RESEARCH

The genomic architecture of introgression among sibling species of bacteria

Maria Izabel A Cavassim[1], Sara Moeskjær[2], Camous Moslemi[2], Bryden Fields[3], Asger Bachmann[1], Bjarni Vilhjálmsson[1], Mikkel H Schierup[1], J Peter W Young[3]* and Stig U Andersen[2]*

*Correspondence: sua@mbg.au.dk and peter.young@york.ac.uk
[2]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark
[3]Department of Biology, University of York, York, United Kingdom
Full list of author information is available at the end of the article

Abstract

Background: Gene transfer between bacterial species is an important mechanism for adaptation. For example, sets of genes that confer the ability to form nitrogen-fixing root nodules on host plants have frequently moved between *Rhizobium* species. It is not clear, though, whether such transfer is exceptional, or if frequent inter-species introgression is typical. To address this, we sequenced the genomes of 196 isolates of the *Rhizobium leguminosarum* species complex obtained from root nodules of white clover (*Trifolium repens*).

Results: Core gene phylogeny placed the isolates into five distinct genospecies that show high intra-genospecies recombination rates and remarkably different demographic histories. Most gene phylogenies were largely concordant with the genospecies, indicating that recent gene transfer between genospecies was rare. In contrast, very similar symbiosis gene sequences were found in two or more genospecies, suggesting recent horizontal transfer. The replication and conjugative transfer genes of the plasmids carrying the symbiosis genes showed a similar pattern, implying that introgression occurred by conjugative plasmid transfer. The only other regions that showed strong phylogenetic discordance with the genospecies classification were two small chromosomal clusters, one neighbouring a conjugative transfer system. Phage-related sequences were observed in the genomes, but appeared to have very limited impact on introgression.

Conclusions: Introgression among these closely-related species has been very limited, confined to the symbiosis plasmids and a few chromosomal islands. Both introgress through conjugative transfer, but have been subject to different types of selective forces.

Keywords: Rhizobia; white clover; genome assembly; introgression; conjugation

Background

The promiscuity of bacteria, and their ability to rapidly transfer DNA, has in the last years challenged microbiologists and geneticists seeking to integrate prokaryotes into standard models of speciation [1, 2, 3, 4]. The dynamic nature of acquisition, loss and transfer of genes in these organisms goes beyond the recombinational process and vertical inheritance, forcing a redesign of the speciation models for prokaryotes [5, 6, 7].

In contrast to most eukaryotes, which have mutation and meiotic recombination as the main adaptive drivers, bacterial species rapidly adapt through other types of genetic exchange: transformation (through the cell membrane), transduction (through a vector), and conjugation (cell-to-cell contact) [8, 9]. These processes can move adaptive genes between distantly related species, creating regions of high genetic similarity.

When describing prokaryotic genomes, an important distinction must be made between core and accessory genomes. The core genome is the set of ubiquitous genes within a defined group, such as a species. These genes often include housekeeping genes and are generally found in the chromosome. In certain species, core genes are also found on chromids, which are large plasmids that have acquired chromosomal characteristics [10, 11]. The accessory genome is a pool of non-ubiquitous genes that can provide a bacterial strain with adaptive advantages, for instance with respect to host interaction, antibiotic resistance, or heavy metal resistance [12, 13, 14]. The accessory genome is mainly found in the accessory plasmids, but also in islands in the chromosome and chromids.

Genetic divergence among closely related species can arise by ecological and genetic processes. Ecologically distinct niches may select genotypes with different adaptations [15, 16, 17]. This model, known as the ecotype model, is frequently observed in nature. In sympatric populations of the aquatic bacterioplankton of the family *Vibrionaceae* for example, phylogenetic differentiation was observed to be initiated by a change in ecological niche [18, 19].

Another possible factor for the isolation of sibling species is recombinational incompatibility [20, 16]. Multiple experimental studies of bacterial recombination have revealed that homologous recombination between prokaryotes may be restricted by sequence divergence between donor and recipient [21, 22], since sequence mismatches interfere with the process of recombination [23]. The relationship between recombination and sequence divergence produces a feedback loop on speciation: increased sexual isolation increases divergence, and genetic isolation prevents gene flow [24].

The intensity and the rate of homologous recombination during the process of prokaryotic genetic differentiation in prokaryotes is still unclear. While analyzing nucleotide sequences of *E. coli*, Visser and Rossez [25] observed that the spread of alleles through homologous recombination was restricted to small regions of the chromosome that carried advantageous information. These patterns could be explained by periodic selection events (selective sweeps) in the genome.

Another study that compared *Vibrio* species from very different ecological backgrounds [26] also concluded that ecological differentiation among species was driven by gene-specific rather than genome-wide selective sweeps, followed by gradual emergence of barriers to gene flow. The species described in this study were still at an early stage of ecological differentiation, and therefore genetic similarity across species was still high enough that interspecies recombination had not been fully inhibited. There is also extensive literature documenting the sharing of symbiosis-related genes among distinct, and sometimes distant, species of rhizobia, the nitrogen-fixing root-nodule symbionts of legumes [27, 28, 29]. This occurs whether the genes are on plasmids [30, 31, 32, 33] or on conjugative chromosomal islands [34, 35].

These events demonstrate that gene introgression has occurred in symbiotic soil bacteria, but it is not known to what extent the symbiosis genes, which are under strong selection because of the interaction with the plant host (reviewed by [36]), reflect the general behaviour of accessory genes. To address this question and obtain a more general understanding of introgression characteristics and mechanisms, we assembled 196 *R. leguminosarum* genome sequences, which comprised five distinct genospecies, and carried out a comprehensive introgression analysis.

Results

Identification and characterization of five distinct genospecies

A collection of 196 draft genome assemblies of *Rhizobium leguminosarum* is presented here. The strains were isolated from root nodules of white clover (*Trifolium repens*) in three different European countries and under two management regimes:

field trial sites in Denmark (DK), France (F), and the United Kingdom (UK), and organic fields in Denmark (DKO) (Additional file 1: Fig. S1 and S2, Additional file 2: Table S1). The genomes of seven strains were sequenced using PacBio and fully assembled into chromosome and plasmids. All 196 strains were sequenced using Illumina, and the assemblies were optimized using the PacBio complete genomes as references in order to determine, as far as possible, the correct order and orientation of contigs (Fig. 4 and 5, Additional file 2: Table S2).

Pairwise comparisons of average nucleotide identity (ANI) based on 282 bacterial conserved genes ([37], Additional file 2: Table S4) revealed clear clusters of genetic similarity (Fig. 1b). These clusters corresponded to the five genospecies described by Kumar et al., 2015 [38] genospecies (gs) A (33 strains), B (33), C (115), D (5) and E (10) (Additional file 1: Fig. S6). Overall, the pairwise similarity within a genospecies is above 96% and between genospecies below 96%. This is with the exception of genospecies D and E, which are on average 97% similar. Within each genospecies there are subclusters with varying degrees of distinctness, as shown in Fig. 1a. Analysis of types of genetic diversity (SNP, core ANI and gene presence/absence) showed similar patterns of structure, agreeing with the genospecies classification.

A total of 22,115 groups of orthologous genes were identified. Across all strains, a dichotomous pattern was observed: the majority of genes were either rare, shared by maximum 2 strains, or ubiquitous (Fig. 1d). Strains that were genetically close tended to have similar gene content, so that a pairwise comparison of gene sharing (Fig. 1c) resembles the core similarity matrices (Fig. 1a,b).

Even though these strains were collected from different countries (Denmark, United Kingdom and France) and soil managements (field trial sites and organic fields), the genetic diversity could not be fully explained by sample location (Fig. 1a-c).

In a Principal Component Analysis (PCA) of SNP variation, 43.97% of the variance was explained by the two first PCs, which separated the five genospecies (Fig. 1e). PC3 and PC4 revealed the genetic substructure within gsC, but also separated gsE and gsD more clearly (Fig. 1f).

Accessory and core genomes

We also assessed the core and accessory gene content (Fig. 2). Almost 20% of the genes (4,204) were shared by all strains (core genes). We observed clusters of genes that were characteristic of a single genospecies but absent elsewhere, as well as clusters confined to groups of related isolates within a genospecies (Fig. 2a).

The abundance of genospecies-private genes and genospecies-accessory genes was estimated (Fig. 2b). Even though gsD and gsE are closely related, only a small number of orthologous genes (116) are exclusive to them. The number of genospecies-private genes correlates with the genospecies sample size: for example, 4,969 genes are only found in gsC, the genospecies with the most members. Furthermore, pangenome analysis based on random addition of genomes showed that the gene pool of these populations can be considered as infinite, and that the inclusion of new genomes in the analysis would probably increase the accessory gene set indefinitely, but would not reduce the core genome significantly (Additional file 1: Fig. S7).

The nucleotide composition of the accessory genome was very distinct from that of the core genome (Fig. 2c). The median GC3 content (GC composition of third bases in codons) of the accessory genome (17,911 genes, 0.5704) was lower and significantly different from that of the core genome (4,204 genes, 0.6148). Differences in accessory and core GC3 content distribution were also observed between the chromosome and the two chromids (Additional file 1: Fig. S8, Additional file 2: Table S5).

Within-species variation

Variation within and between genospecies was investigated by characterizing nucleotide diversity, Site Frequency Spectra (SFS), Tajima's D, and decay of Linkage Disequilibrium (LD) with genomic distance (Fig. 3, see Methods).

The average nucleotide diversity differs by a factor of 5 among genospecies, and is higher for accessory than core genes and slightly higher for genes located on chromids compared to the chromosome (Fig. 3a). This is consistent with stronger purifying selection acting on essential genes.

The site frequency spectra are shown separately for synonymous and non-synonymous sites for genospecies A, B and C (Fig. 3b). Overall, the peaks of intermediate frequency SNPs reflect the population structure within each genospecies. For synonymous SNPs, the shape of the SFS differs among genospecies with genospecies C having a larger proportion of rare variants and genospecies A having a large proportion of intermediate frequency variants. This suggests different population demography of the genospecies, with genospecies C showing a signal of population expansion and genospecies A of population decline. This is reflected in positive values of Tajima's D for genospecies A and negative values for genospecies C (Fig. 3c). Contrasting synonymous and non-synonymous SFS for each genospecies we find a relative excess of rare non-synonymous variants consistent with segregation of non-synonymous variation under weak purifying selection.

We assessed the decay in intragenic linkage disequilibrium with distance using the r^2 measure of LD ([39] see details in Methods). In all genospecies there is a rapid decay of LD within the first 1000 base pairs, suggesting a very high rate of recombination within genospecies. The less dramatic decay in genospecies B may either reflect a lower per generation recombination rate or a lower population size consistent with its low level of nucleotide diversity.

Full PacBio assemblies gave us an opportunity to precisely explore structural variation across genospecies. Multiple alignments of representative strains from each genospecies revealed high chromosomal collinearity (Additional file 1: Fig. S9).

From all 196 genomes, 24 distinct RepA sequence groups were identified. However, four of these correspond to isolated *repA*-like genes that are not part of *repABC* operons, and twelve others are rare (in no more than four genomes), so eight types account for nearly all the plasmids (Fig. 4a). We numbered them Rh01 to Rh08 in order of decreasing frequency in the set of genomes. Of these, Rh01 and Rh02, corresponding to the two chromids pRL12 and pRL11 of the reference strain 3841 [10], are present in every genome. The distribution of the other plasmids shows some dependence on genospecies, but none is confined to a single genospecies. For example, Rh03 is present in all strains of gsA, gsB and gsC, but absent from gsE and in just one gsD strain, while Rh05 is universal in gsA and gsB but absent elsewhere. The phylogeny of *repA* genes within individual plasmid groups sheds light on their history of transfer between and within genospecies. In groups Rh01 to Rh05, each clade in the phylogeny contains strains of a single genospecies, providing no evidence for recent transfer of these plasmids between genospecies.

Symbiosis genes are found on Rh04, Rh06, Rh07 and Rh08 plasmids, depending on genospecies. Not all symbiosis genes are on scaffolds with *repABC* genes, because of incomplete genome assembly, but the overall picture is clear. Genospecies A symbiosis plasmids are all Rh06, in gsB they are Rh07, gsC has mostly Rh04 but some Rh07 and Rh08, gsD has Rh08, gsE has mostly Rh08 but some Rh06 and Rh07. There are striking differences in the apparent mobility of these plasmids. Conjugal transfer genes (*tra* and *trb*) are present in some Rh04 plasmids and all Rh07 and Rh08 plasmids, including those that are symbiosis plasmids. These genes are all located together immediately upstream of the *repABC* replication and partitioning operon, in the same arrangement as in the plasmid p42a of *R. etli* CFN42, which has been classified as a Class I, Group I conjugation system [40].

Interestingly, some *repA* sequences of sym plasmids from strains of different genospecies are identical or almost identical in sequence (Fig. 4b and Additional file 1: Fig. S10). The phylogenies of the corresponding conjugal transfer genes (e.g. *traA*, *trbB* and *traG*) show the same pattern (Additional file 1: Fig. S11), indicating that symbiosis plasmids have introgressed across genospecies boundaries through conjugation.

HGT and intergenic Linkage Disequilibrium

Different modes of genetic exchange are expected for the different genomic compartments (chromosome, chromids and plasmids), so the rates of DNA exchange in the symbiosis plasmid cannot be directly correlated to the rates for other plasmids. Hence, we evaluated patterns of intergenic linkage disequilibrium (LD) in the different compartments as a proxy for recombination. High rates of recombination would reduce the genetic correlations between genes, unless genes or genomic compartments have been recently acquired.

Strong patterns of relatedness in this data can produce biased estimates of LD, so population structure adjusted genotype matrices were used to estimate LD (see details in Methods). Genome-wide pairwise comparisons between all genes ordered by plasmid origin demonstrated different intensities of recombination in the different genomic compartments (Fig. 5a). High intergenic correlations were restricted to genes within each compartment; few inter-compartment interactions were observed. Interestingly, we found that the symbiosis plasmids maintained high levels of intergenic LD, suggesting that this plasmid has been recently acquired (Fig. 5b).

Intergenic LD between all pairs of symbiosis genes showed clear blocks of linkage disequilibrium similar to those that have been previously described [41] (Fig. 5c). The small LD blocks within the symbiosis cluster agree with functionality: nod genes are required for infection and nodule organogenesis, *nifHDKEN* genes encode the nitrogenase enzyme, and the other *nif* and *fix* genes are needed to support symbiotic nitrogen fixation [42]. Intergenic LD before and after correction for population structure showed how structure can introduce noise and overestimate intergenic LD (Additional file 1: Fig S12-S13) [43, 44]. No strong evidence for high LD between symbiosis genes and other genes from different genomic compartments was found.

Evidence for sym-plasmid transfer between genospecies was also observed when analyzing phylogenetic patterns of symbiosis genes in contrast to the species tree (Fig. 6a, Additional file 1: Fig. S14). Certain clades of identical sequences in single gene phylogenies included members of different genospecies (Fig. 6b-d), meaning that these strains shared alleles with strains from other genospecies than their own. Interestingly, the majority of these strains originated from organic fields.

In order to understand if genomic introgression among these sibling species was restricted to the sym plasmid, analysis of the evolutionary history of single genes was conducted. We calculated the discordance between the gene trees and the genospecies classification (discordance score, Additional file 1: Fig. S15; Methods). If a gene tree resembles the genospecies topology of the species tree, where distinct clades of genospecies are observed, then the gene would have a zero discordance score. The results showed that around 20% of the genes have no evidence for transfer between genospecies (discordance equal to zero), 35% have a discordance score of 1, and 16% have a discordance score of 2, indicating that the majority of the genes closely follow the species phylogeny. Symbiosis genes are in the tail of this distribution with a discordance score above 6 (Fig. 7a), in accordance with our expectations based on our observation of sym-plasmid introgression.

Population genetic parameters were contrasted between symbiosis genes and other classes of gene (Table 1, Additional file 2: Table S6). The results show that the level of polymorphism overall is similar for symbiosis genes and other genes but that the diversity is distributed differently. In symbiosis genes, identical or near-identical haplotypes are more often observed even across several genospecies (Fig. 6). However, several distinct groups of haplotypes exist yielding a very high Tajima's

D for symbiosis genes (Additional file 2: Table S7). This suggests either selective sweeps within these groups, some form of balancing selection among groups, or a combination of both.

By plotting discordance scores to gene locations based on a PacBio reference genome (SM3), we observed that highly introgressed genes are concentrated in the smaller plasmids (Fig. 7b). This reflects the most frequent mode of exchange of the symbiosis plasmids, where entire sym-plasmids are transferred through conjugation [45]. On the other hand, patterns of introgression on the chromosome are restricted to small regions, showing evidence of linkage blocks. The functionalities and origin of the chromosomal introgression islands were further investigated.

Chromosomal introgression is restricted to few events

We identified two specific chromosomal regions where introgression events predominantly occur. Cluster 1 (Fig. 8b and c, Additional file 2: Tables S8 and S9) was consistently found in the same region in 87 strains (64 gsC, 23 gsB) downstream of a core phasin gene. The cluster comprises two regions of accessory genes with higher than average discordance scores flanking a region of core genes that probably travels with them and also has elevated discordance (Fig. 8b, Fig. S16).

Cluster 1 encodes a type IV secretion system (T4SS) in many strains, and this T4SS bears a striking resemblance to one of the three T4SSs of *Agrobacterium tumefaciens* C58. Two of these systems, Trb and AvhB, mediate conjugal transfer of Ti and pAtC58 plasmids, respectively, between *Agrobacterium* cells, whereas the third system, VirB, transfers DNA from *Agrobacterium* to host plant cells [46, 47]. The overall structure of the cluster 1 T4SS genes most closely resembles that of the *avhB* system, which includes 10 genes homologous to the *virB* operon and a DNA transfer and replication (Dtr) system comprising *traG*, *traD*, *traC*, and *traA* (Fig. 8a). There is a full *avhB* cluster inserted after the phasin gene in 64 out of 87 strains (for example, SM3 in Fig. 8c), whereas 23 strains lack the *traC* homolog in the Dtr (for example, SM121B). One or two nucleotidyltransferase genes, a traA relaxase gene, and DNA polymerase gene are conserved downstream of the avhB cluster and in synteny within the introgressed region.

Not all strains have *avhB* in cluster 1: 5 strains, including SM170C and SM153D (Fig. 8c) have a DNA rearrangement system that includes an ATP-dependent DNA ligase, a metallophosphatase superfamily gene, and a high number of hypothetical proteins (Additional file 2: Table S9). In 104 strains there was no insert at the start of cluster 1. All strains have a discordant cluster of polysaccharide metabolism genes, which seems to travel with the chromosomal island, but these genes are distinctive in strains without the initial insert, such as SM4 and SM100 (Fig. 8c).

Cluster 2 (Fig. 8d, Table S8) was found in all 196 strains. It contained a large number of hypothetical proteins, many of which contained conserved domains corresponding to transposases and integrases. No obvious DNA transfer mechanism that could mediate the transfer between genospecies was discovered in this island. However, we observed toxin-antitoxin (*VapC*/*YefM*) genes within this cluster; these represent the type II toxin-antitoxin system, which is a homologue of T4 RNase H with a PIN domain [48] and is thought to move from one genome to another by horizontal gene transfer [49].

We have also evaluated population genetic parameters of highly discordant chromosomal genes (Additional file 2: Table S8). In contrast to the symbiosis genes, chromosomal introgressed genes have lower than average Tajima's D values that are not significantly different from zero, which suggests that these genes are evolving as expected under neutrality.

Other modes of genetic exchange

Phage-mediated introgression is another mechanism of horizontal gene transfer that could drive gene introgression between bacterial strains and even genospecies. It is well known that during transduction, bacterial host genetic material can be transported to another bacterium by incorporation into phage vectors [50]. Additionally, a greater similarity between genomes has been suggested to increase the probability of successful introgression by transduction, although both trans-species and trans-genus DNA transduction has been known to occur [51].

In order to evaluate the extent of phage-mediated gene transfer between genospecies, we used PHASTER [52, 53], an online platform for prophage annotation in bacterial genomes. This identified 344 unique homologous phage protein families from our 196 Rhizobium genomes (Additional file 1: Fig. S17a, Additional file 2: Table S10). The most abundant phage protein identified was a putative portal protein homologous to that in *Brucella* phage Pr (gi418487847), which is an essential component of stable DNA encapsidation [54].

Phylogenetic analysis shows that individual homologous phage proteins have the tendency to cluster by genospecies; however, due to high conservation of protein sequences, different genospecies are found in the same clades. We therefore speculate that phages have the ability to transduce between genospecies, but are more often transducing within genospecies where strains are more genetically similar (Additional file 1: Fig. S18b).

Furthermore, to confirm that observed chromosomal gene introgressions were not predominantly a consequence of phage-mediated introgression, we calculated the base pair (bp) distance between phage proteins and the two chromosomal cluster regions. Only six strains (2 gsA, 4 gsC) out of 87 contained phage proteins closer than 15,000bp to the cluster 1 start site. Three gsC and the two gsA strains had phage proteins located upstream of the cluster start site, and only one gsC strain had identified phage proteins downstream. The two gsA strains and one gsC strain incorporated phage proteins 3,000-5,000 bp upstream from the cluster 1 start site. These proteins were identified as transposases (gi209447153, gi26989834, gi17546153, gi209447152, gi209447153). However, cluster and phage presence are not concordant, and 25 of 87 strains possessing the cluster had no identifiable prophage regions in their genomes. Similarly, strains sharing homologous phage proteins did not necessarily have the gene cluster.

Only the two strains from gsA (SM154C and SM163B) showed potential evidence for recent phage introgression near the cluster, with four orthologous phage proteins located exactly the same base pair distance from the cluster start site in both strains.

Discussion

Five related but distinct genospecies can be found in sympatry

We have assembled the genomes of 196 *Rhizobium leguminosarum* strains, which were isolated from root nodules of white clover (*Trifolium repens*) in three different European countries and under two management regimes: field trial sites in Denmark (DK), France (F), and the United Kingdom (UK), and organic fields in Denmark (DKO). Multiple samples from the same field were collected in order to capture as much of the genetic variation as possible. Based on the analysis of SNPs, we observed clear patterns of genomic clustering into five genospecies as previously reported [38] (Figure 1a). The average nucleotide identity of conserved core genes and the number of shared orthologous genes (Fig. 1b and c) also reflected the five distinct genospecies. Multiple genospecies were observed at the same field site, as previously reported [38]. The distinct genospecies thus coexist in sympatry, but remain genetically well separated.

The core genomes of the genospecies are completely diverged

Although sympatry is observed, analysis of individual gene trees showed that horizontal gene transfer has been mainly confined to symbiosis plasmids and two chromosomal islands. The occurrence of HGT of symbiosis genes within and between distant rhizobia genera (*Rhizobium*, *Bradyrhizobium*, *Sinorhizobium*, *Azorhizobium*, and *Mesorhizobium*), nodulating different legume species, has been widely reported [33, 55, 27, 56, 29]. This shows that symbiosis gene transfer is not restricted by genetic divergence and in many cases is not species specific [57]. Studies comparing rhizobial genera have shown that HGT of the symbiosis apparatus occurred through the transfer of symbiosis plasmid (pSym) or genomic islands [58, 59, 60, 61].

The genetic differentiation maintained in the core genome of these genospecies could have been caused by rather high rates of within-genospecies compared to interspecies homologous recombination [62, 33, 4]. Based on intragenic LD analysis (Fig 3c), we observed LD decay that is indicative of fairly high rates of within-genospecies homologous recombination [38, 33, 63]. Interspecies recombination may be restricted by the genetic divergence between strains, and this is an important factor in speciation of many prokaryotes (*Vibrio* [19, 26]; *Rhizobium* [33] and *Salmonella enterica* [64]).

Selection also plays an important role in shaping genospecies divergence. We have shown here that the genospecies have remarkably different demographic histories and, therefore, have been affected differently by purifying selection (Fig 3a and 3b). Despite clear genetic differentiation, these strains have maintained very syntenic chromosomes and chromids (Rh01 and Rh02). The chromids have genomic signatures (GC content, nucleotide diversity composition, low interspecies recombination) that more closely resemble those of the chromosome than of the plasmids [11, 37, 33]. The strong conservation of the genomic organization highlights the essential nature of core genes and the possible selective constraints preventing genomic rearrangements and HGT [65]. By contrast, plasmids are more plastic, with multiple rearrangements and lower median GC content (Fig. 2c). This may reflect differences in selective pressures, with core genes being subject to stronger purifying selection compared to the accessory genome [66].

Symbiosis gene introgression is driven by conjugative plasmids

The genospecies studied here displayed a diverse set of plasmid profiles (Fig. 4a), as has been previously described in these and other *Rhizobium* species [10, 67, 68]. The distribution of these plasmids shows some dependence on genospecies, but no plasmid type is confined to a single species, and plasmids therefore seem to have been transferred among genospecies. Symbiosis plasmids can belong to any of a number of plasmid types (Rh04, Rh06, Rh07 and Rh08), and phylogenetic evidence indicated that some of them have been transferred through conjugation between different genospecies (Fig. 4b). These transfers are likely recent, since the sequences have not yet diverged at all. Because conjugation requires cell-to-cell contact, it is evident that plasmid transfer is not just constrained by genetic similarity [69, 33], but also by the requirement that donor and recipient are found in the same location, again underlying the sympatric nature of these sibling species.

Chromosomal introgression events were detected based on phylogenetic discordance

Evidence for sym-plasmid transfer between genospecies was also observed when analyzing phylogenetic patterns of symbiosis genes in contrast to the species tree (Fig. 6). These results led us to develop a phylogenetic method that calculates discordance scores based on gene tree deviations from the overall genospecies classification. Many phylogenetic [70, 71, 72] and parametric methods [73, 74, 75] have been previously used to detect HGT events. Parametric methods characterize sequence composition (GC content, codon usage, sequence conservation) and search for regions of the genome that significantly deviate from the genomic average [76]. These approaches rely on the uniformity of the host signature and on a relative distant origin of the exogenous sequences [73]. For many HGT events these assumptions are unrealistic, especially when dealing with ancient DNA acquisitions [77, 78]. On the other hand, phylogenetic methods can integrate information from multiple genomes using a specific evolutionary model [76]. The comparison of a large number of genomes, combined with a well-defined species tree and carefully pruned orthologous gene groups, gave us enough power to confidently find genes strongly deviating from the species phylogeny.

Based on our phylogenetic method, we identified two events of chromosomal introgression where clusters of genes were transferred between genospecies. Cluster 1 includes genes that bear a striking resemblance to the *Agrobacterium tumefaciens* AvhB type IV secretion system that mediates the transfer of a small plasmid (pAtc58) to a donor cell [47]. Therefore we hypothesize that the transfer of this chromosomal island is mediated by the combination of a full VirB conjugative system and a *tra* DNA transfer and replication system [79].

The *avhB* gene cassette and the *traG* gene in cluster 1 also show similar organisation to a conjugative transfer system encoded by the *virB/traG* of the plasmid pSymA of *S. meliloti* [80, 81] and to the *virB/virD4* of *Bartonella tribocorum* [82]. However, both T4SSs in *A. tumefaciens* and *S. meliloti* (AvhB and VirB, respectively) mediate the transfer of whole plasmids, whereas we are proposing that the T4SS encoded in cluster 1 mediates the transfer of an integrative conjugative element (ICE). Other integrative and conjugative elements have been observed in the rhizobial genera (*Azorhizobium caulinodans*: [61], *Sinorhizobium*: [83]) and in other species (*Streptococcus agalactiae*: [84], *Bacillus subtilis*: [85], *V. cholerae*: [86]).

In cluster 2 we found toxin-antitoxin (TA) genes located within the cluster, but we could not determine a putative transfer mechanism. The maintenance of integrative conjugative elements (ICE) is in many cases mediated by the presence of functional toxin-antitoxins [87, 88, 89]. The loss of these TA genes causes a post-segregational killing of the bacterial cell by the toxin's destructive effect [88]. Chromosomally-encoded TA systems have been shown to protect against large-scale deletion of genomic islands [90], but have also been reported to have different functions in the host [88].

Mobile genetic elements (MGEs), such as sym plasmids and ICEs, are important for the evolution of bacterial species, since a single event (conjugation of entire mobile plasmids or insertion of gene sets) can introduce a whole set of new functions to the recipient that can drastically change its lifestyle (*e.g.* from free-living bacterium to symbiont) [61]. Many of these genes in the chromosomal islands may not confer any adaptive advantage, and they have possibly hitch-hiked along with proximally located positively selected genes. This could be the reason that we see striking discordance peaks in the two chromosomal islands (Fig. 7b). MGEs can also be viewed as elements with independent evolutionary trajectories to their host. The presence of a toxin-antitoxin system placed close to the second cluster shows one of the possible strategies that these elements deploy to increase their own fitness and vertical propagation.

Our results indicate that conjugation is the predominant mechanism of introgression among the five genospecies, but we also investigated the effect of phage-mediated transduction. Despite the presence of prophage sequences within the majority of the genomes, we found that phage-related proteins were not linked to the chromosomal islands and did not have high discordance scores (Additional file 1: Fig. S18). While genetic transduction is known to be a important mechanism for bacterial adaptation in many different species (*P. aeruginosa*: [91]; *Escherichia coli*: [92], *Staphylococcus aureus* [93]), phages do not appear to play a dominant role in gene introgression for our set of *R. leguminosarum* strains.

Symbiosis genes and genomic islands introgressed independently

Since we found a very limited number of major introgression events, we investigated whether they might all be related to symbiosis gene transfer. We first examined this by exploring intergenic linkage disequilibrium by applying the Mantel test to pairs of gene genetic relationship matrices (GRM's) using population-structure corrected markers, which reduced the overestimation of genetic linkage due to population structure (Additional file 1: Fig. S11-S12). Although we observed high linkage disequilibrium within sym-clusters, symbiosis genes did not appear to be linked to the chromosomal islands (Fig 6).

We found significantly positive values of Tajima's D for the symbiosis genes, which indicates the presence of several distinct groups of haplotypes. This distinguished the symbiosis genes not just from the core genome, but also from most of the accessory gene set (Fig. 7, Table 1). Evidence for similar balancing selection of symbiosis genes was previously reported for *Rhizobium leguminosarum bv. viciae* [94]), whereas purifying selection was observed in the *nod* gene region of *Sinorhizobium medicae* [95]. In contrast, the introgressed chromosomal islands did not seem to have been subject to strong selective pressures, since the majority of the introgressed genes in these regions did not show Tajima's D values significantly different from zero. The lack of genetic linkage and different selection signatures suggest that the symbiosis plasmids and chromosomal islands introgressed independently.

Conclusions

Five genospecies in the *R. leguminosarum* species complex are frequently sympatric but maintain distinct genetic variants of their core genes, demonstrating a lack of significant introgression in the core genome. Many accessory genes are found across two or more genospecies but, surprisingly, their phylogenies indicate that most of them have no recent history of introgression between genospecies. Striking exceptions are the genes sitting in symbiosis plasmids, especially the symbiosis genes, and two small chromosomal islands of unknown function.

Methods

Rhizobium sampling and isolation

White clover (*Trifolium repens*) roots were collected from three different breeding trial sites in the United Kingdom (UK), Denmark (DK), and France (F) (Additional file 1: Fig. S1A), and 50 Danish organic fields (DKO) (Additional file 1: Fig. S1b). Roots were sampled from 40 different plots from each trial site. The total number of plots was 170. The samples were stored at ambient temperature for 1-2 days and in the cold room (2°C) for 2-5 days prior to processing. Pink nodules were collected from all samples, and a single bacterial strain was isolated from each nodule as described by [96]. From each plot, 1 to 4 independent isolates were produced. In total 249 strains were isolated from *T. repens* nodules. For each site the clover varieties were known, and representative soil samples from clover-free patches were collected and sent for chemical analysis. Furthermore, site-specific geographic information system (latitude and longitude) were collected (Additional file 2: Table S1).

Genome assembly

A representative set of of 196 strains was subjected to whole genome shotgun sequencing using 2x250 bp Illumina (Illumina, Inc., USA) paired-end reads by MicrobesNG ([97], IMI - School of Biosciences, University of Birmingham). In addition, 8 out of the 196 strains were re-sequenced using PacBio (Pacific Biosciences of California, Inc., USA) sequencing technology (Additional file 2: Table S2, Additional file 1: Fig. S2). Analysis of 16S rDNA confirmed that all 196 of the strains were *Rhizobium leguminosarum.*

Genomes were assembled using SPAdes (v. 3.6.2) [98]. SPAdes contigs were cleaned and assembled further, one strain at a time, using a custom Python script (Jigome, available at [99]). First, low-coverage contigs were discarded because they were mostly contaminants from other genomes sequenced in the same Illumina run. The criterion for exclusion was a SPAdes k-mer coverage less than 30% of the median coverage of putative single-copy contigs (those $>$ 10kb). Next, putative chromosomal contigs were identified by the presence of conserved genes that represent the syntenic chromosomal backbone common to all *R. leguminosarum* genospecies. A list of 3215 genes that were present, in the same order, in the chromosomal unitigs of all eight of the PacBio assemblies was used to query the Illumina assemblies using *blastn* ($\geq$90% identity over $\geq$90% of the query length). In addition, contigs carrying *repABC* plasmid replication genes were identified using a set of *RepA* protein sequences representing the twenty distinct plasmid groups found in these genomes (*tblastn* search requiring $\geq$95% identity over $\geq$90% of the query length). A 'contig graph' of possible links between neighbouring contigs was created by identifying overlaps of complete sequence identity between the ends of contigs. The overlaps created by SPAdes were usually 127 nt, although overlaps down to 91 nt were accepted. Contigs were flagged as 'unique' if they had no more than one connection at either end, or if they were $>$ 10 kb in length. Other contigs were treated as potential repeats. The final source of information used for scaffolding by Jigome was a reference set of *R. leguminosarum* genome assemblies that included the eight PacBio assemblies and 39 genomes publicly available in GenBank [99]. A 500-nt tag near each end of each contig, excluding the terminal overlap, was used to search this database by blastn; high-scoring matches to the same reference sequence, with the correct spacing and orientation, were subsequently used to choose the most probable connections through repeat contigs. Scaffolding was initiated by placing all the chromosomal backbone contigs in the correct order and orientation, based on the conserved genes that they carried, and extending each of them in both directions, using the contig graph and the pool of remaining non-plasmid contigs, until the next backbone contig was reached or no unambiguous extension was possible. Then each contig carrying an identified plasmid origin was similarly extended as far as possible until the scaffold became circular or no further extension was justified, and unique contigs that remained unconnected to chromosomal or plasmid scaffolds were extended. Finally, scaffolds were connected if their ends had appropriately spaced matches in the reference genomes. Scaffold sequences were assembled using overlap sequences to splice adjacent contigs exactly, or inserting an arbitrary spacer of twenty ”N” symbols if adjacent contigs did not overlap. The *dnaA* gene (which was the first gene in the chromosomal backbone set and is normally close to the chromosomal origin of replication) was located in the first chromosomal scaffold, and this scaffold was split in two, with chromosome-01 starting 127 nt upstream of the ATG of *dnaA* and chromosome-00 ending immediately before the ATG. The remaining chromosomal scaffolds were numbered consecutively, corresponding to their position in the chromosome. Plasmid scaffolds were labelled with the identifier of the *repA* gene that they carried. Scaffolds that could not be assigned to the chromosome or a specific plasmid were labelled ’fragment’ and numbered in order of decreasing size. Subsequent analysis revealed large exact repeats in a few assemblies. These were either internal inverted repeats in the contigs created by SPAdes (5 instances) or large contigs used more than once in Jigome assemblies (18 instances). They were presumed to be artifacts and removed individually.

Assembly statistics were generated with QUAST (v 4.6.3, default parameters) [100]. (Additional file 1: S3). Genes were predicted using PROKKA (v 1.12) [101]. In summary, genomes were assembled into [10-96] scaffolds, with total lengths of [8355366-6967649] containing [6,642-8,074] annotated genes, indicating that we have produced assemblies of reasonable quality, which comprehensively captured the gene content of the sequenced strains (Additional file 2: Table S2 and S3).

Orthologous genes prediction

Orthologous gene groups were identified among a total of 1,468,264 gene products present across all (196) strains. We used two different software packages for ortholog identification: Proteinortho [102] and Syntenizer3000 [103]. The software Proteinortho [102, 104] (v5.16b), was executed with default parameters and the synteny flag enabled, to predict homologous genes while taking into account their physical location. For the analysis in this paper, we were only interested in orthologs and not paralogs. Paralogous genes predicted by Proteinortho were carefully filtered out by analyzing the synteny of homologous genes surrounded by a 40-gene neighbourhood (see Synteny section). After this filtering step, the orthologous gene groups were aligned using ClustalO ([105], v. 1.2.0). Each gene sequence was translated to its corresponding amino acid sequence before alignment and back-translated to the original nucleotides. Each gap was replaced by 3 gaps, resulting in a codon-aware nucleotide alignment. Manual check of highly diverse genes (nucleotide diversity > 0.2) was conducted. We observed that many of these genes were composed of fragmented/partial genes, wrongly assigned orthologous groups, composed of few taxa and were enriched for ”hypothetical proteins” annotation. Therefore, for the population genetic analysis we filtered out these possibly problematic genes with a ANI cutoff equal to 0.65.

Synteny

First, gene groups were aligned with their neighbourhoods (20 genes each side) using a modified version of the Needleman-Wunsch algorithm [106]. We counted the number of gene neighbours that were syntenic across strains before a collinearity break. We used this score to disambiguate gene groups that contain paralogs. Paralogs are the result of gene duplication, and as such one of the paralogs is the original, and the rest are copies. Based on similarity, we kept the least divergent gene inside of the original homology group while removing the copied paralogs, if possible into a new gene group. Orphan genes, that were present only in one strain, were removed from the analysis.

Variant Calling

Codon-aware alignments were used in order to detect single nucleotide polymorphisms (SNPs). For a given gene alignment (individuals as rows and sequence as columns) and position, we first counted the number of unique nucleotides (A, C, T, G). Columns containing 2 unique nucleotides were considered variable sites (bi-allelic SNPs). After finding variable sites, SNP matrices were encoded as follows: major alleles were encoded as 1 and minor alleles as 0. Gaps were replaced by the column mean. Later steps were executed in order to filter out unreliable SNPs. We restricted the analyses to genes found in at least 100 strains. By looking at the variants and their codon context, we excluded SNPs placed in codons containing gaps, or containing more than one SNP, or with multi-allelic SNPs. Based on these criteria we ended up with 6,529 genes and 441,287 SNPs. Scripts and pipelines are available at a github repository [107].

Plasmid replicon groups

Plasmid replication genes (repABC operons) were located in the genome assemblies by *tblastn*, initially using the RepA protein sequences of the reference strain 3841 as queries. Hits covering ≥70% of the query length were accepted as repA genes, and those with ≥90% amino acid identity were considered to belong to the same replication group (putative plasmid compatibility group). Hits with lower identity were used to define reference sequences for additional groups, using sequences from published *Rhizobium* genomes when available, or from strains in this study. Groups were numbered (Rh01, etc) in order of decreasing abundance in the genome set. RepB and RepC sequences corresponding to the same operons as the RepA references were used to check whether the full *repABC* operon was present at each location, requiring ≥85% amino acid identity.

Presence of symbiosis genes in all strains

Since all sequenced strains were isolated from white clover nodules, they are expected to carry the canonical symbiosis genes. One strain, SM168B, carried no symbiosis genes. Subsequent nodulation tests showed that the strain could colonize white clover and produce pink nodules, suggesting that the genes were lost during the pre-sequencing processing. On the other hand, strains SM165B and SM95 were found to have duplicated symbiosis regions.

Average nucleotide identity of core genes

In order to place 196 strains into the previously described genospecies [38], a phylogenetic tree was first constructed based on a single gene (*rpoB*) (Additional file 1: Fig. S6). The tree contained representative genospecies identifiers and the RpoB sequence alignment of each strain member. After classification of genospecies, we calculated pairwise average nucleotide identity (Fig. 1B) based on the concatenation of 282 core bacterial genes (331617 bp) of chromid-bearing bacteria established by Harrison et al. 2010 (Additional file 2: Table S4).

Pangenome

Pangenome analyses were based on comparisons of orthologous gene families by carefully excluding singletons of each strain. A variance measure was added by randomly permuting the order of strains 20 times.

Principal Component Analysis

Principal Component Analysis was based on a total of 6,529 genes that were present in at least 100 strains (441287 SNPs). A minimal minor allele frequency threshold of 0.10 was used to filter out rare variants. Individual gene covariances were then computed as follows:

Let N denote the total number of individuals and M the total number of markers, the full genotype matrix (X) for a given gene has $N \times M$ dimensions with genotypes encoded as $0's$ and $1's$ for the N haploid individuals. Each column S_i $(i = 1, \ldots, M)$ of the X matrix is a vector of SNP information of size N. The first step of the calculation was to apply a Z-score normalization to each SNP vector by subtracting by its mean and dividing it by its standard deviation: $\left(\frac{S_i - \bar{S}_i}{\sqrt{Var(S_i)}}\right)$, this results in a vector with mean 0 and variance 1, where SNPs are assumed to be independently sampled from a distribution with covariance matrix V. We then computed the covariance matrix between individuals as follows:

$$Cov(X_i) = \hat{V} = \tfrac{1}{M-1} \sum_{i=1}^{M} (X_i - \bar{X})(X_i - \bar{X})'$$

$Cov(X)$ can also be computed by the dot product of the full genotype matrix:

$$Cov(X) = \hat{V} = XX'$$

The result is an $N \times N$ matrix, where N is the number of strains. This matrix is also known as the Genomic Relationship Matrix (GRM) [108]. We then decomposed the GRM using the linalg function of scipy (python library).

Population genetic analysis

Population genetic parameters (Tajima's D, nucleotide diversity, average pairwise differences (π) and number of segregating sites) were estimated using the python library dendropy [109].

Intragenic LD

Intragenic linkage disequilibrium (LD) measures the dependence between SNPs within a gene and it was estimated using Pearson's r^2 correlation measure. This analysis was done within each population, therefore, we did not use the corrected genotype matrices.

Each individual genotype matrix (containing a minimal set of 3 SNPs) was first normalized as described in the PCA section. After this normalization, each SNP contributes equally to the downstream analysis. LD was then calculated as a function of distance d (maximum 2000 base pairs apart) and was computed as the average LD of SNPs d base pairs away from each other. The calculations were done in the following way:

$$Cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

$$r^2 = Cor(X_i, X_j)^2$$

In which $j > i$ and X_i is composed of the genotypes of all individuals of a given genospecies for position i in the genotype matrix. X_j is composed of the genotypes of all individuals of the same genospecies for position j in the genotype matrix, and $d = j - i$ and d $\leq$ 2000 base pairs. Results were summarized into bins of size 10.

Intergenic Linkage Disequilibrium corrected for population structure

Sample structure or relatedness between genotyped individuals leads to biased estimates of linkage disequilibrium (LD) and increase of type I error. In order to correct for the autocorrelation present in this data, the genotype matrix X (coded as $0's$ and $1's$) was adjusted as exemplified in [110]. The covariance V between individuals was calculated first (as shown in the Principal Component Analysis section). Then the 'decorrelation' of genotype matrix X was done by multiplying X by the inverse of the square root of $\hat{V}$ as follows:

$$T_i = \hat{V}^{-\frac{1}{2}} X_i$$

T is therefore the pseudo SNP matrix, which is corrected for population structure. The correlation between genes matrices was obtained by applying a Mantel test to the GRM (genetic distances) between pairs of genes:

For a data set composed of a distance matrix of gene X (D^x_{ij}) and a genetic distance matrix of gene Y (D^y_{ij}), the scalar product of these matrices was computed, adjusted by the means and the variances ($Var(X)$ and $Var(Y)$) of the matrices X and Y:

$$r_{cor} = \frac{\sum(D^x_{ij} - \bar{X})(D^y_{ij} - \bar{Y})}{\sqrt{Var(X)Var(Y)}}$$

The standardized Mantel test is actually the Pearson correlation between the elements of genes X and Y.

Discordance Score

Individual gene trees were first constructed using the neighbour-joining clustering method (software RapidNJ version 2.3.2) [111]. Each tree was traversed based on depth first traversal algorithm, by visiting each node after visiting its left child and before visiting its right child, searching deeper in the tree whenever possible. When the leaf of the tree was reached, the strain number and its genospecies origin were extracted. A list containing the genospecies was stored for the entire tree. The discordance score was computed as following:

$$\text{Discordance score} = \#\text{shifts -set(genospecies)} + 1$$

The discordance score evaluates the number of times a shift (from one genospecies to another) is observed in a branch. The minimum possible is the total number of genospecies -1 shifts. A tree congruent to the species tree must have a discordance score equal to zero. (Additional data 1: Fig. S15).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceptualization: MIAC, JPWY, SM, MHS and SUA; Methodology: MIAC, JPWY and SM; Software: MIAC, AB, BV, JPWY and CM; Validation: MIAC, CM, SM, JPWY; Formal Analysis: MIAC, JPWY, CM, SM, AB, BV and BF; Investigation: SM; Resources: SUA, JPWY and MHS; Data Curation: MIAC, CM, JPWY, SM, SUA and MHS; Writing - Original Draft: MIAC; Writing - Review and Editing: MIAC, JPWY, SUA, MHS, SM, BV; Visualization: MIAC, SM, JPWY; Supervision: SUA, JPWY, MHS; Project Administration: SUA; Funding Acquisition: SUA.

Acknowledgements

This work was funded by grant no. 4105-00007A from Innovation Fund Denmark (S.U.A.). Genome sequencing was provided by MicrobesNG [97], which is supported by the BBSRC (grant number BB/L024209/1). The authors would also like to thank industrial partners DLF Trifolium, SEGES and Legume Technology Ltd. for their contribution to the field trials.

Author details

[1]Bioinformatics Research Center, Aarhus University, Aarhus, Denmark. [2]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. [3]Department of Biology, University of York, York, United Kingdom.

References

1. Lawrence, J.G.: Gene transfer in bacteria: speciation without species? Theoretical population biology **61**(4), 449–460 (2002)
2. Gogarten, J.P., Doolittle, W.F., Lawrence, J.G.: Prokaryotic evolution in light of gene transfer. Molecular biology and evolution **19**(12), 2226–2238 (2002)
3. Cohan, F.M.: Bacterial species and speciation. Systematic biology **50**(4), 513–524 (2001)
4. Cohan, F.M.: Bacterial speciation: genetic sweeps in bacterial species. Current Biology **26**(3), 112–115 (2016)

5. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., Hanage, W.P.: The bacterial species challenge: making sense of genetic and ecological diversity. science **323**(5915), 741–746 (2009)
6. Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L., *et al.*: Re-evaluating prokaryotic species. Nature Reviews Microbiology **3**(9), 733 (2005)
7. Shapiro, B.J., Leducq, J.-B., Mallet, J.: What is speciation? PLoS genetics **12**(3), 1005860 (2016)
8. Ochman, H., Lawrence, J.G., Groisman, E.A.: Lateral gene transfer and the nature of bacterial innovation. nature **405**(6784), 299 (2000)
9. Hanage, W.P.: Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harbor perspectives in biology, 018069 (2016)
10. Young, J.P.W., Crossman, L.C., Johnston, A.W., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Wexler, M., Curson, A.R., Todd, J.D., Poole, P.S., *et al.*: The genome of rhizobium leguminosarum has recognizable core and accessory components. Genome biology **7**(4), 34 (2006)
11. diCenzo, G.C., Finan, T.M.: The divided bacterial genome: Structure, function, and evolution. Microbiology and Molecular Biology Reviews **81**(3) (2017)
12. Jiao, J., Ni, M., Zhang, B., Zhang, Z., Young, J.P.W., Chan, T.-F., Chen, W.X., Lam, H.-M., Tian, C.F.: Coordinated regulation of core and accessory genes in the multipartite genome of sinorhizobium fredii. PLoS genetics **14**(5), 1007428 (2018)
13. von Wintersdorff, C.J., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H., Wolffs, P.F.: Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. Frontiers in microbiology **7**, 173 (2016)
14. Sobecky, P.A., Coombs, J.M.: In: Gogarten, M.B., Gogarten, J.P., Olendzenski, L.C. (eds.) Horizontal gene transfer in metal and radionuclide contaminated soils, pp. 455–472. Springer, Switzerland (2009)
15. Ward, D.M.: A macrobiological perspective on microbial species. Microbe-american society for microbiology **1**(6), 269 (2006)
16. Cohan, F.M.: Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Philosophical Transactions of the Royal Society B: Biological Sciences **361**(1475), 1985 (2006)
17. Shapiro, B.J., Polz, M.F.: Ordering microbial diversity into ecologically and genetically cohesive units. Trends in microbiology **22**(5), 235–247 (2014)
18. Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J., Polz, M.F.: Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science **320**(5879), 1081–1085 (2008)
19. Polz, M.F., Alm, E.J., Hanage, W.P.: Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics **29**(3), 170–175 (2013)
20. Cohan, F.M.: Does recombination constrain neutral divergence among bacterial taxa? Evolution **49**(1), 164–175 (1995)
21. Vulić, M., Dionisio, F., Taddei, F., Radman, M.: Molecular keys to speciation: Dna polymorphism and the control of genetic exchange in enterobacteria. Proceedings of the National Academy of Sciences **94**(18), 9763–9767 (1997)
22. Majewski, J., Cohan, F.M.: Dna sequence similarity requirements for interspecific recombination in bacillus. Genetics **153**(4), 1525–1533 (1999)
23. Cohan, F.M.: Sexual isolation and speciation in bacteria. In: blaaaaaaaa (ed.) Genetics of Mate Choice: From Sexual Selection to Sexual Isolation, pp. 359–370. Springer, Dordrecht (2002)
24. Zawadzki, P., Roberts, M.S., Cohan, F.M.: The log-linear relationship between sexual isolation and sequence divergence in bacillus transformation is robust. Genetics **140**(3), 917–932 (1995)
25. de Visser, A., Rozen, D.E.: Clonal interference and the periodic selection of new beneficial mutations in escherichia coli. Genetics (2006)
26. Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., Alm, E.J.: Population genomics of early events in the ecological differentiation of bacteria. science **336**(6077), 48–51 (2012)
27. Rogel, M.A., Ormeno-Orrillo, E., Romero, E.M.: Symbiovars in rhizobia reflect bacterial adaptation to legumes. Systematic and Applied Microbiology **34**(2), 96–104 (2011)

28. Remigi, P., Zhu, J., Young, J.P.W., Masson-Boivin, C.: Symbiosis within symbiosis: evolving nitrogen-fixing legume symbionts. Trends in microbiology **24**(1), 63–75 (2016)
29. Andrews, M., De Meyer, S., James, E., Stepkowski, T., Hodge, S., Simon, M., Young, J.: Horizontal transfer of symbiosis genes within and between rhizobial genera: occurrence and importance. Genes **9**(7), 321 (2018)
30. Segovia, L., Young, J.P.W., Martínez-Romero, E.: Reclassification of american rhizobium leguminosarum biovar phaseoli type i strains as rhizobium etli sp. nov. International Journal of Systematic and Evolutionary Microbiology **43**(2), 374–377 (1993)
31. Haukka, K., Lindström, K., Young, J.P.W.: Three phylogenetic groups of noda and nifhgenes in sinorhizobium and mesorhizobium isolates from leguminous trees growing in africa and latin america. Applied and Environmental Microbiology **64**(2), 419–426 (1998)
32. Laguerre, G., Nour, S.M., Macheret, V., Sanjuan, J., Drouin, P., Amarger, N.: Classification of rhizobia based on nodc and nifh gene analysis reveals a close phylogenetic relationship among phaseolus vulgaris symbionts. Microbiology **147**(4), 981–993 (2001)
33. Pérez Carrascal, O.M., VanInsberghe, D., Juárez, S., Polz, M.F., Vinuesa, P., González, V.: Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing rhizobium species associated with phaseolus vulgaris. Environmental microbiology **18**(8), 2660–2676 (2016)
34. Sullivan, J.T., Patrick, H.N., Lowther, W.L., Scott, D.B., Ronson, C.W.: Nodulating strains of rhizobium loti arise through chromosomal symbiotic gene transfer in the environment. Proceedings of the National Academy of Sciences **92**(19), 8985–8989 (1995)
35. Nandasena, K.G., O'Hara, G.W., Tiwari, R.P., Howieson, J.G.: Rapid in situ evolution of nodulating strains for biserrula pelecinus l. through lateral transfer of a symbiosis island from the original mesorhizobial inoculant. Applied and environmental microbiology **72**(11), 7365–7367 (2006)
36. Friesen, M.L.: Widespread fitness alignment in the legume-rhizobium symbiosis. New Phytologist **194**(4), 1096–1111 (2012)
37. Harrison, P.W., Lower, R.P., Kim, N.K., Young, J.P.W.: Introducing the bacterial 'chromid': not a chromosome, not a plasmid. Trends in microbiology **18**(4), 141–148 (2010)
38. Kumar, N., Lad, G., Giuntini, E., Kaye, M.E., Udomwong, P., Shamsani, N.J., Young, J.P.W., Bailly, X.: Bacterial genospecies that are not ecologically coherent: population genomics of rhizobium leguminosarum. Open biology **5**(1) (2015)
39. Clark, A., Zheng, Y.: Dynamics of linkage disequilibrium in bacterial genomes undergoing transformation and/or conjugation. Journal of Evolutionary Biology **10**(4), 663–676 (1997)
40. Wetzel, M.E., Olsen, G.J., Chakravartty, V., Farrand, S.K.: The repabc plasmids with quorum-regulated transfer systems in members of the rhizobiales divide into two structurally and separately evolving groups. Genome biology and evolution **7**(12), 3337–3357 (2015)
41. Parker, M.A.: Legumes select symbiosis island sequence variants in bradyrhizobium. Molecular ecology **21**(7), 1769–1778 (2012)
42. Freiberg, C., Fellay, R., Bairoch, A., Broughton, W.J., Rosenthal, A., Perret, X.: Molecular basis of symbiosis between rhizobium and legumes. Nature **387**(6631), 394 (1997)
43. Guillot, G., Rousset, F.: Dismantling the mantel tests. Methods in Ecology and Evolution **4**(4), 336–344 (2013)
44. Harmon, L.J., Glor, R.E.: Poor statistical performance of the mantel test in phylogenetic comparative analyses. Evolution: International Journal of Organic Evolution **64**(7), 2173–2178 (2010)
45. Perez-Mendoza, D., Domínguez-Ferreras, A., Munoz, S., Soto, M.J., Olivares, J., Brom, S., Girard, L., Herrera-Cervera, J.A., Sanjuán, J.: Identification of functional mob regions in rhizobium etli: evidence for self-transmissibility of the symbiotic plasmid pretcfn42d. Journal of bacteriology **186**(17), 5753–5761 (2004)
46. von Bodman, S.B., McCutchan, J., Farrand, S.K.: Characterization of conjugal transfer functions of agrobacterium tumefaciens ti plasmid ptic58. Journal of bacteriology **171**(10), 5281–5289 (1989)
47. Chen, L., Chen, Y., Wood, D.W., Nester, E.W.: A new type iv secretion system promotes conjugal transfer in agrobacterium tumefaciens. Journal of bacteriology **184**(17), 4838–4845 (2002)
48. Masuda, H., Inouye, M.: Toxins of prokaryotic toxin-antitoxin systems with sequence-specific endoribonuclease activity. Toxins **9**(4), 140 (2017)

49. Leplae, R., Geeraerts, D., Hallez, R., Guglielmini, J., Dreze, P., Van Melderen, L.: Diversity of bacterial type ii toxin–antitoxin systems: a comprehensive search and functional analysis of novel families. Nucleic acids research **39**(13), 5513–5525 (2011)

50. Novick, R.P., Christie, G.E., Penadés, J.R.: The phage-related chromosomal islands of gram-positive bacteria. Nature Reviews Microbiology **8**(8), 541 (2010)

51. Wendling, C.C., Goehlich, H., Roth, O.: The structure of temperate phage–bacteria infection networks changes with the phylogenetic distance of the host bacteria. Biology letters **14**(11), 20180320 (2018)

52. Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., Wishart, D.S.: Phast: a fast phage search tool. Nucleic acids research **39**(suppl_2), 347–352 (2011)

53. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D.S.: Phaster: a better, faster version of the phast phage search tool. Nucleic acids research **44**(W1), 16–21 (2016)

54. Isidro, A., Henriques, A.O., Tavares, P.: The portal protein plays essential roles at different steps of the spp1 dna packaging process. Virology **322**(2), 253–263 (2004)

55. Hirsch, P., Van Montagu, M., Johnston, A., Brewin, N., Schell, J.: Physical identification of bacteriocinogenic, nodulation and other plasmids in strains of rhizobium leguminosarm. Microbiology **120**(2), 403–412 (1980)

56. Lemaire, B., Dlodlo, O., Chimphango, S., Stirton, C., Schrire, B., Boatwright, S., Honnay, O., Smets, E., Sprent, J., James, E., *et al.*: Symbiotic diversity, specificity and distribution of rhizobia in native legumes of the core cape subregion (south africa). FEMS microbiology ecology **91**, 2–17 (2015)

57. Provorov, N., Andronov, E., Onishchuk, O.: Forms of natural selection controlling the genomic evolution in nodule bacteria. Russian Journal of Genetics **53**(4), 411–419 (2017)

58. Cervantes, L., Bustos, P., Girard, L., Santamaría, R.I., Dávila, G., Vinuesa, P., Romero, D., Brom, S.: The conjugative plasmid of a bean-nodulating sinorhizobium fredii strain is assembled from sequences of two rhizobium plasmids and the chromosome of a sinorhizobium strain. BMC microbiology **11**(1), 149 (2011)

59. Li, X., Tong, W., Wang, L., Rahman, S.U., Wei, G., Tao, S.: A novel strategy for detecting recent horizontal gene transfer and its application to rhizobium strains. Frontiers in Microbiology **9**, 973 (2018). doi:10.3389/fmicb.2018.00973

60. Sullivan, J.T., Trzebiatowski, J.R., Cruickshank, R.W., Gouzy, J., Brown, S.D., Elliot, R.M., Fleetwood, D.J., McCallum, N.G., Rossbach, U., Stuart, G.S., *et al.*: Comparative sequence analysis of the symbiosis island of mesorhizobium loti strain r7a. Journal of bacteriology **184**(11), 3086–3095 (2002)

61. Ling, J., Wang, H., Wu, P., Li, T., Tang, Y., Naseer, N., Zheng, H., Masson-Boivin, C., Zhong, Z., Zhu, J.: Plant nodulation inducers enhance horizontal gene transfer of azorhizobium caulinodans symbiosis island. Proceedings of the National Academy of Sciences **113**(48), 13875–13880 (2016)

62. Bailly, X., Olivieri, I., Brunel, B., Cleyet-Marel, J.-C., Béna, G.: Horizontal gene transfer and homologous recombination drive the evolution of the nitrogen-fixing symbionts of medicago species. Journal of bacteriology **189**(14), 5223–5236 (2007)

63. Klinger, C.R., Lau, J.A., Heath, K.D.: Ecological genomics of mutualism decline in nitrogen-fixing bacteria. Proc. R. Soc. B **283**(1826), 20152563 (2016)

64. Falush, D., Torpdahl, M., Didelot, X., Conrad, D.F., Wilson, D.J., Achtman, M.: Mismatch induced speciation in salmonella: model and data. Philosophical Transactions of the Royal Society of London B: Biological Sciences **361**(1475), 2045–2053 (2006)

65. Rocha, E.P.: The organization of the bacterial genome. Annual review of genetics **42**, 211–233 (2008)

66. Bohlin, J., Eldholm, V., Pettersson, J.H., Brynildsrud, O., Snipen, L.: The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC genomics **18**(1), 151 (2017)

67. Reeve, W., O'Hara, G., Chain, P., Ardley, J., Bräu, L., Nandesena, K., Tiwari, R., Malfatti, S., Kiss, H., Lapidus, A., *et al.*: Complete genome sequence of rhizobium leguminosarum bv trifolii strain wsm2304, an effective microsymbiont of the south american clover trifolium polymorphum. Standards in genomic sciences **2**(1), 66 (2010)

68. Servín-Garcidueñas, L.E., Rogel, M.A., Ormeño-Orrillo, E., Delgado-Salinas, A., Martínez-Romero, J., Sánchez, F., Martínez-Romero, E.: Genome sequence of rhizobium sp. strain ccge510, a symbiont isolated from nodules of the endangered wild bean phaseolus albescens. Journal of bacteriology **194**(22), 6310–6311 (2012)

69. Silva, C., Vinuesa, P., Eguiarte, L.E., Martínez-Romero, E., Souza, V.: Rhizobium etli and rhizobium gallicum nodulate common bean (phaseolus vulgaris) in a traditionally managed milpa plot in mexico: population genetics and biogeographic implications. Applied and environmental microbiology **69**(2), 884–893 (2003)
70. Jeong, H., Sung, S., Kwon, T., Seo, M., Caetano-Anollés, K., Choi, S.H., Cho, S., Nasir, A., Kim, H.: Hgtree: database of horizontally transferred genes determined by tree reconciliation. Nucleic acids research **44**(D1), 610–619 (2015)
71. Beiko, R.G., Harlow, T.J., Ragan, M.A.: Highways of gene sharing in prokaryotes. Proceedings of the National Academy of Sciences **102**(40), 14332–14337 (2005)
72. Gophna, U., Ron, E.Z., Graur, D.: Bacterial type iii secretion systems are ancient and evolved by multiple horizontal-transfer events. Gene **312**, 151–163 (2003)
73. Lawrence, J.G., Ochman, H.: Reconciling the many faces of lateral gene transfer. Trends in microbiology **10**(1), 1–4 (2002)
74. Azad, R.K., Lawrence, J.G.: Detecting laterally transferred genes: use of entropic clustering methods and genome position. Nucleic acids research **35**(14), 4629–4639 (2007)
75. van Passel, M.W., Bart, A., Thygesen, H.H., Luyf, A.C., van Kampen, A.H., van der Ende, A.: An acquisition account of genomic islands based on genome signature comparisons. BMC genomics **6**(1), 163 (2005)
76. Ravenhall, M., Škunca, N., Lassalle, F., Dessimoz, C.: Inferring horizontal gene transfer. PLoS computational biology **11**(5), 1004095 (2015)
77. Lawrence, J.G., Ochman, H.: Amelioration of bacterial genomes: rates of change and exchange. Journal of molecular evolution **44**(4), 383–397 (1997)
78. Becq, J., Churlaud, C., Deschavanne, P.: A benchmark of parametric methods for horizontal transfers detection. PLoS One **5**(4), 9989 (2010)
79. Alt-Mörbe, J., Stryker, J.L., Fuqua, C., Li, P.-L., Farrand, S.K., Winans, S.C.: The conjugal transfer system of agrobacterium tumefaciens octopine-type ti plasmids is closely related to the transfer system of an incp plasmid and distantly related to ti plasmid vir genes. Journal of bacteriology **178**(14), 4248–4257 (1996)
80. Galibert, F., Finan, T.M., Long, S.R., Pühler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P., *et al.*: The composite genome of the legume symbiont sinorhizobium meliloti. Science **293**(5530), 668–672 (2001)
81. Barnett, M.J., Fisher, R.F., Jones, T., Komp, C., Abola, A.P., Barloy-Hubler, F., Bowser, L., Capela, D., Galibert, F., Gouzy, J., *et al.*: Nucleotide sequence and predicted functions of the entire sinorhizobium meliloti psyma megaplasmid. Proceedings of the National Academy of Sciences **98**(17), 9883–9888 (2001)
82. Schulein, R., Dehio, C.: The virb/vird4 type iv secretion system of bartonella is essential for establishing intraerythrocytic infection. Molecular microbiology **46**(4), 1053–1067 (2002)
83. Zhao, R., Liu, L.X., Zhang, Y.Z., Jiao, J., Cui, W.J., Zhang, B., Wang, X.L., Li, M.L., Chen, Y., Xiong, Z.Q., *et al.*: Adaptive evolution of rhizobial symbiotic compatibility mediated by co-evolved insertion sequences. The ISME journal **12**(1), 101 (2017)
84. Rosini, R., Rinaudo, C.D., Soriani, M., Lauer, P., Mora, M., Maione, D., Taddei, A., Santi, I., Ghezzo, C., Brettoni, C., *et al.*: Identification of novel genomic islands coding for antigenic pilus-like structures in streptococcus agalactiae. Molecular microbiology **61**(1), 126–141 (2006)
85. Merkl, R.: Sigi: score-based identification of genomic islands. BMC bioinformatics **5**(1), 22 (2004)
86. Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L., *et al.*: Dna sequence of both chromosomes of the cholera pathogen vibrio cholerae. Nature **406**(6795), 477 (2000)
87. Engelberg-Kulka, H., Amitai, S., Kolodkin-Gal, I., Hazan, R.: Bacterial programmed cell death and multicellular behavior in bacteria. PLoS genetics **2**(10), 135 (2006)
88. Van Melderen, L., De Bast, M.S.: Bacterial toxin–antitoxin systems: more than selfish entities? PLoS genetics **5**(3), 1000437 (2009)
89. Hayes, C.S., Sauer, R.T.: Toxin-antitoxin pairs in bacteria: killers or stress regulators? Cell **112**(1), 2–4 (2003)
90. Rowe-Magnus, D.A., Guerout, A.-M., Biskri, L., Bouige, P., Mazel, D.: Comparative analysis of superintegrons: engineering extensive genetic diversity in the vibrionaceae. Genome research **13**(3), 428–442 (2003)
91. Davies, E.V., James, C.E., Williams, D., O'Brien, S., Fothergill, J.L., Haldenby, S., Paterson, S., Winstanley,

C., Brockhurst, M.A.: Temperate phages both mediate and drive adaptive evolution in pathogen biofilms. Proceedings of the National Academy of Sciences **113**(29), 8266–8271 (2016)

92. Volkova, V.V., Lu, Z., Besser, T., Gröhn, Y.T.: Modeling infection dynamics of bacteriophages in enteric escherichia coli: estimating the contribution of transduction to antimicrobial gene spread. Applied and environmental microbiology, 00446 (2014)
93. Chen, J., Quiles-Puchalt, N., Chiang, Y.N., Bacigalupe, R., Fillol-Salom, A., Chee, M.S.J., Fitzgerald, J.R., Penadés, J.R.: Genome hypermobility by lateral transduction. Science **362**(6411), 207–212 (2018)
94. Van Cauwenberghe, J., Verstraete, B., Lemaire, B., Lievens, B., Michiels, J., Honnay, O.: Population structure of root nodulating rhizobium leguminosarum in vicia cracca populations at local to regional geographic scales. Systematic and applied microbiology **37**(8), 613–621 (2014)
95. Bailly, X., Olivieri, I., De Mita, S., CLEYET-MAREL, J.-C., Béna, G.: Recombination and selection shape the molecular diversity pattern of nitrogen-fixing sinorhizobium sp. associated to medicago. Molecular Ecology **15**(10), 2719–2734 (2006)
96. Bailly, X., Giuntini, E., Sexton, M.C., Lower, R.P., Harrison, P.W., Kumar, N., Young, J.P.W.: Population genomics of sinorhizobium medicae based on low-coverage sequencing of sympatric isolates. The ISME Journal **5**(11), 1722 (2011)
97. Microbes NG Service. https://microbesng.uk/
98. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., *et al.*: Spades: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology **19**(5), 455–477 (2012)
99. Github Repository of Jigome. https://github.com/jpwyoung/genomics
100. Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G.: Quast: quality assessment tool for genome assemblies. Bioinformatics **29**(8), 1072–1075 (2013)
101. Seemann, T.: Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**(14), 2068–2069 (2014)
102. Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J.: Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC bioinformatics **12**(1), 124 (2011)
103. Github Repository of Syntenizer 3000. https://github.com/kamiboy/Syntenizer3000/
104. Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., Hartmann, R.K., Prohaska, S.J., Stadler, P.F.: Orthology detection combining clustering and synteny for very large datasets. PLoS One **9**(8), 105015 (2014)
105. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., *et al.*: Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Molecular systems biology **7**(1), 539 (2011)
106. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48**(3), 443–453 (1970)
107. Github Repository for Rhizobium Analysis. https://github.com/izabelcavassim/Rhizobium_analysis
108. VanRaden, P.M.: Efficient methods to compute genomic predictions. Journal of dairy science **91**(11), 4414–4423 (2008)
109. Sukumaran, J., Holder, M.T.: Dendropy: a python library for phylogenetic computing. Bioinformatics **26**(12), 1569–1571 (2010)
110. Long, Q., Rabanal, F.A., Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B.J., Korte, A., Nizhynska, V., *et al.*: Massive genomic variation and strong selection in arabidopsis thaliana lines from sweden. Nature genetics **45**(8), 884 (2013)
111. Simonsen, M., Pedersen, C.N.: Rapid computation of distance estimators from nucleotide and amino acid alignments. In: Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 89–93 (2011). ACM

Figures

Figure 1 Genetic divergence across 196 rhizobium strains. Pairwise comparisons of genetic diversity were analyzed at three different levels. **(a)** Proportion of shared single nucleotide polymorphisms (SNPs) in genes that were present in at least 100 strains and that passed filtering criteria (6,529 genes, 441,287 SNPs). Clusters of strains with SNP identity above 96% were recognised as 5 genospecies: gsA (blue), gsB (salmon), gsC (green), gsD (purple), gsE (pink) as indicated in the legend. **(b)** Average nucleotide identity for concatenated sequences of 282 housekeeping genes. **(c)** Number of shared genes. Strains were ordered by clustering of the SNP data. Strain origins are indicated by coloured bars at the left (DKO in red, DK in purple, F in yellow, and UK in green). **(d)** Histogram showing the distribution of shared genes across strains, with a total of 22,115 orthologous genes. **(e)** Principal component analysis (PCA) of the covariance matrix based on the allelic variation of 6,529 genes that were present in at least 100 strains (see Methods). The colours correspond to the genospecies and the shapes to the origin of the sample. PC1 and PC2. **(f)** PC3 and PC4 of the PCA.

Figure 2 Accessory and core genome. (a) Matrix of the presence (dark) and absence (light) of all 22,115 orthologous gene groups. Strains (y-axis) are clustered by similarity as in Fig. 1a, and genes (x-axis) are clustered by similarity in distribution. **(b)** Venn diagram of the shared orthologous genes across the 5 genospecies; the outermost numbers represent the number of genes that are private to the genospecies. **(c)** GC3 content distribution across accessory and core genes; dashed lines represent the median GC3 of each category.

Figure 3 Population genetic characteristics of the genospecies. (a) Nucleotide diversity of core and accessory genes on the chromosome and the chromids (Rh01 and Rh02). **(b)** Tajima's D distribution for each replicon. Both statistics (nucleotide diversity and Tajima's D) were computed within genospecies and only genes present in all genospecies are shown. **(c)** Site frequency spectrum of each of the three largest genospecies. **(d)** Intragenic Linkage Disequilibrium (LD) decay for these genospecies.

Figure 4 Distribution of plasmid types and evidence of Sym-plasmid introgression through conjugation. (a) The distribution of plasmid groups, which were defined based on the genetic similarity of the RepA plasmid partitioning protein. **(b)** Phylogenetic analysis of the *repA* gene of plasmid type Rh08. DKO represents strains sampled from Danish organic fields, DK from Danish conventional trials. A complete set of conjugal transfer genes has the following genes upstream of *repA*: *traI,trbBCDEJKLFGHI,traRMHBFACDG,* with the origin of transfer (*oriT*) between *traA* and *traC*. Partial sets are broken by the end of the scaffold, mostly after *traM*.

Figure 5 Different intensities of LD between compartments and evidence of HGT. (a) Intergenic LD was calculated for each genomic compartment of strain SM3 (578, 468, 249, 228, 133 genes are present in plasmids Rh01 Rh02, Rh03, Rh05 and Rh07 respectively). The mean intergenic r^2 is: Rh01=0.11; Rh02=0.15; Rh03=0.11; Rh05=0.14; Rh07=0.15. The colors reflect the pairwise correlation between genes, red patches reveal linkage blocks. **(b)** Intergenic LD across genes of the sym plasmid. **(c)** Strong linkage blocks comprising the symbiosis genes (sorted by physical position).

Figure 6 Evidence of horizontal gene transfer between genospecies. (a) Species phylogeny based on a concatenation of 282 core genes using the neighbor-joining method.Bootstrap values are shown only for the branches separating the genospecies. **(b)-(d)** Examples of symbiosis gene phylogenies, with insets showing clades in which identical alleles are shared across genospecies.

Figure 7 Incongruent genes across compartments. (a) Distribution of discordance scores based on genes present in at least 2 genospecies (13,843).**(b)** Distribution of discordance score in genes present in the strain SM3 (5,920 orthologous genes). Only genes that had at least 18 segregating sites and nucleotide diversity $<$ 0.25 were plotted.

Figure 8 Functionality of chromosomal islands. (a) Gene organization of the *avhB*/tra type IV secretion system from SM3. **(b)** Distribution of discordance scores for cluster 1. Coloured bars above the chart represent the classification of gene groups found in the area. **(c)** Illustration of synteny between gene groups in cluster 1 for strains lacking an insert (SM4, SM100), with the *avhB*/Tra conjugative system (SM3, SM121B), with a DNA rearrangement gene cluster (SM170C, SM153D), and one strain with both inserts (SM113). Dot plots above the gene group lines represent the discordance score for each gene in the gene group. **(d)** Distribution of discordance scores for cluster 2. Bars above the chart represent the classification of gene groups found in the area.

955 **Tables**

Table 1 Contrast of average population genetics parameters. Symbiosis gene values in comparison to the average of core genes and accessory genes placed in four different genomic compartments (chromosome, Rh01, Rh02, Rh03).

Gene type	Replicon	GC	Gene length	Segregating sites	Nucleotide diversity	Pairwise differences	Tajima's D
Symbiosis genes	Sym-plasmid	0.547	951.231	106.769	0.036	33.578	2.544
Accessory	Rh01	0.577	798.334	72.807	0.025	19.813	0.08
Accessory	Rh02	0.566	756.836	78.883	0.036	26.649	-0.006
Accessory	Rh03	0.565	899.796	79.281	0.035	26.311	-0.074
Accessory	Chrm	0.567	733.634	72.686	0.04	27.359	0.263
Core	Rh01	0.603	1076.008	200.658	0.041	42.592	0.758
Core	Rh02	0.611	1030.538	210.462	0.039	39.76	0.309
Core	Rh03	0.604	969.504	188.023	0.038	36.431	0.424
Core	Chrm	0.607	941.889	163.674	0.038	35.574	0.818
Core	All genes	0.607	961	171	0.039	36.5	0.765
Acessory	All genes	0.568	755	73.9	0.037	26.2	0.181

Additional Files

Additional file 1 — Supplementary figures

Figure S1-2. Map of soil sampling locations; Figure S3. Pacbio assembly stats; Figure S4. Spades and Jigome assembly; Figure S5. Overall assembly stats; Figure S6. Phylogenetic tree based on rpoB; Figure S7. Pan genome analysis; Figure S8. Population genetics stats; Figure S9. Structural rearrangements between genospecies; Figure S10. repA phylogeny of plasmid Rh07; Figure S11. Phylogenies of *tra* genes of plasmid Rh08; Figure S12-13. Population structure effects on LD estimates; Figure S14. Species tree; Figure S15. Discordance score scheme; Figure S16. Chromosomal introgression islands; Figure S17. Introgression mediated by phage; Figure S18. Discordance score distribution across genomic compartments.

Additional file 2 — Excel spreadsheet with multiple data

This file is a multi-page table composed of the following information:

- Table S1 - Metadata: information on field trials for each isolate.
- Table S2 - Genome statistics: information on genome assemblies.
- Table S3 - Genes statistics: information on genes and plasmid types for each isolate.
- Table S4 - Conserved genes: list of conserved genes used for species tree construction.
- Table S5 - Gene counts; GC content and Population genetics for each compartment.
- Table S6 - Population genetic parameters: of every orthologous gene.
- Table S7 - Symbiosis genes parameters: population genetic parameters of symbiosis genes in contrast to *recA* and *rpoB*.
- Table S8 - Chromosomal islands: features and gene ordering.
- Table S9 - Inserts description: configuration of avhB in different strains.
- Table S10 - Phage diversity: phage ID's, position and sequence for every isolate.
- Table S11 - Accession numbers of the 196 genomes.

Availability of data and materials

The data that support the findings of this study are available in the INSDC databases under Study/BioProject ID PRJNA510726. Accessions numbers are from SAMN10617942 to SAMN10618137 consecutively and are also provided in the Additional file 2 - Table S11.

Figure 1

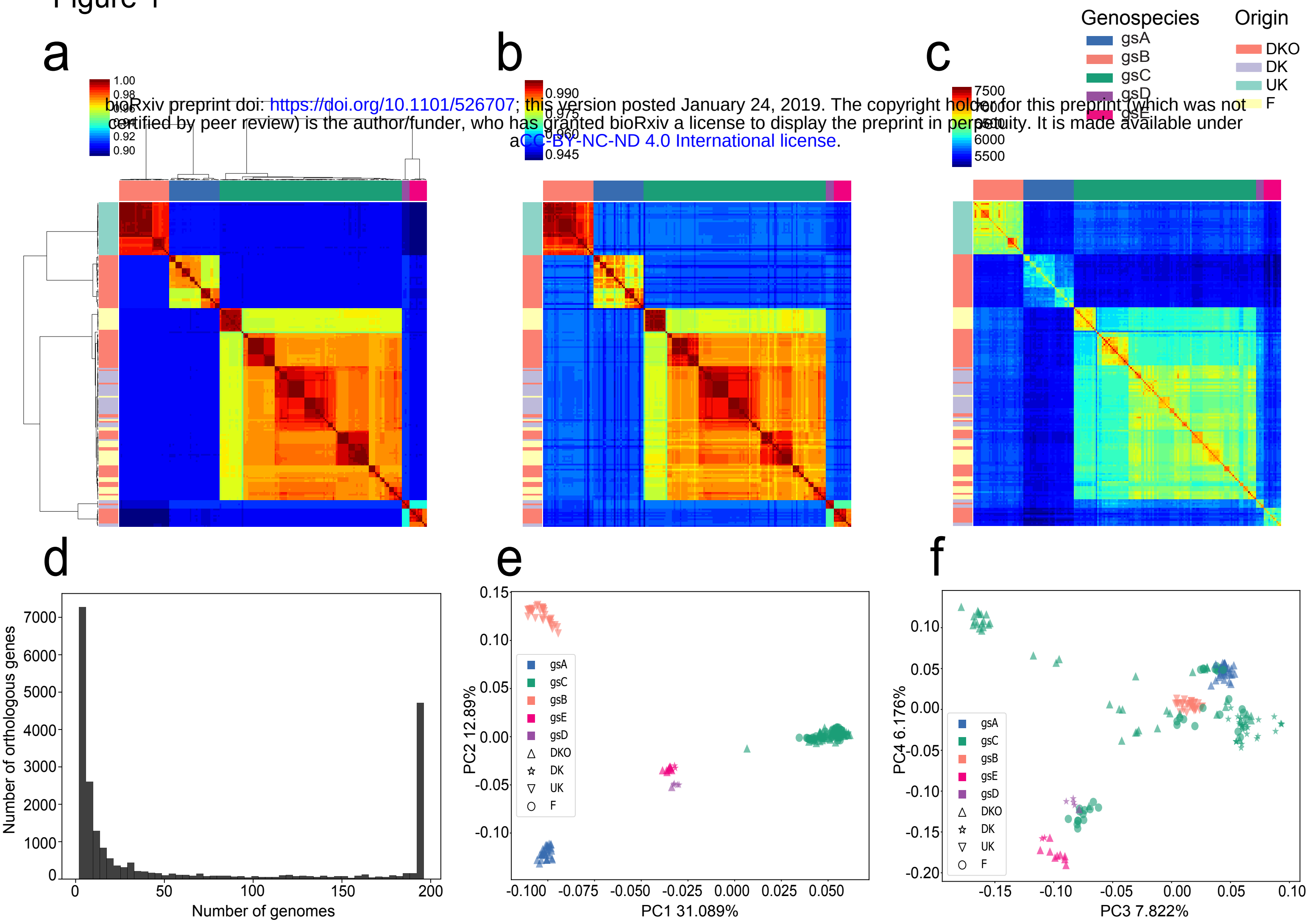

a
1.00
0.98
0.96
0.94
0.92
0.90
b
0.990
0.975
0.960
0.945
c
7500
7000
6500
6000
5500
Genospecies
gsA
gsB
gsC
gsD
gsE
Origin
DKO
DK
UK
F
d
Number of orthologous genes
7000
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
Number of genomes
e
0.15
0.10
0.05
0.00
-0.05
-0.10
PC2 12.89%
-0.100
-0.075
-0.050
-0.025
0.000
0.025
0.050
PC1 31.089%
gsA
gsC
gsB
gsE
gsD
DKO
DK
UK
F
f
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
PC4 6.176%
-0.15
-0.10
-0.05
0.00
0.05
0.10
PC3 7.822%
gsA
gsC
gsB
gsE
gsD
DKO
DK
UK
F

Figure 2

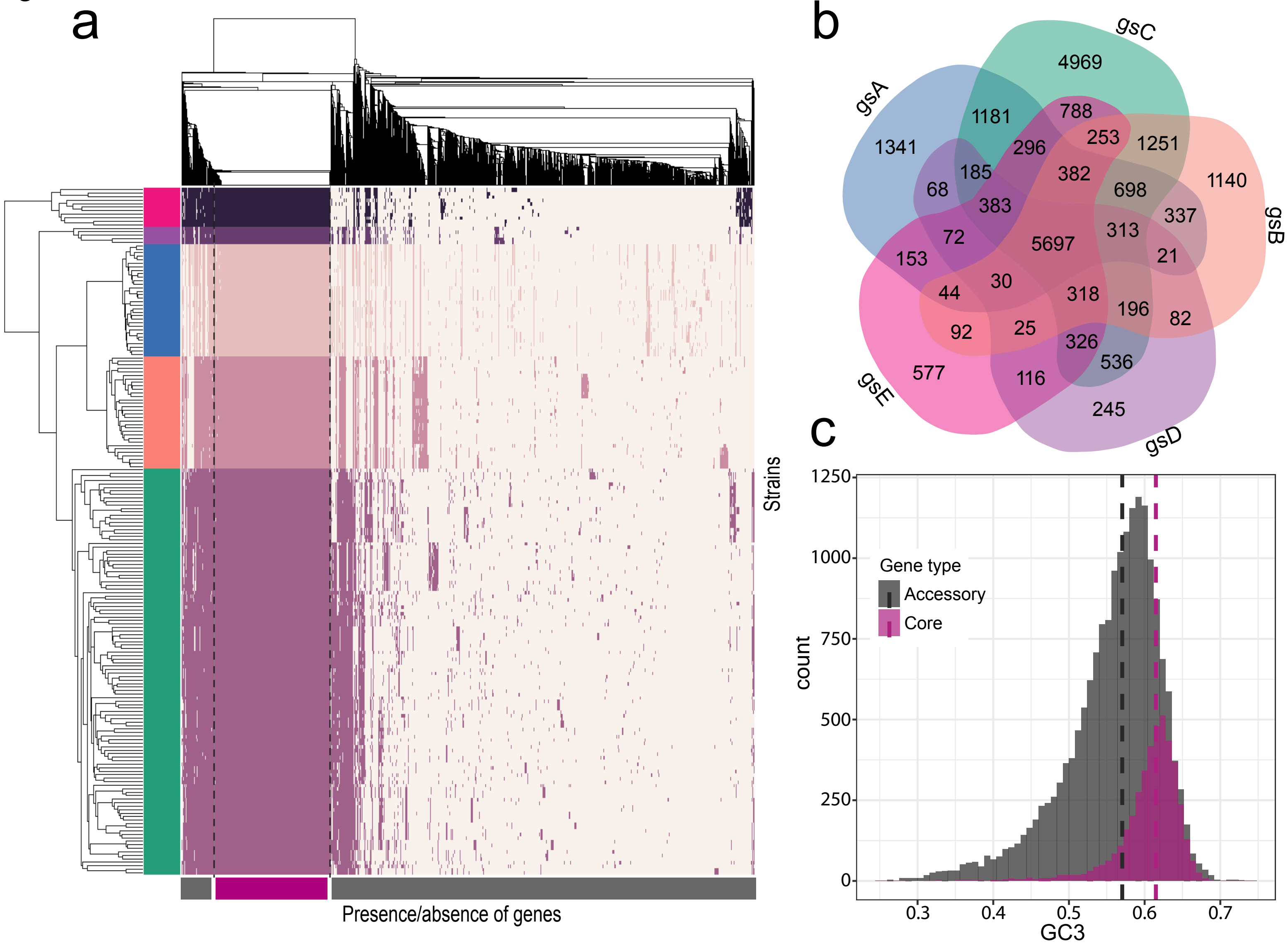

a
Strains
Presence/absence of genes
b
gsA
gsC
gsB
gsD
gsE
4969
1181
788
1341
296
253
1251
185
382
68
698
1140
383
337
313
72
5697
21
153
30
44
318
196
82
92
25
326
536
577
116
245
c
Gene type
Accessory
Core
count
GC3
1250
1000
750
500
250
0
0.3
0.4
0.5
0.6
0.7

Figure 3

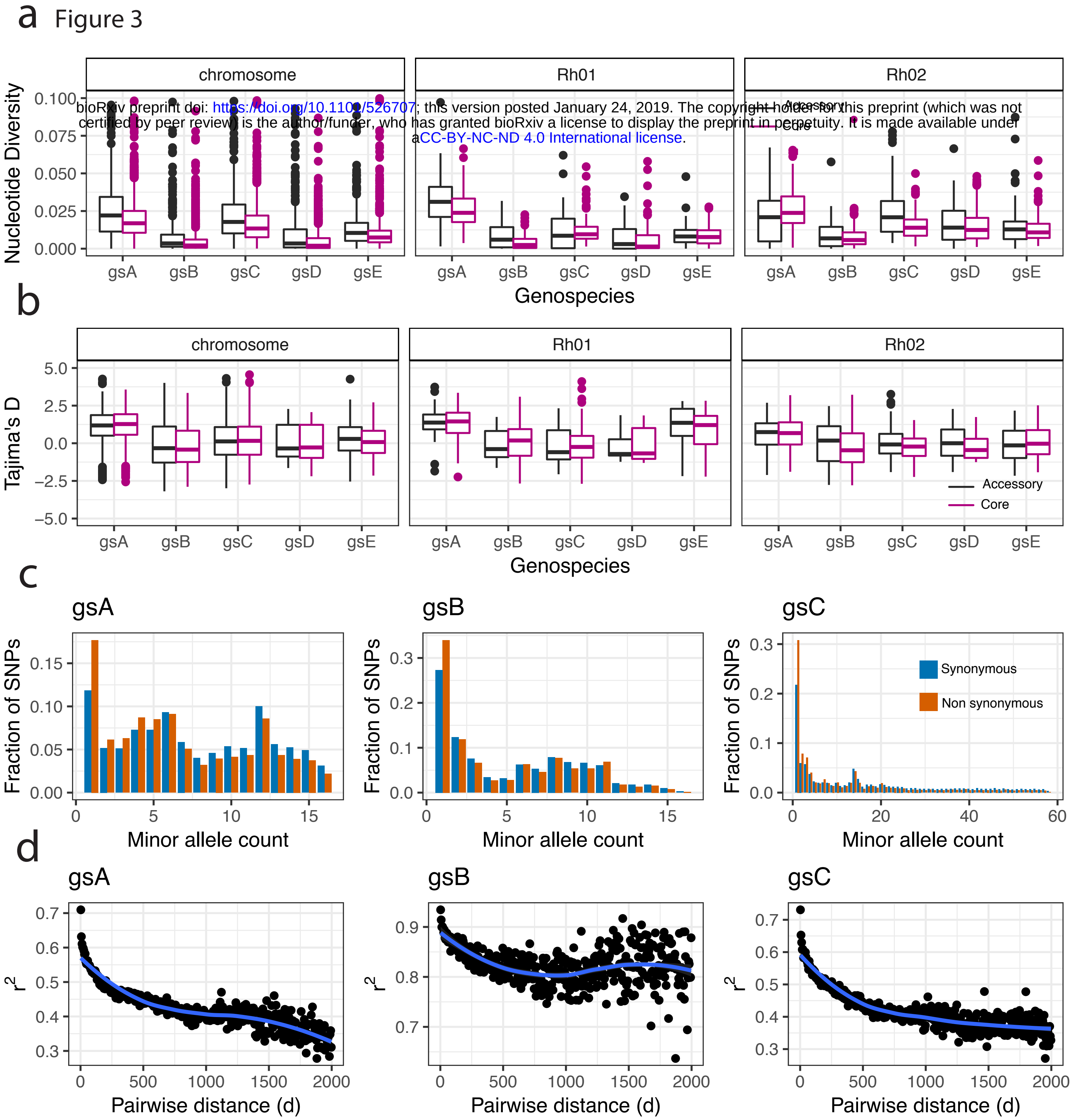

a
chromosome
Rh01
Rh02
Nucleotide Diversity
0.100
0.075
0.050
0.025
0.000
gsA
gsB
gsC
gsD
gsE
Accessory
Core
Genospecies
b
chromosome
Rh01
Rh02
Tajima's D
5.0
2.5
0.0
−2.5
−5.0
gsA
gsB
gsC
gsD
gsE
Accessory
Core
Genospecies
c
gsA
gsB
gsC
Fraction of SNPs
0.15
0.10
0.05
0.00
0.3
0.2
0.1
0.0
0
5
10
15
20
40
60
Synonymous
Non synonymous
Minor allele count
d
gsA
gsB
gsC
r^2
0.7
0.6
0.5
0.4
0.3
0.9
0.8
0
500
1000
1500
2000
Pairwise distance (d)

Figure 4

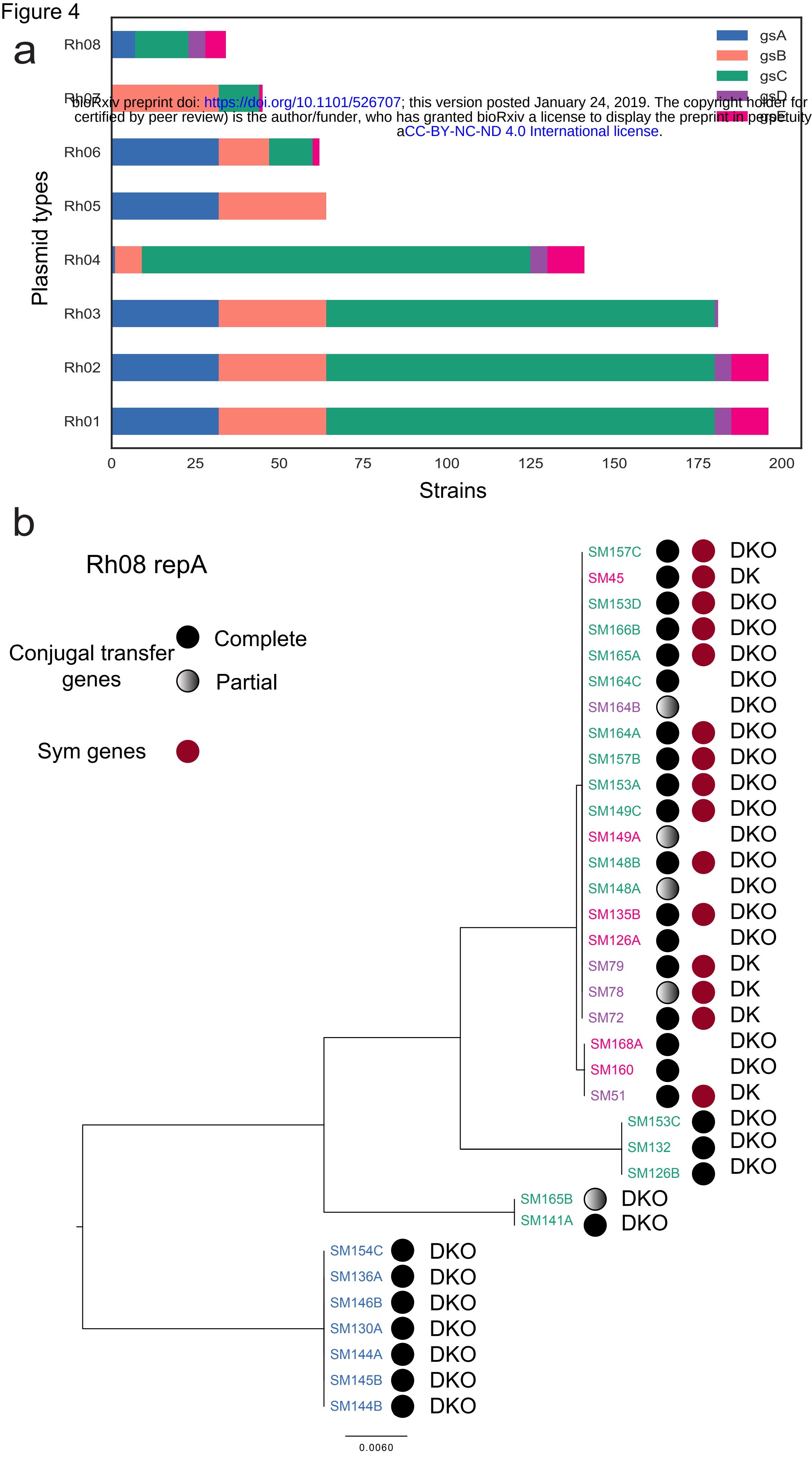

a
Plasmid types
Rh08
Rh07
Rh06
Rh05
Rh04
Rh03
Rh02
Rh01
0
25
50
75
100
125
150
175
200
Strains
gsA
gsB
gsC
gsD
gsE
b
Rh08 repA
Conjugal transfer genes
Complete
Partial
Sym genes
SM157C DKO
SM45 DK
SM153D DKO
SM166B DKO
SM165A DKO
SM164C DKO
SM164B DKO
SM164A DKO
SM157B DKO
SM153A DKO
SM149C DKO
SM149A DKO
SM148B DKO
SM148A DKO
SM135B DKO
SM126A DKO
SM79 DK
SM78 DK
SM72 DK
SM168A DKO
SM160 DKO
SM51 DK
SM153C DKO
SM132 DKO
SM126B DKO
SM165B DKO
SM141A DKO
SM154C DKO
SM136A DKO
SM146B DKO
SM130A DKO
SM144A DKO
SM145B DKO
SM144B DKO
0.0060

Figure 5

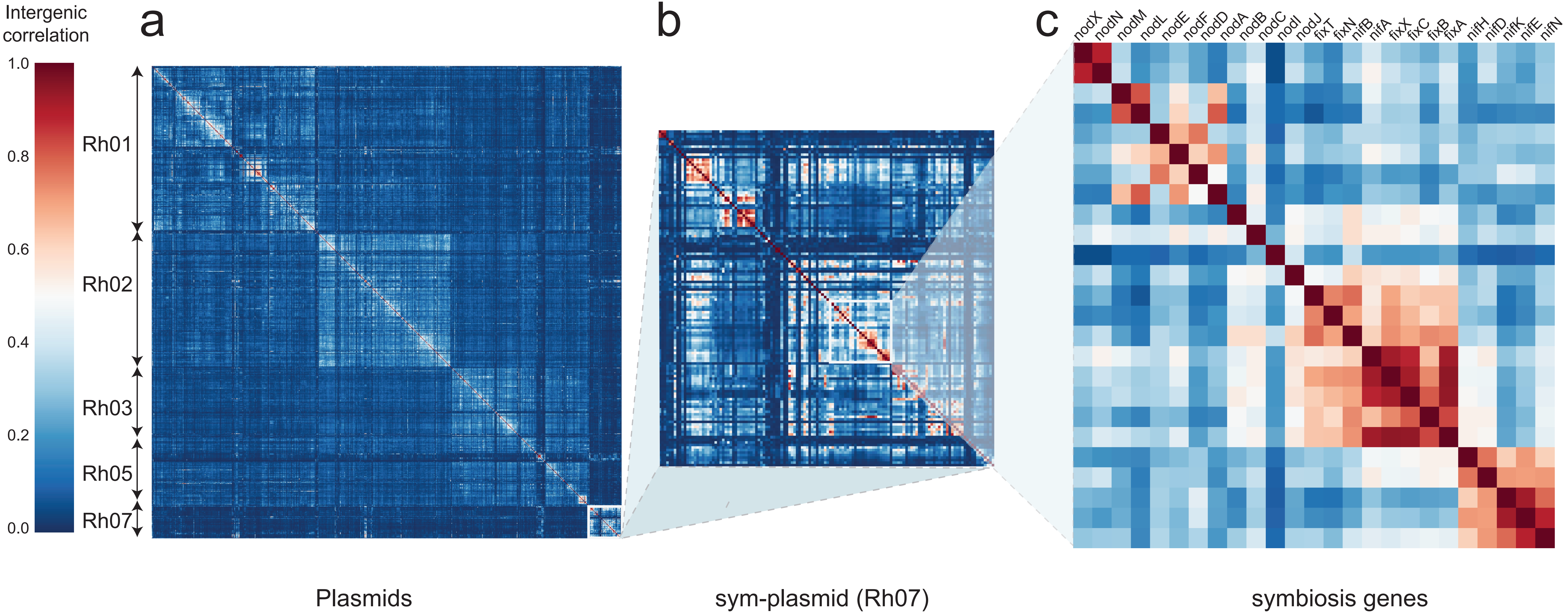

Intergenic correlation
1.0
0.8
0.6
0.4
0.2
0.0
a
Rh01
Rh02
Rh03
Rh05
Rh07
b
c
nodX
nodN
nodM
nodL
nodE
nodF
nodD
nodA
nodB
nodC
nodI
nodJ
fixT
fixN
nifB
nifA
fixX
fixC
fixB
fixA
nifH
nifD
nifK
nifE
nifN
Plasmids
sym-plasmid (Rh07)
symbiosis genes

Figure 6

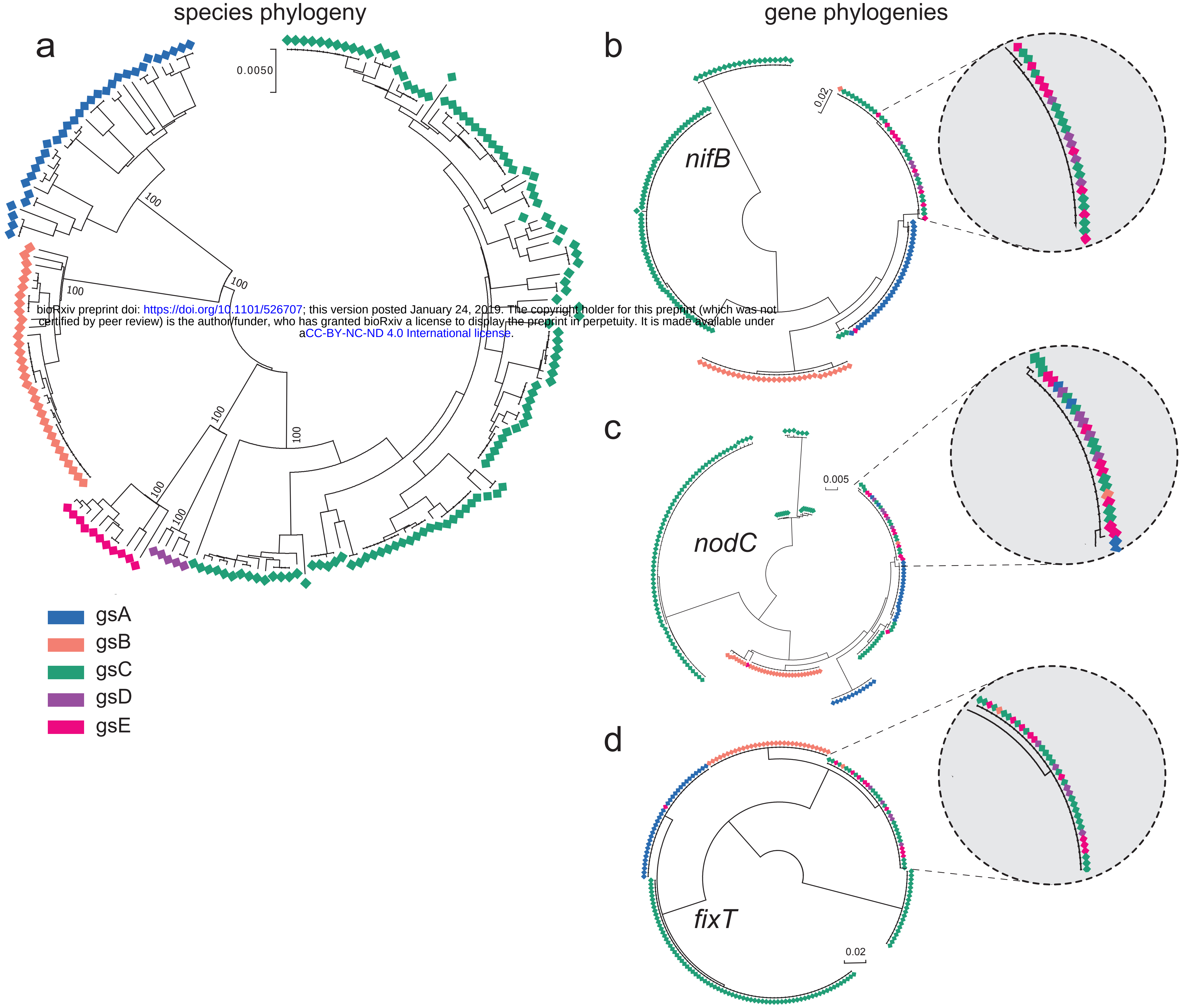

species phylogeny
gene phylogenies
a
0.0050
100
100
100
100
100
100
100
b
nifB
0.02
c
nodC
0.005
d
fixT
0.02
🟦 gsA
🟧 gsB
🟩 gsC
🟪 gsD
🟥 gsE

Figure 7

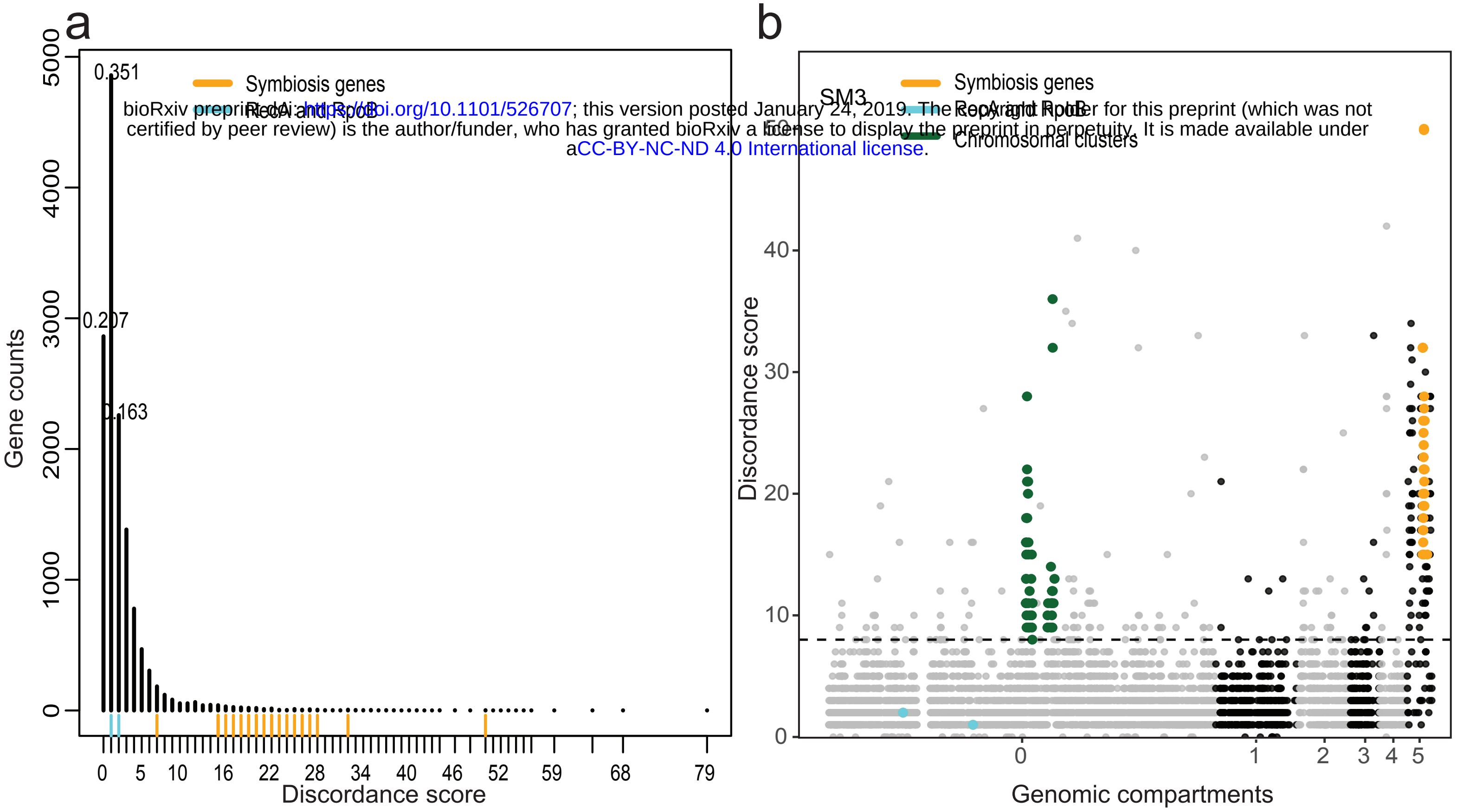
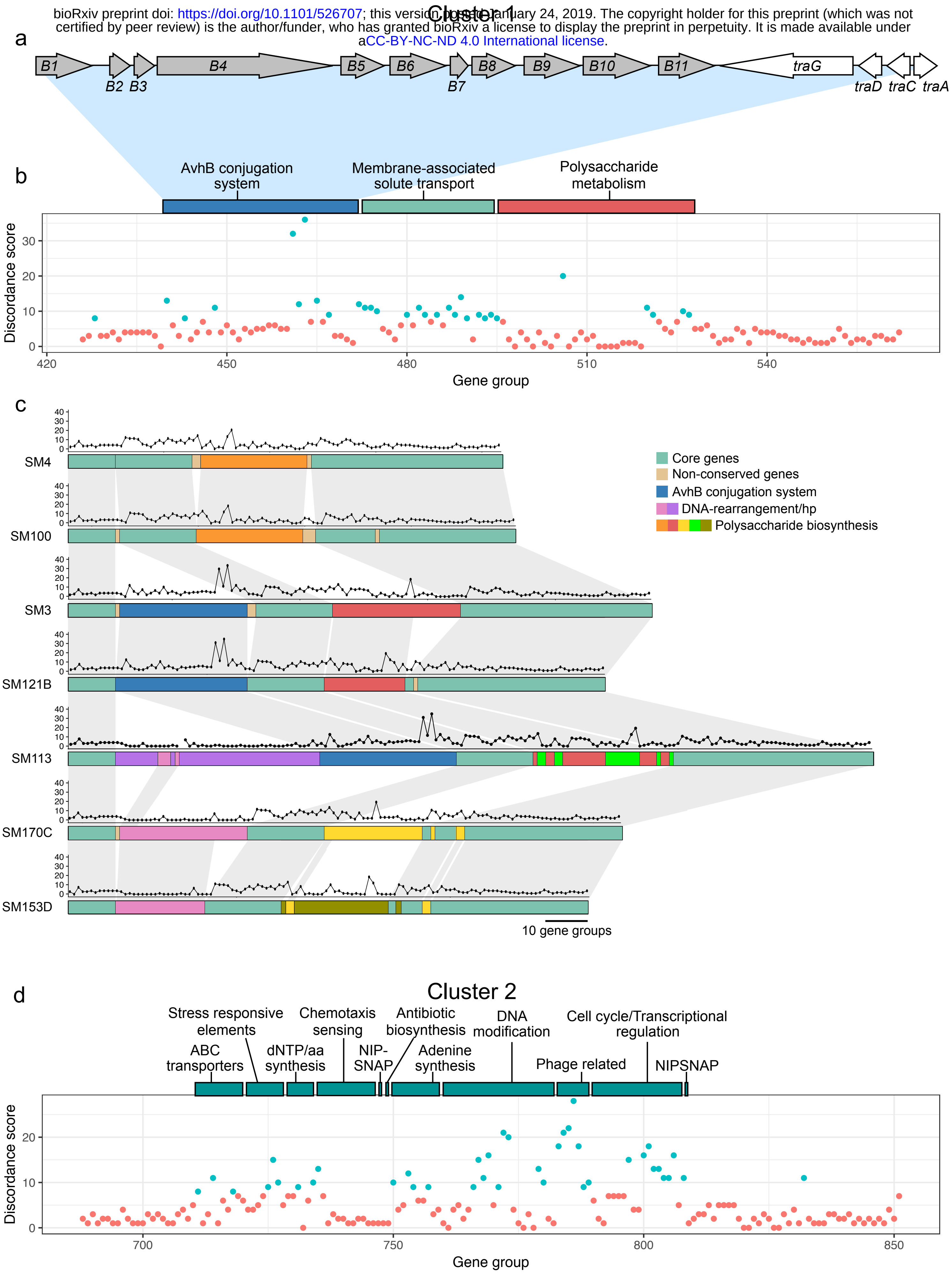

a
Symbiosis genes
RecA and RpoB
0.351
0.207
0.163
Gene counts
0
1000
2000
3000
4000
5000
Discordance score
0
5
10
16
22
28
34
40
46
52
59
68
79
b
SM3
Symbiosis genes
RecA and RpoB
Chromosomal clusters
Discordance score
0
10
20
30
40
50
Genomic compartments
0
1
2
3
4
5

Figure 8


Cluster 1
a
B1
B2
B3
B4
B5
B6
B7
B8
B9
B10
B11
traG
traD
traC
traA
b
AvhB conjugation system
Membrane-associated solute transport
Polysaccharide metabolism
Discordance score
0
10
20
30
420
450
480
510
540
Gene group
c
SM4
SM100
SM3
SM121B
SM113
SM170C
SM153D
0
10
20
30
40
Core genes
Non-conserved genes
AvhB conjugation system
DNA-rearrangement/hp
Polysaccharide biosynthesis
10 gene groups
d
Cluster 2
Stress responsive elements
ABC transporters
dNTP/aa synthesis
Chemotaxis sensing
NIP-SNAP
Antibiotic biosynthesis
Adenine synthesis
DNA modification
Phage related
Cell cycle/Transcriptional regulation
NIPSNAP
Discordance score
0
10
20
700
750
800
850
Gene group