Systematic and Applied Microbiology xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Genomic insight into the taxonomy of *Rhizobium* genospecies that nodulate *Phaseolus vulgaris*

Wenjun Tong[a,1], Xiangchen Li[a,b,1], Yunyun Huo[a], Lu Zhang[a], Ying Cao[a], Entao Wang[c], Weimin Chen[a,*], Shiheng Tao[a,b,*], Gehong Wei[a,*]

[a] State Key Laboratory of Crop Stress Biology in Arid Areas, College of Life Sciences, Northwest A&F University, Yangling, Shaanxi 712100, China
[b] Bioinformatics Center, Northwest A&F University, Yangling, Shaanxi 712100, China
[c] Departamento de Microbiología, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, 11340 México D.F., Mexico

ARTICLE INFO

Article history:
Received 17 December 2017
Received in revised form 5 March 2018
Accepted 7 March 2018

Keywords:
Phaseolus vulgaris
Rhizobium
Symbiovar
Genome-based taxonomy

ABSTRACT

Due to the wide cultivation of bean (*Phaseolus vulgaris* L.), rhizobia associated with this plant have been isolated from many different geographical regions. In order to investigate the species diversity of bean rhizobia, comparative genome sequence analysis was performed in the present study for 69 *Rhizobium* strains mainly isolated from root nodules of bean and clover (*Trifolium* spp.). Based on genome average nucleotide identity, digital DNA:DNA hybridization, and phylogenetic analysis of 1,458 single-copy core genes, these strains were classified into 28 clusters, consistent with their species definition based on multilocus sequence analysis (MLSA) of *atpD*, *glnII*, and *recA*. The bean rhizobia were found in 16 defined species and nine putative novel species; in addition, 35 strains previously described as *Rhizobium etli*, *Rhizobium phaseoli*, *Rhizobium vallis*, *Rhizobium gallicum*, *Rhizobium leguminosarum* and *Rhizobium* spp. should be renamed. The phylogenetic patterns of symbiotic genes *nodC* and *nifH* were highly host-specific and inconsistent with the genomic phylogeny. Multiple symbiovars (sv.) within the *Rhizobium* species were found as a common feature: sv. phaseoli, sv. trifolii and sv. viciae in *Rhizobium anhuiense*; sv. phaseoli and sv. mimosae in *Rhizobium sophoriradicis*/*R. etli*/*Rhizobium* sp. III; sv. phaseoli and sv. trifolii in *Rhizobium hidalgonense*/*Rhizobium acidisoli*; sv. phaseoli and sv. viciae in *R. leguminosarum*/*Rhizobium* sp. IX; sv. trifolii and sv. viciae in *Rhizobium laguerreae*. Thus, genomic comparison revealed great species diversity in bean rhizobia, corrected the species definition of some previously misnamed strains, and demonstrated the MLSA a valuable and simple method for defining *Rhizobium* species.

Introduction

With the development of cost-effective and high-throughput next-generation sequencing, genomics has revolutionized microbial research, especially determination of evolutionary relationships between prokaryotic species [53]. Traditionally, 70% of relatedness in DNA:DNA hybridization (DDH) as the species threshold has been applied as the gold standard for nearly 50 years [69]. However, DDH has some flaws due to its complex, time-consuming, scarcely reproducible procedure [17]. Novel tools such as average nucleotide identity (ANI) and digital DDH (dDDH) based on the genome sequence have been proposed for species delineation [5,63]. Both ANI and dDDH achieve a high correlation with traditional DDH. The 70% DDH threshold is equivalent to ANI values of 95–96% [63]. Integration of genome sequences into the systematics and taxonomy of *Bacteria* and *Archaea* strengthens the believability of microbial systematics [30], and revision of taxonomy with the genomic data has been applied to some bacterial groups such as *Rhizobium etli* [38] and *Bacillus cereus* [36].

The genus *Rhizobium* is affiliated with the class α-*Proteobacteria*, order *Rhizobiales*, and family *Rhizobiaceae*. So far, more than 90 *Rhizobium* species have been described and all of them are Gram-negative, non-spore-forming, aerobic bacteria. Most of the *Rhizobium* species were recovered from legume root nodules, including more than 15 symbiotic species from effective nodules of bean, a cosmopolitan leguminous food crop. The symbiotic bacteria of bean have been widely investigated in both native and non-native countries [42], and they are mainly affiliated to the genus *Rhizobium* [42]. Depending on their specificities of host plants and phylogenetic types of symbiosis genes, rhizobial strains could be

* Corresponding authors.
E-mail addresses: chenwm029@nwafu.edu.cn (W. Chen), shihengt@nwafu.edu.cn (S. Tao), weigehong@nwafu.edu.cn (G. Wei).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.syapm.2018.03.001
0723-2020/

defined as symbiotic variants (symbiovars) despite of their species definition, and a single rhizobial species may contain different symbiovars depending on the presence of distinct symbiosis plasmids or symbiotic islands [21,33]. In our previous studies, 371 rhizobial strains were isolated from bean nodules in four provinces of China (Shaanxi, Heilongjiang, Liaoning, and Jiangxi), and most of them were affiliated to *Rhizobium* based on analysis of 16S rRNA gene sequences [10,80]. Although they mainly corresponded to *R. etli*, *Rhizobium phaseoli*, and *Rhizobium leguminosarum* based on phylogeny of concatenated housekeeping genes (*atpD*, *glnII*, and *recA*) [10,80], some of them displayed distinguishing genomic lineages [1] with unclear species affiliation.

Given the powerful resolution of comparative genome analysis for bacterial systematics, and the existence of many bean-nodulating strains with uncertain species affiliation, in the present study we further clarified the phylogenetic relationships and the taxonomic positions of 19 bean symbionts isolated from China and two bean symbionts from Mexico by comparative genome approaches. Another 48 genomes for related strains assigned to *R. etli*, *R. phaseoli*, and *R. leguminosarum* were also extracted from databases and included as references. A variety of approaches were used in this study, including core-genome phylogeny based on all shared single-copy genes, together with ANI and dDDH analyses, phylogenetic assessment of 16S rRNA genes, and multilocus sequence analysis (MLSA) of three housekeeping genes (*atpD*, *recA*, and *glnII*) and symbiotic genes (*nodC* and *nifH*). The purposes of this study were (i) to establish a reliable and relatively simple protocol for effective species definition among closely related *Rhizobium* species, and (ii) to reconstruct the taxonomic status of some *Rhizobium* strains.

Materials and methods

DNA extraction, genome sequencing, and assembly

Each of the 21 *Proteus vulgaris* microsymbionts (Table 1) was cultured in 5 mL of tryptone yeast broth (tryptone, 5 g; yeast extract, 3 g; $CaCl_2$, 0.6 g; distilled water, 1 L; pH 7.2) to the late log phase. Genomic DNA was extracted from each culture using a commercial DNA extraction kit (Catalogue No. 9763, TaKaRa, Dalian, China), following the manufacturer's instructions. Draft genomes of the 21 isolates were sequenced by massively parallel sequencing using Illumina technology at Novogene Bioinformatics Technology Co., Ltd (Beijing, China). The 500 bp library was sequenced with an Illumina HiSeq 2500 by PE250 strategy. All high-quality paired reads were assembled by SOAPdenovo [40] to generate the scaffolds.

Gene prediction and functional annotation

Transfer RNA (tRNA) genes and ribosomal RNA (rDNA) genes were predicted with tRNAscan-SE [39] and rDNAmmer [32], respectively. A whole-genome BLAST search [37] (E-value $\leq$1e-5, minimal alignment length percentage $\geq$40%) was performed against the Non-Redundant (NR) protein database. Automatic annotation was conducted on the RAST web server [49]. The acquired draft genomes have been deposited in the GenBank database under project accession no. PRJNA403813.

Collection of genome sequence data

A total of 69 *Rhizobium* genomes comprising the 21 draft genomes obtained by our laboratory and 48 complete or draft genomes of related *Rhizobium* strains available in the GenBank database were collected (Table 1). Among them, 54 strains were isolated from *P. vulgaris*, eight strains were from *Trifolium* spp., one strain was isolated from *Mimosa affinis*, two strains were from *Mimosa pudica*, two strain nodulated *Pisum sativum*, and two strain nodulated *Vicia faba*. Previously, these were named *R. etli* (14 strains), *R. phaseoli* (six strains), *R. leguminosarum* (18 strains in three symbiovars, namely phaseoli, trifolii, and viciae), *Rhizobium mesoamericanum* (two strains), *Rhizobium anhuiense*, *Rhizobium vallis* and *Rhizobium gallicum* (one strain each), and 13 *Rhizobium* type strains. The remaining 13 strains were defined as unclassified bacteria in *Rhizobium*.

ANI and dDDH calculation

All pairwise ANI values among the 69 strains were calculated using the MUMmer algorithm in Jspecies software [63]. All pairwise dDDH values were estimated by GGDC 2.1 under the recommended Formula 2 since only ~20% of the genome is needed to obtain the same result as with the full genome [45]. Heatmaps of ANI and dDDH results were built using the pheatmap package [29] in R version 3.2.5.

Correlations among ANI, dDDH, and MLSA similarity (obtained in subsequent analyses) were determined using a nonlinear simulation analysis method with the default option of the Curve Fitting Tool implemented in MATLAB 8.1 (The MathWorks Inc., Natick, MA, USA).

Species tree construction

Sixty-eight *Rhizobium* genomes except the poor-quality genome GR56 were used to retrieve a more accurate reference phylogeny with a pan-genome analysis implemented by using the ITEP pipelines [6]. Ortholog clusters were generated using the Markov Cluster algorithm with an inflation value of 2.0 and a cut-off value of 0.4. Single-copy core genes were aligned with mafft v7.271 [26] and reverse-translated with PAL2NAL v14 [73]. These alignments were concatenated to infer maximum likelihood phylogenies with RAxML v8.2.4 under the GTR model using gamma correction for variable evolutionary rates [72]. The best maximum likelihood phylogenies were generated using autoMRE bootstrap convergence tests. In addition, supertree was constructed using the neighbor-joining method with 1000 bootstrap by MEGA 6.0 with concatenated alignments.

Phylogenetic analysis of 16S rRNA gene, housekeeping genes, and symbiotic genes

For 68 of the tested strains, the sequences of 16S rRNA genes, three housekeeping genes (*recA*, *atpD*, and *glnII*), and two symbiotic genes (*nodC* and *nifH*) were extracted from database or genome sequences using BLAT [27], but this extraction was failed for the GR56 strain due to its poor-quality genome sequence. To investigate the absence of certain genes in some genomes, PCR amplification was performed using genomic DNA as template [41] to confirm that the absence was the result of incomplete genome sequencing. The genes *recA*, *atpD*, and *glnII* were chosen for MLSA since they have been frequently used for *Rhizobium* species definition [12]. To clarify the species definition of the 68 tested strains, 16S rRNA and the three housekeeping genes of type strains for related *Rhizobium* species were also extracted from the GenBank database. All gene sequences were aligned using the Muscle program in MEGA version 6.0 [74]. In the MLSA, the sensitivity and quality of the data were established to assess phylogeny [20]. Quality was evaluated by measuring the index of substitution saturation (*Iss* and *Iss.c*) with DAMBE [84]. The three housekeeping genes were combined and the sensitivity of the concatenated sequences was also analyzed. Incongruence length difference tests [14] were performed with the PAUP*4.0b10 program to evaluate whether datasets containing different evolutionary histories were incongru-

Please cite this article in press as: W. Tong, et al., Genomic insight into the taxonomy of *Rhizobium* genospecies that nodulate *Phaseolus vulgaris*, Syst. Appl. Microbiol. (2018), https://doi.org/10.1016/j.syapm.2018.03.001