 44 legume genera was based on Lewis et al. (2005), emended with data from Kajita et al. (2001) and Wojciechowski et al. (2004). For each branch in the cladogram, bacteria from all descendant host lineages of the branch were aggregated. Bradyrhizobium phylogenetic clustering at the level of each branch was then assessed by permutation analyses as outlined above. To reduce the possibility of spurious results from multiple significance testing, the procedure described by Benjamini et al. (2006) was applied to control the false discovery rate at \( q^* = 0.05 \). This resulted in exclusion of 15 permutation tests with borderline \( P \)-values (\( P = 0.037–0.048 \)) from the set considered to be significant.

**RESULTS**

**Patterns of nucleotide diversity**

Pairs of North American Bradyrhizobium isolates with identical six-locus haplotypes were rare (13 out of 20 503 pairwise comparisons). Each of these pairs of identical haplotypes came from the same region, suggesting that for natural ecosystems, it is likely that cosmopolitan Bradyrhizobium clones do not exist. The median pairwise nucleotide difference was about 10% (373/3674 bp), illustrating the high genetic diversity in the sample. Only 18 strain pairs had one or two nucleotide differences and all of these highly similar pairs were found in the same region, with the exception of one pair sampled in North Carolina and New York from the same host plant.

The gene nifD had approximately double the nucleotide diversity of the other five genes (Table S4). Elevated diversity in nifD was entirely due to synonymous variation, because non-synonymous variation was not higher than other genes (Table S4). High polymorphism in nifD may be a result of genetic hitch-hiking caused by balancing...
selection on linked loci in the SI region arising from symbiotic interactions with host legumes (Parker, 2012).

The isolates collected in Panama had the highest overall nucleotide diversity while those from Washington State had the lowest (Table 1). There was a significant negative correlation between nucleotide diversity and source region latitude (Spearman rank correlation, $r_s = -0.92; P<0.01$). Nucleotide diversity was not significantly correlated with the number of legume host species sampled in a region, the number of legume genera sampled or the average annual rainfall ($r_s$ ranged from 0.51 to 0.56; $P>0.10$).

**Bradyrhizobium phylogenetic relationships**

Bayesian analysis of concatenated sequences for five housekeeping gene loci yielded a well-resolved tree for North American *Bradyrhizobium* strains (Fig. 2). Closely related sets of isolates in 14 terminal clades were grouped prior to plotting to more easily depict the tree topology. These clades varied in size from 4 to 32 isolates. All 14 of the terminal clades had a posterior probability of 1.00.

The most basally divergent lineage included the photosynthetic *Aeschynomene* symbiont ORS278 and two strains from Panama. The placement of ORS278 is not unexpected, since genome analysis has shown that ORS278 and related photosynthetic strains have many unusual traits that set them apart from most other *Bradyrhizobium* lineages (Giraud et al., 2007; Nzoué et al., 2009). Strain aec12.3 in this group phenotypically resembled ORS278 in that it is also a stem-nodulating *Aeschynomene* symbiont. However, the other strain grouped with the *Aeschynomene* symbionts (ma9.4) was a root-nodule symbiont of a different host legume species (*Machaerium*).

The remainder of the strains split into two broad lineages affiliated with two sets of reference taxa: *B. japonicum*/*B. liaoningense*/*B. yuanmingense*/*B. canariense* (upper group; clades 1–8) and *B. elkanii*/*B. pachyrhizii*/*B. jicamae* (lower group; clades 9–14). For brevity, these will be referred to as the 'B. japonicum superclade' and the 'B. elkanii superclade', respectively. These two broad lineages have been detected in most prior phylogenetic analyses of *Bradyrhizobium* (e.g. Moulin et al., 2004; Vinuesa et al., 2008; Ramírez-Bahena et al., 2009; Rivas et al., 2009; Stepkowski et al., 2012). Both superclades were represented in all eight of the regions sampled (Fig. 2, right), indicating widespread geographical coexistence. North American bacteria with specific affinities to each of the seven named *Bradyrhizobium* species were detected, but were rarely predominant in more than one or two locations. For example, almost all isolates sampled from Washington State (24/25) grouped with *B. canariense* (clade 2), but no other samples belonged to this clade.

The eight regions varied substantially in their degree of clade overlap. For example, every clade found in northeastern USA and North Carolina was also present in Chihuahua, Mexico (Fig. 2, right). By contrast, only one of the two *Bradyrhizobium* clades present in the Washington State sample was detected in any other location (clade 14, Texas). Two clades had distributions that extended from the tropics (Costa Rica, Panama) to the mid-latitude temperate zone (clades 1 and 12). However, the predominant tropical clade of *Bradyrhizobium* was not detected in temperate North America or vice versa (compare clades 3 and 9).

Bayesian analysis of the *nifD* sequence yielded a well-resolved tree (Fig. 3). Closely related sets of isolates in 14 terminal clades were again grouped to more easily depict the tree topology (all 14 terminal clades had a posterior probability of 1.00). The three most basally diverging strains (ma9.4, ORS278 and aec12.3) were the same as in the housekeeping gene tree (Fig. 2). Certain sets of clades were highly congruent in the two trees (Table S5), but the topologies of the two trees conflicted in many places. For example, while the housekeeping gene tree was partitioned into two superclades affiliated with *B. japonicum* versus *B. elkanii* (Fig. 2), this structure was not evident in the *nifD* tree. All *B. elkanii* superclade strains had *nifD* sequences placed in a large monophyletic group comprising *nifD* clades 1–8 (Fig. 3; Table S5). However, two of the major

<table>
<thead>
<tr>
<th>Source region</th>
<th>No. of isolates*</th>
<th>Nucleotide differences†</th>
<th>Host species</th>
<th>Host genera</th>
<th>Latitude (°N)</th>
<th>Annual rainfall (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panama</td>
<td>23</td>
<td>0.089 ± 0.007</td>
<td>18</td>
<td>16</td>
<td>9.2</td>
<td>260</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>24</td>
<td>0.080 ± 0.007</td>
<td>16</td>
<td>13</td>
<td>9.2</td>
<td>434</td>
</tr>
<tr>
<td>Oaxaca</td>
<td>35</td>
<td>0.077 ± 0.005</td>
<td>19</td>
<td>17</td>
<td>16.4</td>
<td>102</td>
</tr>
<tr>
<td>Texas</td>
<td>21</td>
<td>0.080 ± 0.004</td>
<td>7</td>
<td>7</td>
<td>26.7</td>
<td>68</td>
</tr>
<tr>
<td>Chihuahua</td>
<td>22</td>
<td>0.076 ± 0.005</td>
<td>6</td>
<td>3</td>
<td>27.5</td>
<td>53</td>
</tr>
<tr>
<td>North Carolina</td>
<td>22</td>
<td>0.047 ± 0.009</td>
<td>16</td>
<td>13</td>
<td>35.5</td>
<td>114</td>
</tr>
<tr>
<td>North-eastern USA</td>
<td>25</td>
<td>0.062 ± 0.008</td>
<td>13</td>
<td>8</td>
<td>42.0</td>
<td>98</td>
</tr>
<tr>
<td>Washington State</td>
<td>25</td>
<td>0.022 ± 0.008</td>
<td>5</td>
<td>2</td>
<td>47.4</td>
<td>48</td>
</tr>
</tbody>
</table>

*Number of isolates studied per region; six isolates from other locations (Alaska, California and Illinois) were excluded due to insufficient regional sampling.
†Mean ± SD pairwise proportion of nucleotide differences among isolates within a region, across six gene loci.
nifD clades within this group (clades 3 and 6) were associated with *B. japonicum* superclade strains (Fig. 3; Table S5). Thus, there were numerous cases where divergent lineages in the housekeeping gene tree shared similar nifD sequences (and vice versa). Phylogenetic incongruence of the two trees is consistent with other research suggesting that lateral transfer of SI-region loci, such as nifD, has been a common event in *Bradyrhizobium* evolution (Parker et al., 2002; Moulin et al., 2004; Stepkowski et al., 2012; Parker, 2012).

For each isolate, a composite clade was defined based on concatenation of its assigned clade for the two trees (Table S6). This partitioned the 203 isolates into 46 groups. Roughly half of the groups were represented by only a single isolate. The largest group included 27 isolates sampled from 12 different legume host genera. The number of associated bacterial clade types per host legume genus varied from 1 to 13 (Table S6). For the 30 legume host genera from which at least two isolates were sampled, the number of bacterial clade types was positively correlated with the sample size of isolates per host genus ($r=0.71$), the number of regions where the genus was sampled ($r=0.70$) and the number of different legume species sampled per genus ($r=0.78$). These variables are all intercorrelated, but a partial correlation analysis showed that the relationship between bacterial clades per host genus and legume species sampled per genus remained significant after removing the influence of other variables ($r=0.40$; $P=0.034$).

### Phylogenetic clustering

Global permutation analysis using Phylcom (Webb et al., 2008) suggested that there was substantial phylogenetic clustering of bacteria according to the geographical source region in both trees (Table 2). Thus, bacteria from each location had genes that were significantly more similar on average, relative to the null distribution generated by permuting across all tips of the tree.

To further analyse geographical clustering, strains were divided into sets according to legume host genus and...
randomization was done only within each set. As a result, permutation for each location only involved sampling of tree tips associated with the same legume genera as in the original sample for that region. Genetic distances in the null distribution were therefore not affected by disparities in how bacteria from different host genera were distributed on the tree. Application of this more conservative method showed that for almost all regional samples in both trees, phylogenetic clustering remained significant, although many probability values were less extreme than in the global permutation analysis (Table 2).

Using global permutation, there was a significant signal of clustering for many host legume genera in both trees (Fig. 4). There were also cases of apparent phylogenetic clustering of *Bradyrhizobium* on some larger host clades at a variety of levels (Fig. 4). For example, the set of bacteria associated with all sampled legumes in the subfamily Mimosoideae (*Pentaclethra*, *Albizia*, *Pithecellobium*, *Inga*, *Cojoba* and *Enterolobium*) exhibited significant clustering in both trees. By contrast, bacteria from another large lineage, the Dalbergioid clade (Wojciechowski et al., 2004, including *Machae-rium*, *Dalbergia*, *Aeschynomene* and *Inga*), showed no signal of phylogenetic clustering in either tree.

To further analyse the putative cases of bacterial clustering on host legume clades (Fig. 4), more focused permutation tests were performed. Strains were divided into sets according to the source region and randomization was done only within each set. As a result, permutation for bacteria from each legume clade only involved randomization of tree tips from the same locations as in the original sample for that clade. In the null distribution, genetic distances were therefore not inflated by disparities in how bacteria from different locations were distributed on the tree. This analysis showed that several apparent cases of bacterial phylogenetic clustering on host taxa in the global analysis became non-significant when the more conservative permutation method was used (Fig. 4). For example, the number of legume genera with significantly clustered symbionts dropped from 19 to 10 for the *nifD* tree and declined from 11 to 8 for the housekeeping gene tree, when within-region permutation was used. Yet cases of bacterial phylogenetic clustering remained evident for legume clades at a variety of levels in the family (Fig. 4). Thus, related sets

---

**Fig. 3.** Bayesian tree for *nifD* sequences (822 bp) from 203 North American *Bradyrhizobium* isolates and eight reference taxa. Outgroups were deleted before plotting to save space and strains in 14 terminal clades were grouped. Within-clade disparity in branch length is indicated by rectangle size, and isolate numbers from each region are shown on the right. Branches with a posterior probability of 0.95 or higher are marked with a dot (see supplementary material for complete topology, strain identity and posterior probability values). See Fig. 2 legend for abbreviations. Scale bar indicates 0.05 nucleotide substitutions per site.
of legumes in some cases harboured related sets of *Bradyrhizobium* symbionts.

**DISCUSSION**

*Bradyrhizobium* biogeography

The results provide some new insights about *Bradyrhizobium* biodiversity on a continental scale. First, at the broadest phylogenetic level, there was little signal of regional endemism, with the *B. japonicum* and *B. elkanii* superclades coexisting in all regions sampled (Fig. 2). For narrower clades within these groups, most still had distributions that extended across multiple regions. For example, only two of 14 clades in the housekeeping gene tree were limited to a single region. Nevertheless, all of the regions sampled were occupied by distinctive bacterial populations, as indicated by the permutation results showing region-based phylogenetic clustering (Table 2). Thus, biogeographic structure in *Bradyrhizobium* involves local variation in prevalence of lineages whose ranges often span several regions.

In both trees, the median number of regions occupied per clade was only three (Figs 2 and 3). This suggests that a variety of constraints act to limit the geographical spread of *Bradyrhizobium* lineages. Adaptation to different abiotic environmental factors, and dispersal limitation, are likely to be important processes in this respect. However, a few clades had surprisingly broad distributions that encompassed both tropical and temperate habitats (Figs 2 and 3). In the housekeeping gene tree, the handful of Alaskan isolates that were sampled were placed in a clade that covered 48.5° of latitude and included strains from Oaxaca and Costa Rica. However, temperate and tropical populations were mostly dominated by different lineages. Intriguingly, there was no simple relationship between inter-population similarity and latitude. For example, sites at similar latitude in Washington State and north-eastern USA exhibited virtually no clade overlap. Yet a relatively high degree of clade overlap was evident for *Bradyrhizobium* populations in north-eastern USA and Chihuahua (Figs 2 and 3), despite their wider geographical and latitudinal separation. The strong affinity between bacterial populations from these two regions was unexpected. Their biotic communities are quite distinct (lowland temperate deciduous forest versus montane evergreen oak pine forest) and they had no legume host species in common. However, palynological studies show that many plant genera of northern temperate origin resided in eastern Mexico in the late Miocene (Graham, 1987), which raises the possibility that *Bradyrhizobium* affinities could be a legacy of earlier biotic interchange. An important priority for future work will be to clarify historical influences on *Bradyrhizobium* biogeography through more intensive region-based sampling.

**Host-related phylogenetic clustering**

It has proven difficult to find general patterns in symbiotic specificity in *Bradyrhizobium*. It is not yet known how common it is that legumes have developed exclusive relationships with certain *Bradyrhizobium* lineages or whether they associate mainly with less specialized symbionts that are shared with other coexisting plants. The current results provide new insights into the relative prevalence of these outcomes.

*Bradyrhizobium* species from about one-third of the legume genera studied (10/31) were significantly clustered in the

---

**Table 2. Phylogenetic clustering of *Bradyrhizobium* isolates in relation to the geographical source region**

Data shown are the net relatedness index (NRI), \((\text{MPD}_{\text{null}} - \text{MPD}_{\text{sample}})/\text{(SD of MPD}_{\text{null}})\) where \(\text{MPD}_{\text{sample}}\) is the mean pairwise patristic distance among strains from one region; positive NRI values indicate that strains from the same region tend to be phylogenetically close, relative to the null distribution generated by random permutation of strains across the tree. Probability of NRI inferred by comparison with null distribution based on 1000 random permutations of strains across tree branch tips: ***\(P<0.001\); **\(P<0.01\); *\(P<0.05\); (ns), \(P>0.05\).

<table>
<thead>
<tr>
<th>Source region</th>
<th>No. of isolates</th>
<th>nifD tree</th>
<th>5-locus housekeeping gene tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Global permutation</td>
<td>Randomization within host genus</td>
</tr>
<tr>
<td>Panama</td>
<td>23</td>
<td>3.04**</td>
<td>3.63**</td>
</tr>
<tr>
<td>Costa Rica</td>
<td>24</td>
<td>3.30**</td>
<td>1.90*</td>
</tr>
<tr>
<td>Oaxaca</td>
<td>35</td>
<td>4.21***</td>
<td>3.11**</td>
</tr>
<tr>
<td>Texas</td>
<td>21</td>
<td>5.33***</td>
<td>1.97*</td>
</tr>
<tr>
<td>Chihuahua</td>
<td>22</td>
<td>5.47***</td>
<td>1.70***</td>
</tr>
<tr>
<td>North Carolina</td>
<td>22</td>
<td>9.40***</td>
<td>3.07**</td>
</tr>
<tr>
<td>North-eastern USA</td>
<td>25</td>
<td>9.00***</td>
<td>3.32**</td>
</tr>
<tr>
<td>Washington State</td>
<td>25</td>
<td>14.82***</td>
<td>3.77**</td>
</tr>
<tr>
<td>California</td>
<td>2</td>
<td>2.36***</td>
<td>0.76 (ns)</td>
</tr>
<tr>
<td>Alaska</td>
<td>3</td>
<td>3.40**</td>
<td>0.90 (ns)</td>
</tr>
<tr>
<td>Combined</td>
<td>202</td>
<td>40.44***</td>
<td>8.82***</td>
</tr>
</tbody>
</table>
Nevertheless, a substantial amount of opportunistic symbiont of legume genera that show significant symbiont clustering. (Table S6). Greater sampling would no doubt add to the list locally coexisting legume taxa that were not close relatives $Zornia$, $Baptisia$ bacteria from many legume genera (including bionts utilized by a particular genus. For example, nodule $\text{nifD}$ tree (Fig. 4). In the housekeeping gene tree, symbionts from fewer legume genera (8/31) showed significant phylogenetic clustering. This is consistent with prior results showing that host-associated differentiation in symbiotic island loci (such as $\text{nifD}$) tends to be more pronounced than for other chromosomal regions (Parker, 2012). However, it was noteworthy that there was no signal of symbiont clustering in either tree for the majority of legume genera. A failure to detect clustering could be due to low statistical power caused by the small sample size for certain legumes or it could indicate a true lack of specificity among the symbionts utilized by a particular genus. For example, nodule bacteria from many legume genera (including $\text{Baptisia}$, $\text{Clitoria}$, $\text{Dalbergia}$, $\text{Desmodium}$, $\text{Lotus}$, $\text{Pentaclethra}$ and $\text{Zornia}$) were observed to be highly similar to those of other locally coexisting legume taxa that were not close relatives (Table S6). Greater sampling would no doubt add to the list of legume genera that show significant symbiont clustering. Nevertheless, a substantial amount of opportunistic symbiont usage appears to characterize many legume genera that associate with $\text{Bradyrhizobium}$.

For certain genera, a failure to detect clustering may also be a consequence of heterogeneity in symbiont utilization among species. If legume species within a genus harbour disparate symbionts, then a permutation test on the aggregated set of bacteria for the genus may fail to show a signal of phylogenetic clustering. In our results, pooling across legume species within a genus did not preclude detection of clustering, since in both of the trees, about half of the legume species within a genus did not preclude detection in either tree for the majority of legume genera.

**Fig. 4.** Outcomes of bacterial phylogenetic clustering tests mapped onto a cladogram for host legumes. Results for the two bacterial trees are separated by a forward slash. For legume genera, clustering results are shown in columns to the right. For larger legume clades, global versus within-region permutation results are shown above and below branches, respectively. Individual permutation $P$-values were screened (Benjamini et al., 2006) to control the false discovery rate at $q^*=0.05$; significant results (*) and non-significant results (-) under this method are shown.
For larger legume clades encompassing multiple genera, aggregation would also be expected to attenuate any signal of phylogenetic clustering seen for symbionts from particular genera. However, significant clustering was detectable for a few large legume clades. For example, bacterial symbionts from a set of five mimosoid genera (Albizia, Pithecellobium, Inga, Cojoba and Enterolobium) exhibited significant clustering in both trees. Two quite different processes could generate these sorts of larger-scale phylogenetic patterns. First, it is possible that legume traits that affect recruitment of certain symbiont lineages may show some degree of phylogenetic conservatism. There is evidence, for example, that some *Bradyrhizobium* lineages associated with genistoid legumes have a characteristic suite of Nod factor-modifying genes that may contribute to specificity with this host lineage (Stepkowski et al., 2007). In our sample, there was a strong signal of symbiont clustering for *nifD* sequences on the clad encompassing four genistoid genera (*Chamacystisus*, *Lupinus*, *Crotalaria* and *Baptisia*).

It is also possible that a signal of higher-level phylogenetic clustering could arise from filtering the effects of environmental variation acting independently on plants and bacteria. For example, a particular legume clade may be non-randomly distributed across habitats as a result of adaptation to soil types, rainfall, temperature regimes or other factors [as a possible example, consider that all of the mimosoid legumes in this study were present only in the three sampled locations closest to the tropics (Oaxaca, Costa Rica and Panama)]. If bacterial lineages differed in their ability to invade and persist in specific environments, this could foster non-random symbiont acquisition by particular legume clades. It would not be easy to identify the environmental filters that are important for plants or for bacteria. Nevertheless, these arguments suggest that a promising direction for future work on *Bradyrhizobium* biogeography and host specificity will be to more explicitly integrate ecology into the phylogenetic framework.

**ACKNOWLEDGEMENTS**

We are grateful to A. Andrus and C. Barrett (Cornell University), J. Bishop (Washington State University, Vancouver), B. Hammond (Spokane, WA), S. Mondlo (Cornell University), M. Monroe (Kodiak, AK), I. Parker (University of California, Santa Cruz), M. B. Parker (Spokane, WA), R. Pratt (University of California, Irvine), R. Rivas and E. Velazquez (Universidad de Salamanca) and P. Vinuesa (UNAM, Cuernavaca) for help with collecting or providing strains. We also thank A. Andrus and C. Barrett (Cornell University), J. H. Koppell and M. A. Parker for financial support from the National Science Foundation grant MCB-0640246.

**REFERENCES**


Edited by: W. Achouak