MICROBIAL GENOMICS

RESEARCH ARTICLE

Bender *et al.*, *Microbial Genomics* 2022;8:000795

DOI 10.1099/mgen.0.000795

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Genetic variation in symbiotic islands of natural variant strains of soybean *Bradyrhizobium japonicum* and *Bradyrhizobium diazoefficiens* differing in competitiveness and in the efficiency of nitrogen fixation

Flavia Raquel Bender[1,2], Sheila Tiemi Nagamatsu[3], Jakeline Renata Marçon Delamuta[2,4], Renan Augusto Ribeiro[4], Marco Antonio Nogueira[2,4] and Mariangela Hungria[1,2,4,*]

Abstract

Soybean is the most important legume cropped worldwide and can highly benefit from the biological nitrogen fixation (BNF) process. Brazil is recognized for its leadership in the use of inoculants and two strains, *Bradyrhizobium japonicum* CPAC 15 (=SEMIA 5079) and *Bradyrhizobium diazoefficiens* CPAC 7 (=SEMIA 5080) compose the majority of the 70 million doses of soybean inoculants commercialized yearly in the country. We studied a collection of natural variants of these two strains, differing in properties of competitiveness and efficiency of BNF. We sequenced the genomes of the parental strain SEMIA 566 of *B. japonicum*, of three natural variants of this strain (S 204, S 340 and S 370), and compared with another variant of this group, strain CPAC 15. We also sequenced the genome of the parental strain SEMIA 586 of *B. diazoefficiens*, of three natural variants of this strain (CPAC 390, CPAC 392 and CPAC 394) and compared with the genome of another natural variant, strain CPAC 7. As the main genes responsible for nodulation (*nod*, *noe*, *nol*) and BNF (*nif*, *fix*) in soybean *Bradyrhizobium* are located in symbiotic islands, our objective was to identify genetic variations located in this region, including single nucleotide polymorphisms (SNPs) and insertions and deletions (indels), that could be potentially related to their different symbiotic phenotypes. We detected 44 genetic variations in the *B. japonicum* strains and three in *B. diazoefficiens*. As the *B. japonicum* strains have gone through a longer period of adaptation to the soil, the higher number of genetic variations could be explained by survival strategies under the harsh environmental conditions of the Brazilian Cerrado biome. Genetic variations were detected in genes enconding proteins such as a dephospho-CoA kinase, related to the CoA biosynthesis; a glucosamine-fructose-6-phosphate aminotransferase, key regulator of the hexosamine biosynthetic pathway; a LysR family transcriptional regulator related to nodulation genes; and NifE and NifS proteins, directly related to the BNF process. We suggest potential genetic variations related to differences in the symbiotic phenotypes.

DATA SUMMARY

The bacterial genomes of *B. japonicum* strain CPAC 15 and *B. diazoefficiens* strain CPAC 7 were previously deposited at the National Center for Biotechnology Information (NCBI) under the accession numbers CP007569 and ADOU00000000, respectively. The genome sequences obtained in this study were deposited at the NCBI. *B. japonicum* strains SEMIA 566, S 204, S 340 and S 370 are under the accession numbers SAMN23829029, SAMN23829998, SAMN2383152 and SAMN23833769, respectively, and *B. diazoefficiens* strains SEMIA 586, CPAC 390, CPAC 392 and CPAC 394 under the accession numbers SAMN23833944, SAMN23834001, SAMN23834014 and SAMN23837688, respectively.

Received 02 January 2022; Accepted 07 February 2022; Published 19 April 2022

Author affiliations: [1]Department of Biotechnology, Universidade Estadual de Londrina, C.P. 10011, 86057-970 Londrina, Paraná, Brazil; [2]Soil Biotechnology Laboratory, Embrapa Soja, C.P. 4006, 86085-981, Londrina-PR, Brazil; [3]Division of Human Genetics, Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA; [4]CNPq, SHIS QI 1 Conjunto B, Blocos A, B, C e D, Lago Sul, 71605-001, Brasília, Federal District, Brazil.

***Correspondence:** Mariangela Hungria, mariangela.hungria@embrapa.br;biotecnologia.solo@hotmail.com

Keywords: nodulation; symbiosis; *Glycine max*; bacteria adaptation; rhizobia; nitrogenase.

Abbreviations: BNF, biological nitrogen fixation; HGT, horizontal gene transfer; SDR, short-chain dehydrogenase (SDR); SNPs, single nucleotide polymorphisms; TCA, tricarboxylic acid; VGT, vertical gene transfer; YM, yeast-mannitol.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.

Impact Statement

There is an increasing demand for improving sustainability in agriculture, and the process of biological nitrogen fixation (BNF) represents a key strategy in this goal. Legumes are important for the possibility of establishing symbiosis with nitrogen-fixing bacteria, partially or fully supplying the plant's nitrogen requirement. Soybean is the legume most cropped worldwide and can largely benefit from BNF. Nowadays, Brazil is the first world soybean producer. A strain-selection programme of more than half a century resulted in strains very effective in fixing nitrogen, competitive, and adapted to the edaphic tropical conditions. Soybean bradyrhizobia are originally exotic to the country. Therefore, the programme searches for natural variability in desirable properties within specific strains. We studied two groups of strains, one of *Bradyrhizobium japonicum* and the other of *Bradyrhizobium diazoefficens*, both consisting of the parental and variant strains different in properties of efficiency of BNF and competitiveness. We sequenced bacterial genomes and compared their symbiotic islands. Single nucleotide polymorphisms (SNPs) and indels were identified, analysed and discussed regarding their putative role in affecting symbiotic properties. The variants that had a longer period of adaptation to the soils presented a higher number of genetic variations. Confirmation of the role of some of the identified proteins may speed up the strain-selection programme.

INTRODUCTION

Biological nitrogen fixation (BNF) is a key process for the maintenance of all ecosystems and a major player in the goal of agricultural sustainability. Archaea and bacteria were first present in free-living style, but about 100 million years ago (MYA), some lineages of prokaryotes started to develop mechanisms of interaction with plants. Some angiosperms evolved a predisposition towards the evolution of a symbiotic relationship with bacteria capable of fixing atmospheric nitrogen. These bacteria are collectively called rhizobia and symbiotic relationships were established, especially with members of the Fabaceae family, in an evolutionary process that involved complex steps of horizontal (HGT) and vertical (VGT) gene transfer [1–4].

For the symbiotic interaction to occur, several molecular signals are exchanged between the rhizobia and the host plants, resulting in infection of plant tissues and the formation of specific structures, the nodules, in general in the roots. In an established symbiosis, the host plant delivers carbon sources, mainly in the form of dicarboxylic acids, to attend to the bacterial energetic demands and receives nitrogen synthesized by the rhizobia. Among the rhizobia, the genus *Bradyrhizobium* might be considered the ancestor of all rhizobia, being a major symbiont of various legumes in the tropics, also standing up for great contributions to the agriculture and the environment [5–7].

For an efficient contribution of BNF in the nutrition of legumes, the first steps include the saprophytic ability of the bacteria in the soil, the competitiveness against the indigenous microbial community, and the compatibility with the host plant, with the recognition of signals that will allow the interaction and establishment of nodules [2, 8]. It is still not clear how natural selection shapes each symbiotic partner, and the extent to which the interaction can be affected by intrinsic and extrinsic factors [9]. Nodulation and BNF require the expression of specific genes in the rhizobia, organized in clusters and located in large symbiotic plasmids (pSym), or inserted in genomic islands, known as symbiosis or symbiotic islands, with profuse evidence of large contribution of events of HGT and VGT [9–11]. In *Bradyrhizobium* spp., genes related to BNF are generally grouped in symbiotic islands, characterized by lower GC content compared to the remaining genome [12, 13].

Brazil is recognized by its leadership in using the benefits of the BNF process with the soybean [*Glycine max* (L.) Merr.] crop [5, 14]. Soybean commercial crops started to expand in the late 1950s, simultaneously to strain-selection programmes aiming to satisfy all plants' requirements of nitrogen [5, 15]. In the 1990s, the selection of strains included, besides the BNF capacity, the adaptation to a new cropping area, the biome of Cerrado, an edaphic type of savannah challenging in abiotic stresses that include very acidic soils, with high aluminium and low phosphorus contents, high temperature and a long dry season period [5, 15]. As soybean is an exotic plant to Brazil, the soils are void of indigenous compatible bradyrhizobia [16]. Therefore, the approach taken has been to search for natural variants of foreigner strains after a period of adaptation to the soils, and selection for higher competitiveness and BNF capacity [15, 17, 18]. The programme resulted in the successful identification of two strains, *Bradyrhizobium japonicum* CPAC 15 (=SEMIA 5079) and *Bradyrhizobium diazoefficiens* CPAC 7 (=SEMIA 5080), largely used in commercial inoculants since 1992 [8, 19, 20]. The parental of CPAC 15 is strain SEMIA 566, isolated in 1966 from a nodule of a soybean inoculated with a North American inoculant and belonging to the same serogroup as the very competitive USDA 123 (15). The parental of CPAC 7 is strain SEMIA 586 (=CB 1809), considered a subculture of USDA 136, derived from USDA 122 in the USA; the strain was sent from USA to Australia and in 1966 from there to Brazil [15].

Other variant strains of SEMIA 566 and SEMIA 586 were obtained using the same approach of searching for differences in competitiveness, BNF efficiency and adaptation to the local soils. Morphological, physiological and genetic differences, and variability in competitiveness and efficiency of the BNF process have been reported in comparisons of parental and variant strains [17, 18, 21–24]. The present study aimed to identify single nucleotide polymorphisms (SNPs) and insertions and deletions (indels)

Table 1. Comparative performance, in relation to the genomes used as reference, of competitiveness and BNF efficiency (total *N* accumulated) of the strains of *B. japonicum* and *B. diazoefficiens* used in this study

B. japonicum*			***B. diazoefficiens**†*		
Strain	**BNF capacity**	**Competitiveness**	**Strain**	**BNF capacity**	**Competitiveness**
CPAC 15 (reference)			CPAC 7 (reference)		
SEMIA 566 (parental)	=	=	SEMIA 586 (parental)	=	<
S 204	<	=	CPAC 390	>	=
S 340	<	>	CPAC 392	=	<
S 370	>	>	CPAC 394	=	=

*After Hungria *et al*. [21].
†After Santos *et al.* [18].

present in symbiotic islands in two groups of strains with parental and variant strains. Group 1 is composed of the parental *B. japonicum* SEMIA 566 and the variant strains CPAC 15, S 204, S 340 and S 370. Group 2 is composed of the parental *B. diazoefficiens* SEMIA 586 and the variant strains CPAC 7, CPAC 390, CPAC 392 and CPAC 394. We searched for genetic variations and possible relationships with symbiotic phenotypes in the two groups of strains.

METHODS

Strains used

Two groups of strains were used in the study. The first is composed of the *B. japonicum* strain SEMIA 566 (parental), the natural variant CPAC 15 (=SEMIA 5079) used in commercial inoculants in Brazil since 1992 [15, 19], and three other natural variants, S 204, S 340 and S 370, characterized as having different properties of competitiveness and BNF efficiency [17, 21]. CPAC 15 was used as the genome of reference for this first group.

The second group is composed of *B. diazoefficiens* parental strain SEMIA 586 (=CB 1809), the natural variant CPAC 7 (=SEMIA 5080) used in commercial inoculants in Brazil since 1992 [19, 25], and three other natural variant strains, CPAC 390, CPAC 392 and CPAC 394, also showing differences in competitiveness and BNF efficiency [18, 26]. CPAC 7 was used as the genome of reference for this second group.

Information about the competitiveness and BNF efficiency within each group is shown in Table 1. All strains are deposited at the 'Diazotrophic and Plant Growth Promoting Bacteria Culture Collection of Embrapa Soja' (WFCC Collection No. 1213, WDCM Collection No. 1054), in Londrina, State of Paraná, Brazil.

Genome sequencing

First, the identity of all strains was confirmed by BOX-PCR, using the methodology as described before [27]. Genomes of the commercial strains *B. japonicum* CPAC 15 and *B. diazoefficiens* CPAC 7 have been sequenced before [24, 28]. Strains sequenced in this study comprise the parental *B. japonicum* SEMIA 566 and the natural variants S 204, S 340 and S 370, and the parental *B. diazoefficiens* SEMIA 586 and the natural variants CPAC 390, CPAC 392 and CPAC 394.

After growing the bacteria in modified-yeast-mannitol (YM) medium [29] till the exponential growth phase, 1.5 ml aliquots of each culture were submitted to the DNA extraction using the DNeasy Blood and Tissue Kit (Qiagen), following the manufacturer's instructions. After the extraction, the DNA of each strain was quantified in a Qubit fluorimeter (Invitrogen), and the concentration adjusted to 0.2 ng μl^{-1}.

The genomic libraries were built with the Nextera XT kit (Nextera DNA Library Preparation Kits), following the manufacturer's recommendations, with the steps of total DNA fragmentation; purification of the fragmented products; linkage of the indexes; and final purification. The validation of the libraries was performed on a 1% agarose gel. After confirming the quality, the libraries were quantified and their concentrations normalized to 4 nM. Genome sequencing was performed on a MiSeq equipment (Illumina) using the MiSeq Reagent v.3 kit (Illumina), at the Laboratory of Soil Biotechnology of Embrapa Soja. The read sequencing generated paired-end reads of ~300 bp.

The genomes were assembled with the A5-MiSeq Pipeline (*de novo*) v. 20140604, and the assembled genomes were deposited in the Rapid Annotation using Subsystem Technology (RAST) v.2.0. The genomes sequenced in this study were deposited at the NCBI. For the group of *B. japonicum* strains, the access numbers are SEMIA 566 (SAMN23829029), S 204 (SAMN23829998),

S 340 (SAMN2383152) and S 370 (SAMN23833769). For the group of *B. diazoefficiens*, the accession numbers are SEMIA 586 (SAMN23833944), CPAC 390 (SAMN23834001), CPAC 392 (SAMN23834014) and CPAC 394 (SAMN23837688).

Bioinformatic analysis

The genomes of *B. japonicum* CPAC 15 (accession number: CP007569) and *B. diazoefficiens* CPAC 7 (accession number: ADOU00000000) previously obtained by Siqueira *et al.* [28] were used as genomes of reference. They were retrieved from the genome database of NCBI in the FASTA format.

The bioinformatics analysis of the variants was performed separately for each *Bradyrhizobium* species, using the respective reference genome. The objective was to identify SNPs and insertions and deletions (indels) present in the symbiotic islands and to analyse the mutations at the protein level. To identify the position of the symbiotic islands in the genomes, the approach of Kaneko *et al.* [12, 13] was used. The authors proposed three regions called symbiotic islands in the genomes of *B. japonicum* USDA 6^T and *B. diazoefficiens* USDA 110^T, called Locus A, B and C.

The bioinformatics pipeline was divided in three steps: (1) reads alignment; (2) variant calling; and (3) variant calling annotation. Raw data in FASTQ format were submitted to the quality-control analysis using the basic metrics of the FastQC tool (v. 0.11.3). The alignment of the sequenced lineages against the respective reference genome was performed using Bowtie2 (v. 2.2.6). After the alignment, files in the SAM format were generated, requiring conversion to BAM format. This conversion was performed with the samtools (v. 0.1.19), subcommand view. After the conversion, all files were treated with Picard (v. 2.21.1), using the AddOrReplaceReadGroups and MarkDuplicates function, according to the GATK best practices. Following, the variant calling was conducted using the HaplotypeCaller function from Genome Analysis Toolkit (GATK; v 4.1.4.0). The SNPs and indels were then filtered using the VariantFiltration tool from the GATK (v 4.1.4.0).

The filtered variants were annotated using the Variant Ensembl Predictor (VEP) (https://www.ensembl.org/info/docs/tools/vep/index.html). SNPs and indels detected in the symbiotic islands were selected and the genetic variations were visualized with the Integrative Genomics Viewer (IGV) (v. 2.8.0). The selected genomic variations were evaluated manually in order to identify those located in repeated regions and to assess the frequency of the mutation in the sequenced reads, approach taken to verify their consistency.

Finally, the genes with selected mutations were annotated using BLASTX tool. The NCBI database allowed us to identify whether the mutations were present within the protein domain.

RESULTS

Genome features of the sequenced strains

Genome sequencing resulted in total coverages of 79, 81, 78, 78, 83, 94, 88 and 77× for strains SEMIA 566, S 204, S 340, S 370, SEMIA 586, CPAC 390, CPAC 392 and CPAC 394, respectively. Estimated genome sizes, number of assembled contigs and GC contents of each strain are shown in Table 2. In *B. japonicum* the genomes of strains S 204 and S 340 (10266898 bp and 10188322 bp, respectively) are larger than the genome of CPAC 15 (9582287 bp) used as reference for this group of strains. They could have acquired genes in the process of adaptation to the Cerrado soils. The other strains (SEMIA 566 and S 370) have genome sizes of 9561677 and 9546788 bp, respectively, very close to the reference genome. In *B. diazoefficiens* the genomes of SEMIA 586, CPAC 390, CPAC 392 and CPAC 394 (9067238, 9066558 bp, 9072300 and 9065038 bp, respectively) have similar sizes to the genome of CPAC 7(9085. 545 bp), used as a reference for this other group of strains.

Table 2. Genome sizes, G+C contents and contigs of the *B. japonicum* and *B. diazoefficiens* strains used in this study

B. diazoefficiens				***B. diazoefficiens***			
Strain	**Genome size (bp)**	**G+C (%)**	**Contigs**	**Strain**	**Genome size**	**G+C (%)**	**Contigs**
CPAC 15*†	9 582 287	63.5	1	CPAC 7*†	9 085 545	64.0	13
566	9 561 677	63.5	145	586	9 067 238	64.0	155
S 204	10 266 898	63.4	283	CPAC 390	9 066 558	64.0	155
S 340	10 188 322	63.4	281	CPAC 392	9 072 300	64.0	170
S 370	9 546 788	63.6	156	CPAC 394	9 065 038	64.0	150

*After Siqueira *et al.* [28].
†Strains used as a reference in the genomes' comparisons.

Analyses of genetic variations

For the analysis of polymorphisms in the symbiotic islands, the groups studied were selected based on the differences in two main properties for the success of the BNF process, competitiveness and the efficiency of BNF. In both groups, there were variants concerning these two properties (Table 1). The results obtained comparing each group of strains with the reference genomes of *B. japonicum* CPAC 15 and *B. diazoefficiens* CPAC 7 are shown in Table 3.

It is worth mentioning that indels are genetic variations that consist on the deletion or addition of one or more nucleotides and SNPs are the exchange of a single nucleotide by another. Table 3, which shows the nucleotides that are present in the reference strains and which codons these nucleotides are part of, also shows the changes resulting from genetic variations in genome position and the codons that are formed from the changes. Codon formation is based on the strand used as a template; therefore, it can be as similar or complementary nucleotides to those shown in the table.

In the *B. japonicum* group, the parental SEMIA 566 presented three genetic variations: two indels in the symbiotic island A, and one SNP in the symbiotic island C. These variations are present in proteins BJS_05223, BJS_05500 and BJS_07962, respectively (Table 3). Strains S 204 and S 340 shared 25 similar genetic variations. In relation to each strain, the S 204 variant presented 33 genetic variations, seven of which are indels present in proteins BJS_05269, BJS_05460, BJS_05577, BJS_05746, BJS_05464, BJS_05558 (island A) and BJS_08099 (island C). Other variations (26) are SNPs present in 18 proteins in island A, and eight in island C. The S 340 variant presented 33 genetic variations, of which six are indels present in proteins BJS_05269, BJS_05460, BJS_05577, BJS_05746 in the symbiotic island A, and BJS_08099 and BJS_08412 in island C. The other 27 variations are SNPs in 22 proteins in island A, and five in island C. No genetic variation was found in the symbiotic islands of S 370.

Of all genetic variations observed in *B. japonicum*, six are in hypothetical proteins, one in SEMIA 566 (BJS_05500), four in both strains S 204 and S 340 (BJS_05269, BJS_05544, BJS_05646 and BJS_08521), and one in S 204 (BJS_08142). No similarity of these proteins was found with any other protein deposited at the NCBI database.

In the *B. diazoefficiens* group, the parental SEMIA 586 showed one SNP in protein BJA5080_04871 present in the symbiotic island A (Table 3). The variant CPAC 394 showed two variations, both SNPs, in proteins BJA5080_04943 and BJA5080_04849 (island A). No genetic variation was found in the symbiotic islands of CPAC 390 and CPAC 392.

Bioinformatics analyses identified the position of the genetic variations in the bacterial genome, the substitution caused in the nucleotides, the position of the nucleotides and amino acids in the protein, and the amino acid resulting from the replacement of nucleotides. The results are shown in Table 3.

After identifying the SNPs and indels present in the symbiotic islands, BLASTX was performed to identify the proteins and search for domains. Table 4 shows the proteins identified and also indicates the mutations within their domains. In the *B. japonicum* group, in 12 out of the 44 polymorphisms we detected no putative conserved domains, six in hypothetical proteins. The putative proteins identified were related to several proteins that could be directly and indirectly related to BNF efficiency and competitiveness, and that will be discussed in details in the discussion section. In the *B. diazoefficiens* group, no putative conserved domain was detected in a PepSY domain-containing protein, while domains were detected in the other two polymorphisms detected, corresponding to the proteins MgtA and ArgE (Table 4).

DISCUSSION

Genome features of *B. japonicum* and *B. diazoefficiens* strains

Biological nitrogen fixation with the soybean crop has been considered as the most successful symbiosis of agronomic importance worldwide, contributing to the acquisition of millions of tons of nitrogen annually [5]. Some soybean *Bradyrhizobium* spp. strains of agronomic importance had their genome sequences, and we may cite the type strain of *B. japonicum* USDA 6^T, the Brazilian commercial strain *B. japonicum* CPAC 15, the type strain *B. diazoefficiens* USDA 110^T, carried in commercial inoculants in several countries [5], and the Brazilian commercial strain CPAC 7 of *B. diazoefficiens*. The genomes of these four strains are composed of a unique circular chromosome and estimated for *B. japonicum* at 9207384 bp for USDA 6^T, and 9582287 bp for CPAC 15, slightly bigger than the genomes of *B. diazoefficens* USDA 110^T, 9105828 bp, and CPAC 7, 9085545 bp. In all four genomes, no plasmids were detected [12, 13, 28].

We confirmed bigger genomes for the *B. japonicum* group in the strains sequenced in our study. It is worth mentioning that the variant strains of *B. japonicum* SEMIA 566 have gone through a longer período of adaptation to the Brazilian Cerrado soils, very acidic, with high aluminium and low phosphorus contents, with high temperature, and a long dry season period [30, 31]. Interestingly, compared to the reference strain CPAC 15, strains S 204 and S 340 have larger genomes, expanded in 684611 bp and 606035 bp, respectively (Table 2), which may result from gene acquisition through HGT transfer, that could be related to an increase in their saprophytic ability. The expansion in the genomes were majorly found outside the symbiotic

Table 3. Description of the position of the mutation in the genome, differences in the nucleotides and the results of the changes in amino acids

Mutation type	Mutation position in the genome (bp)	Nucleotides in reference strains CPAC 15 and CPAC 7	Nucleotides in variants	Symbiotic island	Gene ID	Codons in reference	Codons in variants	Exchange of amino acids	Position of nucleotides in the protein	Position of amino acids in protein
B. japonicum										
SEMIA 566										
indel	8 028 033	TGCG	T	A	BJS_05223	CCG·CAA	CAA	PQ/Q	590–592	197–198
indel	8 320 820	T	TCGCGATG	A	BJS_05500	GTC	GTC·GCG-ATG - C	V/VAMX	809–810	270
SNP	9 129 346	G	A	C	BJS_07962	CAG	TAG	Q/*	1378	460
S 204 and S 340										
indel	8 075 666	GA	G	A	BJS_05269	GAA	GA	E/X	314	105
indel	8 278 099	GCTC	G	A	BJS_05460	CTC	–	L/-	313–315	105
indel	8 420 952	GTGGAAT	G	A	BJS_05577	TGG·AAT	–	WN/-	136–141	46–47
indel	8 599 363	G	GCA	A	BJS_05746	GCC	GTGCC	A/VX	52–53	18
indel	9 284 755	C	CG	C	BJS_08099	CCG	CCCG	P/PX	1496–1497	499
SNP	7 933 105	C	A	A	BJS_05131	CAC	CAA	H/Q	3981	1327
SNP	7 952 042	C	A	A	BJS_05148	AGA	ATA	R/I	326	109
SNP	8 043 911	G	T	A	BJS_05235	CAG	AAG	Q/K	64	22
SNP	8 141 963	G	A	A	BJS_05330	GGA	GAA	G/E	596	199
SNP	8 191 404	C	A	A	BJS_05382	GTG	TTG	V/L	373	125
SNP	8 233 122	G	C	A	BJS_05420	TGG	TCG	W/S	449	150
SNP	8 268 647	C	G	A	BJS_05451	GAG	GAC	E/D	429	143
SNP	8 272 077	T	C	A	BJS_05454	ATG	ACG	M/T	521	174
SNP	8 277 718	A	G	A	BJS_05459	GAC	GGC	D/G	803	268
SNP	8 278 103	G	A	A	BJS_05460	GGC	AGC	G/S	316	106
SNP	8 378 062	T	C	A	BJS_05544	TAC	TGC	Y/C	86	29
SNP	8 452 478	G	C	A	BJS_05609	CTG	GTG	L/V	562	188
SNP	8 489 483	G	C	A	BJS_05642	ACC	AGC	T/S	68	23
SNP	8 494 065	A	G	A	BJS_05646	TAT	CAT	Y/H	22	8
SNP	8 603 761	A	G	A	BJS_05752	ACC	GCC	T/A	184	62
SNP	9 189 349	G	T	C	BJS_08013	CAG	AAG	Q/K	301	101
SNP	9 427 817	A	G	C	BJS_08521	CTT	CCT	L/P	179	60
SNP	9 508 237	A	G	C	BJS_08355	ATC	GTC	I/V	757	253
SNP	9 516 655	A	G	C	BJS_08364	ACC	GCC	T/A	508	170

Continued

Table 3. Continued

Mutation type	Mutation position in the genome (bp)	Nucleotides in reference strains CPAC 15 and CPAC 7	Nucleotides in variants	Symbiotic island	Gene ID	Codons in reference	Codons in variants	Exchange of amino acids	Position of nucleotides in the protein	Position of amino acids in protein
SNP	9 533 906	A	C	C	BJS_09076	ATC	AGC	I/S	1244	415
S 204										
indel	8 283 054	CGACGATCGAGCCACCCAGCTCGA	C	A	BJS_05464	GAC·GAT-CGA-GCC-ACC-CAG-CTC- GAG	G	DDRATQLE/X	292–314	98–105
indel	8 398 040	TG	T	A	BJS_05558	TGG	TG	W/X	1073	358
SNP	8 161 914	T	C	A	BJS_05348	ACC	GCC	T/A	79	27
SNP	8 312 231	T	C	A	BJS_05494	GTG	GCG	V/A	758	253
SNP	8 482 609	C	G	A	BJS_05635	AGC	AGG	S/R	675	225
SNP	9 335 708	G	A	C	BJS_08142	GCG	GTG	A/V	266	89
SNP	9 384 316	A	G	C	BJS_08489	AGG	GGG	R/G	16	6
SNP	9 482 621	G	C	C	BJS_08333	GCT	GGT	A/G	434	145
S 340										
Indel	9 569 452	G	GT	C	BJS_08412	CGT	CGTT	R/RX	158–159	53
SNP	8 125 727	T	G	A	BJS_05315	CTG	CGG	L/R	437	146
SNP	8 127 640	T	G	A	BJS_05317	TTT	TTG	F/L	960	320
SNP	8 264 276	C	G	A	BJS_05446	CGT	CCT	R/P	131	44
SNP	8 321 676	T	C	A	BJS_05501	GAG	GGG	E/G	752	251
SNP	8 457 225	C	T	A	BJS_05614	GCG	GTG	A/V	11	4
SNP	8 605 053	A	G	A	BJS_05753	GAG	GGG	E/G	125	42
SNP	8 606 552	G	A	A	BJS_05753	GAC	AAC	D/N	1624	542
B. diazoefficiens										
SEMIA 586										
SNP	1 988 610	C	T	A	BJA5080_04871	CCG	TCG	P/S	307	103
CPAC 394										
SNP	1 918 866	C	T	A	BJA5080_04943	GAA	AAA	E/K	106	36
SNP	2 014 149	G	A	A	BJA5080_04849	CCG	CTG	P/L	482	161

(*) Indicates the replacement of the amino acid by a stop codon. - (-) Indicates the loss of amino acids. (X) Indicates that it was not possible to identify the amino acid.

Table 4. Description of proteins and domains identified

Gene ID	Protein	Present in the domain	Domain	Access code
B. japonicum				
SEMIA 566				
BJS_05223	Dephospho-CoA kinase	Yes	CoaE	COG0237
BJS_05500	Hypothetical protein	No putative conserved domains have been detected		
BJS_07962	C4-dicarboxylate ABC transporter substrate-binding protein	Yes	PBP2_TAXI_TRAP_like_1	cd13569
S 204 and S 340				
BJS_05269	Hypothetical protein	No putative conserved domains have been detected		
BJS_05460	ABC transporter permeasse	Yes	TauC	COG0600
BJS_05577	Carboxymuconolactone decarboxylase family protein	Not		
BJS_05746	Nucleotidyltransferase	No putative conserved domains have been detected		
BJS_08099	Methionyl-tRNA synthetase	Yes	MetG	COG0143
BJS_05131	AAA family ATPase	Not		
BJS_05148	Nuclear transport factor two family protein	Yes	SnoaL_2	pfam12680
BJS_05235	Leucine--tRNA ligase	Não		
BJS_05330	SDR family oxidoreductase	Yes	FabG	COG1028
BJS_05382	DUF541 domain-containing protein	Yes	YggE	COG2968
BJS_05420	Primosomal protein N'	Yes	PriA	COG1198
BJS_05451	LysR family transcriptional regulator	Yes	LysR	COG0583
BJS_05454	LysR family transcriptional regulator	Yes	LysR	COG0583
BJS_05459	ABC transporter ATP-binding protein	Yes	TauB	COG1116
BJS_05460	ABC transporter permeasse	Yes	TauC	COG0600
BJS_05544	Hypothetical protein	No putative conserved domains have been detected		
BJS_05609	ComF family protein	Yes	ComFC	COG1040
BJS_05642	PilZ domain-containing protein	No putative conserved domains have been detected		
BJS_05646	Hypothetical protein	No putative conserved domains have been detected		
BJS_05752	DUF932 domain-containing protein	No putative conserved domains have been detected		
BJS_08013	Methylenetetrahydrofolate reductase [NAD(P)H]	Yes	MetF	COG0685
BJS_08521	Hypothetical protein	No putative conserved domains have been detected		
BJS_08355	Nitrogenase iron-molybdenum cofactor biosynthesis protein NifE	Yes	NifD	COG2710
BJS_08364	Cysteine desulfurase NifS	Yes	NifS	COG1104
BJS_09076	Transposase	Não		
S 204				
BJS_05464	Proline iminopeptidase	Yes	MhpC	COG0596
BJS_05558	Glycoside hydrolase family 15 protein	Yes	SGA1	COG3387
BJS_05348	Succinate dehydrogenase cytochrome b556 subunit	Yes	SdhC	COG2009

Continued

Table 4. Continued

Gene ID	Protein	Present in the domain	Domain	Access code
BJS_05494	ABC transporter permease	Yes	TauC	COG0600
BJS_05635	Cobaltochelatase subunit CobS	Yes	Glutenin_hmw	pfam03157
BJS_08142	Hypothetical protein	No putative conserved domains have been detected		
BJS_08489	Glucosamine-fructose-6-phosphate aminotransferase	Yes	GlmS	COG0449
BJS_08333	C4-dicarboxylic acid transport protein	Yes	GltP	COG1301
S 340				
BJS_08412	Transposase	Yes	HTH_32	pfam13565
BJS_05315	Alpha/beta fold hydrolase	No putative conserved domains have been detected		
BJS_05317	SDR family oxidoreductase	No putative conserved domains have been detected		
BJS_05446	GNAT family N-acetyltransferase	Yes	RimL	COG1670
BJS_05501	Aminotransferase class V-fold PLP-dependent enzyme	Yes	PucG	COG0075
BJS_05614	Preprotein translocase subunit SecA	Yes	PRK12904	PRK12904
BJS_05753	ParB/RepB/Spo0J family partition protein	Yes	Spo0J	COG1475
BJS_05753	ParB/RepB/Spo0J family partition protein	No putative conserved domains have been detected		
B. diazoefficiens				
SEMIA 586				
BJA5080_04871	HAD-IC family P-type ATPase	Yes	MgtA	COG0474
CPAC 394				
BJA5080_04943	M20/M25/M40 family metallo-hydrolase	Yes	ArgE	COG0624
BJA5080_04849	PepSY domain-containing protein	No putative conserved domains have been detected		

islands, indicating that gene acquisition was probably related to genes to improve saprophytic ability in the soil, but that could also impact competitiveness and efficiency of BNF.

Genomic islands are regions that form syntenic groups of multiple genes, conferring advantages to the bacteria compared to others that colonize the same environment [13]. The symbiosis islands are defined as regions with the integration of genes with specific functions, mostly related to the symbiosis, arising from HGT and presenting a GC content considerably smaller than the remaining genome [32]. In all strains sequenced in our study it was possible to identify three symbiotic islands, as defined by Kaneko *et al.* [12, 13].

It is worth mentioning that in *B. japonicum* USDA 6^T the sizes of symbiotic islands A, B and C were estimated at 694277, 4360 and 518647 bp, respectively, and in *B. diazoefficiens* USDA 110^T at 681342, 44 and 201222 bp, respectively, and that the main nodulation and nitrogen fixation genes are in symbiotic island A. In our study, in total, 44 polymorphisms were found in the *B. japonicum* group and three in the *B. diazoefficiens*. In the *B. japonicum* and *B. diazoefficiens* group, 25 and one polymorphisms, respectively, were found in island A, the main symbiotic island. In addition, eight and two polymorphisms, respectively, were found in island C. Island B is very small in comparison with the two other ones and no polymorphisms were detected in this region. It is worth to hypothesize that the polymorophisms in island A might affect more straightly BNF.

In the next items we will discuss some main polymorphisms found. However, it is worth mentioning that we found no polymorphisms in *B. japonicum* variant strain S 370 and in *B. diazoefficiens* variant strains CPAC 390 and 392. As the variant strains show differences in properties of competitiveness and efficiency of BNF, the result suggests that mutations outside the symbiotic islands may contribute to the differences in symbiotic phenotypes. However, to search for polymorphisms related to the symbiotic performance in the remaining chromosome represents a more difficult task and will be subject of a further study.

Genetic variation in the group of *B. japonicum* strains

Genetic variation in strain SEMIA 566

In the *B. japonicum* group, SEMIA 566 is the parental strain, characterized by a lower efficiency of BNF than CPAC 15 [21, 33]. The indel present in SEMIA 566, at the position 8028,033 bp, resulted in the deletion of the amino acid proline in the CoaE domain of the enzyme dephospho-CoA kinase (DPCK) EC 2.7.1.24, localized in the symbiotic island A. This enzyme uses ATP as a phosphate donor, generating the coenzyme A (CoA) and ADP [34, 35]. An efficient TCA (tricarboxylic acid) cycle in bacteroids requires the essential cofactor CoA, which is synthesized by the enzymatic reaction with pantothenic acid (vitamin B5), L-cysteine and purine/pyrimidine nucleotides as substrates [36]. The last stage of CoA biosynthesis is catalysed by the DPCK enzyme [37], essential for many organisms [35, 38]. To better understand the role of this enzyme, Nurkanto *et al.* [39] obtained a strain of *Entamoeba histolytica* with repressed *dpck* gene expression and verified that growth in culture medium was significantly decreased. Negative changes in the TCA cycle can also compromise the supply of ATP and nitrogenase reducer, affecting the BNF activity. In addition to energy generation, the TCA cycle plays a role in the biosynthesis of amino acids, purines, pyrimidines and vitamins [40, 41]. Therefore, it is feasible to assume that the enzyme DPCK in SEMIA 566 is affected, since it does not have the amino acid proline, responsible for stabilizing its structure. In contrast, the natural variant CPAC 15, isolated after some years of adaptation to the Cerrado soils, evolved to acquire a functional enzyme that may help in its saprophytic ability. The other three variant strains, S 204, S 340 and S370 were also derived from SEMIA 566 after the same period of adaptation to the Cerrado, and presented a functional DPCK enzyme.

Genetic variation in strains S 204 and S 340

The competitiveness of S 204 and S 340 is equal or higher than CPAC 15, respectively, but the efficiency of BNF of both is lower [21] (Table 1). The SNPs present in strains S 204 and S 340 at the position 8141963 bp resulted in the exchange of the amino acid glycine, present in the FabG domain, by the amino acid glutamate, in one short-chain dehydrogenase (SDR), localized in the symbiotic island A. The SDR family encompasses NADP-dependent oxidoreductases (H) [42, 43]. The wide variety of biochemical functions of oxidoreductases include detoxifying ethanol and xenobiotics, regulating hormones and signalling molecules, and detecting the redox state in the metabolism or transcription, all regulating vital cellular processes [44]. In humans, many SNPs have been identified in SDR genes and related to inherited metabolic diseases [43]. Therefore, the SNP in the enzyme SDR in rhizobia might affect the metabolic efficiency, resulting in a lower capacity of competing with other micro-organisms of S 204 and S 340, compared to CPAC 15. The substituted glycine can be part of the structure that follows a sequence pattern rich in highly variable glycine, critical for the structural integrity present in the Rossmann fold [44].

The SNP variation present in S 204 and S 340 at the position 8233122 bp resulted in the exchange of the amino acid tryptophan, present in the PriA domain, by the amino acid serine in the enzyme primosomal protein N', localized in island A. The protein N' is a component of the primosome, an enzyme complex involved in restarting the paralysed replication fork, ensuring complete DNA duplication [45, 46]. DNA duplication can be interrupted by breaks, gaps and/or proteins attached to the replication fork, and if the interruption is not resolved, the end of the DNA can be degraded. To guarantee the resumption of duplication, it is necessary to operate the primossoma complex [47]. Cells lacking the PriA protein suffer severe growth defects [47]. A study performed by Lee and Kornberg [48] showed that a *priA* mutation of *E. coli* (EL500) resulted in slower growth, with a generation time of 70 min, compared to 30 min in the wild-type strain. Furthermore, the PriA mutant had a filamentous morphology, with cells two to 50 times longer than the wild-type, and the chromosomal DNA of the filamentous cells often appeared to be disrupted. Colony formation analysis showed a 10- to 100-fold reduction in the PriA mutant compared to wild-type, showing that the duplication was also interrupted. Sandler *et al.* [49] obtained *priA* mutants of *E. coli* strain K-12. One mutant had a deletion of part of the gene and the insertion of a kanamycin resistance gene (priA1::kan), and the other had a simple insertion of the kanamycin resistance gene at the codon 153 (priA2::kan). They observed that the mutants priA1::kan and priA2::kan were sensitive to UV radiation and defective in recombination. Only the priA1::kan mutant showed a filamentous morphology. These studies show that it is likely that mutations at different points in a gene of a multifunctional protein can result in different phenotypes. McCool and Sandler [50] observed that cells of *E. coli* K-12 mutated for the *priA* gene presented filamentous morphology and hypothesized that the filament formation is a partition defect resulting from the inability to restart DNA replication. From the results obtained with *priA* mutants, we may hypothesize that the lower BNF capacity of strains S 204 and S 340 might result from difficulties in restarting DNA duplication for cell multiplication, resulting in growth defficiency [50].

The SNP polymorphism present in strains S 204 and S 340 at the position 8268647 bp resulted in the exchange of the amino acid glutamate in the LysR domain by the amino acid aspartate of the LysR family transcriptional regulator (ID: BJS_05451). The SNP at the position 8272077 bp resulted in the exchange of the amino acid methionine in the LysR domain by the amino acid threonine, also in an LysR family transcriptional regulator (ID: BJS_05454). Both SNPs are on the symbiotic island A. Although the enzymes BJS_05451 and BJS_05454 differ in size, with 291 and 317 amino acids, respectively, they showed 100% similarity in the LysR family of transcriptional regulators, a family first described in the 1980s by Henikoff *et al.* [51]. This is the largest family of prokaryotic transcriptional binding proteins and can regulate the expression of genes with different functions, involved in bacterial cell division, virulence, metabolism, quorum sensing, motility, antioxidant reaction, nitrogen

fixation [52]. Establishing an efficient nitrogen-fixing symbiosis between rhizobia and legumes depends on the adequate expression of nodulation genes controlled by LysR-type regulators [53]. In *Sinorhizobium* (=*Ensifer*) *meliloti* strain RM1021, two regulatory genes of the LysR type were found and named as *lsrA* and *lsrB*. When mutant strains of both genes and the wild-type strain were inoculated in alfalfa (*Medicago sativa* L.), the mutant strains indicated nitrogen deficiency [54]. Regarding the formation of nodules, plants inoculated with the wild-type strain had cylindrical nodules with pink internal colour, while those of the mutants had irregularly shaped nodules, many without pink internal colour. The results suggest that the growth of plants inoculated with the mutant strains was limited in nitrogen fixation due to a mutation in regulatory genes of the LysR type regulating nodulation genes and possibly other genes including those for nitrogen fixation [54]. Based on these studies, we hypothesize that mutations in strains S 204 and S 340 may help to explain the reported differences in BNF efficiency [21]. Interestingly, downstream the LysR-type transcriptional regulator (BJS_05451) is the BJS_05452 gene that encodes an HlyD family secretion protein. This protein is involved in the efflux pump of the RND family, which participates in the metabolism *Bradyrhizobium* cells within the plant, with a functional role in the specific excretion of 5-hydroxyflavonoids such as genistein [55], a strong *nod*-gene inducer. It is also worth mentioning that in *Bradyrhizobium elkanii* USDA 76^{T} it has been shown that the BNF efficiency with soybeans is related, among other factors, with amino acid metabolism [56]. The enzyme 3-isopropylmalate dehydratase performs the second step of the biosynthesis of the amino acid leucine, catalysing the stereospecific isomerization of 2-isopropylmalate and 3-isopropylmalate, through the formation of 2-isopropylmaleate. This enzyme is present in most prokaryotes and consists of a heterodimer composed of a large subunit (LeuC) and a small subunit (LeuD) [57, 58]. In the *B. japonicum* strains of our study, the genes encoding for LeuC and LeuD are present downstream of the gene encoding the transcriptional regulator of the LysR family (BJS_05454). Considering that the mutation present in the LysR-type transcriptional regulator decreases its regulatory efficiency and assuming that BJS_05454 regulates leuC and LeuD, we may suppose that this mutation interferes in the regulation of genes important for the properties of these strains.

Strains S 204 and S 340 showed SNPs at the positions 9508237 and 9516655 bp in proteins BJS_08355 (NifE) and BJS_08364 (NifS), respectively, both in the symbiotic island C. The *nif* genes are key for the BNF process [59], and studies with the diazotrophic *Klebsiella oxytoca* and *Azotobacter vinelandii* revealed that the nitrogenase biosynthesis requires the products of 16 *nif* genes (*nifHDKYTENXUSVZWMBQ*) [60]. Both *nifE* and *nifS* participate in the biosynthesis and insertion of FeMo-co into the nitrogenase, and *nifS* also participates in the assembling the Fe-S cluster [60–62]. In a study with *Frankia casuarinae*, a symbiont of *Casuarina* and *Allocasuarina*, mutants of *nifE* strains presented reduced growth in N-free medium [63]. In another study with *A. vinelandii* grown in N-free medium with Mo, Va or no metal, the NifEN mutants (Nif $^{-}$, Vnf $^{+}$ and Anf $^{+}$) and the NifS mutants showed poor growth in all conditions, with a deficiency in all three nitrogenase activities [64]. From these reports, and considering the resulting amino acid exchange, we may hypothesize that mutations in NifE and NifS proteins in strains S 204 and S 340 might affect the efficiency of the BNF. However, probably they do not interfere in competitiveness, and for this property it might be necessary to search also for SNPs outside the symbiotic island.

Genetic variation in strain S 204

The genetic variation detected in strain S 204 at the position 9384316 bp, in island C, resulted in the exchange of the amino acid arginine in the GlmS domain, by the amino acid glycine, in the enzyme glucosamine-fructose-6-phosphate aminotransferase (GFAT, EC 2.6.1.16). GFAT is the key regulator of the hexosamine biosynthetic pathway [65]. This cytoplasmic enzyme catalyses the first stage of the pathway, forming glucosamine 6-phosphate from glutamine, the amide donor and fructose-6-phosphate, derived from glucose. GFAT controls the flow of glucose and limits the speed of the hexosamine biosynthetic pathway, thus controlling the formation of the final product [66]. The glucosamine-fructose-6-phosphate aminotransferase is encoded by *nodM*, an important nodulation gene with a role in the biosynthesis and secretion of Nod factors [67]. An experiment with mutant strains in the *nodM* and *nodN* genes of *Rhizobium leguminosarum* inoculated in hairy vetch [*Vicia hirsuta* (L.) Gray] resulted in an average reduction of 30% in the number of nodules; however, no differences were observed in the formation of curled root hairs or infection threads, and the nodules formed fixed normal amounts of nitrogen [68]. In alfalfa inoculated with *S. meliloti* mutants in *nodM* and *nodN* genes, there was a significant delay in nodulation, but no effects were observed in the inoculation of *Melilotus albus Medik* [69]. These results indicate that *nodM* gene, which encodes GFAT, is transcribed in symbiosis with specific legumes, being putatively related to host range specificity. Being a paralogue of the housekeeping enzyme glucosamine synthase, in rhizobia there are reports of *nodM* being localized in symbiotic plasmids or islands. In *B. japonicum* USDA 6^{T} *nodM* is in symbiotic island A, but both in S 204 and in CPAC 15 is in island C.

Genetic variation in strain S 340

The SNP in strain S 340 at the position 8264276 bp resulted in the exchange of the amino acid arginine in the RimL domain by the amino acid proline in the GNAT family N-acetyltransferase enzyme, located in the symbiotic island A. GNATs comprise a large superfamily of enzymes responsible for a variety of biological processes, catalysing the transfer of an acetyl group from acetyl coenzyme A (AcCoA) to substrates such as aminoglycoside antibiotics, amino acids, polyamines, peptides, vitamins, catecholamines and large macromolecules, including proteins. Consequently, GNATs play roles in a large number of biological processes, including aminoglycoside antibiotic resistance, transcriptional regulation, protein acetylation and

stress response, allowing organisms to respond and adapt quickly to changing environmental conditions [70–72]. We found no studies with GNATs in *B. japonicum*. However, to explain the higher competitiveness of strain S 340 compared to S 204 we can hypothesize that the mutation may improve the tolerance to abiotic stresses, as found in the Brazilian Cerrado [31].

Strain S 370 is more competitive and efficient in BNF than both the parental SEMIA 566 and the commercial CPAC 15. However, no mutations were detected in the symbiotic islands of S 370, indicating that other genomic features might be searched for, including symbiotic genes outside the symbiotic islands, as well as other genes not necessarily related to the symbiosis.

Genetic variation in the group of *B. diazoefficiens* strains

Genetic variation in strain SEMIA 586

The SNP present in the parental strain SEMIA 586 at the position 1988610 bp of island A, when compared to the natural variant and more competitive strain CPAC 7, resulted in the exchange of the amino acid proline, present in the MgtA domain, by the amino acid serine in the protein HAD-IC family P-type ATPase. Found in all domains of life, type P ATPases, belonging to the haloacid dehydrogenase (HAD) superfamily are responsible for playing key roles in the transport of ions across biological membranes [73]. In *Listeria monocytogenes*, the P-type ATPase (CadA) is responsible for cadmium (Cd) resistance [74]. Another study identified the Cd-resistant strain F8027 of *L. monocytogenes*, and the inactivation of the *cadA4* gene (which encodes a P-type ATPase) resulted in decreased resistance to Cd, affecting bacterial growth [75]. Another bacterium that exhibits metal resistance conferred by a P-type ATPase is *Pseudomonas aeruginosa*, and the deletion of the *cadA* gene resulted in a susceptible phenotype, with a clear delay in growth kinetics in the presence of zinc [76]. The genome of the plant colonizing bacterium *Pseudomonas fluorescens* SBW25 has a putative P-type ATPase copper (Cu) carrier (CueA), and its inactivation reduced Cu tolerance. In general, the genes responsible for Cu tolerance are homologous to the Pl-type ATPase, and their deletion results in bacterial sensitivity to Cu [77]. Liang *et al.* [78] isolated a Cu-tolerant strain of *Bradyrhizobium liaoningense* from *Vigna unguiculata*. Two mutants were constructed (Bln-d and Bln-163), both with the P-type ATPase gene (*cueA*) interrupted. While the wild-type strain tolerated 2,0 mM of Cu_2^+, the mutants had their tolerance decreased to 0.8 mM. We may hypothesize that the mutation in the P-type ATPase enzyme in SEMIA 586 results in lower tolerance to heavy metals and, consequently, competitiveness compared to CPAC 7. As CPAC 7 has gone through a period of adaptation to the Cerrado soils [5, 15, 18, 19, 31] rich in aluminium [18, 31], the P-type ATPase could be involved in tolerance to this metal.

Final remarks

Native Brazilian soils are originally void of compatible soybean *Bradyrhizobium,* and a strain selection programme started in the 1960s and continues till today, resulting in the most successful contribution of BNF in agriculture worldwide [5]. The approach used in the selection of elite strains used in commercial inoculants in Brazil is based on the search for natural variability within a specific strain, aiming at higher competitiveness and efficiency of BNF, as well as adaptation to the local soil conditions [5, 6, 15, 18–23]. In this programme, natural variants were obtained from the parental strains of *B. japonicum* SEMIA 566 and *B. diazoefficens* SEMIA 586, each group obtained under the same conditions and time, and they had now their genomes sequenced and presented in this study. Natural variants of *B. japonicum* SEMIA 566 passed through a more extended period of adaptation to the harsh conditions of the Cerrado soils than the variant strains of *B. diazoefficens* SEMIA 586, which might explain the greater number of SNPs and indels identified. Many genetic variations may be related to the adaptation to the edaphic conditions, but improving adaptation may also affect competitiveness.

Our study suggests that SNPs and indels associated with missense and frameshift variation detected inside and outside protein domains may affect the structure and function of these proteins. Some of the proteins identified in our study are recognized as being directly or indirectly related to BNF and might affect the efficiency of the process. Additionally, polymorphisms were also found in hypothetical proteins and deserve further investigation. Gene edition is now a promising tool that might help to clarify the relation of the genetic variation detected and the symbiotic performance.

Funding information
Partially financed by INCT - Plant Growth-Promoting Microorganisms for Agricultural Sustainability and Environmental Responsibility (CNPq 465133/2014-4, Fundação Araucária-STI 043/2019, CAPES).

Acknowledgements
Flavia R. Bender acknowledges a PhD fellowship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Finance Code 001). Marco A. Nogueira and Mariangela Hungria are also CNPq research fellows.

Author contributions
Conceptualization: M.H. Methodology: S.T.N., F.R.B., M.H. Formal analysis: F.R.B., S.T.N., J.R.M., R.A.R., M.H. Investigation: F.R.B., S.T.N. Resources: M.H. Data curation: F.R.B., S.T.N. Writing: F.R.B., M.H., S.T.N. Reviewing: All authors. Supervision: M.H. Project administration: M.H.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Ethical statement

All material used in this manuscript does not need any external consent for publication. All authors have read the final version and given their consent for publication.

References

1. Lloret L, Martínez-RomerO E. Evolución y filogenia de rhizobium. *Rev Latinoam Microbiol* 2005;47:43–60.
2. Sachs JL, Skophammer RG, Regus JU. Evolutionary transitions in bacterial symbiosis. *Proc Natl Acad Sci U S A* 2011;108 Suppl 2:10800–10807.
3. Werner GDA, Cornwell WK, Sprent JI, Kattge J, Kiers ET. A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. *Nat Commun* 2014;5:4087.
4. Daubech B, Remigi P, Doin de Moura G, Marchetti M, Pouzet C, *et al*. Spatio-temporal control of mutualism in legumes helps spread symbiotic nitrogen fixation. *elife* 2017;6:e28683.
5. Hungria M, Mendes IC. Nitrogen fixation with soybean: the perfect symbiosis. In: de Bruijn FJ (eds). *Biological Nitrogen Fixation*, vol. 2. New Jersey: John Wiley & Sons; 2015. pp. 1005–1019.
6. Hungria M, Barcellos FG, Mendes IC, Chueire LMO, Ribeiro RA, *et al*. Introdução, Estabelecimento e Adaptação de Bradirrizóbios Simbiontes da Soja em Solos Brasileiros. In: Yamada-Ogata SF, Nakazato G, Furlaneto MC and Nogueira MA (eds). *Tópicos Especiais Em Microbiologia*. Londrina: UEL; 2015. pp. 243–261.
7. Hungria M, Menna P, Delamuta JRM. Bradyrhizobium, the ancestor of all rhizobia: phylogeny of housekeeping and nitrogen-fixation genes. In: de Bruijn FJ (eds). *Biological Nitrogen Fixation*, vol. 1. New Jersey: John Wiley & Sons; 2015. pp. 191–202.
8. de Souza GK, Sampaio J, Longoni L, Ferreira S, Alvarenga S, *et al*. Soybean inoculants in Brazil: an overview of quality control. *Braz J Microbiol* 2019;50:205–211.
9. Jaiswal SK, Dakora FD. Widespread distribution of highly adapted *Bradyrhizobium* species nodulating diverse legumes in Africa. *Front Microbiol* 2019;10:310.
10. MacLean AM, Finan TM, Sadowsky MJ. Genomes of the symbiotic nitrogen-fixing bacteria of legumes. *Plant Physiol* 2007;144:615–622.
11. Menna P, Hungria M. Phylogeny of nodulation and nitrogen-fixation genes in *Bradyrhizobium*: supporting evidence for the theory of monophyletic origin, and spread and maintenance by both horizontal and vertical transfer. *Int J Syst Evol Microbiol* 2011;61:3052–3067.
12. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, *et al*. Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. *DNA Res* 2002;9:189–197.
13. Kaneko T, Maita H, Hirakawa H, Uchiike N, Minamisawa K, *et al*. Complete genome sequence of the soybean symbiont *Bradyrhizobium japonicum* strain USDA6T. *Genes (Basel)* 2011;2:763–787.
14. Santos MS, Nogueira MA, Hungria M. Microbial inoculants: reviewing the past, discussing the present and previewing an outstanding future for the use of beneficial bacteria in agriculture. *AMB Express* 2019;9:205.
15. Hungria M, Campo RJ, Mendes IC, Graham PH. 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 RP, Shankar N and Jaiwal PK (eds). *Nitrogen Nutrition and Sustainable Plant Productivity*. Houston, Texas: Studium Press, LLC; 2006. pp. 43–93.
16. Ferreira MC, Hungria M. Recovery of soybean inoculant strains from uncropped soils in Brazil. *Field Crops Res* 2002;79:139–152.
17. Boddey LH, Hungria M. Phenotypic grouping of Brazilian *Bradyrhizobium* strains which nodulate soybean. *Biol Fertil Soils* 1997;25:407–415.
18. Santos MA, Vargas MAT, Hungria M. Characterization of soybean *Bradyrhizobium* strains adapted to the Brazilian savannas. *FEMS Microbiol Ecol* 1999;30:261–272.
19. Vargas MAT, Mendes IC, Suhet AR, Peres JRR. Duas novas cepas de rizóbio para inoculação em soja. *Brasília:EMBRAPA-SPI* 1992:1–3.
20. Peres JRR, Mendes IC, Suhet AR, Vargas MAT. Eficiência e competitividade de estirpes de rizóbios para soja em solos do cerrado. *R Bras Ci Solo* 1993;17:357–363.
21. Hungria M, Boddey LH, Santos MA, Vargas MAT. Nitrogen fixation capacity and nodule occupancy by *Bradyrhizobium japonicum* and *B. elkanii* strains. *Biol Fertil Soils* 1998;27:393–399.
22. Barcellos FG, Batista J da S, Menna P, Hungria M. Genetic differences between *Bradyrhizobium japonicum* variant strains contrasting in N_2-fixation efficiency revealed by representational difference analysis. *Arch Microbiol* 2009;191:113–122.
23. Batista JS da S, Torres AR, Hungria M. Towards a two-dimensional proteomic reference map of *Bradyrhizobium japonicum* CPAC 15: spotlighting "hypothetical proteins." *Proteomics* 2010;10:3176–3189.
24. Gomes DF, da Silva Batista JS, Rolla AAP, da Silva LP, Bloch C, *et al*. Proteomic analysis of free-living *Bradyrhizobium diazoefficiens*: highlighting potential determinants of a successful symbiosis. *BMC Genomics* 2014;15:643.
25. Hungria M, Franchini JC, Campo RJ, Crispino CC, Moraes JZ, *et al*. Nitrogen nutrition of soybean in Brazil: Contributions of biological N_2 fixation and N fertilizer to grain yield. *Can J Plant Sci* 2006;86:927–939.
26. Barcellos FG, Menna P, da Silva Batista JS, Hungria M. Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous diazotrophs *Sinorhizobium (Ensifer) fredii* and *Bradyrhizobium elkanii* in a Brazilian Savannah soil. *Appl Environ Microbiol* 2007;73:2635–2643.
27. Chibeba AM, Kyei-Boahen S, Guimarães M de F, Nogueira MA, Hungria M. Isolation, characterization and selection of indigenous *Bradyrhizobium* strains with outstanding symbiotic performance to increase soybean yields in Mozambique. *Agric Ecosyst Environ* 2017;246:291–305.
28. Siqueira AF, Ormeño-Orrillo E, Souza RC, Rodrigues EP, Almeida LGP, *et al*. Comparative genomics of *Bradyrhizobium japonicum* CPAC 15 and *Bradyrhizobium diazoefficiens* CPAC 7: elite model strains for understanding symbiotic performance with soybean. *BMC Genomics* 2014;15:1.
29. Hungria M, O'Hara GW, Zilli JE, Araujo RS, Deaker R, *et al*. Isolation and growth of rhizobia. In: Howieson JG and Dilworth MJ (eds). *Working with Rhizobia*. Canberra: Australian Centre for International Agriculture Reserch (ACIAR); 2016. pp. 39–60.
30. Hungria M, Vargas MAT, Andrade DS, Campo RJ, Chueire LMO. *et al*Fixação biológica do nitrogênio em leguminosas de grãos. In: Siqueira JO, Moreira FMS, Lopes AS, Guilherme LR, Faquin V, et al. (eds). *Soil Fertility, Soil Biology and Plant Nutrition Interrelationships*. Lavras: SBCS/UFLA/DCS; 1999. pp. 597–620.
31. Hungria M, Vargas MAT. Environmental factors affecting N_2 fixation in grain legumes in the tropics, with an emphasis on Brazil. *Field Crops Res* 2000;65:151–164.
32. Göttfert M, Röthlisberger S, Kündig C, Beck C, Marty R, *et al*. Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the *Bradyrhizobium japonicum* chromosome. *J Bacteriol* 2001;183:1405–1412.
33. Hungria M, Vargas MAT. Exploring the microbial diversity and soil management practices to optimize the contribution of soil microorganisms to plant nutrition. In: Stacey G, Mullin B and Gresshoff P (eds). *Biology of Plant-Microbe Interactions*. 1996. pp. 493–496.
34. Spry C, Kirk K, Saliba KJ. Coenzyme A biosynthesis: an antimicrobial drug target. *FEMS Microbiol Rev* 2008;32:56–106.

35. Walia G, Kumar P, Surolia A. The role of UPF0157 in the folding of *M. tuberculosis* dephosphocoenzyme A kinase and the regulation of the latter by CTP. *PLoS One* 2009;4:e7645.
36. Genschel U, Powell CA, Abell C, Smith AG. The final step of pantothenate biosynthesis in higher plants: cloning and characterization of pantothenate synthetase from *Lotus japonicus* and *Oryza sativum* (rice). *Biochem J* 1999;341 (Pt 3):669–678.
37. Abiko Y. Pantothenic acid and coenzyme A: dephospho-coa pyrophosphorylase and dephospho-coa kinase as A possible bifunctional enzyme complex 1 (ATP: pantetheine-4'-phosphate adenyltransferase, EC 2.7.7.3 and ATP: dephospho- coa 3'-phosphotransferase, EC 2.7.1.24). *Methods Enzymol* 1970:358–364.
38. Hart RJ, Abraham A, Aly ASI. Genetic characterization of coenzyme a biosynthesis reveals essential distinctive functions during malaria parasite development in blood and mosquito. *Front Cell Infect Microbiol* 2017;7:260.
39. Nurkanto A, Jeelani G, Yamamoto T, Hishiki T, Naito Y, *et al*. Biochemical, metabolomic, and genetic analyses of dephospho coenzyme a kinase involved in coenzyme a biosynthesis in the human enteric parasite *Entamoeba histolytica Front Microbiol* 2018;9:2902.
40. Dunn MF. Tricarboxylic acid cycle and anaplerotic enzymes in rhizobia. *FEMS Microbiol Rev* 1998;22:105–123.
41. Green LS, Li Y, Emerich DW, Bergersen FJ, Day DA. Catabolism of alpha-ketoglutarate by a *sucA* mutant of *Bradyrhizobium japonicum*: evidence for an alternative tricarboxylic acid cycle. *J Bacteriol* 2000;182:2838–2844.
42. Oppermann U, Filling C, Hult M, Shafqat N, Wu X, *et al*. Short-chain dehydrogenases/reductases (SDR): the 2002 update. *Chem Biol Interact* 2003;143–144:247–253.
43. Persson B, Kallberg Y, Bray JE, Bruford E, Dellaporta SL, *et al*. The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem Biol Interact* 2009;178:94–98.
44. Kavanagh KL, Jörnvall H, Persson B, Oppermann U. The SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell Mol Life Sci* 2008;65:3895–3906.
45. Lasken RS, Kornberg A. The primosomal protein n' of *Escherichia coli* is a DNA helicase. *J Biol Chem* 1988;263:5512–5518.
46. Huang Y-H, Lien Y, Huang C-C, Huang C-Y, Korolev S. Characterization of *Staphylococcus aureus* Primosomal DnaD protein: highly conserved C-terminal region is crucial for ssDNA and PriA helicase binding but not for DnaA protein-binding and self-tetramerization. *PLoS ONE* 2016;11:e0157593.
47. Michel B, Sandler SJ. Replication restart in bacteria. *J Bacteriol* 2017;199:e00102-17.
48. Lee EH, Kornberg A. Replication deficiencies in *priA* mutants of *Escherichia coli* lacking the primosomal replication n' protein. *Proc Natl Acad Sci U S A* 1991;88:3029–3032.
49. Sandler SJ, Samra HS, Clark AJ. Differential suppression of priA2::kan phenotypes in *Escherichia coli* K-12 by mutations in *priA*, *lexA*, and *dnaC*. *Genetics* 1996;143:5–13.
50. McCool JD, Sandler SJ. Effects of mutations involving cell division, recombination, and chromosome dimer resolution on a *priA2::kan* mutant. *Proc Natl Acad Sci U S A* 2001;98:8203–8210.
51. Henikoff S, Haughn GW, Calvo JM, Wallace JC. A large family of bacterial activator proteins. *Proc Natl Acad Sci U S A* 1988;85:6602–6606.
52. Yang W, Wang W-Y, Zhao W, Cheng J-G, Wang Y, *et al*. Preliminary study on the role of novel LysR family gene *kp05372* in *Klebsiella pneumoniae* of forest musk deer. *J Zhejiang Univ Sci B* 2020;21:137–154.
53. van Rhijn P, Vanderleyden J. The Rhizobium-plant symbiosis. *Microbiol Rev* 1995;59:124–142.
54. Luo L, Yao S-Y, Becker A, Rüberg S, Yu G-Q, *et al*. Two new *Sinorhizobium meliloti* LysR-type transcriptional regulators required for nodulation. *J Bacteriol* 2005;187:4562–4572.
55. Takeshima K, Hidaka T, Wei M, Yokoyama T, Minamisawa K, *et al*. Involvement of a novel genistein-inducible multidrug efflux pump of *Bradyrhizobium japonicum* early in the interaction with *Glycine max* (L.) Merr. *Microbes Environ* 2013;28:414–421.
56. Cooper B, Campbell KB, Beard HS, Garrett WM, Mowery J, *et al*. A proteomic network for symbiotic nitrogen fixation efficiency in *Bradyrhizobium elkanii*. *Mol Plant Microbe Interact* 2018;31:334–343.
57. Gruer MJ, Artymiuk PJ, Guest JR. The aconitase family: three structural variations on a common theme. *Trends Biochem Sci* 1997;22:3–6.
58. Yasutake Y, Yao M, Sakai N, Kirita T, Tanaka I. Crystal structure of the *Pyrococcus horikoshii* isopropylmalate isomerase small subunit provides insight into the dual substrate specificity of the enzyme. *J Mol Biol* 2004;344:325–333.
59. Olson JW, Agar JN, Johnson MK, Maier RJ. Characterization of the NifU and NifS Fe-S cluster formation proteins essential for viability in *Helicobacter pylori*. *Biochemistry* 2000;39:16213–16219.
60. Li Q, Chen S. Transfer of Nitrogen Fixation (*nif*) Genes to Non-diazotrophic Hosts. *Chembiochem* 2020;21:1717–1722.
61. Aguilar OM, Taormino J, Thöny B, Ramseier T, Hennecke H, *et al*. The *nifEN* genes participating in FeMo cofactor biosynthesis and genes encoding dinitrogenase are part of the same operon in *Bradyrhizobium* species. *Mol Gen Genet* 1990;224:413–420.
62. Hu Y, Fay AW, Lee CC, Yoshizawa J, Ribbe MW. Assembly of nitrogenase MoFe protein. *Biochemistry* 2008;47:3973–3981.
63. Kucho K-I, Tamari D, Matsuyama S, Nabekura T, Tisa LS. Nitrogen fixation mutants of the Actinobacterium *Frankia casuarinae* CcI3. *Microbes Environ* 2017;32:344–351.
64. Kennedy C, Dean D. The *nifU*, *nifS* and *nifV* gene products are required for activity of all three nitrogenases of *Azotobacter vinelandii*. *Mol Gen Genet* 1992;231:494–498.
65. Kato N, Dasgupta R, Smartt CT, Christensen BM. Glucosamine:fructose-6-phosphate aminotransferase: gene characterization, chitin biosynthesis and peritrophic matrix formation in *Aedes aegypti*. *Insect Mol Biol* 2002;11:207–216.
66. McKnight GL, Mudri SL, Mathewes SL, Traxinger RR, Marshall S, *et al*. Molecular cloning, cDNA sequence, and bacterial expression of human glutamine:fructose-6-phosphate amidotransferase. *J Biol Chem* 1992;267:25208–25212.
67. Safronova VI, Kuznetsova IG, Sazanova AL, Kimeklis AK, Belimov AA, *et al*. Extra-slow-growing Tardiphaga strains isolated from nodules of *Vavilovia formosa* (Stev.) Fed. *Arch Microbiol* 2015;197:889–898.
68. Surin BP, Downie JA. Characterization of the *Rhizobium leguminosarum* genes nodLMN involved in efficient host-specific nodulation. *Mol Microbiol* 1988;2:173–183.
69. 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.
70. Dyda F, Klein DC, Hickman AB. GCN5-related N-acetyltransferases: A structural overview. *Annu Rev Biophys Biomol Struct* 2000;29:81–103.
71. Hentchel KL, Escalante-Semerena JC. Acylation of biomolecules in prokaryotes: a widespread strategy for the control of biological function and metabolic stress. *Microbiol Mol Biol Rev* 2015;79:321–346.
72. Shirmast P, Ghafoori SM, Irwin RM, Abendroth J, Mayclin SJ, *et al*. Structural characterization of a GNAT family acetyltransferase from *Elizabethkingia anophelis* bound to acetyl-CoA reveals a new dimeric interface. *Sci Rep* 2021;11:1.
73. Aravind L, Galperin MY, Koonin EV. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem Sci* 1998;23:127–129.
74. Lebrun M, Audurier A, Cossart P. Plasmid-borne cadmium resistance genes in *Listeria monocytogenes* are similar to *cadA* and *cadC* of *Staphylococcus aureus* and are induced by cadmium. *J Bacteriol* 1994;176:3040–3048.

75. Parsons C, Lee S, Jayeola V, Kathariou S. Novel cadmium resistance determinant in *Listeria monocytogenes*. *Appl Environ Microbiol* 2017;83:e02580-16.
76. Ducret V, Gonzalez MR, Leoni S, Valentini M, Perron K. The CzcCBA efflux system requires the CadA P-Type ATPase for timely expression upon zinc excess in *Pseudomonas aeruginosa* *Front Microbiol* 2020;11:911.
77. Zhang X-X, Rainey PB. The role of a P1-type ATPase from *Pseudomonas fluorescens* SBW25 in copper homeostasis and plant colonization. *Mol Plant Microbe Interact* 2007;20:581–588.
78. Liang J, Zhang M, Lu M, Li Z, Shen X, *et al*. Functional characterization of a *csoR-cueA* divergon in *Bradyrhizobium liaoningense* CCNWSX0360, involved in copper, zinc and cadmium cotolerance. *Sci Rep* 2016;6:35155.