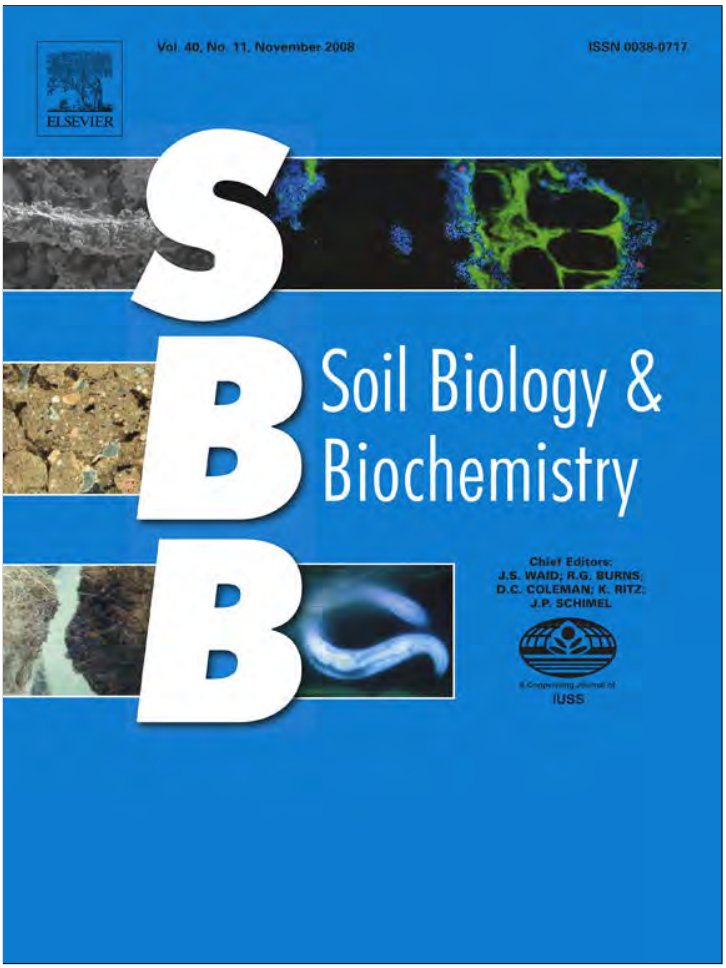

Soil Biology & Biochemistry 40 (2008) 2743–2753

Contents lists available at ScienceDirect

Soil Biology & Biochemistry

journal homepage: www.elsevier.com/locate/soilbio

Genomic panorama of *Bradyrhizobium japonicum* CPAC 15, a commercial inoculant strain largely established in Brazilian soils and belonging to the same serogroup as USDA 123

Leandro Pereira Godoy [a,b], Ana Tereza Ribeiro Vasconcelos [c,d], Ligia Maria Oliveira Chueire [a,d], Rangel Celso Souza [c], Marisa Fabiana Nicolás [c], Fernando Gomes Barcellos [a,b,e], Mariangela Hungria [a,b,d,*]

[a] *Embrapa Soja, Cx. Postal 231, CEP 86001-970 Londrina, Paraná, Brazil*
[b] *Universidade Estadual de Londrina, Department of Microbiology, Londrina, Brazil*
[c] *Laboratório Nacional de Computação Científica, Rua Getúlio Vargas 333, CEP 25651-071, Petrópolis, Rio de Janeiro, Brazil*
[d] *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-MCT), Brasília, Federal District, Brazil*
[e] *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasília, Federal District, Brazil*

ARTICLE INFO

Article history:
Received 28 March 2008
Received in revised form 18 July 2008
Accepted 18 July 2008
Available online 13 August 2008

Keywords:
Biological nitrogen fixation
Bradyrhizobium japonicum
Genome
Glycine max

ABSTRACT

Of the many genomes of prokaryotes that have been sequenced, most are pathogenic organisms and very few of agriculturally beneficial bacteria. Soybean, the most important cash crop in Brazil, can provide its need for nitrogen through a symbiosis with exotic strains of bradyrhizobia. *Bradyrhizobium japonicum* strain CPAC 15 (equates to SEMIA 5079, the same serogroup as USDA 123), which is a highly competitive commercial strain applied to soybean crops since the early 1990s, is now established on several millions of hectares. As financial resources for sequencing genomes are still very limited in developing countries, a panoramic genomic view of CPAC 15 was generated. A total of 4328 shotgun reads resulted in 2,046,740 bp with a phred score ≥ 20; the assemblage resulted in 1106 phrap contigs scattered by 69 scaffolds and 966 isolated contigs, with an average of 2.5 reads per contig, covering approximately 13% of the genome. Annotation identified 1371 coding DNA sequences (CDSs), 53% with putative known functions, 23% encoding conserved hypothetical and 24% hypothetical genes, representing about 16% of the estimated putative genes. Several comparisons – on COG and KEGG databases, tRNAs, transposases, G + C content of CPAC 15 with the complete genome of *B. japonicum* strain USDA 110 indicated a successful coverage of the whole genome. However, the two strains were surprisingly different, as at least 35% of the CDS of CPAC 15 shows higher similarity to microorganisms other than strain USDA 110. Several new putative genes and others with low similarity to USDA 110, were identified. These were related to nodulation, interaction with the host plant and adaptation, e.g., *nodB*, *nodW*, *ndvA*, effector *nopP*, genes of secretion systems, transporters and environmentally related genes.

1. Introduction

Prokaryotes were established on Earth billions of years before the first single-celled eukaryotes evolved, and the long period of adaptation to a variety of environments explains their enormous diversity and versatility. The genomes of almost 700 prokaryotes have been sequenced – with several others in progress or accessible as drafts (GOLD, 2008) – Genomes OnLine Database). Variability in size, number, density and organization of genes in operons has been demonstrated, and each new genome provides unexpected and sometimes intriguing information. For example, small genomes such as that of the thermophile *Nanoarchaeum equitans* strain Kin4-M (490,885 bp; Waters et al., 2003) may reveal the minimum set of essential genes and metabolic processes necessary to guarantee survival (Ochman, 2005). Large genomes such as that of *Burkholderia xenovorans* strain LB400 (9,731,138 bp; Chain et al., 2006) may provide understanding of events such as horizontal gene transfer and reveal functions of large sets of genes (Kaneko et al., 2002; Chain et al., 2006). Even within the same species, remarkable differences have been reported; for example, the genome of *Escherichia coli* strain K12 was estimated at 4,646,332 bp coding for 4337 genes (Riley et al., 2006), whereas a larger genome was reported for strain O157:H7 – 5,498,450 bp coding for 5449

* Corresponding author. Embrapa Soja, Cx. Postal 231, CEP 86001-970 Londrina, Paraná, Brazil. Tel.: +55 43 33716206; fax: +55 43 33716100.
E-mail addresses: godoylp@hotmail.com (L.P. Godoy), atrv@lncc.br (A.T.R. Vasconcelos), ligia@cnpso.embrapa.br (L.M.O. Chueire), rangel@lncc.br (R.C. Souza), marisa@lncc.br (M.F. Nicolás), fgbarcel@yahoo.com.br (F.G. Barcellos), hungria@cnpso.embrapa.br, hungria@pq.cnpq.br (M. Hungria).

0038-0717/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.soilbio.2008.07.016

genes (Perna et al., 2001). Furthermore, the photosynthetic *Bradyrhizobium* sp. strain BTAi1 has more than 1 million bp than the ORFS278 strain (Giraud et al., 2007). However, the large majority of sequenced genomes are of pathogenic microorganisms, with emphasis on agents of human disease (GOLD – Genomes OnLine Database).

Soybean was introduced to Brazil 125 years ago and is now grown on over 22 million hectares, about 45% of the country's cropped land (Hungria et al., 2006). This legume's need for N may be supplied by a symbiosis in which bacteria belonging mainly to the species *Bradyrhizobium japonicum* and *B. elkanii* penetrate the roots resulting in the formation of nodules in which atmospheric N_2 is fixed and passed on to the host plant. Research on N_2 fixation, with emphasis on identification of superior strains of bradyrhizobia, has been a chief contributor to the success of the crop in Brazil, fostering high yields with low input costs (Hungria et al., 2006).

Strain-selection programs for soybean were established in Brazil with the first commercial plantings in the early 1960s, and *B. japonicum* SEMIA 566 – isolated from the nodules of a vigorous soybean plant inoculated with a North American inoculant – was one of the first "selected" strains. It belongs to the same serogroup as USDA 123 – recognized as the most competitive in the USA (Ham et al., 1971; Weber et al., 1989) – and was employed in commercial inoculants from 1966 to 1978, greatly contributing to the successful establishment of the crop in southern Brazil (Hungria et al., 2006). A decade later, higher demands for N by newer, more-productive soybean genotypes necessitated identification of strains with greater capacity for N_2 fixation and tolerance to the more stressful environmental conditions of newly cropped tropical regions in the Brazilian Cerrados. Selection emphasis was then on identification of adapted naturalized strains with relatively high efficiency of N_2 fixation, and CPAC 15 (=SEMIA 5079) was recognized as a very effective variant strain derived from SEMIA 566 (Hungria and Vargas, 2000; Hungria et al., 2006). CPAC 15 has been extensively used in Brazilian commercial inoculants since 1992, and the high competitiveness of both SEMIA 566 and CPAC 15 explain the dominance of this serogroup in practically all areas cropped to soybean, as well as its dispersion and establishment in non-inoculated areas (Freire et al., 1983; Vargas et al., 1993; Ferreira and Hungria, 2002; Mendes et al., 2004; Hungria et al., 2006; Batista et al., 2007).

When considering genome sequencing of prokaryotes, three main points should be taken into account. Firstly, despite the importance of the symbiotic N_2 fixation to the global N cycle (Newton, 2000), only a few complete genomes of rhizobia have been published so far: *Mesorhizobium loti* strain MAFF303099 (Kaneko et al., 2000), *Sinorhizobium meliloti* strain 1021 (Galibert et al., 2001), *B. japonicum* strain USDA 110 (Kaneko et al., 2002), *Rhizobium etli* bv. phaseoli strain CFN 42 (González et al., 2006), *R. leguminosarum* biovar *viciae* strain 3841 (Young et al., 2006), *Bradyrhizobium* sp. strains ORS278 and BTAi1 (Giraud et al., 2007), and *Azorhizobium caulinodans* strain ORS571 (Lee et al., 2008). Secondly, regardless of the economic and environmental importance of biological N_2 fixation to Brazilian agriculture, research funding for genomic studies in the country remains low, as in the great majority of developing countries. Finally, contrary to initial predictions (Chothia, 1992), the number of new protein families grows with each new genome sequenced, particularly among prokaryotes (Kunin et al., 2003), and mathematical models predict that, even with hundreds of genomes per species, new protein families with important and novel biochemical properties may yet be discovered (Tettelin et al., 2005). Therefore, creative, low-cost initiatives are needed to allow prospecting of genes of important microorganisms such as strain CPAC 15. A possibility is the partial sequencing of the genome, as proposed by (Viprey et al., 2000), allowing understanding of the main classes of genes and their distribution on the genome. Therefore, this study aimed at partially covering the genome of *B. japonicum* strain CPAC 15, to detect the main classes of genes and their similarities to those of other rhizobial genomes.

2. Material and methods

2.1. Rhizobial strain and growth conditions

B. japonicum strain CPAC 15 [CPAC refers to Embrapa-Centro de Pesquisa Agropecuária dos Cerrados, Planaltina, Distrito Federal, Brazil; the designation of the strain at the National collection of rhizobia, at FEPAGRO (Fundação Estadual de Pesquisa Agropecuária, Porto Alegre, Rio Grande do Sul, Brazil) is SEMIA 5079; other designations for the strain are 566a and DF 24] was obtained from the "Diazotrophic and Plant Growth Promoting Bacteria Culture Collection" of Embrapa Soja. The strain was grown on Luria–Bertani (LB) medium (Sambrook et al., 1989) for five days at 28 °C, pellets were obtained after centrifugation at 10,000*g* for 20 min and were stored at −70 °C.

2.2. Library construction

Shotgun libraries of strain CPAC 15 were prepared as described before (Vasconcelos et al., 2003) and involved DNA purification and random mechanical shearing by nebulization. Total DNA mixed with glycerol at 50% and NaOAc 3 M was nebulized by 30 s at 2 kgf cm^{-2} to obtain fragments ranging from 1 to 3 kb, that were repaired using T4 DNA polymerase, polynucleotide kinase (PNK), and Klenow polymerase of *E. coli*, and size-fractionated by low melting agarose (Promega) gel electrophoresis. After the extraction from the gel using the "Gel Extraction Kit" from Qiagen, fragments were cloned into the vector pUC18. The vector pUC18 used to clone the DNA fragments was previously digested with *Sma*I and dephosphorylated with the enzyme BAP (*b*acterial *a*lkaline *p*hosphatase), and then the DNA fragments were ligated to pUC18 with the use of T4 DNA ligase (Invitrogen™). For the transformation *E. coli* strain DH10B was used. Transformants were plated on LB medium containing ampicillin (250 μg mL^{-1}), 5-bromo-4-chloro-indoyl-β-D-galactoside (20 μL of an stock of 50 μg μL^{-1}) and isopropyl-β-thiogalactopyranoside (100 μL of 0.1 M IPTG) and grown overnight at 37 °C. Recombinant clones were identified and transferred to 96-well plates containing "Terrific Grow – TB" (Invitrogen™) medium with ampicillin (250 μg mL^{-1}) and after growing were maintained in 80% glycerol (Sigma, >99.5%) at −70 °C.

2.3. Sequencing of shotgun clones

Individual colonies of the libraries were inoculated in 96-well microplates containing "Terrific Grow – TB" medium with ampicillin (250 μg mL^{-1}) and grown at 150 rpm for 16 h at 37 °C. DNA was extracted by the usual method of alkaline lysis (Sambrook et al., 1989), with a modification in the final procedure, passing the supernatant by multiple filters (MultiScreen, Millipore) before the DNA precipitation (Vasconcelos et al., 2003). Purified DNA was resuspended in water and verified in 0.8% (1%) gel agarose, as described by Sambrook et al. (1989). DNA was precipitated with 3 M KOAc and then sequenced utilizing the DYEnamic™ ET dye terminator cycle sequencing (MegaBACE™) kit (Amersham Pharmacia Biotech). The PCR reactions were performed using universal and reverse (Invitrogen™ or RW genes) primers. PCR-products were analyzed on a MegaBACE 1000 capillary sequencer (Amersham Pharmacia Biotech). The quality of a library was checked by sequencing a small number of plasmids, which were assembled with the phrap program (Ewing et al., 1998) and aligned by BLAST (www.ncbi.nlm.nih.gov) to validate the randomness of the library and the proportion of vector sequences.

2.4. Assembling and annotation

Sequences were assembled and annotated using the software System for Automated Bacterial Integrated Annotation (SABIA; Almeida et al., 2004), developed to integrate public-domain software's. For the assembly, SABIA includes the programs Phred, Phrap and Consed (Ewing et al., 1998; Gordon et al., 1998). SABIA identifies the gene coding regions using "Glimmer" (Delcher et al., 1999). The sequences were compared with those deposited at the GenBank database using the comparison with BLAST (http://www.ncbi.nlm.nih.gov) for nucleotides (BLASTN) and for proteins (BLASTP). Metabolic pathways were predicted based on KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000). Protein sequences were compared by COG (Clusters of Orthologous Groups of Proteins) (Tatusov et al., 2000), INTERPRO (families, domains and functional sites of proteins) (Apweiler et al., 2000), PSORT (protein localization) (Nakai and Kanehisa, 1991), TCDB (http://tcdb.org) and UniProt (UniProt, 2007), for protein information. Non-coding regions were identified with a software that searches for ribosomal binding sites and identifies promoters and operators. The size for the each CDS [coding DNA sequence(s)] without database homologues was established at a minimum of 50 bp, and in the case of shorter sequences an observation was included in the CDS description. All CDSs were manually curated using a cutoff E value of $\approx 10^{-10}$ and $\approx 50\%$ of identity considering the BLASTP. The following criteria were applied to the annotation: CDSs were target as hypothetical when no homolog could be detected; conserved hypothetical CDSs were those displaying strong similarity with several hypothetical proteins or weak similarity to known genes; and CDSs with high similarity with known genes were assigned the same name as the matching gene.

Supplementary information is available at http://www.bnf.lncc.br/bj/final/main.html (or http://www.bnf.lncc.br, then *B. japonicum* SEMIA 5079, genome draft).

3. Results and discussion

3.1. General genome features

The panorama of strain CPAC 15 was built with 4328 shotgun reads of clones submitted both in forward and reverse directions, and, excluding vectors, resulted in the deposit of 4,576,417 base pairs (bp), 2,046,740 bp of which with a phred score of $\geq$20, with a final assemblage of phrap contigs of 1,184,994 bp (Table 1). The genome of *B. japonicum* strain USDA 110 has 9,105,828 bp (Kaneko et al., 2002), thus the genome coverage of CPAC 15 is estimated at about 51%, and at 13% considering the coverage with phrap contigs, close to the 10% suggested in the snapshot approach of Viprey et al. (2000). More general information about the genome is given in Table 1, and all supplementary information is available at http://www.bnf.lncc.br/bj/final/main.html (or http://www.bnf.lncc.br, then *B. japonicum* SEMIA 5079, genome draft).

After manual annotation, using the criterion described in Section 2, 1371 CDSs were confirmed and classified as follows: 53% with putative known functions, 23% as conserved hypothetical and 24% as hypothetical genes (Table 1). The percentages were very close to those reported for strain USDA 110, of 52%, 30% and 18%, respectively (Kaneko et al., 2002), indicating that representative coverage of the genome of CPAC 15 was obtained. Those numbers also highlight that there is still a lack of knowledge about *Bradyrhizobium*, as there is no experimental evidence for the function of about half of the predicted proteins of the genome. The great majority of the CDSs (96%) were also categorized after Riley's classification (Riley, 1993), and information about each CDS is also available as supplementary information.

Table 1
General features of the genomic panorama of *Bradyrhizobium japonicum* strain CPAC 15

Reads	
Total number of reads	4328
Number of bases deposited (bp)	4,576,417
Number of bases with quality $\geq$ 20 (bp)	2,046,740
Average read length (bp)	1057.4
Average read length (quality $\geq$ 20) (bp)	472.9
Assembly	
Number of phrap contigs	1106
Average contig length (bp)	1071.4
Average number of reads in a contig	2.49
Coverage by phrap contigs (bp)	1,149,042
Coverage by singletons (bp)	35,952
Coverage	
Estimated genome length (bp)	9,000,000
Genome coverage considering the total # of bp deposited	50.85%
Genome coverage by phrap contigs (bp and %)	1,184,994 (13.17%)
%G + C (considering CDSs of COG catergory J)	64.86%
Annotation [number of "coding DNA sequences" (CDSs)]	
Total number	1371
Assigned function	729 (53%)
Conserved hypothetical	312 (23%)
Hypothetical	330 (24%)
Categories after Monica Riley[a]	1318
Known function	673
Conserved hypothetical	315
Hypothetical	330
Classified in COG[b]	797
Known function	603
General function prediction only	130
Unknown	64
Classified in KEGG[c]	1339
Assigned	486
Unassigned	853

[a] After Riley (1993).
[b] Clusters of Orthologous Groups of Proteins (Tatusov et al., 2000).
[c] Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000).

3.2. Functional classification by COG

Fifty-eight percent ($n = 797$) of the CDSs were assigned into 20 COG categories (Table 2), similar to the percentage reported for USDA 110 (Kaneko et al., 2002). The CDS of CPAC 15 fit into all categories (Table 2, and supplementary information) described in the complete genomes of the other rhizobial strains. The only exceptions were unique CDS with extracellular functions (category W) reported in *M. loti* MAFF303099 and in *S. meliloti* 1021, and not detected in CPAC 15. Finally, 16.3% and 8.0% of the CDSs were classified in categories R and S, of predicted general functions and unknown functions, respectively (Table 2). The mean G + C content of CPAC 15 is of 64.86% (Table 1), determined using the 31 CDSs classified in COG category J (ribosomal structure, biogenesis and translation), and slightly lower (64.66%) if only the ribosomal genes are considered, very close to the 64.1% of USDA 110 (Kaneko et al., 2002).

The comparison of all CDSs of CPAC 15 and USDA 110 within each COG category demonstrates a similar distribution (Fig. 1), an additional supporting evidence that a representative draft of the genome of CPAC 15 was obtained.

3.3. Mobile elements and plasticity of the genome

A total of 167 putative transposases were assigned in USDA 110, representing 2% of all CDSs and indicating high plasticity of the

Table 2
Number of coding DNA sequences (CDSs) of *B. japonicum* strain CPAC 15 classified in each COG [Clusters of Orthologous Groups of Proteins (Tatusov et al., 2000)] category

COG functional category		# CDSs
J	Translation	31
A	RNA processing and modification	0
K	Transcription	63
L	Replication, recombination and repair	27
B	Chromatin structure and dynamics	0
D	Cell cycle control, mitosis and meiosis	2
Y	Nuclear structure	0
V	Defense mechanisms	12
T	Transduction mechanisms	30
M	Cell-wall/membrane biogenesis	29
N	Cell motility	5
Z	Cytoskeleton	0
W	Extracellular structures	0
U	Intracellular trafficking and secretion	9
O	Post-translational modification, protein turnover, chaperones	17
C	Energy production and conversion	43
G	Carbohydrate transport and metabolism	57
E	Amino acid transport and metabolism	105
F	Nucleotide transport and metabolism	3
H	Coenzyme transport and metabolism	25
I	Lipid transport and metabolism	47
P	Inorganic ion transport and metabolism	53
Q	Secondary metabolites biosynthesis, transport and catabolism	45
R	General function prediction only	130
S	Function unknown	64
Total		797

genome (Kaneko et al., 2002). Fourteen transposases (1%) were detected in CPAC 15, the great majority related to insertion sequence elements (IS3, IS4, IS66, IS6501, IS911, IS1111A, IS1328, ISR1 families; supplementary information). These elements might be responsible for the experimental evidence of high rates of horizontal gene transfer (HGT) of symbiotic genes from strains belonging to this serogroup to other indigenous or naturalized rhizobia under field conditions in Brazil (Batista et al., 2007; Barcellos et al., 2007).

The distribution of G + C contents of all 1371 CDSs was as follows: 5.47% with G + C content in the class of 20–55%; 17.67% between 56 and 60%; 49.09% in the 61–65%; 26.04% in the 66–70% class and 1.53 in the 71–80% class. The comparison with the G + C content of the CDS classified in the J category of COG (64.66%) thus indicate that the very low or high contents for about one quarter of the CDS (below 60% and above 70% of G + C) is another evidence of HGT events. Some of those CDS had highest similarity with USDA 110 using BLASTP, indicating HGT events prior to the introduction to Brazil.

3.4. KEGG classification

According to the KEGG functional classification, known functions were attributed to 486 CDS, whereas 853 were unassigned (Table 3). The major categories are shown in Table 3, while the detailed distribution of the CDS within each subcategory is available as supplementary information.

When compared with other sequenced genomes, 1071 CDSs of CPAC 15, representing 78% of the total CDS identified, had similarity with completely sequenced microorganisms, 729 (53%) of which with known function (supplementary information). As expected, most of the CDSs (883) had highest similarity with USDA 110, however, they represented only 64.4% of all CDSs identified in this study, decreasing to 45.1% when considering only the putative genes with known function. The other CDS of CPAC 15 had highest similarity with seventy other microorganisms of the KEGG database (Table 4 and supplementary information), further evidence that the bacterium contains a mosaic of genes received by HGT.

After the comparison with the other two sequenced genomes of *Bradyrhizobium*, 24 CDSs had highest similarity by KEGG with strain BTAi1 (supplementary information). Interestingly, five hypothetical proteins might function as exported proteins of still-unknown function (BS00867, BS05486, BS09519, BS09787, and BS09942). High similarity with BTAi1 was also found with the relevant oxidase protein SoxC (BS08303), and with the intriguing antirestriction *ardC* (BS03887), a gene resembling actins, encoding important cellular

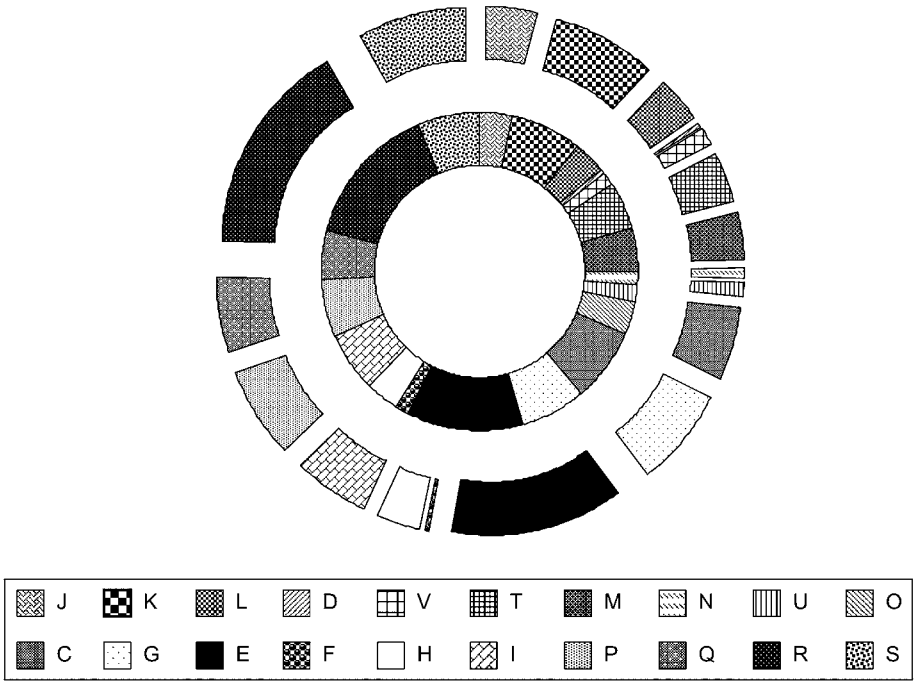

Fig. 1. Comparison of the coding DNA sequences (CDSs) distribution in functional categories of COG [Clusters of Orthologous Groups of Proteins (Tatusov et al., 2000)] of *Bradyrhizobium japonicum* strains USDA 110 (internal circle, 100% of the genome, Kaneko et al., 2002) and CPAC 15 (external circle, 13% of the genome). COG classes as described in Table 2.

Table 3
Number of coding DNA sequences (CDSs) of *B. japonicum* strain CPAC 15 classified in each KEGG [Kyoto Encyclopedia of Genes and Genomes, Kanehisa and Goto (2000)] category

KEGG category	# CDSs
Amino acid metabolism	102
Biosynthesis of polyketides and nonribosomal peptides	1
Biosynthesis of secondary metabolites	16
Carbohydrate metabolism	76
Cell motility	4
Energy metabolism	30
Glycan biosynthesis and metabolism	1
Lipid metabolism	33
Membrane transport	44
Metabolism of cofactors and vitamins	24
Metabolism of complex carbohydrates	11
Metabolism of complex lipids	6
Metabolism of other amino acids	29
Nucleotide metabolism	13
Replication and repair	2
Signal transduction	10
Sorting and degradation	8
Transcription	1
Translation	12
Xenobiotics biodegradation and metabolism	14
Total assigned	486
Unassigned	853

proteins involved in various types of cell motility and ubiquitously expressed in eukaryotic cells (Gonzalez-y-Merchand and Cox, 1988). Furthermore, some of the hypothetical or conserved hypothetical proteins (BS00395, BS00867, BS2617, BS04194, BS09519, BS09549, BS10675) were unique to CPAC 15 and BTAi1. Twenty-one other CDSs had highest KEGG similarity with *Bradyrhizobium* sp. strain ORS278, including a conserved hypothetical protein that might be secreted (BS09243), while others, e.g., BS03141, BS06010, BS06684, BS07516, occurred exclusively in ORS278 and CPAC 15.

In relation to the other sequenced rhizobia, we describe CDS showing high KEGG similarity with epoxide hydrolases of *R. leguminosarum* (Ephx2) (BS05001) and *M. loti* (BS10402), potentially capable of degrading toxic arene and aliphatic epoxides. High similarity was also found with a putative glycosyltransferase involved in cell-wall biogenesis (BS02706), present in several pathogenic prokaryotes such as *Mycobacterium*; in rhizobia the gene has been detected exclusively in *R. etli* and CPAC 15, with high identity also with regulatory proteins of *S. meliloti* (BS04319, BS10828).

We also describe 16 CDSs similar to those of *Nitrobacter hamburgensis* strain X14 that may be related to environmental adaptation, including a cold-shock protein (BS01664) also present in USDA 110, a highly conserved carbohydrate-selective porin (OprB) (BS05388) and the secretion protein HlyD (BS07181).

Table 4
Complete genomes of microorganisms in KEGG [Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000)] database showing highest similarity with the coding DNA sequences (CDSs) of *B. japonicum* strain CPAC 15

Microorganism	Total CDSs			CDSs with known function		
	Number	% in relation to CDSs in KEGG	% of total CDSs	Number of CDSs	% in relation to CDSs in KEGG	% of total CDSs
Bradyrhizobium japonicum USDA 110	883	82.45	64.40	619	84.91	45.14
Bradyrhizobium sp. BTAi1	24	2.24	1.75	11	1.51	0.80
Bradyrhizobium sp. ORS278	21	1.96	1.53	11	1.51	0.80
Nitrosomonas hamburgensis	16	1.49	1.16	5	0.69	0.36
Rhizobium leguminosarum	8	0.75	0.58	7	0.96	0.51
Rhodopseudomonas palustris HaA2	7	0.65	0.51	4	0.55	0.29
R. palustris BisB5	6	0.56	0.43	4	0.55	0.29
R. palustris BisB18	5	0.47	0.36	1	0.14	0.07
Burkholderia pseudomallei 1710b	5	0.47	0.36	0	0.00	0.00
N. winogradskyi	5	0.47	0.36	2	0.27	0.14
61 Other microorganisms in KEGG	91	8.50	8.49	65	8.92	8.00
Other microorganisms not in KEGG			20.07			43.60

3.5. Free-living style and environmental adaptability

Strain CPAC 15 was isolated using soybean as the trap host after a long period of adaptation of the parental strain SEMIA 566 to the stressful conditions (acid soils, aluminum toxicity, high temperatures, drought) of the Brazilian Cerrados, therefore besides showing a higher capacity for N_2 fixation than the parental strains, CPAC 15 is also "adapted" to tropical acid soils (Peres et al., 1993; Hungria and Vargas, 2000; Hungria et al., 2006). Several characteristics in the genome of CPAC 15 are consistent with the saprophytic capability, the adaptation to a wide-range of environmental conditions, and with the nutrition capacity of the strain in low-fertility soils.

Transcriptional regulators or two-component regulatory system proteins are genes with such function. Therefore in versatile microorganisms able to colonize soil and water as well as plant and animal tissues, these genes represent from 5 to 8%, of the genome, decreasing to about 3% or as low as 1% in specialized pathogens (Stover et al., 2000). Eighty-one CDSs (5.9%) of CPAC 15 encode regulatory proteins, 24 of which were related to two-component regulators, including a NodW-like protein (BS01652). ScanProsite search (De Castro et al., 2006) shows that BS01652 contains one HTH-*luxR*-type DNA-binding domain (PS00622 and PS50043) from residue 147 to 212 and one response regulator domain (PS50110) from residue 17 to 131 found in all response regulator members of two-component regulatory systems.

The two-component regulatory system *nodV/nodW*, first described by Göttfert et al. (1990), is localized in the symbiotic island of USDA 110: NodV responds to the isoflavonoid genistein and subsequently phosphorylates its cognate regulator NodW, which in turn may be required to positively regulate the transcription of one or several unknown genes involved in the nodulation of alternative hosts [mung bean (*Vigna radiata*), cowpea (*Vigna unguiculata*), siratro (*Macroptilium atropurpureum*)]. The role of NodW in nodulation has been extended with the report that a defect caused by a mutation in *nwsB* – a response regulator required for full expression of the *nodYABC* operon of *B. japonicum* that shows 65% of identity with *NodW* – could be complemented by over-expression of NodW (Loh et al., 2001; Loh et al., 2002). Recently, *nodW* was also detected in plasmid "c" of *R. etli* (González et al., 2006), and ortholog candidates were proposed in other rhizobial species, all showing BLASTP similarity of 50% or less with *B. japonicum*. The putative *nodW*-like gene of CPAC 15 shows high BLASP similarity (90%) with a general two-component response regulator of USDA 110 (blr0258) (Kaneko et al., 2002), and lower (61%) with the *nodW* described in USDA 110 (Göttfert et al., 1990; Kaneko et al., 2002). Both broader promiscuity and regulation of common *nod* genes may be critical to bacterial survival and maintenance of symbiotic genes even in the absence of host plants, or under adverse environmental conditions, as has been broadly reported for CPAC 15 in Brazilian soils (Hungria et al., 2006). Therefore, this putative *nodW*-like gene of CPAC 15 might represent an additional response regulator that may cross talk with either the NodVW or NwsAB two-component systems.

Forty-three CDSs (5.3%) of CPAC 15 were related to energy production and conversion (COG category C), demonstrating capacity for fermentation, aerobic and anaerobic respiration, thus

indicating a broad capacity to obtain energy under various conditions, from well aerated to flooded soils. Considering all CDSs, at least 62 oxidoreductases and dehydrogenases may help the strain to exploit several energy sources, and in the absence of oxygen, formate (*fdha*, *fdhF*), fumarate or nitrate may be utilized as final electron acceptor, including a putative nitroreductase (BS08886) (present but not assigned in USDA 110). Another important energy-related capability is represented by genes of the *hup/hyp* operon (uptake-hydrogenase system), possibly sufficiently efficient as to allow autotrophic growth of *B. japonicum* using the oxidation of H_2 to provide energy, as described before (Hanus et al., 1979).

Thirty-nine percent of the CDSs of CPAC 15 were related to the transport and metabolism of carbohydrate (COG category G), amino acids (E), lipids (I), inorganic ions (P), and secondary metabolites (Q), therefore directly related to the bacterium's interactions with the environment. Also in other rhizobial genomes, approximately one third of the CDS are within those categories (Tables 2 and 3 and supplementary information). Putative transport genes of CPAC 15 fit into all known transport systems, and many of the CDSs were either absent or have not been assigned in USDA 110.

One transport family poorly described in rhizobia is that of the MFS (major facilitator superfamily), members of which in CPAC 15 include genes for potassium-efflux system proteins *kefA* (component of the MscS mechano-sensitive ion channel) and *kefB*, not assigned in USDA 110, and BS09460, detected in BTAi1 but not USDA 110; *ntrD*, related to the transport of nitrate, showing high similarity (BLASTP, 67%) with *Pseudomonas aeruginosa* was also detected in BTAi1 (63% similarity), but absent in USDA 110. We have also detected in CPAC 15 members of sugar transporters: *exuT* (BS03505), a possible hexuronate transporter not assigned in USDA 110; an alpha-ketoglutarate permease, *kgtP*, similar to marine actinobacterium and not described in rhizobia; and three copies of *shiA*, related to the uptake of shikimate, among others. A putative exported protein of CPAC 15 (BS05941), belonging to the tricarboxylate transporter family, was also interesting as it shows high similarity to USDA 110 and to several *Bordetella*, but not to any other rhizobia.

Four copies of the multidrug resistance *emrB* (=*emrY*), and one of *emrK* were also present in CPAC 15, showing high similarity with USDA 110. The role of these genes in soybean nodulation should be investigated, as in *R. etli* CFN 42 *rmrA* mutant (also an MFS gene) reduced nodulation in common bean (*Phaseolus vulgaris*) by 40%; additionally, *rmrA* and *rmrB* mutants had enhanced sensitivity to phytoalexins, flavonoids, and salicylic acid, but could be complemented by *emrAB* from *E. coli* (González-Pasayo and Martínez-Romero, 2000).

3.6. Other genes related to stress tolerance

Fifty-three CDSs of CPAC 15 fit into the KEGG classification of environmental information and might be important in determining strong saprophytic capacity under the stressful environmental conditions of the Brazilian Cerrados. Adaptation to extreme temperatures may be enabled by five cold/heat-shock proteins: BS01664 (showing high BLASTP similarity with *Nitrobacter hamburgensis*), and four genes showing high similarity with USDA 110, *cspA*, *hspD*, *hslO* and *deaD*; additionally, two different copies of *groEL* are likely to be beneficial under stressful conditions. Protection against oxidative stress may be bestowed by several putative genes encoding glutathione-S-transferases, aldolases, hydrolases, and peroxidases. Also, four copies of *uvrD* are indicative of an effective SOS response mechanism. The Ada-DNA methyltransferase, a regulatory protein of adaptative response (Demple et al., 1985) showing high similarity with *M. loti* and *Brucella*, but not reported in *Bradyrhizobium*, was also detected in CPAC 15. Interestingly, tolerance of acidic soils may be facilitated by the chloride-channel protein ClcA, possibly part of an extreme acid resistance (XAR) response mechanism; the putative gene shows high BLASTP similarity with *Methylococcus capsulatus*, *Yersinia* sp. and *Vibrio* sp., and has never been reported in other rhizobia. Finally, gene duplications have been postulated to assist microbes in adapting to changing environments, and 32 paralogous families were detected in the genome of CPAC 15 (supplementary information).

3.7. Xenobiotic degradation

A large set of genes with xenobiotic potential is important because it reflects both the capacity to survive on a broad-range of carbon sources and potential biotechnological utility for degradation of toxic compounds. In the partial genome of CPAC 15, we identified enzymes participating in sixteen pathways for degradation of xenobiotics. Forty-two CDSs, representing 5.7% of the genome were related to xenobiotics, and the major class included CDS related to the degradation of benzoate via coenzyme-A (17%), and via hydroxylation (10%) (supplementary information).

3.8. Transporters of the RND and RhtB superfamiles, and their probable role in cell density, competitiveness and nodulation

As pointed out by Burse et al. (2004), very few multidrug efflux pumps, classified in the RND (resistance-nodulation-cell division) superfamily, have been described in plant-pathogenic and symbiotic bacteria. Members of this superfamily probably catalyze substrate efflux via an H^+ antiport mechanism and include genes related to the export of lipo-oligosaccharides, involved in nodulation. In CPAC 15, two copies of *acrB* (=*acrE*, =*ragC*, acriflavine resistance protein B), a potent drug efflux protein with broad substrate specificity, were detected. In the plant pathogen *Erwinia amylovora*, AcrBA was induced by plant phytoalexins – including naringenin – contributing to both the colonization of the host and the virulence of the bacterium (Burse et al., 2004). In *B. japonicum*, the genes are co-transcribed with *rpoH*, and were thus named *rag* (*ropH3*-associated-genes; Krummenacher and Narberhaus, 2000). Interestingly, in the organization of *rag* genes of *B. japonicum*, *ragC* corresponds to *nolGHI* of *S. meliloti*. Therefore, although in a unique study (Krummenacher and Narberhaus, 2000) *B. japonicum* mutants in *ragC* apparently were not affected in growth or in the symbiotic properties, new studies on this gene may reveal information about nodulation competitiveness.

CPAC 15 also has *rhtB* and BS02847, both encoding homoserine/homoserine lactone efflux proteins, belonging to the resistance to homoserine/threonine (RhtB) family (Vrljic et al., 1999). These two putative genes in CPAC 15 are present in *Bradyrhizobium*, *Mesorhizobium* and *Agrobacterium*, but have not been reported in other rhizobia, and they may be important in relation to quorum signals associated with *N*-acyl homoserine lactones. A quorum-signal molecule was reported to affect both cell-population density and nodulation in soybean, in a complex mechanism activating some genes and repressing others including *nolA* and *nodD2* (Loh et al., 2001; Loh et al., 2002). Identification of genes related to mechanisms controlling both cell density and expression of *nod* genes are of interest for the inoculant industry.

3.9. Secretion systems

Virulence of bacterial pathogens towards host cells (human, animal, plant) has been extensively characterized. It is related mainly to factors located either on the bacterial surface or secreted into the extracellular space (Hueck, 1998). Only a few pathways are known, actively transporting proteins from the bacterial cytoplasm to the extracellular space, and genome data have revealed that those secretion systems are also present in symbiotic bacteria.

Types II and IV are *sec*-dependent secretion pathways responsible for the transport of proteins from the cytoplasm to the inner and outer membranes (Hueck, 1998). Two Type II secretion proteins were identified in CPAC 15: the protein membrane export SecG and the immunogenic membrane protein YajC, both showing high BLASTP similarity with *Bradyrhizobium*, *Nitrobacter* and *Rhodopseudomonas*, therefore more related to evolution rather than to symbiotic capability. Type IV proteins are usually related to conjugation, mainly on plasmids in *Agrobacterium* (*Rhizobium*) (Chen et al., 2002). Despite not having plasmids, CPAC 15 has a putative *trbG* (=*virB9* family), which participates in the conjugal transfer of the Ti plasmid between *Agrobacterium* and host cells (Alt-Morbe et al., 1996; Chen et al., 2002), and also the conjugal transfer protein TraA/ID523. Another putative Type IV gene is *tapD* (BS05724), a prepilin signal peptidase also present in USDA 110 but not seen in other rhizobia. In *Aeromonas hydrophila*, several proteins associated with virulence are secreted, and *tapD* was found acting as a homolog of *pilD* of *Pseudomonas aeruginosa*, responsible for the cleavage of a Type IV fimbrial leader sequence and the methylation of the N-terminal residual, as well as for the secretion of proteases and of the toxin aerolysin (Pepe et al., 1996).

Types I and III secretion systems are *sec*-independent pathways and, therefore, do not involve amino-terminal processing of the secreted proteins; in addition, secretion occurs in a continuous process without the presence of periplasmic intermediates (Hueck, 1998). Type I secretion proteins identified in CPAC 15 are HylD, also present in USDA 110, *M. loti* and *R. etli*, and two other HylD-family proteins, PrtE, involved in the secretion of proteases, and PrtD, an ATPase, both showing high similarity with *Rhodopseudomonas palustris*.

Type III secretion systems (T3SS) include proteins coding for flagellae of eubacteria and, in some Gram-negative bacteria, for injectisomes, responsible for delivering proteins across the eukaryotic cell membrane – the effectors – which can reprogram the cells to the benefit of the bacterium (Journet et al., 2005). The flagella-related genes detected in CPAC 15 were *fliF* (M-ring protein), *fliR*, *flhB* (flagellum biosynthesis), and two probable polar flagellar motor-switch proteins, MotB and BS02666, all with high BLASTP with USDA 110.

We have also found indications of the presence of effector Hrp (*h*ypersensitive response and pathogenicity)-dependent proteins, BS00574 and BS07765, this last one showing high similarity with *Acidiphilium cryptum* and followed by a cellulose. BS06844, an AraC-family transcriptional regulator, could also be related to T3SS, as described for the specific binding role of HrpX in the plant pathogen *Xanthomonas campestris* pv. *vesicatoria* (Koebnik et al., 2006).

Another key T3SS CDS (BS00587) of CPAC 15 might be involved in bacterial–plant symbiosis. The CDS shows high similarity with USDA 110 (bll1858), and with CDS ID322 of the strain 110spc4, both encoded within a potential symbiotic island (Göttfert et al., 2001). Interestingly, the protein sequence of BS00587 shares homology with the C-terminal of the nodulation outer protein P (NopP) of *Rhizobium* sp. NGR234, which in turn exhibits homology in other rhizobia and bradyrhizobia with either parts of its sequence or with the entire protein sequence (Fig. 2). NopP of NGR234 has been characterized as a flavonoid-induced extracellular protein secreted by a T3SS (Ausmees et al., 2004), which is capable of phosphorylation at the N-terminal by a non-identified host kinase and has various effects on the nodulation of the legumes, ranging from highly beneficial (in *Flemingia congesta* and *Tephrosia vogelii*) (Skorpil et al., 2005), to very weak (*Pachyrhizus tuberosus*) and deleterious (*Vigna unguiculata*) (Ausmees et al., 2004). The exact role of the secreted NopP into the host cell remains to be investigated; however, once phosphorylated into plant cells it might perturb the expression of genes that are activated by kinase pathways, such as plant defense genes. Regarding the NopP homolog found in the genome of strain CPAC 15 we suppose that it might function as an effector protein involved in the host specificity with legumes. Although the protein sequence of BS00587 lacks the N-terminal region of NopP of NGR234 that contains both the signal for T3SS-dependent secretion and sites for phosphorylation, it has the consensus sequence Asn-Ala-Gly-Asp-Ihe-Leu-Leu-Glu (Fig. 2) found in most NopP homologues (Ausmees et al., 2004). Therefore, this founding supports a first report that the genome of strain CPAC 15 might encode a protein of the NopP family.

3.10. Genes related to nodulation, N_2 fixation and N_2 fixation efficiency

The nodulation genes detected in the genome of CPAC 15 show high similarity with USDA 110 (Table 5). However, we have assigned a putative new *nodB*-like (BS08735) gene that in USDA 110 is described as a hypothetical gene. The CDS was identified as belonging to a family of polysaccharide deacetylases, which includes NodB, a chitooligosaccharide deacetylase. The CDS shows similarity with the NodB proteins described in other *Bradyrhizobium*, *Rhizobium* and *Mesorhizobium* species, and, it is interesting to note, it shows 68% identity with BTAi1, although it was not reported in the genome of this strain (Giraud et al., 2007).

Although with lower e-values in relation to the other *nod* genes, a *nodD2*-like putative gene was also assigned in the genome of CPAC 15, based on comparison with the UniProt/Swiss Prot database (UniProt, 2007). NodD binds flavonoids and regulates the expression of the *nodABCFE* genes, which encode other nodulation proteins. In *B. japonicum* there are two NodD proteins (Göttfert et al., 1992): NodD1 is a positive transcriptional activator of *nod* genes, whereas NodD2 represses *nod* gene expression (Loh et al., 2002). Therefore, it remains to be determined if CPAC 15 has another NodD2 showing high similarity with USDA 110, or if BS11187 represents a second copy of NodD2 of unknown function.

NodW and NopP have already been discussed, and another interesting CDS was BS05498, showing similarity to NdvA of *S. meliloti* and with ChvA of *Agrobacterium* (=*Rhizobium*) *tumefasciens*; the gene also shares homology with HlyB of *E. coli*, a protein involved in the export of hemolysin (Stanfield et al., 1988). It has been demonstrated that the *ndvA* of *S. meliloti* encodes a beta-1,2-glucan export ATP-binding protein required for normal nodulation of alfalfa roots, and mutant strains lacking this gene form only "empty" nodules (Dickstein et al., 1988; Stanfield et al., 1988). The protein might also be involved in exporting other classes of oligosaccharides. NdvA has also been reported in the genomes of *M. loti* and *R. etli*, and, although BS05498 shows high similarity with USDA 110, the gene was not assigned in any of the genomes of *Bradyrhizobium*.

CDS for *fixO*, *fixP* and *fixQ*, all encoding cytochrome oxidases, as well as the regulatory protein FixR, were also detected in CPAC 15 (Table 5). In relation to the *nif* genes, NifE, related to the biosynthesis of the nitrogenase iron-molybdenum cofactor, shows high similarity with bradyrhizobia, but not with rhizobia. As it occurs with USDA 110 (Göttfert et al., 2001; Kaneko et al., 2002), all identified *fix* and *nif* genes should be located in the symbiotic island.

Two proteins that are key to the metabolism of N, NtrX and NtrY, are members of the two-component regulatory system *ntrY/ntrX*, involved in the activation of nitrogen-assimilatory genes such as *glnA*; NtrX is probably phosphorylated by *ntrY* and interacts with sigma-54. In CPAC 15, the two putative genes show high BLASTP similarity with *Bradyrhizobium*, *Rhodopseudomonas* and *Nitrobacter* (Table 5).

The uptake-hydrogenase system can contribute to the efficiency of the N_2 fixation, recycling part of the energy (ATP) spent in the

```
MUSCLE (3.6) multiple sequence alignment

BS00587           ------------------------------------------------------------
ID322(110spc4)    ------------------------------------------------------------
ID185(110spc4)    ------------------------------------------------------------
HIA(RHIFR)        ---------------------------------------MHLDRSDSNGGSS---RYTL
Y4yP(RHIFR)       ------MYGRIDSSSDFHYTQSASKQTDAETQEFADTFARMHLDR--SNGGSSSAARYTL
NOPP(RHIFR)       ------MYGRIDSSSDFHYTQSASKQTDAETQEFADTFARMHLDRSDSNGGSS---RYTL
NOPP(NGR234)      ------MYGRIDSSSDFHYTQSASKQMDAETQEFADTFARMHLDR--SNGGSSSAARYTL
ORF3(RHILE)       MDNTERSLRYNISEEDDRTRAMAGWIENMTGAQTLTGMPSRSTSDPHARSPSSPTRPYSL
Blr1752(USDA110)  -------------------------------------------MEPGGSSSETRAYYL
ID84(110spc4)     ------MYGRVVGSSSQSTSANQVDEPGDSPRFA------ETLAGMEPGGSSSETRAYYL

BS00587           ------------------------------------------------------------
ID322(110spc4)    ------------------------------------------------------------
ID185(110spc4)    -------------MSEL--FKG-------------------------------------
HIA(RHIFR)        DHEPPVVPIDLKTFRREIRKFHGKEITDIADNPQEYSDFVSAKARRTADVAQQYGIRR--
Y4yP(RHIFR)       DHEPPVVPIDLKTFRREIRKFHGKEITDIADNPQEYSDFVSAKARRTADVAQQYGIRR--
NOPP(RHIFR)       DHEPPVVPIDLKRFRREIRKFRGKEITDIADNPQEYSDFVSAKARRTADVAQQYGIRR--
NOPP(NGR234)      DHEPPVVPIDLETFRREIRKFHGKEITDIANNPQEYSDFVSAKARRTADVAQQYGIRR--
ORF3(RHILE)       SSEPPIKVLDRSTFARKVREFYGDDIKHISVNPQEYSPFVSSKALRTATIARLYGGTD-N
Blr1752(USDA110)  NSGPPIVEIDQLSFDRGLQRFLGRDIMNIAINPQEYSDFVSKKAERAATVAGSYSATHYD
ID84(110spc4)     NSGPPIVEIDQLSFDRGLQRFLGRDIMNIAINPQEYSDFVSKKAERAATVAGSYSATHYD

BS00587           --------------------------MATGFTSQRWRELFPGRTDVTSIVDFQVAHPLV
ID322(110spc4)    --------------------------MATGFTSQRWRELFPGRTDVTSIVDFQVAHPLV
ID185(110spc4)    -----------------------------NNDWRHQFPGRDEITSTIAFRVTHPLV
HIA(RHIFR)        DSENARYFSYQLGNQCVGLMRTEGGFSMEEEFESKSWRDQFPGHQEITSTVDLQVAHPLV
Y4yP(RHIFR)       DSENARYFSYQLGNQCVGLMRTEGGFSMEEEFESKSWRDQFPGHQEITSTVDLKSPSSL-
NOPP(RHIFR)       DSENARYFSYQLGNQCVGLMRTEGGFSMEEEFESKSWRDQFPGHQEITSTVDLQVAHPLV
NOPP(NGR234)      DSENARYFSYQLGNQCVGLMRTEGGFSMEEEFESKSWRDQFPGHQEITSTVDLQVAHPLV
ORF3(RHILE)       DTPSARYFSYDLGERSVGLLRTEDGFSIE----NEPWRELFPGRKRITSIVDLRVTHPLV
Blr1752(USDA110)  PARPVRFFSYQLGDETVGLLRAGGPVRIK----GETFREKF-GRNDLTSVVDLRVTHPLV
ID84(110spc4)     PARPVRFFSYQLGDETVGLLRAGGPVRIK----GETFREKF-GRNDLTSVVDLRVTHPLV
                                                .: :*  * *.  :** : :. . .*

BS00587           GNAGDILLEHQLGRDGERPLVNWRPANDEARARAEQLGFVHVDPDDMVLDPTQ-SKQW-R
ID322(110spc4)    GNAGDILLEHQLGRDGERPLVNWRPANDEARARAEQLGFVHVDPDDMVLDPTQ-SKQW-R
ID185(110spc4)    ENAGDILLEHQLQIDGNKPLVLSHSSNPEAKARAAALGFLEVSDSMMVLDPTKHDDKW-K
HIA(RHIFR)        ENAGDILLEAPTSEGRRTTVAELARGKPRGESRAAMMGFVEVDDCDMVLDPKQHPDKWTQ
Y4yP(RHIFR)       ENAGDILLEHQLRRDGERPLLNWRAENPEAKARAAMMGFVEVDDCDMVLDPKQHPDKWTQ
NOPP(RHIFR)       ENAGDILLEYQLRRDGERPLLNWRAENPEAKARAAMMGFVEVDDCDMVLDPKQHPDKWTQ
NOPP(NGR234)      ENAGDILLEHQLRRDGERPLLNWRAENPEAKARAAMMGFVEVDDCDMVLDPKQHPDKWTQ
ORF3(RHILE)       ENAGDILLEHPTSTRRRTAVNHVTSRSPEMGARLEQMGFVDVGQNRWVLDPTQHPDKWTK
Blr1752(USDA110)  ENAGDILLEYQLREDGDDPLILSKPGLPGMEPRLAEMGFVHVGRNHWVLDPHQHPEVWTK
ID84(110spc4)     ENAGDILLEYQLREDGDDPLILSKPGLPGMEPRLAEMGFVHVGRNHWVLDPHQHPEVWTK
                   ********         .:           .*   :**: *.    **** :  . * .

BS00587           YRDGEWQRATNSSMYLSKA-----SS------DDEPSSE-P-----DSDGDFM-------
ID322(110spc4)    YRDGEWQRATNSSMYLSKA-----SS------DDEPSSE-P-----DSDGDFM-------
ID185(110spc4)    NEGGNWQRAGKPPLYLAAT-----EDNSGNAKYTESVAK-PANDPYDEEYDFM-------
HIA(RHIFR)        TSAAEWRRKDKPPLYLRKF-----ED-AETAQCSTKAAL-TRLTKMTSCDRILARFGRDG
Y4yP(RHIFR)       TSAAEWRRKDKPPLYLRKF-----ED-AETAQCSTSCSY-E-----TYEDDFM-------
NOPP(RHIFR)       TSAAEWRRKDKPPLYLCKF-----ED-AETAQCSTSCSY-E-----TYEDDFM-------
NOPP(NGR234)      TSAAEWRRKDKPPLYLRKF-----ED-AETAQCSTSCSY-E-----TYEDDFM-------
ORF3(RHILE)       NSKGGWQRAGKPWSYLSKTVDYLSDDEAGVEGHSSEASV-ETDSSDDDPSWYFEHLNLNR
Blr1752(USDA110)  NENDQWQRVGKPTKYLSKV-----EDDDAAAESTVQADYSDEDDPSVYLERVLTGLSME-
ID84(110spc4)     NENDQWQRVGKPTKYLSKV-----EDDDAAAESTVQADYSDEDDPSVYLERVLTGLSME-
                       *.*  :.  **        ..                          :

BS00587           ---------------------------------------------
ID322(110spc4)    ---------------------------------------------
ID185(110spc4)    ---------------------------------------------
HIA(RHIFR)        RAPSAKTGPHVMDRERKSVTNCHALTAIPIRCLGAKELERRVSLRPS
Y4yP(RHIFR)       ---------------------------------------------
NOPP(RHIFR)       ---------------------------------------------
NOPP(NGR234)      ---------------------------------------------
ORF3(RHILE)       E--------------------------------------------
Blr1752(USDA110)  ---------------------------------------------
ID84(110spc4)     ---------------------------------------------
```

Fig. 2. Muscle (3.6) (Edgar, 2004) alignment of NopP homolog (ORF BS00587). NopP of *Rhizobium* sp. strain NGR234, ID322, ID185 and ID84 of *Bradyrhizobium japonicum* 110spc4, HIA of *Sinorhizobium fredii* USDA201, NopP of *S. fredii* HH103, Y4yP of *S. fredii* USDA257, ORF3 of *R. leguminosarum* bv phaseoli and Blr1752 of *B. japonicum* USDA 110 were aligned by using Muscle version 3.6. Amino acids common to all NopP homologues are indicated by * and the consensus sequence Asn-Ala-Gly-Asp-Ihe-Leu-Leu-Glu is shown.

Table 5
Coding DNA sequences (CDSs) related to *nod*, *nif*, *fix* and *ntr* genes detected in the genome of *B. japonicum* strain CPAC 15

CDS	Gene	Gene product	Query coverage (%)	Subject coverage (%)	Identity (%)[a]	e-Value	Size (aa)
BS11185	*fixO*	Protein FixO/cb-type cytochrome *c* oxidase subunit II	100	11	96	$6e^{-07}$	28[b]
BS11156	*fixP*	Protein FixP/cb-type cytochrome *c* oxidase subunit III/cbb3 oxidase, subunit III	66	18	91	$2e^{-40}$	134
BS11179	*fixQ*	Protein FixQ, cytochrome oxidase subunit, small membrane protein/cbb3 oxidase, subunit IV	100	100	94	$2e^{-21}$	165
BS01532	*fixR*	Protein FixR	100	56	98	$1e^{-82}$	159
BS03636	*nifE*	Nitrogenase iron-molybdenum cofactor biosynthesis protein NifE	100	61	99	0.0	338
BS08735	*nodB*	NodB-like protein/chitooligosaccharide deacetylase	95	100	100[c]	$4e^{-38}$	82
BS11187	*nodD2*	NodD2-like protein, transcriptional regulator	93	76	28[c]	$7e^{-13}$	241
BS01652	*nodW*	NodW-like protein, two-component regulatory system	100	100	61[c]	$3e^{-65}$	224
BS00587	*nopP*	Effector protein NopP	100	100	100[c]	$7e^{-71}$	128
BS05498	*ndvA*	Beta-1,2-glucan export ATP-binding protein NdvA	93	32	95[c]	e^{-100}	206
BS01726	*ntrX*	Nitrogen assimilation regulatory protein NtrX	100	17	98	$9e^{-38}$	81
BS05510	*ntrY*	Nitrogen regulation protein NtrY/two-component sensor histidine kinase	89	40	97	e^{-180}	381

[a] Identity considering the genome of *B. japonicum* strain USDA 110.
[b] Despite the size of the gene is lower than what was established in the protocol of annotation, the gene was annotated due to the high similarity with *fixO* in several bacteria.
[c] CDS present but not assigned with this gene in the genome of USDA 110.

process through the oxidation of hydrogen (H_2) produced by the nitrogenase (Brito et al., 2005). For soybean, it has long been known that Hup^+ mutants may result in higher plant-N contents and grain yields (Albrecht et al., 1979; Hanus et al., 1981; Hungria et al., 1989). At least 17 common hydrogenase genes (*hupSLCDFGHIJKhypABFCDEX*) are organized in three operons, and in CPAC 15 we identified four of them: *hupB* (=*hupL*), *hupF*, *hypB* and *hypF* (=*hupY*). Interestingly, as also reported in USDA 110 (Kaneko et al., 2002), CPAC 15 has two copies of *hupB* (=*hupL*), possibly conferring advantage. The two copies of CPAC 15 share approximately 75% of BLASTP identity.

3.11. Other generalities

Some putative genes of CPAC 15 might show biotechnological applicability in areas other than agriculture. For example, BS01192, encoding an *N*-carbamoyl-D-amino acid hydrolase, and BS09595, encoding a polyketide synthase, have potential utility in the antibiotics industry.

Other CDS might help to explain the strong competitiveness of CPAC 15 in the soybean rhizosphere, e.g., two copies of *mocR* (present but not assigned in USDA 110), a gene related to the catabolism of rhizopines, inositol derivates synthesized in legumes in response to rhizobia (Murphy et al., 1995). Another gene is *ooxA*, an opine oxidase, as opine is a compound released by crown-gall tumors produced by *Agrobacterium* (Montoya et al., 1977).

Also, there are genes that may facilitate invasion of the host and the nodulation process, such as *sleB* and BS10711 (cell-wall hydrolases), *iaaH* and BS09332 (related to synthesis of plant-growth promoters such as indole-3-acetic acid), BS07441 and BS04791 (invasion), and BS02453 (virulence). Possible new roles in symbiosis may be revealed by mutations in genes such as BS01394, related to the flavonoid synthesis in eukaryotes and detected in rhizobia.

Finally, it is worthwhile to mention many conserved hypothetical CDSs (e.g., BS01269, BS06276, BS10925, BS00607, BS00601, BS09949, BS04562) highly similar in several rhizobial species, the functions of which remain to be determined.

3.12. Final remarks

Serogroup SEMIA 566 has important ecological implications in Brazilian soils, not only because it predominates in practically all surveys performed so far (in 2008 there are approximately 22 million cropped hectares with the legume), but also because the serogroup was largely dispersed and it is now established even in soils that have never been inoculated, including the Amazon forest (Freire et al., 1983; Vargas et al., 1993; Ferreira and Hungria, 2002; Mendes et al., 2004; Hungria et al., 2006). Therefore, the serogroup may be widespread in other South American countries. Additionally, the competitiveness of strains belonging to the SEMIA 566/CPAC 15 serogroup (the same serogroup as USDA 123) is also a major concern in other countries, as the USA (Ham et al., 1971; Weber et al., 1989), Canada (Semu and Hume, 1979) and Korea (Kang et al., 1991), as it is very hard to introduce new and more effective strains in soils with a dominance of this serogroup.

The genomes of many important bacteria remain to be sequenced, such as that of strain CPAC 15 and of others of economical and ecological relevance in tropical regions. However, financial resources in developing countries are limited, and although new technologies such as pyrosequencing have been developed, completing a genome is still costly. In addition, as pointed out by Simpson (2001), the finishing work may consume 50–60% of the time and costs of sequencing a complete genome. In contrast, in our study we read ≈4% of the sequences deposited for USDA 110 at an estimated cost of ≈2–4%. However, comparisons of CPAC 15 with the complete genome of USDA 110 in several databases and features indicated successful coverage of the whole genome of CPAC 15. With about 13% coverage of the genome (final assemblage with phrap contigs; or 51% considering the number of bp deposited), we have identified approximately 16% of the estimated number of genes, including several new putative genes and some showing very low similarity with strain USDA 110. Indeed, at least one third of the genome of CPAC 15 shows higher similarity to microorganisms other than strain USDA 110. This finding also validates the importance of sequencing many strains within the same species, the "pan-genome" approach (Medini et al., 2005; Tettelin et al., 2005).

The information provided by the panoramic view of the genome of CPAC 15 strongly indicates potential utility in obtaining at least partial genome sequences of other bacteria of agronomic importance in the tropics while broadening understanding of the ecology of CPAC 15 in South American soils.

Acknowledgements

Reagents were financed by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, project # 505499/2004-5); the study is also supported by the Millennium Institute project. The authors thank to Pâmela Menna, Daisy Rickle Binde and Jesiane S. S. Batista (fellows from CNPq # 505946/2004-1) for the laboratory assistance in the sequencing project and Dr. Fábio O. Pedrosa and Dr.

Emanuel M. Souza for scientific advice. L.P. Godoy is an M.Sc. student with a fellowship from Embrapa.

Supplementary material

Supplementary material relating to this work, as cited in the text is available at: http://www.bnf.lncc.br/bj/final/main.html (or http://www.bnf.lncc.br, then *Bradyrhizobium japonicum* SEMIA 5079, genome draft).

References

Albrecht, S.L., Maier, R.J., Hanus, F.J., Russell, S.A., Elmerich, D.W., Evans, H.J., 1979. Hydrogenase of *Rhizobium japonicum* increases nitrogen fixation by nodulated soybeans. Science 203, 1255–1257.

Almeida, L.G., Paixão, R., Souza, R.C., Costa, G.C., Barrientos, J.A., Santos, M.T., Almeida, D.F., Vasconcelos, A.T.R., 2004. A system for automated bacterial (genome) integrated annotation – SABIA. Bioinformatics 20, 2832–2833.

Alt-Morbe, J., Stryker, J.L., Fuqua, C., Li, P.L., Farrand, S.K., Winans, S.C., 1996. The conjugal transfer system of *Agrobacterium tumefaciens* octopine-type Ti plasmids is closely related to the transfer system of an IncP plasmid and distantly related to Ti plasmid vir genes. Journal of Bacteriology 178, 4248–4257.

Apweiler, R., Attwood, T.K., Bairoch, A., 2000. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16, 1145–1150.

Ausmees, N., Kobayashi, H., Deakin, W.J., Marie, C., Krishnan, H.B., Broughton, W.J., Perret, X., 2004. Characterization of NopP, a type III secreted effector of *Rhizobium* sp. strain NGR234. Journal of Bacteriology 186, 4774–4780.

Barcellos, F.G., Menna, P., Batista, J.S.S., Hungria, M., 2007. Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous *Sinorhizobium (Ensifer) fredii* and *Bradyrhizobium elkanii* in a Brazilian savannah soil. Applied and Environmental Microbiology 73, 2635–2643.

Batista, J.S.S., Hungria, M., Barcellos, F.G., Ferreira, M.C., Mendes, I.C., 2007. Variability in *Bradyrhizobium japonicum* and *B. elkanii* seven years after introduction of both the exotic microsymbiont and the soybean host in a Cerrados soil. Microbial Ecology 53, 270–284.

Brito, B., Baginsky, C., Palacios, J.M., Cabrera, E., Ruiz-Argüeso, T., Imperial, J., 2005. Biodiversity of uptake hydrogenase systems from legume endosymbiotic bacteria. Biochemical Society Transactions 33, 33–35.

Burse, A., Weingar,t, H., Ullrich, M.S., 2004. The phytoalexin-inducible multdrug efflux pump AcrAB in the fire blight pathogen, *Erwinia amylovora*. Molecular Plant–Microbe Interactions 17, 43–54.

Chain, P.S.G., Denef, V.J., Konstantinidis, K., Vergez, L.M., Agullo, L., Hauser, L., Cordova, M., Gomez, L., Gonzalez, M., Land, M., Lao, V., Larimer, F., Lipuma, J.J., Mahenthiralingam, E., Malfatti, S.A., Marx, C.J., Jacob, P.J., Ramette, A., Richardson, P., Seeger, M., Smith, D., Spilker, T., Sul, W.J., Tsoi, T., Ulrich, L.E., Zhulin, I.B., Tiedje, J.M., 2006. *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.7 M bp genome shaped of versatility. Proceedings of the National Academy of Sciences of the United States of America 103, 15280–15287.

Chen, L., Chen, Y., Wood, D.W., Nester, E.W., 2002. A new type IV secretion system promotes conjugal transfer in *Agrobacterium tumefaciens*. Journal of Bacteriology 184, 4838–4845.

Chothia, C., 1992. One thousand families for the molecular biologist. Nature 357, 543–544.

De Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N., 2006. ScanProsite: detection of PROSITE signature matches and ProRel-associated functional and structural residues in proteins. Nucleic Acids Research 34, W362–W365.

Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L., 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Research 27, 4636–4641.

Demple, B., Sedgwick, B., Robins, P., Totty, N., Waterfield, M.D., Lindahl, T., 1985. Active site and complete sequence of the suicidal methyltransferase that counters alkylation mutagenesis. Proceedings of the National Academy of Sciences of the United States of America 82, 2688–2692.

Dickstein, R., Bisseling, T., Reinhold, V.N., Ausubel, F.M., 1988. Expression of nodule-specific genes in alfalfa root nodules blocked at an early stage of development. Genes and Development 2, 677–687.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1782–1797.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8, 186–194.

Ferreira, M.C., Hungria, M., 2002. Recovery of soybean inoculant strains from uncropped soils in Brazil. Field Crops Research 79, 139–152.

Freire, J.R.J., Kolling, J., Vidor, C., Pereira, J.S., Kolling, I.G., Mendes, N.G., 1983. Sobrevivência e competição por sítios de nodulação de estirpes de *Rhizobium japonicum* na cultura da soja. Revista Brasileira de Ciências do Solo 7, 47–53.

Galibert, F., Finan, T.M., Long, S.R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R.W., Dreano, S., Federspiel, N.A., Fisher, R.F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernandez-Lucas, I., Hong, A., Huizar, L., Hyman, R.W., Jones, T., Kahn, D., Kahn, M.L., Kalman, S., Keating, D.H., Kiss, E., Komp, C., Lelaure, V., Masuy, D., Palm, C., Peck, M.C., Pohl, T.M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thebault, P., Vandenbol, M., Vorholter, F.J., Weidner, S., Wells, D.H., Wong, K., Yeh, K.C., Batut, J., 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science 293, 668–672.

Giraud, E., Moulin, L., Vallenet, D., Barbe, V., Cytryn, E., Avarre, J.C., Jaubert, M., Simon, D., Cartieaux, F., Prin, Y., Bena, G., Hannibal, L., Fardoux, J., Kojadinovic, M., Vuillet, L., Lajus, A., Cuveiller, S., Rouy, Z., Mangenot, S., Segurens, B., Dossat, C., Franck, W.L., Chang, W.S., Saunders, E., Bruce, D., Richardson, P., Normand, P., Dreyfus, B., Pignol, D., Stacey, G., Emerich, D., Verméglio, A., Médigue, C., Sadowsky, M., 2007. Legumes symbioses: absence of *nod* genes in photosynthetic bradyrhizobia. Science 316, 1307–1312.

GOLD(™) (Genomes OnLine Database), 15 June 2008. Genomes online database v. 2.0. www.genomesonline.org/.

Gonzalez-y-Merchand, J.A., Cox, R.A., 1988. Structure and expression of an actin gene of *Physarum polycephalum*. Journal of Molecular Biology 202, 161–168.

González, V., Santamaria, R.I., Bustos, P., Hernández-González, I., Medrano-Soto, A., Moreno-Hagelsieb, G., Janga, S.C., Ramírez, M.A., Jiménez-Jacinto, V., Collado-Vides, J., Dávila, G., 2006. The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. Proceedings of the National Academy of Sciences of the United States of America 103, 3834–3839.

González-Pasayo, R., Martínez-Romero, E., 2000. Multiresistance genes of *Rhizobium etli* CFN42. Molecular Plant–Microbe Interactions 13, 572–577.

Gordon, D., Abajian, C., Green, P., 1998. *Consed*: a graphical tool for sequence finishing. Genome Research 8, 195–202.

Göttfert, M., Grob, P., Hennecke, H., 1990. Proposed regulatory pathway encoded by the *nodV* and *nodW* genes, determinants of host specificity in *Bradyrhizobium japonicum*. Proceedings of the National Academy of Sciences of the United States of America 87, 2680–2684.

Göttfert, M., Holzhäuser, D., Bäni, D., Hennecke, H., 1992. Structural and functional analysis of two different *nodD* genes in *Bradyrhizobium japonicum* USDA110. Molecular Plant–Microbe Interactions 15, 257–265.

Göttfert, M., Roethlisberger, S., Kuendig, C., Beck, C., Marty, R., Hennecke, H., 2001. Potential symbiosis-specific genes uncovered by sequencing a 410-kb DNA region of the *Bradyrhizobium japonicum* chromosome. Journal of Bacteriology 183, 1405–1412.

Ham, G.E., Caldwell, V.B., Johnson, H.W., 1971. Evaluation of *Rhizobium japonicum* inoculants in soil containing naturalized populations of rhizobia. Agronomy Journal 63, 301–303.

Hanus, F.J., Albrecht, S.L., Zablotowicz, R.M., Elmerich, D.W., Russell, S.A., Evans, H.J., 1981. Yield and N content of soybean seed as influenced by *Rhizobium japonicum* inoculants possessing the hydrogenase characteristic. Agronomy Journal 73, 368–371.

Hanus, F.J., Maier, R.J., Evans, H.J., 1979. Autotrophic growth of H_2-uptake positive strains of *Rhizobium japonicum* in an atmosphere supplied with hydrogen gas. Proceedings of the National Academy of Sciences of the United States of America 76, 1788–1792.

Hueck, C.J., 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. Molecular Microbiology 62, 379–433.

Hungria, M., Campo, R.J., Mendes, I.C., Graham, P.H., 2006. Contribution of biological nitrogen fixation to the N nutrition of grain crops in the tropics: the success of soybean (*Glycine max* L. Merr.) in South America. In: Singh, R.P., Shankar, N., Jaiwal, P.K. (Eds.), Nitrogen Nutrition and Sustainable Plant Productivity. Studium Press, Houston, Texas, pp. 43–93.

Hungria, M., Neves, M.C.P., Döbereiner, J., 1989. Relative efficiency, ureide transport and harvest index in soybeans inoculated with isogenic HUP mutants of *Bradyrhizobium japonicum*. Biology and Fertility of Soils 7, 325–329.

Hungria, M., Vargas, M.A.T., 2000. Environmental factors affecting N_2 fixation in grain legumes in the tropics, with an emphasis on Brazil. Field Crops Research 65, 151–154.

Journet, L., Hughes, K.T., Cornelis, G.R., 2005. Type III secretion: a secretory pathway serving both motility and virulence (review). Molecular Membrane Biology 522, 41–50.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 29–34.

Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y., Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M., Tabata, S., 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Research 7, 331–338.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., Watanabe, A., Idesawa, K., Iriguchi, M., Kawashima, K., Kohara, M., Matsumoto, M., Shimpo, S., Tsuruoka, H., Wada, T., Yamada, M., Tabata, S., 2002. Complete genome sequence of the nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA 110. DNA Research 9, 189–197.

Kang, U.G., Somasegaran, P., Hoben, H.J., Bohlool, B.B., 1991. Symbiotic potential, competitiveness, and serological properties of *Bradyrhizobium japonicum* indigenous to Korean soils. Applied and Environmental Microbiology 57, 1038–1045.

Koebnik, R., Krüger, A., Thieme, F., Urban, A., Bonas, U., 2006. Specific binding of the *Xanthomonas campestris* pv. vesicatoria AraC-Type transcriptional activator HrpX to plant-inducible promoter boxes. Journal of Bacteriology 188, 7652–7660.

Krummenacher, P., Narberhaus, F., 2000. Two genes encoding a putative multidrug efflux pump of the RND/MFP family are cotranscribed with an *rpoH* gene in *Bradyrhizobium japonicum*. Gene 241, 247–254.

Kunin, V., Cases, I., Enright, A.J., Lorenzo, V., Ouzouni, C.A., 2003. Myriads of proteins families, and still counting. Genome Biology 4, Article 401.

Lee, K.-B., De Backer, P., Aono, T., Liu, C.-T., Suzuki, S., Suzuki, T., Kaneko, T., Yamada, M., Tabata, S., Kupfer, D.M., Najar, F.Z., Wiley, G.B., Roe, B., Binnewies, T.T., Ussery, D.W., D'Haeze, W., Den Herder, J., Gevers, D., Vereecke, D., Holsters, M., Oyaizu, H., 2008. The genome of the versatile nitrogen fixer *Azorhizobium caulinodans* ORS571. BMC Genomics 9, 271.

Loh, J., Lohar, D.P., Andersen, B., Stacey, G., 2002. A two-component regulator mediates population-density-dependent expression of the *Bradyrhizobium japonicum* nodulation genes. Journal of Bacteriology 184, 1759–1766.

Loh, J.T., Yuen-Tsa, J.P.Y., Stacey, M.G., Lohar, D., Welborn, A., Stacey, G., 2001. Population density-dependent regulation of of the *Bradyrhizobium japonicum* nodulation genes. Molecular Microbiology 42, 37–46.

Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R., 2005. The microbial pan-genome. Current Opinion in Genetics and Development 15, 589–594.

Mendes, I.C., Vargas, M.A.T., Hungria, M., 2004. Establishment of *Bradyrhizobium japonicum* and *B. elkanii* in a Brazilian Cerrados oxisol. Biology and Fertility of Soils 40, 28–35.

Montoya, A.L., Chilton, M.D., Gordon, M.P., Sciaky, D., Nester, E.W., 1977. Octopine and nopaline metabolism in *Agrobacterium tumefaciens* and crown gall tumor cells: role of plasmid genes. Journal of Bacteriology 129, 101–107.

Murphy, P.J., Wexler, W., Grzemski, W., Rao, J.P., Gordon, D., 1995. Rhizopines – their role in symbiosis and competition. Soil Biology and Biochemistry 27, 525–529.

Nakai, K., Kanehisa, M., 1991. Expert system for predicting protein localization sites in Gram-negative bacteria. Proteins – Structure Function and Genetics 11, 95–110.

Newton, W.E., 2000. Nitrogen fixation in perspective. In: Pedrosa, F.O., Hungria, M., Yates, M.G., Newton, W.E. (Eds.), Nitrogen Fixation: From Molecules to Crop Productivity. Kluwer, Dordrecht, pp. 3–8.

Ochman, H., 2005. Genomes on the shrink. Proceedings of the National Academy of Sciences of the United States of America 102, 11959–11960.

Pepe, C.M., Eklund, M.W., Strom, M.S., 1996. Cloning of an *Aeromonas hydrophila* type IV pilus biogenesis gene cluster: complementation of pilus assembly functions and characterization of a type IV leader peptidase/N-methyltransferase required for extracellular protein secretion. Molecular Microbiology 19, 857–869.

Peres, J.R.R., Mendes, I.C., Suhet, A.R., Vargas, M.A.T., 1993. Eficiência e competitividade de estirpes de rizóbio para a soja em solos de Cerrados. Revista Brasileira de Ciências do Solo 17, 357–363.

Perna, N.T., Plunkett, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., Blattner, F.R., 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529–533.

Riley, M., 1993. Functions of the gene products of *Escherichia coli*. Microbiological Reviews 57, 862–952.

Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., Mori, H., Perna, N.T., Plunkett, G., Rudd, K.E., Serres, M.H., Thomas, G.H., Thomson, N.R., Wishart, D., Wanner, B.L., 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005. Nucleic Acids Research 5, 1–9.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, second ed. Cold Spring Harbor Press, New York.

Semu, E., Hume, D.J., 1979. Effects of inoculation and fertilizer N levels on N_2 fixation and yields of soybeans. Canadian Journal of Plant Science 59, 1129–1137.

Simpson, A.J.G., 2001. Genome sequencing network. Nature Reviews Genetics 2, 979–983.

Skorpil, P., Saad, M.M., Bukll, N.M., Kobayashi, H., Ares-Orpel, F., Broughton, W.J., Deakin, W.J., 2005. NoP, a phosphorylated effector of *Rhizobium* sp. strain NGR234, is a major determinant of nodulation of tropical legumes *Flemingia congesta* and *Tephrosia vogelii*. Molecular Microbiology 57, 1304–1317.

Stanfield, S.W., Ielpi, L., O'Brochta, D., Helinski, D.R., Ditta, G.S., 1988. The *ndvA* gene product of *Rhizobium meliloti* is required for beta-(1–>2)glucan production and has homology to the ATP-binding export protein HlyB. Journal of Bacterioloy 170, 3523–3530.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 406, 959–964.

Tatusov, R., Galperin, M., Natale, D., Koonin, E., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28, 33–36.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit, Y., Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome. Proceedings of the National Academy of Sciences of the United States of America 102, 13950–13955.

Uniprot (The UniProt Consortium), 2007. The Universal protein resource. Nucleic Acids Research 35, D193–D197.

Vargas, M.A.T., Mendes, I.C., Suhet, A.R., Peres, J.R.R., 1993. Serological distribution of *Bradyrhizobium japonicum* from Brazilian "Cerrados" areas under soybean cultivation. Revista de Microbiologia 24, 239–243.

Vasconcelos, A.T.R., Almeida, D.F., Hungria, M., Guimarães, C.T., Antônio, R.V., Almeida, F.C., Almeida, L.G.P., Almeida, R., Alves-Gomes, J.A., Andrade, E.M., Araripe, J., Araújo, M.F.F., Astolfi-Filho, S., Azevedo, V., Baptista, A.J., Bataus, L.A.M., Batista, J.S., Beló, A., Van Den Berg, C., Bogo, M., Bonatto, S., Bordignon, J., Brigido, M.M., Brito, C.A., Brocchi, M., Burity, H.A., Camargo, A.A., Cardoso, D.D.P., Carneiro, N.P., Carraro, D.M., Carvalho, C.M.B., Cascardo, J.C.M., Cavada, B.S., Chueire, L.M.O., Creczynski-Pasa, T.B., Cunha-Junior, N.C., Fagundes, N., Falcão, C.L., Fantinatti, F., Farias, I.P., Felipe, M.S.S., Ferrari, L.P., Ferro, J.A., Ferro, M.I.T., Franco, G.R., Freitas, N.S.A., Furlan, L.R., Gazzinelli, R.T., Gomes, E.A., Gonçalves, P.R., Grangeiro, T.B., Grattapaglia, D., Grisard, E.C., Hanna, E.S., Jardim, S.N., Laurino, J., Leoi, L.C.T., Lima, L.F.A., Loureiro, M.F., Lyra, M.C.C.P., Madeira, H.M.F., Manfio, G.P., Maranhão, A.Q., Martins, W.S., Mauro, S.M.Z., Medeiros, S.R.B., Meissner, R.V., Moreira, M.A.M., Nascimento, F.F., Nicolás, M.F., Oliveira, J.G., Oliveira, S.C., Paixão, R.F.C., Parente, J.A., Pedrosa, F.O., Pena, S.D.J., Pereira, J.O., Pereira, M., Pinto, L.S.R.C., Pinto, L.S.P., Porto, J.I.R., Potrich, D.P., Ramalho-Neto, C.E., Reis, A.M.M., Rigo, L.U., Rondinelli, E., Santos, E.B.P., Santos, F.R., Schneider, M.P.C., Seuanez, H.N., Silva, A.M.R., Silva, A.L.C., Silva, D.W., Silva, R., Simões, I.C., Simon, D., Soares, C.M.A., Soares, R.B.A., Souza, E.M., Souza, K.R.L., Souza, R.C., Steffens, M.B.R., Steindel, M., Teixeira, S.R., Urmenyi, T., Vettore, A., Wassem, R., Zaha, A., Simpson, A.J.G., 2003. The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. Proceedings of the National Academy of Sciences of the United States of America 100, 11660–11665.

Viprey, V., Rosenthal, A., Broughton, W.K., Perret, X., 2000. Genetic snapshots of the *Rhizobium* species NGR234 genome. Genome Biology 1, 1–17.

Vrljic, M., Garg, J., Bellman, A., Wachi, S., Freudl, R., Malecki, M.J., Sahm, H., Kozina, V.J., Eggeling, L., Saier, M.H., 1999. The LysE superfamily: topology of the lysine exporter LysE of *Corynebacterium glutamicum*, a paradigm for a novel superfamily of transmembrane solute translocators. Journal of Molecular Microbiology and Biotechnology 1, 327–336.

Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G.G., Simon, M., Soll, D., Stetter, K.O., Short, J.M., Noordewier, M., 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proceedings of the National Academy of Sciences of the United States of America 100, 12984–12988.

Weber, D.F., Keyser, H.H., Uratsu, S.L., 1989. Serological distribution of *Bradyrhizobium japonicum* from U.S. soybean production areas. Agronomy Journal 81, 786–789.

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