Plasmids Related to the Symbiotic Nitrogen Fixation Are Not Only Cooperated Functionally but Also May Have Evolved over a Time Span in Family *Rhizobiaceae*

Ling-Ling Yang[1], Zhao Jiang[1], Yan Li[2], En-Tao Wang[3], and Xiao-Yang Zhi[1,*]

[1]Yunnan Institute of Microbiology, School of Life Sciences, Yunnan University, Kunming, Yunnan, PR China

[2]Key Laboratory of Coastal Biology and Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, Shandong, PR China

[3]Departamento de Microbiología, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, Mexico City D.F., México

*Corresponding author: E-mail: xyzhi@ynu.edu.cn.

Accepted: 15 July 2020

Abstract

Rhizobia are soil bacteria capable of forming symbiotic nitrogen-fixing nodules associated with leguminous plants. In fast-growing legume-nodulating rhizobia, such as the species in the family *Rhizobiaceae*, the symbiotic plasmid is the main genetic basis for nitrogen-fixing symbiosis, and is susceptible to horizontal gene transfer. To further understand the symbioses evolution in *Rhizobiaceae*, we analyzed the pan-genome of this family based on 92 genomes of type/reference strains and reconstructed its phylogeny using a phylogenomics approach. Intriguingly, although the genetic expansion that occurred in chromosomal regions was the main reason for the high proportion of low-frequency flexible gene families in the pan-genome, gene gain events associated with accessory plasmids introduced more genes into the genomes of nitrogen-fixing species. For symbiotic plasmids, although horizontal gene transfer frequently occurred, transfer may be impeded by, such as, the host's physical isolation and soil conditions, even among phylogenetically close species. During coevolution with leguminous hosts, the plasmid system, including accessory and symbiotic plasmids, may have evolved over a time span, and provided rhizobial species with the ability to adapt to various environmental conditions and helped them achieve nitrogen fixation. These findings provide new insights into the phylogeny of *Rhizobiaceae* and advance our understanding of the evolution of symbiotic nitrogen fixation.

Key words: Rhizobiaceae, phylogenomics, symbiotic plasmid, accessory plasmid, genome expansion, symbiotic nitrogen fixation.

Introduction

Rhizobia are soil bacteria capable of forming nitrogen-fixing symbioses with legumes and sustaining the latter's growth in poor nitrogen soils. Legumes are the third-largest group of angiosperms and the second largest group of food and feed crops grown globally. They have significant potential for sustainable production without the economic and environmental costs of chemical fertilization (Ferguson et al. 2010). Therefore, rhizobia, and especially rhizobia–legume symbiosis (Oldroyd et al. 2011; Udvardi and Poole 2013), which is one of the most significant ecological services that prokaryotes offer eukaryotes, have attracted much scientific attention. Phylogenetically, rhizobia are polyphyletic and distributed mainly in the alpha-Proteobacteria (Shamseldin et al. 2017). Among them, *Rhizobiaceae* represents the most cohesive and widely distributed group, and it contains seven genera with symbiotic members: *Rhizobium*, *Ensifer* (formerly *Sinorhizobium*), *Agrobacterium*, *Allorhizobium*, *Neorhizobium*, *Pararhizobium*, and *Shinella*. Previous phylogenomic analyses revealed two major super clades within *Rhizobiaceae* that corresponded to the *Rhizobium–Agrobacterium* and *Shinella–Ensifer* groups (Ormeño-Orrillo et al. 2015).

Although symbiotic nitrogen fixation (SNF) is a unique feature that defines rhizobia, they also have saprophytic and endophytic lifestyles in soil and nonsymbiotic hosts (such as

like rice, potato, and maize), respectively (Ji et al. 2010). In different environments, the rhizobia survive in complex microbial communities and compete with other members of the microbiota. During initiation of the *Rhizobium*–legume symbiosis, rhizobia enter the host plant cells and differentiate into nitrogen-fixing bacteroids (Mergaert et al. 2006). Rhizobia have achieved ecological and evolutionary successes that have reshaped our biosphere (Masson-Boivin and Sachs 2018). The dual lifestyles and developmental changes are supported by large (~5–10 Mb) and highly plastic genomes that are richly endowed with transport, regulatory, and stress-related functions (Remigi et al. 2016; Poole et al. 2018). Furthermore, rhizobial genomes often possess complex architectures that consist of a chromosome plus one or more additional replicons. For free-living growth, the nonchromosomal replicons can be essential (e.g., chromids, which are hybrid replicons with both plasmid and chromosome features) (Harrison et al. 2010; diCenzo et al. 2013) or nonessential (e.g., symbiotic plasmid). In *Rhizobiaceae*, SNF-related genes (*nod*, *nif*, and *fix*) are usually located on plasmids (so defined as symbiotic plasmid), or occasionally on chromids (e.g., in *Neorhizobium galegae*). Studies on sympatric populations of rhizobia confirmed greater genetic diversity levels in chromosomes and accessory plasmids (replicons other than symbiotic plasmid, but including chromids) than in symbiotic plasmids (Guo et al. 2014; Kumar et al. 2015; Pérez Carrascal et al. 2016; Wang et al. 2018). These findings indicated that SNF-related genes are frequently exchanged between different rhizobial genotypes in the natural environment. However, these conclusions were mainly based on investigations that focused on several commonly studied rhizobial species, such as *Rhizobium leguminosarum*, *Rhizobium etli*, *Rhizobium phaseoli*, and *Rhizobium gallicum*. The breadth of symbiotic plasmid transfers within a lineage like *Rhizobiaceae* is still unclear. Thus, comparative genomics at the family level is required to further understand the cross-species transfer of symbiotic plasmids.

In addition to the pivotal role of symbiotic plasmids in SNF, both experimental observation (Ramachandran et al. 2011) and metabolic modeling (diCenzo et al. 2016) have demonstrated that adaption to the nodule environment depends on accessory plasmid genes (Barreto et al. 2012; Zahran 2017). From an evolutionary perspective, accessory plasmids might be of more significance, given that frequent horizontal transfer has obscured the evolutionary vertical signal of symbiotic plasmids. However, accessory plasmid evolution in *Rhizobiaceae* remains mostly unknown, except that they have similar genetic diversity to chromosomes (Pérez Carrascal et al. 2016; Wang et al. 2018). In these nonchromosomal replicons, even though chromids are often separately analyzed because they carry essential genes, chromids may have originated from a megaplasmid that subsequently gained a few core genes because of the severe underrepresentation of essential genes and functional bias (Harrison et al. 2010; diCenzo and Finan 2017). Recently, comparative genomics and ancestral state reconstruction analyses indicated that all chromids of family *Burkholderiaceae* arose from plasmid acquisition events; chromids likely increased the genetic malleability and then there was accumulation of niche-specialized functions (diCenzo et al. 2019). Therefore, chromid evolution in *Burkholderiaceae* can help elucidate accessory plasmid evolution in *Rhizobiaceae*.

To reveal the genome evolution within the family *Rhizobiaceae* (rhizobia-related genera), in this work, we collected the type strains of species in this family, and filled the phylogenetic gaps using genomic data. Phylogenomic analysis showed that the main clades of family *Rhizobiaceae* were relatively consistent between supermatrix and gene-content methods. Additionally, ancestral state reconstruction of gene content indicated that the dramatic changes in gene number occurred among the different common ancestors of symbiotic species. Moreover, the pan-genome has expanded throughout the whole evolutionary history of *Rhizobiaceae*. Although genomic expansion of chromosomes occurred throughout the family, accessory plasmid acquisition expanded the gene content of some ancestors, especially the ancestors of symbiotic species. In contrast, the effect of symbiotic plasmids on genome expansion was less than that of accessory plasmids. Evidently, diverse symbiotic plasmids exist in the *Rhizobiaceae* pan-genome, and although some species such as phylogenetically close ones have similar genomic backgrounds compatible with the same symbiotic plasmid, the nonbiological factors, such as physical isolation of their hosts may also impede the widespread transfer of a symbiotic plasmid. These findings indicated that the symbiotic species of *Rhizobiaceae* underwent genome expansion by successively acquiring accessory and symbiotic plasmids, which allowed them to undergo dramatic lifestyle changes and SNF.

Materials and Methods

Collection and Genome Sequencing of Type Strains

The type strains of 39 species belonging to the genera *Rhizobium* (32 species), *Agrobacterium* (two species), *Allorhizobium* (four species, two of which were previously published), and *Pararhizobium* (one species) were collected from the culture collection centers CCTCC, DSM, CCBAU, and LMG. All the strains were cultivated on nutrient agar medium (peptone, 10 g; beef extract, 3 g; sodium chloride, 5 g; agar, 15 g; distilled water, 1 l; pH 7.2) at 28 °C for biomass collection. Genomic DNA was isolated from cell pellets collected in a 60-ml culture (28 °C for 24 h) following the CTAB bacterial genomic DNA isolation protocol (version 3) provided by the DOE JGI (https://jgi.doe.gov/user-programs/pmo-overview/protocols-sample-preparation-information/).

Sequencing was carried out using the Illumina MiSeq platform, and paired-end reads (average, 250 bp) were generated

from 460-bp insert libraries. High-quality paired-end reads were assembled using SOAPdenovo2 and optimized using GapCloser v1.12 (Luo et al. 2012). In addition, 53 reference genome sequences were retrieved from the NCBI Assembly (https://www.ncbi.nlm.nih.gov/assembly). The genomes sequenced in this study are available from the JGI IMG database (Markowitz et al. 2012). In total, genome sequences of 92 strains used in this study were annotated using Prokka v1.11 (Seemann 2014). The detailed for the 92 strains information, including isolation sources and taxonomic references, are shown in supplementary table S1, Supplementary Material online.

Phylogenomic Analysis

Homologous protein families (PFs) were determined using the program OrthoMCL v2.0.9 (Li et al. 2003). The BLAST reciprocal best-hit algorithm (Moreno-Hagelsieb and Latimer 2008) was employed, and Markov cluster algorithms (Enright et al. 2002) were applied with an inflation index of 1.5. All PFs were divided into two categories: core PFs, which included paralogs and orthologs shared by all species, and species group-specific (GS) PFs, which only included those found in a subset of species. The GS PFs were further divided into ten groups (GS0–GS9) based on their distributions and proportions in the 92 genomes. For example, group GS9 included all PFs that could be found in $\geq$90% (83) of the 92 genomes, whereas group GS0 included all PFs that could be found at least in one genome but fewer than nine genomes. Each PF contained at least two sequences. Besides PFs, in the pan-genome of *Rhizobiaceae*, there were many genes present as a single copy (singletons). Therefore, the whole pan-genome could be divided into 12 portions: core (paralogs and orthologs), ten GSs, and singletons. In this work, all singleton proteins of $<$100 amino acids (aa) were ignored because of their functional uncertainty and the potential inaccuracy of gene prediction.

Both supermatrix and gene-content methods were applied to infer phylogenetic trees (Delsuc et al. 2005). Before constructing the supermatrix tree, the phylogenies of the individual orthologous proteins were inferred based on the maximum likelihood (ML) method by PhyML 3.0 (Guindon et al. 2010). To estimate the incongruence of the phylogenies of the individual orthologous proteins, principal component analysis of the likelihood values estimated for each tree's topologies, based on every protein data set, was used (Brochier and Philippe 2002). This method allowed the simultaneous analysis of the congruence between many proteins within a reasonable computational time (Brochier et al. 2002). For the supermatrix phylogenetic tree, we selected all of the orthologous PFs shared by the 92 genomes. However, the PFs that had shorter alignment lengths ($\leq$100 sites) and were incongruent with other PFs based on phylogeny, were excluded. The detailed methods for building supermatrix and gene-content trees are the same as described in Zhi et al. (2017).

Gene Family Turnover Rate Estimates

In addition to phylogenomic trees, an ultrametric tree was constructed using program r8s (Sanderson 2003) based on the supermatrix tree. In this process, two calibration points, including the *R. leguminosarum* and *Rhizobium pisi* at 14.94 Ma, and the node of *R. leguminosarum* and *Rhizobium endophyticum* at 71.5 Ma, were introduced and used to estimate the ages of the remaining nodes (Kumar et al. 2017). The gene family turnover rates and family sizes of the ancestors (internal nodes) were estimated using BadiRate (Librado et al. 2012). Four turnover rate models (Birth, Death and Innovation rates, Lambda and Innovation rates, Gain and Death rates, and Birth and Death rates) and two different branch-specific models (all branches have the same turnover rates, and some particular branches have different turnover rates) were used, and the best model was selected according to the Akaike information criterion (AIC).

Plasmid Sequence Identification and Comparative Genome Analysis

To discriminate the sequences that belong to plasmids in a draft genome, we used the method of homologous fragments counting. The scaffold sequence was fragmented into 100-bp subsequences (without overlap), and then each subsequence (fragment) was mapped onto the reference plasmids using BlastN in BLAST+ (key parameters: "-evalue 0.00001 -max_target_seqs 1 -perc_identity 50") (Camacho et al. 2009). The percentage of fragments that could be matched to plasmid references was calculated (count only once for fragment with multiple hits), and the scaffold that contained $\geq$50% fragments homologous to plasmids was classified as a plasmid sequence. Moreover, the scaffold that contained 10 kb of plasmid-homologous fragments were also treated as a plasmid sequence. To build the plasmid reference databases, 497 complete genome sequences (including 583 plasmid sequences) were collected from NCBI GenBank that belonged to the genera *Rhizobium*, *Agrobacterium*, *Aminobacter*, *Burkholderia*, *Cupriavidus*, *Devosia*, *Ensifer*, *Methylobacterium*, *Microvirga*, *Ochrobactrum*, and *Phyllobacterium*. The plasmids, including symbiotic and accessory plasmids, were separated from chromosomes. Finally, two independent BlastN databases were built for accessory plasmids (497 sequences without SNF-related genes) and symbiotic plasmids (86 sequences that contained *nod*, *nif*, and *fix* genes), respectively. An in-house python script was used to fragment scaffolds, calculate the percentages of homologous fragments, and split the scaffolds identified as plasmids.

The pairwise average nucleotide identity (ANI) within the genome compartments: chromosomal regions, accessory

plasmid regions (APRs), and symbiotic plasmid regions (symPRs) were estimated using FastANI (Jain et al. 2018). Based on the ANI matrix, the genome compartments were clustered using the UPGMA algorithm in the pheatmap package of the R program. Additionally, we determined the nucleotide diversity for each PF in different genome compartments using VariScan v2.0.3 with the default parameters (Vilella et al. 2005). The gene set enrichment analysis of different genome compartments was performed using Kobas v3.0 (Xie et al. 2011). The significantly enriched pathways (Kanehisa et al. 2017) were identified using Fisher's exact test. Then, the Benjamini–Hochberg procedure for multiple tests was used to control the false discovery rate.

Results and Discussion

Genomic Features of Tested Strains

For the 39 genome sequences obtained in this study, the coverage of clean reads was >161-fold (averaging 279.9-fold coverage, supplementary table S2, Supplementary Material online). The number of scaffolds in each draft genome ranged from 40 (*Agrobacterium pusense* and *Pararhizobium herbae*) to 346 (*Rhizobium sullae*), with an average of 105.9. The N50 value (the sequence length of the shortest contig at 50% of the total genome length) had a mean of 538,491 bp (supplementary table S3, Supplementary Material online). The average number of coding DNA sequences (CDSs) was 5,929 (SD: 1,053), ranging from 3,925 to 7,828 (supplementary table S4, Supplementary Material online). When including the additional 53 reference genomes, the average genome size of the symbiotic species was 6.7 Mb, which was significantly (Student's *t* test, *P* value <0.01) different from that of the nonsymbiotic species (5.4 Mb, supplementary fig. S1, Supplementary Material online). As discussed in the Introduction, the significant difference in genome size might be related to the features that allow symbiotic species to adapt to more complicated habitats (in nodules as symbionts, in root tissues as endophytes, and in soil as saprophytes) (Nadarajah 2017) compared with the nonsymbiotic bacteria (endophytes and/or saprophytes).

Pan-Genome of Family *Rhizobiaceae*

In the pan-genome of the 92 tested strains, 514,196 proteins were predicted, of which 25,310 proteins were present as singletons (including 13,168 ones with length >100 aa). In total, 23,557 PFs, including 247 paralogs (with at least one copy in each genome), 992 orthologs (with only one copy in each genome), and 22,318 species GS PFs (including 518 single species-specific PFs), were identified from all predicted proteomes. All GS PFs were further divided into ten groups based on their distributions in all 92 predicted proteomes. The histogram of the number of PFs versus the groups of the pan-genome was approximating bimodal (fig. 1*A*); PFs with intermediate breadths tended to be lower in abundance compared with PFs that make up the core and highly flexible genome. That is a common pattern for prokaryote pan-genomes (Lapierre and Gogarten 2009). However, interestingly, the bottom of the U-shape deflected to the left (near to the pan-genome core); this indicated that there was a high proportion of low-frequency flexible gene families in *Rhizobiaceae* pan-genome. Moreover, the abundance gradually increased as the PFs' distribution narrowed.

In general, we thought that the core genes are crucial for all individuals of a population, and the genes near the core (absent in a few of species) might be the result of gene loss events, or caused by the incompleteness of genome sequences. For a species of *Rhizobiaceae*, only 35.9% (mean percentage) of its genes were shared with nearly all neighbor species (core and GS9, fig. 1*B*), and its identity as an independent species was mainly characterized by 12.5% of its genes (GS0, and singletons excluding sequences of <300 bp, which accounted for 2.3% of the genome). Additionally, 16.8% of its genes (GS5–GS8) were shared with the majority of neighbors (≥60%); 19.5% of its genes (GS1–GS4) were shared with the minority of neighbors (<40%). Although the proportion of low-frequency (GS0–GS4) flexible genes in a single genome was not much higher than the proportion of high-frequency (GS5–GS8) flexible genes, the number of low-frequency flexible gene families in the pan-genome was clearly higher than that of high-frequency flexible gene families, because of the cumulative counting.

Theoretically, the evolutionary events that affect the high-frequency gene family in a pan-genome mainly include ancient gene innovation (e.g., gene gain, and gene duplication followed by neo- or subfunctionalization), and/or the recent gene loss (including the complete loss of functionality and the functional replacement by nonhomologous genes). Both mechanisms can seemingly lead to the same distribution of the high-frequency gene family. The distribution of the low-frequency gene family might be associated with ancient gene loss and/or recent gene innovation events. Certainly, the actual evolutionary process was much more complicated than theoretically speculated. Furthermore, there were usually different rates of change in gene content among sublineages, consequently the pan-genome might have a corresponding structural difference. An extreme example is that plasmid acquisition would instantaneously increase the gene content of descendants. In prokaryotes, although new genes with novel functions can evolve from extra copies of duplicated genes (Näsvall et al. 2012), the expansion of gene families is mainly driven by horizontal gene transfer (HGT) (Treangen and Rocha 2011). During plasmid acquisition, HGT can effectively increase the genetic diversity of the receptor genome in a short time. Therefore, considering the above-mentioned significant differences in genome size between the symbiotic and nonsymbiotic species, we speculated that the significant differences in genome size and pan-genome architecture are

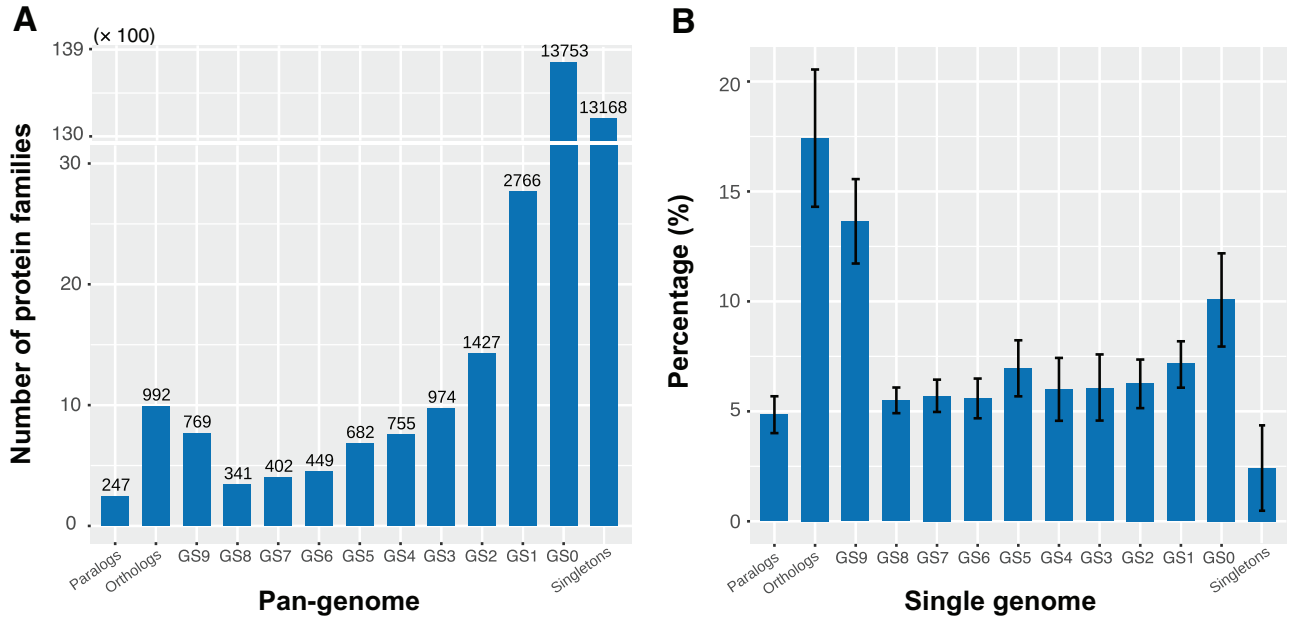

Fig. 1.—Pan-genome structure of family *Rhizobiaceae*. (*A*) Composition of the *Rhizobiaceae* pan-genome. (*B*) Proportion of different protein family categories in a single genome. Bars indicate the SDs of the proportions in the different genomes. GS PFs, species group-specific protein families.

probably related to the differential rates of gene content change among *Rhizobiaceae* sublineages.

Phylogenomics-Based Species Phylogeny

In total, 992 orthologous PFs (with a single copy in each predicated proteome) were identified by OrthoMCL. Every ortholog was aligned, and the resultant ambiguous sites were blocked. Then, 87 orthologous PFs were excluded because their alignments contained <100 sites. Principal component analysis based on the ML (905 × 905 matrix) of each data set (ortholog) to different topologies (trees) revealed that 28 orthologs (points with outlier distances; supplementary fig. S2, Supplementary Material online) had different phylogenetic topologies compared with most of the other orthologs. Detailed information on the phylogenetically incongruent orthologs is supplied in supplementary table S5, Supplementary Material online. An ultimate supermatrix tree, which included 877 orthologs (237,925 amino acids; supplementary table S6, Supplementary Material online) was reconstructed (fig. 2*A*). All 92 genomes could be divided into the following three clades: 1) *Rhizobium* sensu stricto (*Rss*), which included the main symbiotic nitrogen-fixing species within *Rhizobium* represented by the type species *R. leguminosarum*; 2) *Neorhizobium–Agrobacterium–Allorhizobium* (*NAA*), which included members of these three genera and some *Rhizobium* species; and 3) *Ensifer–Pararhizobium–Shinella* (*EPS*), which included the species of these three genera and several *Rhizobium* species. Compared with the previous phylogenomic analysis of family *Rhizobiaceae* (Ormeño-Orrillo et al. 2015), most of the species were successfully separated into the defined genera, except for 22 currently named *Rhizobium* species, five *Allorhizobium* species and three *Neorhizobium* species, which were grouped with members of other genera or formed independent clades. Obviously, further revision of *Rhizobiaceae* taxonomy is needed. In particular, the taxonomic status of genera *Allorhizobium* and *Neorhizobium* should be re-examined.

In the supermatrix tree, 64.4% of the internal nodes were consistent with the gene content's tree (fig. 2*B*). Most of the consistent nodes were near the terminals of the trees. Of the three clades, clade *Rss* had revealed a relatively greater consistency (70.3%) between the two approaches. On the contrary, there were certain changes in the other two clades, which mainly concentrated on branches near to their ancestral nodes. And, most of the inconsistent branches were with low bootstrap support. To a single species, the significant changes occurred in *Pararhizobium haloflavum*, "*Rhizobium rhizosphaerae*," and *Allorhizobium oryzae* in clade *EPS*. This result indicated that the information provided by the gene content was not sufficient to resolve the relationships presented by middle and near-to-root nodes. Alternatively, the complexity of evolutionary events went beyond the resolution of the clustering method based on gene content dissimilarity. Gene content is a comprehensive result of genetic reduction and expansion throughout evolution. Gene loss is a pervasive source of genetic reduction, and exogenous and endogenous gene gain is the main driver of genetic expansion. However, gain and/or loss of the same gene in distant species could also cause the incongruence between supermatrix and gene content trees. In summary, although there was a certain degree of incongruence between the phylogenies based on

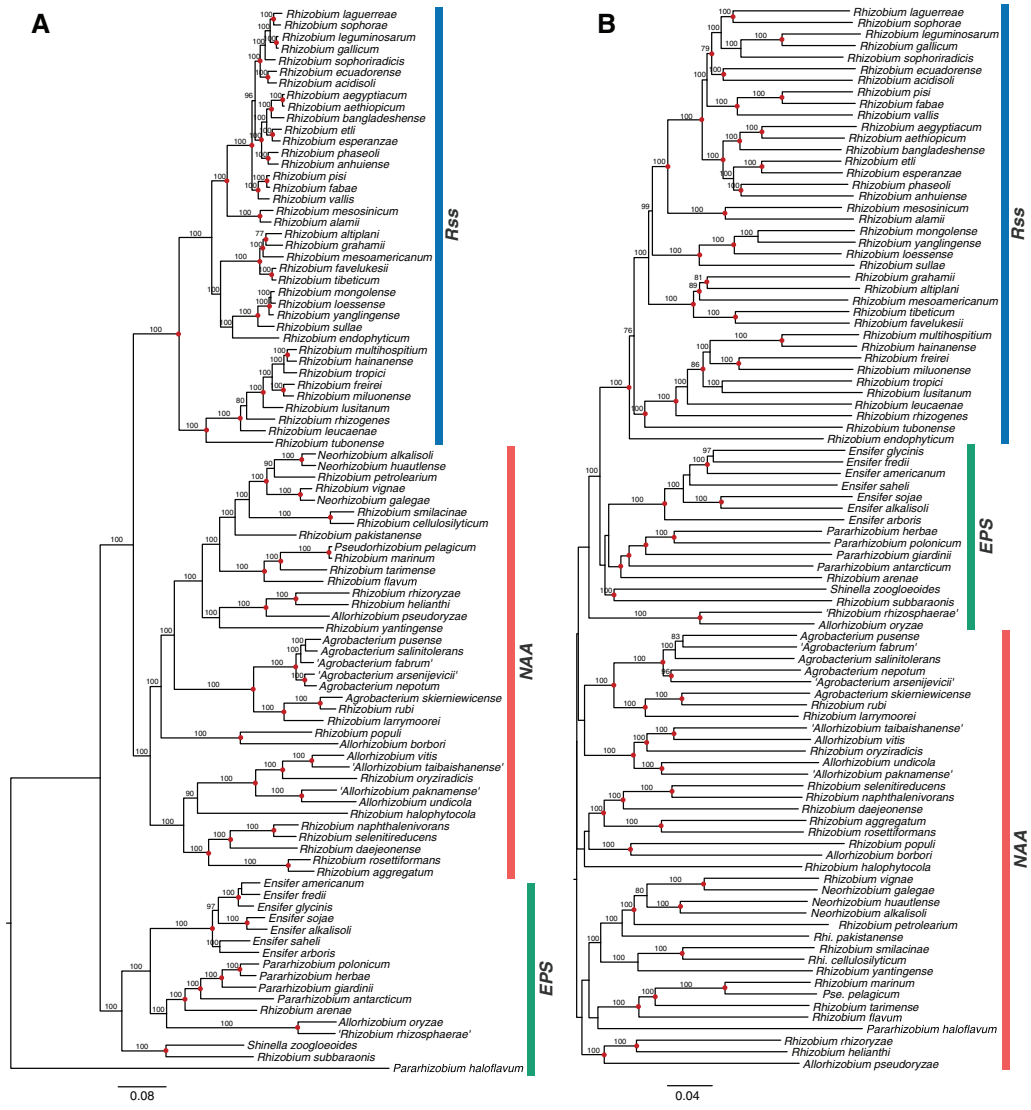

Fig. 2.—Phylogenies of family *Rhizobiaceae* based on the supermatrix method (*A*) and gene content (*B*). These trees were rooted as midpoints. The scale bars denote the number of substitutions per site (*A*), and gene content dissimilarity (*B*). Nonvalidly published species names are in indicated single quotation marks. The nodes (ancestors) that accommodated same species in these two trees are labeled with red circles. *Rss*, *Rhizobium* sensu stricto; *NAA*, *Neorhizobium–Agrobacterium–Allorhizobium*; *EPS*, *Ensifer–Pararhizobium–Shinella*.

concatenated alignment and gene content, three main clades within *Rhizobiaceae* were stable basically.

Gene Family Evolution

The gene family turnover rates were estimated based on an ultrametric tree (fig. 3*A*) built by introducing two divergence time calibration points. The origin and speciation of *Rhizobiaceae* was traced back to at least ~218.4 Ma. Interestingly, the ancestor of clade *Rss* originated between 159.3 and 102.8 Ma, which is earlier than the origin of their host plants in order *Fabales* (75–87 Ma) (Fiz-Palacios et al. 2011; Särkinen et al. 2012) and even earlier than the occurrence of nodulation (~100 Ma) in a common ancestor of nodulating plants (Griesmann 2018; van Velzen et al. 2018). The divergence time of genus *Ensifer* was estimated at ~200 Ma based on a comparison of glutamine synthetase genes (Turner and Young 2000). *Rhizobium* began diversifying in the Cretaceous (145.5–65.5 Ma), with the *Agrobacterium* complex split occurring at ~149–100 Ma (Chriki-Adeeb and Chriki 2016); the estimation based on genomic data in this study is consistent with those results. The ancestor of clade *Rss* started to differentiate during the late Cretaceous to Neogene and might have coevolved with the ancestor of nodulating plants.

In gene family evolution analysis, the best model (Birth, Death, and Innovation–different turnover rates in particular branches–ML) was selected based on AIC (supplementary

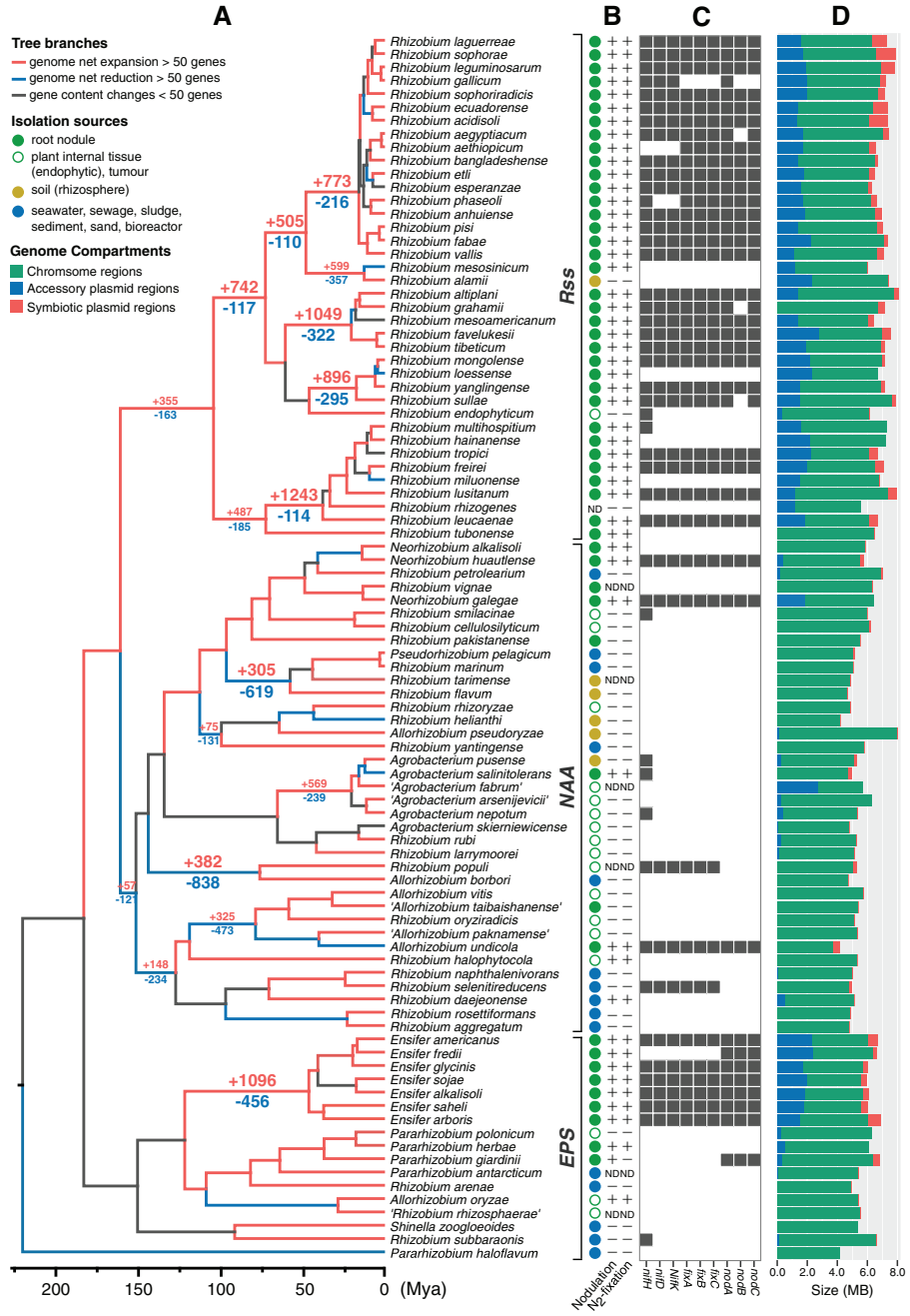

Fig. 3.—Diagram of the correlations between phenotype and genotype for symbiotic nitrogen fixation and the species phylogeny. (*A*) Ultrametric tree that illustrates the divergence times among species. The evaluated numbers of gene gains (red) and losses (blue) are labeled on/under some internal branches. (*B*) Isolation source, and nodulation and nitrogen fixation phenotypes (+, positive; −, negative). (*C*) Key symbiosis-related genes. (*D*) Percentages of chromosomal, accessory plasmid (APR), and symbiotic plasmid (symPR) regions in each genome.

table S7, Supplementary Material online). This model assumed that the three clades (*Rss*, *NAA*, and *EPS*) had different gene family turnover rates, which were then estimated. The gene birth (i.e., gene gain through unequal crossing-over and/or gene duplication) rates of clades *Rss* and *EPS* (3.5×10^{-4} and 1.1×10^{-4}, respectively) were relatively greater than that of clade *NAA* (3×10^{-5}). The family sizes at internal nodes (supplementary fig. S3, Supplementary Material online) and the minimum number of gains/losses in each lineage were also estimated. The changes in gene family sizes at internal nodes are shown in figure 3*A*. Remarkably, several internal nodes underwent genome net expansions (mainly in clades *Rss* and *EPS*) and reductions (mainly in clade *NAA*). Gene family evolution eventually resulted in significant

differences (Student's *t* tests, *P* value <0.05) in the gene contents among different clades: *Rss* ($6{,}774.9 \pm 514.2$ CDSs), *NAA* ($5{,}103.9 \pm 710.8$ CDSs), and *EPS* ($5{,}517.1 \pm 673.7$ CDSs). The gene family turnover rate estimation confirmed that there were differential rates of change in gene content among different sublineages within *Rhizobiaceae*.

Based on these differences and the gene content tree (fig. 2*B*), we hypothesized that the three clades within *Rhizobiaceae* and the genera in each clade underwent different genome expansion processes. In addition, the lower gene contents in the *NAA* and *EPS* clades were consistent with their smaller genome sizes compared with that of clade *Rss*. In total, 92.1% of *Rss*, 18.9% of *NAA*, and 70.7% of *EPS* were the symbiotic species, and average (range) genome sizes of clades *Rss*, *NAA*, and *EPS* were ~6.9 Mb (5.9–7.9), 5.3 Mb (4.1–7.8), and 5.7 Mb (4.1–6.7), respectively (fig. 3). In most cases, the presence of SNF-related genes (*nifHDK*, *fixABC*, and *nodABC*) was correlated with the nodulation and nitrogen fixation capabilities of the strains (fig. 3*B* and *C*). Additionally, the sizes of genomes with SNF-related genes (5.9 ± 0.6 Mb) were significantly (Student's *t* test, *P* value <0.01) larger than that of genomes without SNF-related genes (4.8 ± 0.69 Mb). However, the SNF-related genes were not detected partially or totally in some nitrogen-fixing/nodulating strains (eight and ten strains, respectively), which indicates that some or all of these genes were not covered by the draft genome sequences, or the symbiotic plasmid has been lost in some strains owing to its genetic instability.

Plasmid Sequence Identification and Comparison

In the present study, the reliability of the sequence identification method was verified using the complete genome of *Neorhizobium galegae* HAMBI 540^T as the control. Strain HAMBI 540^T harbors only one megaplasmid (1.81 Mb, Österman et al. 2014), or so-called chromid, that contains SNF-related genes. Here, 9.8% (176 kb, discrete regions) of this megaplasmid matched the referenced accessory plasmids; and 113-kb regions (6.3% of 1.81 Mb) matched the referenced symbiotic plasmids. Therefore, this megaplasmid was classified as an APR, because this megaplasmid had more regions related to accessory plasmid references. Similarly, in the genome of "*Agrobacterium fabrum*" C58, the linear chromosome was also classified as an APR, due to 9.4% of this replicon matched the referenced accessory plasmids and 5.1% regions matched the referenced symbiotic plasmids. Comparison with the complete sequence data (*R. etli* CFN 42^T and *Rhizobium tropici* CIAT 899^T) also showed that this method effectively discriminated APRs from symPRs. Additionally, symPRs were further confirmed by individual identification of the *nodABC*, *nifHDK*, and *fixABC* genes (fig. 3*C*). Figure 3*D* shows that the proportions of chromosomal regions, APRs and symPRs in each genome. APRs and symPRs were detected in most of the symbiotic strains, especially those in the clade *Rss*. Only symPRs were detected in the genomes of *Rhizobium grahamii*, *Allorhizobium undicola*, and *Agrobacterium salinitolerans*. Conversely, only APRs were detected in the genomes of *Rhizobium mesosinicum*, *Rhizobium alamii*, *Rhizobium loessense*, *R. endophyticum*, *Rhizobium multihospitium*, *Rhizobium hainanense*, *Rhizobium miluonense*, and *Rhizobium rhizogenes* in clade *Rss*. In the *Rhizobium tubonense* genome, neither an APR nor symPR were identified. Among the strains in which we failed to detect a symPR, *R. alamii*, *R. endophyticum*, and *R. rhizogenes* do not have a symbiotic phenotype. The absence of symPRs in other *Rhizobium* strains may result from missing data in the draft sequences or loss of the plasmid.

Previously, through pan-genome analysis and gene family turnover rate estimation, we found that the high proportion of low-frequency flexible gene families might be directly related to the differential rates of gene content change. However, although plasmid acquisition is an effective way for genome expansion, genetic innovation in chromosomes can also lead to increased proportions of low-frequency gene families. Therefore, to determine the contribution of different genome components to the pan-genome architecture, we separated the PFs that can appear on different genome components (i.e., chromosome regions, APRs, and symPRs) (fig. 4). As expected, although the plasmid system plays an obvious role in genome expansion, the gene innovation in chromosomes cannot be ignored. Moreover, the effect of APRs on genome expansion was more profound than that of symPRs. APRs can even accommodate the genes of gene families near the genome core. Furthermore, the ancestral state reconstruction of gene content demonstrated that APRs gave rise to more genome expansion by introducing more genes than symPRs and chromosome, at the common ancestors of symbiotic species (fig. 5). Additionally, the ancestral nodes in the ultrametric tree were also more affected by APRs than symPRs. This finding indicates that the introduction of accessory plasmids might predate that of symbiotic plasmids in *Rhizobiaceae* and supports the hypothesis that accessory plasmids coevolved with chromosome long before the introduction of symbiotic plasmids (Harrison et al. 2010; Harrison and Brockhurst 2012; diCenzo and Finan 2017).

To reveal the evolutionary relationships among the chromosome regions, APRs, and symPRs, 35 strains (28 from clade *Rss* in *Rhizobium* and seven from *Ensifer*) positive for nodulation and nitrogen fixation were subjected to the ANI value calculation and clustering analysis (fig. 6). The clustering relationships expressed by the ANI based on the whole genome (fig. 6*A*), chromosome regions (fig. 6*B*), and APRs (fig. 6*C*) were similar. Intriguingly, the clustering relationships of these 35 genomes based on the symPRs (fig. 6*D*) were entirely different from those based on the whole genome, chromosome

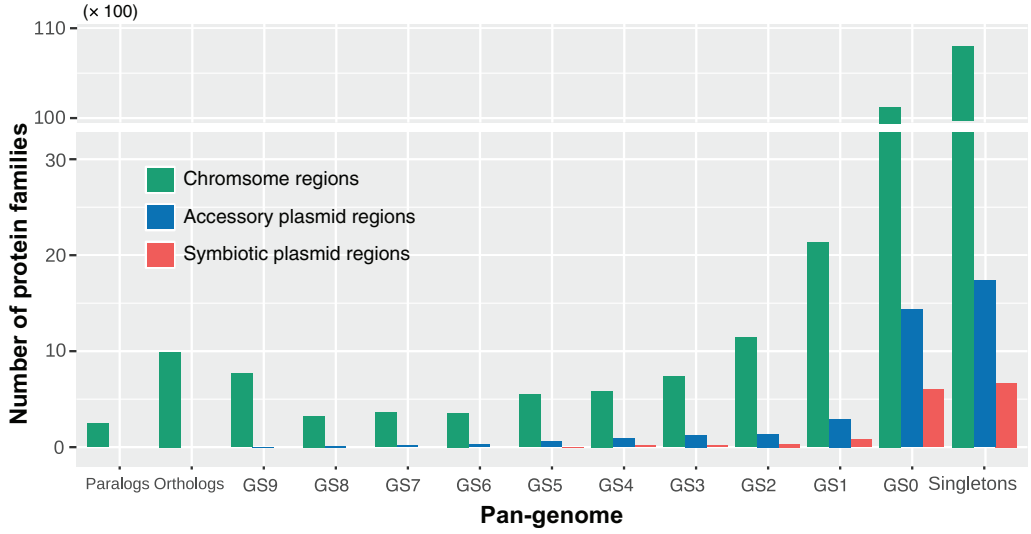

Fig. 4.—Composition of family *Rhizobiaceae* pan-genome based on separated protein families separated according to the genome components where they can be found.

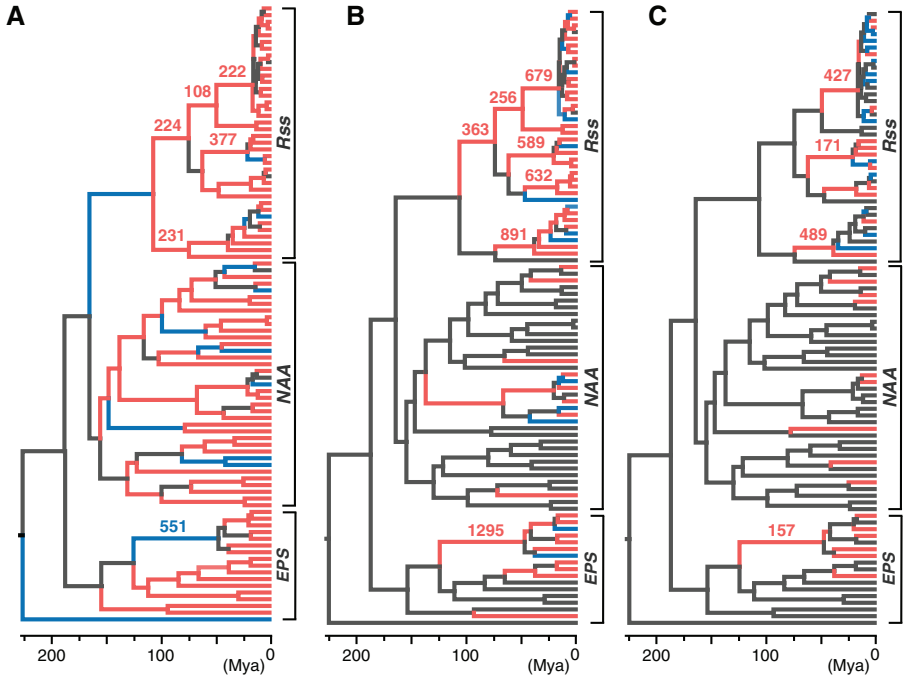

Fig. 5.—Ancestral state reconstruction of gene content based on the gene families that can be found in chromosome regions (*A*), accessory plasmid regions (*B*), and symbiotic plasmid regions (*C*). Red branches, genome net expansion >50 genes; blue branches, genome net reduction >50 genes; gray branches, gene content changes <50 genes. *Rss*, *Rhizobium* sensu stricto; *NAA*, *Neorhizobium–Agrobacterium–Allorhizobium*; *EPS*, *Ensifer–Pararhizobium–Shinella*.

regions, and APRs, which were consistent with the phylogenetic relationships estimated from the proteins FixABC, NifKDH, and NodABC (supplementary fig. S4, Supplementary Material online). The subclade that included 17 very closely related *Rhizobium* species (fig. 6*A*–*C*) was split into two apparent clustering groups that should include several different kinds of symbiotic plasmids (fig. 6*D*). If the highly conserved symPRs in different species are the result of HGT, then HGT did not result in a certain type of symbiotic plasmid that spread throughout the family. Our findings indicated that the genetic diversity of symPRs was not dramatically reduced by HGT, among these type strains of *Rhizobiaceae*, even between very closely related species, such as *Rhizobium aegyptiacum* and *R. aethiopicum*, and *Ensifer americanum* and

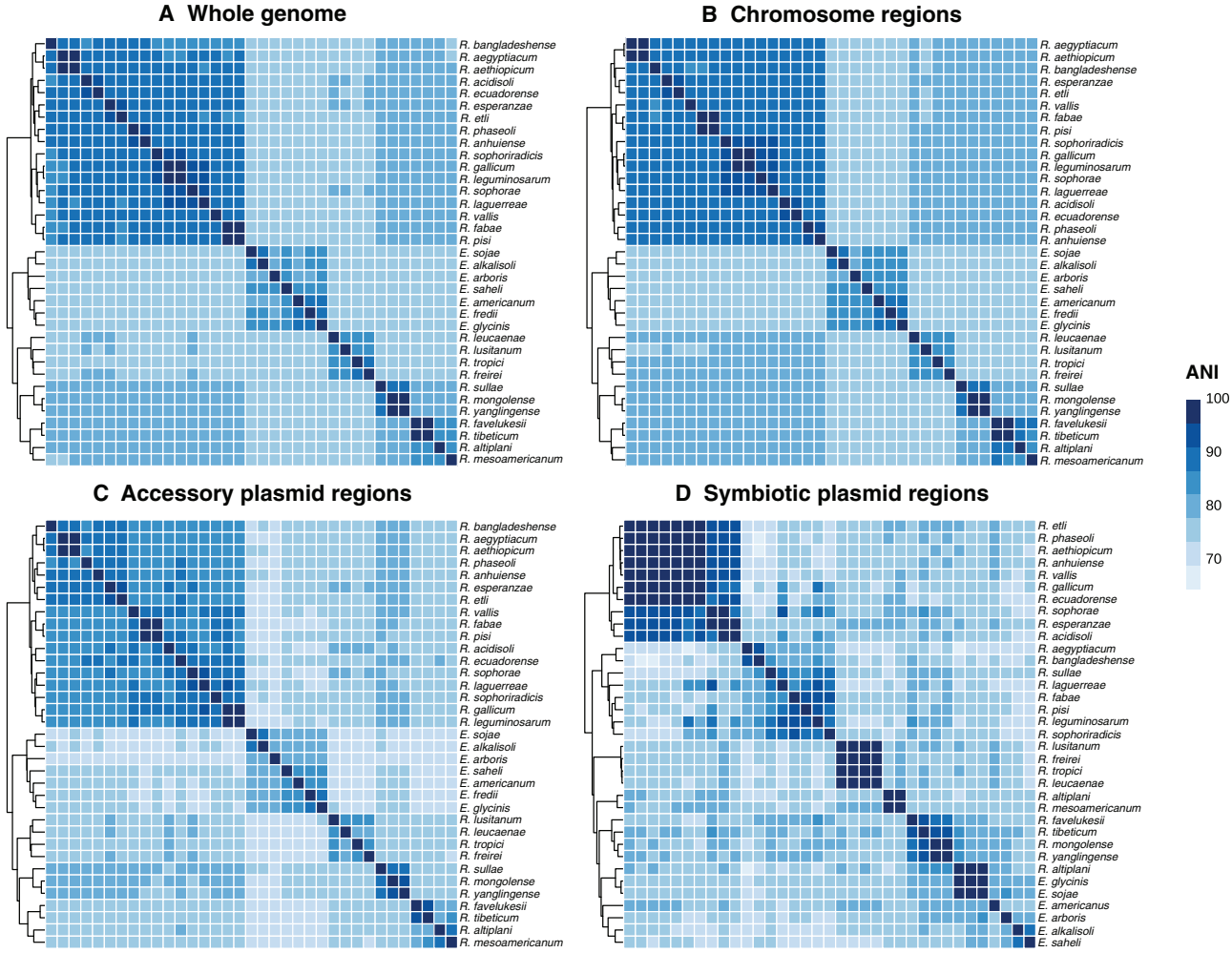

Fig. 6.—Evaluation and clustering of average nucleotide identity (ANI) values within genome compartments. ANI values from 60% to 100% are indicated by the changing intensity of the blue coloration (light to dark).

Ensifer fredii (figs. 2*A* and 6*D*). Moreover, the nucleotide diversities of symPRs gene families were clearly different from those of chromosome and APRs gene families (supplementary fig. S5, Supplementary Material online).

In general, the transfer of symbiosis-related genes often occurs within species and between closely related species. Successful transfer depends on both environmental and genetic factors, including the availability of symbiosis-related genes in the microenvironment and the compatibility of the recipient's genomic background (Remigi et al. 2016). Triple selection based on soil conditions, plant host, and rhizobial species results in the followings: 1) the same rhizobial species may harbor different types of symbiotic plasmids that allow the nodulation of different hosts, such as *R. leguminosarum* symbiovars viciae and trifolii (supplementary fig. S6, Supplementary Material online); 2) various rhizobial species may harbor the same type of symbiosis-related genes that allow nodulation of the same host (fig. 6*D* and supplementary fig. S6, Supplementary Material online); and 3) different species or genera of rhizobia may harbor divergent symbiosis-related genes that allow nodulation of the same host in distinct regions, such as the common bean-nodulating *Rhizobium* and *Ensifer* (supplementary fig. S6, Supplementary Material online). These various combinations demonstrated that the HGT of symbiosis-related genes was influenced by comprehensive factors. However, the species that shared highly conserved symPRs were from the same isolation source (supplementary fig. S7, Supplementary Material online). Therefore, the host's physical isolation and/or soil conditions might impede HGT among phylogenetically close species. In addition to horizontal transfer (species to species), the vertical evolution (generation to generation) of symbiotic plasmids, including symbiosis-related genes, was also observed (fig. 6*D* and supplementary fig. S5, Supplementary Material online, Zhang et al. 2011).

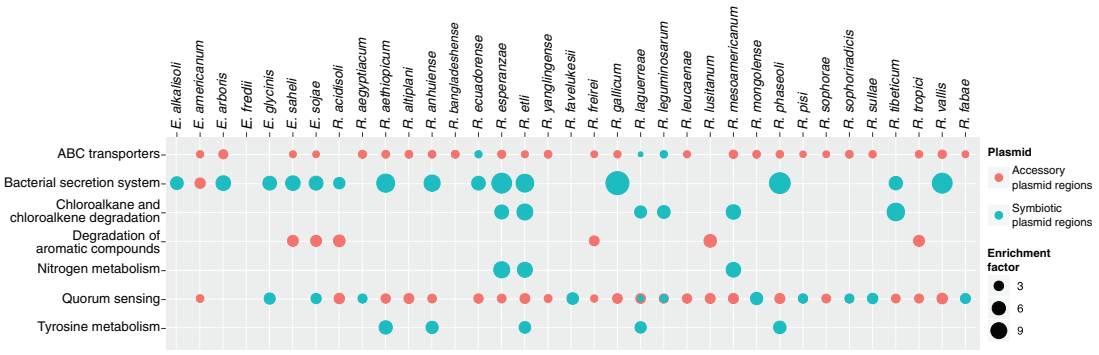

Fig. 7.—Pathways significantly (corrected *P* value <0.05) enriched in accessory plasmid and symbiotic plasmid regions.

Functional Enrichment Analysis

The experimental genome reduction of *Sinorhizobium meliloti* (*Ensifer meliloti*) revealed that the nonchromosomal replicons carry genes with more specialized functions, such as growth in the rhizosphere and interactions with plants (diCenzo et al. 2014). Functional biases exist between each compartment in a multipartite genome (diCenzo and Finan 2017). For testing if the functional bias is common among species with nonchromosomal replicons, we performed a metabolic pathway enrichment analysis based on the above-mentioned 35 symbiotic species, and found that the ATP-binding cassette (ABC) transporter was significantly enriched (corrected *P* value <0.05) in the APRs of most tested genomes, the bacterial secretion system was a significantly enriched pathway in symPRs, and quorum sensing was significantly enriched in APRs and/or symPRs (fig. 7). Interestingly, species with an enriched quorum-sensing system in symPRs included *R. aegyptiacum*, *Rhizobium laguerreae*, *R. leguminosarum*, *R. pisi*, *Rhizobium sophoriradicis*, and *Rhizobium fabae*. This further indicates that phylogenetically closely related species can carry different types of symbiotic plasmids.

The ABC transporter substrates include mineral and organic ions, oligosaccharides, monosaccharides, phosphate, and amino acids (supplementary fig. S8*A*, Supplementary Material online). The transport of such materials from the host-associated habitat might be crucial for rhizobial symbiosis, nodulation, or nitrogen fixation (Ding et al. 2012; Garcia-Fraile et al. 2015; Cheng et al. 2016). *Rhizobium*–legume symbiotic interactions are complex processes that involve several of signal transduction pathways. Many rhizobial species use complex N-acyl-homoserine lactone-based quorum-sensing systems to monitor their population densities and regulate their symbiotic interactions with their plant hosts (González and Marketon 2003; Schmeisser et al. 2009; Palmer et al. 2016). Unlike ABC transporter and quorum-sensing systems (supplementary fig. S8*B*, Supplementary Material online), bacterial secretion systems were mainly enriched in symPRs (supplementary fig. S8*C*, Supplementary Material online). Although the rhizobial strains could possess a remarkable number of secretion systems (Schmeisser et al. 2009), only T3SS and T4SS were enriched. T3SS, T4SS, and a portion of the nodulation region are located in the highly variable regions of symbiotic plasmids (Pérez Carrascal et al. 2016). T3SS is involved in the delivery of effectors associated with virulence from bacterial cells to eukaryotic cells, whereas T4SS participates in the conjugative transfer of plasmids or protein export (Deakin and Broughton 2009; Wang et al. 2012).

The successful completion of SNF requires the mobilization of multiple physiological mechanisms in rhizobia and their legume hosts. These mechanisms would have evolved over time. For fast-growing legume-nodulating rhizobia, although symbiosis-related genes are located on the symbiotic plasmid, other accessory plasmids are equally important (Barreto et al. 2012; Stasiak et al. 2014; Price et al. 2015; Zahran 2017). The functional enrichment analysis at the family level confirmed that transport systems, signal transduction (including quorum-sensing), and bacterial secretion systems are crucial for SNF; they work with SNF-related genes to accomplish biological nitrogen fixation. Based on the functional division of accessory plasmid and symbiotic plasmids, and their sequential appearance in the evolution of *Rhizobiaceae*, we believed that the role of accessory plasmid(s) in the evolution of SNF cannot be underestimated.

Conclusions

Our evidence supports that the divergence of symbiotic species within family *Rhizobiaceae* occurred prior to the origin of their host plants, and even earlier than the occurrence of nodulation; this provided the foundation for the hypothesis that the nitrogen-fixing symbiosis arose from polyphyletic origins and convergently evolved within family *Rhizobiaceae*. In this study, we collectively analyzed the pan-genome of *Rhizobiaceae* and definitely found that, although the genetic expansions in chromosomal regions were pervasive within this family, gene gain events associated with accessory plasmids brought more genes into the genomes of symbiotic nitrogen-fixing species. Although HGT reduced the genetic diversity of symbiotic plasmids to some extent, the transfer was probably impeded by nonbiological factors like host's physical isolation,

even among phylogenetically close species. The plasmid system, which includes accessory and symbiotic plasmids, may have evolved over a time span in rhizobial species with the ability to adapt to the various environmental conditions and helped them achieve nitrogen fixation. Therefore, in family *Rhizobiaceae*, the multipartite genome is most likely the result of adaptation to different environmental niches under the synergistic effects of nodulating plants.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31500007 and 31560309).

Literature Cited

Barreto EF, Straliotto R, Baldani JI. 2012. Curing of a non-symbiotic plasmid of the *Rhizobium tropici* strain CIAT 899 affected nodule occupancy and competitiveness of the bacteria in symbiosis with common beans. Eur J Soil Biol. 50:91–96.

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18(1):1–5.

Brochier C, Philippe H. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature 417(6886):244–244.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Cheng G, Karunakaran R, East AK, Poole PS. 2016. Multiplicity of sulfate and molybdate transporters and their role in nitrogen fixation in *Rhizobium leguminosarum* bv. *viciae* Rlv3841. Mol Plant Microbe Interact. 29(2):143–152.

Chriki-Adeeb R, Chriki A. 2016. Estimating divergence times and substitution rates in Rhizobia. Evol Bioinform Online. 12:87–97.

Deakin WJ, Broughton WJ. 2009. Symbiotic use of pathogenic strategies: rhizobial protein secretion systems. Nat Rev Microbiol. 7(4):312–320.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6(5):361–375.

diCenzo G, Milunovic B, Cheng J, Finan TM. 2013. The $tRNA^{arg}$ gene and *engA* are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. J Bacteriol. 195(2):202–212.

diCenzo GC, et al. 2016. Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. Nat Commun. 7(1):12219.

diCenzo GC, Finan TM. 2017. The divided bacterial genome: structure, function, and evolution. Microbiol Mol Biol Rev. 81(3):e00019–17.

diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. 2014. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet. 10(10):e1004742.

diCenzo GC, Mengoni A, Perrin E. 2019. Chromids aid genome expansion and functional diversification in the family *Burkholderiaceae*. Mol Biol Evol. 36(3):562–574.

Ding H, Yip CB, Geddes BA, Oresnik IJ, Hynes MF. 2012. Glycerol utilization by *Rhizobium leguminosarum* requires an ABC transporter and affects competition for nodulation. Microbiology 158(5):1369–1378.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575–1584.

Ferguson BJ, et al. 2010. Molecular analysis of legume nodule development and autoregulation. J Integr Plant Biol. 52(1):61–76.

Fiz-Palacios O, Schneider H, Heinrichs J, Savolainen V. 2011. Diversification of land plants: insights from a family-level phylogenetic analysis. BMC Evol Biol. 11(1):341.

Garcia-Fraile P, et al. 2015. Arabinose and protocatechuate catabolism genes are important for growth of *Rhizobium leguminosarum* biovar *viciae* in the pea rhizosphere. Plant Soil 390:251–264.

González JE, Marketon MM. 2003. Quorum sensing in nitrogen-fixing rhizobia. Microbiol Mol Biol Rev. 67(4):574–592.

Griesmann M. 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science 361:eaat1743.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.

Guo HJ, et al. 2014. Replicon-dependent differentiation of symbiosis-related genes in *Sinorhizobium* strains nodulating *Glycine max*. Appl Environ Microbiol. 80(4):1245–1255.

Harrison E, Brockhurst MA. 2012. Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. 20(6):262–267.

Harrison PW, Lower RP, Kim NK, Young JP. 2010. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends Microbiol. 18(4):141–148.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 9(1):5114.

Ji KX, et al. 2010. Movement of rhizobia inside tobacco and lifestyle alternation from endophytes to free-living rhizobia on leaves. J Microbiol Biotechnol. 20(2):238–244.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45(D1):D353–D361.

Kumar N, et al. 2015. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. Open Biol. 5(1):140133.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34(7):1812–1819.

Lapierre P, Gogarten JP. 2009. Estimating the size of the bacterial pan-genome. Trends Genet. 25(3):107–110.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.

Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28(2):279–281.

Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1(1):18.

Markowitz VM, et al. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. 40(D1):D115–D122.

Masson-Boivin C, Sachs JL. 2018. Symbiotic nitrogen fixation by rhizobia-the roots of a success story. Curr Opin Plant Biol. 44:7–15.

Mergaert P, et al. 2006. Eukaryotic control on bacterial cell cycle and differentiation in the *Rhizobium*-legume symbiosis. Proc Natl Acad Sci U S A. 103(13):5230–5235.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24(3):319–324.

Nadarajah KK. 2017. Rhizobium in rice yield and growth enhancement. In: Hansen A, Choudhary D, Agrawal P, Varma A, editors. Vol. 50. Rhizobium biology and biotechnology. Soil biology. Switzerland: Springer Cham. p. 83–103.

Näsvall J, Sun L, Roth JR, Andersson DI. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. Science 338(6105):384–387.
Oldroyd GE, Murray J, Poole PS, Downie JA. 2011. The rules of engagement in the legume-rhizobial symbiosis. Annu Rev Genet. 45(1):119–144.
Ormeño-Orrillo E, et al. 2015. Taxonomy of rhizobia and agrobacteria from the *Rhizobiaceae* family in light of genomics. Syst Appl Microbiol. 38(4):287–291.
Österman J, et al. 2014. Genome sequencing of two *Neorhizobium galegae* strains reveals a *noeT* gene responsible for the unusual acetylation of the nodulation factors. BMC Genomics 15(1):500.
Palmer AG, et al. 2016. Interkingdom responses to bacterial quorum sensing signals regulate frequency and rate of nodulation in legume-Rhizobia symbiosis. Chembiochem 17(22):2199–2205.
Pérez Carrascal OM, et al. 2016. Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing *Rhizobium* species associated with *Phaseolus vulgaris*. Environ Microbiol. 18(8):2660–2676.
Poole P, Ramachandran V, Terpolilli J. 2018. Rhizobia: from saprophytes to endosymbionts. Nat Rev Microbiol. 16(5):291–303.
Price PA, et al. 2015. Rhizobial peptidase HrrP cleaves host-encoded signaling peptides and mediates symbiotic compatibility. Proc Natl Acad Sci U S A. 112(49):15244–15249.
Ramachandran VK, East AK, Karunakaran R, Downie JA, Poole PS. 2011. Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. Genome Biol. 12(10):R106.
Remigi P, Zhu J, Young JPW, Masson-Boivin C. 2016. Symbiosis within symbiosis: evolving nitrogen-fixing legume symbionts. Trends Microbiol. 24(1):63–75.
Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19(2):301–302.
Särkinen T, Pennington T, Lavin M, Simon MF, Hughes CE. 2012. Evolutionary islands in the Andes: persistence and isolation explain high endemism in Andean dry tropical forests. J Biogeogr. 39(5):884–900.
Schmeisser C, et al. 2009. *Rhizobium* sp. strain NGR234 possesses a remarkable number of secretion systems. Appl Environ Microbiol. 75(12):4035–4045.
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14):2068–2069.
Shamseldin A, Abdelkhalek A, Sadowsky MJ. 2017. Recent changes to the classification of symbiotic, nitrogen-fixing, legume-associating bacteria: a review. Symbiosis 71(2):91–109.
Stasiak G, et al. 2014. Functional relationships between plasmids and their significance for metabolism and symbiotic performance of *Rhizobium leguminosarum* bv. *trifolii*. J Appl Genet. 55(4):515–527.
Treangen TJ, Rocha EP. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7(1):e1001284.
Turner SL, Young JP. 2000. The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. Mol Biol Evol. 17(2):309–319.
Udvardi M, Poole PS. 2013. Transport and metabolism in legume-rhizobia symbioses. Annu Rev Plant Biol. 64(1):781–805.
van Velzen R, et al. 2018. Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. Proc Natl Acad Sci U S A. 115(20):E4700–E4709.
Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. 2005. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21(11):2791–2793.
Wang D, Yang S, Tang F, Zhu H. 2012. Symbiosis specificity in the legume: rhizobial mutualism. Cell Microbiol. 14(3):334–342.
Wang X, et al. 2018. Comparative analysis of rhizobial chromosomes and plasmids to estimate their evolutionary relationships. Plasmid 96–97:13–24.
Xie C, et al. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 39(Suppl 2):W316–W322.
Zahran HH. 2017. Plasmids impact on rhizobia-legumes symbiosis in diverse environments. Symbiosis 73(2):75–91.
Zhang YM, et al. 2011. Biodiversity and biogeography of rhizobia associated with soybean plants grown in the North China Plain. Appl Environ Microbiol. 77(18):6331–6342.
Zhi XY, Jiang Z, Yang LL, Huang Y. 2017. The underlying mechanisms of genetic innovation and speciation in the family *Corynebacteriaceae*: a phylogenomics approach. Mol Phylogenet Evol. 107:246–255.

Associate editor: Tal Dagan