Supplementary figures of the paper: Symbiosis genes show a unique pattern of introgression and selection within a *Rhizobium leguminosarum* species complex

Cavassim et al.

Contents

List of Figures

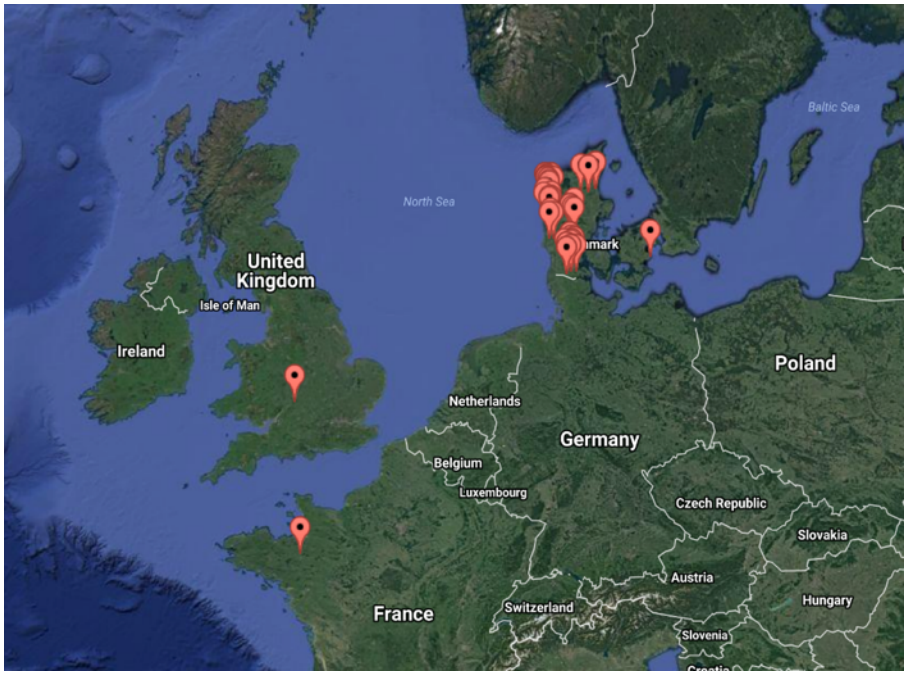

Figure S1: White clover roots were collected from three different DLF trials sites: United Kingdom (UK), Denmark (DK) and France (F).

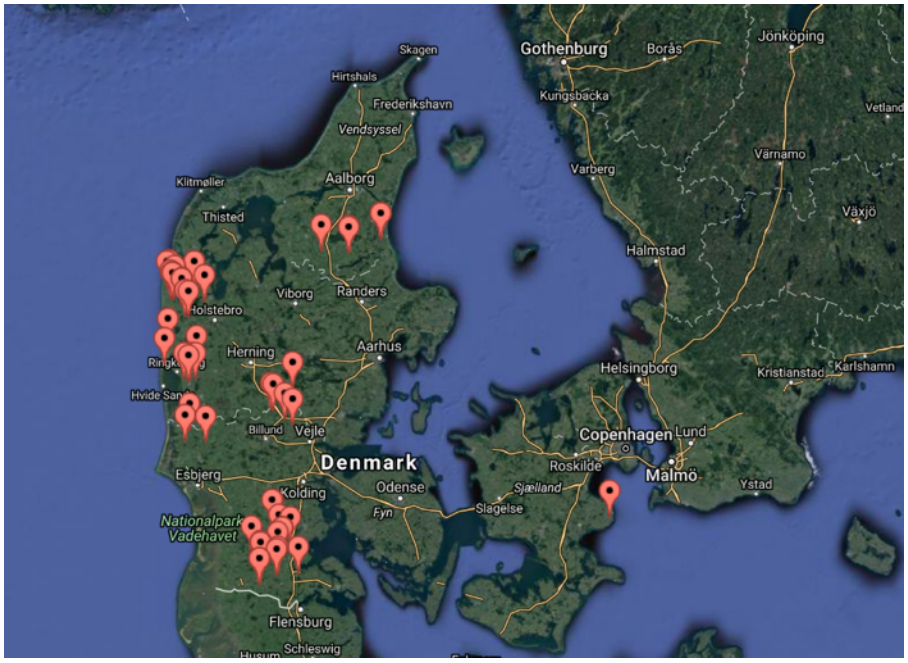

Figure S2: Soil samples were also collected from 50 Danish organic fields (DKO). Geographic information system (SIS) data is attached in supplementary table 1.

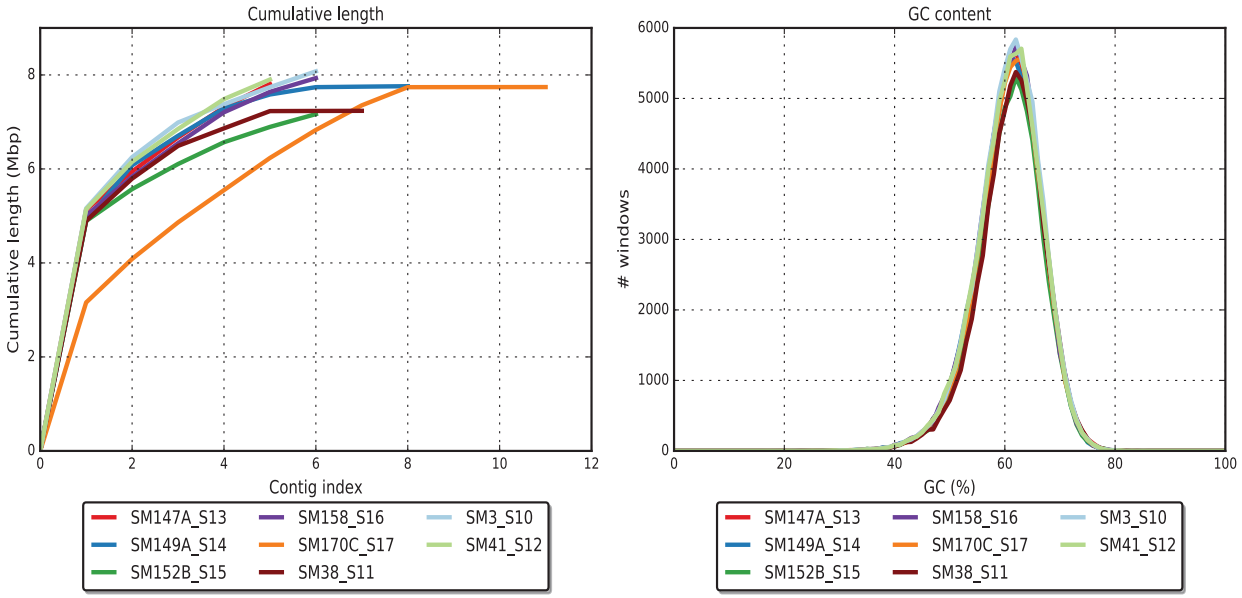

Figure S3: Number of contigs and GC content in each pacbio assembly. These strains were used in order to improve the illumina assemblies. Strain SM170C was excluded from the re-assembly analysis.

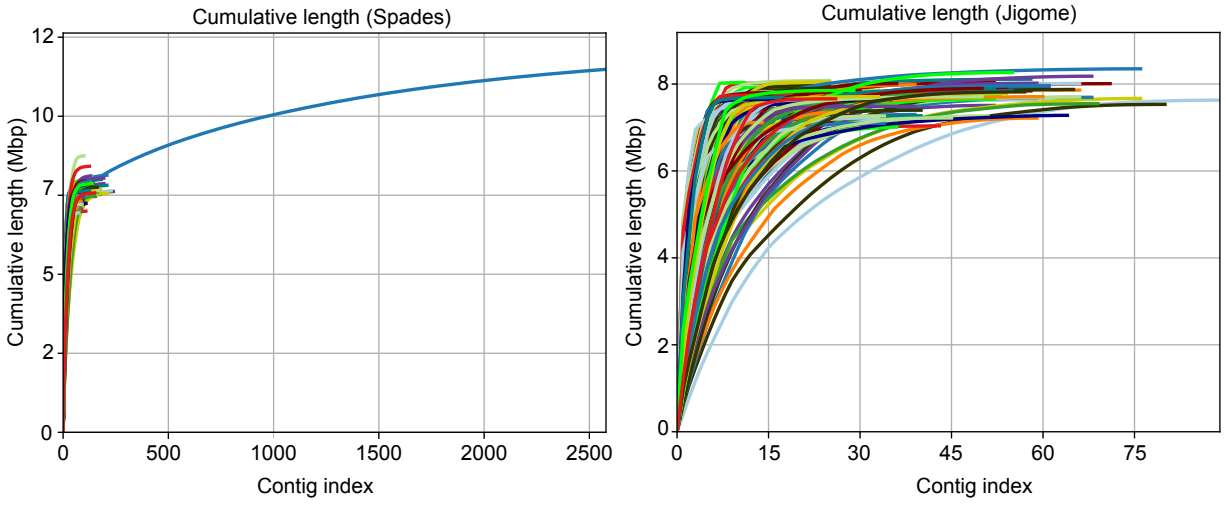

Figure S4: Number of contigs per strain using Spades and later Jigome. A fixed threshold for a minimum contig length of 200 bp was used.

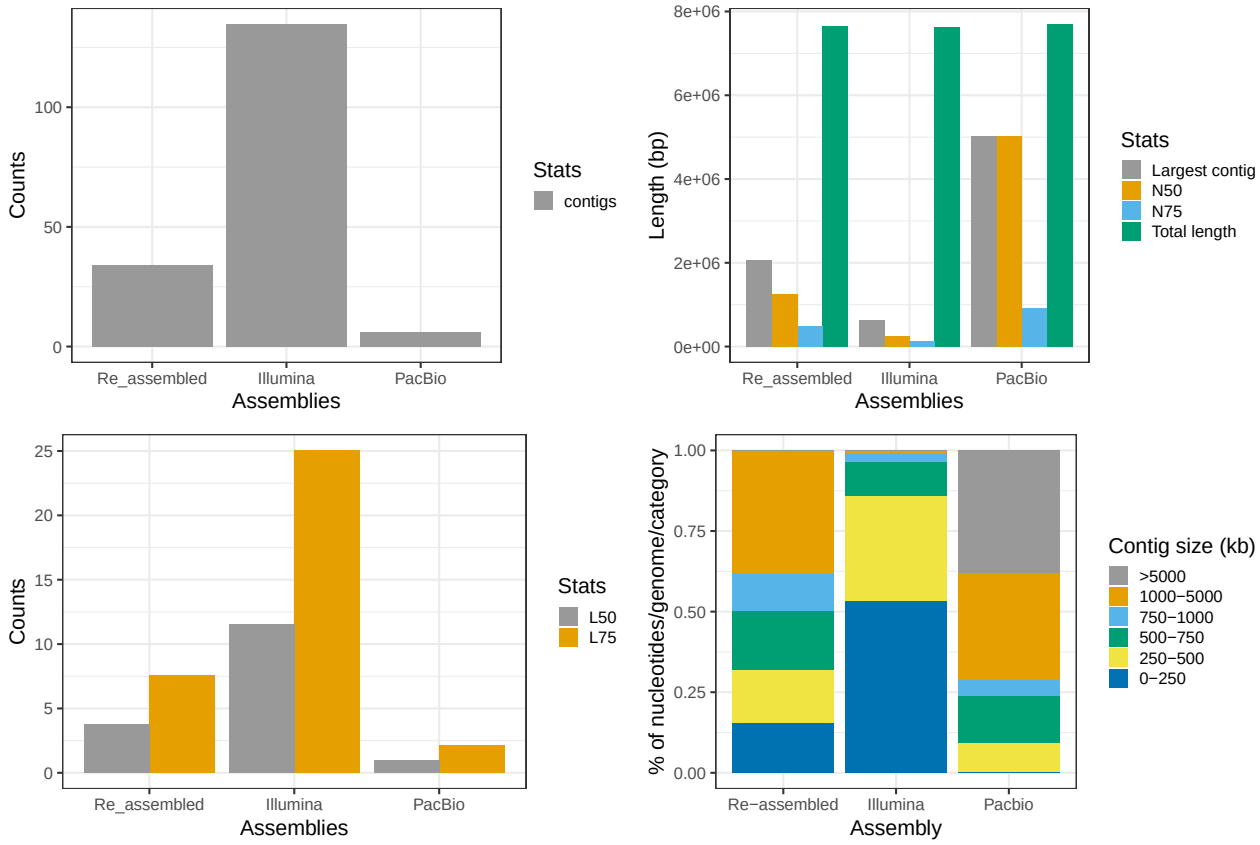

Figure S5: Different statistics across the 3 assemblies: Illumina (Spades assembly), Pacbio (HGAP.3 assembly) and Re-assembled (Illumina re-assembled with Jigome). Re-assembled and Pacbio were used in these analysis.

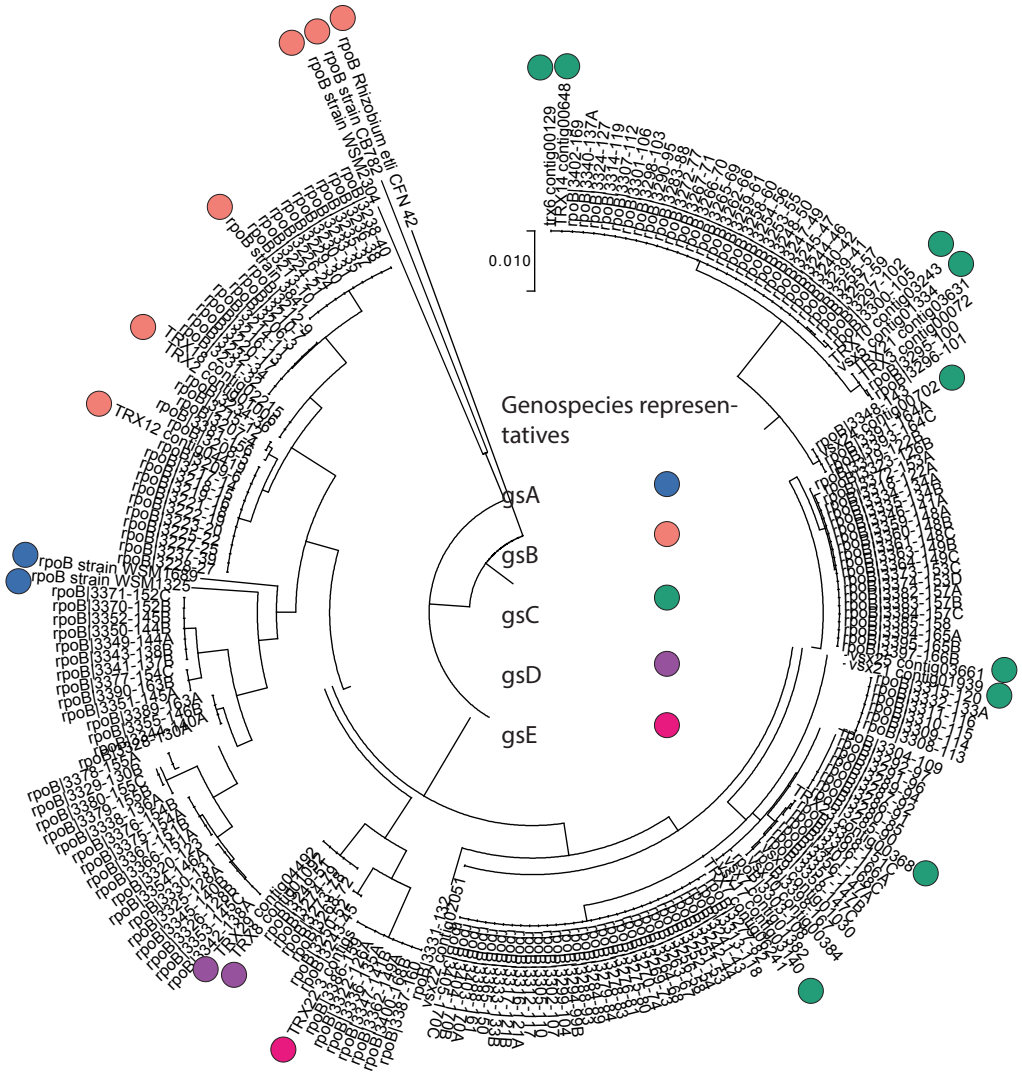

Figure S6: (a) RpoB phylogenetic tree and *rpoB* sequences of representatives of each genospecies (circles). These sequences were previously classified by Kumar et al., 2015.

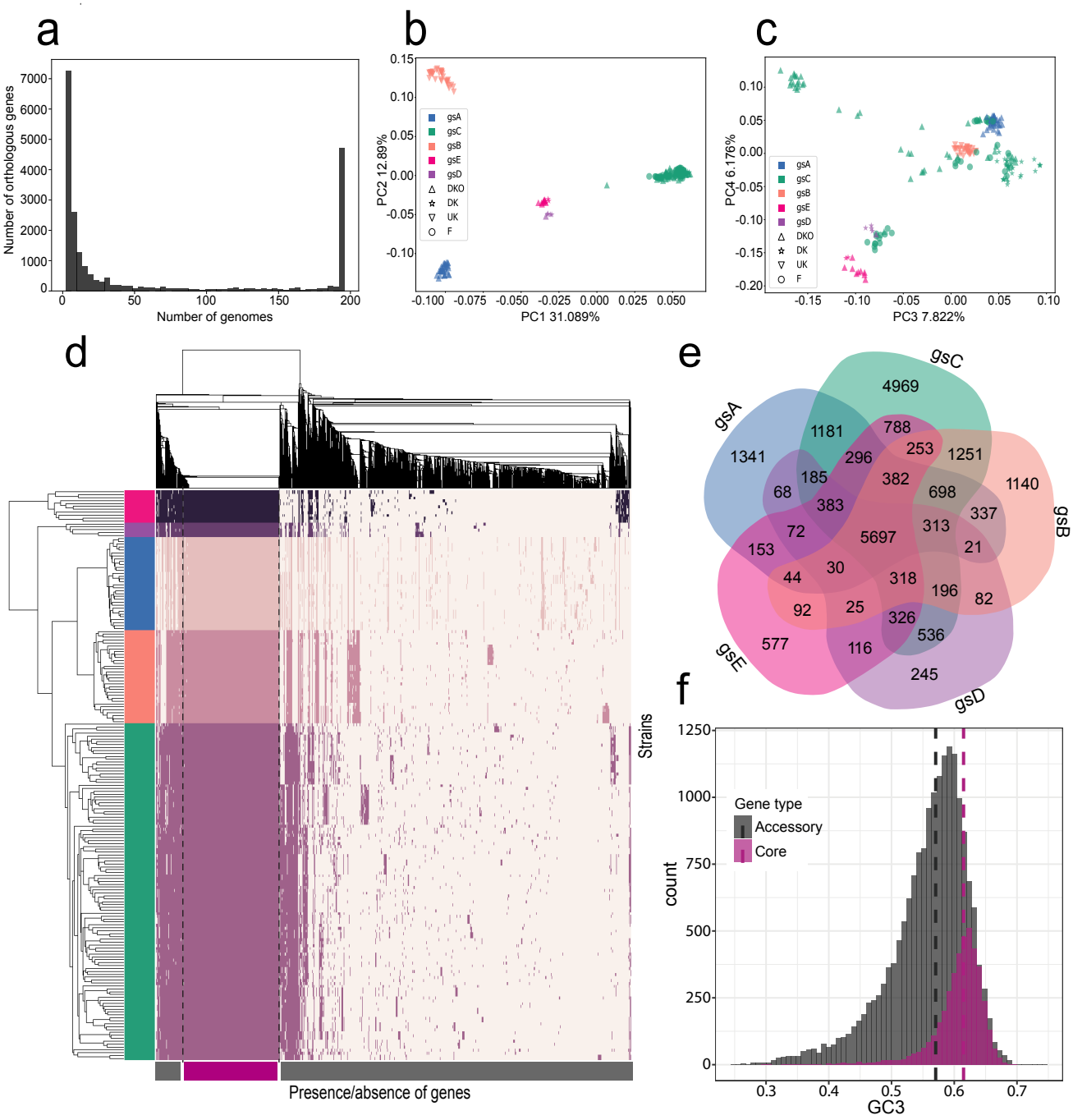

Figure S7: (a) Histogram showing the distribution of shared genes across strains, with a total of 22,115 orthologous genes. (b) Principal component analysis (PCA) of the covariance matrix based on the allelic variation of 6,529 genes that were present in at least 100 strains (see Methods). The colours correspond to the genospecies and the shapes to the origin of the sample. PC1 and PC2. (c) PC3 and PC4 of the PCA. (d) Matrix of the presence (dark) and absence (light) of all 22,115 orthologous gene groups. Strains are clustered by similarity (y-axis), and genes are clustered by their patterns of presence and absence (y-axis). (e) 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. (f) GC3 content distribution across accessory and core genes; dashed lines represent the median GC3 of each category.

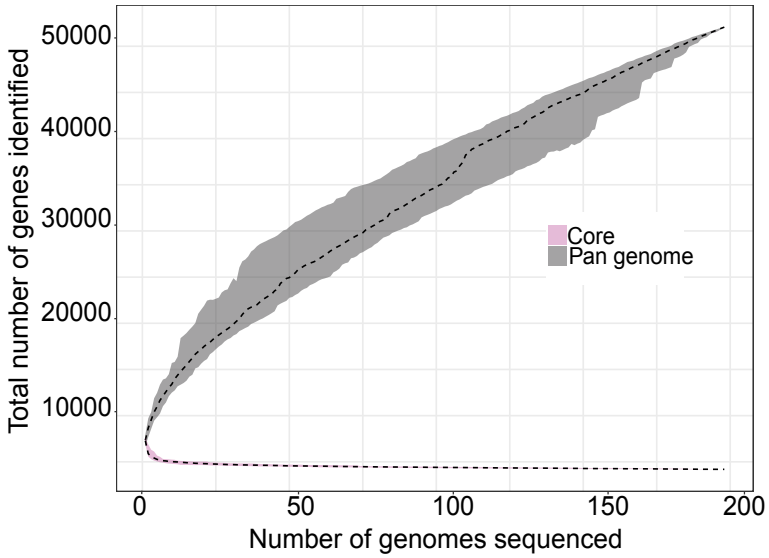

Figure S8: Pan-genome profiles of the Rhizobium species complex. Pan- and core-genome size prediction with all combinations of studied strains (196). It demonstrates that this species has a 'open' pan genome: the number of genes included in the pan genome increases with the number of additional sequenced strains.

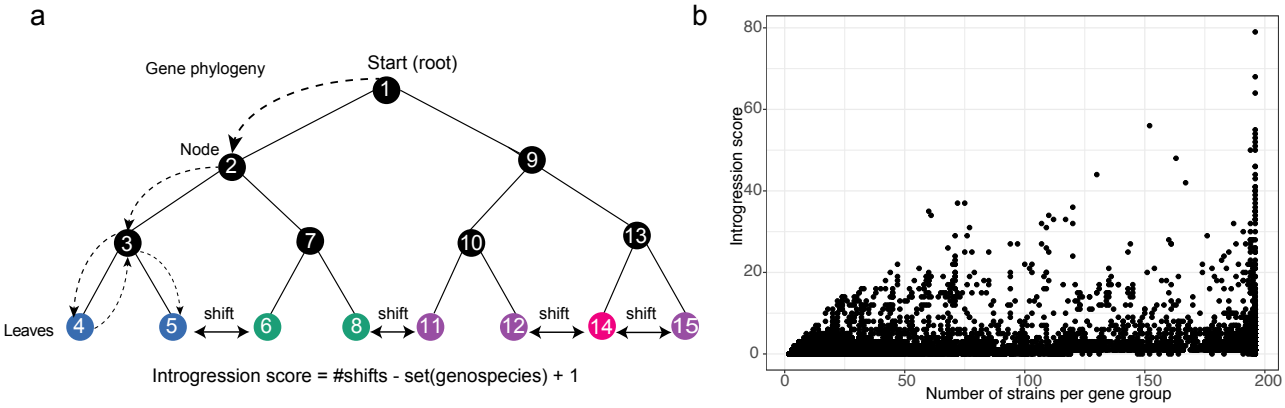

Figure S9: Illustration of the approach for detecting gene introgression (a), and its dependency on the number of members in each orthologous gene (b).

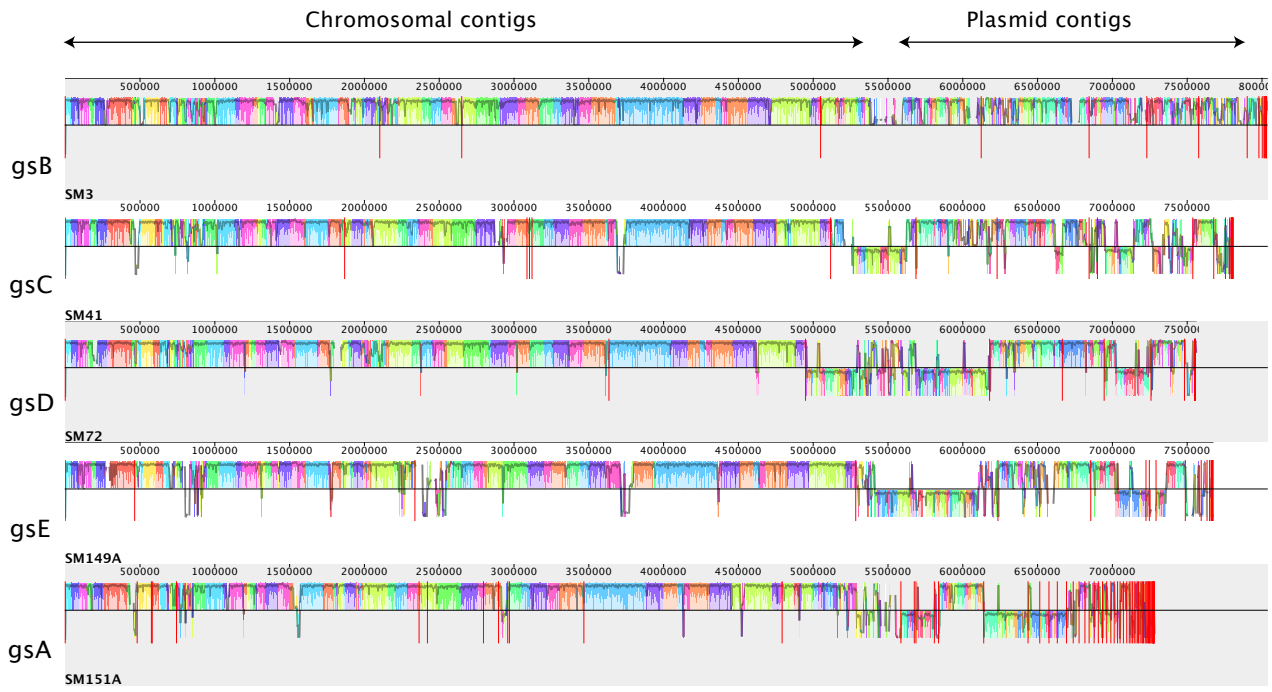

Figure S10: Structural rearrangements and gene interactions of *Rhizobium leguminosarum* bv. *trifolii*. High chromosomal collinearity and distribution of plasmid types. Multiple alignment across one strain from each genospecies, plasmids and chromosomal contigs are distinguished. The coloured blocks correspond to local collinear blocks that are detected by Mauve alignment and are internally free from genome rearrangements.

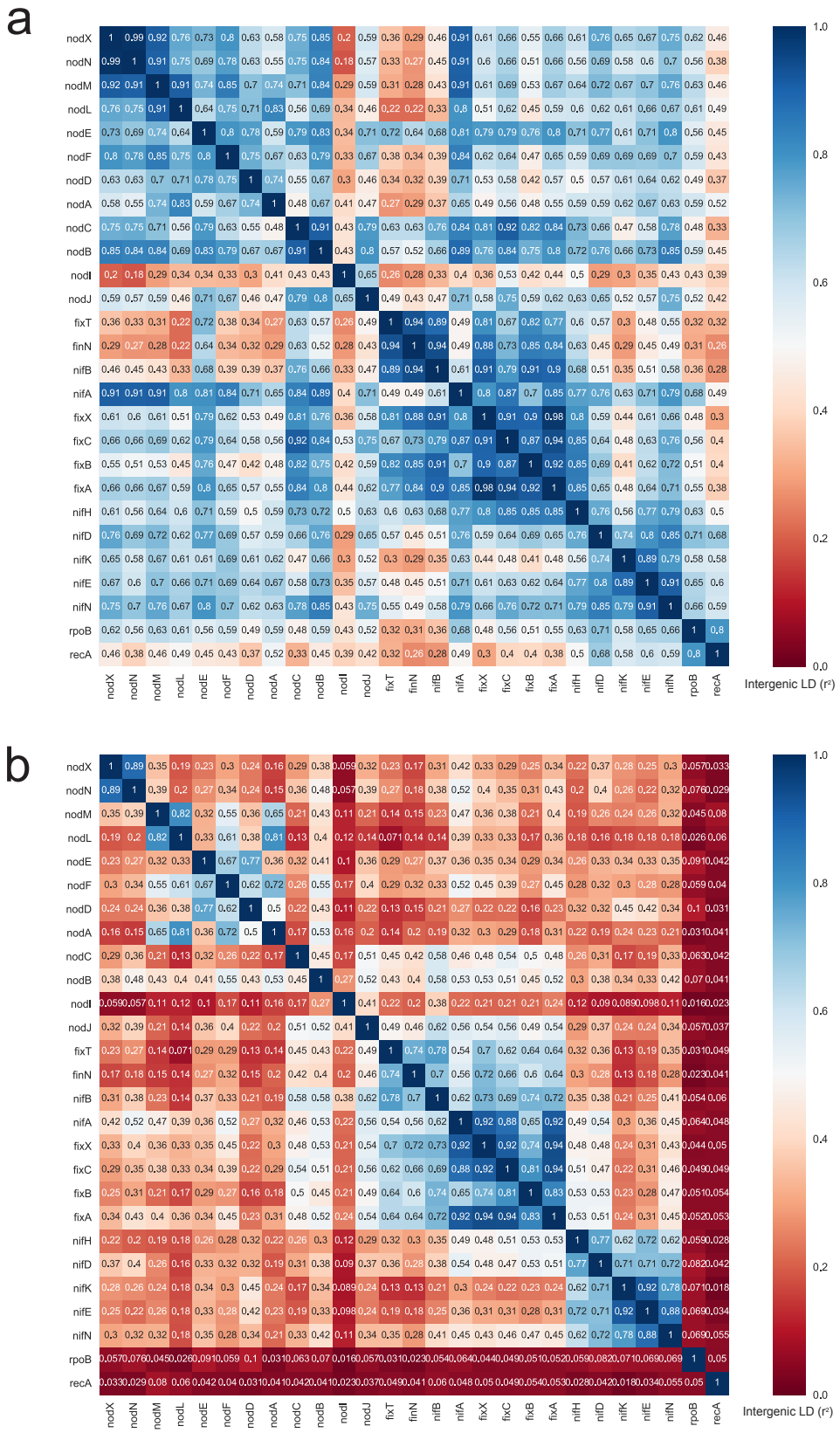

Figure S11: Intergenic linkage disequilibrium before (a) and after (b) population structure correction. The top 24 genes displayed in the matrix are plasmid-borne symbiosis genes, the two last genes (*recA* and *rpoB*), are highly conserved chromosomal genes: part of the DNA recombination and repair system; and part of beta subunit in RNA polymerase, respectively.

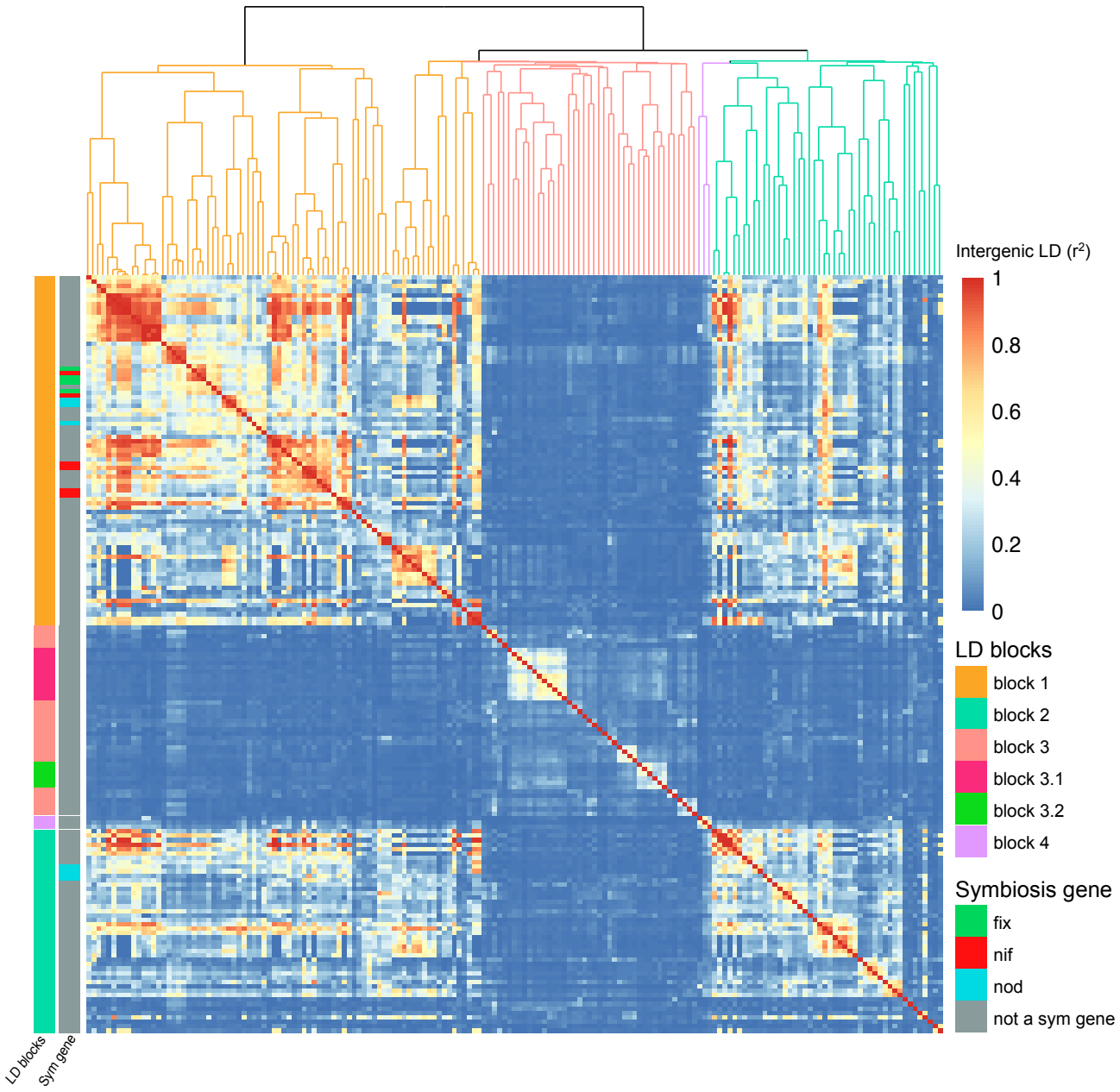

Figure S12: Introgressed genes were clustered based on the method complete of the R function hclust. Dendrogram indicates the hierarchical dependence between gene correlations (r2). The central block was manually flipped to produce the main Figure 3b.

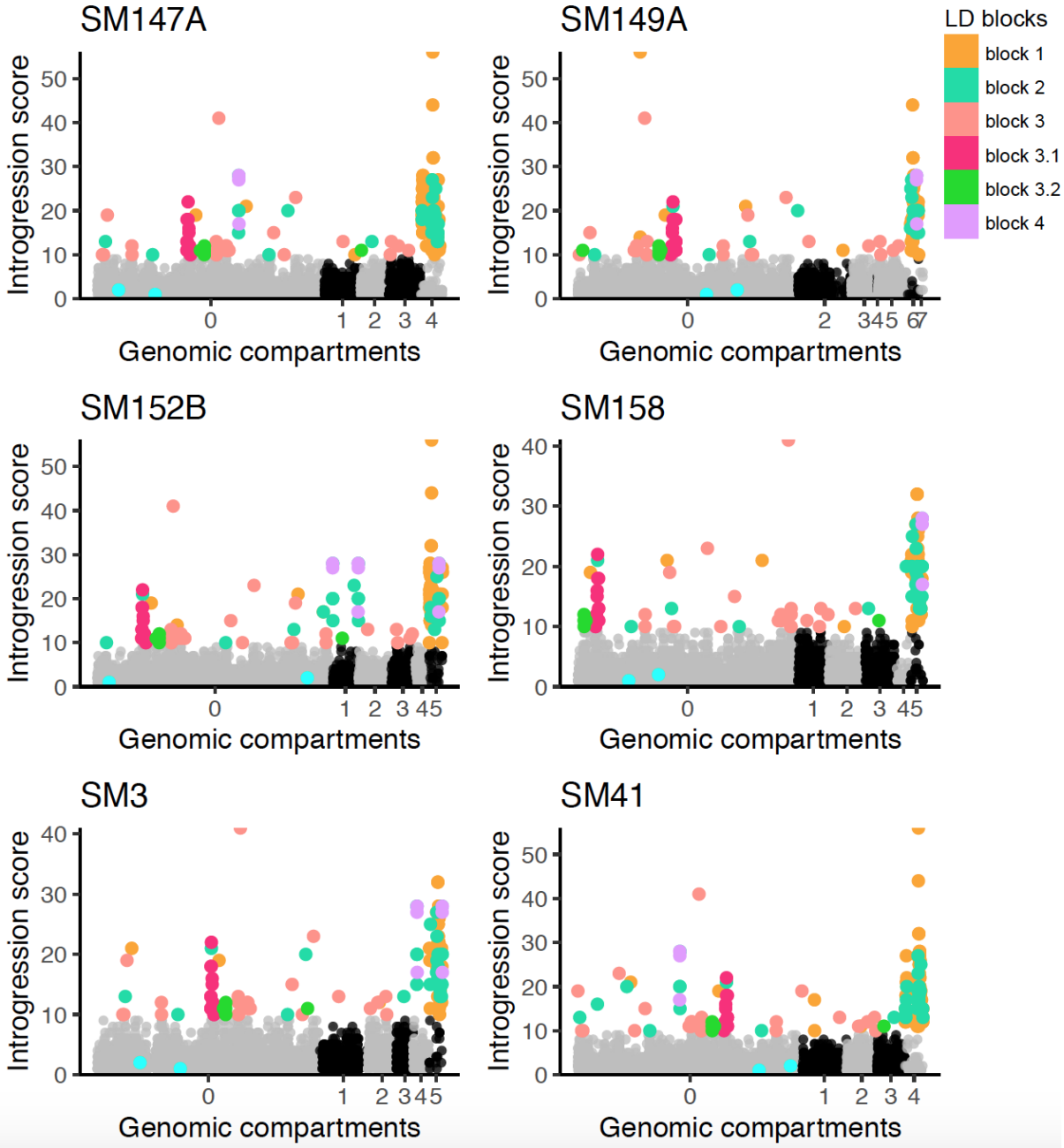

Figure S13: Orthologous gene groups were blast against pacbio assemblies and introgression score (y-axis) is plotted against genomic positions (x-axis). Grey and black dots represents the genes distributed in the different compartments (chromosome = 0, chromid = 1 and 2, >2 = plasmids). Light blue are the two conserved genes (*recA* and *rpooB*), all the other colors correspond to the linkage blocks classified by the intergenic LD analysis.

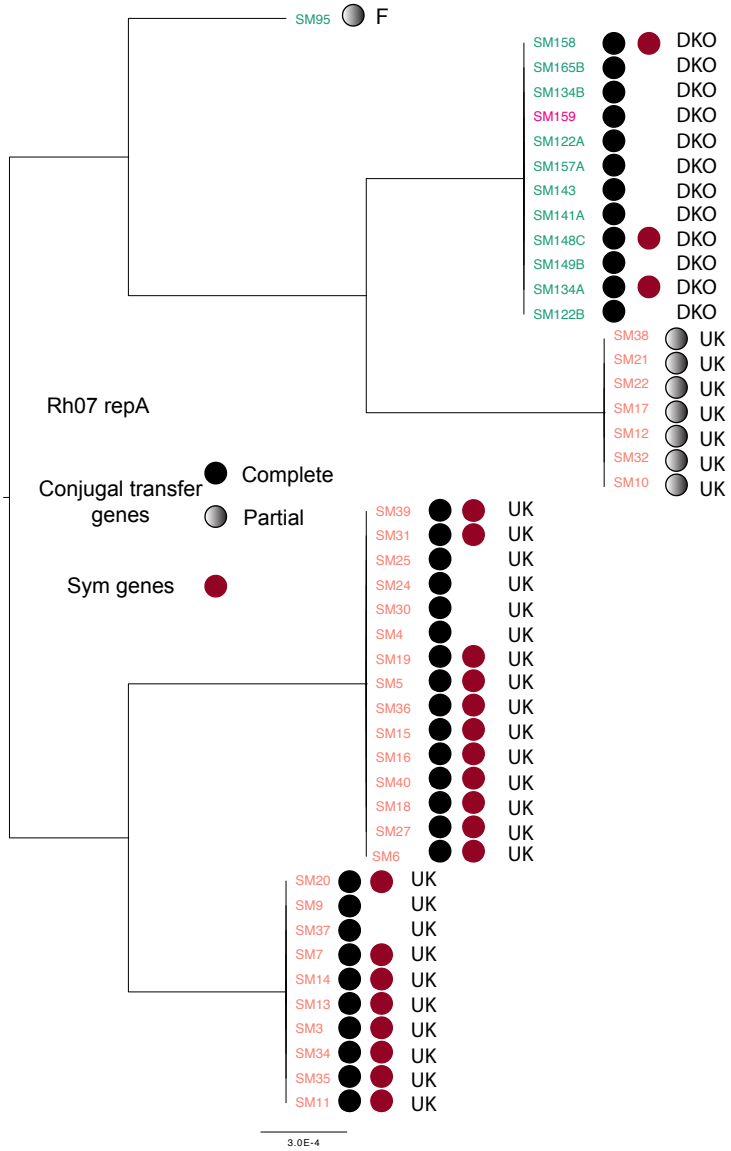

Figure S14: Phylogenetic analysis of the *repA* gene of plasmid type Rh07. DKO represents strains sampled from Danish organic fields, F from France and UK from United Kingdom. A complete set of conjugal transfer genes has the following genes upstream of *repA*: *traI*,*trbBCDEJKLFGHI*,*traRMHBFACDG*, with the origin of transfer (*oriT*) between *traA* and *traC*. Partial sets are broken by the end of the scaffold, mostly after *traM*.

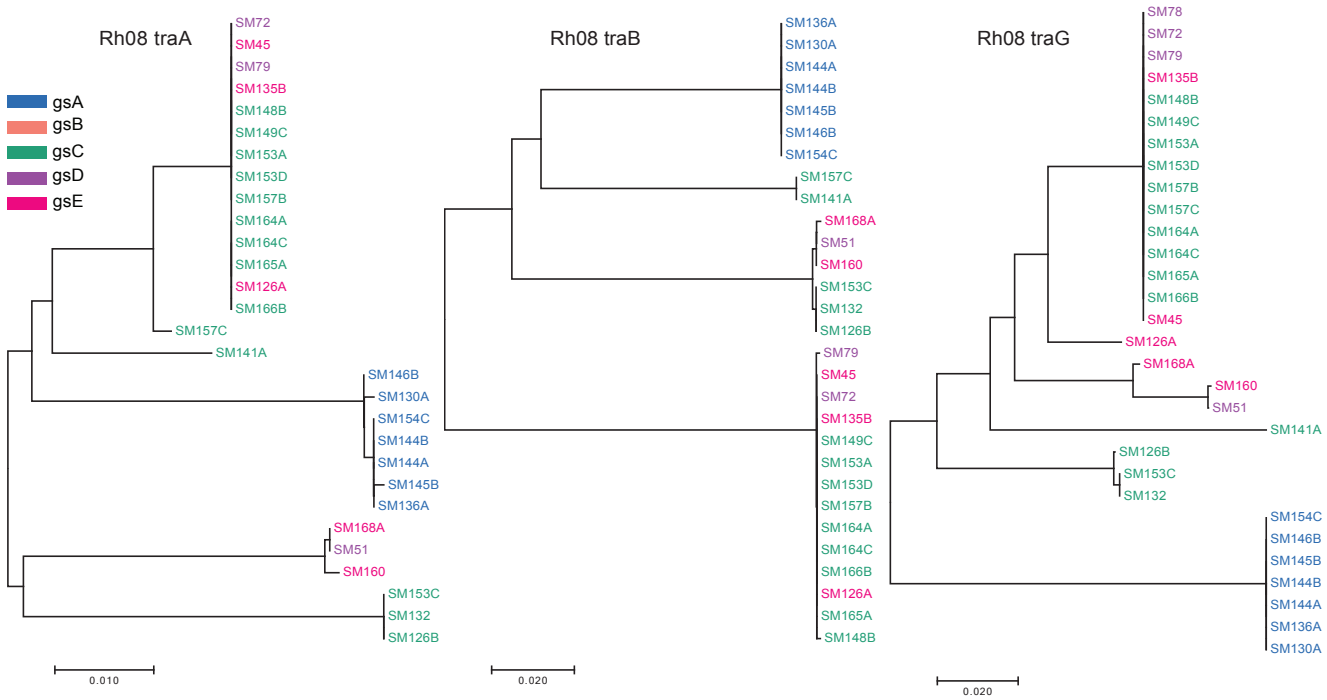

Figure S15: Phylogenetic trees of tranfer genes (*tra*) essential for the conjugation process. These genes are found in strains containing the plasmid Rh08, which is the sym-plasmid for some of the strains. A complete set of conjugal transfer genes has the following genes upstream of *repA*: *traI*,*trbBCDEJKLFGHI*,*traRMHBFACDG*.)

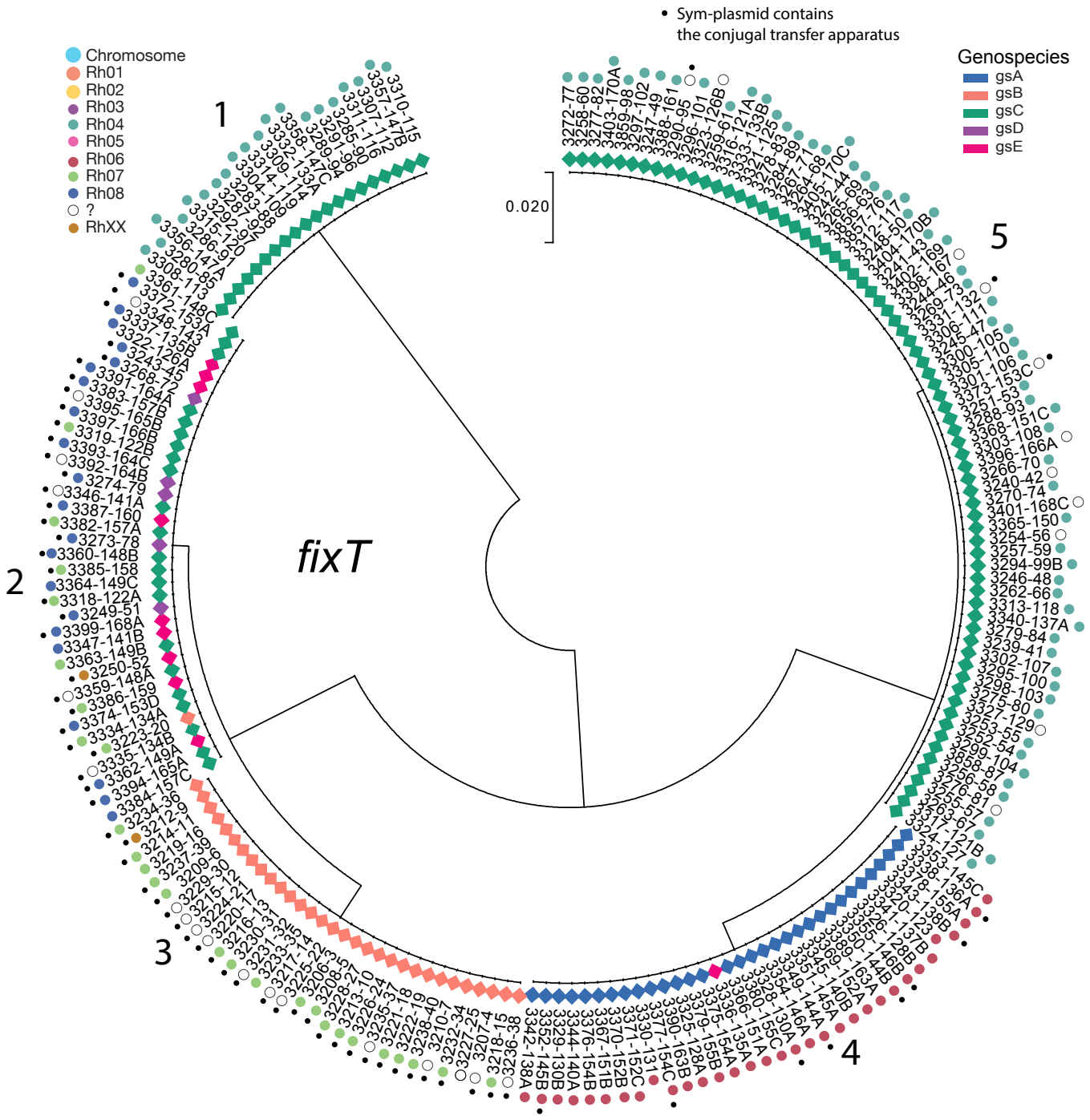

Figure S16: Phylogenetic tree of *fixT* and sym-plasmid classification. Dots correspond to strains containing a mobile sym-plasmid, with conjugal transfer system. With the exception of gsB clade (all strains from UK), no other clade is confined to a specific country of origin. All the numbers following the dash corresponds to the SM strain name.