

Research

## The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments

Víctor González, Patricia Bustos, Miguel A Ramírez-Romero, Arturo Medrano-Soto, Heladia Salgado, Ismael Hernández-González, Juan Carlos Hernández-Celis, Verónica Quintero, Gabriel Moreno-Hagelsieb, Lourdes Girard, Oscar Rodríguez, Margarita Flores, Miguel A Cevallos, Julio Collado-Vides, David Romero and Guillermo Dávila

Address: Centro de Investigación Sobre Fijación de Nitrógeno, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México 62210.

Correspondence: Guillermo Dávila. E-mail: [davila@cifn.unam.mx](mailto:davila@cifn.unam.mx).

Published: 13 May 2003

*Genome Biology* 2003, 4:R36

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2003/4/6/R36>

Received: 1 November 2002

Revised: 6 March 2003

Accepted: 2 April 2003

© 2003 González *et al.*; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** Symbiotic bacteria known as rhizobia interact with the roots of legumes and induce the formation of nitrogen-fixing nodules. In rhizobia, essential genes for symbiosis are compartmentalized either in symbiotic plasmids or in chromosomal symbiotic islands. To understand the structure and evolution of the symbiotic genome compartments (SGCs), it is necessary to analyze their common genetic content and organization as well as to study their differences. To date, five SGCs belonging to distinct species of rhizobia have been entirely sequenced. We report the complete sequence of the symbiotic plasmid of *Rhizobium etli* CFN42, a microsymbiont of beans, and a comparison with other SGC sequences available.

**Results:** The symbiotic plasmid is a circular molecule of 371,255 base-pairs containing 359 coding sequences. Nodulation and nitrogen-fixation genes common to other rhizobia are clustered in a region of 125 kilobases. Numerous sequences related to mobile elements are scattered throughout. In some cases the mobile elements flank blocks of functionally related sequences, thereby suggesting a role in transposition. The plasmid contains 12 reiterated DNA families that are likely to participate in genomic rearrangements. Comparisons between this plasmid and complete rhizobial genomes and symbiotic compartments already sequenced show a general lack of synteny and colinearity, with the exception of some transcriptional units. There are only 20 symbiotic genes that are shared by all SGCs.

**Conclusions:** Our data support the notion that the symbiotic compartments of rhizobia genomes are mosaic structures that have been frequently tailored by recombination, horizontal transfer and transposition.

## Background

Nitrogen-fixing symbiotic bacteria grouped within the Rhizobiaceae, Phyllobacteriaceae and Bradyrhizobiaceae families are widespread in nature [1]. Ordinarily known as rhizobia, these organisms contain genomes of one or two chromosomes and several large plasmids ranging in size from about 100 kilobases (kb) to more than 2 megabases (Mb). A common feature of the genomes of rhizobia is that the genes involved in the symbiotic process are located in specific symbiotic genome compartments (SGCs), either as independent replicons known as symbiotic plasmids (pSym) or as symbiotic islands or regions within the chromosome. Complete genome sequences have been recently reported for *Mesorhizobium loti* MAFF303099 [2], *Sinorhizobium meliloti* 1021 [3–6], *Bradyrhizobium japonicum* USDA110 [7] and the non-nitrogen-fixing close relative *Agrobacterium tumefaciens* C58 [8,9]. In addition, the sequence of the pSym of *Rhizobium* species NGR234 - pNGR234a [10] - as well as that of the chromosomal symbiotic regions of *B. japonicum* USDA110 and *M. loti* R7A have been reported [11,12]. Genomic comparisons reveal that the chromosomes of *S. meliloti*, *M. loti*, and the circular chromosome of *A. tumefaciens* have more than 50% of orthologous genes in common [6]. A clear syntenic relationship is observed between the circular chromosomes of *S. meliloti* and *A. tumefaciens* [8,9] and albeit to a lesser extent, synteny is also apparent when both are compared to the chromosome of *M. loti* [6,8,9]. These results lead to the hypothesis that rhizobial chromosomes have a common ancestral origin [6,8,9]. Other genome constituents of rhizobia (that is, other chromosomes and plasmids) are thought to be the result of subsequent events of genomic rearrangements and horizontal transfer [6,8,9], but the precise mechanisms involved in their generation have not been elucidated so far.

Here we report the complete DNA sequence of the pSym (p42d) of *Rhizobium etli* CFN42 and its comparative analysis with other rhizobial SGCs. *R. etli* is the symbiont of the common bean *Phaseolus vulgaris* and has been widely used as model for metabolic and genome dynamics studies [13,14]. Its genome is composed by one chromosome and six plasmids ranging in size from 184 kb to about 600 kb [15]. The physical map of p42d was previously determined and was the basis for obtaining the entire sequence [16]. In this study we show that the SGCs are heterogeneous in sequence, gene composition and gene order. There are only 20 symbiotic genes that are shared by all SGCs. There are also some conserved gene clusters of related function that are present in some SGCs, but absent in others. Besides genes unique to a particular SGC, several orthologous genes are located in different genome contexts in other rhizobia. Other common features to all SGCs, such as reiterated genes, pseudogenes, and a large amount of insertion sequences (ISs), support the view that p42d, as well as other SGCs, is a mosaic structure that may have assembled from different genome contexts, either chromosomal or plasmidic.

## Results and discussion

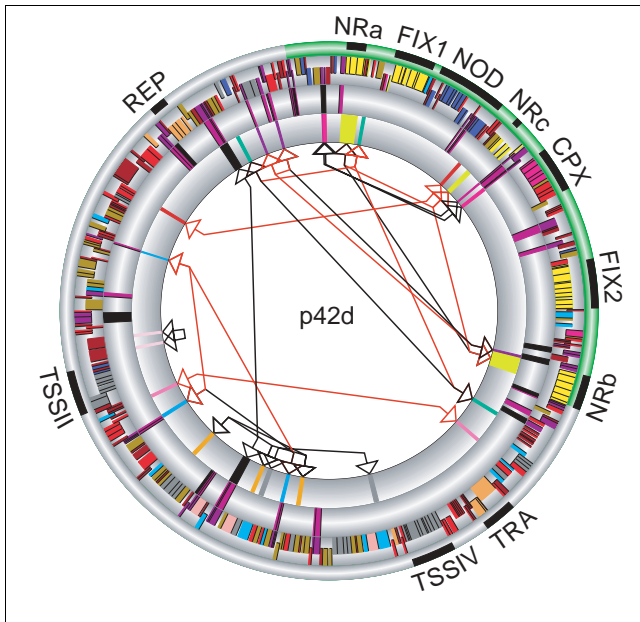
### General features of p42d

The symbiotic plasmid p42d is a circular molecule of 371,255 base-pairs (bp) (Figure 1) that belongs to the *repABC* type of replicator [17]. We identified 359 coding sequences (CDS), of which 63% have an assigned function, 17% have homologs in databases without an assigned function, and 20% are orphan (Figure 1, see also Additional data file 1). The CDS distribution between the two strands is asymmetrical, with 61% of them located in the minus strand. The plus and minus strands were defined according to the previously reported physical map [16]. Moreover, the plus strand contains two reiterated *nifHDK* gene clusters in a clockwise orientation (NRa and NRb, Figure 1). The main functional classes of genes identified are: transport, nitrogen fixation, nodulation and transcriptional regulation. Ten pseudogenes related to known genes were identified that carry deletions and frameshifts at their amino or carboxyl termini. The plasmid also contains many reiterated sequences and a large number of elements related to insertion sequences (ERIS) accounting for 10% of the entire sequence. The major reiterations (28 elements) were grouped into 12 families on the basis of their sequence similarity (see below).

The average GC content of the plasmid is 58.1%. When genes were classified as low, average or high GC content (using the mean GC  $\pm$  1 standard deviation as thresholds), we observed a clear distinction between high or low GC in some gene clusters (Figure 2a). Several hypothetical genes and the *nod* genes show the lowest GC values (< 55%), whereas the highest GC values (> 62%) were displayed by the genes for cytochrome P450 (CPX), *tra* genes (TRA), and the genes for type III (TSSIII) and type IV (TSSIV) transport secretion systems. Similarly, when the genes were classified according to poor, typical or rich codon usage (CU) (see Materials and methods for details), genes with high GC also exhibited a rich CU (Figure 2b), whereas the GC-CU correlation was found to be lower for other genes. For example, the regions that contain the nitrogenase structural genes, other *nif* genes (NRa, b and c, see below), and the *fixNOQPGHIS* genes (FIX1) showed average GC content but rich CU. The variable correlation between GC content and CU levels reveals sequence heterogeneity within p42d and suggests a dynamic structure for this plasmid, presumably as a consequence of extensive genomic rearrangements, recombination rates, lateral transfer, and relaxation or intensification of selective pressures.

### Organization of genes involved in nodulation

Most *nod* genes present in p42d are clustered in a region of 16 kb (NOD); however, *nodA* is separated from *nodBC* by 27 kb. The Nod factor backbone of *R. etli* CFN42 is an *N*-acetylglucosamine pentasaccharide synthesized by the common *nodA* and *nodBC* gene products. Modifications to this backbone consist of methyl and carbamoyl groups at the non-reducing end, while the reducing one is modified by the addition of a fucosyl group that is in turn acetylated [18]. The



**Figure 1**  
 Structure of the symbiotic plasmid p42d of *R. etli* CNF42. The structure of p42d is represented in five concentric circles. Outermost circle, relevant regions referred to in the text: NRa, b and c, regions containing nitrogenase structural genes; FIX1 and FIX2, clusters containing nitrogen-fixation genes; NOD, major cluster of nodulation genes; CPX, cluster for cytochrome P450; TRA, cluster for *tra* genes; REP, replicator region; TSSIII and IV, clusters for transport secretion system genes. The 125 kb region that contains most of the symbiotic genes, described in the text as a putative mobile element, is shown in green. Second circle, organization of predicted CDSs located according to the direction of transcription color-coded as below; those transcribed on the plus strand are shown in the outer half of the circle. For each class, the number of CDSs and the percentage of the total are: hypothetical (70) 19.5% (dark red); hypothetical conserved (62) 17.3% (red); integration recombination (55) 15.3% (purple); various enzymatic functions (45) 12.3% (khaki); transport secretion systems (37) 10.3% (gray); nitrogen fixation (35) 9.8% (yellow); nodulation (18) 5% (dark blue); transcriptional regulation (15) 4.2% (light blue); plasmid maintenance (10) 2.8% (orange); electron transfer (7) 2.1% (magenta); chemotaxis (3) 0.8% (pink); and polysaccharide synthesis (2) 0.6% (green). Third circle, elements related to insertion sequences (ERIS). Putative partial ISs (purple), and putative complete ISs (black). Fourth circle, reiterated DNA families. The major reiterated families (see text) are shown in different colors. Innermost circle, potential genomic rearrangements. Arrowheads indicate the sites for homologous recombination leading to genomic rearrangements. Black lines connect sites for amplification or deletion events; red lines connect sites for inversion.

methyltransferase, fucosyltransferase and acetyltransferase activities required for these modifications are encoded by *nodS*, *nodZ* and *noII*, respectively. It is unclear, however, which gene product is responsible for the carbamoylation of the Nod factor, as *nodU*, the most likely gene to carry out this function, is a pseudogene. The two membrane proteins encoded by *nodI* and *nodJ* (located downstream of *nodBCSU*), participate in the transport of the Nod factor to the outside of the cell [19]. Other genes present in p42d whose homologs in other rhizobia have a role in nodulation are *nolO*, *nolE*, *nolT*

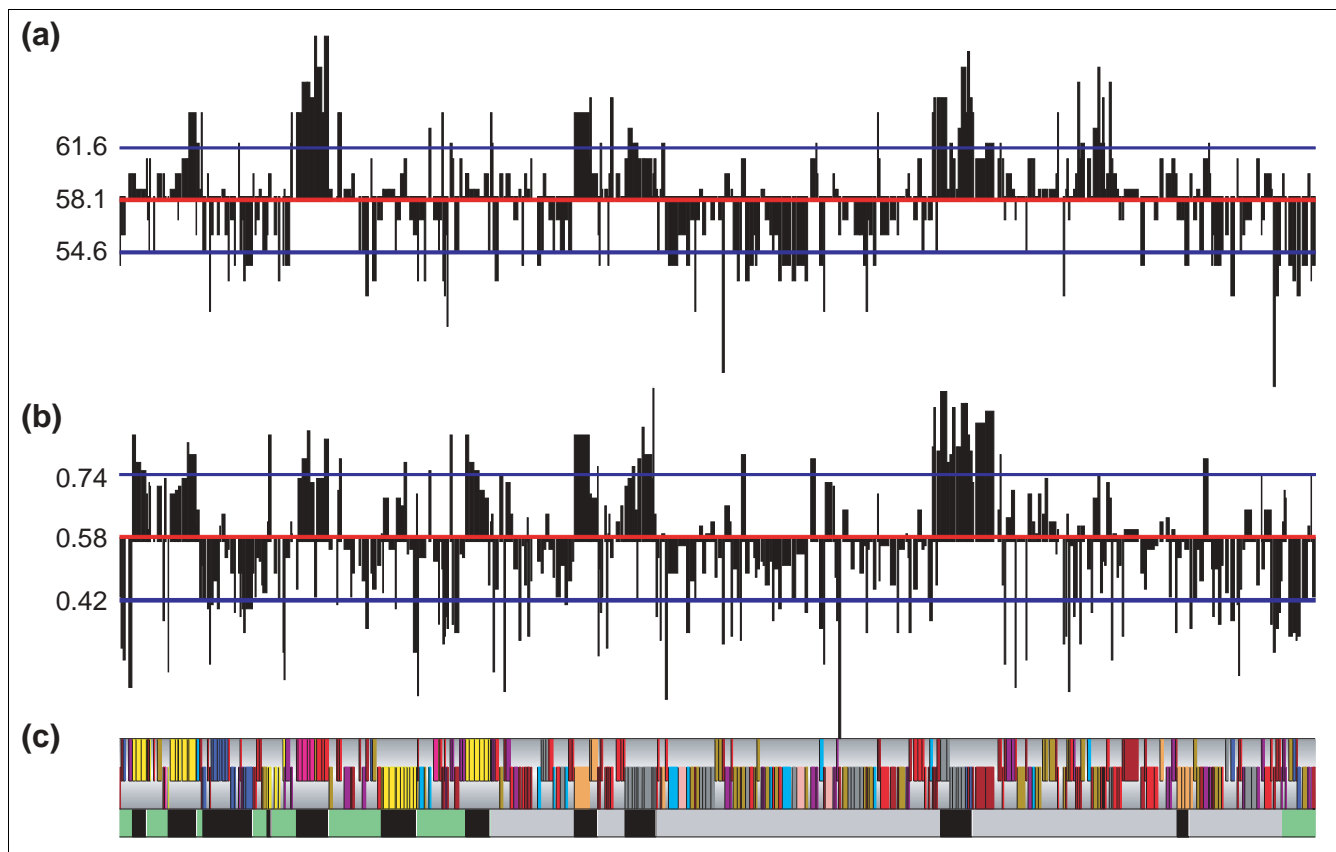
and *noIV*, the last two being part of the TSSIII system (see below). In addition to *nodU*, two other pseudogenes, *noeI* and *nodQ*, were identified.

The expression of *nod* genes depends on the activity of NodD proteins [20], which interact with specific sites known as *nod* boxes located upstream of the *nod* operons [21]. The sequence of the p42d revealed three *nodD* genes; *nodD*<sub>1</sub> is present in the NOD region while *nodD*<sub>2</sub> and *nodD*<sub>3</sub> are 50 kb apart. We also predict 15 potential *nod* boxes (see Materials and methods and Additional data file 2), seven of which are associated with almost all *nod* genes: *nodA*, *nodZ*, *nodBCSU*, *nolE*, *nodD*<sub>1</sub>, *nodD*<sub>2</sub>, and *nodD*<sub>3</sub>. The rest of the *nod* boxes are located proximal to genes so far unrelated to the nodulation process; namely, the genes *bglS* ( $\beta$ -glucosidase), *yp108* (putative monooxygenase), and the orphans *yho05*, *yho07* and *yho50*. There is also a putative *nod* box upstream of the gene encoding NifA, the major transcriptional regulator of the nitrogen-fixation genes. Even though the regulation of *nifA* is variable among rhizobia, dependence on flavonoid induction is unknown.

**Organization of the genes involved in nitrogen fixation**

The *nif* and *fix* genes are distributed in five regions spanning a total of 125 kb (Figure 1); the NOD cluster mentioned previously maps within this section as well. There are three copies of the nitrogenase reductase gene *nifH* [22], defining the three *nif* regions (NR) a, b and c (Figure 1). NRa contains *nifHDK* genes and a truncated *nifE* pseudogene; NRb contains the *nifHDKENX* genes; NRC contains *nifH* and a truncated *nifD* pseudogene. The largest reiterated regions found in p42d correspond to NRA and NRb regions that share 4,470 identical nucleotides. The NRC region of 1,131 nucleotides is identical to sequences within NRA and NRb. The orientation of NRC is inverted in relation to the direction of NRA and NRb. Recent duplications of these NR regions might underlie the unusually high sequence identity between them. Alternatively, a mechanism of 'copy-correction' resembling gene conversion may be involved in maintaining nucleotide identity [23].

The highest density of *nif* and *fix* genes in p42d occurs 10 kb upstream of NRb, in the FIX2 region. This contains the *fixA-BCX* genes that encode a flavoprotein [24] (see below); *nifB*, which is needed for the synthesis of the iron-molybdenum cofactor [25]; *nifW* and *nifZ*, whose products may be required for protection of the nitrogenase from oxygen [26,27]; and the genes for the regulatory proteins NifA and RpoN2. The genes *nifU*, *nifS* and *hesB* (also named *iscN*) also map in the FIX2 region. The products of these genes have been implicated in the formation of the Fe-S cluster required for nitrogenase complex function [28]. In *R. etli* CNPAF512, the inactivation of *hesB* (*iscN*) results in a Fix<sup>-</sup> phenotype [28]. Other genes commonly found in *nif* regions of rhizobia were also identified in the FIX2 region. These are the ferredoxin gene *fdxN*, which is essential for nitrogen fixation in *S. meliloti* [29], and the gene for the anaerobic transcriptional regulator FnrNd

**Figure 2**

Compositional features of the coding sequences (CDS) of p42d. **(a)** GC content, and **(b)** CU of the 359 CDS of p42d. Red lines indicate the average in GC (58.1%) and CU (0.58). Blue lines indicate 1 standard deviation of GC  $\pm$  3.5% and CU  $\pm$  0.16. Highest and lowest percentage values of GC are 69.4 and 45.8 respectively. The CU limit values varies from 0.11 to 1.00. **(c)** CDS distribution with the color codes for functional classes and the relevant regions described in Figure 1.

(see below). The products of *nifV* and *nifQ* have been involved in the synthesis of the iron-molybdenum cofactor [30,31]; nevertheless, *nifV* is absent in the p42d and *nifQ* is located upstream of *nifHc* in the NRc region.

### RpoN regulation

The RpoN ( $\sigma^N$ , also known as  $\sigma^{54}$ ) subunit of the RNA polymerase, encoded by *rpoN*, and the transcriptional activator NifA protein, encoded by *nifA* (both present in the FIX2 region, Figure 1), participate in the regulation of *nif* genes. RpoN binds to specific promoter regions and interacts with the NifA protein that binds to specific upstream activator sequences (UAS) [32]. In *R. etli* CNPAF512, two *rpoN* genes have been described [33], one located in the chromosome (*rpoN*<sub>1</sub>), and the other in the pSym (*rpoN*<sub>2</sub>). The *rpoN* gene found in the p42d is orthologous to *rpoN*<sub>2</sub>. Regulation by RpoN and NifA has been demonstrated for *nifH a, b* and *c* [34]. We predicted, as described in Materials and methods, potential RpoN-binding sites and UAS for NifA in the upstream region of several genes (see Additional data file 3). Both types of sites were also identified upstream of other genes; the reiterated *yp003*, *yp021* and *yp099* genes that

encode the recently described BacS protein, highly expressed in nodules [35]; *yp010* in the putative operon for terpenoid synthesis [36]; the *fixA*, *hesB*, and *cpxA5* genes; and *yp104*, which encodes a toxin-transport-related protein. The expression of *yp003* (*bacS*) and *hesB* (*iscN*) has recently been shown to depend on NifA [28,35].

RpoN-like promoters were also predicted upstream of several genes for which no associated NifA-binding sites could be detected (see Additional data file 3). Among them are the nitrogen-fixation genes *fixO*, *nifQ* and *nifB*; the genes for the putative decarboxylase, *pcaC1*, and alcohol dehydrogenase, *xyLB2*. Furthermore, potential sites for RpoN were also found in several genes of unknown function. Recently, Dombrecht *et al.* [37] predicted RpoN promoter sites in all complete rhizobial genomes and p42d; we report here a larger set of genes potentially regulated by RpoN in p42d. It includes genes for nitrogen fixation, electron transfer, transport, and several of unknown function. The genes reported by Dombrecht *et al.* are mainly in *nif* and *fix* genes, the ferredoxins (*fdxB* and *N*; not predicted by us), and some genes of unknown function [37]. The differences between their results and ours

may be explained in part by the different strategies used to construct the weight-matrices in both studies, which in our case includes only 85 RpoN promoters whose transcription start sites have been experimentally determined, instead of the whole set of 186 promoters used by Dombrecht *et al.* ([37], see also [38]); see Materials and methods for details.

### Energy supply and anaerobic regulation

The electron flux and supply of energy for the reduction of molecular nitrogen requires the flavoprotein encoded by the *fixABCX* genes mentioned above (FIX2 region, Figure 1), a specific cytochrome oxidase encoded by *fixNOQP* genes, and a cation pump encoded by *fixGHIS* genes. The latter clusters in the region FIX1 (Figure 1). A second copy of *fixNOQP* and *fixG* genes has been found in the plasmid p42f [39].

In the symbiotic state, the cytochrome terminal oxidases encoded by the *fixNOQP* operon provide the energy required to fix nitrogen [32]. The cytochrome production is regulated in response to oxygen concentration and the products of *fixLJ*, *fixK* and *fnrNd* genes are also known to be involved in such regulation [40]. In *R. etli*, the duplicated *fixNOQP* operons are differentially regulated and only the *fixNOQPd* is required for symbiosis [39]. An inactive *fixKd* is present in p42d but no *fixJ* genes have been found in *R. etli* CFN42 [39]. It has been shown that FixKf controls both *fixNOQP* operons; loss of FixL (presumably a fusion protein of FixL and FixJ) suppresses *fixNOQPf* expression, but has only a moderate effect on that of *fixNOQPd* [39]. Two *fnrN* genes have been described in *R. etli* CFN42; one is chromosomal (*fnrNchr*), and the other is on p42d (*fnrNd*). Both regulators participate in the activation of the operon *fixNOQPd* [41].

In *Escherichia coli*, Fnr is an oxygen-responsive global transcriptional regulator that binds to conserved boxes upstream of several genes (anaeroboxes) [42]. By computational methods we predict 45 possible anaeroboxes in p42d (see Materials and methods). In some cases there are pairs of anaeroboxes in the same region. For example, two anaeroboxes lie within the intergenic region of the divergent operons *fixKd* and *fixNOQPd* and two more were detected upstream of *fixG*, *nocR*, *nodD<sub>3</sub>*, and *fnrNd*. Other genes that display single anaeroboxes are *fixX*, *nifW*, *nifU*, *hemN<sub>2</sub>*, *psiB*, *hesB*, *mcpC*, *teuB<sub>1</sub>* and some other genes of unknown function. Although there is no direct transcriptional evidence about the expression of these genes in microaerobic conditions, previous observations suggest that several regions of p42d are activated under these conditions [43].

### Complex systems for macromolecular transport

A variety of transporters, which account for 10% of the CDSs, are scattered throughout p42d. These include several partial and complete ABC transporters for sugars, as well as the type III (TSSIII), and type IV (TSSIV) large-molecule secretion systems (Figure 1).

In several pathogenic bacteria, the TSSIII translocate virulence factors into eukaryotic cells [44]. In *Rhizobium* this system was first found in pNGR234a [10] and it has been shown to have a role in nodulation efficiency in some host plants [45]. Genes that encode proteins implicated in this system are also present in some of the SGCs [2] and were detected in the sequence of p42d. Interestingly, a gene homologous with an elicitor of the hypersensitive response in plants, *hrpW*, is exclusively present in p42d. This gene might form an operon with *pcrD*, which encodes a calcium-binding membrane protein that is also part of the type-III secretion system.

The TSSIV encoded by the *virB* genes has been described in several  $\alpha$ -proteobacterial pathogens and plant symbionts [46] (see below). It consists of a membrane channel for delivering proteins or DNA into eukaryotic cells. In p42d, a complete set of *virB* genes, from *virB<sub>1</sub>* to *virB<sub>11</sub>*, is present (Figure 1). Other TSSIV correspond to the *tra* genes that participate in bacterial conjugation. Although p42d is not a self-conjugative plasmid [47], it contains the *traACDG* genes, an *oriT* and a truncated *traI* pseudogene (*yp096*), suggesting that p42d might have lost its self-conjugative capability.

### Other functions

In addition to *nifA*, *fnrN*, *rpoN<sub>2</sub>*, and *nodD<sub>1-3</sub>*, 12 predicted genes encoding potential transcriptional regulators are present in p42d. They belong to different families, including LysR, AraC, Crp and GntR. The plasmid also encodes other functions including plasmid-maintenance, electron transfer, polysaccharide biosynthesis, melanin synthesis and secondary metabolism. The sequence of p42d revealed a putative methionyl-tRNA synthetase that could represent a reiterated gene or could have another functional role (for example in antibiotic resistance) [48].

### Elements related to insertion sequences (ERIS)

In general, large numbers of ERIS have been found in the symbiotic compartments of rhizobia [2,7,10,11]. The genome of *S. meliloti*, however, contains a relatively low abundance of these elements and their distribution is asymmetric; that is ERIS are more abundant in the pSymA, especially near symbiotic genes [3]. In p42d, ERIS belonging to 12 known IS families comprise 10% of the entire DNA sequence. The great majority of them belong to the IS3 and IS66 families. Although most ERIS represent incomplete, presumably inactive, IS sequences, some of them are organized in complete IS elements (Figure 1).

The positions of some ERIS might suggest a role in plasmid shuffling. The 125 kb region that contains most of the symbiotic genes (Figure 1) is flanked by two complete IS elements. Both elements share identical 30 bp direct repeats at their borders, suggesting a potential transposition capability. The presence of the gene for an integrase-like protein (*yp018*) and the fact that the 125 kb region separates the *repABC* and the *tra* genes, has prompted the idea that the entire symbiotic

region could be a mobile element. Furthermore, some groups of genes flanked by ERIS might have arrived in p42d as part of composite transposons, such as the cytochrome P450 cluster (see below), the NRb region, and a putative ATPase of an ABC transporter.

### Reiterated DNA families and genomic rearrangements

It has previously been shown that p42d contains several reiterated DNA sequences [16] that can recombine, leading to genomic rearrangements [14,49,50]. The sequence of the plasmid revealed a large amount of DNA reiteration. The major reiterated families were defined by containing a continuous stretch of at least 300 nucleotides with identical sequence. There are 12 such families, with two or three members each (Figure 1). In addition to the *nif* family described above, five families are related to ERIS and the rest are various genes such as those that encode the BacS protein [35], or gene fragments.

As previously shown with pNGR234a [51], the DNA sequence allows prediction, identification and isolation of the potential rearrangements that may be generated by homologous recombination. The complete sequence of p42d will allow the identification of the precise sites of previously identified genomic rearrangements [50]. In the present study we have predicted the major potential rearrangements in p42d as it was previously described [51]; these include amplifications, deletions and inversions such as those illustrated in Figure 1.

In other SGCs the differences in number, organization, orientation and length of the reiterated elements predict specific genome rearrangements, as exemplified by the rearrangements that involve the *nifH* reiteration of p42d and pNGR234a [50,51].

### Genetic information of p42d in the context of other genomes

The putative protein sequences of p42d were compared to the proteomes of several complete bacterial genomes extracted from GenBank [52] (see Materials and methods) as well as to the SGC sequences available to date (Figure 3). We identified all pairs of potential orthologs between p42d and each of the genomes analyzed, following the strategy and definition described in Materials and methods. As expected, the highest percentage of orthologs common to p42d and to any other bacterial genome was found among the nitrogen-fixing symbiotic bacteria. *S. meliloti* and *M. loti* have, respectively, 51% and 45% of the orthologs found in p42d (see Additional data file 5). Members of the  $\alpha$ -proteobacterial subclass such as *Caulobacter crescentus*, *Brucella melitensis* and *A. tumefaciens* (a plant pathogenic member of the Rhizobiaceae) have from 25% to 32% of the orthologs present in p42d. The percentage of p42d orthologs within the genomes of plant pathogens varies from 17% for *Xyllella fastidiosa* to 31% for *Ralstonia solanacearum*. Human bacterial pathogens such as *Haemophilus influenzae* and *Helicobacter pylori*, those

with small genomes as *Rickettsia prowazekii* and *Mycoplasma genitalium*, and the archaea compared here, display the lowest number of shared orthologs. Instances of putative orthologs found in p42d and some complete bacterial genomes are shown in Figure 3a. In general, a collection of orthologs involved in diverse enzymatic activities is present in p42d and in most genomes compared here. They include the genes *hemN*<sub>1</sub>, *hemN*<sub>2</sub>, *ctrE*, *hisC*, *icfA*, *pgmV*, *aatC*, *pcaC*<sub>1</sub>, *adhE*, *ribAB*, *bglS*, *kprS*, *mcpG*, *mcpA* and *mmsB* (see Additional data file 1, for the assigned function). Their identity is in most cases 50% or lower.

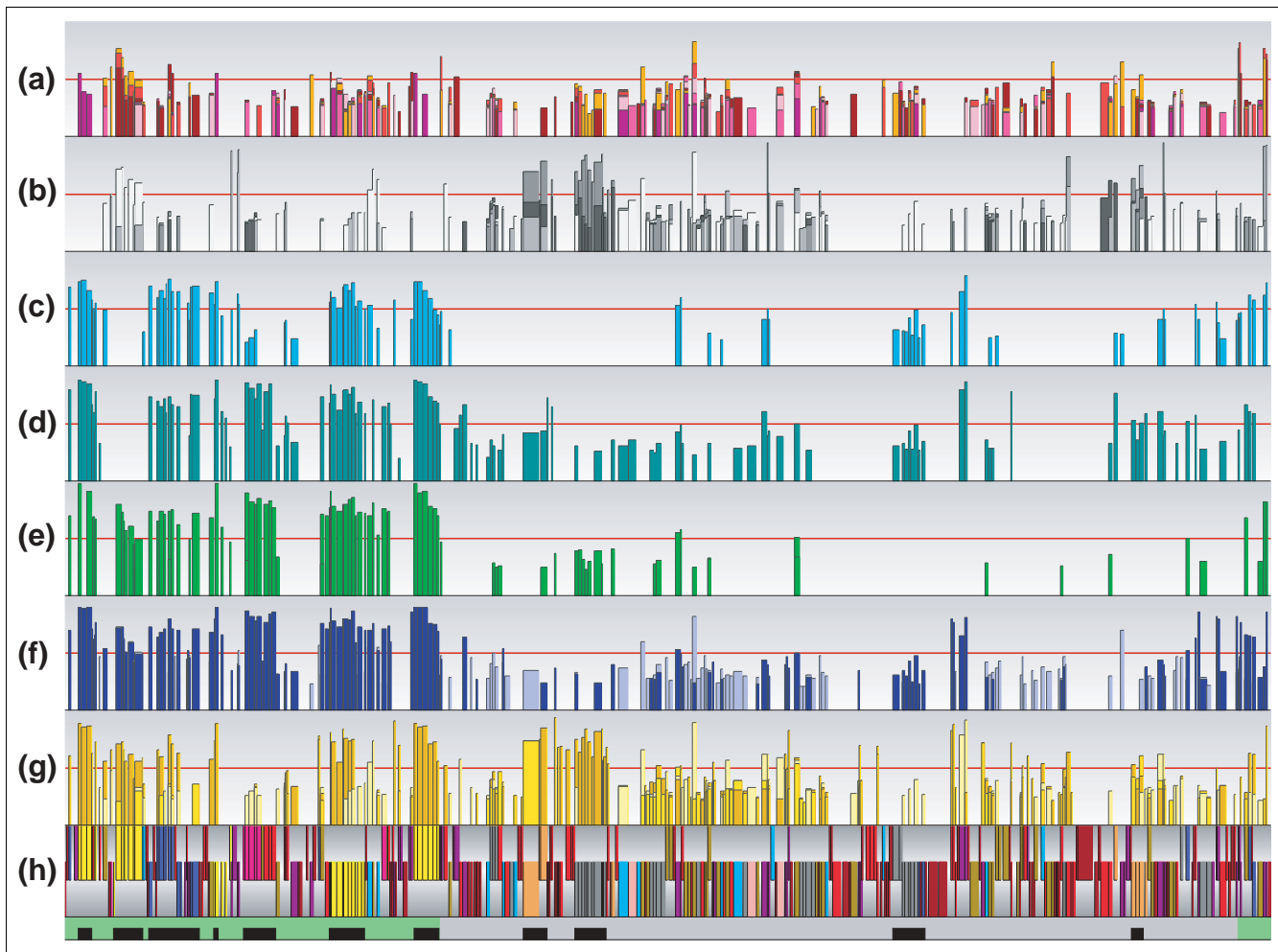
When we examined the distribution of orthologs in the six SGCs (see above), including p42d and using the genomes of *M. loti* and *S. meliloti* as reference, it was found that half of the hits lie in the respective SGCs and the rest are dispersed among other replicons, including the chromosomes (Table 1, Figure 3f,3g). In general, the genes for symbiosis are very well conserved in the SGCs, whereas the orthologs of genes not involved in symbiosis are distributed in nonsymbiotic plasmids and in the chromosomes (Figure 3c,3d,3e).

A total of 177 p42d CDSs (49%) have orthologs at least in one SGC. A subset of these (80 CDSs) belongs to the symbiotic region of 120 kb (Figure 3c,3d,3e,3f,3g, from NRA to NRb regions; Table 1) and the rest are interspersed in the remaining 251 kb of the plasmid. Among the SGCs compared, pNGR234a shares the highest percentage of orthologs (30%) with p42d (Table 1, Figure 3d), followed by the pSymA (28%) and the SGC of *M. loti* R7A (Table 1, Figure 3g and 3e, respectively). The SGC of *M. loti* MAFF303099 and *B. japonicum* share the fewest orthologs (24%) with p42d (Table 1, Figure 3f and 3c, respectively). The *A. tumefaciens* plasmids display the highest similarity with the TSSIV, TRA, and REP regions of p42d; the rest of the matches are distributed in the circular and the linear chromosomes (Figure 3b).

There are 20 genes common to all SGCs. These correspond exclusively to symbiotic genes including both nitrogen fixation (*nifHDKENXAB*, *fixABCX*, *fdxN*, *fdxB*) and nodulation (*nodABCJJD*) genes. The essential *nodBC* genes, however, have possible paralogs in some plant pathogens such as *A. tumefaciens*, *Ralstonia solanacearum* and *Xanthomonas*. Possible paralogs of the transport genes *nodIJ* are present in all the genomes analyzed. In these bacterial species, putative paralogs of *nod* genes might participate in the synthesis and secretion of outer membrane lipopolysaccharides [53].

### Conserved gene clusters in SGCs and other genomes

The *fixNOQPGHIS* common to different nitrogen-fixing symbiotic rhizobia are not always confined to the SGCs [10,11]. As mentioned above, in *R. etli* CFN42, the *fix* genes are distributed in two replicons, p42d and p42f, and some of them are reiterated, as is frequently observed in other genomes [2,3]. In *S. meliloti*, these genes are reiterated three times in pSymA [3] and in *M. loti* there are two copies of the entire operon [2].



**Figure 3**

Comparison of predicted proteins from p42d with those from other genomes and SGCs. Bidirectional best hits (BDBHs) between p42d and other genomes are shown. The bars in all rows represent the percentage identity (number of identities/length of the alignment) of BDBHs between p42d and the indicated genome (see below for color code). The horizontal red line in each row indicates 50% of similarity. A color code is shown for each genome or compartment. **(a)** Different organisms: *Bacillus subtilis* (dark magenta); *Brucella melitensis* (yellow); *Caulobacter crescentus* (red); *Escherichia coli* K12 (light magenta), *Methanobacterium thermoautotrophicum* (dark purple), and *Ralstonia solanacearum* (purple). **(b)** *A. tumefaciens* C58 circular chromosome (white), linear chromosome (pale gray), pAT (gray), and pTi (dark gray). **(c)** *B. japonicum* USDA110 SGC (turquoise). **(d)** pNGR234a (blue green). **(e)** *M. loti* R7A SGC (green). **(f)** *M. loti* MAFF303099 SGC (dark blue), and the rest of the chromosome (light blue). **(g)** *S. meliloti* pSymA (pale yellow), pSymB (yellow), and the chromosome (dark yellow). **(h)** CDS distribution for p42d with the color codes for functional classes and the relevant regions as indicated in Figure 1.

In *B. japonicum* they lie outside of the SGC (410 kb) determined by Gottfert *et al.* [11] but are included in the equivalent 681 kb SGC of the complete genome [7]. Moreover, in *Rhizobium* sp. NGR234 they are chromosomal [54]. The *fixNO-QPGHIS* cluster was identified in the genome of the plant pathogen *A. tumefaciens* (circular chromosome), the intracellular parasite *Brucella melitensis* (chromosome I), and in the free-living aquatic bacterium *C. crescentus*; all of them belonging to the  $\alpha$ -proteobacterial subdivision. Among  $\gamma$ -proteobacteria, the plant pathogen *Pseudomonas aeruginosa* has this *fix* cluster, which is absent in *E. coli*. Also, orthologs of this gene cluster are conserved in *R. solanacearum*, a plant pathogen that belongs to the  $\beta$ -proteobacteria.

The *fixABCX* operon is highly conserved in diazotrophs as well as in a wide variety of other bacterial and archaeal species such as *E. coli*, *Mycoplasma genitalium*, *Bacillus subtilis*, *Thermotoga maritima* and *Archeoglobus fulgidus*. In *E. coli* these *fix* genes are related to the carnitine pathway, but their function is unknown in the other species [55]. The ferredoxins FdxN and FdxB are always linked to *nif* genes in symbiotic as well as nonsymbiotic organisms. In *S. meliloti*, mutations in *fdxN* significantly impair the nitrogen-fixation process [29].

As mentioned above, the CPX cluster (9 kb, 15 CDSs) in p42d exhibits GC and CU profiles that diverge from the rest of the

**Table 1****Number of bidirectional best hits between pairs of SGCs or complete genomes**

	p42d									
pNGR234a	120	pNGR23a								
SGCBj	88*	133	SGCBj							
SmChr	63	86	59	SmChr						
pSymA	100	77	47	ND	pSymA					
pSymB	20	43	17	ND	ND	pSymB				
MIChr	62	127	66	2367	321	613	MIChr			
pMLa	15	29	8	18	17	10	ND	pMLa		
pMLb	5	12	11	8	23	12	ND	ND	pMLb	
SGCMI	81	116	79	ND	65	ND	ND	ND	ND	SGCMI
SGCR7A	101	135	89	86	73	54	30	21	2	240

Bidirectional best hits (BDBHs) were calculated in pairwise comparisons using BLASTP. All reciprocal matches with e-value up to  $1e^{-04}$  and a coverage of at least 50% on the length of the shorter CDS were collected. p42d, the symbiotic plasmid of *R. etli*, 371 kb, 359 CDS; pNGR234a, the symbiotic plasmid of *Rhizobium* sp. 536 kb, 416 CDS; SGCBj, *B. japonicum* USDA110 symbiotic chromosomal region, 410 kb, 388 CDS; SmChr, *S. meliloti* chromosome, 3,600 kb, 3396 CDS; pSymA, *S. meliloti* symbiotic plasmid A, 1,354 kb, 1,295 CDS; pSymB, *S. meliloti* symbiotic plasmid B, 1,683 kb, 1,571 CDS; MIChr, *M. loti* MAFF303099 chromosome without the symbiotic island, 6,425 kb, 6,172 CDS; pMLa, *M. loti* MAFF303099 cryptic plasmid a, 351 kb, 320 CDS; pMLb, *M. loti* MAFF303099 cryptic plasmid b, 208 kb, 209 CDS; SGCMI, *M. loti* MAFF303099 symbiotic island, 611 kb, 580 CDS; SGCR7A, *M. loti* R7A symbiotic island, 502 kb, 414 CDS. ND, not determined. \*The number of BDBHs with the complete genome of *B. japonicum* USDA110 is 150.

genes; CPX gene function is not known and no symbiotic role has so far been assigned to them. In *B. japonicum* some of these genes might participate in terpenoid synthesis [36]. The genes included in the CPX region showed similar organization in the SGC of *M. loti* (strains MAFF303099 and R7A), in pNGR234a, and in p42d. In *B. japonicum*, the CPX cluster was not located in the 410 kb SGC [11,36] but is present in the 680 kb SGC determined by Kaneko *et al.* [7]. In pSymA of *S. meliloti*, this cluster is partially represented by homologs of *cpxP2*, *cpxP4*, *ctrE* and some conserved hypothetical genes, *yp013-yp015*. Homologs of IS are located at the right border of the CPX region in pNGR234a and pSymA, while they are to the left of the SGC in *M. loti* R7A. The CPX region in p42d is flanked by ERIS, highlighting its potential for transposition.

A common feature in the SGCs is the presence of either the TSSIII or the TSSIV transport secretion systems. The TSSIII is found in pNGR234a, in the SGCs of *B. japonicum*, and in *M. loti* MAFF303099. The TSSIV is located in pSymA of *S. meliloti* and the symbiotic island of *M. loti* R7A. Both transport systems are present in p42d. In the pTi and pRi plasmids of *A. tumefaciens* C58, the TSSIV system is used for transferring the T-DNA to plant cells. In the absence of T-DNA in the SGCs, the precise function of these systems is not clear. Furthermore, both TSSIII and TSSIV are found in bacterial pathogens of plants and animals as well as in some  $\alpha$ -proteobacteria. Complete or partial TSSIII or TSSIV are present in *Brucella melitensis*, *C. crescentus*, *X. citri* and *X. campestris*, whereas in *Rickettsia prowazekii*, some *virB* genes are conserved. *P. aeruginosa* contains a complete

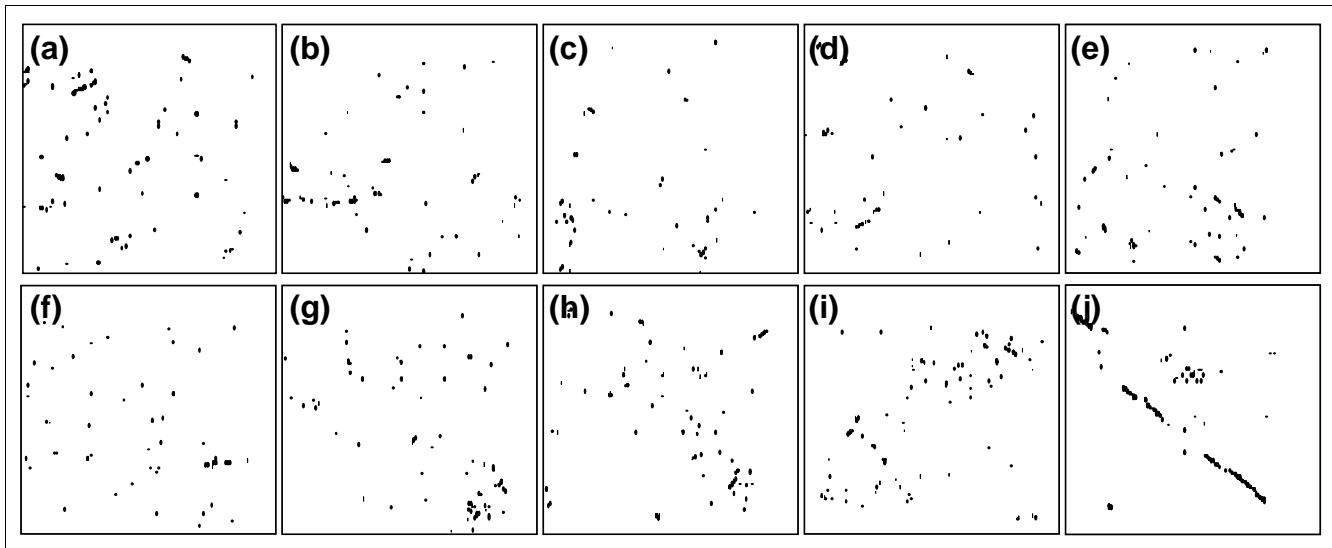
TSSIII but lacks homologs of the TSSIV, while in *Xyllella fastidiosa*, nine putative conjugative proteins of the plasmid pXF41 are clearly orthologs of the corresponding *virB* gene set found in other microorganisms.

### Absence of synteny among SGCs

It is generally known that gene order is conserved in closely related strains and species. The six SGCs compared here, except the SGCs of the two *M. loti* strains, have 20-30% of genes in common according to our estimates (Table 1). Most of these genes are located within the conserved clusters described above. Furthermore, genes unique to each of the individual genomes are interspersed among genes present in all in SGCs. For example, p42d contains 71 orphan genes throughout its structure.

The SGCs in *M. loti* strains MAFF303099 and R7A share large conserved segments that contain all the symbiotic genes [12] (Figure 4, panel 10). The colinearity is disrupted by genes unique to either of the SGCs. The smallest region that encloses the 20 common orthologous genes (essentially *nod* and *nif* genes) can be delimited to about 50 kb in pSymA, 120 kb in p42d, 250 kb in pNGR234a, 300 kb in the SGC of *B. japonicum*, and 320 kb in the two SGCs of *M. loti* (Figure 5). Such variability in gene order suggests that the SGCs have recombined frequently with other genome elements.

Several transcriptional units that are conserved in some SGCs appear to have undergone rearrangements in others. Examples taken from the *nif*, *fix* and *nod* operons are illustrated in



**Figure 4**

Analysis of synteny among the SGCs. Pairs of orthologous proteins among different genomes or SGCs are plotted. Each protein pair is shown according to the location of the corresponding coordinate of the predicted translation start of the gene on the DNA region. The axes correspond to the total length of the respective DNA region: p42d 371,255 bp; *M. loti* MAFF303099 symbiotic island 610,975 bp; *M. loti* R7A symbiotic island 502,000 bp; *S. meliloti* pSymA 354,226 bp; pNGR234a 536,165 bp and *B. japonicum* symbiotic region 410,573 bp. For each group the first region mentioned corresponds to the x-axis. (a) p42d vs pNGR234a; (b) p42d vs pSymA; (c) p42d vs *B. japonicum* symbiotic region; (d) p42d vs *M. loti* MAFF303099 symbiotic island; (e) p42d vs *M. loti* R7A, symbiotic island; (f) pNGR234a vs *S. meliloti* pSymA; (g) pNGR234a vs *B. japonicum* symbiotic region; (h) pNGR234a vs *M. loti* 303099 symbiotic island; (i) pNGR234a vs *M. loti* R7A symbiotic island; (j) *M. loti* MAFF303099 symbiotic island vs *M. loti* R7A symbiotic island.

Additional data file 6. The *nifHDK* and *nifENX*, are neighboring conserved transcriptional units in p42d, in the two SGCs of *M. loti*, and in pNGR234a. However, *nifH* and *nifN* are separated from their respective operons in the SGC of *B. japonicum* and in pSymA of *S. meliloti*, respectively. Similarly, *nodA* is located away from the *nodBC* genes in p42d, and *nodB* is distant in the SGC of *M. loti* strains. The operon *fixABCX* is disrupted in the SGC of *B. japonicum*, where *fixA* is in an operon with *nifA*. In turn, in other SGCs, *nifA* is commonly found in an operon with *nifB* and *fdxN*. Phylogenetic analyses of the 20 common genes in the six SGCs result in nonequivalent trees, even for genes that are organized in operons (data not shown). For example, trees derived from the genes of the operons *nifHDK* and *fixABCX* are incongruent, indicating that intraoperon recombination has been also frequent.

## Conclusions

Our results indicate that p42d contains several regions that significantly deviate from the average GC content and typical CU. The plasmid harbors a large amount of ERIS and several reiterated DNA families. In addition, it contains 10 pseudogenes. These features resemble those found in other SGCs. All SGCs sequenced so far are heterogeneous regarding their gene content, and the genes common to most of them are mainly those involved in nodulation and nitrogen fixation. Other common genes are present either in SGCs or in other genome locations (see above). The lack of synteny between p42d and the different SGCs analyzed gives further support to

the notion that the symbiotic compartments of rhizobial genomes are mosaic structures [10], presumably assembled from regions derived from diverse genomic contexts, that might have been frequently modified as a consequence of transposition, recombination and lateral transfer events.

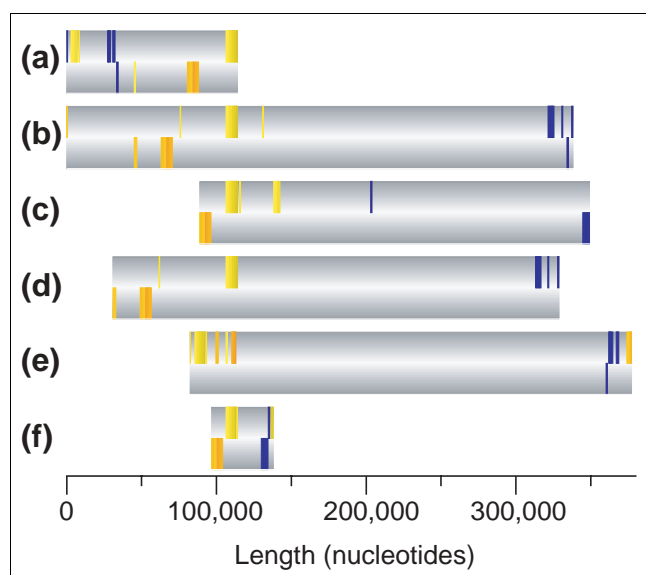
## Materials and methods

### Sequencing strategy

A minimal set of cosmids that covers the entire p42d [16] were used to generate shotgun libraries (1-2 kb mean insert size) cloned in M13 or pUC19 vectors. DNA sequencing reactions were performed using the Big-Dye Terminator kit in an automatic 373A DNA Sequencer (Applied Biosystems, Foster City, CA). Gaps were filled by a primer-walking strategy as well as by sequencing appropriate clones from pBR328 and pSUP202 libraries. A total of 6,210 readings of 450 bases in average were collected to achieve a coverage of 7× for the entire p42d.

### Assembly

Base calling was done using the program PHRED and the assembly was obtained by PHRAP [56,57]. Graphic representation and edition of the assembly were accomplished using the CONSED program [58]. Low-quality and single-stranded regions were located, and further sequencing was done to cover these areas. An error rate of less than 1 per 10,000 bases was estimated using base qualities determined by the PHRAP assembler. To confirm the assembly, pairs of



**Figure 5**  
Distribution of the 20 genes common to all the SGCs analyzed. (a) p42d; (b) *M. loti* MAFF303099 SGC; (c) pNGR234a; (d) *M. loti* R7A SGC; (e) *B. japonicum* SGC; (f) *S. meliloti* pSymA. The color bars indicate the position of the genes. The nodulation genes *nodABC DIJ* are represented in blue, and the nitrogen-fixation genes *nifHDKNEXAB*, *fixABCX* and *fdxBN* are represented in yellow.

forward and reverse primers were designed and used to raise overlapping PCR products with an average size of 5 kb, covering the entire plasmid in a single circular contig. The PCR products obtained agree well with the determined sequence (data not shown).

### CDS prediction and annotation

The coding capacity of p42d was determined by applying GLIMMER 2.02 [59,60] iteratively to enhance the overall prediction efficiency. Given the evidence indicating that GLIMMER-based predictions are less effective in plasmids [2], our approach also took into consideration the existence of several gene classes with different codon-usage (CU) patterns [61], and a potential ribosome-binding site (RBS) specific to p42d to aid GLIMMER in the selection of start codons.

#### RBS prediction

An initial set of presumably functional genes with a corresponding upstream RBS was detected by running BLASTX [62] comparisons (using a maximum e-value cutoff of 0.001) of the entire plasmid against the nonredundant (nr) database [52] at the National Center for Biotechnology Information (NCBI). All matches with hypothetical or putative proteins as well as those with an upstream neighbor hit closer than 50 bp were discarded to avoid genes within operons. We took into consideration only hits displaying an identity  $\geq 40\%$ , starting at the first amino acid, and alignment coverage of at least 80% of the matched protein. We then extended the selected hits

towards the 5' terminus and kept those with an upstream in-frame stop codon before any other possible start codon. This procedure left 21 hits. From the p42d sequence we extracted 20-bp regions upstream of the start codons of these hits and inferred the most probable RBS (6 bp in length) by applying the CONSENSUS program [63]. The resulting consensus matrix supported the sequence GGAGAG with an expected frequency of  $2.034 \times 10^{-8}$ .

#### CDS prediction

To train GLIMMER, we took the initial output of the BLASTX comparison detailed above, and selected as training set all hits with an alignment length  $\geq 100$  amino acids. Again, all matches with hypothetical/putative proteins were discarded. Overlapping hits matching the same protein were merged into a single larger hit, generating a total of 183 DNA segments. The RBS and the training set obtained were then used to run GLIMMER, yielding a prediction of 460 CDSs that included 93% of the 183 segments in the training set. However, we noticed that running GLIMMER iteratively yielded better results, because it produced a lower number of predicted CDSs and a greater number of segments in the training set mapping within predicted CDSs. We applied the method of A.M.-S., G.M.-H., A. Christen and J.C.-V. (unpublished work) to split the initial set of 460 CDSs into three groups displaying poor, typical and rich codon usage. Essentially, this method quantifies the extent to which individual genes use the most abundant codons in the plasmid. Each group was used as a training set and GLIMMER was run for 20 iterations to predict CDSs  $\geq 300$  bp (CDSs  $\geq 500$  bp in the first iteration composed the training set for the second iteration, and so forth). The best prediction for each CU group was selected, and the three resulting predictions were incorporated into a single one that produced 396 CDSs and recovered 97.75% of the initial training set.

#### Annotation

All CDSs were manually curated using BLASTX comparisons (e-value  $\leq 0.001$ ) against the nr database. The following criteria were applied to annotate the CDSs: CDSs were tagged as hypothetical (*yh*) when no homolog could be detected; hypothetical conserved CDSs (*yp*) were those displaying strong similarity to hypothetical proteins or weak similarity to known genes; CDSs with similarity  $\geq 50\%$  along the entire length of known genes were assigned the same name as the matching gene; CDSs related to insertion sequences (IS) and transposons (*yi*) were compared with BLASTN and BLASTX against the IS database [64] to identify the family they belong to. These elements were also analyzed for the presence of inverted repeats at their borders applying OLIGO 6.4 [65] and BLAST2 programs. Functional classification was carried out following the categories proposed in Freiberg *et al.* [10]. Additional support for annotation was obtained by searching for protein domains and motifs with the Interpro suite [66]. Transmembrane domains and leader peptides were searched using the PSORT program [67]. A relational database that

compiles all this information is available at [68], and Additional data file 1, which shows the set of 359 annotated CDS.

### Transcription units, RpoN promoters and regulatory binding sites

We predicted that all CDS in p42d are organized in 235 transcription units (TUs) by applying a previously reported distance-based methodology [69]. Binding sites were detected using upstream regions of variable length (but properly specified in each case) for all annotated CDS in the pSym. To identify genes potentially expressed by RpoN promoters, we compiled an initial training set containing 85 prokaryotic promoters for which the transcription start site has been experimentally mapped [38]. The CONSENSUS/PATSER set of programs [63] was then used to predict promoters 16 bp long in upstream regions of 250 bp. A final set of 37 RpoN promoters was obtained using as PATSER threshold the mean ( $\mu$ ) minus one standard deviation ( $\sigma$ ) estimated from the set of 85 promoters ( $\mu - 1\sigma = 6.33$ ). Binding sites for NifA or UAS were predicted using seven reported sites [27,70–73] as the training set. CONSENSUS/PATSER was run to predict sites of 16 bp in length within -400 to +50 bp regions, yielding 21 sites with PATSER score  $\geq 8.03$  ( $\mu - 1\sigma$ ); if a more stringent threshold is used instead, several known sites are undetected. We further discarded all predicted UAS without an associated RpoN promoter. In the case of *nod* boxes, we applied the dyad-sweeping method [74] to a set of six reported sites [75–78] in order to pinpoint the location of potential *nod* boxes in the p42d (as a conglomerate of five or more dyads), and then used CONSENSUS/PATSER to determine 47-bp sites within -600 to +50 bp regions.

Seven putative *nod* boxes were found by these approaches; however, several known functional sites were still undetected, and thus we trained CONSENSUS/PATSER again with the seven p42d sites found. Given that the mean PATSER score for these sites is too high (21.48), the usual threshold ( $\mu - 1\sigma$ ) is also correspondingly high (16.21), and thus we could not predict any additional sites. For these reasons, we decided to use as threshold the lowest PATSER score (9.15) obtained from the seven training sequences, in this way we finally predicted 15 *nod* boxes. CONSENSUS/PATSER programs were also applied to identify regulatory motifs for Fnr based on 30 known binding sites in *E. coli* extracted from RegulonDB [79]. Predictions were carried out in the -400 to +50 bp regions using as threshold the PATSER score  $\geq 6.2$  ( $\mu - 1\sigma$ ), which yielded 45 potential Fnr binding sites in p42d. If we get stricter and use the mean PATSER score (9.77) as threshold, only eight sites are detected. Nonetheless, given the evidence suggesting there is high transcriptional activity in the p42d under low-oxygen conditions [43], we decided to relax the score to allow more Fnr sites.

### Genome comparisons

Protein sequences from different genomes or symbiotic compartments were obtained from GenBank [52]: pNGR234a

U00090; *B. japonicum* USDA110 symbiotic region AF322012 and AF322013; *S. meliloti* AL591688; *M. loti* MAFF303039 NC\_002678; *M. loti* R7A symbiotic island AL672111; *A. tumefaciens* C58 (U. Washington) AE008688 and AE008689; *A. tumefaciens* C58 (Cereon) AE007869 and AE007870; *Ralstonia solanacearum* AL646052; *C. crescentus* AE005673; *Rickettsia prowazekii* AJ235269; *E. coli* O157:H7 BA000007; *E. coli* K12 U00096; *Brucella melitensis* AE008917; *P. aeruginosa* AE004091; *Xanthomonas citri* AE008923; *Nostoc* NC\_003272; *Xanthomonas campestris* AE008922; *Synechocystis* PCC6803 AB001339; *Xylella fastidiosa* AE003851; *Borrelia burgdorferi* AE000783; *Buchnera* sp. APS AP000398; *Mycoplasma genitalium* L43967; *Thermotoga maritima* AE000512; *Aquifex aeolicus* AE000657; *Archeoglobus fulgidus* AE000782; *Aeropyrum pernix* BA000002; *Methanobacterium thermoautotrophicum* AE000666; *Methanococcus jannaschii* L77117; *Methanopyrus kandleri* AE009439. Most probable orthologs were detected applying a previously reported method [69]. Essentially, the method performs BLASTP pairwise comparisons against the protein sequences of p42d, and bidirectional best hits (BDBHs) were used to define the most likely orthologous genes. All BDBHs with an e-value  $\leq 0.0001$  and alignment coverage of at least 50% of the smaller CDS were taken into consideration.

### Nucleotide sequence accession number

The nucleotide sequence reported here has been deposited in GenBank under the accession number U80928.

### Additional data files

The most relevant features of the functional annotation of the p42d can be found in Additional data file 1 available with the online version of this paper. It contains the name, the predicted protein size, the best nr-matching homolog, and the percentage of similarity/identity. Lists of predicted binding sites are shown in Additional data file 2 (*nod* boxes), Additional data file 3 (RpoN promoters and NifA UAS) and Additional data file 4 (anaeroboxes). The number of BDBHs between several complete genomes and the p42d is given in Additional data file 5. The topological representation of the 20 common genes in the six SGCs is detailed in Additional data file 6, A, p42d; B, *M. loti* MAFF303099; C, pNGR234a; D, *M. loti* R7A SGC; E, *B. japonicum* SGC; F, *S. meliloti* pSymA.

### Acknowledgements

We dedicate this paper to Rafael Palacios and Jaime Mora in gratitude for their support and stimulating critical discussions. We are grateful for the skillful technical support and advice given by J.A. Gama, R.E. Gómez, R.I. Santamaría, S. Caro, J. Espiritu, D. García, F. Sánchez, E. Díaz, E. Pérez-Rueda, V. del Moral, K.D. Noel, J. Sanjuan, M. Rosenblueth, P. Gaytán, E. López, P. Rabinowicz, and P.M. Reddy. This work was partially supported by a public grant from CONACyT (México) under the Program for Emerging Areas (N-028).

## References

- Garrity GM, Johnson KL, Bell JA, Searles DB: **Taxonomic outline of the Prokaryotes. Release 3.0.** In *Bergey's Manual of Systematic Bacteriology*. New York: Springer-Verlag 2002.
- Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K, Ishikawa A, Kawashima K, et al.: **Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*.** *DNA Res* 2000, **7**:331-338.
- Barnett MJ, Fisher RF, Jones T, Komp C, Abola AP, Barloy-Hubler F, Bowser L, Capela D, Galibert F, Gouzy J, et al.: **Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid.** *Proc Natl Acad Sci USA* 2001, **98**:9883-9888.
- Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J, Boistard P, Becker A, Boutry M, Cadieu E, et al.: **Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021.** *Proc Natl Acad Sci USA* 2001, **98**:9877-9882.
- Finan TM, Weidner S, Wong K, Buhrmester J, Chain P, Vorholter FJ, Hernández-Lucas I, Becker A, Cowie A, Gouzy J, et al.: **The complete sequence of the 1,683-kb pSymB megaplasmid from the N<sub>2</sub>-fixing endosymbiont *Sinorhizobium meliloti*.** *Proc Natl Acad Sci USA* 2001, **98**:9889-9894.
- Galibert F, Finan TM, Long SR, Pühler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, et al.: **The composite genome of the legume symbiont *Sinorhizobium meliloti*.** *Science* 2001, **293**:668-672.
- Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S, Watanabe A, Idesawa K, Iriguchi M, Kawashima K, et al.: **Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110.** *DNA Res* 2002, **9**:189-197.
- Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Quorllo B, Goldman BS, Cao Y, Askenazi M, Halling C, et al.: **Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58.** *Science* 2001, **294**:2323-2328.
- Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, et al.: **The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58.** *Science* 2001, **294**:2317-2323.
- Freiberg C, Fellay R, Bairoch A, Broughton WJ, Rosenthal A, Perret X: **Molecular basis of symbiosis between *Rhizobium* and legumes.** *Nature* 1997, **387**:394-401.
- Gottfert M, Rothlisberger S, Kundig C, Beck C, Marty R, Hennecke H: **Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the *Bradyrhizobium japonicum* chromosome.** *J Bacteriol* 2001, **183**:1405-1412.
- Sullivan JT, Trzebiatowski JR, Cruickshank RW, Gouzy J, Brown SD, Elliot RM, Fleetwood DJ, McCallum NG, Rossbach U, Stuart GS, et al.: **Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A.** *J Bacteriol* 2002, **184**:3086-3095.
- Encarnación S, Dunn M, Willsms K, Mora J: **Fermentative and aerobic metabolism in *Rhizobium etli*.** *J Bacteriol* 1995, **177**:3058-3066.
- Romero D, Brom S, Martínez-Salazar J, Girard ML, Palacios R, Dávila G: **Amplification and deletion of a *nod-nif* region in the symbiotic plasmid of *Rhizobium phaseoli*.** *J Bacteriol* 1991, **173**:2435-2441.
- Brom S, García de los Santos A, Stepkowsky T, Flores M, Dávila G, Romero D, Palacios R: **Different plasmids of *Rhizobium leguminosarum* bv. *phaseoli* are required for optimal symbiotic performance.** *J Bacteriol* 1992, **174**:5183-5189.
- Girard ML, Flores M, Brom S, Romero D, Palacios R, Dávila G: **Structural complexity of the symbiotic plasmid of *Rhizobium leguminosarum* bv. *phaseoli*.** *J Bacteriol* 1991, **173**:2411-2419.
- Ramírez-Romero MA, Bustos P, Girard L, Rodríguez O, Cevallos MA, Dávila G: **Sequence, localization and characteristics of the replicator region of the symbiotic plasmid of *Rhizobium etli*.** *Microbiology* 1997, **143**:2825-2831.
- Poupot R, Martínez-Romero E, Gautier N, Prome JC: **Wild type *Rhizobium etli*, a bean symbiont, produces acetyl-fucosylated, N-methylated, and carbamoylated nodulation factors.** *J Biol Chem* 1995, **270**:6050-6055.
- Cárdenas L, Domínguez J, Santana O, Quinto C: **The role of the *nodI* and *nodJ* genes in the transport of Nod metabolites in *Rhizobium etli*.** *Gene* 1996, **173**:183-187.
- Schlaman HR, Phillips D, Kondorosi E: **Genetic organization and transcriptional regulation of rhizobial nodulation genes. In *The Rhizobiaceae*.** Edited by: Spink HP, Kondorosi A, Hoykaas PJJ. Dordrecht: Kluwer 1998, 361-386.
- Rostas K, Kondorosi E, Horváth B, Simoncsits A, Kondorosi A: **Conservation and extended promoter regions of nodulation genes in *Rhizobium*.** *Proc Natl Acad Sci USA* 1986, **83**:1757-1761.
- Quinto C, de la Vega H, Flores M, Fernández L, Ballado T, Soberón G, Palacios R: **Reiteration of nitrogen fixation gene sequences in *Rhizobium phaseoli*.** *Nature* 1982, **299**:724-726.
- Rodríguez C, Romero D: **Multiple recombination events maintain sequence identity among members of the nitrogenase multigene family in *Rhizobium etli*.** *Genetics* 1998, **149**:785-794.
- Earl CD, Ronson CW, Ausubel FM: **Genetic and structural analysis of the *Rhizobium meliloti* *fixA*, *fixB*, *fixC*, and *fixX* genes.** *J Bacteriol* 1987, **169**:1127-1136.
- Kaminski PA, Batut J, Boistard P: **A survey of symbiotic nitrogen fixation by *Rhizobia*.** In *The Rhizobiaceae*. Edited by: Spink HP, Kondorosi A, Hoykaas PJJ. Dordrecht: Kluwer 1998, 431-460.
- Kim S, Burgess BK: **Evidence for the direct interaction of the *nifW* gene product with the MoFe protein.** *J Biol Chem* 1996, **271**:9764-9770.
- Lee HS, Berger DK, Kustu S: **Activity of purified NIFA, a transcriptional activator of nitrogen fixation genes.** *Proc Natl Acad Sci USA* 1993, **90**:2266-2270.
- Dombrecht B, Tesfay MZ, Verreth C, Heusdens C, Napoles MC, Vanderleyden J, Michiels J: **The *Rhizobium etli* gene *iscN* is highly expressed in bacteroids and required for nitrogen fixation.** *Mol Genet Genomics* 2002, **267**:820-828.
- Klipp W, Reilander H, Schluter A, Krey R, Pühler A: **The *Rhizobium meliloti* *fdxN* gene encoding a ferredoxin-like protein is necessary for nitrogen fixation and is cotranscribed with *nifA* and *nifB*.** *Mol Gen Genet* 1989, **216**:293-302.
- Hoover TR, Robertson AD, Cerny RL, Hayes RN, Imperial J, Shah VK, Ludden PW: **Identification of the V factor needed for synthesis of the iron-molybdenum cofactor of nitrogenase as homocitrate.** *Nature* 1987, **329**:855-857.
- Imperial J, Ugalde RA, Shah VK, Brill WJ: **Role of the *nifQ* gene product in the incorporation of molybdenum into nitrogenase in *Klebsiella pneumoniae*.** *J Bacteriol* 1984, **158**:187-194.
- Fischer HM: **Genetic regulation of nitrogen fixation in *rhizobia*.** *Microbiol Rev* 1994, **58**:352-386.
- Michiels J, Van Soom T, D'Hooghe I, Dombrecht B, Benhassine T, de Wilde P, Vanderleyden J: **The *Rhizobium etli* *rpoN* locus: DNA sequence analysis and phenotypal characterization of *rpoN*, *ptsN*, and *ptsA* mutants.** *J Bacteriol* 1998, **180**:1729-1740.
- Valderrama B, Dávalos A, Girard L, Morett E, Mora J: **Regulatory proteins and cis-acting elements involved in the transcriptional control of *Rhizobium etli* reiterated *nifH* genes.** *J Bacteriol* 1996, **178**:3119-3126.
- Jahn OJ, Dávila G, Romero D, Noel KD: **BacS: An abundant bacteroid protein in *Rhizobium etli* whose expression *ex planta* requires *NifA*.** *Mol Plant-Microbe Interact* 2003, **16**:65-73.
- Tully RE, Keyster DL: **Cloning and mutagenesis of a cytochrome p-450 locus from *Bradyrhizobium japonicum* that is expressed anaerobically and symbiotically.** *Appl Environ Microbiol* 1993, **59**:4136-4142.
- Dombrecht B, Marchal K, Vanderleyden J, Michiels J: **Prediction and overview of the RpoN-regulon in closely related species of the Rhizobiales.** *Genome Biol* 2002, **3**:research0076.1-0076.11.
- Barríos H, Valderrama B, Morett E: **Compilation and analysis of sigma(54)-dependent promoter sequences.** *Nucleic Acids Res* 1999, **27**:4305-4313.
- Girard L, Brom S, Dávalos A, López O, Soberón M, Romero D: **Differential regulation of *fixN*-reiterated genes in *Rhizobium etli* by a novel *fixL* - *fixK* cascade.** *Mol Plant Microbe Interact* 2000, **13**:1283-1292.
- Batut J, Daveran-Mingot ML, David M, Jacobs J, Garnerone AM, Kahn D: ***fixK*, a gene homologous with *fnr* and *crp* from *Escherichia coli*, regulates nitrogen fixation genes both positively and negatively in *Rhizobium meliloti*.** *EMBO J* 1989, **8**:1279-1286.
- López O, Morera C, Miranda-Ríos J, Girard L, Romero D, Soberón M: **Regulation of gene expression in response to oxygen in *Rhizobium etli*: role of FnrN in *fixNOQP* expression and in symbiotic nitrogen fixation.** *J Bacteriol* 2001, **183**:6999-7006.
- Lynch AS, Lin ECC: **Responses to molecular oxygen.** In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. Edited by: Neidhardt FC. Washington, DC: American Society of Microbiology 1996, 1526-1538.

43. Girard ML, Valderrama B, Palacios R, Romero D, Dávila G: **Transcriptional activity of the symbiotic plasmid of *Rhizobium etli* is affected by different environmental conditions.** *Microbiology* 1996, **142**:2847-2856.
44. Cornelis GR, Van Gijsegem F: **Assembly and function of type III secretory systems.** *Annu Rev Microbiol* 2000, **54**:735-774.
45. Viprey V, Del Greco A, Golinowski W, Broughton WJ, Perret X: **Symbiotic implications of type III protein secretion machinery in *Rhizobium*.** *Mol Microbiol* 1998, **28**:1381-1389.
46. Christie PJ: **Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines.** *Mol Microbiol* 2001, **40**:294-305.
47. Brom S, García-de los Santos A, Cervantes L, Palacios R, Romero D: **In *Rhizobium etli* symbiotic plasmid transfer, nodulation competitiveness and cellular growth require interaction among different replicons.** *Plasmid* 2000, **44**:34-43.
48. Kitabatake M, Ali K, Demain A, Sakamoto K, Yokoyama S, Soll D: **Indolmycin resistance of *Streptomyces coelicolor* A3(2) by induced expression of one of its two tryptophanyl-tRNA synthetases.** *J Biol Chem* 2002, **277**:23882-23887.
49. Flores M, Brom S, Stepkowski T, Girard ML, Dávila G, Romero D, Palacios R: **Gene amplification in *Rhizobium*: identification and *in vivo* cloning of discrete amplifiable DNA regions (amplicons) from *Rhizobium leguminosarum biovar phaseoli*.** *Proc Natl Acad Sci USA* 1993, **90**:4932-4936.
50. Romero D, Martínez-Salazar J, Girard L, Brom S, Dávila G, Palacios R, Flores M, Rodríguez C: **Discrete amplifiable regions (amplicons) in the symbiotic plasmid of *Rhizobium etli* CFN42 *J Bacteriol* 1995, **177**:973-980.**
51. Flores M, Mavingui P, Perret X, Broughton WJ, Romero D, Hernández G, Dávila G, Palacios R: **Prediction, identification, and artificial selection of DNA rearrangements in *Rhizobium*: toward a natural genomic design.** *Proc Natl Acad Sci USA* 2000, **97**:9138-9143.
52. **GenBank** [<http://www.ncbi.nlm.nih.gov/Genbank>]
53. Vázquez M, Santana O, Quinto C: **The NodL and NodJ proteins from *Rhizobium* and *Bradyrhizobium* strains are similar to capsular polysaccharide secretion proteins from gram-negative bacteria.** *Mol Microbiol* 1993, **8**:369-377.
54. Viprey V, Rosenthal A, Broughton WJ, Perret X: **Genetic snapshots of the *Rhizobium* species NGR234 genome.** *Genome Biol* 2000, **1**:research0014.1-0014.17.
55. Buchet A, Nasser W, Eichler K, Mandrand-Berthelot MA: **Positive co-regulation of the *Escherichia coli* carnitine pathway *cai* and *fix* operons by CRP and the *CaiF* activator.** *Mol Microbiol* 1999, **34**:562-575.
56. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
57. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
58. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8**:195-202.
59. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27**:4636-4641.
60. Salzberg SL, Delcher AL, Kasif S, White O: **Microbial gene identification using interpolated Markov models.** *Nucleic Acids Res* 1998, **26**:544-548.
61. Hayes WS, Borodovsky M: **How to interpret an anonymous bacterial genome: machine learning approach to gene identification.** *Genome Res* 1998, **8**:1154-1171.
62. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
63. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**:563-577.
64. **IS database** [<http://www-is.biotoul.fr>]
65. **Oligo 6.0 primer analysis software** [<http://www.oligo.net>]
66. **InterPro database** [<http://www.ebi.ac.uk/interpro/scan.html>]
67. **Psort Software** [<http://psort.nibb.ac.jp/index.html>]
68. **R. *etli* database** [<http://www.cifn.unam.mx/retlidb>]
69. Moreno-Hagelsieb G, Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes.** *Bioinformatics* 2002, **18**(Suppl 1):S329-S336.
70. Barrios H, Grande R, Olvera L, Morett E: ***In vivo* genomic footprinting analysis reveals that the complex *Bradyrhizobium japonicum* *fixRnifA* promoter region is differently occupied by two distinct RNA polymerase holoenzymes.** *Proc Natl Acad Sci USA* 1998, **95**:1014-1019.
71. Charlton W, Cannon W, Buck M: **The *Klebsiella pneumoniae* *nif* promoter: analysis of promoter elements regulating activation by the *NifA* promoter.** *Mol Microbiol* 1993, **7**:1007-1021.
72. Morett E, Buck M: ***NifA*-dependent *in vivo* protection demonstrates that the upstream activator sequence of *nif* promoters is a protein binding site.** *Proc Natl Acad Sci USA* 1988, **85**:9401-9405.
73. Preker P, Hubner P, Schmehl M, Klipp W, Bickle TA: **Mapping and characterization of the promoter elements of the regulatory *nif* genes *rpoN*, *nifA1* and *nifA2* in *Rhodobacter capsulatus*.** *Mol Microbiol* 1992, **6**:1035-1047.
74. Benítez-Bellón E, Moreno-Hagelsieb G, Collado-Vides J: **Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA.** *Genome Biol* 2002, **3**:research0013.1-0013.16.
75. Baev N, Endre G, Petrovics G, Banfalvi Z, Kondorosi A: **Six nodulation genes of nod box locus 4 in *Rhizobium meliloti* are involved in nodulation signal production: *nodM* codes for D-glucosamine synthetase.** *Mol Gen Genet* 1991, **228**:113-124.
76. Fisher RF, Long SR: **DNA footprint analysis of the transcriptional activator proteins *NodD1* and *NodD3* on inducible nod gene promoters.** *J Bacteriol* 1989, **171**:5492-5502.
77. Suominen L, Paulin L, Saano A, Saren AM, Tas E, Lindstrom K: **Identification of nodulation promoter (*nod*-box) regions of *Rhizobium galegae*.** *FEMS Microbiol Lett* 1999, **177**:217-223.
78. Wang SP, Stacey G: **Studies of the *Bradyrhizobium japonicum* *nodD1* promoter: a repeated structure for the nod box.** *J Bacteriol* 1991, **173**:3356-3365.
79. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millán-Zárate D, Díaz-Peredo E, Sánchez-Solano F, Pérez-Rueda E, Bonavides-Martínez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12.** *Nucleic Acids Res* 2001, **29**:72-74.