environmental microbiology

Environmental Microbiology (2019) **00**(00), 00–00 doi:10.1111/1462-2920.14762

Exploring the evolutionary dynamics of *Rhizobium* plasmids through bipartite network analysis

Xiangchen Li,[1,2†] Hao Wang ,[1,2†] Wenjun Tong,[1] Li Feng,[3] Lina Wang,[1,2] Siddiq Ur. Rahman,[1,2,4] Gehong Wei[1*] and Shiheng Tao[1,2*]

[1]*State Key Laboratory of Crop Stress Biology in Arid Areas, Shaanxi Key Laboratory of Agricultural and Environmental Microbiology, College of Life Sciences, Northwest A&F University, Yangling, Shaanxi, 712100, China.*

[2]*Bioinformatics Center, Northwest A&F University, Yangling, Shaanxi, 712100, China.*

[3]*College of Enology, Northwest A&F University, Yangling, Shaanxi, 712100, China.*

[4]*Department of Computer Science and Bioinformatics, Khushal Khan Khattak University, Karak, Khyber Pakhtunkhwa, 27200, Pakistan.*

Summary

The genus *Rhizobium* usually has a multipartite genome architecture with a chromosome and several plasmids, making these bacteria a perfect candidate for plasmid biology studies. As there are no universally shared genes among typical plasmids, network analyses can complement traditional phylogenetics in a broad-scale study of plasmid evolution. Here, we present an exhaustive analysis of 216 plasmids from 49 complete genomes of *Rhizobium* by constructing a bipartite network that consists of two classes of nodes, the plasmids and homologous protein families that connect them. Dissection of the network using a hierarchical clustering strategy reveals extensive variety, with 34 homologous plasmid clusters. Four large clusters including one cluster of symbiotic plasmids and two clusters of chromids carrying some truly essential genes are widely distributed among *Rhizobium*. In contrast, the other clusters are quite small and rare. Symbiotic clusters and rare accessory clusters are exogenetic and do not appear to have co-evolved with the common accessory clusters; the latter ones have a large coding potential and functional complementarity for different lifestyles in *Rhizobium*. The bipartite network also provides preliminary evidence of *Rhizobium* plasmid variation and formation including genetic exchange, plasmid fusion and fission, exogenetic plasmid transfer, host plant selection, and environmental adaptation.

Received 25 February, 2019; revised 24 June, 2019; accepted 25 July, 2019. *For correspondence. E-mail weigehong@nwafu.edu.cn. E-mail shihengt@nwafu.edu.cn; Tel. (+86) 29 87091060; Fax. (+86) 29 87091060. †These two authors contributed equally to this work.

Introduction

The genus *Rhizobium* is a well-known group of rhizobia that have established two distinct lifestyles in nature, as free-living soil bacteria or as symbionts of legume plants (Masson-Boivin *et al*., 2009). Most of the *Rhizobium* genomes contain multiple replicons: one indispensable chromosome plus a variable number of non-dispensable or dispensable extra-chromosomal plasmids (González *et al*., 2006; Mazur and Koper, 2012). Some megaplasmids denoted as 'chromids' for carrying some essential (core) genes have been observed in several *Rhizobium* species, e.g., p42e and p42f in *Rhizobium etli*. CFN42 (Harrison *et al*., 2010). The plasmids are effective components because they are independently replicated based on the *repABC* module (Cevallos *et al*., 2008). Owing to their separate replication genes and mobilization ability, the plasmids contribute significantly to the dynamics of the host genome (Cevallos *et al*., 2002; Palacios and Flores, 2005; Ding *et al*., 2013). Furthermore, the plasmids are crucial mediators of horizontal gene transfer (HGT), permitting spread of symbiotic ability (Li *et al*., 2018). The prominent characters of plasmids, such as plasticity, dynamics and diversity, owing to transposition, HGT, plasmid-DNA recombination and genome rearrangement are well documented (Freiberg *et al*., 1997; Zhang *et al*., 2001; González, 2003; Brom *et al*., 2004; López-Guerrero *et al*., 2012; Pérez Carrascal *et al*., 2016). For example, the genome of *Rhizobium* sp. NGR234 (also called *Sinorhizobium fredii* NGR234) undergoes large-scale DNA rearrangements involving replicon fusions and excisions, promoted mainly by homologous recombination between insertion sequence elements (Mavingui *et al*., 2002). It is proposed that the non-random organization of bacterial genomes is shaped by selective pressures to facilitate host interactions and

niche adaptation (diCenzo *et al*., 2016). However, the evidence to support such a hypothesis in rhizobia is lacking (MacLean and San Millan, 2015). Hence, *Rhizobium* is a good candidate for studies on the function and evolution of multipartite bacterial genomes.

The plasmids in rhizobia are typically divided into two main groups: the symbiotic plasmid (pSym) carrying the symbiosis module in which genes essential for nodulation, infection, and nitrogen fixation are clustered, and the non-symbiotic/accessory plasmids (Remigi *et al*., 2016). Unlike the pSym encoding for symbiotic activity, the accessory plasmids enable the host rhizobia to survive in diverse habitats and under stress conditions (diCenzo *et al*., 2014; Zahran, 2017). Comparatively, the accessory plasmids have gained limited attention and experimental investigations have been difficult to perform because of their high variability and ability to easily exchange genetic material between and within species (Dresler-Nurmi *et al*., 2009). Although some studies have been conducted comparing the rhizobial replicons based on their genomic content, traditional technologies of phylogenetics and phylogenomics have limited applicability in the study of the evolutionary relationships and boundaries of these coexisting plasmids due to an absence of universal plasmid genes (González *et al*., 2006, 2019; Orlandini *et al*., 2014; Pérez Carrascal *et al*., 2016; Orlek *et al*., 2017).

Owing to the expansion of network-oriented representation of sequence similarity, graph theory measures have been applied to better describe the gene flow across diverse microbial communities, paving the way for large-scale comparative analyses (Tamminen *et al*., 2011; Orlandini *et al*., 2014; Corel *et al*., 2016). Reticulated evolution is increasingly studied using sequence similarity networks (Yamashita *et al*., 2014; Fondi *et al*., 2016; Méheust *et al*., 2016; Iranzo *et al*., 2017; Bernard *et al*., 2018). Recently, a gene families-genomes bipartite network was introduced for visualization analysis of microbiome data to uncover genome–proteome relationships associated with the adaptive attributes of single strains and/or specialized populations (Jaffe *et al*., 2016; Sedlar *et al*., 2016; Lanza *et al*., 2017). Bipartite networks could facilitate the analysis of large collections of plasmids, which is especially useful in the comprehensive characterization of conjugative elements on plasmids or integrative conjugative elements (including surveillance of antibiotic resistance, highly pathogenic species, and characterization of emerging lineages) (Lanza *et al*., 2017; Corel *et al*., 2018). Consequently, the ever-increasing diversity of network tools provides an accurate multilevel framework to study the 'web of life' (Soucy *et al*., 2015). In this context, an operational plasmid classification system could generalize and analyse the most relevant aspects of plasmid biology, such as host range, transfer rates, and functional interactions; this would permit overlap with species boundaries (Fernandez-Lopez *et al*., 2017).

Given the growing number of complete genomes and network tools, a large-scale comparative genomic analysis of plasmids is feasible. Here, we focused on the extrachromosomal replicons of available complete genomes of *Rhizobium*. By adopting a bipartite network analysis, we determined the complex evolution of 216 *Rhizobium* plasmids and subsequently investigated their intrinsically genomic, genetic and functional characteristics. We obtained evidence of *Rhizobium* plasmid variation and formation including (i) plasmid replication and conjugation, (ii) plasmid fusion and fission, (iii) genetic diversity, recombination, and codon usage bias features of plasmids, and (iv) host-plant specificity and functional enrichment of plasmids. Together, these provide a deep insight into the evolution and function of these multipartite genomes.

Results

An overview of Rhizobium plasmids

A total of 216 plasmids were retrieved from 49 complete genomes of *Rhizobium* (Table S1). The *Rhizobium* strains were mainly recovered from the root nodules of legume plants, except for *Rhizobium* sp. NT-26 from an arsenic-containing gold mine and *Rhizobium phaseoli* R744 from the rhizosphere of *Phaseolus vulgaris*. Both of these strains lacked the pSyms and were consequently non-symbiotic. Based on a threshold of 95% average nucleotide identity, all 49 strains were divided into 19 genospecies (Fig. S1). Most of the genomes were composed of a single circular chromosome and several plasmids (mean: 4.4 ± 1.2), except for *Rhizobium* sp. IRBG74 containing a circular chromosome, a linear chromosome, and a pSym. The remaining 46 genomes were found to each comprise a single pSym. For convenience, we renamed all plasmids by combining the original GenBank definitions of both strains and plasmids (Table S1). For example, CFN42_p42d_pSym denotes the symbiotic plasmid p42d in *R. etli* CFN42.

According to their annotation files, all plasmids encoded a total of 101,206 proteins. Some plasmids were very small, with the smallest being BIHB1148_pSK06, which had a molecular size of 12 kb and carried 19 genes. In contrast, some other plasmids were considerably large, such as four plasmids with a molecular size of up to 2 Mb and carrying more than 2000 genes. The largest plasmid, R602_pRgaIR602c carrying 2339 genes, was nearly 57% of its associated chromosome. In addition, the plasmids were found to have lower coding densities than the chromosomes (895 genes/Mb vs. 939 genes/Mb; Wilcoxon test; $P < 0.001$). Larger plasmids showed lower coding density than smaller ones (Spearman's ρ between gene

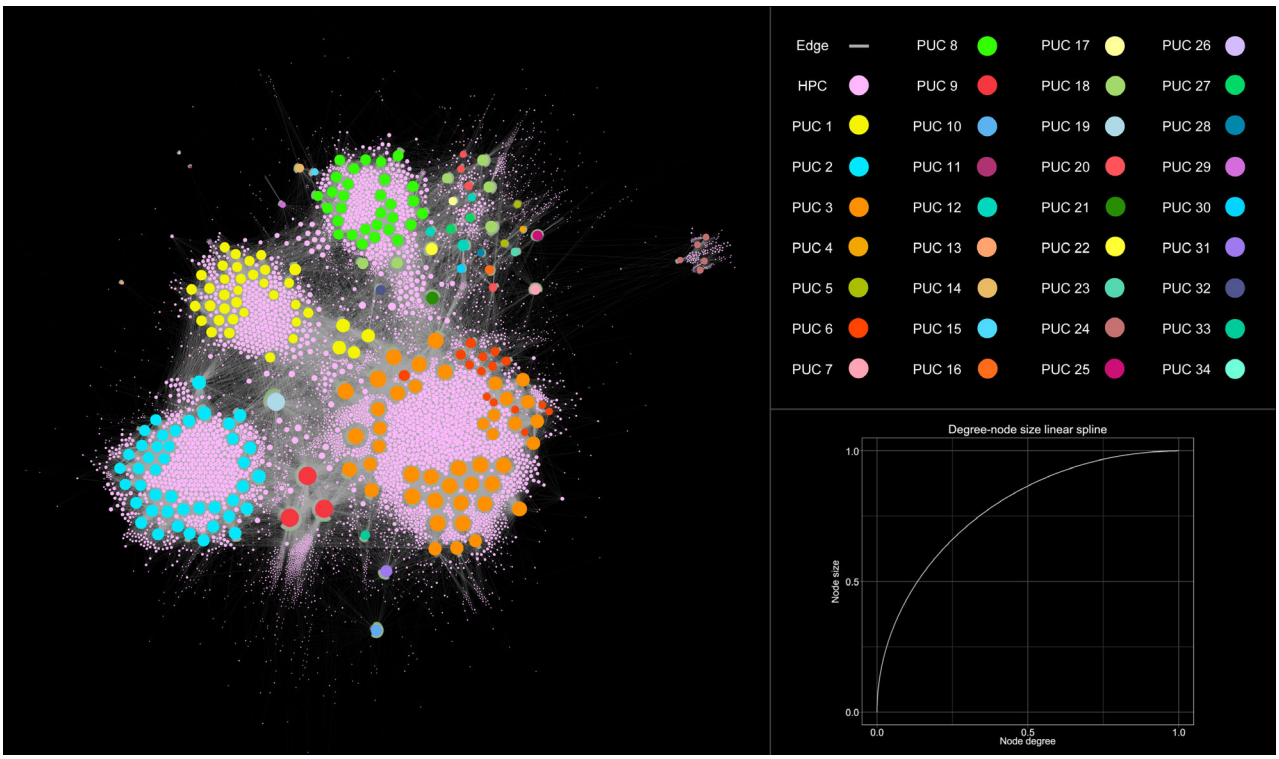

Fig. 1. The PU-HPC bipartite network visualization of the *Rhizobium* plasmids' pan-genome. Plasmid units (PUs, 216) and homologous protein clusters (HPCs, 13,057) are represented as coloured circles. The size of a circle is ranked as the degree followed by a non-linear spline (bottom right) that changes node size more rapidly in the lower ranges and less so as the nodes approach the maximum degree in the graph. The colour of a PU node is automatically set according to its PU cluster (PUC) assignment (top right). PU-specific HPCs are shown in light green, while other HPCs are shown in pink. Edge-weight values of the HPCs are normalized to the phylogenetic distance and shown in grey (70% transparency).

density and plasmid size = 0.45; $P = 0.02$), but no such trend was found in the main chromosomes ($\rho = 0.07$; $P = 0.63$). These observations suggest that the coding potential of the *Rhizobium* plasmids is both large and diverse.

The bipartite network of plasmid-protein families

To develop a network representation of the evolutionary structure among all *Rhizobium* plasmids, we clustered their annotated proteins into families of homologues by sequence similarity. The bipartite network consisted of two classes of nodes: 216 plasmid units (PUs, the vehicles or containers of genes) and 13,057 homologous protein clusters (HPCs, protein families according to amino acid sequence identity, coverage, and E-value). Edges connected every PU with the HPC that it contained. Edge-weight values reflecting the phylogenetic distance of the corresponding protein in each PU and the consensus HPC were used to construct the layout. The result was a connected PU-HPC bipartite network characterizing the whole pan-genome of plasmid sequences: PUs were connected only through HPCs, while different HPCs were connected through PUs in which they were jointly represented. The eventual layout of the bipartite network reflected a balanced state of the mechanical system (Fig. 1). The network directly showed that (i) most of the PUs gathered into several separate regions and formed the main structure of the network, in addition to some peripheral PUs and (ii) the HPCs reflected a complex pan-genome of *Rhizobium* plasmids.

A standard topological analysis of the bipartite network showed that the degree distribution of HPCs approximately followed a continuous power law distribution, with the fewest nodes having the greatest number of connections (Fig. S2). Notably, the maximum degree was 165 indicating that there were no universally shared HPCs among all 216 PUs. Additionally, the degree distribution of PUs revealed that four mega-plasmids (R602_pRgalR602c, IE4872_pRgalIE4872d, 8C3_pRsp8C3c, and CIAT899_pRtrCIAT899c) had extremely high degrees, reflecting prominent hubs on the network. The squared clustering coefficient (SCC) distribution of PUs was characterized by multiple peaks, indicating a hierarchical structure of the PU organization. Thus, the bipartite network appeared to be held together primarily by both hallmark HPCs and giant PUs. The degree and SCC values of all PUs had a non-significant negative correlation (Spearman's $\rho = -0.13$, $P = 0.053$). Such a negative correlation may be due to some PUs containing many unique HPCs, which largely increased the degrees but had little influence

upon the SCCs. In contrast, there was a significant positive correlation between the degrees and SCCs of all HPCs (Spearman's $\rho = 0.86$, $P < 0.001$). Together, the PU-HPC bipartite network represented a complex evolutionary structure of the *Rhizobium* plasmids.

Clustering homologous PUs on the bipartite network

To unveil the evolutionary relationships of all *Rhizobium* plasmids, we applied a hierarchical clustering analysis based on the network adjacency matrix. A total of 34 homologous PU clusters (PUCs) were retrieved at an 85% distance threshold (Fig. 1; Table S2). On the PU-HPC bipartite network, we found a high modularity score of 0.814 with a resolution of 1.0 and a total of 32 modular classes, which revealed an overall coherence with the network clustering result (Fig. S3).

We verified that the clustering results were almost consistent at different amino acid sequence identities (Table S3). Besides, as large plasmids could tend to cluster and to connect directly together, due to their large number of genes, it is necessary to measure its influence on the clustering. The jackknife resampling analyses showed that the PUC assignments were robust with respect to the size of the plasmid genome (Table S4). To test the potential clustering confusion by the HGT of the plasmids, we also performed the same network analysis at 95% and 99% sequence similarity cut offs, leading to a characterization of the putative recent HGTs. We found that the only 15% and 10% of HGTs occurred across different PUCs at 95% and 99% cut offs respectively, indicating little influence of these events upon network clustering (Fig. S4). Hence, the network clustering result is robust in reflecting the evolutionary relationships of the plasmids.

We found that 10 PUCs were shared by multiple genospecies while the other 24 PUCs were genospecies/strain specific, reflecting a complex evolutionary structure of the *Rhizobium* plasmids relative to the associated chromosomes (Table S5). Moreover, the PUs in each strain were entirely separated into different PUCs (Table 1). Notably, there were only five large PUCs (1, 2, 3, 6 and 8), each comprising more than 10 PUs. These large PUCs accounted for 79% of all PUs and connected to 58% of all HPCs. Both PUCs 9 and 19 seemed to be hubs for PUCs 1–3. According to the knowledge of the well-studied *R. etli* CFN42, there are some key biological/genetic features of the large PUCs: CFN42_p42c (PUC 1) is reported to be required for an efficient nodule occupancy; CFN42_p42e (PUC 2) and CFN42_p42f (PUC 3) are chromids; CFN42_p42b (PUC 6) encodes LPS O-antigen biosynthesis genes for successfully establishing symbiosis with *P. vulgaris*; CFN42_p42d_pSym (PUC 8) is a typical pSym for the symbiosis with *P. vulgaris* (known features of all 34 PUCs are listed in Table S2) (Brom *et al*., 1992; Harrison *et al*., 2010). Additionally, 20 PUCs harbouring only one PU were sparsely distributed on the periphery of the network. In contrast, some PUCs, such as PUCs 10 and 24, were extremely distant to the others. Based on the network clustering, a total of 26 different PUC compositions were found among the 49 strains (Table 1). Despite these distinct PUC compositions, we found an apparent universality where the four common PUCs (1–3 and 8) coexisted in 39 strains.

Table 1. Plasmid unit cluster (PUC) composition of the 49 *Rhizobium* strains used in this study.

Strain abbreviation	PUC composition	Genospecies
IE4771	1, 2, 3, 18*, 20	3
R744	1, 2, 3, 20	14
3841	1*, 2, 3, 4, 5, 6	8
BIHB1217	1, 2, 3, 5, 12*, 13	12
BIHB1148	1, 2, 3, 6, 10, 11	10
Vaf108	1, 2, 3, 6, 12*, 27, 28, 29	10
Vaf10	1, 2, 3, 6, 12*, 27, 30	10
NXC12	1, 2, 3, 6, 18*	1
Mim1	1, 2, 3, 6, 22*, 23	1
WSM1325	1, 2, 3*, 6, 31	11
WSM1689	1, 2, 3, 6, 32*	9
N1314, N324, N731, TAL182	1, 2, 3, 6, 8*	7, 13
CFN42	1, 2, 3, 6, 8*, 15	1
Bra5, IE4803, Kim5, N113, N161, N261, N561, N621, N6212, N831, N871, N931, R611, R620, R630, R650, R723	1, 2, 3, 8*	3, 13, 14
N1341, N671, N741, N771, N841	1, 2, 3, 8*, 24	13, 14
CB782	2, 3, 14*	4
NXC14	2, 3, 18*	2
CIAT894	2, 3, 4, 6, 8*, 16	6
WSM2304	2, 3*, 6, 33	5
CIAT652	2, 3, 8*	14
8C3	7, 8*, 9	19
R602	9, 18*, 20	15
IE4872	9, 18*, 20, 34	15
CIAT899	17, 18*, 19	16
IRBG74	21*	18
NT26	25, 26	17

Symbiotic plasmids are marked with an asterisk.

The 46 symbiotic PUs on the bipartite network were clustered into nine PUCs (1, 3, 8, 12, 14, 18, 21, 22 and 32). Although most of the symbiotic PUs were clustered into their exclusive PUCs, only three symbiotic PUs (3841_pRL10_pSym, WSM1325_pR132501_pSym and WSM2304_pRLG201_pSym) were unexpectedly shared in mixed PUCs 1 and 3 since most PUs in these two PUCs were non-symbiotic. The comparison of nine representative symbiotic PUs in the above PUCs revealed that their genomic backbones were obviously different, only except some regions containing symbiosis genes for nodulation, infection and nitrogen fixation (*nod*/*nif*/*fix*) were found in all PUs (Fig. 2). Notably, the symbiotic PUs in

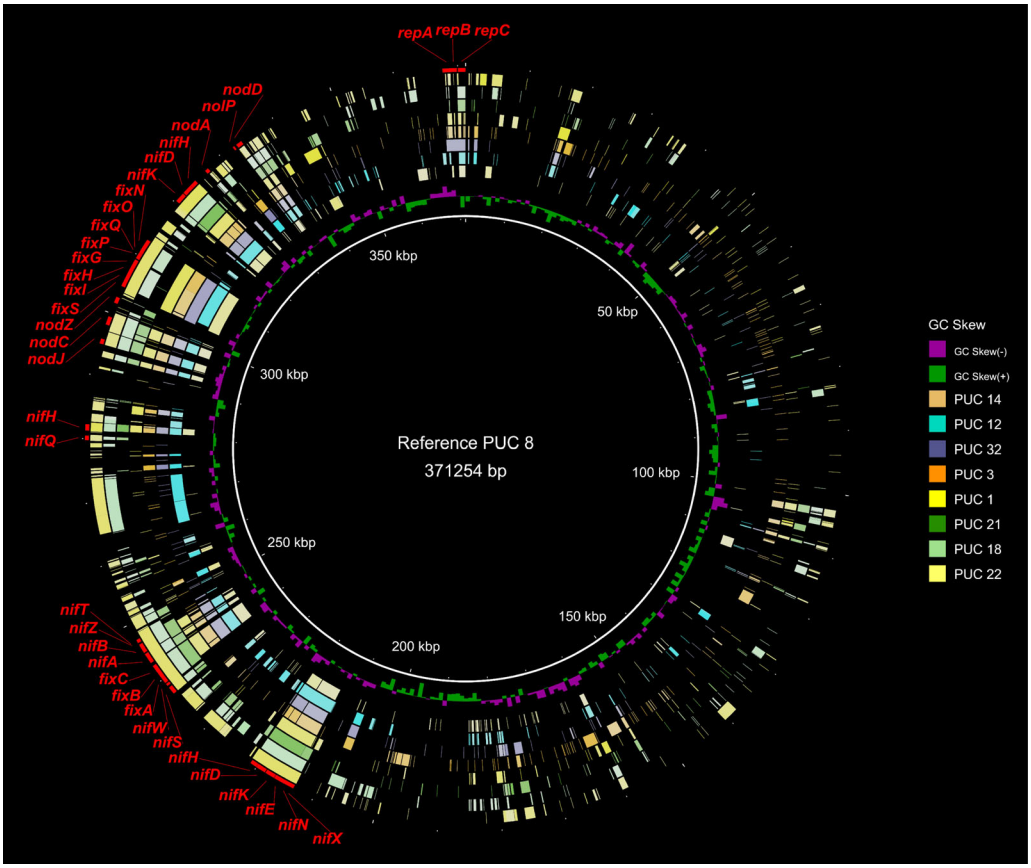

Fig. 2. Genome comparison of nine representative pSyms in different PUCs. The inner ring denotes the reference CFN42_p42d_pSym in PUC 8 with corresponding genetic coordinates. From inner to outer ring: CB782_unnamed3_pSym in PUC 14, Vaf10_unnamed3_pSym in PUC 12, WSM1689_unnamed3_pSym in PUC 32, WSM2304_pRLG201_pSym in PUC 3, 3841_pRL10_pSym in PUC 1, IRBG74_III_pSym in PUC 21, CIAT899_pRtrCIAT899b_pSym in PUC 18 and Mim1_pRetMIM1e_pSym in PUC 22. Regions in the query plasmids that are also present in the reference plasmid are displayed. The colour of each ring accords with the assignment in Fig. 1.

PUCs 1 and 3 were formed by the integration of symbiosis genes with respect to their homologous PUs expanding the clusters of the *Rhizobium* pSyms (Fig. S5).

The rare PUCs appeared to be important in the remarkable increase in plasmid complexity and they all connected to a large number of specific HPCs. To find possible exogenetic sources of the rare PUCs, we constructed sequence alignments of the HPCs connecting to all PUCs on the network. We found that most of the small PUCs contained abundant homologous genes from other closely related bacterial genera (Table S6). For instance, more than 70% of the HPCs connecting to PUCs 19–21 had homologues among the genus *Sinorhizobium* or *Agrobacterium*. In stark contrast, the proportions of external homologous genes in the common PUCs were all less than 10%. Moreover, only 3% of the genes in PUC 10 had exogenetic homologues. To further explore the source of the rare PUCs, we compared all plasmids with the chromosomes by constructing their bipartite network. We observed that PUC 10 was strongly connected to the chromosomes (Fig. S6). Although it has a close relationship with chromosomes, it lacks integrase, type IV secretion system and type IV secretion system coupling protein indicating it is not the integrative and conjugative element (Wozniak and Waldor, 2010).

Less effective traditional phylogenetic-based classifications compared with bipartite network analysis

According to genome annotations, a total of 209 PUs were found to harbour the complete *repABC* operon and thus could be regarded as self-replicating plasmids (Fig. S7A). Each phylogram of the *repABC* operon was generally consistent with the network clustering result (Fig. S8). However, 57 PUs contained multiple (complete or partial) *repABC* regions (Table S7), and these PUs were mainly found in PUC 3, indicating the presence of several incompatibility groups in this PUC. In addition, the *repC* genes on four mega-plasmids showed a different phylogeny compared with their associated *repA* and *repB*. These findings suggest that the phylogenetic analysis based on the *repABC* operon is not sufficient to reveal the evolutionary relationships of the *Rhizobium* plasmids.

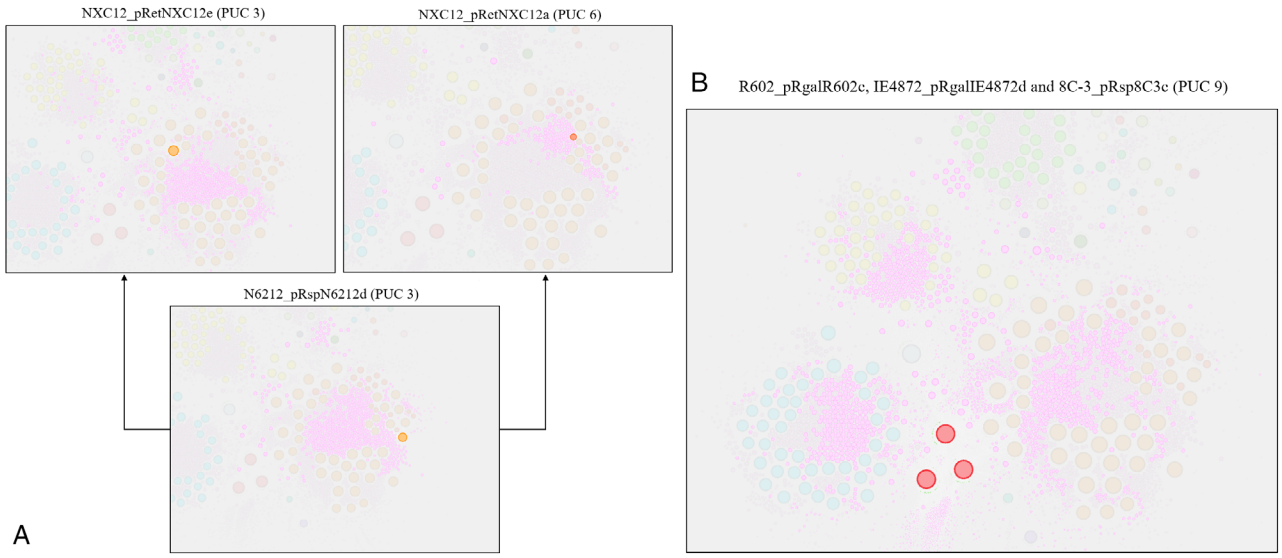

Fig. 3. Typical cases of plasmid fusion and fission on the bipartite network of all *Rhizobium* plasmids. *A*. A small PU in PUC 3 and its coexisting PU in PUC 6, for example, NXC12_pRetNXC12e (PUC 3) and NXC12_pRetNXC12a (PUC6) were formed by a fission event of a common larger plasmid in PUC 3, such as N6212_pRspN6212d. The three PUs and the HPCs they connect to are highlighted. Proteins coded by the two smaller PUs are merely intersected and they unite to consist of the entire proteome coded by the larger plasmid. *B*. Plasmids R602_pRgalR602c, IE4872_pRgalIE4872, and 8C3_pRsp8C3c in PUC 9 were formed by the fusion event(s) from PUCs 1–3.

Apart from the *repABC* operon, we also classified the *Rhizobium* plasmids based on their conjugative relaxases, the proteins required to initiate plasmid mobilization through conjugation. We retrieved a total of 100 PUs containing six types of conjugative relaxases (Table S8). Phylogenetic analysis of the conjugative relaxases revealed an overall consensus with the network clustering result (Fig. S9). However, this approach omitted the remaining 116 non-mobilizable PUs. Moreover, there were conflicts between the mobility and PUCs of some PUs. Furthermore, we assessed the mobility of the *Rhizobium* plasmids by identifying three components of a conjugative apparatus: a conjugative relaxase (R), a type IV coupling protein (T4CP) and a type IV secretion system (T4SS). Eventually, 85 conjugative PUs and 15 mobilizable PUs were assigned (Fig. S7B). Notably, most of the pSyms and the PUs on the peripheral network were conjugative, in contrast to the common accessory PUCs (1–3 and 6) that were rarely conjugative or mobilizable.

Evidence of plasmid fission and fusion from the bipartite network

The network perspective revealed the occurrence of large genetic exchange events between some PUs. It is worth noting that PUC 6 was closely connected to PUC 3 and also assigned to the same modular class with PUC 3. Each PU in PUC 6 had a coexisting PU in PUC 3, which was clearly smaller than other common PUs in PUC 3. The two coexisting PUs had an approximate mean complementary composition of 60% for all HPCs connected to PUC 3 (Table S9). The maximum complementary proportion was 80.6% for NXC12_pRetNXC12e (PUC 3) plus NXC12_pRetNXC12a (PUC 6) compared with N6212_pRspN6212d (PUC 3; Fig. 3A). In addition, the above-mentioned phylograms of the *repABC* operons showed that PUC 6 was close to a PUC 3 clade, indicating that these two PUCs were evolutionarily related. Therefore, it would appear that PUC 6 originated from fission of PUC 3.

PUCs 9 and 19 included the four mega-plasmids and seemed to be hubs in connecting PUCs 1–3 and 6. In total, 51.8% and 49.6% of the HPCs connecting to PUCs 9 and 19, separately, were also shared by PUCs 1–3 and 6 (Fig. 3B). In addition, a syntenic comparison of the samples in these PUCs showed that PUCs 1–3 had an obvious partial synteny with PUCs 9 and 19 (Fig. S10). In contrast, HPCs 24118 and 36504 were members of the ATP-binding cassette transporter system and had considerably high degrees in the central position of the network (Fig. S11), shared by most of PUCs 1–3, 9 and 19. Notably, there were multiple copies of both HPCs connecting to the PUs in PUC 9. Phylogenetic analysis revealed that the copies of the two HPCs were separated into three clades reflecting PUCs 1–3 respectively (Fig. S12). Together, these results indicate that both PUCs 9 and 19 were formed by the fusion of PUCs 1–3. Furthermore, the phenomenon of plasmid fusion was also observed for some large PUs in the other central areas of the network. For example, four PUs in PUC 1 (IE4771_pRetIE4771e, BIHB1217_pPR3, Kim5_pRetKim5d and IE4803_pRetIE4803d) were fused by the PUs in both PUCs 1 and 6; two PUs in PUC 3 (CB782_unnamed1 and WSM2304_pRLG201_pSym)

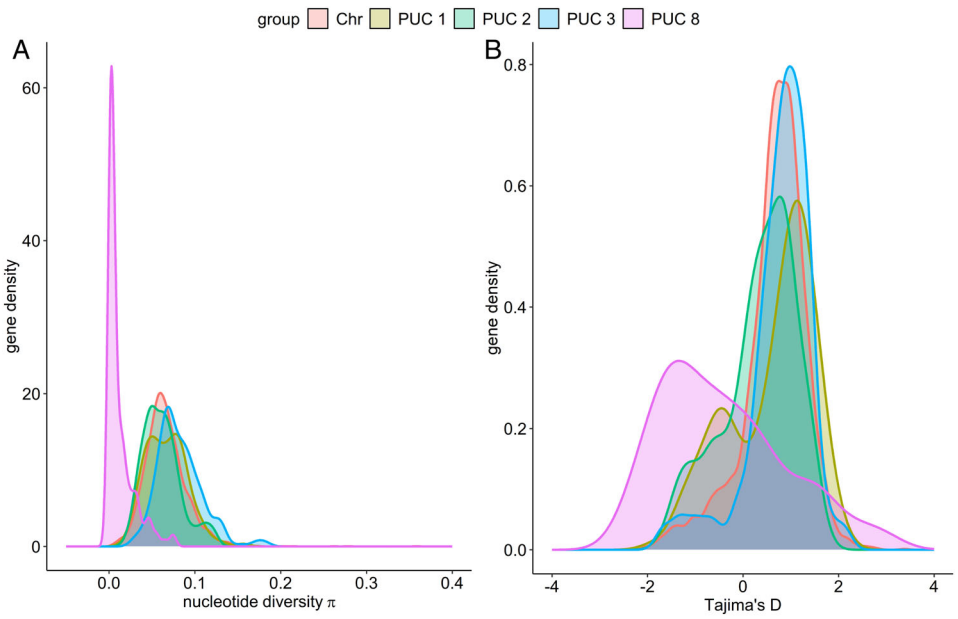

Fig. 4. Population genetics of the common PUCs (1, 2, 3 and 8) and the main chromosomes (Chr).
A. Distribution of nucleotide diversity (π).
B. Distribution of Tajima's *D* values. The plots show the genetic density for PUC 1 (136 genes), PUC 2 (247 genes), PUC 3 (210 genes), PUC 8 (139 genes) and the Chr (3006 genes).

were fused by the PUs in both PUCs 1 and 3; and one PU in PUC 2 (NXC14_pRspNXC14b) was fused by the PUs in both PUCs 1 and 2.

Different evolutionary characteristics of PUCs

To discover the evolutionary characteristics of different PUCs, especially the common PUCs, we selected 27 strains carrying the plasmids in PUCs 1–3 and 8 together as a genetic population for subsequent analyses. To obtain a more accurate measure of the degree of genetic diversity in the four PUCs and their associated chromosomes, nucleotide diversity per site (π) and Tajima's *D* values were calculated. We found that PUCs 1–3 all had similar π distributions to the chromosomes, whereas PUC 8 had a lower genetic diversity (Fig. 4A). Tajima's *D* values showed a similar distribution pattern with π values. In particular, PUC 8 had a peak *D* value around −2.0, indicating that this PUC may be restricted in its genetic diversity due to sweeping selection (Fig. 4B).

To investigate the extent of recombination in the four PUCs with their associated chromosomes, we estimated the relative effect of recombination versus mutation (*r*/*m*) and the ratio of recombination to mutation rate (ρ/θ). In general, the *r*/*m* estimates were higher within the four PUCs than the chromosomes (Table 2). These estimates indicated that recombination resulted in over twice as many substitutions per event on average than mutation in the chromosomes, while the estimates were much higher for PUCs 1 and 8. In addition, the ρ/θ estimates within the four PUCs ranged from 0.1 to 0.3, indicating that the rate of mutation was predominant over that of recombination.

Next, we calculated relative synonymous codon usage (RSCU) of all PUCs and the chromosomes to measure their codon usage bias differences. The dendrogram based on Spearman's correlational distance of RSCU values revealed two major clades (Fig. S13). The larger clade mainly consisted of the accessory PUCs and the chromosomes. Notably, 151 PUs on the network had an RSCU value close to that of the chromosomes.

Table 2. Population genetic statistics of the common PUCs (1, 2, 3 and 8) and the main chromosomes (Chr).

Group	ρ/θ	δ (bp)	ν	*r*/*m*	π	Tajima's *D*
PUC 1	0.278	288	0.0522	4.18	0.0657	0.627
PUC 2	0.233	234	0.0475	2.60	0.0627	0.450
PUC 3	0.155	260	0.0590	2.38	0.0814	0.866
PUC 8	0.260	565	0.0321	4.71	0.0110	−0.159
Chr	0.175	201	0.0505	1.77	0.0667	0.747

ρ/θ: The ratio of rates of recombination and mutation.
δ: The average length of recombined fragments.
υ: The mean divergence of imported DNA.
r/*m*: The ratio of effects of recombination and mutation.
π: Nucleotide diversity.

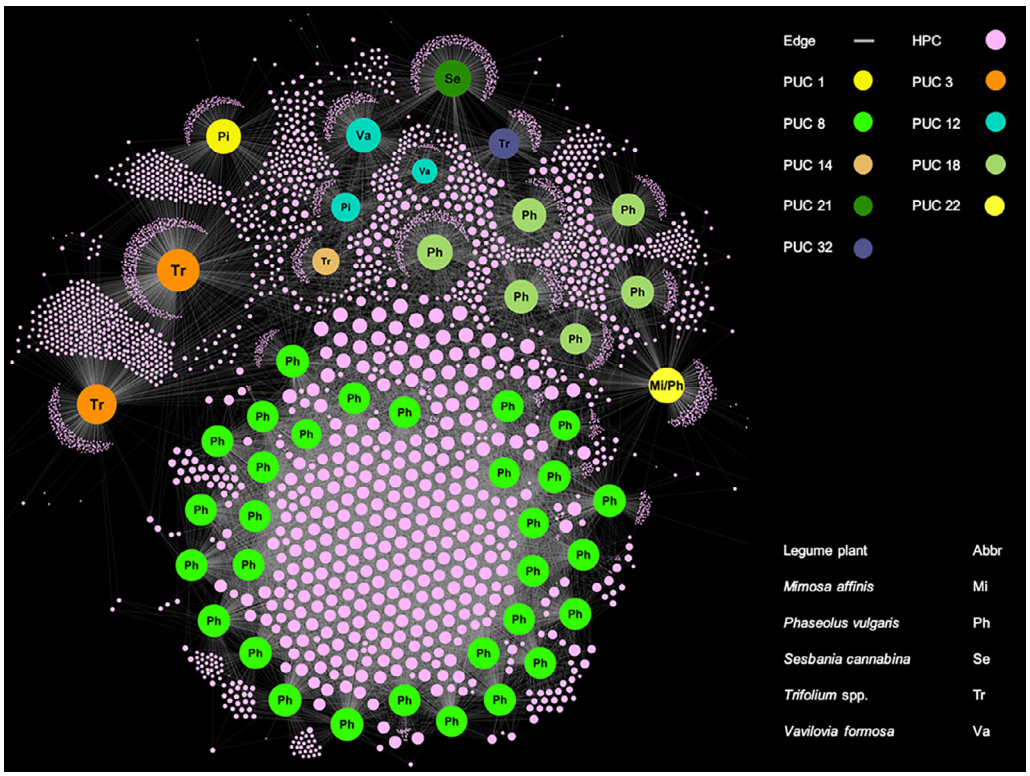

Fig. 5. The bipartite network of the symbiotic plasmids with their host legume plants. This network is derived from the main PU-HPC bipartite network (Fig. 1) by removing the non-symbiotic plasmids and adding the host plant information to the remaining PU nodes.

Additionally, 93 of these PUs among PUC 1–3, 9, 10 and 19 were regarded as 'chromids' based on their clearly distinct characteristics from chromosomes and plasmids, including plasmid-type maintenance and replication systems, genomic sizes, G + C contents and encoding core genes (see details in Experimental procedures, Table S10 and Fig. S14). In contrast, the smaller clade comprised most of the pSyms and some peripheral PUs on the network.

Higher host-specificity of symbiotic PUCs compared with accessory ones

As many PUCs have been recovered, it is of interest to determine whether the host plants have a selection effect on the clustering result. Thus, the bipartite network was modified by adding the host plant information to all PUs. This modified network showed that several accessory PUCs, such as PUCs 1–3, were independent of host plant selection (Fig. S15). However, a few symbiotic PUCs, such as the largest symbiotic PUC 8, showed a strong specificity to *Phaseolus vulgaris*, the most common host-plant for the *Rhizobium*.

Next, we reconstructed the bipartite network by removing all accessory PUs. The resulting network showed a complex diversity and specificity of the pSyms to five legume plants (Fig. 5). Despite PUCs 18 and 22 also being related to *P. vulgaris-Rhizobium* symbiosis, PUC 22 was not closely related to the other two PUCs (8 and 18) and showed a broad symbiosis with both *P. vulgaris* and *Mimosa affinis*. Taken together, these three PUCs (8, 18 and 22) revealed three forms of the pSyms for *P. vulgaris-Rhizobium* symbiosis. Similarly, the three pSyms in PUC 12 had a broad host range, such as *Vavilovia formosa* and *Pisum sativum*. In addition, 3841_pRL10_pSym in PUC 1 and BIHB1217_pPR4_pSym in PUC 12 were also represented by two forms of the pSyms for *P. sativum-Rhizobium* symbiosis. In contrast, four pSyms for *Trifolium-Rhizobium* symbiosis were found in three different PUCs (3, 14 and 32), separately. These pSyms may have evolved to suit different *Trifolium* species and the symbiotic genes were even integrated into common non-pSym PUCs such as PUC 3. Such a phenomenon was also observed for 3841_pRL10_pSym in PUC 1. Moreover, IRBG74_III_pSym in PUC 21 was distantly related to all other pSyms and specific for *Sesbania cannabina-Rhizobium* symbiosis.

Functional bias and complementarity of PUCs

The functional enrichment of all PUCs was analysed based on Clusters of Orthologous Groups (COG) annotations (Fig. S16). The four common PUCs retrieved the most COG categories compared with other PUCs in accordance

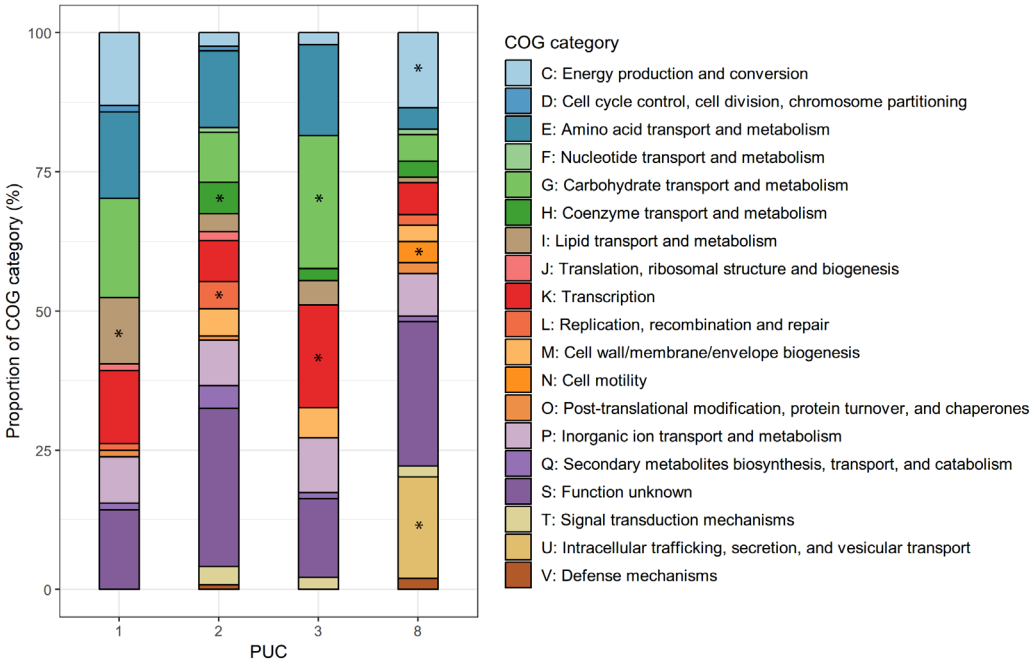

Fig. 6. Relative abundance of protein identifications in the common PUCs by Clusters of Orthologous Groups (COG) functional category. Stacked bar chart shows the proportion of total protein identifications in each PUC for general COG categories. Each significantly enriched COG category (Fisher's exact test, $P < 0.05$) is marked with an asterisk.

with their genomic sizes. The HPCs connecting to PUC 8 were found to have significantly more enriched COG categories than PUCs 1–3 (Fig. 6; Fisher's exact test; $P < 0.05$), while each PUC had distinct enriched COG categories. These results suggest that the common PUCs in a *Rhizobium* multipartite genome have a strong functional bias. According to the COG enrichment results, PUCs 1–3 may improve the capability of the host cell in utilizing lipids, coenzymes and carbohydrates respectively. Moreover, the COG category K (Transcription) was found to be significantly enriched in PUC 3 because many HPCs connecting to this PUC were transcriptional regulators. According to the 'chromid' results, PUCs 2 and 3 are clearly important in the host genome. For example: CFN42_p42f (PUC 3) encodes *panBC* genes (3-methyl-2-oxobutanoate hydroxymethyltransferase and pantoate-beta-alanine ligase respectively) which are both essential for synthesis of pantothenate (Villaseñor *et al*., 2011), and two core genes *groES* and *groEL* (molecular chaperone); CFN42_p42e (PUC 2) encodes an essential *panB* gene, in addition to three core genes, *actP* (copper-translocating P-type ATPase), *pcaB* (3-carboxy-*cis*,*cis*-muconate cycloisomerase) and *folD* (methylenetetrahydrofolate dehydrogenase).

To further determine the PUCs involved in symbiosis, we conducted sequence similarity searches of all HPCs in the NodMutDB database. A total of 149 HPCs were identified as having an influence on *Rhizobium*-legume symbiosis. By calculating the frequency of these HPCs in each PUC (Table S11), we found that in addition to the symbiotic PUCs, other accessory PUCs also encoded several symbiosis-related genes. Indeed, PUCs 1 and 3 had the greatest influence on symbiosis compared with PUC 2 and other small accessory PUCs. In summary, the four common PUCs showed both functional bias and complementarity in the two distinct lifestyles of *Rhizobium*.

Discussion

Although rhizobial plasmids can vary among different strains, it is often challenging to accurately demonstrate meaningful classification due to the lack of universal genes or universal marker genes suitable for traditional phylogenetic approaches. Given an increasing number of completely sequenced *Rhizobium* species obtained from various legumes, the complex evolutionary relationships of plasmids could be defined through a comparative genomic DNA analysis and network analysis approach. Here, we propose a clear pan-genomic structure of *Rhizobium* plasmids as a fully connected PU-HPC bipartite network. The HPCs in the network satisfy the topological scale-free criterion of the biological networks, indicating a large number of plasmid genes with highly variable abundances and often complex evolutionary histories (Albert, 2005). The PUs show strong modularity of the network, providing a complementary and more comprehensive account of the deep evolutionary connections within the rhizobial plasmids (Iranzo *et al*., 2016). We believe these findings are useful for

conducting higher-level functional and evolutionary analysis of multipartite rhizobial genomes.

Compared with traditional genetic markers, the PU-HPC bipartite network presented here shows great advantage in permitting the rapid and systematic resolution of a reliable classification of *Rhizobium* plasmids with simple and specific bioinformatics processing. In total, 34 homologous PUCs and 26 PUC compositions are retrieved from 216 plasmids of 49 *Rhizobium* strains. Meanwhile, the bipartite network analyses in this study reflect the genetic material exchanges among different PUCs and chromosomes. This observation shows an instance of the process called 'gene externalization' that gene sharing between extra-chromosomal elements and chromosomes in the microbial world, which has been comprehensively illuminated by a recent bipartite network analysis (Corel *et al*., 2018). These indicate the *Rhizobium* plasmids significantly contributed to the dynamics of their host genomes.

Despite this, it is clear that a major extra-chromosomal composition is carried by PUCs 1–3, 6 and 8. Additionally, the four mega-plasmids in PUCs 9 and 19 are found to be formed by fusions of PUCs 1–3, while PUC 6 appears to be derived from PUC 3 fission. The existence of the four mega-plasmids may be explained by the incompatibility of the source plasmids (Velappan *et al*., 2007). If so, the mega-plasmids essentially retain all the genes necessary for main accessory plasmid properties. Therefore, we suspect that the functions of PUCs 1–3 are quite complementary and allow for interaction (Snel *et al*., 2000). In summary, we assume that a common extra-chromosomal genome including the core/essential HPCs connects to the four common PUCs (1–3 and 8). Only two strains, *Rhizobium* sp. IRBG74 and NT-26, are found to be exceptions to this phenomenon. *Rhizobium* sp. IRBG74 is a special symbiont nodulating the aquatic legume *Sesbania* sp. and also a growth-promoting endophyte of wetland rice; however, its genome is composed of a circular chromosome, a linear chromosome and a pSym (Crook *et al*., 2013). *Rhizobium* sp. NT-26 has lost the major colonizing capabilities needed for symbiosis with legumes; nonetheless, it has a plasmid NT-26_p1 containing the various genes to metabolize arsenite, which enables it to live in an arsenic-containing goldmine (Andres *et al*., 2013). Therefore, we suspect that the distinct plasmid compositions of these two strains are the results of adaptive evolution to fit their natural environment and ecologic niches.

As the most important extra-chromosomal replicon, the pSyms found on the bipartite network are clustered into nine PUCs including seven exclusive PUCs and two mixed PUCs. This reveals their large genomic diversity and complex evolutionary history. The advanced network of pSyms with interactive visualization can be simply modified by adding their host legume plants. Such a modified network shows that the complex evolution among most of the pSyms are selected by the host plants. Indeed, one legume species can be related to two or three different PUCs, while some PUCs also have a broad range of host plants. These results lend support to the hypothesis that host plant selection contributes to the adaptive evolution of pSyms and structuring of rhizobial communities on host legume rhizosperes, root surfaces and nodules (Miranda-Sánchez *et al*., 2016; Remigi *et al*., 2016). In addition, most of the pSyms are conjugative and have lower diversity, higher recombination rates and different codon usage bias compared with both the common accessory PUCs and the main chromosomes (Pérez Carrascal *et al*., 2016). These evolutionary characteristics indicate that most of the pSyms do not appear to have co-evolved with the chromosomes or accessory plasmids. Additionally, the three unanticipated pSyms in the mixed PUCs (1 and 3) were formed by the integration of symbiosis modules into their non-symbiotic genomic backbones. These occurrences seem to be similar to the chromosomal islands in other rhizobia such as *Mesorhizobium loti* R7A (Sullivan *et al*., 2002). Importantly, these findings suggest that any analysis of comparative genomics of pSyms should be performed with caution. In summary, the pSyms in *Rhizobium* bacteria closely co-evolve with their host plants by replaying HGT-driven evolution, and the extra-chromosomal replicons are pre-adapted for transfer and integration of the symbiosis module.

It is worth noting that the accessory PUCs and their HPCs contribute greatly to the modularity of the bipartite network. Meanwhile, these PUCs can be divided into two groups. One group includes the common and large accessory PUCs (e.g., 1–3, 6, 9 and 19), which are closely related to the main chromosomes according to the evolutionary characteristics, and are independent of host plant selection. This reflects the fact that these non-pSyms have co-evolved with the chromosome in *Rhizobium* over a long duration. Moreover, most of their members can be regarded as the 'chromids' for carrying the core genes found on the chromosome in other bacterial species (Harrison *et al*., 2010). These chromids are important and may confer particular advantages to *Rhizobium* bacteria; for example, they may enable a strain to have a larger genome while keeping the chromosome small. This could allow the bacteria to replicate faster and adapt to a specialist niche better (Harrison *et al*., 2010). The other group includes those rare and small accessory PUCs, which are mostly located on the periphery of the network. These PUCs vary and appear to be dispensable compared with the common PUCs because they are rare and connect to few HPCs. Moreover, most PUCs are specific to the host legume plant or environment from

which they were isolated, while they are similar to the symbiotic PUCs in both codon usage bias and mobility. After conducting an in-depth survey of their genes in the database, most of these can be retrieved from the genus *Sinorhizobium* or *Agrobacterium*, which are closely related to *Rhizobium*. Taken together, we postulate that the common accessory PUCs have spent a long time in the same cellular environment as the chromosomes they are associated with. In contrast, the rare accessory PUCs have transferred from other related soil bacteria much more recently. Among these accessory PUCs, PUC 10 containing the plasmid BIHB1148_pSK03 deserves attention due to its closely evolutionary connection to the chromosome and the absence of a *repABC* replicon system. They may use the replication enzymes from the host-cell for making copies, by inserting themselves into the main chromosome (Hajare and Ade, 2012). This observation suggests a type of newly formed plasmid deriving from a fragment of the main chromosome (Kado, 1998). In summary, all observations related to the genetic exchanges and innovations of the accessory PUCs synthetically elucidate the dynamic evolutionary formations of various plasmid compositions among *Rhizobium* bacteria that contribute to their adaptive evolution.

As the fundamental structure and genetic content of *Rhizobium* plasmids have been elucidated, it is of interest to explore the functional characteristics of the different PUCs. Currently, it remains a challenge to investigate this area because a large percentage of the plasmid genes have not yet been clearly annotated. Herein, we have taken two steps to clarify the role of the PUCs, especially the common PUCs. First, based on COG enrichment analysis, we find that each of the common PUCs has its distinct functional bias and all of these biased functions are important for the ability to adapt to the environment and be competitive with other bacteria. Second, by conducting sequence similarity searches, the common PUCs are found to encode some symbiosis-related genes, indicating their influence on symbiotic activity. The impact of all *R. etli* CFN42 plasmids on fitness and nodulation competitiveness have been well experimentally validated (Brom *et al*., 1992, 2000). Therefore, the functions of the common PUCs are complementary and interact with one another. This is also in accordance with the visualization of the fully connected bipartite network. Additionally, our conclusion is consistent with previous finding in the reference *R. etli* CFN42 genome, where the plasmids p42b (PUC 6), p42c (PUC 1), p42e (PUC 2) and p42f (PUC 3) are strongly related functionally and have co-evolved with the chromosomes over a long period, while p42a (PUC 15) and p42d_Sym (PUC 8) were only recently acquired (González *et al*., 2006; Pérez Carrascal *et al*., 2016; Bañuelos-Vazquez *et al*., 2019). Increasing evidence has been mounting regarding the accessory plasmids' influence on legume-rhizobia symbiosis under stress, such as resistance to heat, acid, antibiotics, heavy metals, pesticides and oxidative stress protection (Kurchak *et al*., 2001; Streit *et al*., 2004, 234; Anjum *et al*., 2011; Vercruysse *et al*., 2011; Naamala *et al*., 2016). Therefore, it is of significance to uncover novel characteristics and functions of different PUCs and then determine their influence on the symbiotic interactions between legumes and rhizobia, in niches with diverse and harsh environmental conditions (Zahran, 2017).

All findings in this study from the PU-HPC bipartite network analysis underscore the importance of undertaking subsequent sequencing and experiment projects in resolving the issue concerning genomic annotation and interaction of rhizobial plasmids (diCenzo *et al*., 2014). Further work aimed at identifying the selective forces acting on both symbiotic and accessory plasmids and responsible for variation in nitrogen fixation efficiency should also help in understanding their co-evolutionary framework ruling the evolution of *Rhizobium*-legume symbiosis (Remigi *et al*., 2014). Given the increasing availability and utilization of multiomics sequencing with the rapid development of network analysis approaches, such as multipartite network and multilayer network, we can expect to better understand these intriguing genetic elements in various lifestyles of all rhizobial bacteria (diCenzo *et al*., 2016; Marx *et al*., 2016).

Experimental procedures

Data collection

All 49 available complete genomes of *Rhizobium* were downloaded from the NCBI GenBank database (January 2018; ftp://ftp.ncbi.nlm.nih.gov/genomes) and the information of their nodulation abilities with legume species was collected from related studies (Table S1). Among them, 37 strains were isolated from the rhizosphere and root nodules of *Phaseolus vulgaris*, four strains were from *Trifolium* spp., three strains were from *Pisum sativum*, two strains were from *Vavilovia formosa*, one strain was from *Mimosa affinis*, one strain was from *Sesbania cannabina*, and one strain was from an arsenic-containing gold mine. Only two strains, *R. phaseoli* R744 and *Rhizobium* sp. NT-26, were reported to be non-symbiotic and missing the pSyms (Andres *et al*., 2013; Pérez Carrascal *et al*., 2016).

In total, 216 plasmid proteomes of the 49 *Rhizobium* strains were extracted from the genome annotation files using an in-house Python script. A plasmid was regarded as a symbiotic plasmid if it contained genes essential for nodulation (*nod*/*nol*/*noe*) and nitrogen fixation (*nif*/*fix*/*fdx*) (Remigi *et al*., 2016); otherwise, it was classified as an accessory plasmid. Due to several incomplete or ambiguous plasmid naming conventions in NCBI, we renamed every plasmid here using the 'strain abbreviation_GenBank definition' format. The reference genomes of *Sinorhizobium*

meliloti 1021 and *Bradyrhizobium japonicum* USDA 110 were downloaded from NCBI GenBank database and the sequences of the symbiosis-related genes were retrieved from these two genomes according to the NodMutDB database with BLASTp (Mao *et al*., 2005).

Bipartite network construction, clustering and visualization

The bipartite network was constructed based on all *Rhizobium* plasmid proteomes using the AccNet software (Lanza *et al*., 2017). Specifically, all homologous proteins over 40% amino acid sequence identity were clustered using kClust with parameters of '-s 1.73 -c 0.8 -e 1e-10'. These parameters mean identity, coverage and e-value respectively (Hauser *et al*., 2013). The network was hierarchically clustered at a practical similarity threshold of 85% with the hclust method in R software. Network modularity was determined using Vincent's heuristic algorithm to measure the strength of division of the network into modules (Newman, 2006). Distinct communities were identified using network modularity with weighted edges and a resolution of 1.0 (Barber, 2007).

To assess stability of the bipartite network analyses at different sequence identities, we reconstructed the bipartite network at 30%–90% sequence identities and redone the hierarchical clustering on these result networks. Next, to test the influence of the plasmid size on the clustering result, we jackknifed the first half of all plasmids according to their genome sizes by randomly sampling 20%–90% of all associated genes for 100 times respectively. Then, we redone all the network construction and clustering analyses on the generated data sets.

The obtained network files including nodes, edges and clusters were then imported into the Gephi software (Bastian *et al*., 2009). We displayed the relative genomic content of each plasmid by making the diameter of each node proportional to its degree. Next, we constructed a visualization of the network using Gephi's built-in ForceAtlas2 algorithm, which is a force-directed layout that simulates a physical system to spatialize a network (Jacomy *et al*., 2014). The bipartite graph was generated using default parameters of the layout algorithm except for the following: approximate speed of 1.0, scaling of 200, gravity of 1.0 and 'prevent overlap' option.

In the context of networks, the degree of a node represents the number of edges connected to that given node, while the SCC measures the degree to which nodes in a network tend to cluster together and is considered an indicator of hierarchical modular organization (Zhang *et al*., 2008). These two topological features of the bipartite graph were analysed by using the NetworkX package in Python and the iGraph package in R respectively (Csardi and Nepusz, 2006; Hagberg *et al*., 2008). For fitting a power law to empirical degree distribution data, we used the poweRlaw package in R (Gillespie, 2015). Furthermore, bacterial plasmids are considered crucial mediators of HGT and can confound the phylogeny between donors and recipients (Yamashita *et al*., 2014). To characterize the putative recent HGT or plasmid transfer, the PU-HPC bipartite network was reconstructed by using both 95% and 99% amino acid sequence identity cut offs respectively (Li *et al*., 2018).

Plasmid genome comparisons

Visual comparisons of plasmid genome homology were done by using BRIG (BLAST Ring Image Generator) v0.95 (Alikhan *et al*., 2011). BRIG program can generate circular comparison images for genomes and display similarity between a reference plasmid genome in the centre and other query plasmid genomes. As the similarity is calculated in regard to the reference genome, regions that are absent from the reference genome but present in one or more of the query genomes will not be displayed. The BRIG method used the software BLASTn v2.60 for the searches. All comparisons were done with default parameters.

Large-scale exogenetic analysis of all PUCs

The example PU in each PUC was randomly selected and its protein sequences were aligned against all available non-redundant protein sequences of bacteria except *Rhizobium* in the NCBI NR database using DIAMOND with an E-value of 1e-5 (Buchfink *et al*., 2015). Then, the matches with less than an identity of 40% or bitscore of 200 were filtered out. The eventual number of one-to-one matches was divided by the number of all proteins in a randomly selected PU to measure the proportion of exogenous source of each PUC.

Replicability and mobility analysis of Rhizobium plasmids

The *repABC* operon is the prevalent replication unit of alphaproteobacterial plasmids and their semi-autonomy is ensured by the essential replicase gene *repC*, as well as the *repAB* partitioning cassette (Cevallos *et al*., 2008). To investigate whether the *Rhizobium* plasmids used in this study were self-replicative, we detected the *repABC* operons in all 216 plasmids. The *repABC* sequences of *R. etli* CFN42 were selected as reference sequences to search the local database containing those *Rhizobium* plasmids using BLASTp with an E-value of 1e-10 and an identity over 40%. The inferred *repABC* homologous families were then manually validated based on the Markov clustering analysis and the annotation of each hit (Li, 2003). Plasmids with at least one complete *repABC* operon were regarded as having the ability to self-replicate.

Mobilizable plasmids encode a minimal mobility (MOB) machinery. Conjugative relaxases are suggested to be a robust, universal evolutionary marker for mobilizable plasmids (Smillie *et al*., 2010). Additionally, genes encoding T4SS-related proteins also endow plasmids with the ability to self-transmit via conjugation. Here, to determine the mobility of each *Rhizobium* plasmid, we detected the existence of MOB relaxase, T4SS, and T4CP components by using BLASTp, mainly following Smillie's pipeline. Representative sequences of 19 relaxase sub-clades were selected from previous studies (Ding and Hynes, 2009; Garcillán-Barcia *et al*., 2009). Any hit with an identity over 40% and its annotation as 'relaxase' or 'Ti-type conjugative transfer relaxase *traA*' was regarded as relaxase. For those hits with lower similarity with the representative relaxase, they often had fuzzy annotations such as 'hypothetical protein' or 'helicase'. Therefore, we validated those sequences by phylogenetic analysis of that putative relaxase with other reported MOB families of relaxases (Table S12). Similarly, T4SS and T4CP were searched against representative mating pair formation sequences in our local database. All putative T4SS and T4CP candidates were manually checked.

Phylogenetic analysis

Average nucleotide identity based on MUMmer (ANIm) analysis of all *Rhizobium* strains was performed using the pyani toolkit (https://github.com/widdowquinn/pyani). Briefly, the nucleotide sequences were collected and pairwise aligned in the ANIm mode. There is no universally conserved gene among all *Rhizobium* plasmids used to construct their single phylogeny. In spite of this, two genetic markers, *repABC* operon and conjugative relaxase (MOB), are usually adopted to define the phylogenetic relationship of plasmids (Fernandez-Lopez *et al*., 2017; González *et al*., 2019). The *repABC* operon is widely conserved among α-proteobacteria and the MOB gene provides a robust, universal evolutionary marker for only mobilizable and conjugative plasmids. The *repABC* and MOB protein sequences were aligned with MAFFT software (Katoh and Standley, 2013) and automatically trimmed using trimAl with the automatic option (Capella-Gutierrez *et al*., 2009). For each alignment, a maximum likelihood tree was constructed in RAxML with 500 bootstraps under the GTR model and gamma correction (GAMMA) for variable evolutionary rates (Stamatakis, 2014). These phylogenetic trees were then visualized using iTOL (Letunic and Bork, 2016).

Population genetic analysis

To measure nucleotide diversity (π) and Tajima's *D* values in the common PUCs 1–3 and 8 as well as their associated chromosomes, we used the PEGAS and APE packages in R (Paradis *et al*., 2004; Paradis, 2010). Nucleotide alignments for all homologous gene families were obtained according to their amino acid alignments using the PAL2NAL program (Suyama *et al*., 2006). To illustrate the distribution of π and Tajima's *D* values among the homologues, plots were drawn using the 'scipy.stats.gaussian_kde' module in Python from the SciPy package, which allowed for estimation of the probability density function using Gaussian kernels.

The PUs in PUCs 1–3 and 8 as well as their associated chromosomes were aligned using the Progressive Mauve program to identify the conserved regions within these replicons, with respect to both the number of nucleotide bases and locally collinear blocks. Locally collinear blocks were then retrieved from the Mauve alignment using the program stripSubsetLCBs in Mauve (Darling *et al*., 2010). These coordinates were employed to graph and observe the synteny along the DNA replicons. The sequence alignments were processed with homemade Python scripts to obtain the FASTA file and remove gaps within the sequences. Using the package ClonalFrameML, alignments of the genomic compartments without gaps, within and between clusters, were used to infer the rate at which recombination mutations occur (ρ/θ), the ratio of the probability that a site is altered by recombination or mutation (r/m), and the position of cross-referenced recombination sites (Didelot and Wilson, 2015).

Codon usage bias analysis

To determine the characteristics of codon usage bias, RSCU was calculated for all *Rhizobium* replicons by using CodonW (http://codonw.sourceforge.net). The RSCU value of each codon was used to estimate the similarity influence among the organisms in this study. All codons (except UAG, UGG, UAA, AUG, and UGA) were organized in a matrix of $N \times M$ dimensions, where N is the number of species and M is the number of degenerated codons. Hierarchical clustering of this matrix was conducted based on Spearman's correlational distance of RSCU values using Bioconductor with Ward's method (Gentleman *et al*., 2004). The resulting dendrogram was extracted using the ape package in R and visualized using iTOL (Paradis *et al*., 2004; Letunic and Bork, 2016).

Chromid detection

A plasmid will be regarded as a 'chromid' based on these four characteristics: (i) G + C content similar to the associated chromosome with a$\pm$ 2% cut off, (ii) RSCU value similar to the associated chromosome, (iii) molecular size larger than 0.4 Mb, and (iv) encoding more than one core ortholog found on the chromosome in other species

(Harrison *et al*., 2010). G + C content of plasmids and associated chromosomes are calculated using Biopython (Cock *et al*., 2009). Here, 70 reference chromid-containing genomes were downloaded from the NCBI RefSeq database (Harrison *et al*., 2010). The reference core genes of these genomes were identified through the OrthoMCL algorithm (Li, 2003). The putative orthologous gene pairs between the reference core genes and plasmid genes were identified by an 'all-against-all' BLASTp analysis.

COG annotation and enrichment analysis of common PUCs

Proteins harboured by all plasmids in a particular PU were used to represent the basic function of plasmids in this PU. A randomly selected sequence of each protein family was used to perform next-step COG annotation. COG annotation is performed using the eggNOG database (Huerta-Cepas *et al*., 2019). Fisher's exact test was used to measure the statistical significance of COG enrichment analysis. In further detail, for each COG category in PUCs 1–3, 6 and 8, we counted the (i) protein family number of this COG category in a particular PUC; (ii) protein family number of this COG category not in the selected PUC; (iii) protein family number not belonging to this COG category but in the selected PUC and (iv) protein family number belonging to neither this COG category nor the selected PUC. Those four elements consisted of the contingency tables that were used to perform Fisher's exact test. The *P*-values were used to evaluate the significance of whether the particular COG category was more likely to occur in a particular PUC.

Acknowledgements

This research was supported by the National Science Foundation of China (Grant Nos. 41830755 and 31771474).

References

Albert, R. (2005) Scale-free networks in cell biology. *J Cell Sci* **118**: 4947–4957.

Alikhan, N.-F., Petty, N.K., Zakour, N.L.B., and Beatson, S. A. (2011) BLAST ring image generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**: 402.

Andres, J., Arsène-Ploetze, F., Barbe, V., Brochier-Armanet, C., Cleiss-Arnold, J., Coppée, J.-Y., *et al*. (2013) Life in an arsenic-containing gold mine: genome and physiology of the autotrophic arsenite-oxidizing bacterium *Rhizobium* sp. NT-26. *Genome Biol Evol* **5**: 934–953.

Anjum, R., Grohmann, E., and Malik, A. (2011) Molecular characterization of conjugative plasmids in pesticide tolerant and multi-resistant bacterial isolates from contaminated alluvial soil. *Chemosphere* **84**: 175–181.

Bañuelos-Vazquez, L.A., Torres Tejerizo, G., Cervantes-De La Luz, L., Girard, L., Romero, D., and Brom, S. (2019) Conjugative transfer between *Rhizobium etli* endosymbionts inside the root nodule. *Environ Microbiol*. [Epub ahead of print]

Barber, M.J. (2007) Modularity and community detection in bipartite networks. *Phys Rev E* **76**: 066102.

Bastian, M., Heymann, S., and Jacomy, M. (2009) Gephi: an open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*; *Third International AAAI Conference on Weblogs and Social Media*.

Bernard, G., Greenfield, P., Ragan, M.A., and Chan, C.X. (2018) K-mer similarity, networks of microbial genomes and taxonomic rank. *mSystems* **3**: e00257–e00218.

Brom, S., de los Santos, A.G., Cervantes, L., Palacios, R., and Romero, D. (2000) In rhizobium etli symbiotic plasmid transfer, nodulation competitivity and cellular growth require interaction among different replicons. *Plasmid* **44**: 34–43.

Brom, S., de los Santos, A.G., Stepkowsky, T., Flores, M., Dávila, G., Romero, D., and Palacios, R. (1992) Different plasmids of *Rhizobium leguminosarum* bv. *phaseoli* are required for optimal symbiotic performance. *J Bacteriol* **174**: 5183–5189.

Brom, S., Girard, L., Tun-Garrido, C., Garcia-de los Santos, A., Bustos, P., Gonzalez, V., and Romero, D. (2004) Transfer of the symbiotic plasmid of *Rhizobium etli* CFN42 requires Cointegration with p42a, which may be mediated by site-specific recombination. *J Bacteriol* **186**: 7538–7548.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Cevallos, M.A., Cervantes-Rivera, R., and Gutiérrez-Ríos, R. M. (2008) The *repABC* plasmid family. *Plasmid* **60**: 19–37.

Cevallos, M.A., Porta, H., Izquierdo, J., Tun-Garrido, C., de los Santos, A.G., Dávila, G., and Brom, S. (2002) *Rhizobium etli* CFN42 contains at least three plasmids of the *repABC* family: a structural and evolutionary analysis. *Plasmid* **48**: 104–116.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., *et al*. (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016) Network-thinking: graphs to analyze microbial complexity and evolution. *Trends Microbiol* **24**: 224–237.

Corel, E., Méheust, R., Watson, A.K., McInerney, J.O., Lopez, P., and Bapteste, E. (2018) Bipartite network analysis of gene sharings in the microbial world. *Mol Biol Evol* **35**: 899–913.

Crook, M.B., Mitra, S., Ane, J.-M., Sadowsky, M.J., and Gyaneshwar, P. (2013) Complete genome sequence of the *Sesbania* symbiont and rice growth-promoting endophyte *Rhizobium* sp. strain IRBG74. *Genome Announc* **1**: e00934-13.

Csardi, G., and Nepusz, T. (2006) The igraph software package for complex network research. *Int J Complex Syst* **1695**: 1–9.

Darling, A.E., Mau, B., and Perna, N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**: e11147.

diCenzo, G.C., MacLean, A.M., Milunovic, B., Golding, G.B., and Finan, T.M. (2014) Examination of prokaryotic multipartite genome evolution through experimental genome reduction. *PLoS Genet* **10**: e1004742.

diCenzo, G.C., Checcucci, A., Bazzicalupo, M., Mengoni, A., Viti, C., Dziewit, L., *et al*. (2016) Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. *Nat Commun* **7**: 12219.

Didelot, X., and Wilson, D.J. (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* **11**: e1004041.

Ding, H., and Hynes, M.F. (2009) Plasmid transfer systems in the rhizobia. *Can J Microbiol* **55**: 917–927.

Ding, H., Yip, C.B., and Hynes, M.F. (2013) Genetic characterization of a novel rhizobial plasmid conjugation system in *Rhizobium leguminosarum* bv. *viciae* Strain VF39SM. *J Bacteriol* **195**: 328–339.

Dresler-Nurmi, A., Fewer, D.P., Räsänen, L.A., and Lindström, K. (2009) The diversity and evolution of rhizobia. In *Prokaryotic Symbionts in Plants*, Pawlowski, K. (ed). Berlin, Heidelberg: Springer, pp. 3–41.

Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M.P., and de la Cruz, F. (2017) Towards a taxonomy of conjugative plasmids. *Curr Opin Microbiol* **38**: 106–113.

Fondi, M., Karkman, A., Tamminen, M.V., Bosi, E., Virta, M., Fani, R., *et al*. (2016) "Every gene is everywhere but the environment selects": global geolocalization of gene sharing in environmental samples through network analysis. *Genome Biol Evol* **8**: 1388–1400.

Freiberg, C., Fellay, R., Bairoch, A., Broughton, W.J., Rosenthal, A., and Perret, X. (1997) Molecular basis of symbiosis between *Rhizobium* and legumes. *Nature* **387**: 394–401.

Garcillán-Barcia, M.P., Francia, M.V., and de La Cruz, F. (2009) The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol Rev* **33**: 657–687.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**: R80.

Gillespie, C.S. (2015) Fitting heavy tailed distributions: the poweRlaw package. *J Stat Softw* **64**: 1–16.

González, V. (2003) The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. *Genome Biol* **4**: 54–56.

González, V., Santamaría, R.I., Bustos, P., Hernándezgonzález, I., Medranosoto, A., Morenohagelsieb, G., *et al*. (2006) The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci U S A* **103**: 3834–3839.

González, V., Santamaría, R.I., Bustos, P., Pérez-Carrascal, O.M., Vinuesa, P., Juárez, S., *et al*. (2019) Phylogenomic *Rhizobium* species are structured by a continuum of diversity and genomic clusters. *Front Microbiol* **10**: 910.

Hagberg, A., Los, A.N.L., Swart, P., Los, A.N.L., Chult, D.S., and Colgate, U. (2008) *Exploring network structure, dynamics, and function using networkx*. Los Alamos, NM: Research Org., Los Alamos National Laboratory (LANL).

Hajare, B., and Ade, A. (2012) Confirming location of nitrogen fixing genes on plasmids in *Rhizobium* isolated from *Pisum sativum*. *Biosci Discov* **3**: 160–164.

Harrison, P.W., Lower, R.P.J., Kim, N.K.D., and Young, J.P. W. (2010) Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol* **18**: 141–148.

Hauser, M., Mayer, C.E., and Söding, J. (2013) kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics* **14**: 248.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., *et al*. (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**: D309–D314.

Iranzo, J., Koonin, E.V., Prangishvili, D., and Krupovic, M. (2016) Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. *J Virol* **90**: 11043–11055.

Iranzo, J., Krupovic, M., and Koonin, E.V. (2017) A network perspective on the virus world. *Commun Integr Biol* **10**: e1296614.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014) ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**: e98679.

Jaffe, A.L., Corel, E., Pathmanathan, J.S., Lopez, P., and Bapteste, E. (2016) Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins: LGT among ultrasmall prokaryotes. *Environ Microbiol* **18**: 5072–5081.

Kado, C.I. (1998) Origin and evolution of plasmids. *Antonie Van Leeuwenhoek* **73**: 117–126.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kurchak, O.N., Provorov, N.A., and Simarov, B.V. (2001) Plasmid pSym1-32 of *Rhizobium leguminosarum* bv. *viceae* controlling nitrogen fixation activity, effectiveness of symbiosis, competitiveness, and acid tolerance. *Russ J Genet* **37**: 1025–1031.

Lanza, V.F., Baquero, F., de la Cruz, F., and Coque, T.M. (2017) AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. *Bioinformatics* **33**: 283–285.

Letunic, I., and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**: W242–W245.

Li, L. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Li, X., Tong, W., Wang, L., Rahman, S.U., Wei, G., and Tao, S. (2018) A novel strategy for detecting recent horizontal gene transfer and its application to *Rhizobium* strains. *Front Microbiol* **9**: 973.

López-Guerrero, M.G., Ormeño-Orrillo, E., Acosta, J.L., Mendoza-Vargas, A., Rogel, M.A., Ramírez, M.A., *et al*. (2012) Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. *Plasmid* **68**: 149–158.

MacLean, R.C., and San Millan, A. (2015) Microbial evolution: towards resolving the plasmid paradox. *Curr Biol* **25**: R764–R767.

Mao, C., Qiu, J., Wang, C., Charles, T.C., and Sobral, B.W. S. (2005) NodMutDB: a database for genes and mutants involved in symbiosis. *Bioinformatics* **21**: 2927–2929.

Marx, H., Minogue, C.E., Jayaraman, D., Richards, A.L., Kwiecien, N.W., Sihapirani, A.F., *et al*. (2016) A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. *Nat Biotechnol* **34**: 1198–1205.

Masson-Boivin, C., Giraud, E., Perret, X., and Batut, J. (2009) Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? *Trends Microbiol* **17**: 458–466.

Mavingui, P., Flores, M., Guo, X., Davila, G., Perret, X., Broughton, W.J., and Palacios, R. (2002) Dynamics of genome architecture in *Rhizobium* sp. strain NGR234. *J Bacteriol* **184**: 171–176.

Mazur, A., and Koper, P. (2012) Rhizobial plasmids—replication, structure and biological role. *Open Life Sci* **7**: 571–586.

Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P., and Bapteste, E. (2016) Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc Natl Acad Sci U S A* **113**: 3579–3584.

Miranda-Sánchez, F., Rivera, J., and Vinuesa, P. (2016) Diversity patterns of *Rhizobiaceae* communities inhabiting soils, root surfaces and nodules reveal a strong selection of rhizobial partners by legumes: community ecology of *Rhizobiaceae*. *Environ Microbiol* **18**: 2375–2391.

Naamala, J., Jaiswal, S.K., and Dakora, F.D. (2016) Antibiotics resistance in *Rhizobium*: type, process, mechanism and benefit for agriculture. *Curr Microbiol* **72**: 804–816.

Newman, M.E. (2006) Modularity and community structure in networks. *Proc Natl Acad Sci U S A* **103**: 8577–8582.

Orlandini, V., Emiliani, G., Fondi, M., Maida, I., Perrin, E., and Fani, R. (2014) Network analysis of plasmidomes: the *Azospirillum brasilense* Sp245 case. *Int J Evol Biol* **2014**: 1–14.

Orlek, A., Stoesser, N., Anjum, M.F., Doumith, M., Ellington, M.J., Peto, T., *et al*. (2017) Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front Microbiol* **8**: 182.

Palacios, R., and Flores, M. (2005) Genome dynamics in rhizobial organisms. In *Genomes and Genomics of Nitrogen-Fixing Organisms*, Palacios, R., and Newton, W. E. (eds). Dordrecht: Springer Netherlands, pp. 183–200.

Paradis, E. (2010) Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**: 419–420.

Paradis, E., Claude, J., and Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Pérez Carrascal, O.M., VanInsberghe, D., Juárez, S., Polz, M.F., Vinuesa, P., and González, V. (2016) Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing *Rhizobium* species associated with *Phaseolus vulgaris*. *Environ Microbiol* **18**: 2660–2676.

Remigi, P., Capela, D., Clerissi, C., Tasse, L., Torchet, R., Bouchez, O., *et al*. (2014) Transient hypermutagenesis accelerates the evolution of legume endosymbionts following horizontal gene transfer. *PLoS Biol* **12**: e1001942.

Remigi, P., Zhu, J., Young, J.P.W., and Masson-Boivin, C. (2016) Symbiosis within symbiosis: evolving nitrogen-fixing legume symbionts. *Trends Microbiol* **24**: 63–75.

Sedlar, K., Videnska, P., Skutkova, H., Rychlik, I., and Provaznik, I. (2016) Bipartite graphs for visualization analysis of microbiome data. *Evol Bioinforma* **12**: 17–23.

Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E. P.C., and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol Mol Biol Rev* **74**: 434–452.

Snel, B., Bork, P., and Huynen, M. (2000) Genome evolution: gene fusion versus gene fission. *Trends Genet* **16**: 9–11.

Soucy, S.M., Huang, J., and Gogarten, J.P. (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**: 472–482.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Streit, W.R., Schmitz, R.A., Perret, X., Staehelin, C., Deakin, W.J., Raasch, C., *et al*. (2004) An evolutionary hot spot: the pNGR234b replicon of *Rhizobium* sp. strain NGR234. *J Bacteriol* **186**: 535–542.

Sullivan, J.T., Trzebiatowski, J.R., Cruickshank, R.W., Gouzy, J., Brown, S.D., Elliot, R.M., *et al*. (2002) Comparative sequence analysis of the symbiosis Island of *Mesorhizobium loti* strain R7A. *J Bacteriol* **184**: 3086–3095.

Suyama, M., Torrents, D., and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609–W612.

Tamminen, M., Virta, M., Fani, R., and Fondi, M. (2011) Large-scale analysis of plasmid relationships through gene-sharing networks. *Mol Biol Evol* **29**: 1225–1240.

Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P., and Bradbury, A.R.M. (2007) Plasmid incompatibility: more compatible than previously thought? *Protein Eng Des Sel* **20**: 309–313.

Vercruysse, M., Fauvart, M., Jans, A., Beullens, S., Braeken, K., Cloots, L., *et al*. (2011) Stress response regulators identified through genome-wide transcriptome analysis of the (p)ppGpp-dependent response in Rhizobium etli. *Genome Biol* **12**: R17.

Villaseñor, T., Brom, S., Dávalos, A., Lozano, L., Romero, D., and los Santos, A. (2011) Housekeeping genes essential for pantothenate biosynthesis are plasmid-encoded in Rhizobium etli and Rhizobium leguminosarum. *BMC Microbiol* **11**: 66.

Wozniak, R.A.F., and Waldor, M.K. (2010) Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Microbiol* **8**: 552–563.

Yamashita, A., Sekizuka, T., and Kuroda, M. (2014) Characterization of antimicrobial resistance dissemination across plasmid communities classified by network analysis. *Pathogens* **3**: 356–376.

Zahran, H.H. (2017) Plasmids impact on rhizobia-legumes symbiosis in diverse environments. *Symbiosis* **73**: 75–91.

Zhang, P., Wang, J., Li, X., Li, M., Di, Z., and Fan, Y. (2008) Clustering coefficient and community structure of bipartite networks. *Phys. A* **387**: 6869–6875.

Zhang, X.-X., Kosier, B., and Priefer, U.B. (2001) Symbiotic plasmid rearrangement in *Rhizobium leguminosarum* bv. *viciae* VF39SM. *J Bacteriol* **183**: 2141–2144.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Heat maps of average nucleotide identity based on MUMmer (ANIm) for the 49 compete genomes of *Rhizobium*. Hierarchical clustering of 49 strains is indicated in rows. At a threshold of 95% average nucleotide identity for species delineation, as shown in red, 19 clusters or genospecies are classified.

Fig. S2. Topological features of the PU-HPC bipartite network based on 216 *Rhizobium* plasmids. (A) Degree distribution of the plasmid units (PUs). (B) Degree distribution of the homologous protein clusters (HPCs). (C) Squared clustering coefficient (SCC) distribution of the PUs. (D) SCC distribution of the HPCs. Both degree and SCC distributions of HPCs approximately conformed to a continuous power law distribution. The degree distribution of PUs shows four mega-plasmids (R602_pRgalR602c, IE4872_pRgalIE4872d, 8C3_pRsp8C3c and CIAT899_pRtrCIAT899c) with high degrees. The SCC distribution of PUs has multiple peaks which reflected the hierarchical structure of PUs in the network.

Fig. S3. The modularity of the PU-HPC bipartite network. The colour of a node accords with its module assignation (right panel). There are a total of 32 modular classes of both PUs and HPCs at a high modularity score of 0.814 with a resolution of 1.0. The modular classes of PUs are generally in accordance with the PUCs shown in Fig. .

Fig. S4. The PU-HPC bipartite network constructed at (A) 99% and (B) 95% amino acid identity cut-offs for characterization of the possible horizontal gene transfer (HGT) or plasmid transfer. The colour and size of a node accords with the PUC assignment in Fig. . The HPCs connecting multiple PUs represent the possible HGTs. Only about 10% and 15% of the HGTs occurred between different PUCs at 99% and 95% amino acid identity cut-offs respectively. These indicate the putative HGTs have little influence on the primary network clustering.

Fig. S5. Genome comparison to show the integration of symbiosis modules at different plasmid backbones. (A) 3841_pRL10_pSym and CFN42_p42c (as reference) in PUC 1, (B) WSM1325_pR132501_pSym, IE4771_pRetIE4771d, CFN42_p42f and WSM2304_pRL201_pSym (as reference) in PUC 3. The inner white ring denotes the reference plasmid genome. Regions in the query plasmids that are also present in the reference plasmid are displayed. Positions of symbiosis-related and *repABC* genes on the reference plasmid are marked in red.

Fig. S6. The bipartite network of all *Rhizobium* replicons showing their evolutionary relationships. The colour of PUs accord with the assignment in Fig. . The colour of main chromosomes (Chrs) are white. Common symbiotic and rare PUCs are more distantly related to the main Chrs than common accessory PUCs. Notably, PUC 10 locates nearby the main Chrs and its label is marked as red.

Fig. S7. The (A) replicability and (B) mobility of all PUs on the bipartite network. (**A**) This graph is derived from Fig. by setting two different colours (bottom right) on the PU nodes to represent the self-replicating PUs and those PUs without a complete *repABC* replication system. Almost all the PUs (209 of 216) encode at least one complete *repABC* cassette indicating they are self-replicative. In contrast, only 7 PUs are found without a complete *repABC* replication system and most of these PUs are on the periphery of the network. (**B**) Similarly, the colour of a node represents PU mobility (bottom right). In 216 PUs, 85 are conjugative, 15 are mobilizable and 116 are non-mobilizable. Most of symbiotic PUs and the PUs on the periphery of the network are conjugative while common accessory PUCs (1–3 and 6) are rarely conjugative or mobilizable.

Fig. S8. Phylograms of (A) *repA*, (B) *repB*, and (C) *repC* genes. The range of each node is coloured using the same assignment in the PU-HPC bipartite network (Fig.). The phylogram of *repA*, *repB* and *repC* were generally consistent with the network clustering result. 57 PUs have multiple (complete or partial) *repABC* cassettes. These PUs are mainly from PUC 3 indicating that several incompatibility groups exist in PUC 3.

Fig. S9. The phylogram of conjugative relaxases in all *Rhizobium* plasmids. The range of each node is coloured using the same assignment in the PU-HPC bipartite network (Fig.). A total of six types of conjugative relaxases are found in 100 *Rhizobium* plasmids and labelled beside the grey arcs.

Fig. S10. Synteny of PUCs 9, 19, and 1–3. Collinear and homologous regions are represented by similarly coloured blocks and connected by lines. The genomic synteny between three large PUCs (1–3) and PUCs 9 and 19 are obvious.

Fig. S11. Two highlighted HPCs located in the central position of the bipartite network. (**A**) HPC24118: peptide ABC transporter permease, and (**B**) HPC 36504: ABC transporter permease. The PUs connected to those two HPCs are highlighted in the figure. Those two HPCs are shared by most PUs in PUCs 1–3, 9 and 19.

Fig. S12. Phylograms of the two HPCs in the central position of the network. (A) HPC 24118 and (B) HPC 36504. The range of a plasmid is coloured using the same assignment in the PU-HPC bipartite network (Fig.). All three PUs in PUC 9 encode three copies of the two HPCs and each copy belongs to a monophyletic clade of PUCs 1–3. This indicates the PUs in PUC 9 are possibly the mergers of PUs in PUCs 1–3.

Fig. S13. The dendrogram of relative synonymous codon usage for all *Rhizobium* plasmids and the chromosomes based on Spearman's correlational distance. The range of a

plasmid is coloured using the same assignment in the PU-HPC bipartite network (Fig.) and the range of a chromosome is coloured light grey. There are three obviously monophyletic clades on the dendrogram which are marked by different grey strips. This indicates that common PUCs have a long co-evolved history with respect to the main chromosomes while most of symbiotic and rare PUCs are recently acquired into their cells. Moreover, all plasmids belonging to the medium grey strip were used as the candidates for subsequent chromids detection.

Fig. S14. The *Rhizobium* chromids detected on the bipartite network. This graph is derived from Fig. by setting two different colours (bottom right) on the PU nodes to represent the chromids and other plasmids. Here, 93 PUs are chromids having these four characteristics: (1) G + C content similar to the associated chromosome with a $\pm$ 2% cut-off, (2) RSCU value similar to the associated chromosome, (3) molecular size larger than 0.4 Mb, and (4) encoding more than one core ortholog found on the chromosome in other species.

Fig. S15. Host-plant specificity of all PUs on the bipartite network. The colour of a PU node denotes the host legume plant (bottom right) from which the strain was isolated. Here, only symbiotic PUs show host-plant specific. For example, the PUs in PUC 8 are all related to the symbiosis with *Phaseolus vulgaris*.

Fig. S16. Relative abundance of protein identifications in all PUCs by Clusters of Orthologous Groups (COG) functional category. Stacked bar chart shows the proportion of total protein identifications in each PUC. Kinds and proportions of the COG categories are different among all PUCs: some PUCs have few COG categories, e.g. PUC 11, while some PUCs have considerably large functional richness, e.g. PUC 8.

Table S1. The basic information and statistics of the 216 *Rhizobium* plasmids used in this study.

Table S2. The hierarchical clustering result from the bipartite network of all *Rhizobium* plasmids.

Table S3. The hierarchical clustering result from the bipartite network of all *Rhizobium* plasmids at seven amino acid sequence identity thresholds (30%, 40%, 50%, 60%, 70%, 80% and 90%).

Table S4. The hierarchical clustering result from the bipartite network of all *Rhizobium* plasmids at 40% amino acid sequence identity threshold and 8 jackknifed proportions (20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%).

Table S5. The genospecies distribution of 34 PUCs.

Table S6. Homologous gene detection result for showing possible exogenetic sources of the rare PUCs.

Table S7. The *repABC* cassette detection result of all *Rhizobium* plasmids.

Table S8. The mobility assignment of all *Rhizobium* plasmids.

Table S9. Statistics of the co-connected HPCs between PUCs 3 and 6.

Table S10. Detection of *Rhizobium* chromids based on GC content, genomic size, RSCU and the number of ortholog core genes on the chromosome in other 70 bacterial species.

Table S11. A matrix of the proportion of the PUs where the symbiosis-related genes in NodMutDB were detected in a PUC.

Table S12. Different types of previously reported conjugative relaxase families used to detect the mobility of all *Rhizobium* plasmids.