The partitioned *Rhizobium etli* genome: Genetic and metabolic redundancy in seven interacting replicons

Víctor González*, Rosa I. Santamaría, Patricia Bustos, Ismael Hernández-González, Arturo Medrano-Soto, Gabriel Moreno-Hagelsieb[†], Sarath Chandra Janga, Miguel A. Ramírez, Verónica Jiménez-Jacinto, Julio Collado-Vides, and Guillermo Dávila*

Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, AP565-A Cuernavaca, Morelos, 62210, México

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved January 6, 2006 (received for review September 28, 2005)

We report the complete 6,530,228-bp genome sequence of the symbiotic nitrogen fixing bacterium *Rhizobium etli*. Six large plasmids comprise one-third of the total genome size. The chromosome encodes most functions necessary for cell growth, whereas few essential genes or complete metabolic pathways are located in plasmids. Chromosomal synteny is disrupted by genes related to insertion sequences, phages, plasmids, and cell-surface components. Plasmids do not show synteny, and their orthologs are mostly shared by accessory replicons of species with multipartite genomes. Some nodulation genes are predicted to be functionally related with chromosomal loci encoding for the external envelope of the bacterium. Several pieces of evidence suggest an exogenous origin for the symbiotic plasmid (p42d) and p42a. Additional putative horizontal gene transfer events might have contributed to expand the adaptive repertoire of *R. etli*, because they include genes involved in small molecule metabolism, transport, and transcriptional regulation. Twenty-three putative sigma factors, numerous isozymes, and paralogous families attest to the metabolic redundancy and the genomic plasticity necessary to sustain the lifestyle of *R. etli* in symbiosis and in the soil.

multireplicon genome | rhizobiales | symbiosis | horizontal transfer

The genomes of several α-proteobacteria are partitioned into replicons of variable size (1). Circular chromosomes and large plasmids are common, with the addition of a linear chromosome in *Agrobacterium tumefaciens* (2). Accessory replicons generally have the RepABC replication system, but their origin remains largely unknown (3). The α-proteobacteria subdivision comprises several nitrogen-fixing symbiotic species, commonly known as rhizobia, grouped in different families (4). Complete genome sequences are currently available for the symbionts *Mesorhizobium loti* MAFF3030 (Phylobacteriaceae), *Sinorhizobium meliloti* 1021 (Rhizobiaceae), and *Bradyrhizobium japonicum* (Bradyrhizobiaceae) (5–7) and for the closely related plant pathogen *A. tumefaciens* C58 (Rhizobiaceae) (8, 9). Less than 1% of the total gene content of these rhizobia has been associated with symbiosis, and few of them are ubiquitous (5–7, 10–13). The majority of the symbiosis-related genes are located in plasmids or in chromosomal islands, probably acquired by horizontal transfer (10–14).

One of the most important crops in Latin America is the common bean, which is mainly nodulated by *Rhizobium etli* (15). Previously, we reported the full sequence of the symbiotic plasmid of *R. etli* CFN42 (10). In this work, we report the complete genome sequence of this bacterium that includes a circular chromosome and six plasmids. Despite the structural partition of the *R. etli* genome, we infer functional relationships among replicons, indicating that plasmids are not completely dispensable elements. Multipartite genomes like that of *R. etli* might enhance the adaptive potential of the bacterium, allowing the reassortment of essential, nonessential, and redundant functions to contend with challenging environments.

Results and Discussion

General Features. The genome of *R. etli* CFN42 (hereafter referred to as *R. etli*) consists of a circular chromosome and six large plasmids. The size and main characteristics of each replicon are shown in Table 1. The average guanine–cytosine (GC) content is 61.5%, but the plasmids p42a and p42d (the symbiotic plasmid) show lower GC values (58%). The chromosome presents several regions with low or high GC content that harbor insertion sequences (ISs), phages, or genes from plasmid origin (Fig. 1). Three low GC regions are occupied by identical ribosomal RNA operons that include genes for two tRNAs (tRNA-ala and tRNA-ile). Forty-four additional genes dispersed throughout the chromosome provide the complete set of tRNAs. Approximately 71% of the 6,034 predicted coding sequences (CDS) have homologs with known or putative function, whereas the remaining 29% represent hypothetical-conserved (23%) and ORFan genes. The average length of genes with annotated function is 1,125 bp, whereas that of hypotheticals is 600 bp. The chromosomal GC skew suggests probable sites for the initiation and termination of replication. The *parAB-gidB* locus was located at the putative *Ori*, whereas *dnaA* maps 300 kb downstream. Several probable binding sites for proteins involved in replication (i.e., CtrA, Fis, IHF, and DnaA) are present within this region, supporting the allocation of the origin of replication. Plasmids show no detectable GC skew.

DNA Reiterations and Paralogous Families. On the basis of DNA-DNA hybridization experiments, it was predicted that in *R. etli*, there are ≈200 reiterated DNA families (17) that can recombine, leading to genomic rearrangements (18). The complete genome sequence of *R. etli* reveals 133 families of identical repeats that are >100 nucleotides. Most of these repeats lie in the plasmids p42a and p42d, whereas there are few in plasmids p42b, p42c, and p42e (Fig. 3, which is published as supporting information on the PNAS web site). As previously shown with pNGR234a, several potential rearrangements may be generated by homologous recombination among these reiterated sequences (19). The 27 major reiterated families found in *R. etli* are composed of two to six identical elements with a minimal length of 534 nucleotides. In comparison, *S. meliloti* shows 24 families of 2–15 elements, and *A. tumefaciens* contains 7 families of 2–4 elements.

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: BDBH, bidirectional best hit; CDS, coding sequences; COG, clusters of orthologous groups; EPS, exopolysaccharide; GC, guanine–cytosine; IS, insertion sequence.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accessions nos. CP000133–CP000138).

*To whom correspondence may be addressed. E-mail: vgonzal@ccg.unam.mx or davila@ccg.unam.mx.

[†]Present address: Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada N2L 3C5.

Table 1. General features of the *R. etli* genome

Features	p42a	p42b	p42c	p42d	p42e	p42f	Chr	Genome
Size, bp	194,229	184,338	250,948	371,254	505,334	642,517	4,381,608	6,530,228
GC average, %	58.00	61.81	61.52	58.35	61.67	61.22	61.27	60.54
rRNA operons	—	—	—	—	—	—	3	3
tRNAs	—	—	—	—	—	—	50	50
Total CDS	182	165	234	354	459	573	4,067	6,034
CDS in functional classes	125	138	178	237	317	403	2,892	4,287 (71%)
Hypothetical CDS	44	23	45	60	118	134	962	1,389 (23%)
Orphans	13	4	11	57	24	36	213	358 (6%)
Transcriptional regulators	5	10	24	14	49	75	359	536
Transporters	25	45	65	37	81	108	476	837
External origin	52	2	1	58	—	—	44	157
Sigma subunits	—	—	1	1	2	4	15	23

In both *R. etli* and *S. meliloti*, the largest fraction of these repeats consist of ISs. The complete ISs of *R. etli* belong to the IS66 (12 copies), IS630 and IS5 (5 copies each), and IS211 (4 copies) families. Other families such as IS3, IS4, and IS110 account for the rest of the ISs (13 copies). We also found 42 incomplete ISs belonging to diverse families. There are no ISs interrupting ORFs, and the 53 pseudogenes found are either truncated or contain frameshifts. The prevalence of IS66 in *R. etli* contrasts with its poor representation in other rhizobia such as *S. meliloti* (two copies) and *B. japonicum* (three copies). Other identical reiterations include the genes *ccmF*, *hemA*, *purU*, *adhC*, *etfA*, *tufA*, *tufB*, *nifHDK*, and the three ribosomal operons, which are the largest identical repeats found in this genome.

R. etli contains 1,652 paralogous genes sorted out in 462 families, which range from 2 to 129 members and comprise 27% of the total gene content. The families with the largest number of members belong to ABC transporters of carbohydrates and amino acids [clusters of orthologous groups (COGs) E and G], transcription factors (COG K), and a variety of genes involved in small molecule metabolism (Fig. 4, which is published as supporting information on the PNAS web site). These families are distributed throughout the chromosome and plasmids, but some paralogous are restricted to single replicons.

COG Distribution Among Replicons. To make a functional distinction among *R. etli* replicons, we classified protein-coding genes into COGs (20). We provide a COG for 4,425 of 6,034 genes. Similar to *S. meliloti* and *A. tumefaciens*, in *R. etli*, several COG categories are overrepresented, namely carbohydrate transport and metabolism, amino acid transport and metabolism, and transcription (COGs G, E, and K, respectively). Conversely, genes related to nucleotide transport and metabolism; translation, ribosomal structure, and biogenesis; cell wall and membrane biogenesis; and posttranslational modification (COGs F, J, M, and O) are absent from plasmids (Table 4, which is published as supporting information on the PNAS web site). Genes related to defense mechanisms (COG V), expected to predominate in plasmids, are most common in the chromosome. This latter category includes several ABC transporters, efflux pumps, and diverse proteins for antibiotic resistance and other toxic compounds. Furthermore, genes coding for ISs (COG L) and for transport systems type III and IV (TSSIII and IV) (COG U) are mainly located in p42a and p42d (pSym). The chromosome also harbors many ISs but such elements are absent in the other four plasmids. No essential genes were located in plasmids, except for the *minCDE* operon, located in p42e, whose protein products are involved in cell division (COG D). This observation explains the unsuccessful recovery of *R. etli* strains cured from p42e (21).

Comparative Genomics. Extensive synteny at the nucleotide level was found among the *R. etli* chromosome and the chromosome of *S. meliloti*, the circular chromosome of *A. tumefaciens*, the chromosome I of *Brucella* species, and in lesser extent, with the chromosomes of *M. loti* and *B. japonicum* (Fig. 5, which is published as supporting information on the PNAS web site). To measure more precisely the genomic similarity between *R. etli* and the rhizobial species sequenced so far, we identified the set of probable orthologs shared between pairs of genomes as bidirectional best hits (BDBHs) (Table 5, which is published as supporting information on the PNAS web site). Most protein-coding genes (>43%) have orthologs with *S. meliloti*, *A. tumefaciens*, *M. loti*, and *B. japonicum*. The great majority of these genes are located in the chromosome of *R. etli* with <24% distributed in the six plasmids. An important part of the *R. etli* chromosomal orthologs (62–72%) are organized in syntenic arrays of two or more consecutive genes (Table 6, which is published as supporting information on the PNAS web site) showing an alternate pattern of microsyntenic, variable, and specific regions in relation to the chromosomes of Rhizobiales (Fig. 1). Housekeeping genes are located into conserved blocks, whereas the variable and species-specific regions contain genes encoding outer-surface components and mobile elements, among others. Common syntenic blocks across the genomes compared likely reflect the evolutionary backbone preserved since the original branching of the Rhizobial clades.

Genomic comparisons among the three Rhizobiacea *R. etli*, *S. meliloti*, and *A. tumefaciens* indicate that 23.3% of the combined set of proteins (10,877) are shared by the three species, an extra 14.4% are present in only two of three, and the rest are unique to each compared species, namely 18%, 21%, and 16%, respectively (Fig. 2). The intersection between *R. etli* and *S. meliloti* contains several known symbiotic proteins and other gene products (e.g., a family of 15 adenylate guanylate cyclases, a duplication of the NADH-ubiquinone oxidoreductase complex, the alternative terminal oxidase *coxMN0P*, several sigma factors, and a family of eight glutathione-S-reductase genes). These proteins are not present in *A. tumefaciens*, which shares a very different set of orthologs with *R. etli* and *S. meliloti* (Table 7, which is published as supporting information on the PNAS web site). The orthologs distribution among these Rhizobiacea suggests that severe differential DNA losses, duplications, and acquisitions have played a major role in their evolution.

Plasmid Comparisons. The six plasmids of *R. etli* CFN42 have a RepABC replication system. Plasmids p42a and p42f have an additional copy of the *rep* operon. Most of the orthologs found in the *R. etli* plasmids are shared with the accessory replicons pSymA and pSymB of *S. meliloti*, the linear chromosome of *A. tumefaciens*, and the chromosome II of *Brucella* (Fig. 1, red

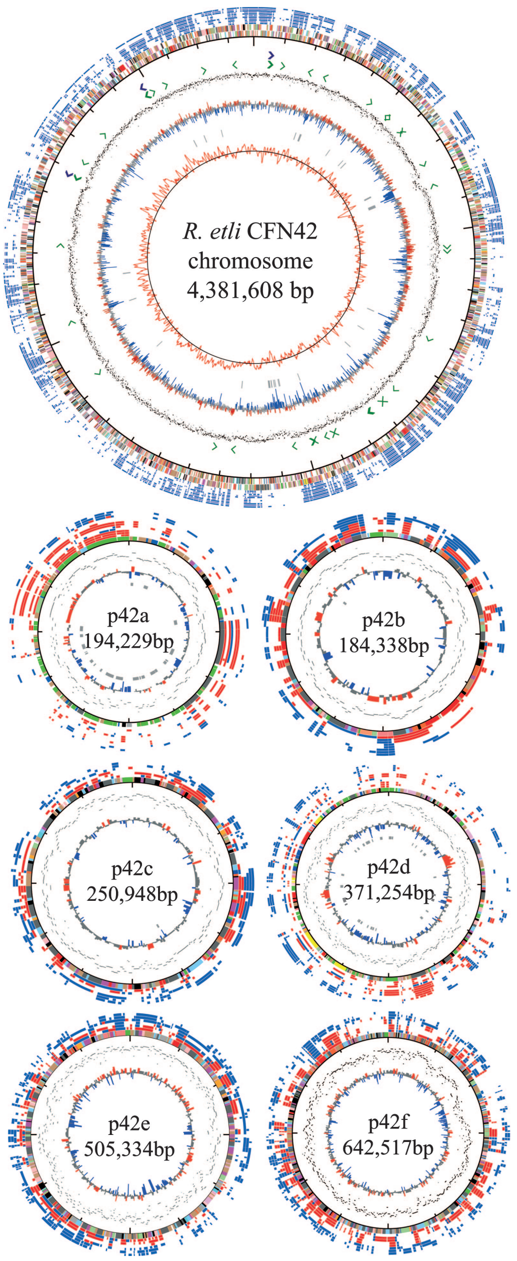


Fig. 1. General features and comparison of the *R. etli* genome with complete genomes of species of the order Rhizobiales. For the chromosome, only matches involving two or more adjacent BDBHs against the chromosome of the other species are shown. For plasmids, all of the BDBHs are shown. In the comparisons, blue indicates chromosomal matches and red designates the matches with accessory elements (plasmids or secondary chromosomes). Descriptions for the chromosome are presented from the innermost circle outward: GC skew, IS and phage integrases, GC content (blue, low GC; gray, medium GC; red, high GC), codon usage as measured by CRI (16), tRNAs, rRNAs, scale (each segment represents 182,331 bp), and predicted CDSs on the reverse strand and forward strand in color code (see below). The outermost blue rings show the shared BDBHs between *R. etli* and the circular chromosome of the following Rhizobial species: *S. meliloti*; *A. tumefaciens* C58 U Washington; *A. tumefaciens* C58 Cereon; *B. melitensis*; *B. suis*; *B. abortus*, *M. loti*; *B. japonicum*, *R. palustris*, and *B. quintanae*; and *B. henselae* in concentric circles from the innermost circle. In p42a and p42b, from the innermost circles outward: IS and phage integrases; GC content; codon usage (CRI); scale; color-coded CDS; shared BDBHs with complete sequences of plasmids pRi1724 (*A. rhizogenes*), pNGR234a (*Rhizobium* spp), pTi Sakura (*A. tumefaciens*), and symbiotic island of *M. loti* R7A; shared BDBHs with the complete genomes of *S. meliloti*, *A. tumefaciens* C58 U Washington, *A. tumefaciens* C58 Cereon, *B. melitensis*, *B. suis*, *M. loti*, *B. japonicum*, and *R. palustris*. Circles in p42b, p42c, p42e, and p42f, from the innermost outward, describe the following: GC; codon usage (CRI); scale; color-coded CDS; and *S. meliloti*, *A. tumefaciens* C58 U Washington, *A. tumefaciens* C58 Cereon, *B. melitensis*, *B. suis*, *M. loti*, *B. japonicum*, and *R. palustris*. Color codes for the CDS according to their functional category are as follows: orange, amino acid biosynthesis; light red, biosynthesis of cofactors; pale green, macromolecule biosynthesis; mid red, nucleotide biosynthesis and central intermediary metabolism; cyan, global functions; yellow; nitrogen fixation; red, energy transfer; magenta, degradation; pink, structural elements; pale pink, cell processes; dark gray, transport; light gray, adaptations; blue, nodulation; green, elements of external origin; sky blue, transcriptional regulators; brown, hypotheticals.

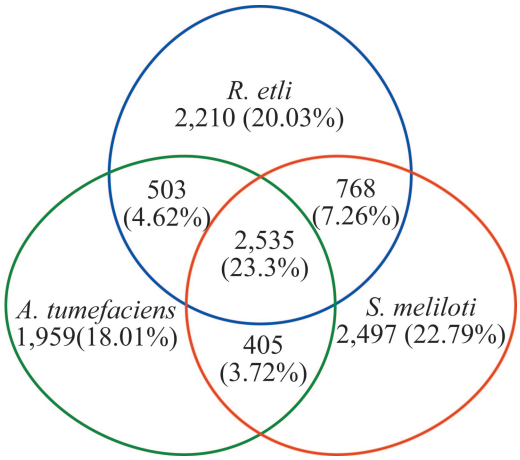


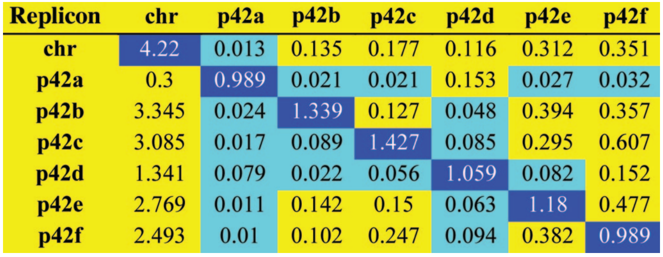
Fig. 2. Shared orthologous proteins among the currently available Rhizobiacea complete genomes. The BDBH criterion was used for ortholog identification. The total number of different proteins is 10,877. The number of proteins is 6,034 for *R. etli*, 6,205 for *S. meliloti* (7), and 5,402 for *A. tumefaciens* (U. Washington) (9).

circles). Other orthologs from plasmids have counterparts in the main chromosome of multipartite genomes or in the unique chromosome of some other species (e.g., *B. japonicum*; Fig. 1, blue circles). In contrast to the chromosome, plasmids lack synteny. The current gene composition of the *R. etli* plasmids p42b, p42c, p42e, and p42f, and the ubiquitous presence of orthologs in both plasmids and chromosome, suggests that these plasmids might have been part of the ancestral genome.

The plasmids p42a and p42d are the poorest conserved replicons of the *R. etli* genome but show some highly conserved regions shared with plasmids of *Agrobacterium* (i.e., pTiC58, pTi-Sakura, and pRi1724), *Rhizobium* spp (pNGR234a), and *M. loti* (pMLa) (5, 11, 22, 23) (Fig. 1). These conserved segments contain the *vir*, *tra*, and *sym* regions required for T-DNA transfer, conjugation and symbiosis, respectively. Plasmid p42a has a complete set of *vir* genes (*virB1–11*, *virD1-D4*, *virA-G*, *virE2*, *virH*, and *virF*) but it lacks the T-DNA region. There is no evidence suggesting that the *vir* proteins play a role in transport of macromolecules in *R. etli*, but it would be interesting to test whether they are able to mobilize T-DNA and transform plant cells.

Transcriptional Regulation. We found 23 putative sigma subunits in *R. etli*, *M. loti*, and *B. japonicum* contain a similar number of sigma factors (25 and 26, respectively), whereas *S. meliloti* (16 sigmas) and *A. tumefaciens* (10 sigmas) (5–9) contain fewer. Fifteen sigma genes are chromosomal, and 7 are found in plasmids. They belong to the sigma-70 family that includes *sigA*, two copies of *rpoH*, and 18 genes of the extracitoplasmic factor group. In addition, there are two copies of sigma-54. The roles

Table 2. Predicted functional relationships among the *R. etli* replicons



Replicon	chr	p42a	p42b	p42c	p42d	p42e	p42f
chr	4.22	0.013	0.135	0.177	0.116	0.312	0.351
p42a	0.3	0.989	0.021	0.021	0.153	0.027	0.032
p42b	3.345	0.024	1.339	0.127	0.048	0.394	0.357
p42c	3.085	0.017	0.089	1.427	0.085	0.295	0.607
p42d	1.341	0.079	0.022	0.056	1.059	0.082	0.152
p42e	2.769	0.011	0.142	0.15	0.063	1.18	0.477
p42f	2.493	0.01	0.102	0.247	0.094	0.382	0.989

Values represent the fraction of gene products with at least one functional link in other replicon. Fractions represent data normalized by the number of genes in each replicon (number of predicted links in each replicon divided by the number of genes in the replicon). Diagonal dark blue boxes denote the predicted relationships within the replicon. Light blue boxes indicate the lowest values for interactions among replicons.

of most of these sigmas are not known, but they might be needed for gene expression under variable environmental conditions.

We also found 536 transcriptional regulators, 65% of which are in the chromosome and the rest are distributed in plasmids. There are 129 two-component regulators: 49 correspond to sensor histidine-kinase, 68 are response proteins, and 12 are fusions between sensor and response domains. The 331 one-component regulators belong to the LysR (the most represented), TetR, AraC, LacI, and GntR families. The great majority (62%) of the one-component regulators are nearby neighbors of ABC transporters or other permease genes. They may constitute specific regulatory circuits activated in response to environmental challenges that *R. etli* faces in the soil.

Functional Relationships Among Replicons. We used NEBULON (24), a recently described method based on operon rearrangements and other genomic-context features, to predict functional relationships between gene products. Of the 6,034 protein coding genes in *R. etli*, 4,785 have at least one predicted functional link in NEBULON. Genes with the maximum number of putative functional relationships (66 links) are related to the translational machinery and ribosomal structure. The chromosome and plasmids p42b, p42c, p42e, and p42f, have an average connectivity (number of interactions per gene) of 0.8. Plasmids p42a and p42d have smaller connectivity values (0.63 and 0.74, respectively) Likewise, the links among replicons pointed out that p42b, p42c, p42e, and p42f are more interconnected among themselves and to the chromosome than to either p42a or p42d (Table 2). The poor functional connectivity of p42a and p42d further supports the suggestion that these plasmids are horizontal acquisitions.

Nodulation. Previously, we reported that p42d (pSym) contains most of the genes needed for symbiosis (10). Here, we identify homologs for nodulation genes in other replicons of the genome. Among them there are genes whose protein products participate in the biosynthesis and modification of fucose and mannose (*nodL*, *noeL*, *nolK*, *noeK*, *noeJ*, and *nodN*), the efflux transporter (*nodT*), the two-component regulators *nodVW* and *nfeD*, the suppressor of *nodVW*, and the two-component regulator *nwsAB*. The role of these genes has not yet been established in *R. etli*, but in other related organisms (i.e., *S. fredii*, *S. meliloti*, *M. loti*, and *B. japonicum*), they are not essential for symbiosis, instead they are involved in competition and nodulation efficiency in some host plants (25).

Additionally, we found 27 genes in different genomic locations functionally linked to nodulation genes, as predicted by NEBULON (Table 8, which is published as supporting information on the PNAS web site). These nodulation-associated genes are immersed into large clusters for exo and lipopolysaccharide biosynthesis and transport (COG M). For instance, the *wzm*, *wzt*, *ypch00257*, *yhch00181*, *ypch00258.1*, *ypch00259*, and *ypch00259.1* genes are in the already described *α-lps* region (26). Some genes in this region have been involved in the synthesis, modification, and transport of the O-antigen and *lps* modification in response to pH or antocianins (27). Mutations on some of them, like *wzm*, have a negative effect on the symbiotic capabilities of *R. etli* CE3 (28).

Exopolysacharide Biosynthesis. We annotated 222 *R. etli* genes as implicated in the synthesis of the extracellular envelope. Most of them are dispersed in the genome, except for three major exopolysacharide (EPS) clusters of genes located in the chromosome. Two clusters with 20 and 18 genes, respectively, encode for several sugar-glycosyl and acetyl transferases, and genes distantly related to *exoQ*, *exoF*, and *exsH* of other rhizobia. The function of these EPS clusters is unknown.

The largest gene cluster for EPS biosynthesis includes 43 genes, most of them previously characterized in *R. leguminosarum* bv. *vicea* and bv. *trifolii* (29, 30). This cluster contains the genes *pssV*, *U*, *T*, *S*, *R*, *M*, *L*, *K*, *J*, *I*, *H*, *G*, *F*, *C*, *D*, *E*, *P*, *O*, and *N*, and other genes encoding epimerases, deacetylases, and glycotransferases, whose role in EPS synthesis has not been studied. In *R. leguminosarum*, mutations in some of these genes yield a variety of phenotypes related to the amount and kind of EPS, including deficiencies in nodule formation with concomitant loss of bacteroid differentiation, and the early release of the bacteria from the infection thread (29, 30). The *pss* locus is highly conserved between *R. etli* and *R. leguminosarum* but absent in *S. meliloti*, *M. loti*, *B. japonicum*, and *A. tumefaciens* (Fig. 6, which is published as supporting information on the PNAS web site). Therefore, the *pss* locus might encode for an alternate pathway for the synthesis of the extracellular polysaccharides. Notably, *R. etli* lacks *exo*, *exp*, and *rkp* clusters homologous to those in other rhizobia. For instance, it has been shown that the *exoPNOMAL* region is present in *Rhizobium* sp. NGR234, *S. meliloti* 1021, *M. loti* MAFF303099, and *A. tumefaciens* C58 (31). This region is present in the pSymB of *S. meliloti* 1021 together with 10 additional gene clusters encoding for the synthesis of EPS. Some of them participate in the synthesis of succinoglycan (EPS I) and galactoglucan (EPS II) having a role in symbiosis (32). *R. etli* lacks most of these gene clusters that include the *exoQXUVTIHKLAO* and *expE1* to *expE8* and *expPGC* and *expA1* to *expA5* genes. Taking these data together, they convey that *R. etli* has a very different kind of external envelope perhaps with distinct polysaccharides, some of them might represent another strategy for nodule formation.

Metabolism. *R. etli* must contend with the free-living environments and the differentiated bacteroid state. Thus, greater metabolic plasticity would be expected in *R. etli* than in organisms with more stable niches. To assess this idea, we modeled the *R. etli* metabolism by using PATHWAY TOOLS (33) and KEGG pathways maps (34). We predicted 263 metabolic pathways comprising 1,340 enzymatic reactions. There are more putative pathways in the *R. etli* chromosome than are currently annotated in *Escherichia coli* (35). A few other pathways reside in plasmids (i.e., protecatechuate degradation, glycerol metabolism, thiamine and cobalamine biosynthesis, and the incomplete denitrification pathway). The main differences between the *R. etli* and *E. coli* metabolic models reside in the elevated number of putative fermentation pathways, degradation and assimilation of amino acids, aromatic compounds, carboxylates, sugars, and polysaccharides. Many of the enzymes identified in *R. etli* pathways are putative isozymes that represent 42% (455) of the whole enzyme set (Table 3), whereas in *E. coli* K12, there are

Table 3. Comparative ECs annotations between *E. coli* and Rhizobeaceas according to KEGG database

Features	*E. coli* K12	*R. etli*	*S. meliloti*	*A. tumefaciens*
EC numbers	693	627	625	584
Proteins with EC numbers	963	1061	1029	882
Monomeric enzymes	538	423	435	429
Complexes	40	42	36	33
Protein in complexes	134	183	129	106
Isozyme families	115	162	154	122
Proteins in isozyme families	291	455	465	340

Monomeric enzymes correspond to unique proteins (without possible isozymes or not forming complex with other different proteins).

only 291 annotated isozymes (30%). As an example, in *R. etli*, the complete enzyme set for amino acid biosynthesis can be mapped in the chromosome; however, 50 additional isozymes are distributed between the chromosome and the plasmids (Table 9, which is published as supporting information on the PNAS web site). In *E. coli*, there are few isozymes for the same pathways. Such high enzyme redundancy in *R. etli* might correlate with the different degrees of metabolic responses and alternative regulation necessary to cope with a challenging environment without compromising the integrity of the pathways.

Horizontal Gene Transfer (HGT). The incidence of horizontal gene transfer in *R. etli* is evidenced by the presence of some phage and plasmid-related genes in the chromosome. To our knowledge, there is no description of bacteriophages for *R. etli* in the literature. However, a GC-rich region of ≈36 kb contains genes encoding for the small and large subunits of phage terminal transferase and lysozime related to bacteriophage Mx3 from *Myxococcus xantus*. Some other phage footprints were identified in distinct chromosomal locations (e.g., genes encoding for uracyl-DNA glycosilase, several phage recombinases, and the Clp protease). Similarly, some genes of putative plasmid origin were recognized as encoding for plasmid stabilization proteins (*stdB* and *ypch000647*) and the relaxase conjugal transfer protein TraA (*traAch* and *ypch000639*). These genes are found within a variable region of ≈67 kb that probably represents a plasmid insertion.

By using a conservative approach of phylogenetic congruence, we identified 109 potential HGT events in *R. etli* (Table 10, which is published as supporting information on the PNAS web site). They consist of genes that belong to the small molecule metabolism, transport, and transcriptional regulation functional classes. Remarkably, almost a half of these genes could compose operons like the genes *appABCDF* that encode for an oligopeptide transporter, the *glms2nagA* for glucosamine biosynthesis, the *sfuAB* for iron transport, among others. It is likely that horizontal gene transfer might have contributed to expand the metabolic repertoire of *R. etli*.

Conclusion

The genome of *R. etli* has more replicons than any other completely sequenced nitrogen-fixing symbiotic bacterium. The RepABC replicator, present in the plasmids, confers great stability to the genome by the use of distinct initiators (the RepC protein) and origins of replication. Cointegration of replicons have been shown to occur in *S. meliloti*, but cointegrates are unstable and revert to the original state (36). One advantage of such partitioned genomes could be faster duplication times to replicate the entire genome, as Guo *et al.* (36) have shown. Absence of homologous plasmids among the species compared suggests independent evolutionary origins or, alternatively, higher rates of variation compared with the chromosome. Even though plasmids have been considered accessory elements, our analysis indicates that their gene products might work together. Previous results on the symbiotic and growth properties of plasmid-cured strains derived from *R. etli* CFN42 support this conclusion (21, 37). Moreover, several genes implicated in symbiosis have been located in p42b (*lps* β loci), p42f (*fix* genes), and, as we have shown here, in the chromosome as well (38, 39). Plasmids p42a and p42d are atypical molecules in the context of the rest of the genome. Both plasmids might have been acquired at some point during the divergence of *R. etli*. The other four plasmids have coherent attributes among themselves and with the chromosome, indicating long-term coevolution. The structural characteristics of the *R. etli* genome highlight the important evolutionary roles of horizontal gene transfer, duplications, gene loss, and genomic rearrangements. The assessment of the specific contribution of these processes to genome differentiation and speciation should await further comparisons with closely related species.

Materials and Methods

DNA Sequencing. The complete sequence of *R. etli* was determined from the wild-type strain *R. etli* CFN42 and the derivate strain *R. etli* CFN42ΔE (21). The native strain *R. etli* CFN42 was collected 25 years ago from red nodules of *Phaseolus vulgaris* in an agricultural field in Guanajuato, México. Small insert shotgun libraries were constructed for the native strain and CFN42ΔE. A bacterial artificial chromosome (BAC) library was made for *R. etli* CFN42. A total of 117,596 high-quality readings from the shotgun libraries were collected by using ABI3700 and MEGABACE-1000 sequencers. One hundred eighty-two pairs of BAC-end sequences covering the entire genome were obtained and used to guide the assembly. Assemblages were obtained by the PHRED-PHRAP-CONSED software (40, 41). For plasmids p42e and p42f, a physical map was developed based on BACs. Final assembly reaches a quality of <1 error per 100,000 bases and an average coverage of 9.1× (Table 11, which is published as supporting information on the PNAS web site).

Bioinformatics. ORF prediction was done by following a reported Glimmer-based iterative strategy in refs. 10 and 42. Annotation was carried out with the help of BLASTX comparisons against the GenBank nonredundant database (43), INTERPRO (44) searches, and manual curation by using ARTEMIS (45). Annotation of COGs, gene ontologies, and EC numbers was performed by using SWISSPROT (46) and KEGG (34). We used PATHWAY TOOLS to model metabolism (33).

Orthologs were defined by the BDBH criterion (47). Paralogous families were identified as sets of proteins showing maximum BLAST alignments with a cutoff of 1×10^{-6} and minimum coverage of 80%. Functional relationships among the *R. etli* predicted proteins were deduced by using Nebulon functional relationships among the *R. etli* predicted proteins were deduced with NEBULON (24). Horizontal gene transfer predictions were carried out by protein-based phylogenetic congruency tests, taking all of the orthologs in the set of 127 nonredundant genomes (48). According to Medrano-Soto *et al.* (16), phylogenies were built only for genes with 10 or more orthologs and horizontal gene transfer predictions involving *R. etli* genes were kept. The annotated *R. etli* genome is available at www.ccg.unam.mx/retlidb.

V.G. and G.D. thank Rafael Palacios for his continuous support to this project; S. Brom (Universidad Nacional Autónoma de México) for providing the *R. etli* CFN42ΔE strain; M. Canedo, J. Martínez, J. Gama, R. E. Gómez, J. L. Fernández, R. Díaz, D. García, F. Sánchez, E. Díaz, H. Salgado, V. del Moral, J. Espíritu, and C. Bonavides for technical and computational assistance; P. Vinuesa for reviewing the manuscript; and L. Lozano for IS identification. Financial support was obtained from Consejo Nacional de Ciencia y Technología Grant N-028 for emergent areas (2001).

1. Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., Ramuz, M. & Allardet-Servent, A. (1998) *J. Bacteriol.* **180,** 2749–2755.
2. Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L. & Ramuz, M. (1993) *J. Bacteriol.* **175,** 7869–7874.
3. Egan, E. S., Fogel, M. A. & Waldor, M. K. (2005) *Mol. Microbiol.* **56,** 1129–1138.
4. Garrity, G. M., Jhonson, K. L., Bell, J. A. & Searles, D. B. (2002) in *Bergey's Manual of Systematic Bacteriology* (Springer, New York), pp. 365.
5. Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., *et al.* (2000) *DNA Res.* **7,** 331–338.
6. Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., Watanabe, A., Idesawa, K., Iriguchi, M., Kawashima, K., *et al.* (2002) *DNA Res.* **9,** 189–197.
7. Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., *et al.* (2001) *Science* **293,** 668–672.
8. Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C., *et al.* (2001) *Science* **294,** 2323–2328.
9. Wood, D. W., Setubal, J. C., Kaul, R., Monks, D. E., Kitajima, J. P., Okura, V. K., Zhou, Y., Chen, L., Wood, G. E., Almeida, N. F., Jr., *et al.* (2001) *Science* **294,** 2317–2323.
10. González, V., Bustos, P., Ramírez-Romero, M. A., Medrano-Soto, A., Salgado, H., Hernández-González, I., Hernández-Celis, J. C., Quintero, V., Moreno-Hagelsieb, G., Girard, L., *et al.* (2003) *Genome Biol.* **4,** R36.
11. Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A. & Perret, X. (1997) *Nature* **387,** 394–401.
12. Gottfert, M., Rothlisberger, S., Kundig, C., Beck, C., Marty, R. & Hennecke, H. (2001) *J. Bacteriol.* **183,** 1405–1412.
13. 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. (2002) *J. Bacteriol.* **184,** 3086–3095.
14. Sullivan, J. T. & Ronson, C. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5145–5149.
15. Segovia, L., Young, J. P. & Martínez-Romero, E. (1993) *Int. J. Syst. Bacteriol.* **43,** 374–377.
16. Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A. & Collado-Vides, J. (2004) *Mol. Biol. Evol.* **21,** 1884–1894.
17. Flores, M., González, V., Brom, S., Martínez, E., Piñero, D., Romero, D., Dávila, G. & Palacios, R. (1987) *J. Bacteriol.* **169,** 5782–5788.
18. Flores, M., González, V., Pardo, M. A., Leija, A., Martínez, E., Romero, D., Piñero, D., Dávila, G. & Palacios, R. (1988) *J. Bacteriol.* **170,** 1191–1196.
19. Flores, M., Mavingui, P., Perret, X., Broughton, W. J., Romero, D., Hernández, G., Dávila, G. & Palacios, R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 9138–9143.
20. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278,** 631–637.
21. Brom, S., García-de los Santos, A., Cervantes, L., Palacios, R. & Romero, D. (2000) *Plasmid* **44,** 34–43.
22. Suzuki, K., Hattori, Y., Uraji, M., Ohta, N., Iwata, K., Murata, K., Kato, A. & Yoshida, K. (2000) *Gene* **242,** 331–336.
23. Moriguchi, K., Maeda, Y., Satou, M., Hardayani, N. S., Kataoka, M., Tanaka, N. & Yoshida, K. (2001) *J. Mol. Biol.* **307,** 771–784.
24. Janga, S. C., Collado-Vides, J. & Moreno-Hagelsieb, G. (2005) *Nucleic Acids Res.* **33,** 2521–2530.
25. Gottfert, M., Grob, P. & Hennecke, H. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2680–2684.
26. Carlson, R. W., Reuhs, B., Chen, T. B., Bhat, U. R. & Noel, K. D. (1995) *J. Biol. Chem.* **270,** 11783–11788.
27. Duelli, D. M., Tobin, A., Box, J. M., Kolli, V. S., Carlson, R. W. & Noel, K. D. (2001) *J. Bacteriol.* **183,** 6054–6064.
28. Lerouge, I., Verreth, C., Michiels, J., Carlson, R. W., Datta, A., Gao, M. Y. & Vanderleyden, J. (2003) *Mol. Plant–Microbe Interact.* **16,** 1085–1093.
29. Wielbo, J., Mazur, A., Krol, J., Marczak, M., Kutkowska, J. & Skorupska, A. (2004) *Arch. Microbiol.* **182,** 331–336.
30. Sadykov, M. R., Ivashina, T. V., Kanapin, A. A., Shliapnikov, M. G. & Ksenzenko, V. N. (1998) *Mol. Biol. (Mosk.)* **32,** 797–804.
31. Streit, W. R., Schmitz, R. A., Perret, X., Staehelin, C., Deakin, W. J., Raasch, C., Liesegang, H. & Broughton, W. J. (2004) *J. Bacteriol.* **186,** 535–542.
32. Charles, T. C. & Finan, T. M. (1991) *Genetics* **127,** 5–20.
33. Karp, P. D., Paley, S. & Romero, P. (2002) *Bioinformatics* **18,** Suppl. 1, S225–S232.
34. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004) *Nucleic Acids Res.* **32,** D277–D280.
35. Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M. & Karp, P. D. (2005) *Nucleic Acids Res.* **33,** D334–D337.
36. Guo, X., Flores, M., Mavingui, P., Fuentes, S. I., Hernández, G., Dávila, G. & Palacios, R. (2003) *Genome Res.* **13,** 1810–1817.
37. Brom, S., García de los Santos, A., Stepkowsky, T., Flores, M., Dávila, G., Romero, D. & Palacios, R. (1992) *J. Bacteriol.* **174,** 5183–5189.
38. García-de los Santos, A. & Brom, S. (1997) *Mol. Plant–Microbe Interact.* **10,** 891–902.
39. Girard, L., Brom, S., Dávalos, A., López, O., Soberón, M. & Romero, D. (2000) *Mol. Plant–Microbe Interact.* **13,** 1283–1292.
40. Gordon, D., Abajian, C. & Green, P. (1998) *Genome Res.* **8,** 195–202.
41. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8,** 175–185.
42. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998) *Nucleic Acids Res.* **26,** 544–548.
43. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. (2005) *Nucleic Acids Res.* **33,** D34–D38.
44. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., *et al.* (2001) *Nucleic Acids Res.* **29,** 37–40.
45. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000) *Bioinformatics* **16,** 944–945.
46. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.* (2003) *Nucleic Acids Res.* **31,** 365–370.
47. Moreno-Hagelsieb, G. & Collado-Vides, J. (2002) *Bioinformatics* **18,** Suppl. 1, S329–S336.
48. Moreno-Hagelsieb, G. & Collado-Vides, J. (2002) *In Silico Biol. (Gedrukt)* **2,** 87–95.