Article

Defining the *Rhizobium leguminosarum* species complex

J. Peter W. Young [1,*], Sara Moeskjær [2], Alexey Afonin [3], Praveen Rahi [4], Marta Maluk [5], Euan K. James [5], Maria Izabel A. Cavassim [6], M. Harun-or Rashid [7], Aregu Amsalu Aserse [8], Benjamin J. Perry [9], En Tao Wang [10], Encarna Velázquez [11], Evgeny E. Andronov [12], Anastasia Tampakaki [13], José David Flores Félix [14], Raúl Rivas González [11], Sameh H. Youseif [15], Marc Lepetit [16], Stéphane Boivin [16], Beatriz Jorrin [17], Gregory J. Kenicer [18], Álvaro Peix [19], Michael F. Hynes [20], Martha Helena Ramírez-Bahena [21], Arvind Gulati [22] and Chang-Fu Tian [23]

[1] Department of Biology, University of York, York YO10 5DD, UK
[2] Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark; sm.moeskjaer@gmail.com
[3] Laboratory for genetics of plant-microbe interactions, ARRIAM, Pushkin, 196608 Saint-Petersburg, Russia; AAfonin@ARRIAM.ru
[4] National Centre for Microbial Resource, National Centre for Cell Science, Pune, India; praveen@nccs.res.in
[5] Ecological Sciences, The James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK; Marta.Maluk@hutton.ac.uk (M.M.); Euan.James@hutton.ac.uk (E.K.J.)
[6] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA; izabelcavassim@gmail.com
[7] Biotechnology Division, Bangladesh Institute of Nuclear Agriculture (BINA), Bangladesh; mhrashid08@gmail.com
[8] Ecosystems and Environment Research programme , Faculty of Biological and Environmental Sciences, University of Helsinki, FI-00014 Finland; aregu.aserse@helsinki.fi
[9] Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand; benjamin.perry@postgrad.otago.ac.nz
[10] Departamento de Microbiología, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, Cd. México, 11340, Mexico; entaowang@yahoo.com.mx
[11] Departamento de Microbiología y Genética, Universidad de Salamanca, Instituto Hispanoluso de Investigaciones Agrarias (CIALE), Unidad Asociada Grupo de Interacción planta-microorganismo (Universidad de Salamanca-IRNASA-CSIC), Salamanca, Spain; evp@usal.es (E.V.); raulrg@usal.es (R.R.G.)
[12] Department of Microbial Monitoring, ARRIAM, Pushkin, 196608 Saint-Petersburg, Russia; eeandr@gmail.com
[13] Department of Crop Science, Agricultural University of Athens, Iera Odos 75, Votanikos, 11855 Athens, Greece; tampakaki@aua.gr
[14] CICS-UBI–Health Sciences Research Centre, University of Beira Interior, Covilhã, Portugal; jdflores@usal.es
[15] Department of Microbial Genetic Resources, National Gene Bank (NGB), Agricultural Research Center (ARC), Giza 12619, Egypt; samehheikal@hotmail.com
[16] Laboratoire des Symbioses Tropicales et Méditerranéennes, UMR INRAE-IRD-CIRAD-UM2-SupAgro, Campus International de Baillarguet, TA-A82/J, 34398 Montpellier cedex 5, France; marc.lepetit@inrae.fr (M.L.); stephane.boivin@cirad.fr (S.B.)
[17] Department of Plant Sciences, University of Oxford, OX1 3RB, Oxford, UK; beatriz.jorrin@plants.ox.ac.uk
[18] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK; gkenicer@rbge.org.uk
[19] Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA-CSIC). Unidad Asociada Grupo de Interacción Planta-Microorganismo (Universidad de Salamanca-IRNASA-CSIC), Salamanca, Spain; alvaro.peix@csic.es
[20] Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary AB, Canada, T2N 1N4; hynes@ucalgary.ca
[21] Departamento de Didáctica de las Matemáticas y de las Ciencias Experimentales. Universidad de Salamanca. Salamanca, Spain; mh.ramirezb@usal.es
[22] Microbial Prospection, CSIR-Institute of Himalayan Bioresource Technology, Palampur (H.P.) 176 061, India; gal_arvind@yahoo.co.in

 doi:10.20944/preprints202012.0297.v1

[23] State Key Laboratory of Agrobiotechnology, Rhizobium Research Center, and College of Biological Sciences, China Agricultural University, Beijing, China; cftian@cau.edu.cn
* Correspondence: peter.young@york.ac.uk

Abstract: Bacteria currently included in *Rhizobium leguminosarum* are too diverse to be considered a single species, so we can refer to this as a species complex (the Rlc). We have found 429 publicly available genome sequences that fall within the Rlc and these show that the Rlc is a distinct entity, well separated from other species in the genus. Its sister taxon is *R. anhuiense*. We constructed a phylogeny based on concatenated sequences of 120 universal (core) genes, and calculated pairwise average nucleotide identity (ANI) between all genomes. From these analyses, we concluded that the Rlc includes 18 distinct genospecies, plus 7 unique strains that are not placed in these genospecies. Each genospecies is separated by a distinct gap in ANI values, usually at around 96% ANI, implying that it is a 'natural' unit. Five of the genospecies include the type strains of named species: *R. laguerreae, R. sophorae, R. ruizarguesonis, "R. indicum"* and *R. leguminosarum* itself. The 16S ribosomal RNA sequence is remarkably diverse within the Rlc, but does not distinguish the genospecies. Partial sequences of housekeeping genes, which have frequently been used to characterise isolate collections, can mostly be assigned unambiguously to a genospecies, but alleles within a genospecies do not always form a clade, so single genes are not a reliable guide to the true phylogeny of the strains. We conclude that access to a large number of genome sequences is a powerful tool for characterising the diversity of bacteria, and that taxonomic conclusions should be based on all available genome sequences, not just those of type strains.

Keywords: *Rhizobium*; species complex; bacterial taxonomy; core genes; housekeeping genes; average nucleotide identity; speciation; genospecies

1. Introduction

The increasing availability of genome-scale DNA sequencing is transforming the practice of bacterial taxonomy. Until recently, bacterial species have been described using an eclectic mixture of phenotypic characteristics plus a limited amount of DNA-based information – an approach called 'polyphasic taxonomy' [1,2]. Over time, the DNA component has come to be considered the most critical information, with the sequence of small subunit ribosomal RNA (16S rRNA) providing reliable placement down to genus level, and DNA-DNA hybridisation (DDH) used to distinguish species [3–5]. Most journals that publish new species descriptions now require a genome sequence for the proposed type strain, and there are initiatives to provide genome sequences for species described in the past [6–8]. These developments recognise the power of genome information [9–14], although genome sequences are not a formal requirement of the code for bacterial nomenclature [15]. So far, many authors have simply used the genome sequence to provide average nucleotide identity (ANI) values that are a more convenient and accurate substitute for the outdated DNA-DNA hybridisation (DDH) laboratory technique [16,17], and perhaps to extract the sequence of 16S rRNA and a few housekeeping genes that would otherwise have required separate amplification and sequencing. These are used for comparison to related species, but usually only to the type strains, i.e. a single strain representing each named species. Genome sequences are capable of providing much more information, though, particularly if they are available for multiple strains of a species and not just for the type strain. They allow core genes, which are present in all strains and experience relatively limited lateral gene transfer, to be distinguished from accessory genes, which are more sporadic in distribution, often transferred, and responsible for many of the phenotypic differences among strains and species [18]. Multiple core genes can be used to construct very robust phylogenies because discrepancies affecting individual genes are averaged out [19,20], while the distribution of accessory genes may also help to distinguish species [21].

In 1987, a committee for bacterial systematics declared that "the complete deoxyribonucleic acid (DNA) sequence would be the reference standard to determine phylogeny and that phylogeny should determine taxonomy" [4]. It would be another eight years before the first bacterial genome was sequenced [22], so the committee could only recommend DDH as the best available substitute at

the time, and defined a genospecies as including those strains with approximately 70% or greater DDH relatedness and less than 5 °C melting temperature difference. It is essentially in this sense that we use the term 'genospecies', although of course we can now use complete DNA sequences to assess relatedness much more accurately and in multiple ways. In fact, the term 'genospecies' had already been defined for bacteria by Ravin in 1963: "When their respective genotypes permit inter-bacterial genetic transfer and recombination, we may say that they belong to the same genospecies" [23]. This is closer to the widely used biological species concept as defined by Mayr [24], and suggests that recombination could provide genetic cohesion within a species. It is known that sequence divergence can prevent homologous recombination, so diverged sequences will continue to diverge further, and this could create genetic barriers between species [25,26]. While natural selection and ecology are undoubtedly also important in bacterial speciation, it is indeed observed that, in many bacterial groups, there are 'gaps' in genome-based distance measures, roughly around 70% DDH or 95-96% ANI, that allow the definition of species without too many ambiguous intermediates [27,28]. For our purposes here, we can say that a genospecies is a discrete cluster in the sequence space of core genes [29].

There have been some broad-scale initiatives to establish taxonomic databases for all bacteria and archaea based on genome sequences. The Type Strain Genome Server (TYGS) includes only type strains, but provides tools to compare any query genome with the most relevant type strains using a whole-genome similarity metric called dDDH, which is designed to correlate well with the laboratory measurements of DDH that were used in the past [30]. On the other hand, the Genome Taxonomy Database (GTDB) uses a 95% ANI threshold to define species and aims to include and classify all available genomes [20], including many that do not fit into species that have been formally described. By necessity, the project generates temporary names to accommodate these genomes. These initiatives are very welcome and potentially provide a universal framework, but the proof of their utility will come from thorough investigations at a much finer-grained level.

In the present study, we explore the potential of genome sequence data to illuminate the diversity of an important and well-studied bacterial group, *Rhizobium leguminosarum* in the broad sense, the archetype of rhizobia. Rhizobia are bacteria that induce the formation of nodules on the roots of legume plants (Fabaceae). The rhizobia colonise cells within the nodule and use energy supplied by the plant to fix nitrogen (i.e. reduce atmospheric N_2 to ammonia, NH_3) and make organic nitrogen compounds available to the plant. This is a mutually beneficial symbiosis, and its importance for agriculture has stimulated research for more than a century [31,32].

The name *Rhizobium leguminosarum* was proposed by Frank in 1889 for these root nodule bacteria [33]. In those early days, there was considerable confusion about their affinities, or even whether they were bacteria at all, but by 1926 the taxonomy had settled down and a number of species were recognised on the basis of host specificity and cultural properties: *R. leguminosarum* and *R. japonicum* [34], and *R. meliloti*, *R. trifolii* and *R. phaseoli* [35]. The taxonomy remained stable until 1982, when a new species, *R. loti*, was described [36] and, in the same year, *R. japonicum* was considered sufficiently distinct to merit a new genus, *Bradyrhizobium* [37]. Eventually, *R. meliloti* and *R. loti* were also transferred to new genera, *Sinorhizobium* and *Mesorhizobium* [38,39]. On the other hand, in the 1984 Bergey's Manual, Jordan amalgamated *R. trifolii* and *R. phaseoli* into *R. leguminosarum* because the only consistent difference among the three was their host specificity [40], which was plasmid-encoded and could be transferred among strains [41–43]. Jordan proposed three biovars, *viciae*, *trifolii* and *phaseoli*, to accommodate the three distinct host-specificity types within *R. leguminosarum* [40]. Since then, there has been a proliferation of new species described within the genus *Rhizobium*, as well as new genera for more distant species of root-nodule bacteria [44]. Many strains originally described as *R. leguminosarum* have been moved to new species, and it is clear that the three biovars, now called symbiovars [45], are not confined to *R. leguminosarum*, because almost identical nodulation genes (host-specificity determinants) can be found in other *Rhizobium* species [46]. Ramírez-Bahena et al. [47] reconsidered the taxonomic status of *R. leguminosarum* with the benefit of gene sequence data and concluded that, while *R. trifolii* was indeed a synonym of *R. leguminosarum*, the type strain of *R. phaseoli* defined a separate species (although some other strains of symbiovar

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

phaseoli do belong to *R. leguminosarum*). They also discovered that two different bacteria were being distributed as the type strain of *R. leguminosarum*, decided which was the true USDA 2370^T, and defined the other (DSM 30132^T) as the type of a new species, *R. pisi*.

Despite the description of many new species, the genetic diversity that remains within *Rhizobium leguminosarum*, as currently defined, is still greater than is considered typical for a single bacterial species. Kumar et al. [29] noted that the genome-based diversity of *R. leguminosarum* isolates from legume nodules at a single site was ten times higher than that of *Sinorhizobium medicae* from the same site [48]. They divided the isolates into five genospecies (A to E) that were separated by ANI values below 95%, a level that is widely considered to be evidence that they represent separate species [49]. They also considered the small number of other published *R. leguminosarum* genomes available at that time, and found that some could be accommodated within the five genospecies, while others potentially represented additional genospecies. A larger survey of almost 200 genomes from three north European countries found only the same five genospecies [21], but another study with genomes from a wider European geographic range included some that fell into two novel genospecies that were closely related to each other [50]. The authors called these gsF-1 and gsF-2, and the second of these included strain FB206T, the type strain of *R. laguerreae*, a species that was described in 2014 [51]. This description was based on six isolates that were distinct from USDA 2370^T, the type strain of *R. leguminosarum*, in housekeeping-gene sequences and by DNA-DNA hybridisation, even though their 16S rRNA sequences were all identical to that of USDA 2370^T. As measured by ANI, strains of gsB and gsC are just as divergent from USDA 2370^T as are those of *R. laguerreae* [50], so it seems likely that these genospecies could also be described formally as distinct species using conventional taxonomic criteria. The type strain USDA 2370^T is in gsE, so this genospecies remains *R. leguminosarum* in the narrow sense.

Recently, several published [21,50,52] and unpublished [53] studies have augmented the public databases with a large number of new genome sequences assigned to, or related to, *R. leguminosarum*. In addition, there have been numerous contributions of individual genomes [54–62]. Here, we examine this wealth of genomic information in order to establish whether *R. leguminosarum*, in the broad sense, is a distinct entity, and whether it can be subdivided into clearly defined groups that might, in future, be named as separate species within the overall *R. leguminosarum* species complex (Rlc).

2. Materials and Methods

2.1. The set of genome sequences

All genome sequences for the genus *Rhizobium* (NCBI:txid379) were downloaded in fasta format from the NCBI database (www.ncbi.nlm.nih.gov) using the script genbank_get_genomes_by_taxon.py that is distributed as part of the pyani package (http://widdowquinn.github.io/pyani/). Initial analysis was based on the 834 genomes available on 25 July 2020 (Table S2). All later analyses were confined to the genomes that belonged to the Rlc or to its sister taxon *R. anhuiense*, using the genomes that were available on 28 August 2020. After eliminating duplicate sequences of the same strain, there were 429 genomes in the Rlc and 11 in *R. anhuiense* (440 genomes altogether). These are listed in Table S1, together with additional relevant information. Files were renamed to include the strain name as well as the accession name, and also the known genospecies in the case of the strains previously described by Cavassim et al. [21].

2.2. The core-gene phylogeny

A core gene phylogeny was constructed from the bac120 set of 120 universal bacterial core genes used by Parks et al. [63]. Whereas Parks et al. used protein sequences for a phylogeny encompassing all bacteria, we used the DNA sequences to gain maximum resolution of closely related strains. The list of protein IDs in Supplementary Table 6 of Parks et al. [63] was used to identify these proteins in the annotated complete genome sequence of WSM1325 (GCF_000023185.1; [64]), or another Rlc strain if not annotated in this genome (Table S11), and the set of protein sequences from this strain was used to find

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

the corresponding genes in each genome using tblastn [65] with an E-value cutoff of 1E-10. Hits were extended at each end by the expected number of nucleotides to obtain the full gene sequence, or to the end of the contig, if nearer. All 120 genes were located in the chromosome in the fully-assembled genome of strain 3841 [66]. Sequences for each gene were aligned using clustalo 1.2.3 [67], then all 120 genes were concatenated before constructing an approximately-maximum-likelihood phylogeny using fasttree 2.1.10 [68], with local bootstrap branch support values. Of the initial 834 genomes (Table S2), 37 lacked some or all of the core genes (Table S3; presumably these assemblies were incomplete), while 797 had all 120 genes and were included in Figure 1. The trees were displayed using Dendroscope 3.7.2 [69] for initial exploration, FigTree 1.4.3 (https://github.com/rambaut/figtree/releases) for Figure 1, and iTOL [70] for other figures. Python scripts are available at https://github.com/jpwyoung/Rlc and the iTOL trees at https://itol.embl.de/shared/rhizobium.

2.3. Average nucleotide identity (ANI)

Pairwise ANI was calculated using fastANI 1.31 [71] with the default settings (kmer = 16, fragment length = 3000, minimum shared fraction = 0.2), and displayed using the seaborn library with custom python scripts available at https://github.com/jpwyoung/Rlc. Within the Rlc plus *R. anhuiense*, the number of sequence fragments per genome ranged from 2189 to 3829, and the number matched in pairwise comparisons ranged from 1640 to 3825.

2.4. Analysis of 16S and nodulation genes

The 16S rRNA sequence of USDA 2370^T was used as the query to find the most similar sequence in each genome using blastn (default parameters). To assign the symbiovar, NodC, NodA and NodD protein sequences of strains representing the three symbiovars (3841 for *viciae*, WSM1325 for *trifolii*, *R. etli* CFN42 for *phaseoli*) were used as queries to search each genome using tblastn (E-value 1E-5) to find the corresponding genes, and the symbiovar determined by the best match. Each of the three genes gave the same result, except that a few SEMIA strains had 100% match to the CFN42 NodA sequence, but no match to NodC or NodD, so were recorded as non-nodulating. The 16S and *nodC* sequences were extracted and aligned for phylogeny as described for core genes. Three assemblies had no 16S sequence (128C53, Ps8, Vh3), and a further three (FB403, SP4, FA23) were omitted because their 16S sequence was incomplete.

2.5. Housekeeping genes for genospecies assignment

Three housekeeping genes that have been widely used to characterise isolates, *recA* [72], *atpD* [72] and *gyrB* [73–75], were located in each genome by tblastn (E-value 1E-10) using the corresponding protein sequences annotated in an arbitrary Rlc strain (SM130B, gsA). To simulate the procedure of a typical application, the gene sequences were trimmed to the expected informative part of the amplicon (excluding primers) and aligned separately using ClustalW [76] before concatenation in the order *atpD-gyrB-recA* and phylogenetic analysis using the neighbor-joining method [77] in the MEGA X software [78]. Branch support was assessed by bootstrap values based on 1000 replications.

Sequences corresponding to the shorter amplicons used by Fields et al. [79] for high-throughput sequencing of parts of *recA* and *rpoB* from bulk nodule DNA were located by blastn (E-value 10) using sequences of strain SM3 (gsB) from [79] and aligned with clustalo for phylogeny with fasttree, as for core genes (above).

2.6. ANI of chromosomal and plasmid compartments

The scaffolds of each genome assembly were separated into chromosomal and nonchromosomal (plasmid) compartments. Scaffolds were classified as chromosomal if they carried one or more of the 3215 genes identified by Cavassim et al. [21] as normally chromosomal in Rlc genomes, as assessed by blastn (E-value 1E-10, only the best hit in each genome). The average plasmid compartment was 2.496 Mb (range 0.957 – 4.436 Mb), making up 32.8% (12.6 – 47.0%) of the total genome. ANI was calculated

 doi:10.20944/preprints202012.0297.v1

separately for the two compartments using fastANI 1.31 [70] and displayed using the seaborn library with custom python scripts available at https://github.com/jpwyoung/Rlc.

2.7. Identification and analysis of ortholog sets

The aim here was to assess the degree to which accessory genes were shared between strains of the same or different genospecies. Putative genes were identified in each genome assembly using Prodigal 2.6.3 [80]. The predicted proteins were sorted into groups of orthologs using Orthofinder 2.4.0 [81]. The number of orthogroups shared between each pair of genomes, and the total number of orthogroups in each genome, were extracted from the output file Orthogroups.GeneCount.tsv. Genes in single copy in a single genome were not included. Orthogroups with more than two copies in any one genome were excluded, as these often had ambiguous or obscure orthology. A normalised gene sharing index between two strains was calculated as the average fraction of the orthogroups present in a strain that were shared with the other strain. The computing for this part of the project was performed on the Aarhus University GenomeDK cluster. It was not computationally feasible to run Orthofinder on the full set of 440 genomes, so the Rlc was split into four major clades (gsL+M+C, D+E, H+A, the rest) and each was run separately. In addition, a set of 100 genomes that included examples from each genospecies, and nine additional smaller sets, were used to sample values for more distantly related strain pairs. Across all runs, the average number of orthogroups was 12,829 (range 10,314-16,965) before filtering on copy number, and 12,198 (range 10,019-15,811) after filtering.

3. Results and Discussion

3.1. The genus Rhizobium *and related genera*

All genomes identified as the genus *Rhizobium* and available at NCBI were downloaded on 25 July 2020 (834 genomes, Table S2). The sequences of 120 core genes were extracted from each genome assembly. The full set of core genes was found in most assemblies, but 37 were incomplete (Table S3). A phylogeny (not shown) based on all assemblies that were complete enough to have at least 100 of the genes showed that GCF_003001755.1, annotated as *R. tropici* NFR14, was an outlier. Analysis of this assembly by the TYGS server (https://tygs.dsmz.de/) indicated that it actually belonged to the genus *Bradyrhizobium*, so this was used as an outgroup to root the tree, together with *R.* sp. FKL33, a distant member of the *Rhizobiaceae* according to the Genome Taxonomy Database (GTDB, https://gtdb.ecogenomic.org). The tree for genomes that had a complete set of 120 core genes is shown in Figure 1.

Many strains in the basal clades of this phylogeny have featured in previous studies that classified them in genera related to, but distinct from *Rhizobium*, and this allowed us to identify the main clades as *Allorhizobium* [82,83], *Neorhizobium* [83,84], *Pseudorhizobium* [85,86], *Agrobacterium* [87], *Mycoplana* [88], *Pararhizobium* [83], and an unnamed genus-level clade identified in GTDB as g__Rhizobium_A, which includes the type strains of *Rhizobium rhizoryzae* and *Allorhizobium pseudoryzae*. It should be emphasised that the only strains of these genera that are shown in Figure 1 are those that are misclassified as *Rhizobium* in the NCBI database. The database also includes a much larger number of related strains that are correctly classified. The systematic use of ANI can be expected to improve the NCBI classification over time [89]. In Table S4 we list the genera to which these strains appear to belong. Clearly, there is scope to use genome sequences to improve our understanding of all these genera, but that is outside our present focus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

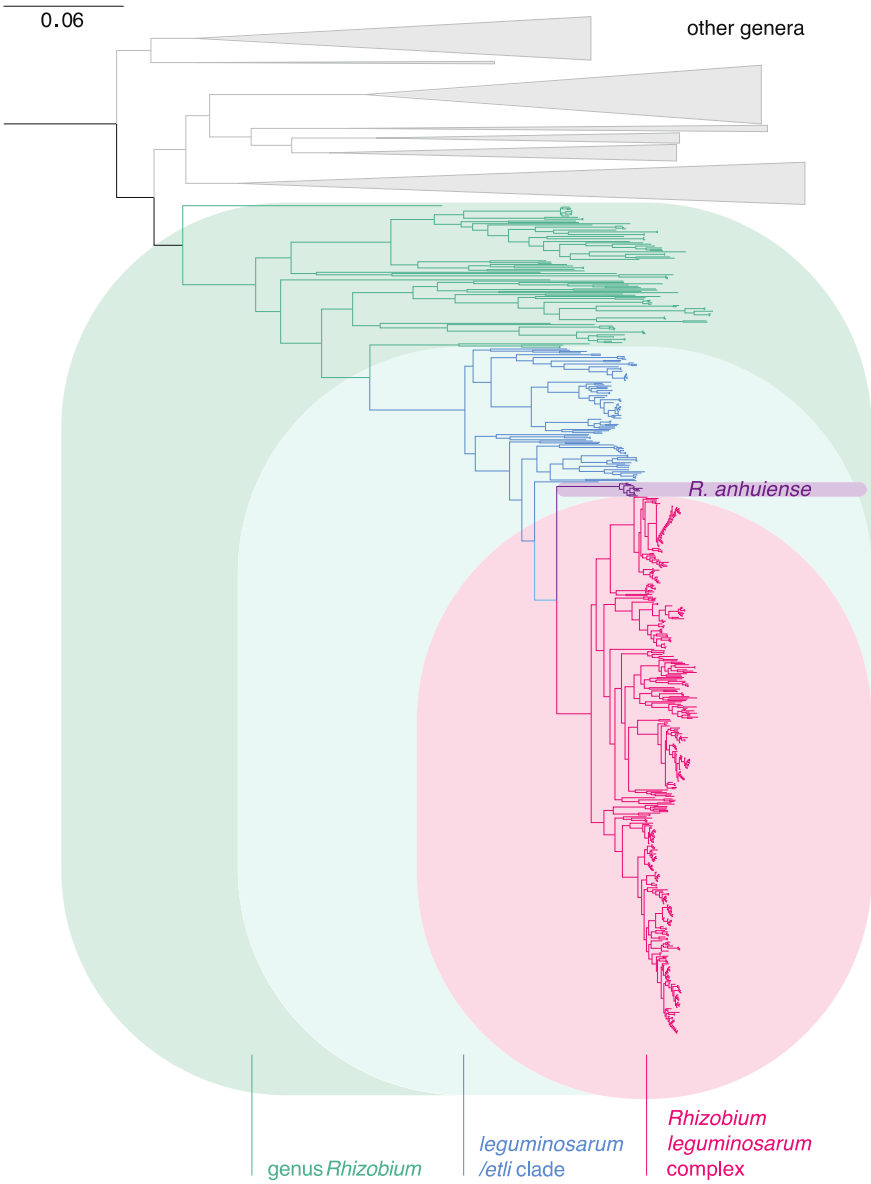

Figure 1. Phylogeny, based on 120 core genes, of genomes assigned to the genus *Rhizobium* by NCBI (NCBI:txid379). Only the 797 genomes that had all 120 genes were included. Genomes that are not currently included in the genus *Rhizobium* are indicated in the basal clades, collapsed by genus: *Pararhizobium, Mycoplana, Agrobacterium,* GTDB g__Rhizobium_A [20], *Pseudorhizobium, Neorhizobium, Allorhizobium* (top to bottom). The tree was rooted using strain FKL33 (not shown).

The genus *Rhizobium*, as currently understood, is separated by a long branch from other genera. Its sister clade is a single strain, "*R. album*" NS-104^T [90], that has no sequenced relatives and is not clearly affiliated with any known genus. The genomes that are in *Rhizobium* but not in the Rlc are listed in Table S5. Within *Rhizobium*, a very long branch distinguishes a clade that includes *R. leguminosarum*, *R. etli*, and a number of more recently described species, including *R. laguerreae, sophorae, indicum, ruizarguesonis, anhuiense, ecuadorense, acidisoli, chutanense, hidalgonense, vallis, pisi, fabae, bangladeshense, sophoriradicis, phaseoli, esperanzae, aethiopicum, aegyptiacum, binae* and *lentis*.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

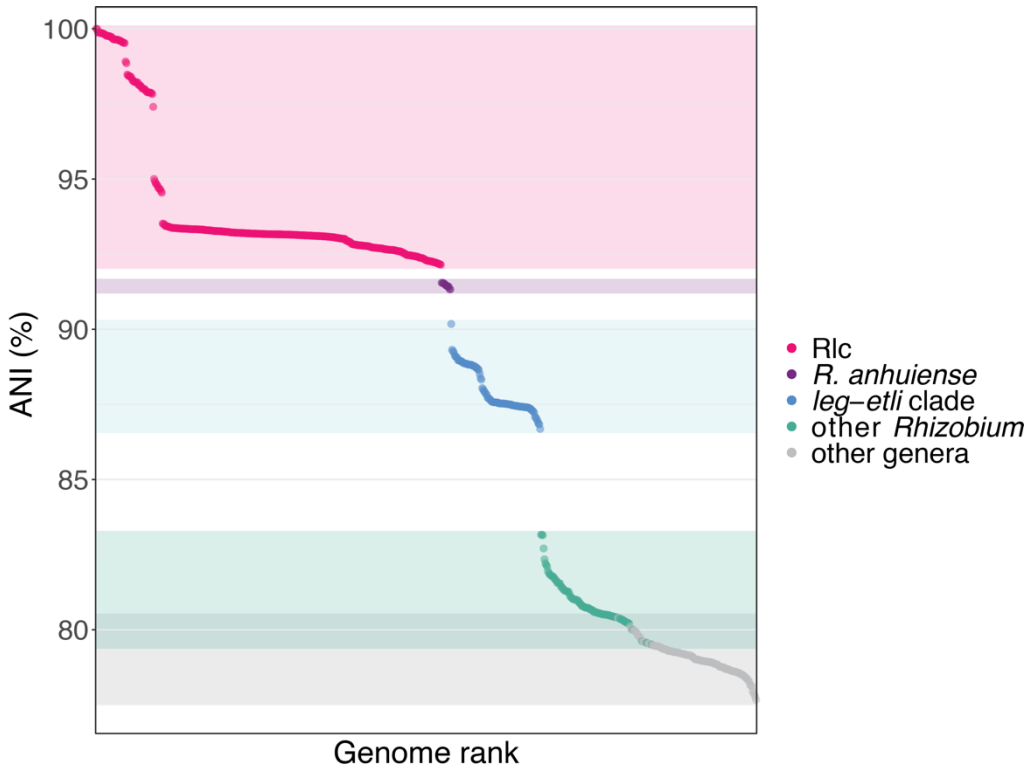

Figure 2. Average nucleotide identity (ANI) to the type strain of *R. leguminosarum* (USDA 2370T) of 766 of the genomes shown in Figure 1, in rank order of ANI. Some more distant genomes were omitted because ANI could not be calculated.

3.2. The Rlc has a clear boundary

The group that we identify as the Rlc forms a distinct subclade within the *leguminosarum-etli* clade, and its sister taxon is *R. anhuiense* (Figure 1). The clear boundary between the Rlc and other *Rhizobium* species was further supported by an analysis of genome-wide ANI, using the type strain of *R. leguminosarum* (USDA 2370^T) as the reference (Figure 2). All genomes within the Rlc have ANI values of 92.15% or higher with respect to USDA 2370^T. Values for *R. anhuiense* are tightly grouped in the range 91.33-91.55%, while the rest of the *leguminosarum-etli* clade starts at 90.18% and ranges down to 86.68%. After that, there is a large drop in ANI to *R. alamii* (83.16%), the sister taxon of the *leguminosarum-etli* clade, reflecting a long branch in the tree. There are no further breaks in ANI, even at the boundary of the genus *Rhizobium*. Indeed, there is some overlap in ANI values between comparisons within *Rhizobium* and those with related genera. This seems a little surprising, considering that the phylogeny shows a long, well supported branch at the base of the genus, but ANI (and especially FastANI) is not very sensitive at values below 80% [71]. Overall, there is very good agreement between the boundaries seen in the phylogeny and those detected by ANI, and there is no ambiguity about which genomes belong within the Rlc and which are outside it. However, since strains within the Rlc can have ANI values as low as 92.15% with the type strain of *R. leguminosarum*, it is evident that the diversity within the Rlc is greater than can be encompassed within a single bacterial species (for which a threshold is usually considered to be around 95-96%). An important question is whether the structure of the Rlc is amorphous or can be broken down into a set of clearly defined species-level units. That is the main goal of our study.

There were genomes of 429 strains within the Rlc available from NCBI by 28 August 2020, and these genomes, together with the 11 genomes of *R. anhuiense* as an outgroup, were used for all subsequent analyses. They are listed in Table S1, while Table S6 lists a few additional Rlc genomes that were not used because they were duplicates.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

3.3. Genospecies can be defined within the Rlc

A core-gene phylogeny of the Rlc and its sister clade *R. anhuiense* is shown in Figure 3. The Rlc is highly diverse, with many well-supported clades at various depths. To help us to decide which clades were sufficiently distinct to be considered genospecies, we also considered ANI values. In Figure 4a, which shows ANI values for all pairwise comparisons among the 440 strains, the five genospecies (A-E) defined previously [29] are clearly visible as red squares (ANI > 96%). There are many strains that fall outside these five genospecies, and most of them are also in clusters that are potential new genospecies (identified by the letters G-S) because within-cluster ANI values are above 96%.

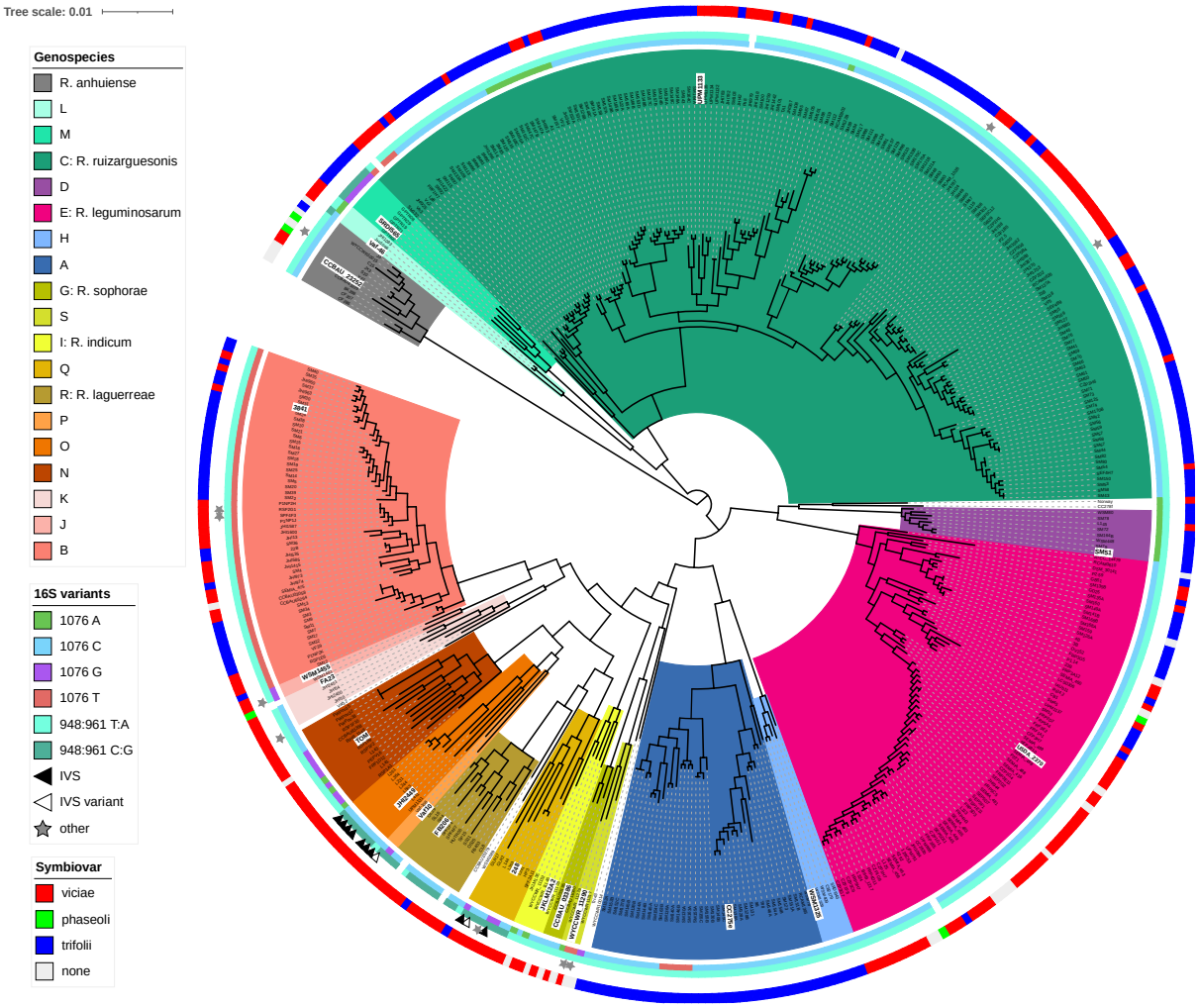

Figure 3. Phylogeny of the Rlc based on 120 core genes. The tree is rooted using *R. anhuiense* as the outgroup. The 18 proposed genospecies are indicated by coloured segments. Inner circles and symbols indicate variation in the sequence of the 16S rRNA gene (see text for details). The outer circle indicates the symbiovar inferred from the sequence of Nod genes (if any are present in the assembly). The representative strain for each genospecies is highlighted in white. An interactive version is available at https://itol.embl.de/shared/rhizobium.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

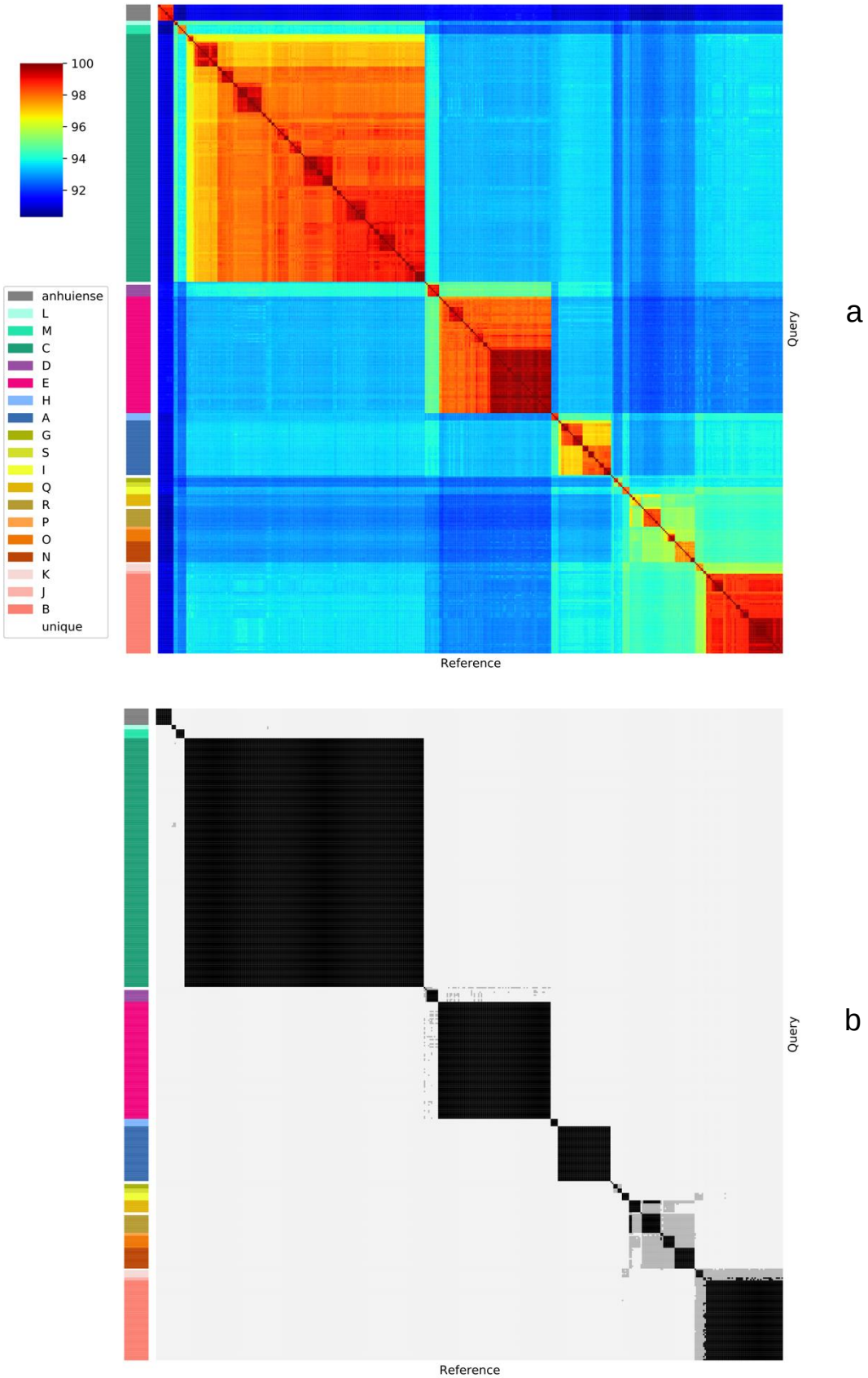

Figure 4. Pairwise average nucleotide identity (ANI) between all genomes in the Rlc and *R. anhuiense*. The bars at the left indicates genospecies of each genome. (a) Continuous colour scale. (b) The same data, but with thresholds to indicate values over 95% (grey) or over 96% (black).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

Published studies of ANI usually suggest a threshold for species designation of 95-96% [16,17,27]. To see the effect of these thresholds, the ANI data of Figure 4a are plotted again in Figure 4b with cutoffs at 95% and 96%. For most clades, there is no ambiguity if a 96% threshold is used. The exceptions are the two strains in genospecies J (gsJ), which have ANI values above 96% with some, but not all, strains of gsB, and the clade that includes gsN, O, P, Q and R, which are all closely related. Collectively, we call this the F-clade, since two potential new genospecies in this group were previously called F-1 and F-2 [50]. Our new analysis suggests that there are several genospecies in this group so, to avoid confusion, we have not used F as a genospecies name. Lowering the threshold to 95% (grey areas in Figure 4b) would lead to increased ambiguity for some of the genospecies.

Altogether, a 96% ANI threshold allows us to define 18 potential genospecies within the Rlc that each have at least two genome sequences, plus 7 single strains that have no close relatives. If we want to define genospecies that reflect real biological units, we have to recognise that a single arbitrary threshold for ANI may not be applicable. We need to assess each potential group for coherence and distinctness. For each of our proposed genospecies, we selected a representative strain, using the type strain where taxonomic species have already been defined, or a representative strain that has been well studied, has a good-quality genome, and is reasonably central within the genospecies. Using each of these 18 strains in turn as the reference, ANI values for all 440 strains are plotted in the panels of Figure 5. These plots provide justification for each of the genospecies we are proposing. In each plot, there are gaps in the ANI values, including a gap at around 96% ANI that we take to be the boundary of the genospecies. This gap is sometimes very large, e.g. more than 2 percentage points for gsE (between 95.00% and 97.40%), but small for some species in the F-clade, e.g. the gap for gsO is between 95.96% and 96.30%.

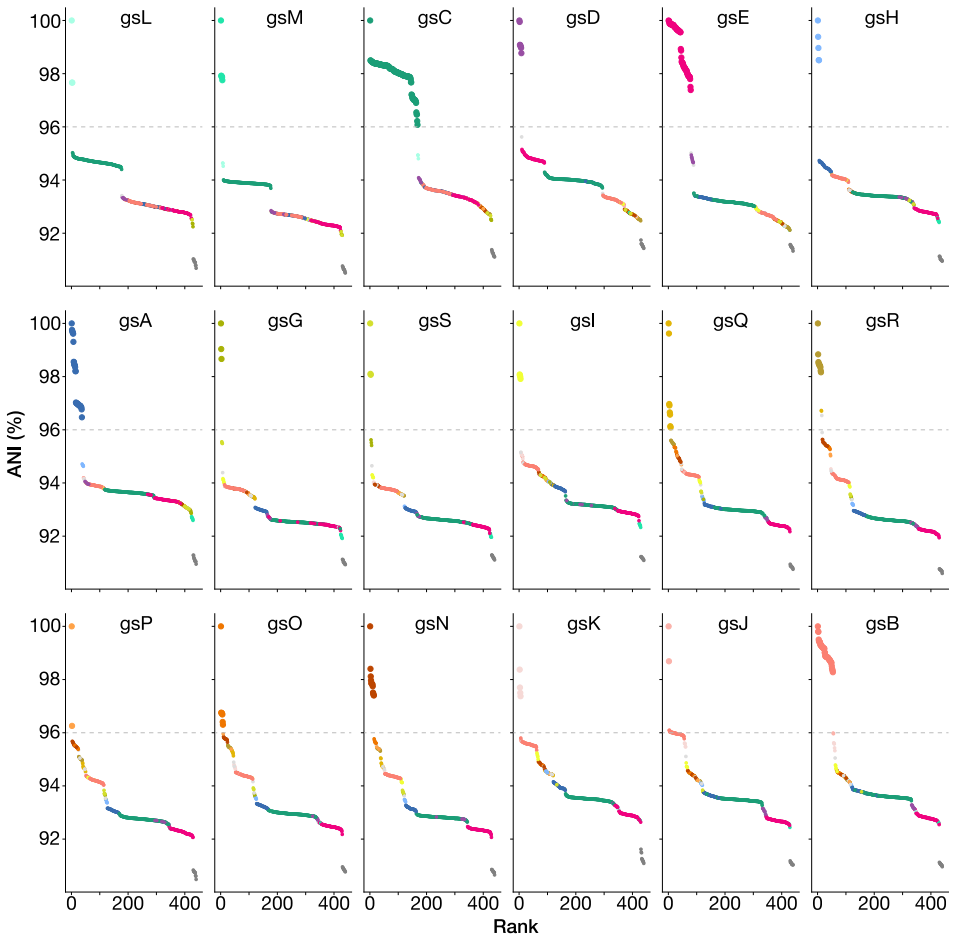

Figure 5. ANI plots using the representative strains of each of the 18 genospecies as reference. Points are coloured by genospecies. Strains that are conspecific with the reference are shown with larger symbols. Dashed line at 96% ANI indicates conventional species boundary.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

Genospecies vary in their compactness: all gsB strains have ANI values above 98% (not only with the reference strain, but also in all pairwise combinations), whereas values for gsA and gsC extend closer to 96%. We consider that the compactness of gsB justifies keeping gsJ as a separate species, even though it currently has just two strains and their ANI to some gsB strains is slightly above 96%, because it is likely that all the gsB strains share a large set of characteristics that may not be present in gsJ.

The F-clade (gsN, O, P, Q, R) is a part of the Rlc that seems to have a less clearly-defined structure than the rest. Most, though not all, of the pairwise ANI values within the F-clade are above 95% (Figure 4b). At a 96% threshold, the five proposed genospecies are generally well defined. However, there are two strains, SPF2A11 and HP3, that have ANI values above 96% to all members of gsQ, but also to most members of gsR (visible as black bars in Figure 4b). We have assigned them to gsQ on the basis of the phylogeny (Figure 3). *R. laguerreae sensu stricto* (gsR) is very compact, with all ANI values above 98%, so we have not included the related strain CCBAU10279 within it, even though its ANI to the type strain is 96.54%.

In the core gene phylogeny (Figure 3), every genospecies is a clade with 100% local bootstrap support for its subtending branch, except gsP, which has a support of just 30.3%. There are only two genomes in gsP, Vaf10 and Vaf108, and it is notable that they have long individual branches in the phylogeny, although their pairwise ANI is 96.25%, which is consistent with including them in the same species and higher than their ANI with any other genome.

The core gene phylogeny is based on a set of 120 genes. To test its robustness, we split the genes arbitrarily (by TIGR number, Table S11) into two independent sets of 60 genes and repeated the analysis. The resulting phylogenies (Figure S1) are highly congruent with the original phylogeny. All 18 genospecies are supported by 100% bootstraps, except that gsP has just 91.3% support in the set B tree and the two strains of gsP in are not grouped together in set A, with Vaf108 becoming internal to gsO. This is further evidence that gsP (Vaf10 and Vaf108) may not be a robust grouping. There are some minor differences between the three trees in the relative placement of a few of the genospecies. We can conclude that the set of 120 core genes is a sufficiently large sample to provide a reliable phylogeny for defining the genospecies.

Out of the 429 Rlc genomes currently available, just seven do not fit into any of the 18 genospecies that we have defined: CC278f and Norway that are related to gsD, the deeply branching WYCCWR10014, Tri-43 related to gsS and gsG, WSM1689 and CCBAU10279 related to gsR, and Vaf12 related to gsK. Apart from CCBAU10279 (discussed above), none of these have ANI above 96% with any other genome. It is possible that each of these is the first sequenced example of a novel genospecies, but there might also be technical reasons for their divergence from other genomes, so they are best left until additional examples are discovered.

Five of the genospecies include the type strains of species whose names have been effectively published. Genospecies E includes *R. leguminosarum* USDA 2370^T, so this genospecies is the species *R. leguminosarum* in the narrow sense (*sensu stricto*). Similarly, gsR is *R. laguerreae*, gsG is *R. sophorae*, gsC is *R. ruizarguesonis*, and gsI is *"R. indicum"*. In each case, we have used the published type strain as the representative of the genospecies in our analyses.

3.4. The genospecies are consistent with genomic taxonomy databases

When the genomes of the 18 strains representing the proposed genospecies were submitted to TYGS [30], those representing gsC, gsE, gsG, gsI and gsR were correctly identified as *R. ruizarguesonis*, *R. leguminosarum*, *R. sophorae*, *"R. indicum"* and *R. laguerreae*, respectively, while the remaining 13 genomes were all described as 'potential new species' (Table S7). Furthermore, the similarity score dDDH(d_4), which should be above 70% for conspecific strains, was no greater than 66.1% for any of the pairwise comparisons among the 18 representative strains, supporting our proposition that they represent 18 distinct genospecies. All of the seven unique strains that did not fall within these genospecies had dDDN(d_4) values below 70% with the representative strains of their closest genospecies (Table S8). Most were significantly below (65.8% or less), but CCBAU10279 scored 69.8% (C.I. 66.8-72.6%) against FB206, the type strain of *R. laguerreae* (gsR). This borderline value is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

consistent with the ANI value of 96.54 and our decision to leave this strain outside gsR despite this. Overall, then, we can conclude that our proposal to split the Rlc into 18 genospecies and 7 unique strains is completely supported by the dDDH-based species definition used by TYGS.

The Genome Taxonomy Database (GTDB) is another resource for genome-based taxonomy that provides a comprehensive genome-based taxonomy of all prokaryotes [20]. The database includes 287 of the 329 Rlc strains that we have studied, and they are assigned to ten species plus two single isolates. These species designations are consistent with the genospecies we have defined, but with some 'lumping' (Table 1), as can be expected because GTDB uses a fixed 95% ANI threshold. All strains in the F-clade (gsN, gsO, gsP, gsQ, gsR and two unique strains) are assigned to a single species, s__Rhizobium laguerreae. The isolates CC278f and Norway are included within gsD as s__Rhizobium leguminosarum_K. The three related genospecies gsB, gsJ and gsK are combined in s__Rhizobium leguminosarum_L. The three small genospecies gsG, gsS and gsI do not yet have any representatives in GTDB. In other respects, the GTDB definitions of species within the Rlc are the same as ours.

Table 1. Equivalence between our genospecies and the species-level taxa defined in the Genome Taxonomy Database (https://gtdb.ecogenomic.org).

Genospecies or strain	GTDB species
R. anhuiense	s__Rhizobium anhuiense
L	s__Rhizobium leguminosarum_D
M	s__Rhizobium leguminosarum_I
C	s__Rhizobium leguminosarum_C
D + CC278f + Norway	s__Rhizobium leguminosarum_K
E	s__Rhizobium leguminosarum
H	s__Rhizobium leguminosarum_J
A	s__Rhizobium leguminosarum_E
WYCCWR10014	s__Rhizobium sp001657485
Tri-43	s__Rhizobium leguminosarum_M
G	not represented
S	not represented
I	not represented
Q, WSM1689, CCBAU10279, R, P, O, N	s__Rhizobium laguerreae
Vaf12	s__Rhizobium sp005860925
K, J, B	s__Rhizobium leguminosarum_L

3.5. The genospecies could be the basis for new formal taxonomic names

Several valid species names already exist within the Rlc. The type strain of *R. leguminosarum*, USDA 2370^T, is centrally placed within gsE (Figure 3), and this genospecies has a very clear boundary in the region of 96% ANI (Figure 5), so there is no doubt that gsE is synonymous with *R. leguminosarum sensu stricto*, providing a new, narrower definition of the species. Two much smaller clades are also readily equated with named species: gsG is *R. sophorae* [91] and gsI is "*R. indicum*" [59] (not yet a validated name). Another recent name is *R. ruizarguesonis* [58], described on the basis of four closely-related strains, but clearly embedded within the large and diverse gsC, which we have concluded is best considered a single genospecies. Hence, gsC is *R. ruizarguesonis*, although the species description may need to be revisited to encompass this much greater diversity.

Finally, there is *R. laguerreae*, the first of these 'new' species to be named [51]. The description included six strains in this species, but no genome sequences were provided. Two strains have subsequently been sequenced, the type strain FB206^T and FB403; both are in gsR. The species description provides a phylogeny based on three housekeeping genes, *rpoB*, *recA* and *atpD*, which shows that three of the other strains are close enough that it is safe to conclude that they are also gsR.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

The sixth strain, CVIII4, was assigned to *R. laguerreae* because it had 82% DNA-DNA hybridisation with the type strain. The evidence from housekeeping gene sequences is more ambiguous, though. The *atpD* sequence places it closest to FB403 and HUTR05 in gsR, but the closest *rpoB* match is GLR2 (gsQ) followed by UPM1131 (gsO), while the *recA* sequence is highly divergent and closest to the three strains in gsL. This explains the diverged position of this strain in the concatenated phylogeny of the three genes, and illustrates once again the hazards of basing taxonomic assignments on single genes. In summary, it is not clear whether all the strains included in the description of *R. laguerreae* are in gsR. We have already commented that deciding on appropriate genospecies boundaries in the F-clade (gsN-R) is not straightforward. If genospecies are defined narrowly, as we have done, some pairs of strains in different genospecies have high similarity (>96% ANI), whereas if the whole F-clade is treated as a single species, many pairs are more distant that is usual within a species (<95% ANI). In the GTDB classification, based on a 95% ANI threshold, the whole F-clade is included in the species *R. laguerreae* [20]. It is evident from the phylogeny (Figure 3) that branch lengths are long in the F-clade, indicating a relatively rapid rate of evolution. The treatment of this clade is perhaps the most debatable taxonomic issue within the Rlc.

Some of the other genospecies seem much clearer candidates for the assignment of new species names. Genospecies B is notably compact, despite including strains from different studies, host plants and geographic areas. A well-known member is 3841, the first Rlc strain to have a published genome sequence [66]. Genospecies A is another clade with a clear boundary, though all the genomes are symbiovar *trifolii* so far, and the majority from a single study [21]. Genospecies D has eight genomes that are all very similar, despite diverse origins (Denmark, France and Australia), and they are well separated from anything else. There are other genospecies that also look clear-cut, but so far they have fewer genomes, so their full extent may not have been sampled.

3.6. Housekeeping gene amplicons can identify isolates to genospecies

Most of our current knowledge of rhizobial diversity comes from studies in which substantial numbers of strains were isolated from root nodules. A fairly standard approach to characterising these strains is to amplify and sequence part of the 16S rRNA gene and of a few (typically three) housekeeping genes. As discussed below, 16S sequence is not able to distinguish the genospecies of the Rlc, so we sought to establish whether individual housekeeping genes might be more informative.

We extracted, from the genomes, the sequences that would be expected using published primers for three commonly used housekeeping genes: *atpD*, *gyrB* and *recA*. The informative parts of these amplicons are 534 bp, 719 bp and 602 bp long, respectively. The concatenated sequence of all three genes was sufficient to resolve all 18 genospecies (Figure S2). Most genospecies formed a single clade, but gsO and gsP strains were scattered across the phylogeny. No exact concatenated sequence was shared between genospecies, though. For a large strain collection, it would be convenient to sequence just a single gene, so the individual phylogenies are shown in Figure 6. Any of the three genes, considered singly, would suffice to identify most of the isolates correctly, but a few sequences would be ambiguous. For example, three strains of gsH (WSM1325, WSM1328 and WSM409) and three of gsC (SM47, SM49 and SM60) share an identical *recA* allele, WSM1481 (gsJ) shares an *atpD* allele with gsQ, and the Vaf10 (gsP) *gyrB* allele is embedded within gsN. In each case, there are also some alleles in different genospecies that are so close that new isolates with related sequences could not be classified with certainty. Overall, though, nearly all isolates could be classified correctly using just one housekeeping gene; ambiguous cases could then be resolved using additional genes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

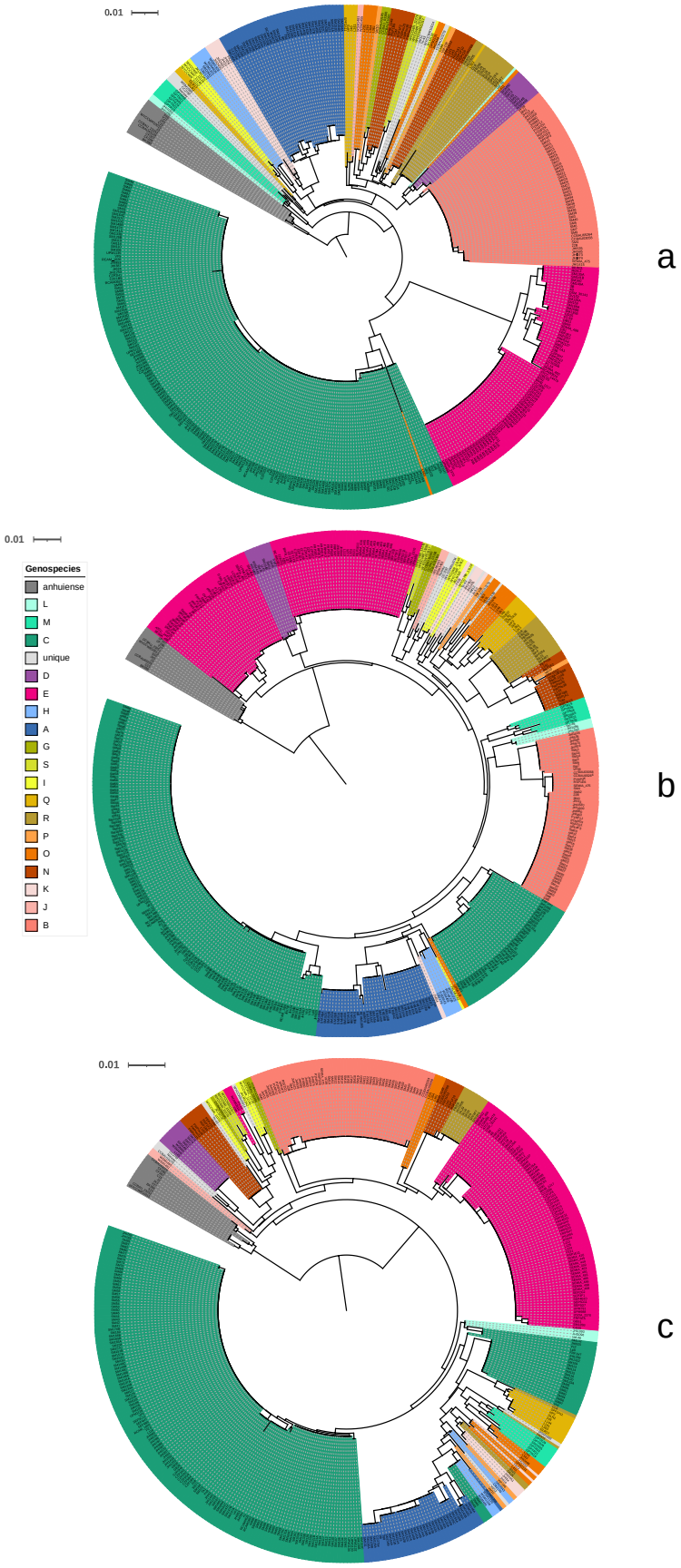

Figure 6. Phylogenies of individual housekeeping gene amplicons. a. *atpD*. b. *gyrB*. c. *recA*. Colours indicate genospecies. Scale bars indicates 1% sequence divergence. Interactive versions are available at https://itol.embl.de/shared/rhizobium.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

Two further points of interest are evident in these single-gene phylogenies (Figure 6). The first is that some genospecies are not monophyletic in the single-gene trees, but have two or more distantly related alleles. For example, gsC has several distinct clades of *recA* sequences, including the one that is shared with some gsH strains. The implication is that these genospecies have experienced introgression of new alleles from other genospecies. Occasionally, this was a recent event and the alleles have not diverged, as in the gsC-gsH *recA* case, but most of the introgression appears to be much more ancient because alleles have diversified separately within each genospecies. The second point, somewhat related, is that each gene has a unique phylogeny. Furthermore, even if all three gene sequences are concatenated, the resulting tree (Figure S2) is significantly different from the reference phylogeny based on 120 genes (Figure 3). This demonstrates that using just three genes does not provide sufficient resolution for a secure phylogenetic placement of potentially novel genospecies. For example, Youseif et al. [92] characterised a number of isolates using amplified sequences of the 16S, *glnA*, *rpoB* and *pgi* genes. Some of the isolates appear to be in or around the F-clade, but their exact identity remains uncertain.

As the technology becomes more accessible, whole genome sequencing will probably replace the sequencing of single genes, even for large samples of isolates [21,50]. Single genes will, however, still be important for high-throughput amplicon sequencing (HTAS), an approach that characterises the diversity of a microbial community by the diversity of sequences amplified from community DNA. While 16S rRNA is often used for general bacterial communities, protein-coding genes are more informative when a narrow taxonomic range is of interest. Partial sequences of *recA* and *rpoB* genes have been used to characterise the diversity of clover root-nodule samples, successfully identifying genospecies A to E [79]. We assessed the sequence diversity of these two amplicons across all 440 genomes. Because HTAS relies on high-throughput sequencing, the length of the amplicons is limited (in this case, 251 informative nucleotides for *recA*, 254 for *rpoB*). Nevertheless, most strains can be identified to the correct genospecies (Figure S3), although there are a few ambiguous sequences, particularly in the F-clade (gsN-R).

3.7. 16S rRNA sequence is not indicative of genospecies

The sequence of the small-subunit ribosomal RNA (16S rRNA) gene has been a valuable marker for bacterial taxonomy that has helped to establish the relationships among higher taxa. However, it is often too conserved to provide useful discrimination among closely-related species. In fact, the full-length 16S sequences of all five type strains within the Rlc (those of *R. leguminosarum, R. laguerreae, R. sophorae, R. ruizarguesonis and "R. indicum"*) are identical. The *R. anhuiense* sequence is also identical, and even that of the more distantly related *R. acidisoli*. It would be reasonable to expect that there would be no variation in 16S sequence within the Rlc, but this is far from the truth. This 'standard' sequence is indeed the most frequent, found in 286 of the 440 genomes (including all *R. anhuiense*), but there is variation in several parts of the 16S rRNA, as indicated in Figure 3 and detailed in Tables S9 and S10. Kumar et al. [29] reported a single nucleotide that varied in genospecies A to E (T in gsA and gsB, C or A in gsC, A in gsD, C in gsE), and this polymorphism is widespread across the genospecies of the Rlc. The fourth possible nucleotide (G) is also found. While there is some association with genospecies (e.g., all gsB strains have T, all gsM have G), every variant is found in more than one genospecies. In the 16S rRNA secondary structure, this polymorphic site is an unpaired base in a loop (position 1076 in the standard sequence). Another polymorphism that is shared by gsM and some, but not all, strains in the F-clade affects a pair of bases in a stem (positions 948 and 961) that are T and A in the standard sequence, C and G in the variant. A third variant is found in eleven strains of gsO, gsP and gsQ. This is the insertion of an intervening sequence (IVS) 78 nt in length in place of the normal 4-base loop at positions 73-76. This IVS has been described previously in a number of Rlc strains, including three of those reported here [93]. Two strains have a single-nucleotide variant within the IVS (white triangles in Figure 3). There are a number of other 16S sequence variants that are confined to one or two strains each. Excluding five single-nucleotide variants that were each found only in a single strain, and might represent sequencing errors, the combined effect of all this polymorphism is that we found 18 distinct 16S sequences within the Rlc.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

Most of the variation is, however, distributed across several genospecies and hence not useful for identifying genospecies. The sequences characterised here are those with the best match to the type-strain sequence. Since *Rhizobium* genomes normally have three copies of the ribosomal RNA operon, within-strain variation is possible in principle, but cannot be investigated thoroughly with this set of genomes because most are not fully assembled and include only a single consensus 16S sequence.

3.8. Nodulation specificity is not a useful taxonomic character

Host specificity for nodulation was abandoned long ago as a taxonomic character because of the evidence for widespread mobility of nodulation genes within and between species [42,43,45,46,94]. This is very evident from the Rlc genome sequences. Nearly all the strains were originally isolated from legume root nodules, although some were not, but even among nodule isolates there are a few genomes that have no nodulation genes, probably because of loss of the symbiosis plasmid before sequencing. In some cases, e.g. SM168B [21], the original culture was shown to nodulate and had nodulation genes detectable by PCR. A coloured circle in Figure 3 shows the symbiovar of each strain, as determined from the sequence of its *nodC* gene. Symbiovar *viciae* strains (nodulating plants in the genera *Vicia, Lathyrus, Pisum, Lens* and *Vavilovia*) and *trifolii* strains (nodulating *Trifolium*) are well mixed in most genospecies; even closely related strains may differ in symbiovar. The distribution is not completely random – all the gsA strains are *trifolii* and most F-clade strains are *viciae*, for example – but it is clear that symbiovar is not a useful character for distinguishing genospecies.

The nodulation genes of the three symbiovars, *viciae*, *trifolii* and *phaseoli*, are very distinct in sequence, but there is also polymorphism within each symbiovar. For example, the phylogeny of *nodC* (Figure S4) shows multiple alleles within each of the three symbiovar-specific clades. Symbiovar *trifolii* shows the pattern reported previously [21], in that most alleles are confined to a single genospecies, but a few are more promiscuous, showing high levels of introgression. Intriguingly, most symbiovar *viciae* alleles are found in more than one genospecies (Figure S4), suggesting that *viciae* symbiosis genes may be more mobile than their *trifolii* relatives. We note, however, that the *viciae* isolates have been sampled from a more diverse set of studies, locations and hosts than the *trifolii* isolates, which may be reflected in the observed patterns of diversity.

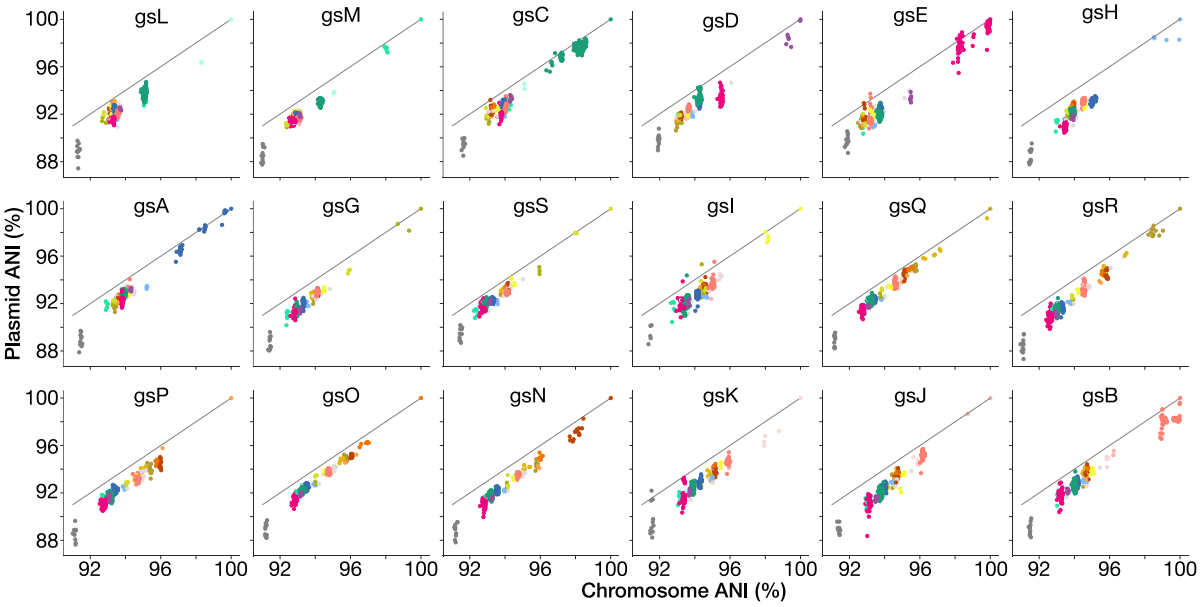

Figure 7. ANI of plasmid and chromosome genome compartments. Each point represents the values for the two compartments in a single pairwise comparison between genomes. Each panel shows values for every strain (coloured by genospecies) compared to the representative strain for the indicated genospecies.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 December 2020 doi:10.20944/preprints202012.0297.v1

3.9. Genospecies have distinct plasmid-borne sequences

Rlc genomes typically have a chromosome of around 5 Mb plus four to six plasmids totalling another 2.5-3 Mb. Thus, around a third of the genome is extrachromosomal. The larger plasmids are chromids that resemble the chromosome in G+C composition and in having a largely stable set of genes [95], while the smaller plasmids mostly carry accessory genes, i.e., genes that are present in some strains but not others, and these genes tend to have a lower G+C content [66]. The bac120 core genes are all chromosomal, and ANI will be dominated by chromosomal matches, but we wondered whether plasmid-borne genes were also characteristic of their genospecies. We sorted the scaffolds of each genome into chromosomal and nonchromosomal (i.e., plasmid) compartments and calculated ANI separately for each compartment. Figure 7 shows that the ANI of plasmid DNA is strongly correlated with that of chromosomal DNA, but that the values are usually lower and are more variable. We can conclude that the fraction of the plasmid-borne genes that is in common between two strains has a strong genospecies-specific signature, but tends to evolve in sequence faster than the chromosome. This may reflect lower selective constraints, as most essential genes are on the chromosome.

3.10. Genospecies have distinct gene complements

While variation in the sequence of shared genes may contribute to functional divergence of the genospecies, it is likely that important differences are also conferred by genes that are present in some strains but absent from others. We therefore sorted all the protein-coding genes found in strains of the Rlc into sets of orthologs, and explored the distribution of these ortholog sets across the genospecies of the Rlc. The level of gene sharing is higher between strains within a genospecies than between those in different genospecies (Figure S5), confirming a previous study that was restricted to five genospecies [21]. Gene sharing is strongly correlated with ANI both within and between genospecies (Figure 8). This demonstrates that each genospecies does, indeed, have a characteristic set of accessory genes. The fact that more closely related strains share more accessory genes could be explained by vertical inheritance of these genes from a common ancestor, but it is also plausible that the rates of horizontal gene transfer are higher between more similar strains and gene sharing is a more dynamic process. Phylogenetic analysis of the shared genes would shed light on this question, but this is beyond the scope of the present study.

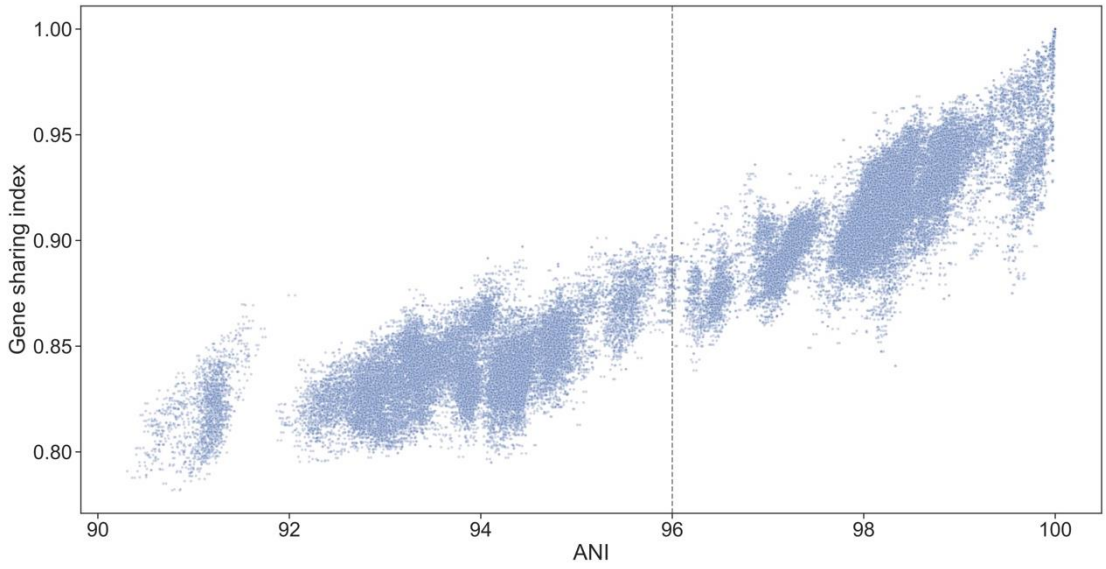

Figure 8. Sharing of accessory genes between strains compared to ANI. Gene sharing index is based on the number of orthogroups shared between two strains, normalised by total orthogroup content, for selected comparisons. See Figure S5. Dashed line at 96% ANI indicates approximate genospecies boundary.

 doi:10.20944/preprints202012.0297.v1

4. Conclusions

We have shown that the *R. leguminosarum* species complex (Rlc) forms a distinct clade and is clearly separated from other species in the genus by a long branch in the core-gene phylogeny and a gap in ANI values (Figures 1 and 2). We have also shown that the Rlc can be subdivided into 18 clusters that are sufficiently different to be considered separate genospecies, plus a further 7 single strains that might be the first representatives of additional genospecies (Figures 3, 4, 5). It is not our purpose here to propose formal species names for these genospecies, but the evidence provided here is certainly a good starting point for such proposals in the future.

Five years ago, many bacterial species descriptions were still being published without the genome sequence of the proposed type strain, but a genome sequence is now almost universally required. Often, though, taxonomists are not fully exploiting the power of genomes. The first important point is that phylogenetic inferences should be based on the sequences of a large number of core genes, such as the set of 120 genes used here. In the past, taxonomists have often relied on just a couple of kilobases of sequence from two or three housekeeping genes, but our analyses show just how unreliable this can be. It is clear that there has been substantial introgression of alleles between genospecies affecting housekeeping genes such as *recA, atpD, rpoB* and *gyrB*, such that a single genospecies may have alleles that are not closely related (Figures 6, S4). In most cases, this introgression appears to have been ancient and followed by sequence divergence, since it is rare to find exactly the same allele in different genospecies. As a result, the sequence of a single gene may identify a known genospecies unambiguously, but give a very misleading indication of the phylogenetic relationships among strains.

A second point is that insightful taxonomy cannot be based entirely on type strains. If we had confined our study to type strains, we would have had just five data points and would have learned nothing except that *R. laguerreae*, *R. sophorae* and *"R. indicum"* were more closely related to each other than to *R. leguminosarum* and *R. ruizarguesonis*. Yet, when each of those species was established, it was based on a handful of strains that were compared only with the existing type strains. The consequence is most dramatically illustrated in the case of *R. ruizarguesonis*. It was described, perfectly correctly, using four similar strains [58]. However, a search of GenBank would have revealed that there were already well over a hundred available genomes of the same species (gsC), greatly expanding the known diversity and geographic range of the species. Until recently, taxonomists have tended to consider only a small number of strains when describing a new species, limited by the practicality of obtaining strains and characterising them in the laboratory. The increasing availability of large numbers of genome sequences gives access to a much richer picture of the extent of variability within each species, potentially leading to more robust and useful species definitions.

There have been recent initiatives to harness the power of genome sequences in the service of bacterial taxonomy. Two are particularly noteworthy for their ambitious scope, covering all bacteria and archaea, and for their contrasting philosophies. The Type Strain Genome Server (TYGS) aims to identify any submitted genome by comparison to the genomes of all published type strains [30]. Strains that do not fit within the species that have been formally named are simply returned as 'potential new species'. By contrast, the developers of the Genome Taxonomy Database (GTDB) take a more inclusive approach, using a 95% ANI threshold to place all available genomes into species-level groups [20].The result is a parallel nomenclature that sometimes diverges from the formal taxonomy but is arguably a better and more consistent reflection of biological reality. Almost two-thirds of the species-level taxa in GTDB have no formal name yet. These include, of course, some of the genospecies we have identified in the Rlc. Our analyses suggest that, within the Rlc, the natural boundaries are mostly rather higher than 95% ANI, centred around 96%, so GTDB does not recognise some of the splits we have discussed. Of course, it cannot be expected that a fixed threshold will coincide with natural species boundaries across the whole of the prokaryotes. The rates of sequence evolution vary across lineages [13], as do population sizes, ecological opportunities, and so on. We do not fully understand the process of speciation in bacteria, but a stable and useful taxonomy needs to reflect the resulting pattern of clusters and gaps that is being revealed by genome sequencing. The mechanisms that create and maintain these gaps in bacterial 'genomic space' could be genetic,

ecological, or both. These ideas have been explored from various perspectives [25,96–100], and we can expect that high-throughput genome sequencing will contribute to a new level of understanding of bacterial evolution and speciation. For our purpose here, though, it is enough to note that bacterial species exist in nature, and genome sequencing enables us to discover and describe them.

Supplementary: **Rlc_supplementary_tables.xlsx**. Table S1: The 440 genomes of the Rlc and *R. anhuiense*; Table S2: Genomes assigned to *Rhizobium* downloaded 25 Jul 2020; Table S3: Genomes with missing bac120 genes; Table S4: Genomes in other genera but assigned to *Rhizobium* by NCBI; Table S5: Genomes in other *Rhizobium* species; Table S6: Duplicate Rlc genomes (not used); Table S7: TYGS results for genospecies representative genomes; Table S8: TYGS results for unique genomes; Table S9: 16S rRNA alleles; Table S10: 16S rRNA variant summary; Table S11: The bac120 protein set with Rlc orthologs. **Rlc_supplementary_figures.pdf**. Figure S1. Phylogeny of the Rlc based on sets of 60 core genes; Figure S2. Phylogeny of concatenated *atpD-gyrB-recA* amplicon sequences; Figure S3. Phylogeny of *recA* and *rpoB* based on short sequences; Figure S4. Phylogeny of *nodC;* Figure S5. Sharing of accessory genes between strains.

Author Contributions: Conceptualization, J.P.W.Y, S.M., A.A., P.R., M.I.A.C., A.A.A, B.J.P., E.T.W., E.V., E.E.A., A.T., J.D.F.F., S.H.Y., M.L., S.B., B.J., G.J.K., A.P., M.F.H., M.H.R.-B., C.F.T; Methodology, J.P.W.Y., S.M., A.A., P.R.; Software, J.P.W.Y., S.M., A.A.; Validation, J.P.W.Y., S.M., P.R., A.A.; Formal Analysis, J.P.W.Y., S.M., A.A., P.R.; Investigation, J.P.W.Y., S.M., A.A., P.R.; Resources, J.P.W.Y., M.M., E.K.J., M.I.A.C., M.H.R., B.J.P., E.T.W., E.V., E.E.A., J.D.F.F., R.R.G., M.L., S.B., B.J., A.P., C.F.T.; Data Curation, J.P.W.Y., S.M., A.A., P.R., M.M.; Writing – Original Draft Preparation, J.P.W.Y.; Writing – Review & Editing, all authors; Visualization, J.P.W.Y., S.M., A.A.; Supervision, J.P.W.Y., E.K.J., E.E.A., R.R.G., M.L., A.G., C.F.T.; Project Administration, J.P.W.Y.; Funding Acquisition, J.P.W.Y., A.A., P.R., E.K.J., E.E.A., R.R.G., M.L., B.J.P., A.G.

Funding: Funding for genome sequencing and analysis was received from Innovation Fund Denmark (4105-00007A, led by S. U. Andersen) to J.P.W.Y.; European Community FP7 'Legumes for the Agriculture of Tomorrow' (LEGATO, FP7-613551) to M.L. and J.P.W.Y.; Agence Nationale de la Recherche (GrasP) to M.L.; Rural and Environment Science and Analytical Services (RESAS, Scotland) and 'Transition paths to sustainable legume based systems in Europe' (TRUE, EC Horizon 2020, 727973 led by P. Iannetta) to M.M.; Genomia Fund to E.K.J., RSF 17-76-30016 to A.A.; RSF 19-16-00081 to E.E.A.; VA2I/463AC06 and CLU-2018-04 to R.R.G.; SERB-DST, India (YSS/2015/000149) to P.R.; a University of Otago Research Grant to B.J.P.; CSIR, India (BSC0117) and CSIR-HRDG, India (21(1023)/16/EMR-II) to A.G. The majority of the genomes used in this study were sequenced by MicrobesNG (http://www.microbesng.uk) which was supported by the BBSRC (grant number BB/L024209/1).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vandamme, P.; Pot, B.; Gillis, M.; Vos, P.D.; Kersters, K.; Swings, J. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol Rev* **1996**, *60*, 407–438.

2. Vandamme, P.; Peeters, C. Time to revisit polyphasic taxonomy. *Antonie van Leeuwenhoek* **2014**, *106*, 57–65, doi:10.1007/s10482-014-0148-x.

3. Brenner, D.J.; Fanning, G.R.; Rake, A.V.; Johnson, K.E. Batch procedure for thermal elution of DNA from hydroxyapatite. *Anal Biochem* **1969**, *28*, 447–459, doi:10.1016/0003-2697(69)90199-7.

4. Wayne, L.G.; Brenner, D.J.; Colwell, R.R.; Grimont, P.A.D.; Kandler, O.; Krichevsky, M.I.; Moore, L.H.; Moore, W.E.C.; Murray, R.G.E.; Stackebrandt, E.; et al. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int J Syst Bact* **1987**, *37*, 463–464, doi:10.1099/00207713-37-4-463.

5. Stackebrandt, E.; Goebel, B.M. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bact* **1994**, *44*, 846–849, doi:10.1099/00207713-44-4-846.

6. Mukherjee, S.; Seshadri, R.; Varghese, N.J.; Eloe-Fadrosh, E.A.; Meier-Kolthoff, J.P.; Göker, M.; Coates, R.C.; Hadjithomas, M.; Pavlopoulos, G.A.; Paez-Espino, D.; et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* **2017**, *35*, 676–683, doi:10.1038/nbt.3886.

7. Chun, J.; Oren, A.; Ventosa, A.; Christensen, H.; Arahal, D.R.; Costa, M.S. da; Rooney, A.P.; Yi, H.; Xu, X.-W.; Meyer, S.D.; et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* **2018**, *68*, 461–466, doi:10.1099/ijsem.0.002516.

8. Wu, L.; Ma, J. The Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation. *Int J Syst Evol Microbiol* **2019**, *69*, 895–898, doi:10.1099/ijsem.0.003276.

9. Chun, J.; Rainey, F.A. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *Int J Syst Evol Microbiol* **2014**, *64*, 316–324, doi:10.1099/ijs.0.054171-0.

10. Whitman, W.B. Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Syst Appl Microbiol* **2015**, *38*, 217–222, doi:10.1016/j.syapm.2015.02.003.

11. Thompson, C.C.; Amaral, G.R.; Campeão, M.; Edwards, R.A.; Polz, M.F.; Dutilh, B.E.; Ussery, D.W.; Sawabe, T.; Swings, J.; Thompson, F.L. Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Arch Microbiol* **2015**, *197*, 359–370, doi:10.1007/s00203-014-1071-2.

12. Garrity, G.M. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? *J Clin Microbiol* **2016**, *54*, 1956–1963, doi:10.1128/jcm.00200-16.

13. Parks, D.H.; Chuvochina, M.; Waite, D.W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.-A.; Hugenholtz, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **2018**, *36*, 996–1004, doi:10.1038/nbt.4229.

14. Murray, A.E.; Freudenstein, J.; Gribaldo, S.; Hatzenpichler, R.; Hugenholtz, P.; Kämpfer, P.; Konstantinidis, K.T.; Lane, C.E.; Papke, R.T.; Parks, D.H.; et al. Roadmap for naming uncultivated Archaea and Bacteria. *Nature Microbiology* **2020**, *5*, 987–994, doi:10.1038/s41564-020-0733-x.

15. Parker, C.T.; Tindall, B.J.; Garrity, G.M. International Code of Nomenclature of Prokaryotes. *Int J Syst Evol Microbiol* **2015**, *69*, S1–S111, doi:10.1099/ijsem.0.000778.

16. Konstantinidis, K.T.; Tiedje, J.M. Genomic insights that advance the species definition for prokaryotes. *P Natl Acad Sci USA* **2005**, *102*, 2567–2572, doi:10.1073/pnas.0409727102.

17. Palmer, M.; Steenkamp, E.T.; Blom, J.; Hedlund, B.P.; Venter, S.N. All ANIs are not created equal: implications for prokaryotic species boundaries and integration of ANIs into polyphasic taxonomy. *Int J Syst Evol Microbiol* **2020**, ijsem004124, doi:10.1099/ijsem.0.004124.

18. Young, J.P.W. Bacteria Are Smartphones and Mobile Genes Are Apps. *Trends Microbiol* **2016**, *24*, 931–932, doi:10.1016/j.tim.2016.09.002.

19. Klenk, H.-P.; Göker, M. En route to a genome-based classification of Archaea and Bacteria? *Syst Appl Microbiol* **2010**, *33*, 175–182, doi:10.1016/j.syapm.2010.03.003.

20. Parks, D.H.; Chuvochina, M.; Chaumeil, P.-A.; Rinke, C.; Mussig, A.J.; Hugenholtz, P. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* **2020**, *12*, e1001920-8, doi:10.1038/s41587-020-0501-8.

21. Cavassim, M.I.A.; Moeskjaer, S.; Moslemi, C.; Fields, B.; Bachmann, A.; Vilhjálmsson, B.J.; Schierup, M.H.; Young, J.P.W.; Andersen, S.U. Symbiosis genes show a unique pattern of introgression and selection within a *Rhizobium leguminosarum* species complex. *Microbial Genomics* **2020**, 89, doi:10.1099/mgen.0.000351.

22. Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.F.; Dougherty, B.A.; Merrick, J.M.; et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **1995**, *269*, 496–512, doi:10.1126/science.7542800.

23. Ravin, A.W. Experimental Approaches to the Study of Bacterial Phylogeny. *Am Nat* **1963**, *97*, 307–318, doi:10.1086/282282.

24. Mayr, E. *Systematics and the origin of species from the viewpoint of a zoologist*; Harvard University Press.: Cambridge, MA, 1942.

25. Hanage, W.P.; Spratt, B.G.; Turner, K.M.E.; Fraser, C. Modelling bacterial speciation. *Phil Trans Roy Soc B* **2006**, *361*, 2039–2044, doi:10.1098/rstb.2006.1926.

26. Fraser, C.; Hanage, W.P.; Spratt, B.G. Recombination and the Nature of Bacterial Speciation. *Science* **2007**, *315*, 476–480, doi:10.1126/science.1127573.

27. Richter, M.; Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Nat Acad Sci USA* **2009**, *106*, 19126–19131, doi:10.1073/pnas.0906412106.

28. Varghese, N.J.; Mukherjee, S.; Ivanova, N.; Konstantinidis, K.T.; Mavrommatis, K.; Kyrpides, N.C.; Pati, A. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **2015**, *43*, 6761–6771, doi:10.1093/nar/gkv657.

29. 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* **2015**, *5*, 140133–140133, doi:10.1098/rsob.140133.

30. Meier-Kolthoff, J.P.; Göker, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications* **2019**, *10*, 1–10, doi:10.1038/s41467-019-10210-3.

31. Lindström, K.; Mousavi, S.A. Effectiveness of nitrogen fixation in rhizobia. *Microb Biotechnol* **2020**, *13*, 1314–1335, doi:10.1111/1751-7915.13517.

32. Mahmud, K.; Makaju, S.; Ibrahim, R.; Missaoui, A. Current Progress in Nitrogen Fixing Plants and Microbiome Research. *Plants* **2020**, *9*, 97, doi:10.3390/plants9010097.

33. Frank, A.B. Über die Pilzsymbiose der Leguminosen. *Berichte der Deutschen Botanischen Gesellschaft* **1889**, *7*, 332–346.

34. Buchanan, R.E. What Names Should Be Used for the Organisms Producing Nodules on the Roots of Leguminous Plants? *Proceedings of the Iowa Academy of Science* **1926**, *33*, 81–90.

35. Dangeard, P.A. *Recherches sur les tubercles radicaux des Légumineuses*; Ed. du Botaniste, 1926; Vol. 16;.

36. Jarvis, B.D.W.; Pankhurst, C.E.; Patel, J.J. *Rhizobium loti*, a New Species of Legume Root Nodule Bacteria. *Int J Syst Bact* **1982**, *32*, 378–380, doi:10.1099/00207713-32-3-378.

37. Jordan, D.C. Transfer of *Rhizobium japonicum* Buchanan 1980 to *Bradyrhizobium* gen. nov., a Genus of Slow-Growing, Root Nodule Bacteria from Leguminous Plants. *Int J Syst Bact* **1982**, *32*, 136–139, doi:10.1099/00207713-32-1-136.

38. Lajudie, P. de; Willems, A.; Pot, B.; Dewettinck, D.; Maestrojuan, G.; Neyra, M.; Collins, M.D.; Dreyfus, B.; Kersters, K.; Gillis, M. Polyphasic Taxonomy of Rhizobia: Emendation of the Genus *Sinorhizobium* and Description of *Sinorhizobium meliloti* comb. nov., *Sinorhizobium saheli* sp. nov., and *Sinorhizobium teranga* sp. nov. *Int J Syst Evol Microbiol* **1994**, *44*, 715–733, doi:10.1099/00207713-44-4-715.

39. Jarvis, B.D.W.; Berkum, P.V.; Chen, W.X.; Nour, S.M.; Fernandez, M.P.; Cleyet-Marel, J.C.; Gillis, M. Transfer of *Rhizobium loti*, *Rhizobium huakuii*, *Rhizobium ciceri*, *Rhizobium mediterraneum*, and *Rhizobium tianshanense* to *Mesorhizobium* gen. nov. *Int J Syst Evol Micr* **1997**, *47*, 895–898, doi:10.1099/00207713-47-3-895.

40. Jordan, D.C. Genus I. *Rhizobium* Frank 1889, 338AL. In *Bergey's Manual of Systematic Bacteriology*; Krieg, N.R., Holt, J.G., Eds.; Williams & Wilkins: Baltimore, 1984; Vol. 1, pp. 235–242.

41. Johnston, A.W.B.; Beynon, J.L.; Buchanan-Wollaston, A.V.; Setchell, S.M.; Hirsch, P.R.; Beringer, J.E. High frequency transfer of nodulating ability between strains and species of Rhizobium. *Nature* **1978**, *276*, 634–636, doi:10.1038/276634a0.

42. Young, J.P.W. *Rhizobium* population genetics: enzyme polymorphism in isolates from peas, clover, beans and lucerne grown at the same site. *Journal of General Microbiology* **1985**, *131*, 2399–2408, doi:10.1099/00221287-131-9-2399.

43. Young, J.P.W.; Wexler, M. Sym Plasmid and Chromosomal Genotypes are Correlated in Field Populations of *Rhizobium leguminosarum*. *Journal of General Microbiology* **1988**, *134*, 2731–2739, doi:10.1099/00221287-134-10-2731.

44. Lajudie, P.M. de; Andrews, M.; Ardley, J.; Eardly, B.; Jumas-Bilak, E.; Kuzmanović, N.; Lassalle, F.; Lindström, K.; Mhamdi, R.; Martínez-Romero, E.; et al. Minimal standards for the description of new genera and species of rhizobia and agrobacteria. *Int J Syst Evol Microbiol* **2019**, *67*, 2485, doi:10.1099/ijsem.0.003426.

45. Rogel, M.A.; Ormeño-Orrillo, E.; Martínez-Romero, E. Symbiovars in rhizobia reflect bacterial adaptation to legumes. *Syst Appl Microbiol* **2011**, *34*, 96–104, doi:10.1016/j.syapm.2010.11.015.

46. Remigi, P.; Zhu, J.; Young, J.P.W.; Masson-Boivin, C. Symbiosis within symbiosis: evolving nitrogen-fixing legume symbionts. *Trends Microbiol* **2016**, *24*, 63–75, doi:10.1016/j.tim.2015.10.007.

47. Ramirez-Bahena, M.H.; Garcia-Fraile, P.; Peix, A.; Valverde, A.; Rivas, R.; Igual, J.M.; Mateos, P.F.; Martinez-Molina, E.; Velazquez, E. Revision of the taxonomic status of the species *Rhizobium leguminosarum* (Frank 1879) Frank 1889AL, *Rhizobium phaseoli* Dangeard 1926AL and *Rhizobium trifolii* Dangeard 1926AL. *R. trifolii* is a later synonym of *R. leguminosarum*. Reclassification of the strain *R. leguminosarum* DSM 30132 (=NCIMB 11478) as *Rhizobium pisi* sp. nov. *Int J Syst Evol Microbiol* **2008**, *58*, 2484–2490, doi:10.1099/ijs.0.65621-0.

48. 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. *ISME J* **2011**, *5*, 1722–1734, doi:10.1038/ismej.2011.55.

49. Goris, J.; Konstantinidis, K.T.; Klappenbach, J.A.; Coenye, T.; Vandamme, P.; Tiedje, J.M. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **2007**, *57*, 81–91, doi:10.1099/ijs.0.64483-0.

50. Boivin, S.; Lahmidi, N.A.; Sherlock, D.; Bonhomme, M.; Dijon, D.; Heulin-Gotty, K.; Le-Queré, A.; Pervent, M.; Tauzin, M.; Carlsson, G.; et al. Host-specific competitiveness to form nodules in *Rhizobium leguminosarum* symbiovar *viciae*. *New Phytol* **2020**, *226*, 555–568, doi:10.1111/nph.16392.

51. Saïdi, S.; Ramírez-Bahena, M.-H.; Santillana, N.; Zúñiga, D.; Álvarez-Martínez, E.; Peix, A.; Mhamdi, R.; Velázquez, E. *Rhizobium laguerreae* sp. nov. nodulates *Vicia faba* on several continents. *Int J Syst Evol Microbiol* **2014**, *64*, 242–247, doi:10.1099/ijs.0.052191-0.

52. Seshadri, R.; Reeve, W.G.; Ardley, J.K.; Tennessen, K.; Woyke, T.; Kyrpides, N.C.; Ivanova, N.N. Discovery of Novel Plant Interaction Determinants from the Genomes of 163 Root Nodule Bacteria. *Scientific Reports* **2015**, *5*, 16825, doi:10.1038/srep16825.

53. Boivin, S.; Mahé, F.; Pervent, M.; Tancelin, M.; Tauzin, M.; Wielbo, J.; Mazurier, S.; Young, J.P.; Lepetit, M. Genetic variation in host-specific competitiveness of the symbiont *Rhizobium leguminosarum* symbiovar *viciae*. *Authorea* **2020**, doi:10.22541/au.159237007.72934061.

54. Afonin, A.; Sulima, A.; Zhernakov, A.; Zhukov, V. Draft genome of the strain RCAM1026 *Rhizobium leguminosarum* bv. *viciae*. *Genomics Data* **2017**, *11*, 85–86, doi:10.1016/j.gdata.2016.12.003.

55. Sánchez-Cañizares, C.; Jorrin, B.; Durán, D.; Nadendla, S.; Albareda, M.; Rubio-Sanz, L.; Lanza, M.; González-Guerrero, M.; Prieto, R.; Brito, B.; et al. Genomic Diversity in the Endosymbiotic Bacterium *Rhizobium leguminosarum*. *Genes* **2018**, *9*, 60, doi:10.3390/genes9020060.

56. Liang, J.; Hoffrichter, A.; Brachmann, A.; Marín, M. Complete genome of *Rhizobium leguminosarum* Norway, an ineffective *Lotus* micro-symbiont. *Standards in Genomic Sciences* **2018**, *13*, 1–11, doi:10.1186/s40793-018-0336-9.

57. Chirak, E.R.; Kimeklis, A.K.; Karasev, E.S.; Kopat, V.V.; Safronova, V.I.; Belimov, A.A.; Aksenova, T.S.; Kabilov, M.R.; Provorov, N.A.; Andronov, E.E. Search for Ancestral Features in Genomes of *Rhizobium leguminosarum* bv. *viciae* Strains Isolated from the Relict Legume *Vavilovia formosa*. *Genes* **2019**, *10*, 990, doi:10.3390/genes10120990.

58. Jorrin, B.; Palacios, J.M.; Peix, Á.; Imperial, J. *Rhizobium ruizarguesonis* sp. nov., isolated from nodules of *Pisum sativum* L. *Syst Appl Microbiol* **2020**, *43*, 126090, doi:10.1016/j.syapm.2020.126090.

59. Rahi, P.; Giram, P.; Chaudhari, D.; diCenzo, G.C.; Kiran, S.; Khullar, A.; Chandel, M.; Gawari, S.; Mohan, A.; Chavan, S.; et al. Rhizobium indicum sp. nov., isolated from root nodules of pea (Pisum sativum) cultivated in the Indian trans-Himalayas. *Syst Appl Microbiol* **2020**, *43*, 126127, doi:10.1016/j.syapm.2020.126127.

60. Ayuso-Calles, M.; García-Estévez, I.; Jiménez-Gómez, A.; Flores-Félix, J.D.; Escribano-Bailón, M.T.; Rivas, R. *Rhizobium laguerreae* Improves Productivity and Phenolic Compound Content of Lettuce (*Lactuca sativa* L.) under Saline Stress Conditions. *Foods* **2020**, *9*, 1166, doi:10.3390/foods9091166.

61. Afonin, A.M.; Gribchenko, E.S.; Sulima, A.S.; Zhukov, V.A.; Newton, I.L.G. Complete Genome Sequence of an Efficient *Rhizobium leguminosarum* bv. *viciae* Strain, A1. *Microbiology Resource Announcements* **2020**, *9*, 143, doi:10.1128/mra.00249-20.

62. Perry, B.J.; Ferguson, S.; Laugraud, A.; Wakelin, S.A.; Reeve, W.; Ronson, C.W. Complete Genome Sequences of *Trifolium* spp. Inoculant Strains *Rhizobium leguminosarum* sv. *trifolii* TA1 and CC275e: Resources for Genomic Study of the *Rhizobium* - *Trifolium* Symbiosis. *Mol Plant-Microbe Interactions* **2020**, MPMI-08-20-0220, doi:10.1094/mpmi-08-20-0220-a.

63. Parks, D.H.; Rinke, C.; Chuvochina, M.; Chaumeil, P.-A.; Woodcroft, B.J.; Evans, P.N.; Hugenholtz, P.; Tyson, G.W. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2017**, *2*, 1533–1542, doi:10.1038/s41564-017-0012-7.

64. Reeve, W.; O'Hara, G.; Chain, P.; Ardley, J.; Bräu, L.; Nandesena, K.; Tiwari, R.; Copeland, A.; Nolan, M.; Han, C.; et al. Complete genome sequence of *Rhizobium leguminosarum* bv. *trifolii* strain WSM1325, an effective microsymbiont of annual Mediterranean clovers. *Stand Genomic Sci* **2010**, *2*, 347–356, doi:10.4056/sigs.852027.

65. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: architecture and applications. *BMC Bioinformatics* **2009**, *10*, 421, doi:10.1186/1471-2105-10-421.

66. 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* **2006**, *7*, R34, doi:10.1186/gb-2006-7-4-r34.

67. Sievers, F.; Higgins, D.G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **2018**, *27*, 135–145, doi:10.1002/pro.3290.

68. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One* **2010**, *5*, e9490, doi:10.1371/journal.pone.0009490.

69. Huson, D.H.; Scornavacca, C. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biol* **2012**, *61*, 1061–1067, doi:10.1093/sysbio/sys062.

70. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **2019**, *47*, gkz239-, doi:10.1093/nar/gkz239.

71. Jain, C.; Rodriguez-R, L.M.; Phillippy, A.M.; Konstantinidis, K.T.; Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **2018**, *9*, 5114, doi:10.1038/s41467-018-07641-9.

72. Gaunt, M.W.; Turner, S.L.; Rigottier-Gois, L.; Lloyd-Macgilp, S.A.; Young, J.P.W. Phylogenies of *atpD* and *recA* support the small subunit rRNA-based classification of rhizobia. *Int J Syst Evol Microbiol* **2001**, *51*, 2037–2048, doi:10.1099/00207713-51-6-2037.

73. Martens, M.; Dawyndt, P.; Coopman, R.; Gillis, M.; Vos, P.D.; Willems, A. Advantages of multilocus sequence analysis for taxonomic studies: a case study using 10 housekeeping genes in the genus *Ensifer* (including former *Sinorhizobium*). *Int J Syst Evol Microbiol* **2008**, *58*, 200–214, doi:10.1099/ijs.0.65392-0.

74. Marek-Kozaczuk, M.; Leszcz, A.; Wielbo, J.; Wdowiak-Wróbel, S.; Skorupska, A. *Rhizobium pisi* sv. *trifolii* K3.22 harboring nod genes of the *Rhizobium leguminosarum* sv. *trifolii* cluster. *Syst Appl Microbiol* **2013**, *36*, 252–258, doi:10.1016/j.syapm.2013.01.005.

75. Efstathiadou, E.; Savvas, D.; Tampakaki, A.P. Genetic diversity and phylogeny of indigenous rhizobia nodulating faba bean (*Vicia faba* L.) in Greece. *Syst Appl Microbiol* **2020**, *43*, 126149, doi:10.1016/j.syapm.2020.126149.

76. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **1994**, *22*, 4673–4680, doi:10.1093/nar/22.22.4673.

77. Saitou, N.; Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **1987**, *4*, 406–25, doi:10.1093/oxfordjournals.molbev.a040454.

78. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **2018**, *35*, 1547–1549, doi:10.1093/molbev/msy096.

79. Fields, B.; Moeskjær, S.; Friman, V.; Andersen, S.U.; Young, J.P.W. MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction. *Mol Ecol Resour* **2020**, doi:10.1111/1755-0998.13294.

80. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **2010**, *11*, 119, doi:10.1186/1471-2105-11-119.

81. Emms, D.M.; Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **2019**, *20*, 1–14, doi:10.1186/s13059-019-1832-y.

82. Lajudie, P. de; Laurent-Fulele, E.; Willems, A.; Torek, U.; Coopman, R.; Collins, M.D.; Kersters, K.; Dreyfus, B.; Gillis, M. *Allorhizobium undicola* gen. nov., sp. nov., nitrogen-fixing bacteria that efficiently nodulate *Neptunia natans* in Senegal. *Int J Syst Evol Microbiol* **1998**, *48*, 1277–1290, doi:10.1099/00207713-48-4-1277.

83. Mousavi, S.A.; Willems, A.; Nesme, X.; Lajudie, P. de; Lindström, K. Revised phylogeny of *Rhizobiaceae*: Proposal of the delineation of *Pararhizobium* gen. nov., and 13 new species combinations. *Syst Appl Microbiol* **2015**, *38*, 84–90, doi:10.1016/j.syapm.2014.12.003.

84. Mousavi, S.A.; Österman, J.; Wahlberg, N.; Nesme, X.; Lavire, C.; Vial, L.; Paulin, L.; Lajudie, P. de; Lindström, K. Phylogeny of the *Rhizobium–Allorhizobium–Agrobacterium* clade supports the delineation of *Neorhizobium* gen. nov. *Syst Appl Microbiol* **2014**, 1–8, doi:10.1016/j.syapm.2013.12.007.

85. Kimes, N.E.; López-Pérez, M.; Flores-Félix, J.D.; Ramírez-Bahena, M.-H.; Igual, J.M.; Peix, A.; Rodriguez-Valera, F.; Velázquez, E. *Pseudorhizobium pelagicum* gen. nov., sp. nov. isolated from a pelagic Mediterranean zone. *Syst Appl Microbiol* **2015**, *38*, 293–299, doi:10.1016/j.syapm.2015.05.003.

86. Lassalle, F.; Dastgheib, S.M.M.; Zhao, F.-J.; Zhang, J.; Verbarg, S.; Frühling, A.; Brinkmann, H.; Osborne, T.H.; Sikorski, J.; Balloux, F.; et al. Phylogenomic analysis reveals the basis of adaptation of *Pseudorhizobium* species to extreme environments. *bioRxiv* **2019**, *54*, 690347, doi:10.1101/690347.

87. Ramírez-Bahena, M.H.; Vial, L.; Lassalle, F.; Diel, B.; Chapulliot, D.; Daubin, V.; Nesme, X.; Muller, D. Single acquisition of protelomerase gave rise to speciation of a large and diverse clade within the *Agrobacterium/Rhizobium* supercluster characterized by the presence of a linear chromid. *Mol Phylogenet Evol* **2014**, *73*, 202–207, doi:10.1016/j.ympev.2014.01.005.

88. Hördt, A.; López, M.G.; Meier-Kolthoff, J.P.; Schleuning, M.; Weinhold, L.-M.; Tindall, B.J.; Gronow, S.; Kyrpides, N.C.; Woyke, T.; Göker, M. Analysis of 1,000+ Type-Strain Genomes Substantially Improves Taxonomic Classification of *Alphaproteobacteria*. *Front Microbiol* **2020**, *11*, 3156, doi:10.3389/fmicb.2020.00468.

89. Ciufo, S.; Kannan, S.; Sharma, S.; Badretdin, A.; Clark, K.; Turner, S.; Brover, S.; Schoch, C.L.; Kimchi, A.; DiCuccio, M. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol* **2018**, *68*, 2386–2392, doi:10.1099/ijsem.0.002809.

90. Hang, P.; Zhang, L.; Zhou, X.-Y.; Hu, Q.; Jiang, J.-D. *Rhizobium album* sp. nov., isolated from a propanil-contaminated soil. *Antonie Van Leeuwenhoek* **2019**, *112*, 319–327, doi:10.1007/s10482-018-1160-3.

91. Jiao, Y.S.; Yan, H.; Ji, Z.J.; Liu, Y.H.; Sui, X.H.; Wang, E.T.; Guo, B.L.; Chen, W.X.; Chen, W.F. *Rhizobium sophorae* sp. nov. and *Rhizobium sophoriradicis* sp. nov., nitrogen-fixing rhizobial symbionts of the medicinal legume *Sophora flavescens*. *Int J Syst Evol Microbiol* **2015**, *65*, 497–503, doi:10.1099/ijs.0.068916-0.

92. Youseif, S.H.; El-Megeed, F.H.A.; Mohamed, A.H.; Ageez, A.; Veliz, E.; Martínez-Romero, E. Diverse *Rhizobium* strains isolated from root nodules of *Trifolium alexandrinum* in Egypt and symbiovars. *Syst Appl Microbiol* **2020**, 126156, doi:10.1016/j.syapm.2020.126156.

93. Flores-Félix, J.D.; Sánchez-Juanes, F.; García-Fraile, P.; Valverde, A.; Mateos, P.F.; Gónzalez-Buitrago, J.M.; Velázquez, E.; Rivas, R. *Phaseolus vulgaris* is nodulated by the symbiovar viciae of several genospecies of *Rhizobium laguerreae* complex in a Spanish region where *Lens culinaris* is the traditionally cultivated legume. *Syst Appl Microbiol* **2019**, *42*, 240–247, doi:10.1016/j.syapm.2018.10.009.

94. Mutch, L.A.; Young, J.P.W. Diversity and specificity of *Rhizobium leguminosarum* biovar *viciae* on wild and cultivated legumes. *Molecular Ecology* **2004**, *13*, 2435–2444, doi:10.1111/j.1365-294x.2004.02259.x.

95. Harrison, P.W.; Lower, R.P.J.; Kim, N.K.D.; Young, J.P.W. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol* **2010**, *18*, 141–148, doi:10.1016/j.tim.2009.12.010.

96. Lawrence, J.G. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr Opin Microbiol* **1999**, *2*, 519–523, doi:10.1016/s1369-5274(99)00010-7.

97. Gevers, D.; Cohan, F.M.; Lawrence, J.G.; Spratt, B.G.; Coenye, T.; Feil, E.J.; Stackebrandt, E.; Peer, Y.V. de; Vandamme, P.; Thompson, F.L.; et al. Re-evaluating prokaryotic species. *Nat Rev Microbiol* **2005**, *3*, 733–739, doi:10.1038/nrmicro1236.

98. Barraclough, T.G.; Balbi, K.J.; Ellis, R.J. Evolving Concepts of Bacterial Species. *Evolutionary Biology* **2012**, *39*, 148–157, doi:10.1007/s11692-012-9181-8.

99. Shapiro, B.J.; Polz, M.F. Microbial Speciation. *CSH Perspect Biol* **2015**, *7*, a018143, doi:10.1101/cshperspect.a018143.

100. Arevalo, P.; VanInsberghe, D.; Elsherbini, J.; Gore, J.; Polz, M.F. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. *Cell* **2019**, *178*, 820-834.e14, doi:10.1016/j.cell.2019.06.033.