Published paper

Lozano, L; Hernandez-Gonzalez, I; Bustos, P, et al (2010)
Evolutionary Dynamics of Insertion Sequences in Relation to the Evolutionary Histories of the Chromosome and Symbiotic Plasmid Genes of Rhizobium etli Populations
APPLIED AND ENVIRONMENTAL MICROBIOLOGY
76 (19) 6504-6513
http://dx.doi.org/10.1128/AEM.01001-10

Evolutionary Dynamics of Insertion Sequences in Relation to the Evolutionary Histories of the Chromosome and Symbiotic Plasmid of *Rhizobium etli* Populations

Luis Lozano[1]*, Ismael Hernández-González[1], Patricia Bustos[1], Rosa I. Santamaría[1], Valeria Souza[2], J. Peter W. Young[3], Guillermo Dávila[1] and Víctor González[1].

[1]Centro de Ciencias Genómicas, UNAM, Av. Universidad N/C Col. Chamilpa, Apdo. Postal 565-A Cuernavaca, Morelos, México.

[2]Departamento de Ecología Evolutiva, Instituto de Ecologia, Universidad Nacional Autónoma de México, AP 70-275, CU, Coyoacán 04510, Mexico DF, México.

[3]Department of Biology, University of York, PO Box 373, YO10 5YW, York, UK.

*Corresponding Author: Luis Lozano

llozano@ccg.unam.mx

Phone: 01 777 3291690

Fax: 01 777 3175581

Centro de Ciencias Genomicas, UNAM. Av. Universidad N/C, Col. Chamilpa. Apdo. Postal 565-A. Cuernavaca, Morelos, Mexico

Running Head: IS and pSym Dynamics in *R. etli* Populations

Keywords: insertion sequence, symbiotic plasmid, genome context

ABSTRACT (235 words)

Insertion sequences (IS) are mobile genetic elements that are distributed in many prokaryotes. In particular, in the genomes of the symbiotic nitrogen fixing bacteria collectively known as rhizobia, IS are fairly abundant in plasmids or chromosomal islands that carry the genes needed for symbiosis. Here, we report an analysis of the distribution and genetic conservation of the IS found in the genome of *Rhizobium etli* CFN42, in a collection of 87 *Rhizobium* strains belonging to populations of different geographical origins. We used PCR to generate presence/absence profiles of the 39 IS found in *R. etli* CFN42, and evaluated whether the IS were located in consistent genomic contexts. We found that the IS from the symbiotic plasmid were frequently present in the analyzed strains, whereas the chromosomal IS were observed less frequently. We then examined the evolutionary dynamics of these strains based on a population genetic analysis of two chromosomal housekeeping genes (*glyA* and *dnaB*) and three symbiotic sequences (*nodC* and the two IS elements). Our results indicate that the IS contained within the symbiotic plasmid have a higher degree of genomic context conservation and a lower nucleotide diversity, genetic differentiation and fewer recombination events compared with the chromosomal housekeeping genes. These results suggest that the *R. etli* populations diverged recently in Mexico, that the symbiotic plasmid also had a recent origin, and that the IS elements have undergone a process of infection and expansion cycle.

Introduction

Insertion sequences (IS) are the smallest transposable elements found in prokaryotes (usually less than 3 kb in size). They encode a transposase and may also encode small hypothetical proteins (4, 9). IS are distinguished by their ability to move within prokaryotic replicons, including both the chromosome and plasmids, and copy themselves into various genomic sites. In this manner, IS elements can inactivate or alter the expression of adjacent genes (4). When IS occur in two or more identical copies within a genome, they can participate in various types of genetic rearrangements (e.g., duplications, inversions and deletions) suggesting that IS may play an important role in the evolution of their hosts by promoting genomic plasticity (34). Due to these evolutionary dynamics, the diversity and distribution of IS elements differs greatly between taxa and even within strains of the same species (27).

Various theories have been proposed to explain the evolution of IS elements in laboratory model strains and environmental bacterial populations (8, 18, 25, 29). Two main hypotheses seek to explain how these elements are maintained over the long term in their host genomes: The first proposes that they occasionally generate beneficial mutations, and therefore may represent a selective advantage to their hosts (34). The second suggests that IS elements are genomic parasites that are maintained by their high rate of transposition and might be disseminated among different bacterial lineages by horizontal gene transfer (HGT). Data supporting the second hypothesis has shown that some IS elements may transpose at high rates upon entering a new host (43). After the initial infection, however, purifying selection may continuously remove these elements from the genome. Thus, IS may be under an infection-expansion-extinction cycle that allows them to remain in different bacterial populations within the gene pool (43).

These two hypotheses are not contradictory, and the evolutionary dynamics and distribution of IS may differ greatly depending on several factors, including (most notably) the rate of transposition and HGT, as well as selective pressures, population size and the host's habitat (18, 25, 27, 29, X, Y).

In the nitrogen-fixing symbiotic bacteria of the genera *Rhizobium*, *Sinorhizobium*, *Mesorhizobium*, *Bradyrhizobium* (of the α-proteobacteria), *Cupriavidus*, and *Burkholderia* (of the β-proteobacteria), IS are particularly abundant in symbiotic plasmids (pSym) and symbiotic chromosomal islands (SI) (2, 12, 14, 19, 20, 42). SI and pSym encode most of the genes needed to establish symbiosis in the roots of leguminous plants through nodule formation and nitrogen fixation (11). It is generally believed that these elements entered the rhizobial genomes through HGT (39, 40). Comparative genomic analyses have shown that both pSym and SI are highly variable with the exception of a common set of genes encoding factors critical to nitrogen fixation (*nif*) and nodulation (*nod*) (5, 14). SI and pSym have been found to have lower GC contents and different codon usages than the corresponding chromosomal and non-symbiotic plasmid sequences, suggesting that they were recently acquired by HGT.

Some of these symbiotic elements, such as in the cases the pSym of *R. etli* CFN42 and the SI of *Mesorhizobium loti*, are conjugative and mobile (30, 32). Genomic analysis of *R. etli* CFN42 revealed the presence of 39 IS belonging to different families (14); these were found in the chromosome (11 IS), the 371-kb symbiotic plasmid (14 IS), the smaller 192-kb conjugative plasmid, p42a (13 IS) and two other plasmids, p42b and p42c (2 IS). Interestingly, this particular strain shows no evidence of IS disrupting ORFs or having transpositional activity. However, another 42 incomplete IS may be found in the chromosome, pSym, and the conjugative plasmid; these incomplete sequences are truncated or contain stop codons in their coding sequences.

Here, we focused on the dynamics and distribution of IS in different populations of the nitrogen-fixing symbiont *Rhizobium etli*. Since the maintenance of IS in bacterial species might depend on their transpositional activity and horizontal transfer rate, the identification of IS in the same genomic contexts across different strains of the same species could provide new insights into their persistence and divergence over short evolutionary periods. To examine the evolutionary dynamics of IS in natural populations of *R. etli*, we characterized the distributions, genomic contexts and sequence variations of IS in isolates of *R. etli* from three populations of different origins, as well as in some other *Rhizobium* species. More specifically, we used PCR to generate presence/absence profiles of the 39 IS found in *R. etli* CFN42, in a collection of 87 strains representing different geographical sites and a gradient of domestication of the bacterial host, the common bean (*Phaseolus vulgaris*). We also evaluated whether the IS were conserved in the same genomic context relative to their position in *R. etli* CFN42, and determined the nucleotide sequence of two IS found in most of the isolates. Several population genetic tests applied to these IS, another pSym gene (*nodC*) and two chromosomal housekeeping genes (*dnaB* and *glyA*) suggest that these two IS elements have been inherited vertically and represent recent components of the *R. etli* gene pool. Finally, the present study strongly suggests that symbiotic plasmids have a recent origin within the *R. etli* populations.

MATERIALS AND METHODS

Bacterial strains, growth conditions and DNA extraction

The 87 *Rhizobium* strains used in this study correspond to three different collections (Table 1). Two of the strain collections are from Mexico. The first collection

(36) was derived from two plots in a traditional milpa system of native bean landraces (San Miguel Acuexcomac, Puebla); the collection consists of 30 strains that were the dominant strains for several years. The second collection (13) comes from the Michoacan-Guanajuato area, which is the reported center of origin of bean domestication (24); the 30 strains in this collection include a bean plant domestication gradient from wild non-domesticated *Phaseolus vulgaris* to milpa landrace cultivars, as well as wild bean plants that are probably the descendants of cultivated plants. The third collection (33) represents *R. etli* from Spain and includes 27 strains obtained from 21 soil samples collected along the Guadalquivir River Valley. These are believed to represent either original native rhizobia or a sub-sample of New-World rhizobia that travel along the original bean seeds (Table 1). Some strains from the Puebla and Spanish collections were previously characterized as *R. gallicum*, *R. giardini*, *R. fredii* and *R. leguminosarum* (Table 1).

The various *Rhizobium* strains were grown in LY medium for 24-48 h. Genomic DNA was extracted with a GenomicPrep DNA Isolation Kit (Amersham Biosciences) and the DNA concentration was determined with a DyNA Quant 200 fluorometer (Hoefer). Two complete sequenced *R. etli* strains, CFN42 and CIAT652, were used as references strains.

PCR amplification and DNA sequencing

We first localized each of the 39 selected IS within the CFN42 genome, thereby "anchoring" our examination of whether these IS conserved their genomic locations (i.e., their synteny) across the wild *R. etli* isolates from different geographical regions. These 39 IS elements represent 11 different famlies and some of these families have identical or near identical copies: 6 and 4 copies for 2 different elements of the IS66

family, 5 copies for IS630, 2 for IS1111 and 2 for IS21. We then designed specific PCR primers spanning the immediate neighborhoods of the 39 studied IS. These primers allowed us to test for conservation (i.e. amplification of a fragment of similar size to the one predicted in CFN42). In cases where we found two contiguous IS elements, we designed primers in the neighboring genes of both IS. There were three possible results: i) no PCR product, indicating that there were no identical priming sites in the wild isolate (this was not an absolute positive or negative result for the presence of the IS); ii) a small DNA fragment, equivalent to the distance between the two sequences near the insertion site but lacking the IS (a negative result for the IS); and iii) a large DNA fragment of similar size to that in CFN42 (a positive result for the IS). The PCR reactions were performed in a DNA thermal cycler (Gene Amp 9700; Applied Biosystems) in a 25 µl reaction mixture containing approximately 10 ng genomic template DNA, 2 mM $MgCl_2$, 1 mM dNTPs, 5 pmol of each primer, and 2 U of TAQ Polymerase (AltaEnzymes). The mixtures were subjected to 5 min denaturation at 94°C followed by 30 cycles of 1 min at 94ºC, 1 to 3 min at 58 to 62ºC, and 3 min at 72ºC.

For a comparison of evolutionary dynamics, we performed direct sequencing on the following: two IS elements, ISRel4 and ISRel2 from pSym, which were present in the highest proportion of test samples (see Results and Fig. 2); *nodC*, which is a pSym gene encoding an N-acetylglucosaminyltransferase that participates in the nodulation process; and two chromosomal housekeeping genes, *glyA* (serine hydroxymethyltransferase) and *dnaB* (replicative DNA helicase), which have been proposed to serve as predictors of genome relatedness because their sequence divergence rates reflect the overall rate of genome divergence (44). Internal PCR primers were designed for the latter three genes to obtain partial sequences of each gene. The sequencing reaction mixtures contained approximately 10 ng genomic

template DNA, 1 mM $MgCl_2$, 1 mM dNTPs, 5 pmol of each primer, and 1U of rTth Polymerase XL (Applied Biosystems). The mixtures were subjected to 5 min denaturation at 92°C followed by 30 cycles of 30 sec at 92ºC and 1 to 6 min at 58 to 62ºC. The PCR products were purified with an Exo/SAP kit (Affymetrix) and sequenced using a Dye-terminator cycle sequencing kit (Perkin Elmer Applied Biosystems). The sequencing reactions were run on an ABI 3730 sequencer (Applied Biosystems).

Sequence analysis, alignments and phylogenetic reconstruction

Sequence quality analysis, assembly and comparison were performed using the SeqScape software V2.5 (Applied Biosystems). For IS characterization, we compared all of the putative transposases in the complete annotated genome sequence of *R. etli* CFN42, to the nr Database of the NCBI and the Insertion Sequence Database (35), with BLASTp (1).

To define a complete and (most likely) functional IS, we used the same criteria applied by the authors of the Insertion Sequence Database, as follows: (i) similar organization of the genes; (ii) sequence similarity of the transposases and inverted repeats; and (iii) the presence of direct repeats (14). Gene alignments for the assessed genes and IS elements were done using the MUSCLE program (7). For each alignment, the best substitution model was determined using FindModel (31). Phylogenetic reconstruction was performed using the maximum likelihood method implemented in PHYML (16). The analysis was carried out with a nonparametric bootstrap analysis of 100 replicates for each alignment. The complete chromosome and plasmid sequences of *R. etli* CFN42 and CIAT652 were compared with the Mummer program (23). The genomic contexts of the IS in the two strains were examined using Perl programs. The

rate of synonymous and non-synonymous substitutions, dn/ds ratio, was evaluated using SNAP (22).

Population genetic parameter estimation

DNASP version 4 (26) was used to assess the following parameters: the average nucleotide divergence between populations (D_{xy}); the average nucleotide diversity within populations; the average nucleotide divergence per site (π); the average nucleotide diversity at synonymous (π_s) and non-synonymous sites (π_{ns}); gene flow estimates (F_{st} and Nm); genetic differentiation estimates (K_s, K_{st}, Z and S_{nn}); the numbers of shared haplotypes, fixed and shared polymorphisms; and Tajima's D neutrality test.

Recombination analysis

Recombination tests were performed with RDP3 (28) and SplitsTree4 (17) for the five genes studied in each of the collections. The RDP3 software was applied to detect and analyze recombination signals using eight different methods (RDP, GENECONV, Bootscan, MaxChi, Chimaera, SiScan, 3SEQ and LARD), with 100 permutation steps, a Bonferroni correction for multiple tests, and a P-value threshold of 0.05. Recombination events and breakpoints were accepted if they were detected by two or more methods that used different approximations to detect recombination. The recombination events were then confirmed independently for each recombinant strain, as suggested in the literature (RDP website user manual; http://darwin.uvigo.es/rdp/rdp.html). Specific parameter modifications were used for each gene alignment depending the number of sequences in the analysis. Split decomposition and neighbor net analyses were performed with the SplitsTree4 software. Both phylogenetic networks represent incompatibilities within and between datasets with the use of splits. For the neighbor net

analysis, we applied a split filter based on the network weight and a threshold value of 95% (3).

Nucleotide sequence accession numbers

The relevant sequence data have been deposited in the GenBank database under accession numbers GUO84443 to GUO84796.

RESULTS

Differentiation of *R. etli* populations

To study the evolutionary dynamics of the IS of *R. etli*, we first asked if the three studied collections of *R. etli* strains, which were obtained from the root nodules of *P. vulgaris* plants located in different areas, were geographically structured. The phylogenies constructed based on the chromosomal housekeeping genes, *glyA* and *dnaB*, were relatively similar to each other, with a few differences in the groupings of some strains (data not shown). In order to obtain a phylogenetic reconstruction that more accurately represent the evolutionary relationships among the 87 strains contained within the three collections, we concatenated the sequence alignments (2238 bp) of both genes. We obtained three large *R. etli* clusters that broadly corresponded to the three different geographical areas from which the strains were collected. The Puebla collection formed one group with some intermingled strains from the Michoacan-Guanajuato and Spanish collections (Fig. 1). The second group consisted of the Michoacan-Guanajuato strains with three intermingled sequences belonging to the Spanish collection. The third *R. etli* group represented the Spanish collection with a single intermingled sequence from the Puebla collection. The Michoacan-Guanajuato and Spanish collections were more closely related to one another than to the Puebla

collection (Fig. 1). A distant fourth group, which was later used as outgroup, contained several *R. gallicum* strains from Puebla and Spain. Since this phylogenetic reconstruction clearly separates the three different collections according to their geographic origins, we considered them to be different populations in the following analyses.

IS profiles of the *R. etli* populations

We next examined which IS from *R. etli* CFN42 were also maintained in the same genomic context among the *R. etli* isolates from the three different populations. Our PCR-based profiling (see Methods) yielded either positive or negative results for 24 of the 39 IS, and showed marked differences in the conservation of the various IS among the tested populations (Fig 2). ISRel4 and ISRel2 (from pSym of *R. etli* CFN42) were the most common IS, present in more than 50% of the individual strains. Sequencing experiments showed that ISRel9 (also from pSym) in most cases was actually a hypothetical protein (hyp304) not related to any IS element and similar in size to ISRel9. Only five isolates (all from Puebla) had an intact ISRel9 in the expected genomic context; these elements had an average identity of 100%. ISRel5 and ISRel12 of pSym were relatively frequent, as they were present in 25% of the tested strains. The other seven IS from pSym were not common. The IS from p42a were found in the Puebla and Michoacan-Guanajuato populations, but not in the collection from Spain, while the chromosomal IS were found only in the Puebla strains. Interestingly, the CFN42 reference strain was found to be close to the Michoacan-Guanajuato population (Fig. 1).

DNA sequence divergence of ISRel4 and ISRel2

The observation that some IS occur in the same genomic context suggests that these IS may have been present since the arrival of the pSym in the populations, as it is far less likely that independent transposition events could have inserted the IS into the same genomic sites. It is important to mention that the ISRel4 and ISRel2 belong to IS families, ISAs1 and IS481 respectively, that do not have specific target sequences for transposition. To investigate the evolutionary dynamics of the two IS elements, ISRel4 and ISRel2, that show conservation of their genomic contexts, we looked for differences in the degree of diversification and selection pressures among these IS elements, as well as in other symbiotic and chromosomal housekeeping genes (Supp. Table 1). The DNA sequence conservation was very high for ISRel4 and ISRel2, with an average identity of 98.7% and 99.4%, respectively. This is surprisingly high compared to the sequence identities for the housekeeping genes *glyA* and *dnaB*, which were 75.3% and 76.4%, respectively. Given that the diversification of chromosomal genes could differ from that on the symbiotic plasmid, we sequenced the *nodC* gene, which encodes an N-acetylglucosaminyltransferase involved in the synthesis of sugar backbones for the production of lipo-oligosaccharides (critical to nodulation signaling). This was done in order to have another gene to trace the diversification differences between the pSym and the chromosome. The average nucleotide identity among the 44 analyzed *nodC* sequences was 96.7%. The nucleotide diversity (π) values for the pSym sequences (ISAs1, IS481 and *nodC*) were very low compared with those of *glyA* and *dnaB* (Table 2). The total average π values of *dnaB* and *glyA* were 0.066 and 0.088, respectively, whereas those for ISRel4, ISRel2, and *nodC* were 0.002, 0.002, and 0.014, respectively (Table 2). Since the number of polymorphic sites and the π values for ISRel4, ISRel2, and *nodC* were very low, we hypothesize that these sequences may have entered the gene pool of the three populations relatively recently.

Recent origins of ISRel4, ISRel2, and pSym

To evaluate the hypothesis that ISRel4 and ISRel2 may have originated recently, we measured their average nucleotide diversities at synonymous sites (πs) and non-synonymous sites (πns), and compared these values with those for *nodC*, *glyA*, and *dnaB*. The πs value reflects the age of an allele in the genetic pool; genes with lower πs values are generally considered to have recent common ancestor. Our results revealed that the πs values were low for ISRel4, ISRel2, and *nodC*, but high for *dnaB* and *glyA*, whereas the πns values were low for all five sequences (Table 2). Similar results were obtained when these πs and πns values were measured within each population (Supp. Table 2). These findings indicate that both the IS elements and *nodC* might have entered recently in the gene pool of the three populations.

Overall, the dn/ds ratio was lower for the chromosomal genes than for the pSym genes. To test what this might mean in terms of pSym evolution, we compared the shared haplotypes for chromosomal and pSym genes among the three populations. We did not find any shared *dnaB* and *glyA* (i.e. chromosomal) haplotypes between the populations. In contrast, we identified a number of shared haplotypes for the pSym genes (*nodC*, ISRel2 and ISRel4) (data not shown). These results suggest either that pSym has a relatively recent origin, or that there is high gene flow among pSym sequences; regardless of either option, it seems that the pSym sequences have a different evolutionary history than the chromosomal genes. To further test the possible recent origin of pSym sequences in the gene pool of the three populations, we compared the mean genetic divergence values between populations (Dxy) to the π values describing the mean genetic divergence within the populations. Higher values of Dxy indicate increasing time since population divergence. Table 3 shows that the Dxy values were

higher than the mean within-group divergence (π) for the chromosomal genes, potentially reflecting phylogenetic differentiation among the three populations. In the case of the pSym sequences, however, the Dxy and π values were very low and did not show clear phylogenetic differentiation between the populations. This further supports the idea that the pSym sequences have a relatively recent origin in the gene pool of the three populations, and have followed a different evolutionary history than the chromosomal genes.

We applied additional genetic differentiation tests (Ks*, Kst, Z, Z*, Snn) to the pSym and chromosomal sequences to test whether the two IS and *nodC* have a lower level of genetic differentiation than the chromosomal sequences which would further support a recent origin within the three populations. The results of the genetic differentiation tests for the two housekeeping genes were highly significant ($p<0.001$), supporting the idea of genetic differentiation among the three populations (Table 4 and Supp. Table 3). In contrast, the three pSym sequences had much lower levels of genetic differentiation across the three populations. For example, comparison of the pSym sequences from the two Mexican populations (those from Puebla and Guanajuato-Michoacan) yielded Kst values that were either non-significant or were just barely significant at $p<0.05$ (Table 4). Although the chromosomal housekeeping genes showed evidence of genetic differentiation, we did not find evidence of fixed polymorphisms; this may indicate that the three populations also diverged recently. Given that the pSym sequences had low πs values, their Dxy and π values were not well differentiated, had several shared alleles, and the genetic differentiation tests were not significant, it is not surprising that these sequences also lacked fixed polymorphisms.

Next, we analyzed whether there was some degree of gene flow across the three populations. If a Fst value above 0.25 and a Nm value > 1 were taken to represent a

significant between-population gene flow (37), most of the analyzed genes could be considered to have evidence of gene flow. The pSym sequences often had higher Fst values than the chromosomal genes (Table 4). A comparison between the Mexican and Spanish populations also yielded values indicative of gene flow. Given the geographical distance between these populations, it seems conceivable that these values could reflect a recent divergence of the three populations and a recent origin for the pSym genes.

There is published evidence indicating that genetic exchange occurred within *R. etli* populations from a single field; this process involved both plasmid and chromosomal loci (36). Accordingly, we hypothesized that the number of recombination events would be different between the chromosomal and pSym genes if their divergence times were clearly distinct. To assess this, we used the RDP3 software package to implement eight different recombination tests to search for recombination events across the three populations (Supp. Table 4). The chromosomal genes, *dnaB* and *glyA*, showed evidence of four and two intragenic recombination events, involving ten and five strains, respectively. In contrast, only one pSym sequence (ISRel4) showed a recombination event, even though the low sequence divergence make it harder to detect recombination. Interestingly, some of the same recombination events were found in strains of both *R. etli* and *R. gallicum*; moreover, four recombination events found between populations could explain some intermingled strains in the phylogeny of the three populations (Fig. 1 and Supp. Table 4).

Next, we used a phylogenetic network approach, consisting of split decomposition and neighbor net analyses to test whether the inconsistencies in the phylogenetic reconstructions could be due to recombination events. The split decomposition analysis of the two chromosomal genes showed a partition in the data between sequences from the *R. etli* and *R. gallicum* species. In the case of *glyA*, we

found some splits within the *R. etli* strains; these could come from the detected recombination events. ISRel4 was the only pSym sequence for which we obtained a split that clearly represents the unique recombination event we detected in the *R. etli* and *R. gallicum* strains. The neighbor net analysis, which is more sensitive to the conflicting sequence data that may represent recombination events, showed that the two chromosomal genes had several splits, some of which were related to the detected recombination events (Supp. Fig. 1). ISRel4 yielded the same split we had found in the split decomposition analysis. The other two pSym sequences failed to reveal splits with either analysis; this was consistent with the lack of any recombination events detected by our RDP3 analysis of these sequences. The relative lack of splits and recombination events in the pSym sequences supports our hypothesis that these genes (and probably the symbiotic plasmid itself) originated relatively recently in the three populations.

Another possibility that could explain the high DNA conservation of the pSym sequences and the conserved genomic contexts of ISRel2 and ISRel4 is that selection pressures may inhibit variation at these sequences. However, Tajima's D neutrality tests showed that most of the observed nucleotide substitution patterns in the chromosomal and pSym sequences from the three populations were under a neutral equilibrium model (i.e., all statistical results were non-significant; Table 2 and Supp. Table 2). The neutral equilibrium model was rejected only for *nodC*. However, this could be due to the effect of a decrease in population size and/or balancing selection.

IS elements in *Rhizobium etli* symbiotic plasmids

We recently sequenced another *Rhizobium etli* strain, the Costa Rican strain CIAT652, which has one chromosome, three plasmids, and 22 IS elements distributed mainly in the chromosome and the symbiotic plasmid (15). Comparative analysis of the

CFN42 and CIAT652 symbiotic plasmids showed that their sequences were highly conserved, with a nucleotide identity value of 99% (Fig. 3). In contrast, the identities of the chromosome and other plasmid sequences ranged from 90 to 95%. The two pSyms differed mainly in the presence/absence of certain genomic regions, the diversity of IS families, and presence/absence the IS elements in a given genomic context. This indicates that the major evolutionary events modifying these two symbiotic plasmids were the gains/losses of genomic regions and the losses/translocations of IS elements. In contrast, the chromosomes and other plasmids did not contain IS in the same genomic contexts. Comparison of the IS elements in the chromosomes and plasmids indicated that the symbiotic plasmids may have originated externally, and harbored IS that thereafter transposed to the chromosomes. Both CFN42 and CIAT652 harbored IS elements that were 100% identical in sequence between the chromosome and pSym. If these plasmids have a recent origin (as suggested by our population genetic analyses), this would seem to indicate that the IS had recently moved to the chromosome in both strains, in the manner of an infection-expansion-extinction cycle (43).

DISCUSSION

In order to examine the evolutionary dynamics of IS elements in relation to the evolutionary history of the chromosomal and symbiotic plasmid genes, we herein compared the distribution and genetic diversity of these sequences in three *R. etli* populations. First, we used the concatenated alignment of the chromosomal *glyA* and *dnaB* genes to perform a phylogenetic reconstruction, and found that the three tested *R. etli* populations were almost completely differentiated according to their geographic origins. Moreover, they were clearly differentiated from the *R. gallicum* strains, which

formed a fourth clade. The close phylogenetic relationship between the Michoacan-Guanajuato and Spanish populations was highlighted by the individual phylogenies of *dnaB* and *glyA* especially the one based on *glyA*, where several strains appeared to be intermingled in both populations. This pattern might be explained by recombination events, which were assessed using the RDP3 software (Supp. Table 4). However, another possibility is that the intermingled strains from Guanajuato could be closely related to migrant strains isolated from the Spanish population. The phylogeny made with the concatenated alignment of *dnaB* and *glyA* was used as the reference for mapping the IS distributions.

Comparison of the IS presence/absence profiles of the *R. etli* populations to that of the CFN42 reference strain showed that the most conserved IS were on the symbiotic plasmid. Because the majority of the IS elements did not maintain their genomic context, the presence/absence profiles were inconsistent with the housekeeping genes phylogeny; some strains that were closely related in the phylogeny had very different IS profiles. For example, most of the Puebla strains had more similar IS profiles to that of the model strain, CFN42, which belonged to the Guanajuato-Michoacan clade. These data demonstrate that the distribution of IS elements between and within populations is strain-specific in concordance with the HGT and transpositional dynamics of IS (4).

We analyzed two of the IS elements, ISRe12 and ISRel4 (both from pSym) in detail, as they were the most conserved, in terms of their genomic context, across the three *R. etli* populations. This pattern may indicate that some selective advantage is conferred by the presence of these IS elements, or it could be a consequence of the apparently recent origin of the pSym plasmid. Other studies on the evolutionary dynamics of IS elements have offered different explanations for the maintenance of such elements in populations or strains of the same species. For example, a report on IS

in strains of *Helicobacter pylori* from different geographical origins suggested that IS are ancient components of the *H. pylori* gene pool, and they evolve at approximately the same rate as normal chromosomal genes (18). A paper on IS within a single population of the hyperthermophilic archaeon, *Pyrococcus*, suggested that the high frequencies of some IS were due to genetic drift (8). Neither of these explanations seems to be applicable to the IS in *R. etli* examined herein. Instead, we think that the IS have undergone rapid turnover in the population, and that the contextual and sequence conservation of ISRel2 and ISRel4 should be viewed within the evolutionary history of pSym.

Due to the transpositional nature of IS elements, they are not expected to be found at the same genomic site in populations of different geographical origin. However, we herein observed such conservation for ISRel2 and ISRel4. One explanation for our observation is that that the IS might have been frequently found in the same genomic context as consequence of the small population size of the *R. etli* collections analyzed in this work. The probability of finding several strains with the same IS in the same location is lower for large populations than for small populations (38). In our case, however, this may not be the best explanation given the characteristics of the three populations (see Methods) and the geographical distances separating the collection sites. Another possibility is that these IS elements were recently acquired by the studied *R. etli* strains. Comparisons among these IS elements, *nodC* and the chromosomal genes suggested that the pSym sequences recently entered the *R. etli* gene pool of the three populations. Population genetic analyses of the pSym sequences showed that they have the following features: i) low π values; ii) Dxy and π values thet imply no clear differentiation; iii) several different shared alleles; and iv) non-significant results from their genetic differentiation tests. In contrast, opposing results

were obtained for the chromosomal genes, *glyA* and *dnaB*. Thus, the differences noted in our population genetic analyses suggest that the chromosome and pSym have different evolutionary histories in the studied *R. etli* populations. These findings, along with the high sequence conservation found in the pSym sequences of CFN42 and CIAT652, support the proposition that pSym could have a recent origin in *Rhizobium etli* (14). However, the pSym sequences from CFN42 and CIAT652 have clear differences in terms of the presence/absence of certain genomic regions (some as large as 50 kb) and IS elements; notably, the gains and losses of IS elements appear to be the main events differentiating these two pSyms.

A prior analysis of the complete genomes of *R. etli* CFN42 and *R. etli* CIAT652 demonstrated that the IS do not disrupt genes (with one exception), and these strains appear to harbor the same pSym plasmid (15). Horizontal transfer of this plasmid to *R. etli* CFN42 and *R. etli* CIAT652 could explain the asymmetric distribution of IS, which were only found in these plasmids and in the chromosome. Some IS elements probably moved from the plasmids to the chromosome, but they do not appear to have moved to other plasmids. The asymmetrical distribution of IS in the chromosome and plasmids could be due the size of the different replicons in the *R. etli* CFN42 genome (41). In the chromosome of CFN42 there are six probable recent transposition events of IS elements (100% identity between the chromosomal IS and another IS in the pSym or conjugative plasmid). This implies a transposition rate of one IS per 730 kb, which is higher than the size of any other replicon of CFN42 (14), so the asymmetrical distribution could be the product of the chromosome size.

The complete genomes of the two strains of *R. etli,* CFN42 and CIAT652, revealed the presence of different IS family members in different copy numbers. It has been proposed that when new IS elements enter the genome, they actively transpose to

different genomic positions (43). This hypothesis may help explain the behavior of some of the IS in *R. etli* CFN42. For instance, members of the IS66 family, which is a very common IS family in rhizobial species (27), were found in several copies in the pSym plasmids of CFN42 and CIAT652, and IS66 copies with 100% nucleotide sequence identity can also be found in the chromosome. The most plausible explanation for this finding is that the IS originally entered the genome via plasmids and were then transposed to the chromosome.

The present work provides useful new insights into differences of the distribution and genomic context maintenance of IS elements in natural populations of *Rhizobium etli*. Although the pSym in these populations seems to be of relatively recent origin, the harbored IS appear to be active elements that may participate in the genomic plasticity of these organisms. Our comparisons of the two genomes of *R. etli* and across the natural populations provides further evidence that different IS can rapidly expand when they arrive in a new host genome, as recently proposed by Wagner (2006).

AKNOWLEDGMENTS

We thank Santiago Castillo for critical reading of the mnuscript. We thank Susana Brom (Universidad Nacional Autónoma de México) for providing the Spainish strains; Valeria Souza for providing those from Puebla and Michoacan-Guanajuato; and J. Espíritu for technical and computational assistance. This work was supported by the grants from CONACyT (grant U4633), PAPIIT-UNAM (grant IN223005) and the Natural Environment Research Council.

L.L. was responsible for data analysis and manuscript preparation. L.L. and V.G. were responsible for the experimental design. I.H-G was responsible for bioinformatic

analysis. P.B. and R.I.S. were responsible for DNA sequencing. L.L., V.S., J.P.W.Y., G.D. and V.G. were responsible for discussion of the data.

REFERENCES

1. **Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Millar, W. and D. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-402.
2. **Amadou, C., Pascal, G., Mangenot, S., Glew, M., Bontemps, C., Capela, D., Carrere, S., Cruveiller, S., Dossat, C., Lajus, A., Marchetti, M., Poinsot, V., Rouy, B. Servin, Z., Saad, M., Schenowitz, C., Barbe, V., Batut, J., Medigue C. and C. Masson-Boivin.** 2008. Genome sequence of the beta-rhizobium Cupriavidus taiwanensis and comparative genomics of rhizobia. Genome Res. 18(9): 1472-83.
3. **Bailly, X., Olivieri, I., De Mita, S., Cleyet-Marel J. C. and G. Bena.** 2006. Recombination and selection shape the molecular diversity pattern of nitrogen-fixing Sinorhizobium sp. associated to Medicago. Mol. Ecol. 15(10): 2719-34.
4. **Chandler, M. and J. Mahillon**. 2002. Insertion Sequences revisited. Pp.305-366. In Craig, L. et al. (Eds). Mobile DNA II, ASM Press, Washington D. C.
5. **Crossman L. C., Castillo-Ramírez S., McAnnula C., Lozano L., Vernikos G. S., Acosta J. L., Ghazoui Z. F., Hernández-González I., Meakin G., Walker A. W., Hynes M. F., Young J. P., Downie J. A., Romero D., Johnston A. W., Dávila G., Parkhill J. And V. González.** 2008. A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria. PloS ONE 3(7): e2567.

6. **Doolittle, W. F. and C. Sapienza.** 1980. Sefish genes, the phenotype paradigm and genome evolution. Nature. 284:601-603.
7. **Edgar, R**. 2004.MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.
8. **Escobar-Páramo, P., Ghosh, S. and J. DiRuggiero**. 2005. Evidence for genetic drift in the diversification of a geographically isolated population of the hyperthermophilic archaeon *Pyrococcus*. Mol. Biol. Evol. 22 (11):2297-2303.
9. **Filée, J., Siguier, P. and M. Chandler**. 2007. Insertion sequence diversity in Archea. Microbiol. Mol. Biol. Rev. 71 (1): 121-157.
10. **Flores, M., Morales, L., Avila, M., Bustos, P., González, V., Garcia, D., Mora, Y., Guo, X., Collado-Vides, J., Piñero, D., Davila, G Mora, J. and R. Palacios**. 2005. Diversification of DNA sequences in the symbiotic genome of Rhizobium etli. J. Bacterial. 187(21):7185-92.
11. **Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal A. and X. Perret.** 1997. Molecular basis of symbiosis between Rhizobium and legumes. Nature 387(6631): 394-401.
12. **Galibert, F., Finan, T. M., Long, S. R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R. W., Dreano, S., Federspiel, N. A., Fisher, R. F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R. W., Jones, T., Kahn, D., Kahn, M. L., Kalman, S., Keating, D. H., Kiss, E., Komp, C., Lelaure, V., Masuy, D., Palm, C., Peck, M. C., Pohl, T. M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thebault, P., Vandenbol, M., Vorholter, F. J.,**

Weidner, S., Wells,D. H., Wong, K., Yeh K. C. and J. Batut. 2001. The composite genome of the legume symbiont Sinorhizobium meliloti. Science 293(5530): 668-72.

13. **Gasca, J**. Genética de poblaciones de *Rhizobium etli* asociado a *Phaseolus vulgaris* en dos localidades del bajío mexicano. Tesis de Licenciatura. Istituto de Ecología. UNAM. 102 pag.
14. **González, V., Santamaría, R., Bustos, P., Hernandez-González, I., Medrano-Soto, A., Moreno-Hagelsieb, G., Janga, S., Ramirez, M., Jiménez-Jacinto, V., Collado-Vides, J. and G. Davila**. 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. PNAS 103(10):3834-3839.
15. **González, V., Acosta, J.L., Santamaría, R., Bustos, P., Fernandez, J.L., Hernandez-González, I., Diza, R., Flores, M., Palacios, R., Mora, J. and G. Davila**. 2010. Conserved symbiotic plasmid DNA sequences in the Multireplicon pangenomic structure of Rhizobium etli. Appl. Environ. Microbiol. 76(5):1604-1614.
16. **Guindon, S. and O. Gascuel**. 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol Biol Evol. (4):534-43.
17. **Huson, D. H. and D. Bryant.** 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23(2): 254-67.
18. **Kalia, A., A. K. Mukhopadhyay, A. K., Dailide, G., Ito, Y., Azuma, T., Wong B. C. and D. E. Berg.** 2004. Evolutionary dynamics of insertion sequences in Helicobacter pylori. J. Bacteriol. 186(22): 7508-20.
19. **Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T.,**

Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada M. and S. Tabata. 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA. Res. 7(6): 331-8.

20. **Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., Watanabe, A., Idesawa, K., Iriguchi, M., Kawashima, K., Kohara, M., Matsumoto, M., Shimpo, S., Tsuruoka, H., Wada, T., Yamada M. and S. Tabata.** 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110. DNA. Res. 9(6): 189-97.
21. **Kidwell, M.G., and D. R. Lisch.** (2002) Transposable elements as sources of genomic variation. Pp.305-366. In Craig, L. et al. (Eds). Mobile DNA II, ASM Press, Washington D. C.
22. **Korber B**. 2000. HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers.
23. **Kurtz, S., Phullippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C. and S. Salzberg**. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.
24. **Kwak, M. and P. Gepts.** 2009. Structure of genetic diversity in the two major gene pools of common bean (Phaseolus vulgaris L., Fabaceae). Theor. Appl. Genet. 118(5): 979-92.

25. **Lenski, R., Winkwirth, C., and M. Riley**. 2003. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations, J. Mol. Evol. 56:498-508.
26. **Librado, P. and J. Rozas.** 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11): 1451-2.
27. **Mahillon, J., Léonard, C., and M. Chandler**. 1999. IS elements as constituents of bacterial genomes. Res Microbiol. 150:675-687.
28. **Martin, D. P., Williamson, C. and D. Posada.** 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21(2): 260-2.
29. **Papadopoulus. D., Schneider, D., Meier-Eis, J., Arber, W., Lenski, R. E. and Blot, M**. et al. 1999. Genomic evolution during a 10,000-generation experiement with bacteria. Proc. Natl. Acad. Sci. 96:3807-3812.
30. **Perez-Mendoza, D., Dominguez-Ferreras, A., Munoz, S., Soto, M. J., Olivares, J., Brom, S., Girard, L., Herrera-Cervera J. A. and J. Sanjuan.** 2004. Identification of functional mob regions in Rhizobium etli: evidence for self-transmissibility of the symbiotic plasmid pRetCFN42d. J. Bacteriol. 186(17): 5753-61.
31. **Posada, D. and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9): 817-8.
32. **Ramsay, J. P., Sullivan, J. T., Stuart, G. S., Lamont I. L. and C. W. Ronson.** 2006. Excision and transfer of the Mesorhizobium loti R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS. Mol. Microbiol. 62(3): 723-34.

33. **Rodríguez-Navarro, D**., Buendia, A. Camacho, M., Lucas M. and C. Santamaria. 2000. Characterization of *Rhizobium* spp. Bean isolates from south-west Spain. Soil Biol. Biochem. 32:1601-1613.

34. **Schneider, D., and R. Lenski**. 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria. Res. Microbiol. 155:319-327.

35. **Siguier, P., Filée, J., and M. Chandler**. 2006. Insertion sequences in prokaryotic genomes. Curr. Opin. Microbiol. 9:1-6.

36. **Silva, C., Vinuesa, P., Eguiarte, L. Martínez-Romero, E. and V. Souza**. 2003. *Rhizobium etli* and *Rhizobium gallicum* nodulate common bean (*Phaseolus vulgaris*) in a traditionally managed milpa plot in Mexico: Population genetics and Biogeographic implications. Appl. Env. Microbiol. 69:884-893.

37. Silva, C., Vinuesa, P., Eguiarte, L., 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.

38. Slatkin, M. 1985. Genetic differentiation of transposable elements under mutation and unbiased gene conversion. Genetics 110(1): 145-58.

39. Sullivan, J. T. and C. W. Ronson. 1998. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc. Natl. Acad. Sci. U S A 95(9): 5145-9.

40. Sullivan, J. T., Trzebiatowski, J. R., Cruickshank, R. W., Gouzy, J., Brown, S. D., Elliot, R. M., Fleetwood, D. J., McCallum, N. G., Rossbach, U., Stuart, G. S., Weaver, J. E., Webby, R. J., De Bruijn F. J. and C. W. Ronson. 2002.

Comparative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A. J. Bacteriol. 184(11): 3086-95.

41. **Touchon, M., and Rocha, E**. 2007. Causes of insertion sequences abundance in prokaryotic genomes. Mol. Biol. Evol. 4:969-981.

42. **Young, J. P., Crossman, L. C., Johnston, A. W., Thomson, N. R., Ghazoui, Z. F., Hull, K. H., Wexler, M., Curson, A. R., Todd, J. D., Poole, P. S., Mauchline, T. H., K. East, M. A., Quail, A., Churcher, C., Arrowsmith, C., Cherevach, I., Chillingworth, T., Clarke, K., Cronin, A., Davis, P., Fraser, A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Sanders, M., Simmonds, M., Whitehead S. and J. Parkhill.** 2006. The genome of Rhizobium leguminosarum has recognizable core and accessory components. Genome Biol 7(4): R34.

43. **Wagner, A**. 2006. Periodic extinctions of transposable elements in bacterial lineales: Evidence from intragenomic variation in multiple genomes. Mol. Biol. Evol. 23 (4):723-733.

44. **Zeigler, D. R.** 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int. J. Syst. Evol. Microbiol. 53(Pt 6): 1893-900.

Figure 1. Phylogenetic reconstruction representing the evolutionary relationships across the three collections. The circles inside the branches represent bootstrap support >70%. The external circles represent *R. etli* strains intermingled from one of the other collections. The arrows represent other species intermingled (4872, GR42 and GR60 are *R. gallicum*; GR93 is *R. giardinii*; and GR84 is *R. leguminosarum*). The underlined strains represent recombinant strains for dnaB or glyA between the three populations and the *R. gallicum* strain.

Figure 2. Presence/absence profile of the IS elements present in the chromosome, conjugative and symbiotic plasmids across the three populations. The bottom line represents CFN42 control strain. Red bars represent positive confirmation of the IS element (a DNA fragment of similar size to that in CFN42), blue bars represent a negative result for the presence of the IS element (a DNA fragment equivalent to the distance between the two sequences near the insertion site but lacking the IS), no colour represents no PCR product and yellow bars represent DNA fragments bigger in size than expected.

Figure 3. Dot plot graphic of pSym from *R. etli* CFN42 and CIAT652. Diagonal red lines represent DNA regions that were aligned between the two replicons. The blue and red points represent small DNA regions, in many cases repeated sequences, that were aligned in different regions of each replicon.

Table 1. Rhizobium isolates and reference strains used in this study

Table 2. DNA divergence in *R. etli* IS elements, plasmid genes and chromosomal genes

Table 3. Comparision of mean genetic divergence between populations (Dxy) and mean genetic divergence ($\boldsymbol{\pi}$) within each population

Table 4. Genetic differentiation and gene flow tests in *R. etli* IS elements, plasmid genes and chromosomal genes

Supp. Table 1. Complete set of sequences obtained from PCR reactions

Supp. Table 2. DNA divergence and neutrality test in *R. etli* IS elements, plasmid genes and chromosomal genes within populations

Supp. Table 3. Genetic differentiation analyses of *R. etli* IS elements, plasmid genes and chromosomal genes

Supp. Table 4. Recombination events in plasmid and chromosomal genes between populations. Each row represents a different recombination event for each gene. The underlined strains represent recombinant strains for *dnaB* or *glyA* between the three populations and the *R. gallicum* strains that could explain the intermingled strains in the phylogeny reconstruction of the three populations (Fig. 1).

Supp. Figure 1. Phylognetic network analises with Neighbor Net for a) *dnaB* and b) *glyA*. Underlined labels represent strains participating in recombination events determined with RDP3 (Supp. Table 4): bold labels represent recombinant strains and italic labels represents parental strains.

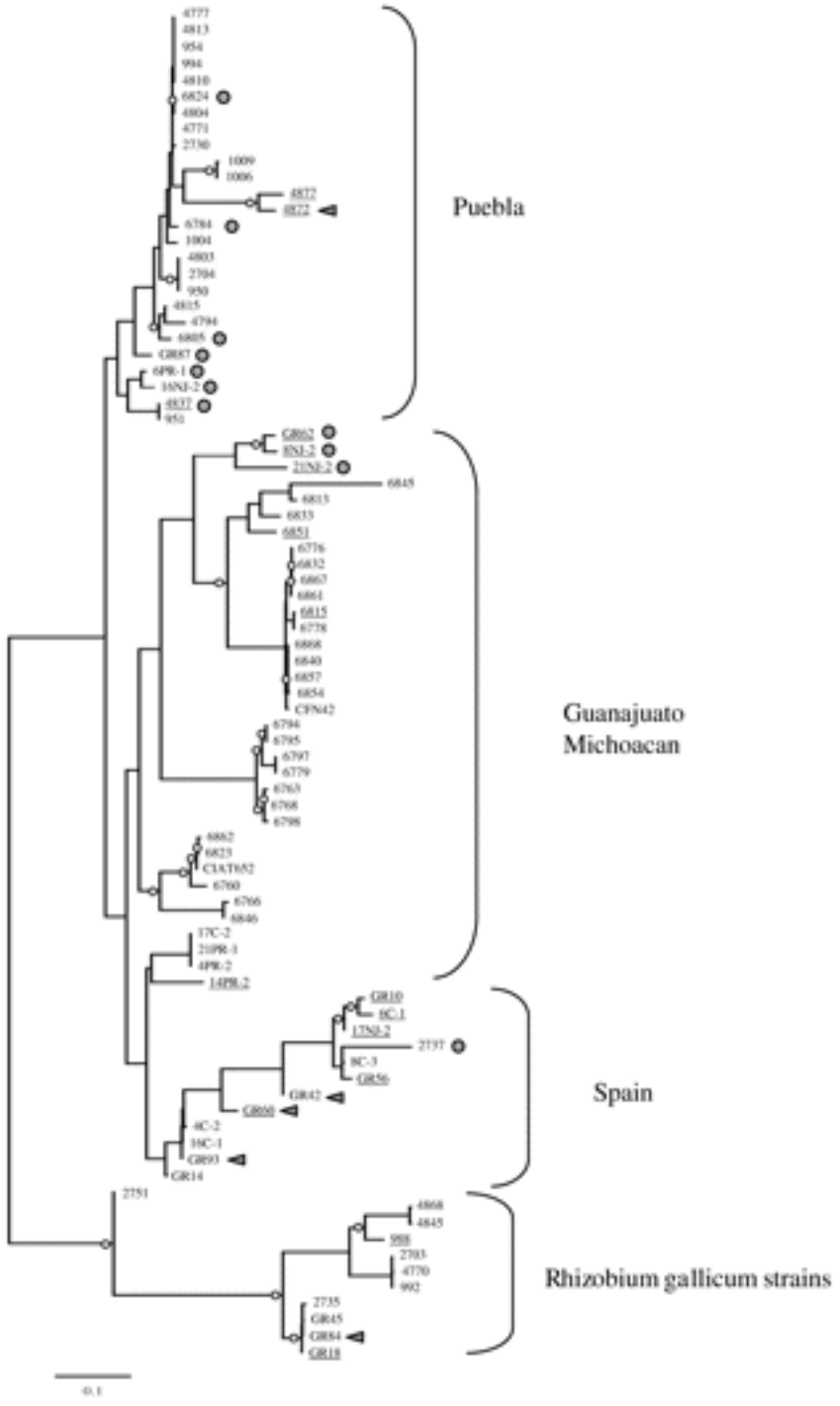
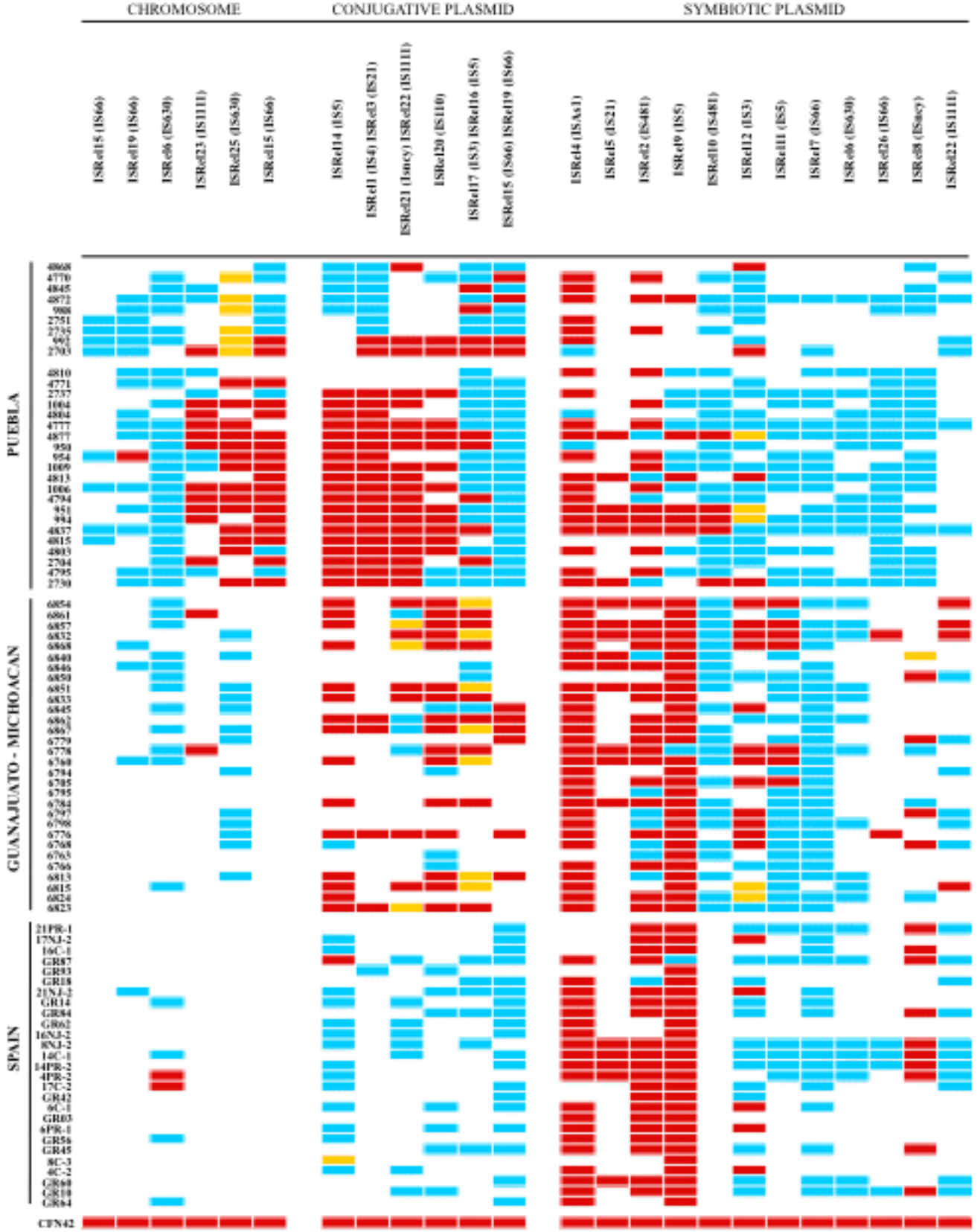

4777
4813
954
994
4810
6824
4804
4771
2730
1009
1006
4877
4822
6784
1004
4803
2704
950
4815
4794
6805
GR87
6PR-1
16N3-2
4837
951
Puebla
GR62
8N3-2
21N2-2
6845
6835
6833
6851
6776
6832
6867
6861
6815
6778
6868
6840
6857
6854
CFN42
6794
6795
6797
6779
6763
6768
6798
Guanajuato
Michoacan
6862
6823
CIAT652
6760
6766
6846
17C-2
21PR-1
4PR-2
14PR-2
GR10
8C-1
17N2-2
2737
8C-3
GR56
GR42
GR66
8C-2
16C-1
GR93
GR14
Spain
2751
4868
4845
958
2703
4770
992
2735
GR45
GR84
GR18
Rhizobium gallicum strains
0.1

Figure 1. Phylogenetic reconstruction representing the evolutionary relationships across the three collections. The circles inside the branches represent bootstrap support >70%. The external circles represent *R. etli* strains intermingled from one of the other collections. The arrows represent other species intermingled (4872, GR42 and GR60 are *R. gallicum*; GR93 is *R. giardinii*; and GR84 is *R. leguminosarum*). The underlined strains represent recombinant strains for dnaB or glyA between the three populations and the *R. gallicum* strain.



Puebla	0	1	0	11	14	18	19	19	17	11	5	0	20	6	14	6	5	4	0	0	0	0	0	0
G-M	0	0	0	2	0	0	15	4	6	16	8	6	28	9	18	29	0	12	7	0	0	2	4	4
Spain	0	0	2	0	0	0	1	0	0	0	0	0	20	5	20	24	0	5	0	0	0	0	10	0
TOTAL	**0**	**1**	**2**	**32**	**14**	**18**	**35**	**23**	**23**	**27**	**13**	**6**	**68**	**20**	**51**	**59**	**5**	**21**	**7**	**0**	**0**	**2**	**14**	**4**

Figure 2. Presence/absence profile of the IS elements present in the chromosome, conjugative and symbiotic plasmids across the three populations. The bottom line represents CFN42 control strain. Red bars represent positive confirmation of the IS element (a DNA fragment of similar size to that in CFN42), blue bars represent a negative result for the presence of the IS element (a DNA fragment equivalent to the distance between the two sequences near the insertion site but lacking the IS), no colour represents no PCR product and yellow bars represent DNA fragments bigger in size than expected.

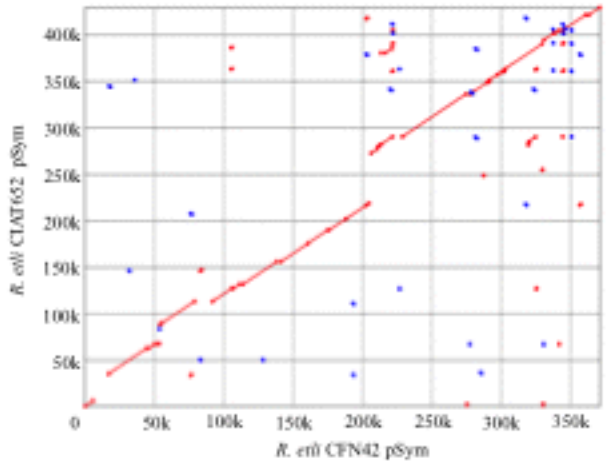

400k
350k
300k
250k
200k
150k
100k
50k
0
50k
100k
150k
200k
250k
300k
350k
R. etli CIAT652 pSym
R. etli CFN42 pSym

Figure 3. Dot plot graphic of pSym from *R. etli* CFN42 and CIAT652. Diagonal red lines represent DNA regions that were aligned between the two replicons. The blue and red points represent small DNA regions, in many cases repeated sequences, that were aligned in different regions of each replicon.

Table 1. Rhizobium isolates and reference strains used in this study

Origin	No. Strains	Species	Identifier	Reference
Rhizobium Isolates				
Puebla, Mexico	30	Rhizobium etli	1009, 1006, 4771, 1004, 2737, 4815, 4803, 4777, 950, 2704, 4813, 4794, 951, 994, 4837, 4804, 954, 4877, 4795, 2730, 4810	(34)
		Rhizobium gallicum	4868, 4872, 4770, 4845, 988, 2751, 2735, 992, 2703	
Guanajuato, Mexico	30	Rhizobium etli	Domesticated: 6854, 6861, 6857, 6832, 6868, 6867, 6840, 6846, 6850, 6851, 6833, 6845, 6862	(12)
			Wild: 6779, 6778, 6760, 6794, 6805, 6795, 6766, 6784, 6797, 6798, 6776, 6768, 6763	
			Weedy: 6815, 6824, 6823, 6813	
Valle Guadalquivir, Spain	27	Rhizobium etli	21PR-1, 16C-1, 21NJ-2, 8C-3, 8NJ-2, 14PR-2, 17NJ-2, 16NJ-2, 14C-1, 4PR-2, 17C-2, 6PR-1, 6C-1, 4C-2, GR10, GR62, GR14, GR87, GR56	(31)
		Rhizobium gallicum	GR18, GR45, GR60, GR42	
		Rhizobium giardinni	GR93, GR03	
		Rhizobium fredii	GR64	
		Rhizobium leguminosarum	GR84	
Rhizobium Reference Strains				
Mexico	1	Rhizobium etli	CFN42	(13)
Costa Rica	1	Rhizobium etli	CIAT652	(9)

Gene	Polymorphic sites[a]	dn/ds	Θ	Nucleotide diversity π	Nucleotide diversity at synonymous sites π_S	Nucleotide diversity at nonsynonymous sites π_{NS}	Tajima´s D
dnaB	263 (23.6)	0.019	0.047	0.066	0.237	0.009	1.39
glyA	278 (24.7)	0.029	0.070	0.088	0.122	0.075	0.84
nodC	19 (3.3)	0.355	0.007	0.014	0.028	0.009	0.72
ISRel2	6 (0.6)	0.267	0.001	0.002	0.001	0.002	2.72
ISRel4	13 (1.3)	0.325	0.002	0.002	0.003	0.007	1.50

[a] The percentages of polymorphic sites per sequence length are shown in parentheses.

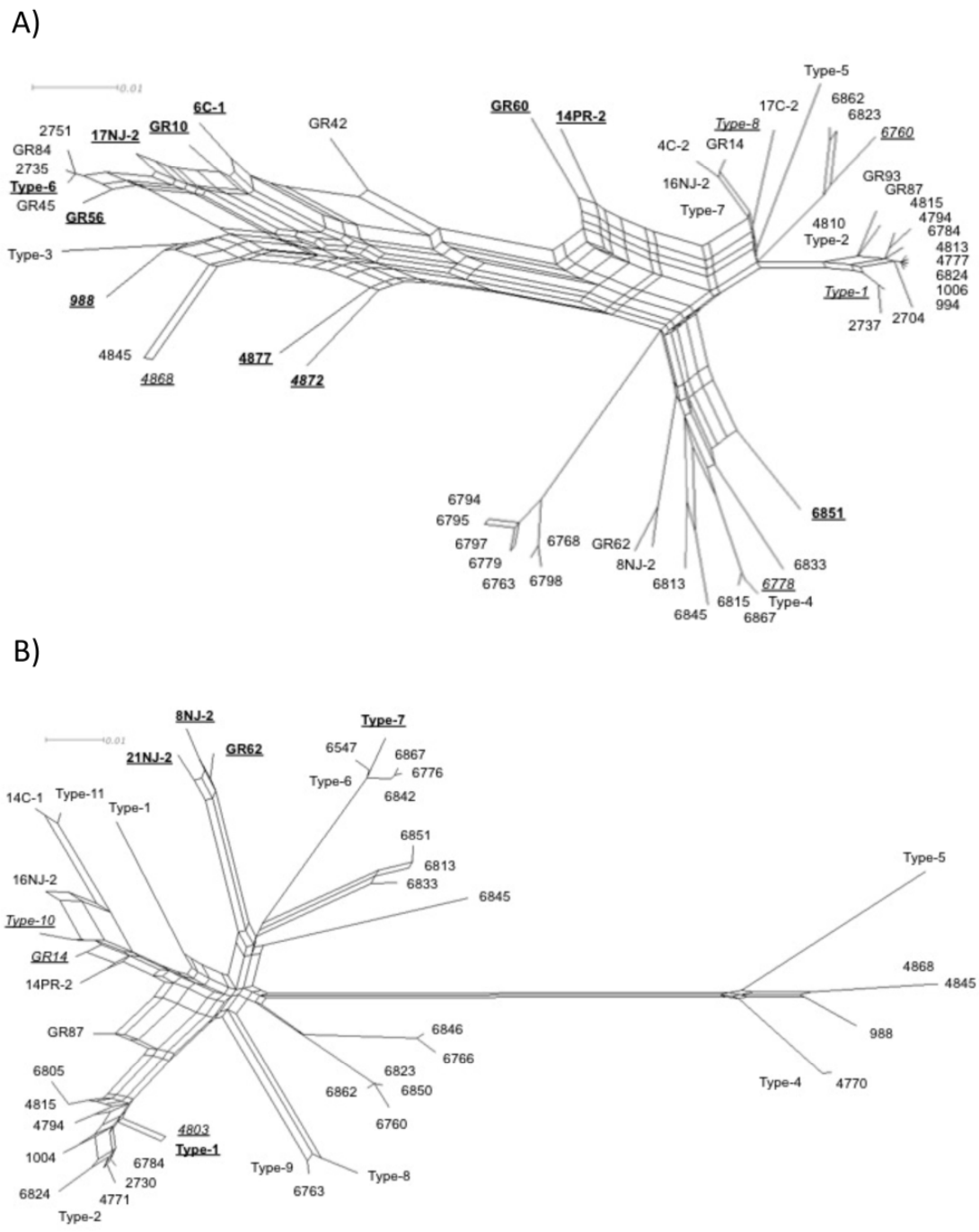
Table 2. DNA divergence in *R. etli* IS elements, plasmid genes and chromosomal genes

	Dxy	π
dnaB		
Pueb - Guanj	0.072	0.056 - 0.052
Pueb - Spain	0.068	0.056 - 0.064
Guanj - Spain	0.073	0.052 - 0.064
glyA		
Pueb - Guanj	0.097	0.080 - 0.066
Pueb - Spain	0.100	0.080 -0.077
Guanj - Spain	0.089	0.066 - 0.077
ISRel2		
Pueb - Guanj	0.003	0.002 - 0.002
Pueb - Spain	0.002	0.002 - 0.001
Guanj - Spain	0.002	0.002 - 0.001
ISRel4		
Pueb - Guanj	0.002	0.001 – 0.003
Pueb - Spain	0.002	0.001 – 0.002
Guanj - Spain	0.002	0.003 - 0.002
nodC		
Pueb - Guanj	0.003	0.001 – 0.005
Pueb - Spain	0.026	0.001 – 0.008
Guanj - Spain	0.025	0.005 - 0.008

Table 3. Comparison of mean genetic divergence between populations (Dxy) and mean genetic divergence (π) within each population

Gene	K_{ST}	F_{ST}	Nm
dnaB			
Pueb - Guanj	0.146***	0.252	1.48
Pueb - Spain	0.064***	0.118	3.72
Guanj - Spain	0.118***	0.208	1.90
glyA			
Pueb - Guanj	0.144***	0.249	1.51
Pueb - Spain	0.142***	0.246	1.54
Guanj - Spain	0.129***	0.226	1.71
ISRel2			
Pueb - Guanj	0.007	0.025	19.75
Pueb - Spain	0.652**	0.845	0.09
Guanj - Spain	0.607**	0.750	0.17
ISRel4			
Pueb - Guanj	0.056	0.217	1.80
Pueb - Spain	0.310**	0.634	0.29
Guanj - Spain	0.156**	0.270	1.35
nodC			
Pueb - Guanj	0.184*	0.374	0.84
Pueb - Spain	0.118*	0.258	1.44
Guanj - Spain	0.225**	0.357	0.90

Table 4. Genetic differentiation and gene flow tests in *R. etli* IS elements, plasmid genes and chromosomal genes. Asterisks represent the probability obtained by the permutation test with 1000 replicates (ns, not significant; *, 0.01<P<0.05; **, 0.001<P<0.01; ***, P<0.001).



Supp. Figure 1. Phylogenetic network analyses with Neighbor Net for a) *dnaB* and b) *glyA*. Underlined labels represent strains participating in recombination events determined with RDP3 (Supp. Table 4): bold labels represent recombinant strains and italic labels represents parental strains.

Gene	No. of sequences	Sequence length (bp)	Puebla	Guanajuato	Spain
dnaB	87	1113	30	30	25
glyA	88	1125	29	31	28
ISRel2	45	981	5	16	20
ISRel4	53	1035	6	23	21
nodC	44	576	7	22	15

Supp. Table 1. Complete set of sequences obtained from PCR reactions

Gene	Θ	Nucleotide diversity π	Nucleotide diversity at synonymous sites π_S	Nucleotide diversity at nonsynonymous sites π_{NS}	Tajima´s D
dnaB	**0.047**	**0.066**	**0.237**	**0.009**	1.39
Puebla	0.040	0.056	0.191	0.011	1.48
Guanajuato	0.040	0.052	0.194	0.005	1.08
Spain	0.042	0.064	0.227	0.010	2.15*
glyA	**0.070**	**0.088**	**0.122**	**0.075**	0.84
Puebla	0.065	0.080	0.118	0.066	0.88
Guanajuato	0.057	0.066	0.091	0.057	0.57
Spain	0.070	0.077	0.098	0.061	0.07
nodC	**0.007**	**0.014**	**0.028**	**0.009**	2.72**
Puebla	0.001	0.001	0.002	0	-1.01
Guanajuato	0.007	0.005	0.012	0.002	-1.16
Spain	0.010	0.008	0.015	0.005	-0.87
ISRel2	**0.001**	**0.002**	**0.001**	**0.002**	0.72
Puebla	0.001	0.002	0.002	0.001	0.70
Guanajuato	0.001	0.002	0.001	0.002	0.29
Spain	0.001	0.001	0	0.001	0.52
ISRel4	**0.002***	**0.002**	**0.003**	**0.002**	-0.62
Puebla	0.001	0.001	0	0.001	-1.13
Guanajuato	0.003	0.003	0.005	0.002	-0.14
Spain	0.002	0.002	0.003	0.001	-0.65

Supp. Table 2. DNA divergence and neutrality test in *R. etli* IS elements, plasmid genes and chromosomal genes within populations.

Gene	K_S	K_S*	Z	Z*	S_{nn}	PM test P-value
dnaB						
Pueb - Guanj	59.84	3.53	786.71	6.22	0.95	<0.001
Pueb - Spain	66.33	3.65	699.60	6.14	0.88	<0.01
Guanj - Spain	63.80	3.79	672.31	6.10	0.93	<0.001
glyA						
Pueb - Guanj	56.90	3.52	752.71	6.18	0.93	<0.001
Pueb - Spain	59.18	3.37	682.83	6.08	0.92	<0.001
Guanj - Spain	53.62	3.52	729.46	6.15	0.86	<0.001
ISRel2						
Pueb - Guanj	2.19	0.58	205.22	5.20	0.62	<0.05
Pueb - Spain	3.05	0.58	79.22	4.22	0.82	<0.05
Guanj - Spain	3.42	0.71	207.62	5.14	0.89	<0.01
ISRel4						
Pueb - Guanj	2.73	1.14	199.90	5.04	0.76	<0.05
Pueb - Spain	2.13	0.90	120.35	4.52	0.81	<0.01
Guanj - Spain	2.82	1.12	404.66	5.64	0.78	<0.01
nodC						
Pueb - Guanj	1.63	0.78	90.11	4.27	0.71	ns
Pueb - Spain	1.13	0.52	138.43	4.76	0.88	<0.001
Guanj - Spain	1.29	0.60	273.51	5.40	0.69	<0.001

Supp. Table 3. Genetic differentiation analyses of *R. etli* IS elements, plasmid genes and chromosomal genes

Gene	Recombinant strains	Parental strains	Methods	P-value
dnaB	4877, 4872, 988	4868 - 4795	3	0.564
	6851	6760 - 6778	4	0.509
	14PR-2, GR60	6PR-1 - 4872	3	0.634
	GR18, 17NJ-2, GR10, 6C-1, GR56	988 – 6PR-1	4	0.621
glyA	4837	4803 – GR42	2	0.479
	6815, 8NJ-2, 21NJ-2, GR62	GR14 - 4803	2	0.510
ISRel4	954	GR87 - 4795	4	0.549

Supp. Table 4.

Recombination events in plasmid and chromosomal genes between populations. Each row represents a different recombination event for each gene. The underlined strains represent recombinant strains for dnaB or glyA between the three populations and the *R. gallicum* strains that could explain the intermingled strains in the phylogeny reconstruction of the three populations (Fig. 1).