Genetic variation in host-specific competitiveness of the symbiont *Rhizobium leguminosarum* symbiovar *viciae*

Stéphane Boivin[1], Frederic Mahé[2], Marjorie Prevent[1], Mathilde Tancelin[1], Marc Tauzin[1], Jerzy Wielbo[3], Sylvie Mazurier[4], Peter Young[5] and Marc Lepetit[1]

[1] Laboratoire des Symbioses Tropicales et Méditerranéennes (LSTM) INRA, IRD, CIRAD, University of Montpellier, Montpellier SupAgro, Montpellier, France

[2] Biologie et Génétique des Interactions Plante-Parasite (BGPI), University of Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

[3] Department of Genetics and Microbiology, Maria Curie-Sklodowska University, Lublin, Poland

[4] Agroecology, AgroSup Dijon, INRA, University of Bourgogne Franche-Comté, F-21000 Dijon, France

[5] Department of Biology, University of York, York, United Kingdom

Corresponding author: Marc Lepetit

Running title: Rhizobial competition for nodule formation

Abstract: 212 words

Introduction: 1164 words

Material and Methods: 942 words

Results: 2617 words

Discussion: 1772 words

Abstract

Legumes in the tribe Fabeae form nitrogen-fixing root nodules resulting from symbiotic interaction with the soil bacteria *Rhizobium leguminosarum* symbiovar *viciae* (*Rlv*). These bacteria are all potential symbionts of the Fabeae hosts but display variable competitiveness to form root nodules (CFN) when co-inoculated in mixture. Because CFN and nitrogen-fixation capacity behave generally as genetically independent traits, the efficiency of symbiosis is often suboptimal when Fabeae legumes are exposed to natural bacterial populations present in soil. A core collection of 32 bacteria was constituted based on the genomic comparison of a collection of 121 genome sequences representative of known worldwide diversity of the rhizobium symbiovar. A variable part of the *nodD* gene sequence was used as a DNA barcode to discriminate and quantify each of the 32 bacteria in a mixture. This core collection was co-inoculated on a panel of 9 genetically diverse *Pisum sativum*, *Vicia faba* and *Lens culinaris* cultivars. We estimated the relative CFN of the bacteria with the diverse hosts by DNA metabarcoding on the nodulated root systems. Comparative genomic analyses within the bacterial core collection allowed us to identify genes associated with host-dependent CFN. These results highlight the emergent properties of rhizobial populations and present a new strategy to identify genetic markers related to important symbiotic traits operating at this level.

Keywords: *Rhizobium leguminosarum* symbiovar *viciae,* competitiveness, core collection, pea, fababean, lentil, Fabeae, DNA metabarcoding, symbiosis.

Introduction

Legumes can escape nitrogen-deficit conditions in soil by interacting with bacteria known as rhizobia to form nitrogen-fixing root nodules. Rhizobia in natural populations are often genetically diverse (Bailly et al., 2011). They are generally gathered in 'symbiovars' based on their host specificity (Triplett & Sadowsky, 1992). The Ability to Form root Nodules (AFN) in mono-inoculation with specific legume tribes defines this host specificity. For most of the legume-rhizobium symbioses, AFN is strongly associated with horizontally transferred symbiosis-related regions of the genome, present on specific plasmids or islands (Young, 2016). These regions carry the *nod* genes involved in the synthesis and secretion of lipo-chito-oligosaccharide Nod Factors (NFs) recognized by the plant. Therefore AFN is associated with the diversity of the *nod* genes rather than with that of the chromosome (Triplett & Sadowsky, 1992; Kumar et al., 2015). *Rhizobium leguminosarum* symbiovar *viciae* (*Rlv*) is the specific symbiont of the Fabeae tribe that includes important crops such as pea (*Pisum sativum* L.), fababean (*Vicia faba* L.) and lentil (*Lens culinaris* L.). Our knowledge of *Rlv* genomic diversity has recently been improved by the release of 50 diverse genome sequences in Genbank (Boivin et al., 2020). *Rlv* is a complex species composed of at least 5 different genospecies that are not symbiovar-specific (Kumar et al., 2015; Boivin et al., 2020; Cavassim et al., 2020). The *nodD* gene encodes a transcriptional regulator of the bacterial symbiosis genes. It has frequently been used as a marker to discriminate *R. leguminosarum* symbiovars (Zézé, Mutch, & Young, 2001; Laguerre, Louvrier, Allard, & Amarger, 2003; Boivin et al., 2020). Although *Rlv* bacteria generally have the capacity to form root nodules with most Fabeae legumes, competition occurs within *Rlv* mixtures and only bacteria displaying the best Competitiveness to Form root Nodules (CFN) finally occupy the roots of the host plants (for review see Boivin &

Lepetit, 2020). Therefore the *Rlv* populations associated with Fabeae root nodules do not necessarily reflect the proportions of bacteria in the soil. CFN varies greatly depending on both the legume host and the *nod* alleles (Boivin et al., 2020). Generally, a poor association was found between CFN and the level of Symbiotic N_2 Fixation (SNF; Bourion et al., 2018; Boivin et al., 2020). This likely explains the frequent failure of inoculation strategies with highly effective *Rlv* strains, due to a higher competitiveness of indigenous ineffective bacteria as compared to inoculated strains (Fesenko, Provorov, Orlova, Orlov, & Simarov, 1995; McKenzie et al., 2001; Laguerre et al., 2003). Early partner choice, AFN and CFN are not the only mechanisms responsible for partner choice between symbiotic partners. As the young nodule becomes N_2-fixing, local and systemic post-infection mechanisms related to the plant nitrogen demand pilot the development of the nascent nodule and may sanction inefficient bacteria (for review, see Boivin and Lepetit, 2020).

Little is known about mechanisms controlling CFN. They might be related either to plant-microbe and/or to microbe-microbe interactions. Antibiosis and quorum-sensing mechanisms modulating the multiplication of free-living *Rlv* bacteria have been reported and are potentially involved in CFN (Robleto, Scupham, & Triplett, 1997; McAnulla, Edwards, Sanchez-Contreras, Sawers, & Downie, 2007; Naamala, Jaiswal, & Dakora, 2016). However, even if the preferential proliferation of certain rhizobium genotypes within host rhizospheres contributes to CFN, the plant-microbe interaction is probably a major driver of CFN (Moawad, Ellis, & Schmidt, 1984; Laguerre et al., 2003; for review see Boivin and Lepetit, 2020). Based on co-inoculation strategies with high densities of *Rlv* mixtures, recent studies showed that pea and fababean preferentially select different *Rlv* genotypes (Boivin et al., 2020). Candidate genes and/or genetic markers associated with pea/fababean CFN were identified. Most of

these genomic sequences belonged to *Rlv* plasmids, in agreement with the hypothesis that they are components of horizontally transferred symbiotic traits. They included some nodulation genes (*nod* genes) such as *nodM*, *nodN*, *nodT* and *nodO.* These genes were previously identified as highly polymorphic (Jorrin & Imperial, 2015), and suspected to be involved in rhizobial host specificity (Djordjevic, Schofield, & Rolfe, 1985; Surin & Downie, 1988; Lewis-Henderson & Djordjevic, 1991; Baev et al., 1992). Rhizobia produce a large diversity of NFs that bind to legume root LysM-RLK receptors (Oldroyd, Murray, Poole, & Downie, 2011). NFs are composed of a chitin-like N-acetyl glucosamine backbone with a fatty acyl chain at the non-reducing end, and carry various substitutions such as glycosylation, acetylation and/or sulfation on the backbone (Mergaert, Montagu, & Holsters, 1997). These modifications influence the binding between NFs and LysM-RLKs and modulate the establishment of the symbiosis (Dénarié, Debellé, & Rosenberg, 1992). For instance, the *nodX* gene has been reported as crucial for the specific partner choice between the *Pisum sativum* cultivar 'Afghanistan', carrying the *SYM2* locus, and the *Rlv* strain TOM (Davis, Evans, & Johnston, 1988). The *nodX* gene encodes an acetyltransferase that modifies the NFs secreted by the bacteria, probably allowing them to bind with a specific LysM-RLK protein encoded within the *SYM2* locus (Firmin, Wilson, Carlson, Davies, & Downie, 1993; Hogg, Davies, Wilson, Bisseling, & Downie, 2002; Sulima et al., 2017). This *SYM2-nodX* association represents a well-documented example of a mechanism restricting AFN. Although genetic association of *nod* gene diversity with CFN argues for the interaction of NFs with LysM-RLK having an important role in CFN, the underlying mechanisms are still unknown. Other mechanisms, related to plant recognition of bacterial surface polysaccharides or bacterial effectors, have been implicated in modulation of the legume-rhizobium interaction and therefore may

contribute to CFN (Janczarek, Rachwał, Marzec, Grządziel, & Palusińska-Szysz, 2015; Miwa & Okazaki, 2017).

In most published co-inoculation experimental strategies, bacteria were inoculated with a reference strain or with a limited number of strains (Triplett & Sadowsky, 1992; Laguerre et al., 2003; Bourion et al., 2018). Recently, a co-inoculation strategy with multiple *Ensifer meliloti* strains was applied on two *Medicago truncatula* genotypes to investigate the impact of the partner choice diversity on symbiotic traits (Epstein et al., 2018). However, phenotyping was done five weeks post-inoculation in this study, which did not allow a focus on early symbiotic traits such as CFN that may potentially be compensated by post-infection mechanisms. Taking the new opportunities offered by both NGS and DNA metabarcoding, we designed a strategy to estimate bacterial CFN in Fabeae roots inoculated with *Rlv* populations, and to identify genes potentially associated with contrasted CFN phenotypes. We defined a core collection representative of the genomic diversity of the symbiovar *viciae*. We identified a *nodD* DNA barcode to discriminate and quantify each rhizobium of the core collection individually within a nodulated root system, using high throughput NGS. We inoculated the core collection in a mixture on 9 diverse genotypes of *Pisum sativum*, *Vicia faba* and *Lens culinaris* in order to compare their CFN with different Fabeae hosts. We used saturating amounts of each strain to focus on plant-microbe interactions and to reduce impacts of differential bacterial growth. We characterized host-specific CFN profiles that varied among bacteria of the core collection. Using a comparative genomic approach we identified bacterial genes associated with host-specific CFN in the various hosts.

Material and methods

Bacterial collection, inoculation and plant growth conditions

Bacteria from different geographical origins isolated from *Pisum sativum*, *Vicia faba*, *Lens culinaris* or *Lathyrus pratensis* root nodules were obtained (Table S1). *Pisum sativum, Vicia faba* and *Lens culinaris* seeds were surface sterilized in 3% calcium-hypochlorite solution for 10 min, washed 5 times in sterilized water, and sown in 2L pots filled with sterilized perlite/sand (3/1). Bacterial strains were grown individually in YEM broth medium. For each culture, the number of colony forming units (CFU) was estimated by dilution plating on YEM medium. The identity of the bacteria was confirmed by PCR amplification and sequencing of the *nodD* gene. Then, bacteria were mixed together in equal amounts (10^7 CFU/mL/strain) to establish the complex inoculum. We inoculated different plant cultivars of *Pisum sativum* (cultivars 'Kayanne', 'Isard' and 'Afghanistan' pea), *Lens culinaris* (cultivars 'Rosana', 'Anicia' and 'Flora') and *Vicia faba* (cultivars 'Diva', 'Organdi' and 'Tiffany'). Seeds were inoculated with the complex inoculum directly after sowing (3 pots of 4 seeds were used for each condition). Plants were grown under high-pressure sodium lamps with a mean photosynthetically active radiation of 250 µmol photons $m^{-2}.s^{-1}$ (16h/8h 22/18°C day/night cycle). They were supplied with N-free nutrient solution (K_2HPO_4 0.8mM, $MgSO_4$ 1 mM, K_2SO_4 0.6 mM, $CaCl_2$ 2 mM, NaCl 0.2 mM adjusted to pH 6.5) twice a week. Plants were harvested 14 days after inoculation for *Pisum sativum* and *Lens culinaris*, and 21 days after inoculation for *Vicia faba*.

DNA extraction, PCR amplification and metabarcoding analysis

The nodulated root systems of the 4 plants in each pot were pooled together to form one DNA metabarcoding biological replicate. Three separate biological replicates

were used to determine each rhizobial CFN profile (12 plants). Whole nodulated root systems were ground in liquid N_2 and DNA extractions were made using the DNeasy Plant Mini Kit (www.qiagen.com). PCR amplifications of the nodD309 barcode sequences were performed using Phusion High-Fidelity DNA Polymerase (www.thermofisher.com) and specific primers and conditions (Table S2). Size of amplicons was checked on agarose gels before sequencing using Illumina MiSeq technology with a 2x250bp paired end protocol, performed at the Genotoul GeT-PlaGe facility (get.genotoul.fr). Single multiplexing was performed using a homemade 6 bp index, which was added to R784 during a second PCR with 12 cycles using specific primers (Table S2). The resulting PCR products were purified and loaded onto the Illumina MiSeq cartridge according to the manufacturer's instructions. Paired Illumina MiSeq reads were assembled with vsearch v2.9.1 (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) using the command fastq_mergepairs and the option fastq_allowmergestagger. Demultiplexing and primer clipping were performed with cutadapt v1.9 (Martin, 2011) forcing a full-length match for sample tags and allowing a 2/3rd-length partial match for forward and reverse primers. Only reads containing both primers were retained. For each trimmed read, the expected error was estimated with vsearch's command fastq_filter and the option eeout. Each sample was then dereplicated (i.e. strictly identical reads were merged) using vsearch's command derep_fulllength, and converted to FASTA format. To prepare for clustering, the samples were pooled and processed with another round of dereplication. Files containing expected error estimates were also dereplicated to retain only the lowest expected error for each unique sequence. To detect potential contaminants, the dereplicated data were further clustered with swarm v2.1.9 (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015), and checked for chimeras using vsearch's command

uchime_denovo (Edgar, Haas, Clemente, Quince, & Knight, 2011). As no significant contamination was detected, downstream analyses and results are based on unclustered data, only retaining reads strictly identical to the 32 expected *Rhizobium* nodD309 reference sequences (Table S2), yielding a total of 511294 reads for all replicates. For each plant cultivar, the read number of the four replicates has been averaged, yielding a mean of 14203 reads per condition (Table S3). We calculated the CFN index (CFNi) of a bacterium in a plant host as the percentage of the nodD309 sequences of the rhizobium of interest to the total number of nodD309 sequences generated by all rhizobia of the sample. The CFN indexes (%) of each rhizobium of the core collection in plant genotypes were the mean of 4 biological repeats (Table S3).

Genome sequencing, genomic and association genetic analysis

Bacterial genomes were sequenced by MicrobesNG (Birmingham, UK, https://microbesng.uk/) on an Illumina HiSeq platform using a 2x250bp paired end protocol. Genomic DNA libraries were prepared using Nextera XT Library Prep Kit (Illumina, San Diego, USA). High-quality paired reads were assembled by the Galaxy/BBRIC pipeline (https://bbric-pipelines.toulouse.inra.fr/) and genome annotations were performed using EuGene-PP (Sallet et al, 2014) and RAST (http://rast.nmpdr.org/). The pairwise ANI values were calculated using the JSpecies software (http://jspecies.ribohost.com/jspeciesws). Heatmaps were built using the pheatmap R package (Kolde, 2015). Presence/absence of genes associated with CFN phenotypes were identified using the Bidirectional Best Hits (BBHs) tool available in RAST (http://rast.nmpdr.org). The presence/absence of the specific genes identified was checked by a BLAST search in the genomes of the core collection (thresholds: nucleotide identity > 70%; sequence coverage > 70%). Kruskal-Wallis tests (eventually followed by a Dunn post-hoc test with correction for multiple testing) were performed

using R software (http://www.R-project.org/.) to test whether the bacterial subgroups displayed equivalent CFN indexes (H_o hypothesis), according to diverse criteria (genospecies/Nod types/groups with specific allele). The nucleotide sequences of the *nodABCDEFIJLMN* genes were concatenated and aligned using ClustalOmega (www.ebi.ac.uk), and a Neighbor-Joining (NJ) phylogenetic tree was built using MEGA v7.0.26 (www.megasoftware.net). The comparisons between reference genomes, and between contigs and the genomic sequence of 3841, for genomic rearrangements/organization (Table S5 and Table S6) were performed using MAUVE software (Darling et al, 2004). Plasmid replicon classes (Rh groups) were identified by sequence similarity of the *repA* gene and presence of cognate *repB* and *repC* (Cavassim et al., 2020).

Results

Constitution of a bacterial core collection of Fabeae rhizobial symbionts

At the beginning of this study we collected the 73 genome sequences of Fabeae rhizobial symbiont available in GenBank. To maximize the diversity, the genomes of 48 additional rhizobia of more diverse geographical origins and/or carrying diverse sequences of the symbiotic marker *nodD* were sequenced (Table S1). They were isolated in previous studies from Fabeae root nodules (*Pisum sativum*, *Vicia faba*, *Lens culinaris* and *Lathyrus pratensis*) in Bangladesh, China, Algeria, Poland, Germany, France, United Kingdom and Italy. As expected most of these bacteria were described as belonging to *Rhizobium leguminosarum* symbiovar *viciae.* Based on the genome comparison and using an Average Nucleotide Identity (ANI) threshold of 95%, we concluded that 117 bacteria belong the *R. leguminosarum* complex species

(ANI>90%). The *R. leguminosarum* complex species is composed of at least 7 genospecies (gs; Figure 1). Four strains were phylogenetically distant from the others and therefore fell outside the *R. leguminosarum* species complex according the criterion of ANI<90%, even though they shared the ability to form root nodules with Fabeae with the other bacteria of the collection. They belonged to other described *Rhizobium* species: *R. binae* (BLR195), *R. pisi* (CZP3G4) and *R. anhuiense* (CCBAU43229), except for SEF4G12 which belonged to an undescribed species close to *R. acidisoli* FH23^{T} (ANIm=94.12%). The other 117 bacteria belonged to the *R. leguminosarum* genospecies (ANI>90%). Most of them are included in the 7 *R. leguminosarum* genospecies described previously (Kumar et al., 2015; Boivin et al., 2020; Cavassim et al., 2020; gsA; gsB; gsC; gsD; gsE, gsF-1 and gsF-2/*R. laguerreae*). A new *R. leguminosarum* genospecies, called gsG (Figure 1; Table S1), was distinguished in accordance with the recently established taxonomic guidance (de Lajudie et al., 2019). This genospecies was defined by grouping two different bacteria isolated in China (CCBAU33195 and CCBAU11080), forming a discrete cluster phylogenetically distant from all others (ANI<94%).

The 121 bacteria of the collection were also discriminated according to their plasmid-borne symbiosis genes. A phylogenetic tree was constructed using the sequence of the 11 conserved *nod* genes located on the symbiosis plasmid (Figure 2; Table S1). In agreement with the *nod* gene phylogeny, the *Rlv* bacteria were separated, independently of their genospecies, into 2 Nod types, named A and B, as previously described (Boivin et al., 2020). These two types were subdivided into 10 Nod groups named A1 to A5 and B1 to B5 respectively. There was considerable sequence variations within the Nod groups, except for the 22 strains belonging to the B1 group, which had few differences. A total of 38 gs/Nod group combinations were

identified within the 121 bacteria, confirming the diversity of the association between the symbiosis plasmid and the genospecies (Table S7). The number of bacteria per combination is variable. For instance, 19 strains carried the combination gsE/A1 whereas only one strain had gsB/A2 (Table S7). The *nodX* gene located within the *nod* gene cluster was present in 27/121 isolates (Figure 2).

We defined a DNA barcode to discriminate and quantify each bacteria in mixture. The 16S or *gyrB* marker sequences, frequently used in DNA metarbarcoding strategies, did not display sufficient intraspecific genetic variation among the 121 *Rlv* bacteria (data not shown). Furthermore, they belonged to the chromosome and were genetically unlinked to nodulation phenotypes. Alternatively, a 309bp fragment of the *nodD* gene located on the symbiosis plasmid allowed 32 *nodD309* alleles to be discriminated among the 121 *Rlv* (Table S2), and covered all Nod groups (Table S7). Based on our ability to discriminate the bacteria in a mixture, as well as criteria of geographical origins, plant hosts used for isolation, diversity of gs/Nod group combinations and presence/absence of *nodX*, 32 representative bacteria were selected to constitute a fabeae symbiont core collection (Table 1; Figure S1; Table S2). Most geographical origins (87%), plant hosts (100%) and gs/Nod group combinations (67%), and bacteria with *nodX* (12/32), were represented (Table S1 and Table S7). Nevertheless, it is noteworthy that, because of the limited variation of the barcode, we could introduce only one strain representing the B1 group.

Competitiveness to Form root Nodules (CFN) varies across host plants

The 32 *Rlv* bacteria of the core collection formed the complex inoculum. We combined saturating densities of each of the 32 bacteria (10^7 CFU/mL for each strain) to minimize the putative impact of bacterial growth on nodulation success. We

inoculated different plant genotypes of *Pisum sativum*, *Lens culinaris* and *Vicia faba* with the core collection. They covered a large genetic diversity of these legume crops. Root systems were harvested once nodules had formed and the nodD309 DNA barcode was PCR-amplified from total DNA. The sequencing of the PCR product yielded numbers of nodD309 sequences, specific to each member of the core collection. Among the 32 bacteria, 6 were removed for the further analyses as they were not detected in any sample from any host (Table S8). We calculated the CFN index (CFNi) for each bacterial strain in each host plant. The CFNi of the 26 bacteria detected in metabarcoding ranged from 0% to 81.9% (Table S8). The mean CFNi on each plant genotype was always around 3.8%, and the mean CFNi per bacterium in all hosts ranged from 0.003% to 32.2%.

A hierarchical clustering, based on the CFNi of the 26 bacteria of the core collection detected in metabarcoding, separated plant hosts clearly into 4 groups with distinct competitiveness profiles (Figure 3). These 4 groups were composed respectively of the *Vicia faba* genotypes, the *Lens culinaris* genotypes, the *Pisum sativum* cultivars 'Kayanne' and 'Isard' (cultivated spring and winter peas), and the *Pisum sativum* cultivar 'Afghanistan' (Figure 3). For bacteria, the clustering was less marked than for plants. However, 4 clusters may be separated according to host-specific CFN profiles of bacteria (Figure 3; Table S9).

NodX is required but not sufficient for nodulation of *Pisum sativum* cv. 'Afghanistan'

The nodulation profile of the *Pisum sativum* cultivar 'Afghanistan' was highly divergent compared to the other pea cultivars (Figure 3). The data confirmed that the bacterial *nodX* gene is required for nodulation with this plant genotype. All competitive

317 strains with this pea genotype, representing 89% of the *Rlv* barcode of the root
318 (BLR195, CCBAU11080 and CCBAU10279), carried the *nodX* gene (Figure 3). These
319 bacteria were not competitive with other pea cultivars. However, when the whole set
320 of bacteria was considered, neither the presence of *nodX*, nor the Nod type, nor the
321 genospecies were solely associated with nodulation success on this host (Table S10
322 and Table S4). Although the 4 most competitive isolates had the *nodX* gene, 8 other
323 isolates of the core collection, belonging to Nod type A or B, also had the *nodX* gene,
324 but were not competitive with this cultivar (Table 1 and Table S4). A significant
325 association with CFN was identified only with bacteria having both the *nodX* gene and
326 Nod type B (Table S4). Mono-inoculation experiments with several strains carrying
327 *nodX* indicated that only strains of Nod type B carrying this gene were able to form
328 nodules with the pea cultivar 'Afghanistan', whereas the strains of Nod type A were
329 unable (Table S11). We concluded that the *nodX* gene was necessary but not sufficient
330 to confer the ability to nodulate this host genotype and genetic determinants related to
331 Nod type B were also required. Nevertheless, there was a wide range of CFN within
332 the Nod type B strains carrying *nodX* and thus able to nodulate the cultivar
333 'Afghanistan'. We concluded that genetic factors independent of NodX, present in Nod
334 type B, were responsible for CFN variations in the context of host-specific ability to
335 form nodule with this cultivar.

336 **Host-specific CFN factors are associated with *Rlv* plasmids and/or genospecies**

337 Beyond the particular case of pea cultivar 'Afghanistan', the three other groups
338 of pea, lentil and fababean cultivars displayed contrasted CFN profiles. The *Rlv* strain
339 3841 was predominant in the three *Vicia faba* genotypes (from 70% to 82% of the total
340 *Rlv* reads in fababean roots), but not at the same level in pea or lentil genotypes. The
341 *Rlv* strain GLR2 was found in nodulated roots of the three *Lens culinaris* genotypes,

but not in other plant species (Figure 3 and Table S8). Behaviors did not depend only on the plant species, and the CFN profiles also varied between genotypes within the three plant species. For example, *Rlv* strains 3841 and L113 were differentially competitive with the pea cultivars ‘Kayanne’ and ‘Isard’ (Figure 3). Concerning the CFN in *Vicia faba* genotypes, the predominance of 3841 confirmed our previous study indicating that the Nod type B (and especially the Nod group B1) is strongly associated with CFN on *Vicia faba*. Unfortunately, as 3841 is the only strain of the core collection belonging to the Nod group B1, it was not possible to extend the comparison to other B1 strains in order, potentially, to associate other genetic variation with CFN. Consequently, further investigations were restricted to pea cultivar ‘Kayanne’ and ‘Isard’ as well as to the three lentil cultivars.

Many bacteria of the core collection displayed a wide range of CFN variation with these hosts. We investigated CFN associations with the Nod type, the Nod group, or the genospecies of bacteria at two levels of plant diversity: the plant cultivar (Table S10) and the plant species (Table S12). Because the core collection had a limited number of bacteria for each category (Nod type, Nod group, and genospecies), and it was tested with a limited number of plant genotypes, the power of the statistical analysis was limited. However, the data showed global effects of bacterial genotype variation on CFN with lentil and/or pea at the level of the cultivar (Table S10). The Nod type was associated with the CFN phenotype in the two pea cultivars ‘Isard’ and ‘Kayanne’, but not in lentil. An association with the Nod group was found with ‘Isard’. The genospecies was significantly associated with the CFN phenotype in the *Pisum sativum* cultivar ‘Isard’ as well as in lentil cultivars ‘Anicia’ and ‘Flora’.

Investigations at the level of the plant species, combining the data obtained on the various cultivars, were more powerful. At this level, it was possible to test more

efficiently the global effect of the bacterial genotype variation, but also to identify by pairwise testing which category of the core collection may explain the variation (Table S12). Nod type A bacteria generally displayed generally higher CFN in the *Pisum sativum* cultivars ‘Kayanne’ and ‘Isard’. Bacteria of the core collection belonging to the Nod groups A1 and A4 had a higher mean CFN than those belonging to Nod groups B3, B4 and B5 (Table S12). The only exception is 3841, the unique representative of the B1 Nod group in the core collection, which displayed high CFN with pea. This particular result was somehow in contradiction with our previous observations indicating that natural isolates of this Nod group are generally good competitor with fababean but poorly competitive with pea (Boivin et al., 2020). Although CFNi of bacteria of the core collection belonging to various Nod groups varied widely in lentil, we got no evidence indicating that the Nod type or the Nod group could predict competitiveness with this plant host.

We found associations of genospecies with CFN in pea (particularly in the ‘Isard’ cultivar) and lentil. Some of them are common to both host species. For instance, bacteria of the core collection belonging to gsB often displayed high competiveness in pea and lentil whereas the bacteria of gsF-1 always performed poorly with these hosts. Nevertheless, there were also host-specific examples. For example, bacteria belonging to gsF-2 were in the medium range of competitiveness in lentil but poor competitors in pea. Despite these global trends, associations were never systematic and exceptions were observed. Genomic investigations will be required to identify which parts of the *Rlv* genomes are responsible for these CFN variations.

Contrasted CFN profiles are associated with the presence of specific bacterial genes

Comparative genomics identified bacterial genetic factors associated with the contrasted nodulation profiles. For each plant genotype, we compared the 4-5 most competitive bacteria to the 4-5 least competitive to identify genes specifically present or absent. Then, we tested whether, in the whole set of 26 bacteria that were detected in the root systems, the presence/absence of the identified genes was significantly associated with the host-specific nodulation phenotype (Table S4). Using this strategy, the number of specific genes identified varied from 2 to 13 depending on the plant species and cultivar (Table S4).

In the particular case of the *Pisum sativum* cultivar 'Afghanistan', only a fraction of strains were able to form nodules with this host. Our first aim was to identify the genetic factors that, together with NodX, allow this host-specific nodulation. We restricted the first genomic comparison to the 8 bacteria of the core collection carrying the *nodX* gene and either Nod type B (these strains could nodulate this host) or Nod type A (these could not). Seventeen genes associated with nodulation of this host were identified by this comparison (Table S13 and Table S14). Our second aim was to identify genetic factors specifically associated with CFN in this particular host by comparing only bacteria able to form nodules with this host, i.e., the 5 bacteria of Nod type B with the *nodX* gene. These displayed contrasted CFN with their host (see above), and genomic comparisons identified a further 6 genes (Table 2; Table S4 and Table S14).

Similar genome comparisons were made for bacteria that displayed contrasted CFN in association with the *Pisum sativum* cultivars 'Kayanne' and 'Isard, as well as with lentil cultivars 'Anicia' and 'Rosana'. This analysis yielded 36 genes potentially involved in the competitive success to form nodules of the core collection bacteria in mixture with pea (except cv 'Afghanistan') and lentil hosts (Table 2 and Table S14).

Homologs of these genes were sought in the 6 fully-assembled *Rlv* genomes described in Genbank: not only the 3841 reference strain but also Vaf10, Vaf108, BIHB1217, UPM791 and TOM. The replication and maintenance of plasmids in *R.leguminosarum* are controlled by the *repABC* genes, and the sequences of these genes fall into distinct clusters that allows 'Rh' incompatibility groups to be defined (Cavassim et al., 2020). The 'Rh' types of the replicons of the *Rlv* strains were characterized, in order to define series of homologous replicons (Table 2, Table S5, Table S6 and Table S14). However, we found plenty of evidence for variations and frequent rearrangements between replicons in these genomes. Most of the Vaf10, Vaf108, BIHB1217, UPM791 and TOM plasmids shared homologies with more than one 3841 plasmid, revealing that replicons (and particularly the distribution of sequences among replicons) vary greatly among the various *Rlv* bacteria (Table S5). For example, in strain BIHB1217, plasmid pPR4 (Rh08) shared sequence homologies with the pRL7 (Rh12), pRL8 (Rh13) and pRL10 (Rh03) plasmids of strain 3841. In the Vaf-108 genome, the chromosome shared homologies with both the 3841 chromosome and the pRL10 plasmid, indicating that rearrangements were probably not restricted to plasmids. This fluidity of the accessory genome was very apparent when we examined the location of the genes that showed significant associations with CFN or AFN (Table 2, Table S14). For example, the genes PI9 and PI10 are always adjacent in the fully assembled genomes, but they are on an Rh01 plasmid in Vaf-108, Rh03 in 3841, and Rh08 in Vaf-10, BIHB1217 and UPM791. In the genomes of BIHB1217 and UPM791, PI12 is on Rh02 and PI13 is on Rh04a, but in 3841 these genes are both on Rh05, just 2.5kb apart. The genes PAFN8-10 and PAFN17 are all on Rh12 in 3841, Vaf-10 and UPM791, but on Rh01 in Vaf-108, and Rh06 in BIHB1217. In TOM, PAFN8-10 are on Rh06, but PAFN17 is on Rh03. These examples

of genome rearrangement, and numerous others, indicated that the replicon location of a gene could not be reliably inferred from its location in another strain.

Discussion

A metabarcoding approach to measure competitiveness to form nodules

Selection pressures related to plant-microbe, microbe-microbe and/or microbe-environment interactions are likely important drivers of rhizobium populations in interaction with their hosts. Previous studies estimated the CFN of *Rlv* isolates individually by co-inoculating a host plant with both a strain of interest and a reference strain (Bourion et al, 2018; Boivin et al., 2020). Selecting antibiotic resistant mutants (Amarger 1981) or introducing fluorescent markers on plasmids has greatly improved the recognition of bacteria (Melkonian et al 2014, Westhoek et al, 2017) but could modify CFN of the bacteria of interest. Nevertheless, these techniques minimize the effect of multiple interactions. Construction of bacterial populations in which each individual can be quantified allows the study of complex interactions that cannot be revealed when single individuals are considered separately (Epstein et al., 2018; Paredes et al., 2018; Carlström et al., 2019; Boivin and Lepetit, 2020). The use of the DNA metabarcoding strategy allowed evaluating CFN after co-inoculation of roots with large populations of diverse unmodified rhizobia and therefore solved this problem. However bacterial competitiveness to form root nodules has generally been estimated by the number of nodules formed with the bacteria of interest as compared to total nodule number on the plant root system. This is consistent with the general hypothesis that CFN is a important symbiotic trait for early partner choice but not for 'post-infection nodule expansion' that mainly involves other preference mechanisms controlled by the

plant (Boivin & Lepetit, 2020). Evaluating CFN by DNA metabarcoding approach quantifies bacteria interacting with the plant and therefore may be biaised as result of differential nodule growth or differential levels of bacteroid endoreduplication of the various bacteria (Kazmierczak et al., 2017). To circumvent this potential bias, we applied DNA metabarcoding on roots at an early stage of interaction before these processes being prominent.

A representative core collection of Fabeae symbiont

Sequencing 48 additional genomes of bacteria isolated from Fabeae root nodules of various geographical origins allowed us to characterize and extend our current knowledge of the Fabeae symbiovar diversity, augmenting the 73 *Rlv* genome sequences available in Genbank at the beginning of the study. It allowed new *Rlv* genospecies to be defined based on established criteria for bacteria in general and *Rhizobium* in particular (Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018; de Lajudie et al., 2019). All the bacteria shared closely related *nod* gene sequences typical of the symbiovar *viciae*. However, several isolates were genetically distant from the main genospecies in the *R. leguminosarum* species complex (ANI<90%), raising the question of the appropriate boundary of this species complex, which has not yet been clearly defined. A set of 32 bacteria representative of the known genomic diversity of Fabeae symbiont was selected to study host-specific dynamics of symbiotic interaction. Individuals within the bacterial mixture have been discriminated and quantified by a DNA barcode located in the *nodD* gene of the symbiosis plasmid. This *nodD* metabarcoding enabled us to explore much of the intraspecific variability of the symbiovar. Nevertheless, for some sequence variants, the barcode has a limit. For example, *Rlv* strains carrying the Nod type B1 shared an identical barcode, so only one could be included in the competition study, but the strains displayed divergent

genome sequences that may potentially be associated with variation of the bacterial phenotype. The design and the high-throughput sequencing of a longer barcode will improve the resolution of the technique in the near future.

Co-inoculation of the core collection of 32 bacteria, in mixture, on the *Pisum sativum*, *Vicia faba* and *Lens culinaris* genotypes revealed host-specific CFN profiles. They varied according to the bacterium, the plant species and the plant genotype, and were probably related to the capacities of the various plant-microbe partnerships. These specificities were unlikely to be due to typical microbe-microbe interactions, generally resulting in differential bacterial multiplication and therefore limiting interaction with the plant, because bacteria were amplified separately and were present at high density in the inoculum (>10^7 CFU/mL). This study allowed us to identify the most competitive bacteria of the *Rlv* core collection with *Vicia faba*, *Pisum sativum* and *Lens culinaris* genotypes. Nevertheless, this study was done on plants cultivated in the same standard substrate and the question of the effect of soil environment on CFN has not been addressed here.

The *nodX* gene is necessary but not sufficient to confer host-specific nodulation of *Rlv* with *Pisum sativum* cultivar 'Afghanistan'

Our data confirmed earlier work indicating that only *Rlv* bacteria carrying the *nodX* gene can nodulate the pea cultivar 'Afghanistan', resulting in a highly specific association profile (Davis et al., 1988; Firmin et al., 1993). This study shed light on unexpected complexities of the NodX/Sym2 interaction. The *nodX* gene was surprisingly found in a large portion of the *Rlv* genomes (22%), including those of both Nod types A and B. The corresponding bacteria were isolated from different Fabeae host plants and have multiple geographical origins around the world, including some

514 locations where relatives of cv. ‘Afghanistan’ are probably not present. The *nodX* gene
515 was not sufficient to allow *Rlv* to form root nodules with this cultivar, and we obtained
516 evidence indicating that other genetic determinants, present only in the Nod type B
517 strains, were required. We suggest that the host-specific nodulation of cv. ‘Afghanistan’
518 or relatives may be not the unique function of NodX. Another unknown function may
519 explain the conservation of this gene within natural *Rlv* populations unable to associate
520 with such hosts. Interestingly, this gene has been also found in *R. leguminosarum*
521 symbiovar *trifolii* strains, also unable to nodulate *P. sativum* cv. ‘Afghanistan’ (Ovtsyna
522 et al., 1999). The control of host-specific nodulation might be more complex, and the
523 function of NodX deserves to be revisited. Nevertheless, NodX alone did not fully
524 explain the specificity of the nodulation profile of cv. ‘Afghanistan’, co-inoculated with
525 the core collection. The 5 bacteria of Nod type B that had the *nodX* gene, though they
526 all had a good capacity to form root nodules with cv. ‘Afghanistan’ in mono-inoculation
527 assays, displayed a wide range of contrasted CFN phenotypes with this host when co-
528 inoculated in mixture. A striking example is the TOM strain, the well-studied reference
529 symbiont of cv. ‘Afghanistan’, which is outcompeted by more competitive Nod type B
530 bacteria carrying *nodX gene*. We identified 6 genomic regions associated with variation
531 of CFN with this host in Nod type B strains with NodX.

532 **Host-specific CFN is associated with different genomic regions depending on**
533 **the host plant and the genetic background of bacteria**

534 We confirmed that pea and fababean differentially select rhizobia depending on
535 the variation of the Nod type/group (Boivin et al., 2020). Host-specific CFN was
536 associated with the Nod types A and B in *Pisum sativum* (except cv. ‘Afghanistan’) and
537 *Vicia faba* cultivars respectively. In contrast, the CFN with lentil was strongly
538 associated with the genospecies, but poorly with the Nod type or Nod group,

539 suggesting that different genetic determinants, possibly located mainly on the
540 chromosome, controlled CFN with this host. Despite these global trends, the genetic
541 control of CFN is likely to be complex as, for all species, several data suggested effects
542 of both Nod groups and genospecies, possibly in interaction. An unexpected result was
543 the observation of a high CFN with pea of 3841, the unique bacterium of Nod group
544 B1 present in the core collection. Our previous investigation on natural isolates using
545 CFN measurements with a reference strain indicated that bacteria of this group B1
546 were generally competitive with fababean but poorly competitive with pea (Boivin et
547 al., 2020). This different behavior of the strain 3841 deserves further investigation with
548 particular emphasis to the association with the pea CFN phenotype. Because the strain
549 3841, initially isolated from pea root nodules, was cultured in the laboratories for many
550 years, it cannot be ruled out that this particular phenotype may be related to
551 spontaneous unknown mutations in the laboratory, in absence of natural selection.

552 The strategy of comparing genomes between competitive and uncompetitive
553 strains allowed us to identify, for each plant genotype, regions of the bacterial genomes
554 associated with host-specific CFN. The assignment of the identified genomic regions
555 to a particular replicon deserves further study. As the knowledge of *R. leguminosarum*
556 genomes is expanding, increasing evidence indicates that plasmid number, size and
557 composition vary greatly in this bacterial species (Mazurier et al., 1997; Laguerre et
558 al., 2003; Kumar et al., 2015; Cavassim et al., 2020). Different bacteria belonging to
559 the same symbiovar share homologous sequences, organized differently in various
560 diverse replicons. There is evidence that genetic rearrangements may occur even
561 between plasmids and chromosome. There is an emerging picture of a high fluidity
562 between *R. leguminosarum* replicons that argues against attempts to assign genomic
563 regions to replicons by homology with a reference genome. Interestingly earlier works

564 have already suggested that large genetic rearrangements between replicons could
565 modify the genetic environment around the *nod* gene cluster, and create and/or disrupt
566 genetic links (Mazurier & Laguerre, 1997; Zhang, Kosier, & Priefer, 2001). This fluidity
567 of the accessory genome is essential for the gene association approach that we have
568 taken in this study, which requires a certain degree of independence between loci in
569 order to associate phenotypes with individual genes rather than with whole replicons
570 or large blocks of coinherited genes.

571 As comparative studies yielded multiple potential genomic regions associated
572 with CFN, it is tempting to hypothesize that multiple genes control this trait. A significant
573 fraction of the protein sequences associated with CFN has known functions, whereas
574 others are only hypothetical (Table 2 and Table S14). Known proteins were involved
575 in amino acid modification (amidinotransferase, homoserine O-acetyltransferase),
576 amino acid transport (L-proline/glycine betaine transporter), nucleic acid
577 repair/modification (DNA/RNA helicase, excinuclease), bacteroid aerotolerance (bat
578 operon), rhizobial nod factor secretion (NodT), or toxin/antitoxin systems (VapC,
579 RelE/StbE). Interestingly, the *nodT* gene has already been associated with CFN in pea
580 and fababean (Boivin et al., 2020). Members of Vap-type toxin/antitoxin system and
581 genes involved in amino acid modifications (aminotransferase) were also associated
582 with CFN in pea and fababean (Boivin et al., 2020) but, as they belong to multigenic
583 families, we cannot confirm that genes identified in both studies are true orthologs.
584 Further investigations are needed to specify this point. Globally, it would be too
585 speculative at this stage to associate these specific functions with CFN. Although the
586 identified genes were statistically associated with the host-specific CFN phenotypes,
587 these associations may be indirect and related to neighboring genetic sequences.
588 Reverse genetics studies with different combinations of alleles will be required to

validate their potential biological role in host-specific CFN in *Rlv*. Nevertheless, these sequences are valuable markers to select competitive *Rlv* strains with pea and/or lentil and they will allow the identification of genes controlling CFN. This knowledge opens new perspectives to select bacteria and plants, by genetic association, for new inoculation strategies that will ultimately improve the agro-ecological services of Fabeae legume holobionts.

Acknowledgements

We acknowledge Markus Braun (University of Heidelberg), Xinhua Sui and Changfu Tian (China Agricultural University, Beijing), Anna Skorupska and Andrzej Mazur (University of Lublin), and Juan Imperial (Instituto de Ciencias Agrarias, Madrid) for providing some of the rhizobial strains included in this study as well as the Genotoul GeT-PlaGe platform of Toulouse for the Illumina MiSeq sequencing. This work was supported by the GRaSP grant of the Agence Nationale pour la Recherche (ANR).

References

Amarger, N. (1981). Competition for nodule formation between effective and ineffective strains of Rhizobium meliloti. Soil Biology and Biochemistry, 13(6), 475–480. doi: 10.1016/0038-0717(81)90037-7

Baev, N., Schultze, M., Barlier, I., Ha, D. C., Virelizier, H., Kondorosi, E., & Kondorosi, A. (1992). Rhizobium nodM and nodN genes are common nod genes: nodM encodes functions for efficiency of nod signal production and bacteroid maturation. *Journal of Bacteriology*, *174*(23), 7555–7565. doi: 10.1128/jb.174.23.7555-7565.1992

Bailly, X., Giuntini, E., Sexton, M. C., Lower, R. P., Harrison, P. W., Kumar, N., & Young, J. P. W. (2011). Population genomics of Sinorhizobium medicae based on low-coverage sequencing of sympatric isolates. *The ISME Journal*, *5*(11), 1722–1734. doi: 10.1038/ismej.2011.55

Bourion, V., Heulin-Gotty, K., Aubert, V., Tisseyre, P., Chabert-Martinello, M., Pervent, M., … Lepetit, M. (2018). Co-inoculation of a Pea Core collection with Diverse Rhizobial Strains Shows Competitiveness for Nodulation and Efficiency of Nitrogen Fixation Are Distinct traits in the Interaction. *Frontiers in Plant Science*, *8*. doi: 10.3389/fpls.2017.02249

Boivin, S., Ait Lahmidi, N., Sherlock, D., Bonhomme, M., Dijon, D., Heulin-Gotty, K., … Lepetit, M. (2020). Host-specific competitiveness to form nodules in Rhizobium leguminosarum symbiovar viciae. The New Phytologist, 226(2), 555–568. doi: 10.1111/nph.16392

Boivin, S., & Lepetit, M. (2020). Partner preference in the legume-rhizobia symbiosis and impact on legume inoculation strategies. In P. Frendo, F. Frugier, & C. Masson-Boivin (Eds.), Advances in Botanical Research (pp. 323–348). doi: 10.1016/bs.abr.2019.09.016

Carlström, C. I., Field, C. M., Bortfeld-Miller, M., Müller, B., Sunagawa, S., & Vorholt, J. A. (2019). Synthetic microbiota reveal priority effects and keystone strains in the Arabidopsis phyllosphere. *Nature Ecology & Evolution*, *3*(10), 1445–1454. doi: 10.1038/s41559-019-0994-z

Cavassim, M.I.A., Moeskjær, S., Moslemi, C., Fields, B., Bachmann, A., Vilhjálmsson, B.J., Schierup, M.H., Young, J.P.W. and Andersen, S.U. (2020). Symbiosis genes show a unique pattern of introgression and selection within a *Rhizobium leguminosarum* species complex. *Microbial Genomics,* 6(4), p.e000351. doi : 10.1099/mgen.0.000351

Darling A. C.E., Mau B., Blattner F. R., & Perna N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Research, 14(7), 1394-1403.

Davis, E. O., Evans, I. J., & Johnston, A. W. B. (1988). Identification of nodX, a gene that allows Rhizobium leguminosarum biovar viciae strain TOM to nodulate Afghanistan peas. *Molecular and General Genetics MGG*, *212*(3), 531–535. doi: 10.1007/BF00330860

de Lajudie, P. M., Andrews, M., Ardley, J., Eardly, B., Jumas-Bilak, E., Kuzmanović, N., … Young, P. (2019). Minimal standards for the description of new genera and species of rhizobia and agrobacteria. *International Journal of Systematic and Evolutionary Microbiology*, *69*(7), 1852–1863. doi: 10.1099/ijsem.0.003426

Dénarié, J., Debellé, F., & Rosenberg, C. (1992). Signaling and host range variation in nodulation. *Annual Review of Microbiology*, *46*, 497–531. doi: 10.1146/annurev.mi.46.100192.002433

642 Djordjevic, M. A., Schofield, P. R., & Rolfe, B. G. (1985). Tn5 mutagenesis of Rhizobium trifolii host-
643 specific nodulation genes result in mutants with altered host-range ability. *Molecular and*
644 *General Genetics MGG*, *200*(3), 463–471. doi: 10.1007/BF00425732
645 Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity
646 and speed of chimera detection. *Bioinformatics (Oxford, England)*, *27*(16), 2194–2200. doi:
647 10.1093/bioinformatics/btr381
648 Epstein, B., Abou-Shanab, R. A. I., Shamseldin, A., Taylor, M. R., Guhlin, J., Burghardt, L. T., … Tiffin,
649 P. (2018). Genome-Wide Association Analyses in the Model Rhizobium Ensifer meliloti.
650 *MSphere*, *3*(5). doi: 10.1128/mSphere.00386-18
651 Fesenko, A. N., Provorov, N. A., Orlova, I. F., Orlov, V. P., & Simarov, B. V. (1995). Selection of
652 Rhizobium leguminosarum bv. viceae strains for inoculation of Pisum sativum L. cultivars:
653 Analysis of symbiotic efficiency and nodulation competitiveness. *Plant and Soil*, *172*(2), 189–
654 198. doi: 10.1007/BF00011321
655 Firmin, J. L., Wilson, K. E., Carlson, R. W., Davies, A. E., & Downie, J. A. (1993). Resistance to
656 nodulation of cv. Afghanistan peas is overcome by nodX, which mediates an O-acetylation of
657 the Rhizobium leguminosarum lipo-oligosaccharide nodulation factor. *Molecular Microbiology*,
658 *10*(2), 351–360. doi: 10.1111/j.1365-2958.1993.tb01961.x
659 Hogg, B., Davies, A. E., Wilson, K. E., Bisseling, T., & Downie, J. A. (2002). Competitive Nodulation
660 Blocking of cv. Afghanistan Pea Is Related to High Levels of Nodulation Factors Made by Some
661 Strains of *Rhizobium leguminosarum* bv. *viciae*. *Molecular Plant-Microbe Interactions*, *15*(1),
662 60–68. doi: 10.1094/MPMI.2002.15.1.60
663 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput
664 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature*
665 *Communications*, *9*(1), 5114. doi: 10.1038/s41467-018-07641-9
666 Janczarek, M., Rachwał, K., Marzec, A., Grządziel, J., & Palusińska-Szysz, M. (2015). Signal molecules
667 and cell-surface components involved in early stages of the legume-rhizobium interactions.
668 *Applied Soil Ecology*, *85*, 94–113.
669 Jorrin, B., & Imperial, J. (2015). Population Genomics Analysis of Legume Host Preference for Specific
670 Rhizobial Genotypes in the Rhizobium leguminosarum bv. viciae Symbioses. *Molecular Plant-*
671 *Microbe Interactions: MPMI*, *28*(3), 310–318. doi: 10.1094/MPMI-09-14-0296-FI
672 Kazmierczak, T., Nagymihály, M., Lamouche, F., Barrière, Q., Guefrachi, I., Alunni, B., … Gruber, V.
673 (2017). Specific Host-Responsive Associations Between Medicago truncatula Accessions and
674 Sinorhizobium Strains. *Molecular Plant-Microbe Interactions: MPMI*, *30*(5), 399–409. doi:
675 10.1094/MPMI-01-17-0009-R
676 Kumar, N., Lad, G., Giuntini, E., Kaye, M. E., Udomwong, P., Shamsani, N. J., … Bailly, X. (2015).
677 Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium
678 leguminosarum. *Open Biology*, *5*(1), 140133. doi: 10.1098/rsob.140133
679 Laguerre, G., Louvrier, P., Allard, M.-R., & Amarger, N. (2003). Compatibility of rhizobial genotypes
680 within natural populations of Rhizobium leguminosarum biovar viciae for nodulation of host
681 legumes. *Applied and Environmental Microbiology*, *69*(4), 2276–2283.

682 Laguerre, G., Depret, G., Bourion, V., & Duc, G. (2007). Rhizobium leguminosarum bv. viciae genotypes
683 interact with pea plants in developmental responses of nodules, roots and shoots. The New
684 Phytologist, 176(3), 680–690. doi: 10.1111/j.1469-8137.2007.02212.x
685 Lewis-Henderson, W. R., & Djordjevic, M. A. (1991). nodT, a positively-acting cultivar specificity
686 determinant controlling nodulation of Trifolium subterraneum by Rhizobium leguminosarum
687 biovar trifolii. *Plant Molecular Biology*, *16*(4), 515–526.
688 Lipuma, J., Cinege, G., Bodogai, M., Oláh, B., Kiers, A., Endre, G., … Dusha, I. (2014). A vapBC-type
689 toxin-antitoxin module of Sinorhizobium meliloti influences symbiotic efficiency and nodule
690 senescence of Medicago sativa. *Environmental Microbiology*, *16*(12), 3714–3729. doi:
691 10.1111/1462-2920.12608
692 Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and
693 high-resolution amplicon clustering. *PeerJ*, *3*. doi: 10.7717/peerj.1420
694 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.
695 *EMBnet.Journal*, *17*(1), 10–12. doi: 10.14806/ej.17.1.200
696 Mazurier, S.-I., & Laguerre, G. (1997). Unusual localization of nod and nif genes in Rhizobium
697 leguminosarum bv. viciae. *Canadian Journal of Microbiology*, *43*(4), 399–402. doi:
698 10.1139/m97-056
699 McAnulla, C., Edwards, A., Sanchez-Contreras, M., Sawers, R. G., & Downie, J. A. (2007). Quorum-
700 sensing-regulated transcriptional initiation of plasmid transfer and replication genes in
701 Rhizobium leguminosarum biovar viciae. *Microbiology (Reading, England)*, *153*(Pt 7), 2074–
702 2082. doi: 10.1099/mic.0.2007/007153-0
703 McKenzie, R. H., Middleton, A. B., Solberg, E. D., DeMulder, J., Flore, N., Clayton, G. W., & Bremer, E.
704 (2001). Response of pea to rhizobia inoculation and starter nitrogen in Alberta. *Canadian*
705 *Journal of Plant Science*, *81*(4), 637–643. doi: 10.4141/P01-006
706 Mergaert, P., Montagu, M. V., & Holsters, M. (1997). Molecular mechanisms of Nod factor diversity.
707 *Molecular Microbiology*, *25*(5), 811–817. doi: 10.1111/j.1365-2958.1997.mmi526.x
708 Melkonian, R., Moulin, L., Béna, G., Tisseyre, P., Chaintreuil, C., Heulin, K., … Laguerre, G. (2014). The
709 geographical patterns of symbiont diversity in the invasive legume Mimosa pudica can be
710 explained by the competitiveness of its symbionts and by the host genotype. Environmental
711 Microbiology, 16(7), 2099–2111. doi: 10.1111/1462-2920.12286
712 Miwa, H., & Okazaki, S. (2017). How effectors promote beneficial interactions. *Current Opinion in Plant*
713 *Biology*, *38*, 148–154. doi: 10.1016/j.pbi.2017.05.011
714 Moawad, H. A., Ellis, W. R., & Schmidt, E. L. (1984). Rhizosphere Response as a Factor in Competition
715 Among Three Serogroups of Indigenous Rhizobium japonicum for Nodulation of Field-Grown
716 Soybeans. *Applied and Environmental Microbiology*, *47*(4), 607–612.
717 Naamala, J., Jaiswal, S. K., & Dakora, F. D. (2016). Antibiotics Resistance in Rhizobium: Type, Process,
718 Mechanism and Benefit for Agriculture. *Current Microbiology*, *72*(6), 804–816. doi:
719 10.1007/s00284-016-1005-0

Oldroyd, G. E. D., Murray, J. D., Poole, P. S., & Downie, J. A. (2011). The rules of engagement in the legume-rhizobial symbiosis. *Annual Review of Genetics*, *45*, 119–144. doi: 10.1146/annurev-genet-110410-132549

Ovtsyna, A. O., Rademaker, G.-J., Esser, E., Weinman, J., Rolfe, B. G., Tikhonovich, I. A., … Spaink, H. P. (1999). Comparison of Characteristics of the nodX Genes from Various Rhizobium leguminosarum Strains. *Molecular Plant-Microbe Interactions*, *12*(3), 252–258. doi: 10.1094/MPMI.1999.12.3.252

Paredes, S. H., Gao, T., Law, T. F., Finkel, O. M., Mucyn, T., Teixeira, P. J. P. L., … Castrillo, G. (2018). Design of synthetic bacterial communities for predictable plant phenotypes. *PLOS Biology*, *16*(2), e2003962. doi: 10.1371/journal.pbio.2003962

Robleto, E. A., Scupham, A. J., & Triplett, E. W. (1997). Trifolitoxin Production in Rhizobium etli Strain CE3 Increases Competitiveness for Rhizosphere Colonization and Root Nodulation of Phaseolus vulgaris in Soil. *Molecular Plant-Microbe Interactions*, *10*(2), 228–233. doi: 10.1094/MPMI.1997.10.2.228

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. doi: 10.7717/peerj.2584

Sallet E, Gouzy J, Schiex T. (2014). EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics. Oxford, England.* 30: 2659–2661.

Sulima, A. S., Zhukov, V. A., Afonin, A. A., Zhernakov, A. I., Tikhonovich, I. A., & Lutova, L. A. (2017). Selection Signatures in the First Exon of Paralogous Receptor Kinase Genes from the Sym2 Region of the Pisum sativum L. Genome. *Frontiers in Plant Science*, *8*. doi: 10.3389/fpls.2017.01957

Surin, B. P., & Downie, J. A. (1988). Characterization of the Rhizobium leguminosarum genes nodLMN involved in efficient host-specific nodulation. *Molecular Microbiology*, *2*(2), 173–183. doi: 10.1111/j.1365-2958.1988.tb00019.x

Triplett, E. W., & Sadowsky, M. J. (1992). Genetics of competition for nodulation of legumes. *Annual Review of Microbiology*, *46*, 399–428. doi: 10.1146/annurev.mi.46.100192.002151

Westhoek, A., Field, E., Rehling, F., Mulley, G., Webb, I., Poole, P. S., & Turnbull, L. A. (2017). Policing the legume-Rhizobium symbiosis: a critical test of partner choice. Scientific Reports, 7(1), 1419. doi: 10.1038/s41598-017-01634-2

Yost, C. K., Del Bel, K. L., Quandt, J., & Hynes, M. F. (2004). Rhizobium leguminosarum methyl-accepting chemotaxis protein genes are down-regulated in the pea nodule. *Archives of Microbiology*, *182*(6), 505–513. doi: 10.1007/s00203-004-0736-7

Young, J. P. W. (2016). Bacteria Are Smartphones and Mobile Genes Are Apps. *Trends in Microbiology*, *24*(12), 931–932. doi: 10.1016/j.tim.2016.09.002

Zézé, A., Mutch, L. A., & Young, J. P. (2001). Direct amplification of nodD from community DNA reveals the genetic diversity of Rhizobium leguminosarum in soil. *Environmental Microbiology*, *3*(6), 363–370.

Zhang, X.-X., Kosier, B., & Priefer, U. B. (2001). Symbiotic Plasmid Rearrangement in Rhizobium leguminosarum bv. viciae VF39SM. *Journal of Bacteriology*, *183*(6), 2141–2144. doi: 10.1128/JB.183.6.2141-2144.2001

Data Accessibility

Rhizobial genome sequences: GenBank assembly accessions (see Table S1)

Author Contributions

ML and SB designed the research. ML, SB, PY, JW and SM provided the rhizobial strains. SB, MP, MatT, FM and MarT performed research. FM performed analysis of metabarcoding data. SB, ML, FM and PY analyzed data. SB and ML wrote the paper with the contribution of PY.

Tables:

Table 1: *Rhizobium leguminosarum* symbiovar *viciae* bacteria included in the core collection

Strains	Isolated from	Locations	Genospecies	Nod groups
3841	*Vicia faba*	United Kingdom	gsB	B1
BLR195	*Lens culinaris*	Bangladesh	*Rhizobium binae*	B4†
CCBAU03058	*Vicia faba*	China	gsB	B5†
CCBAU10279	*Vicia faba*	China	gsF-2 / *Rhizobium laguerreae*	B5†
CCBAU11080	*Vicia faba*	China	gsG	B5†
CCBAU43229	*Vicia faba*	China	*Rhizobium anhuiense*	B3
CCBAU83268	*Vicia faba*	China	gsF-1	B5†
CZP1G9	*Pisum sativum*	Czech Republic	gsE	A4
CZP3G4	*Pisum sativum*	Czech Republic	*Rhizobium pisi*	A1†
CZP3H6	*Pisum sativum*	Czech Republic	gsC	A1
FRP3E11	*Pisum sativum*	France	gsE	B4
FRP3G5	*Pisum sativum*	France	gsE	B3
FRP4D11	*Pisum sativum*	France	gsE	A3†
GB29	*Pisum sativum*	Poland	gsC	A1
GB51	*Pisum sativum*	Poland	gsE	A1†
GD25	*Pisum sativum*	Poland	gsE	A4†
GLR17	*Lens culinaris*	Germany	gsF-2 / *Rhizobium laguerreae*	B4
GLR2	*Lens culinaris*	Germany	gsF-2 / *Rhizobium laguerreae*	B4
HP3	*Pisum sativum*	Algeria	gsF-2 / *Rhizobium laguerreae*	B2
IAUb11	*Pisum sativum*	France	gsE	A4
L113	*Lens culinaris*	France	gsE	A1†
L125	*Lens culinaris*	France	gsE	A1†
L411	*Lens culinaris*	France	gsF-1	B4
Lp8	*Lathyrus pratensis*	United Kingdom	gsC	A4
OS25	*Pisum sativum*	Algeria	gsF-2 / *Rhizobium laguerreae*	B3
P1NP2H	*Pisum sativum*	France	gsB	A4
P1NP2K	*Pisum sativum*	France	gsB	A3
P221	*Pisum sativum*	France	gsE	A1†
SEF5C12	*Vicia faba*	Sweden	gsC	A1
SL16	*Lens culinaris*	Algeria	gsF-2 / *Rhizobium laguerreae*	B3
TOM	*Pisum sativum*	Turkey	gsF-1	B5†
UPM1134	*Pisum sativum*	Italy	gsC	A5

Note: † indicates bacteria carrying the *nodX* gene within the nod gene cluster

774 **Table 2:** Putative functions and replicon location of genes associated with pea/lentil AFN or CFN
775 identified by the comparative genomic analyses.

Gene identifier	Comparison	Putative function	3841	Vaf10	Vaf-108	BIHB1217	UPM791	TOM
PK1	CFN	hypothetical protein	NA	Rh08	Rh08	Rh08	Rh08	NA
PK2	CFN	hypothetical protein	NA	NA	Rh02	Rh02	Rh02	NA
PI1	CFN	putative transmembrane	chr	NA	NA	chr	chr	NA
PI2	CFN	putative transmembrane	chr	NA	NA	chr	chr	NA
PI3	CFN	hypothetical protein	chr	Chr	NA	chr	chr	NA
PI4	CFN	RelE/StbE replicon stabilization toxin	chr	chr	NA	chr	chr	NA
PI5	CFN	hypothetical protein	chr	NA	NA	chr	chr	NA
PI6	CFN	amidinotransferase	chr	NA	NA	chr	chr	NA
PI7	CFN	amidinotransferase	chr	NA	NA	chr	chr	NA
PI8	CFN	miscellaneous	chr	NA	NA	chr	chr	NA
PI9	CFN	putative transcriptional regulator	Rh03	Rh08	Rh01	Rh08	Rh08	NA
PI10	CFN	hypothetical protein	Rh03	Rh08	Rh01	Rh08	Rh08	NA
PI11	CFN	uracil/thymine DNA glycolase	Rh02	NA	NA	Rh02	Rh02	NA
PI12	CFN	L-2-hydroxyglutarate oxidase	Rh05	NA	NA	Rh04a	Rh04a	NA
PI13	CFN	L-proline/glycine betaine transporter ProP	Rh05	NA	NA	Rh01	Rh01	NA
PA1	CFN	hypothetical protein	NA	chr	chr	NA	NA	NA
PA2	CFN	Delta-9 fatty acid desaturase	NA	chr	NA	NA	NA	NA
PA3	CFN	Putative membrane-bound ClpP-class protease	chr	chr	chr	chr	chr	NA
PA4	CFN	hypothetical protein	Rh02	NA	Rh02	Rh02	Rh02	NA
PA5	CFN	hypothetical protein	chr	chr	chr	Rh01	NA	NA
PA6	CFN	hypothetical protein	NA	Rh08	Rh17	NA	NA	NA
LA1	CFN	TPR domain protein in aerotolerance operon	chr	Rh12	NA	chr	NA	NA
LA2	CFN	BatA	chr	Rh12	Rh01	chr	NA	NA
LA3	CFN	Possible Neuromedin U precursor	chr	Rh12	NA	chr	NA	NA
LA4	CFN	Enoyl-acyl-carrier-protein reductase	chr	NA	NA	NA	NA	NA
LA5	CFN	hypothetical protein	Rh01	Rh01	NA	Rh01	NA	Rh01
LA6	CFN	hypothetical protein	chr	NA	NA	chr	chr	NA
LR1	CFN	Cell filamentation protein fic	chr	Rh08	chr	chr	chr	NA
LR2	CFN	VapC toxin protein antagonist	Rh01	NA	Rh01	NA	NA	Rh01
LF1	CFN	Polyhydroxyalkanoic acid synthase	Rh03	Rh17	Rh17	Rh08	NA	NA
LF2	CFN	Acetyl-coenzyme A synthetase	Rh03	Rh01	NA	Rh08	NA	NA
LF3	CFN	Homoserine O-acetyltransferase	Rh03	Rh01	NA	Rh08	NA	NA
LF4	CFN	putative dehalogenase-hydrolase	Rh03	Rh01	NA	Rh08	NA	NA
LF5	CFN	hypothetical protein	Rh03	Rh01	Rh01	Rh08	NA	NA
LF6	CFN	hypothetical protein	Rh03	Rh01	Rh01	Rh08	NA	NA
LF7	CFN	hypothetical protein	Rh03	Rh01	NA	Rh08	NA	NA
PAFN1	AFN	hypothetical protein	NA	Rh01	Rh01	NA	NA	Rh01
PAFN2	AFN	MFS permease	chr	chr	chr	NA	NA	chr
PAFN3	AFN	hypothetical protein	NA	NA	NA	NA	NA	chr
PAFN4	AFN	Beta-galactosidase	chr	NA	chr	NA	NA	chr
PAFN5	AFN	Beta-galactosidase	chr	NA	chr	NA	NA	chr
PAFN6	AFN	CobN component of cobalt chelatase	chr	chr	NA	NA	NA	chr
PAFN7	AFN	hypothetical protein	NA	Rh12	NA	Rh06	NA	Rh06
PAFN8	AFN	hypothetical protein	Rh12	Rh12	Rh01	Rh06	Rh12	Rh06
PAFN9	AFN	Bll0066 protein	Rh12	Rh12	Rh01	Rh06	Rh12	Rh06
PAFN10	AFN	hypothetical protein	Rh12	Rh12	Rh01	Rh06	Rh12	Rh06
PAFN11	AFN	Probably methylase/helicase	Rh12	Rh12	Rh01	Rh06	NA	Rh06
PAFN12	AFN	nodT RND efflux system	Rh03	Rh01	Rh08	Rh08	NA	Rh06
PAFN13	AFN	Alpha-aspartyl dipeptidase Peptidase E	chr	chr	chr	NA	NA	chr
PAFN14	AFN	putative transmembrane protein	chr	chr	chr	chr	chr	chr
PAFN15	AFN	PE-PGRS FAMILY PROTEIN	chr	chr	NA	NA	NA	chr
PAFN16	AFN	hypothetical protein	chr	chr	NA	NA	NA	chr
PAFN17	AFN	Bll0066 protein	Rh12	Rh12	Rh01	Rh06	Rh12	Rh03

Note: Replicons are classified as chromosome 'chr' or 'rh' *repABC* type plasmid (i.e. based on sequence homology to *repABC* genes as described by Cavassim et al., 2020). NA indicates no significant homologous sequence. Gene identifiers referred to the host: pea cv. 'Kayanne' (PK), pea cv. 'Isard' (PI), pea cv. 'Afghanistan' (PA), lentil cv. 'Anicia' (LA), lentil cv. 'Rosana' (LR) and lentil cv. 'Flora' (LF). More details are provided in Table S14.

776

Figure Legends:

Figure 1: Genomic diversity of the 121 fabeae symbiont genomes selected for this study. Hierarchical clustering and heat map were based on the Average Nucleotide Identity (ANI) values between each couple of the 124 bacterial genomes. *Rlv* genospecies classification (gs) has been based on an ANI threshold of 95%. Star gathered Sm1021 and isolates phylogenetically distantly related to *Rlv* (ANI<90%). Reference bacteria *Sinorhizobium meliloti* Sm1021*, Rhizobium leguminosarum symbiovar trifoli* WSM1689, *Rhizobium leguminosarum symbiovar phaseoli* Rlp4292 were also included in the comparison. Additional informations are provided in the Table S1.

Figure 2: *Nod* gene cluster diversity of of the 121 fabeae symbiont genomes selected for this study. Phylogenetic tree was based on the *nodABCDEFIJLMN* concatenated gene sequences of the 123 genomes. Boxes defined Nod groups. The 48 new fabeae symbiont genomes are indicated in italic. The 32 bacteria of the core collection are indicated in red. Stars indicate bacteria carrying the *nodX* gene within the *nod* gene cluster. *Rhizobium leguminosarum* symbiovar *trifolii* WSM1689 and *Sinorhizobium meliloti* 1021 have been used as outgroups.

Figure 3: Competitiveness to Form root Nodule of rhizobia with Fabeae plant species/genotypes. Hierarchical clustering and heat map based on the CFNi for each condition, including the 20 most competitive rhizobia within the core collection. FO: Vicia faba cultivar ‘Organdi’; FD: Vicia faba cultivar ‘Diva’; FT: Vicia faba cultivar ‘Tiffany’; PK: *Pisum sativum* cultivar ‘Kayanne’; PI: *Pisum sativum* cultivar ‘Isard’; PJ: *Pisum sativum* cultivar ‘Afghanistan JI1357’; LF: *Lens culinaris* cultivar ‘Flora’; LR: *Lens culinaris* cultivar ‘Rosana’; LA: *Lens culinaris* cultivar ‘Anicia’.

805 **Figures:**

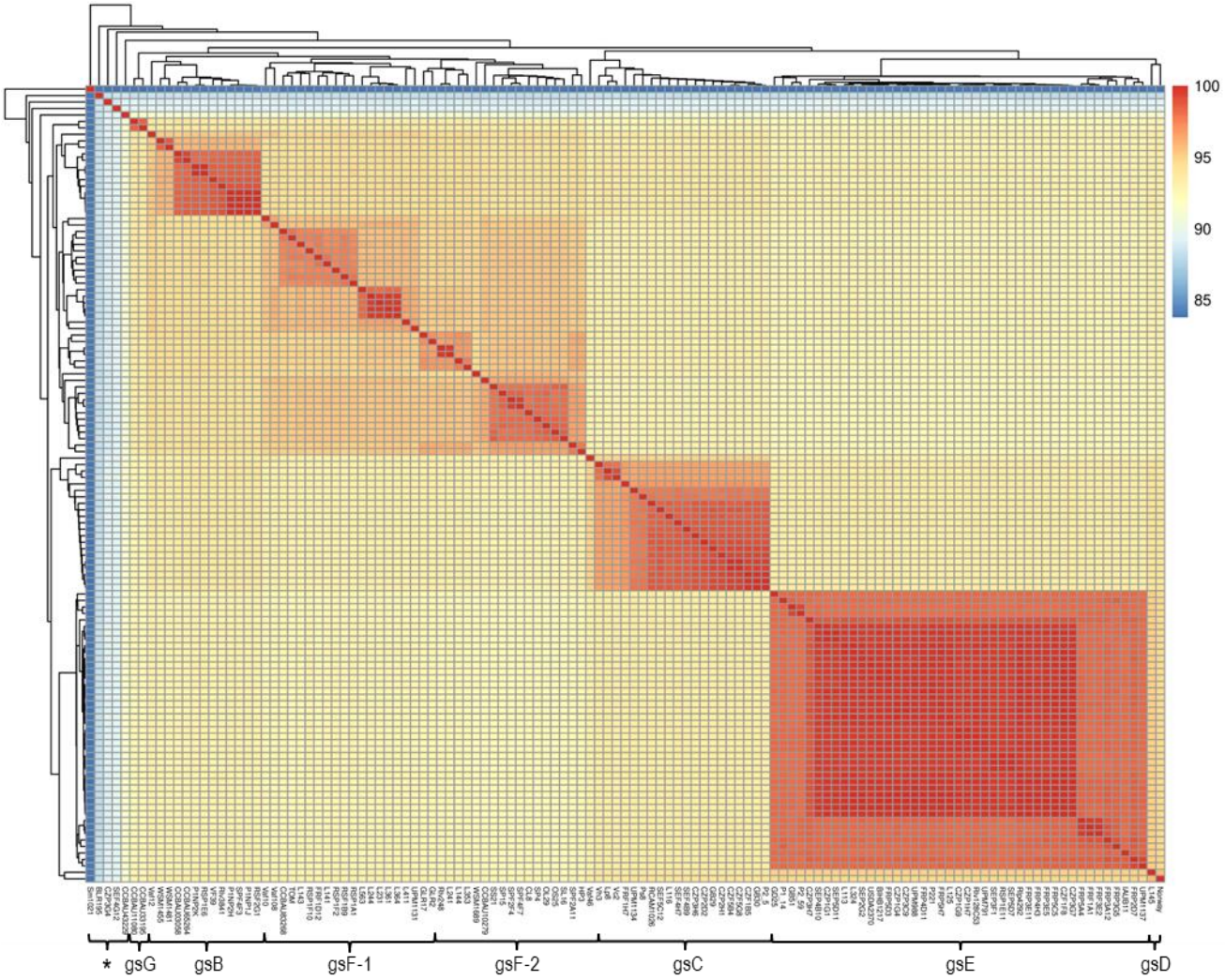

Figure 1: Genomic diversity of the 121 fabeae symbiont genomes selected for this study. Hierarchical clustering and heat map were based on the Average Nucleotide Identity (ANI) values between each couple of the 124 bacterial genomes. *Rlv* genospecies classification (gs) has been based on an ANI threshold of 95%. Star gathered Sm1021 and isolates phylogenetically distantly related to *Rlv* (ANI<90%). Reference bacteria *Sinorhizobium meliloti* Sm1021, *Rhizobium leguminosarum symbiovar trifoli* WSM1689, *Rhizobium leguminosarum symbiovar phaseoli* Rlp4292 were also included in the comparison. Additional informations are provided in the Table S1.

806

807

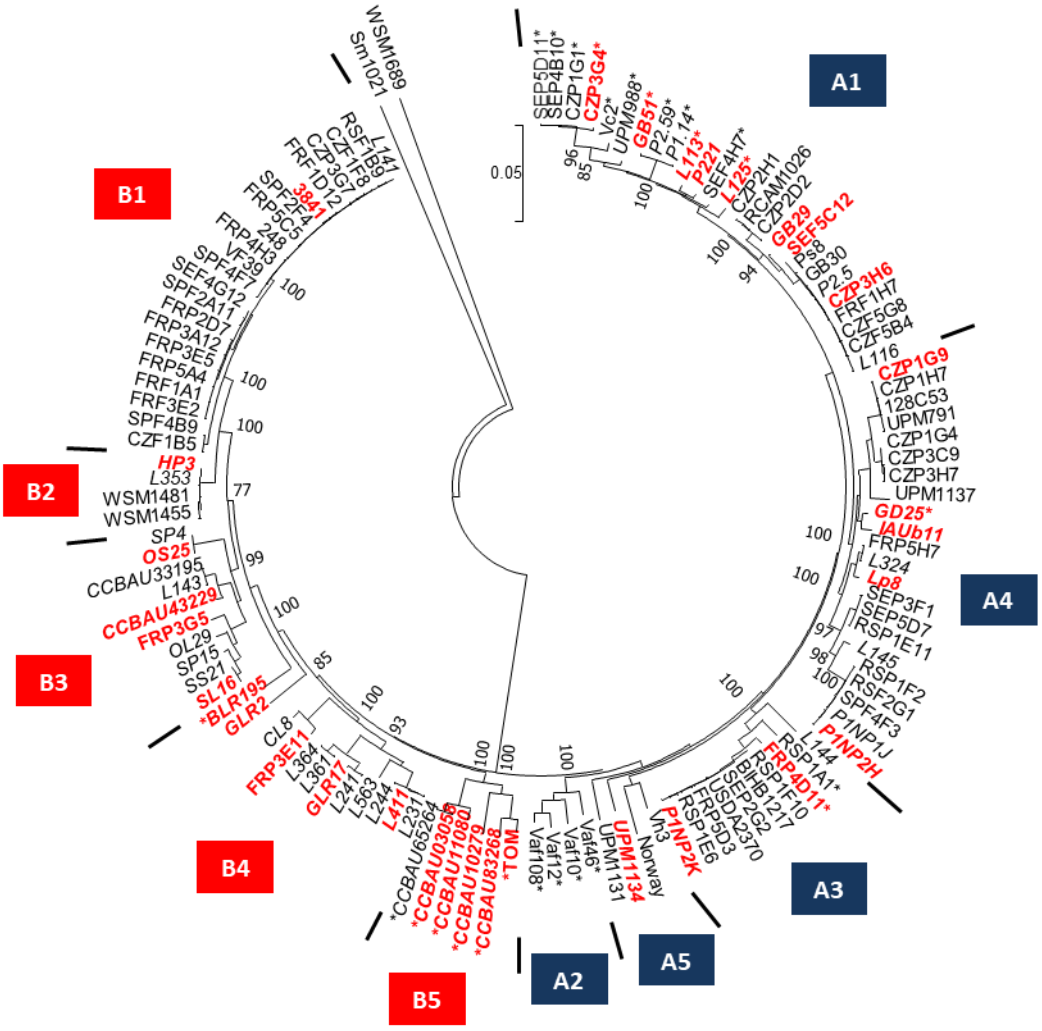

Figure 2: *Nod* gene cluster diversity of of the 121 fabeae symbiont genomes selected for this study. Phylogenetic tree was based on the *nodABCDEFIJLMN* concatenated gene sequences of the 123 genomes. Boxes defined Nod groups. The 48 new fabeae symbiont genomes are indicated in italic. The 32 bacteria of the core collection are indicated in red. Stars indicate bacteria carrying the *nodX* gene within the *nod* gene cluster. *Rhizobium leguminosarum* symbiovar *trifolii* WSM1689 and *Sinorhizobium meliloti* 1021 have been used as outgroups.

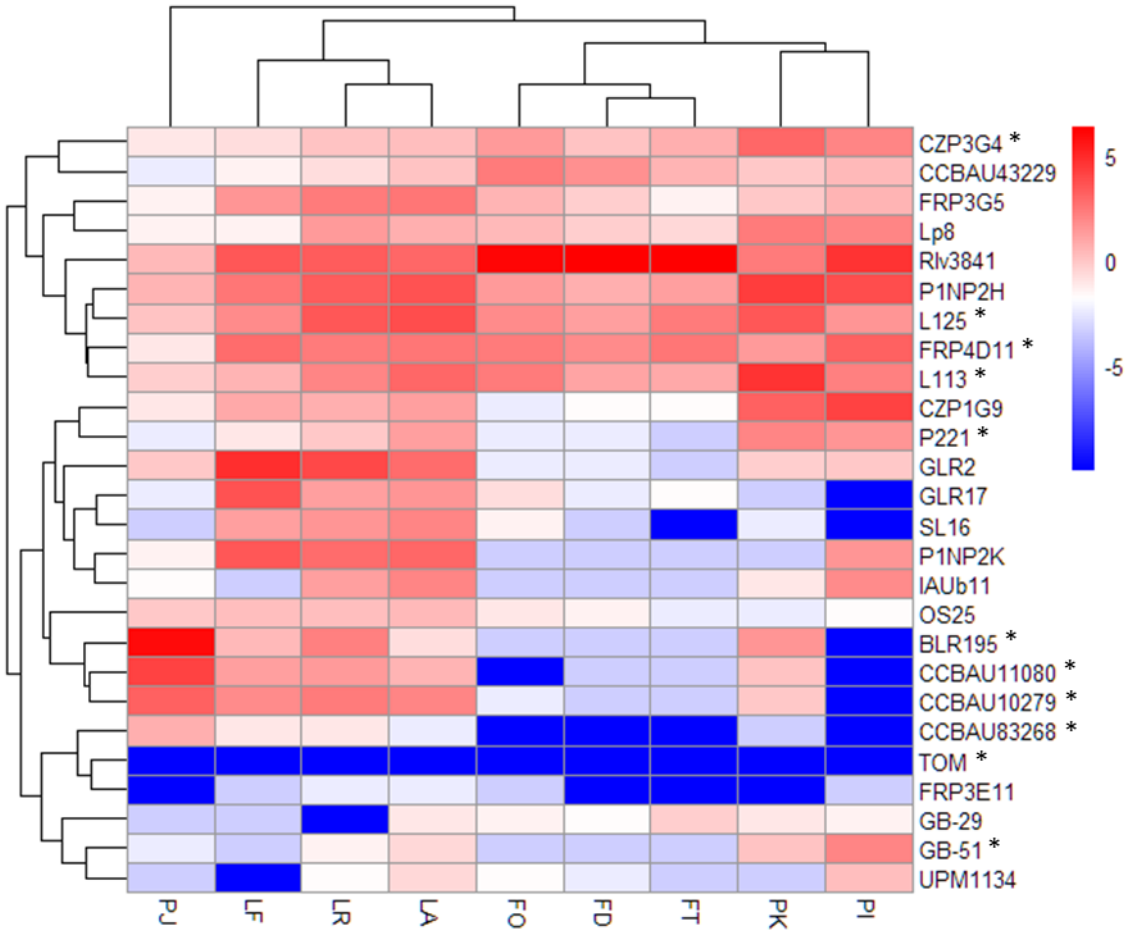

Figure 3: Competitiveness to Form root Nodule of rhizobia with Fabeae plant species/genotypes. Hierarchical clustering and heat map based on the CFNi for each condition, including the 20 most competitive rhizobia within the core collection. Stars indicate bacteria carrying the *nodX* gene within the *nod* gene cluster. FO: Vicia faba cultivar 'Organdi'; FD: Vicia faba cultivar 'Diva'; FT: Vicia faba cultivar 'Tiffany'; PK: *Pisum sativum* cultivar 'Kayanne'; PI: *Pisum sativum* cultivar 'Isard'; PJ: *Pisum sativum* cultivar 'Afghanistan JI1357'; LF: *Lens culinaris* cultivar 'Flora'; LR: *Lens culinaris* cultivar 'Rosana'; LA: *Lens culinaris* cultivar 'Anicia'.

810

Supplemental Tables and Figures

Figure S1: NodD309 barcode diversity among the 121 *Rhizobium leguminosarum* symbiovar *viciae* bacteria of this study.

Table S1: Genomic data of the sequenced *Rhizobium leguminosarum* symbiovar *viciae* bacteria.

Table S2: NodD309 sequences for the 32 *Rlv* bacteria included in the core collection, primers used and PCR conditions.

Table S3: Number of reads in each sample after quality filtering.

Table S4: Presence/absence of genomic regions associated with CFN phenotypes in pea, faba bean and lentil.

Table S5: Global rearrangements of plasmids/contigs between reference strains, in comparison with 3841.

Table S6: Genomic organization of contigs, containing genes associated to CFN, compared to 3841.

Table S7: Genospecies and Nod group combinations among the bacterial collection.

Table S8: Competitiveness to Form Nodule indexes (CFNi) for each of the plant/bacteria couples.

Table S9: Clusters of Rlv bacteria from the core collection, depending on their CFN profiles in the Figure 3.

Table S10: Effect of variation of *Rlv* genospecies, Nod types and Nod groups in the core collection on the CFN in the pea and lentil cultivars.

Table S11: Mono-inoculation of *Rlv* strains included in the core collection with the *Pisum sativum* cultivar 'Afghanistan'.

Table S12: Effect of the plant species on the CFN of the *Rlv* bacteria from the core collection, as function of the Nod type, the Nod group, and the genospecies.

Table S13: Presence/absence of genomic regions associated with AFN in *Pisum sativum* cultivar ‘Afghanistan’

Table S14: Locations of genes significantly associated with AFN and CFN from the comparative genomic analyses