Phylogenomic analysis reveals the basis of adaptation of *Pseudorhizobium* species to extreme environments

Florent Lassalle[*1], Seyed M. M. Dastgheib[2], Fang-Jie Zhao[3], Jun Zhang[3], Susanne Verbarg[4], Anja Frühling[4], Henner Brinkmann[4], Thomas H. Osborne[5#], Johannes Sikorski[4], Francois Balloux[6], Xavier Didelot[1,7], Joanne M. Santini[*5], Jörn Petersen[4].

1. Department of Infectious Disease Epidemiology, Imperial College London, UK; MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, UK.

2. Research Institute of Petroleum Industry, Tehran, Iran.

3. State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing, China.

4. Leibniz Institut DSMZ, Braunschweig, Germany.

5. Institute for Structural & Molecular Biology, University College London, London, UK.

6. UCL Genetics Institute, University College London, London, UK.

7. School of Life Sciences, University of Warwick, Coventry, UK.

#: current address: University of Bedfordshire, Luton, UK.

[*] corresponding authors. Contact: f.lassalle@imperial.ac.uk; j.santini@ucl.ac.uk.

Abstract (250 words)

The family Rhizobiaceae includes many genera of soil bacteria, often isolated for their association with plants. Herein, we investigate the genomic diversity of a group of *Rhizobium* species and unclassified strains isolated from atypical environments, including seawater, rock matrix or polluted soil. Based on whole-genome similarity and core genome phylogeny, we show that they belong to the genus *Pseudorhizobium*. We thus reclassify *Rhizobium halotolerans, R. marinum, R. flavum* and *R. endolithicum* as *P. halotolerans* comb. nov., *P. marinum* comb. nov., *P. flavum* comb. nov. and *R. endolithicum* comb. nov., respectively, and show that *P. pelagicum* is a synonym of *P. marinum*. We also delineate a new chemolithoautotroph species, *P. banfieldii* sp. nov., whose type strain is NT-26^T (= DSM 106348^T = CFBP 8663^T). This genome-based classification was independently supported by a chemotaxonomic comparison, with gradual taxonomic resolution provided by fatty acid, protein and metabolic profiles. In addition, we used a phylogenetic approach to infer scenarios of duplication, horizontal transfer and loss for all genes in the *Pseudorhizobium* pangenome. We thus identify the key functions associated with the diversification of each species and higher clades, shedding light on the mechanisms of adaptation to their respective ecological niches. Respiratory proteins acquired at the origin of *Pseudorhizobium* are combined with clade-specific genes to encode different strategies for detoxification and nutrition in harsh, nutrient-poor environments. Finally, we predict diagnostic phenotypes for the distinction of *P. banfieldii* from other *Pseudorhizobium* species, including autotrophy and sensitivity to the azo dye Congo Red, which we experimentally validated.

Keywords: *Rhizobium* sp. NT-26; genome taxonomy; clade-specific genes; ecological specialisation; phylogenomics; pangenome analysis

Introduction

Bacteria of the family Rhizobiaceae (Alphaproteobacteria) are usually soil-borne and found in association with plant roots, where they mostly rely on a saprophytic lifestyle degrading soil organic compounds and plant exudates, including aromatic compounds [8,9,47]. This particular versatility in using various organic compounds likely stems from the presence of some of the largest known sets of carbohydrate transporter genes in Rhizobiaceae genomes [38,72]. Some members of this taxon sometimes engage in a symbiotic or pathogenic relationship with a specific host plant, with the ability to switch to these lifestyles being determined by the presence of accessory megaplasmids in the bacterium [10,21,72]. It is more unusual to isolate them from an arsenic-containing rock in a sub-surface environment mostly devoid of organic matter such as the Granites goldmine in the Northern Territory, Australia [50]. Rock samples containing arsenopyrite (AsFeS) were used to enrich for and isolate organisms (designated NT-25 and NT-26) capable of using arsenite (oxidation state +3, i.e. As(III)) as the electron donor coupling its oxidation to arsenate (As(V)) with oxygen and using carbon dioxide as the sole carbon source [50]. 16S rRNA gene sequence analysis revealed that these strains were very closely related and likely belonging to the same species in the family *Rhizobiaceae*; they were provisionally named *Rhizobium* sp. [50]. However, recent advances in multi-locus sequence analysis (MLSA) and genome sequencing led to the recognition of the polyphyly of the genus *Rhizobium*. Subsequently, the taxonomy of *Rhizobiaceae* was largely revised, with many *Rhizobium* species being reclassified in newly created genera, including *Neorhizobium*, *Pararhizobium*, *Allorhizobium* and *Pseudorhizobium* [26,40–42], suggesting that the taxonomic status of NT-25 and NT-26 should be re-examined.

The strains NT-25 and NT-26 can withstand very high levels of arsenic (greater than 20 mM arsenite and 0.5 M arsenate), thanks to functions encoded notably by the arsenic-resistance (*ars*) genes [48]. In addition, they can gain energy from the oxidation of arsenite, with the mechanism in NT-26 studied in detail [5,6,16,48,49,51,69]. Analysis of the NT-26 genome sequence revealed other key genes involved in the resistance to arsenic, including the *phn*/*pst* genes, which encode phosphate-specific transporters with high affinity for phosphate but not for its structural analogue arsenate. This likely explains the high tolerance of NT-26 to arsenate [2]. The key determinants of arsenic metabolism and resistance, including the *aio*, *ars* and *phn/pst* genes, were primarily located on an accessory 322-kb megaplasmid distantly related to symbiotic plasmids of rhizobia. In addition, comparative genomics showed that the *aio* operon formed a stable genetic unit that can be found sporadically in a diverse set of bacteria – the closest relatives to these accessory genes residing in genomes of other members of the Rhizobiaceae [2,3]. The presence of the sulphur oxidase genes (*soxXYZABCD*) on the NT-26 chrosomome indicated that sulphur compounds such as sulphite and thiosulphate can also serve as electron donors for chemolithoautotrophic growth, an activity that was experimentally confirmed and found to be down-regulated when the organism was grown in the presence of arsenite [2].

Having identified the genetic features allowing NT-25 and NT-26 to live in harsh environments, we sought to investigate if this combination of adaptive determinants were only present in these ecologically specialized strains, or if they were the trademark of a wider taxon. We thus searched organisms closely related to NT-25/NT-26 based on their 16S rRNA sequence. Among them, several were isolated from polluted environments: *R.* sp. strain Khangiran2 from a soil contaminated with

petroleum, *R.* sp. strain Q54 from an arsenic-contaminated paddy soil; *R. flavum* strain YW14 from organophosphorus insecticide-contaminated soil [22] and *R. halotolerans* AB21 from soil contaminated with the detergent chloroethylene [18]. These unusual environments suggest an ability of strains within this group to cope with chemically harsh environments. Interestingly, several pathways mediating resistance to these toxins or relating to their metabolism rely on the cellular respiration machinery, including oxidative degradation of noxious organic compounds, or the oxidation of arsenite. The closest relative of strains NT-25 and NT-26, strain TCK [17], was not isolated from a toxic environment, but selected for its ability to oxidize sulphur compounds, including hydrogen sulphide, sulphite and thiosulphate, a phenotype shared by NT-25 and NT-26. In a wider phylogenetic perspective, the more distantly related species *R. endolithicum* has the peculiar ability to live within a rock matrix [44], whereas the even more distant relatives *Pseudorhizobium pelagicum* and *R. marinum* live in open sea waters, a quite unusual feature within the Rhizobiaceae [26,37]. Again, these rhizobial species display the rare capacity to live in a habitat depleted in organic nutrients.

We used a combination of long and short read sequencing technology to sequence the genomes of all known bacterial isolates closely related to *Rhizobium* sp. NT-26, resulting in eight complete or almost complete genome sequences. To place these bacterial strains into a broader taxonomic context, we complemented our dataset with all currently available complete or near-complete genome data for the bacterial families Rhizobiaceae and (closest relative) Aurantimonadaceae. We used this extensive dataset to compute a robust phylogenomic tree, which led us to the delineation of a new species, *Pseudorhizobium banfieldii* sp. nov., and the reclassification of four species into the genus *Pseudorhizobium*. Using a phylogenetic framework, we analysed the distribution and history of all pangenome genes in this genus, and revealed key genetic innovations along its diversification history. Crucially, a large repertoire of respiratory chain proteins was acquired by the ancestor of *Pseudorhizobium* and later expanded in descendant lineages. The diversification of *Pseudorhizobium* species was then marked by their respective acquisition of unique metabolic pathways, providing each species with some specific detoxification mechanisms. Finally, we predicted and experimentally validated new phenotypic traits that characterize and distinguish the studied species, providing at least one new diagnostic phenotype (sensitivity of *P. banfieldii* to the azo dye Congo Red).

Methods

DNA extraction and genome sequencing

Two independent projects were conducted at University College London (UCL; London, UK) in collaboration with the Earlham Institute (Norwich, UK) and at the Leibniz Institute-DSMZ (Braunschweig, Germany) for the genome sequencing of strains *Rhizobium* sp. NT-25 (= DSM 106347) [50], *R. flavum* YW14^T (= DSM 102134^T = CCTCC AB2013042^T = KACC 17222^T) [22], *R.* sp. Q54 (= DSM 106353), *R.* sp. TCK (= DSM 13828) [17] and Khangiran2 (= DSM 106339 = IBRC-M 11174). In addition, strains *R. halotolerans* AB21^T (= DSM 105041^T = KEMC 224-056^T = JCM 17536^T) [18], *R. endolithicum* JC140^T (= DSM 104972^T = KCTC32077^T = CCUG64352^T = MTCC11723^T = HAMBI 2447^T) [44] and *R.* sp. P007 [20] were sequenced only at the DSMZ.

For long-read sequencing, cells were cultured at UCL until stationary phase in a minimal salts medium (MSM) containing 0.08% yeast extract (YE) at 28°C [50]. Genomic DNA was extracted using the Wizard DNA Purification kit (Promega, Madison, Wisconsin) according to the manufacturer's instructions. Quality of the genomic DNA was assessed as described in the Supplementary Methods. Genomic DNA was sent to the Earlham Institute for sequencing, where quantity and quality of the DNA was determined using a Qubit fluorometer (Invitrogen, Waltham, Massachusetts) and fragment length was controlled with Tapestation 2200 (Agilent, Santa Clara, California). DNA libraries were prepared for sequencing on the PacBio RSII platform using C4-P6 chemistry with one SMRT cell per genome. This generated 82–174 $x10^3$ long reads per genome (mean: 139 $x10^3$), representing 0.332–1.26 $x10^9$ bp (mean: 0.982 $x10^9$).

Genome assembly

Illumina short-read sequencing and short read-only genome assembly was conducted at the DSMZ, as previously described [71]. Hybrid assembly of short and long reads was performed using the Unicycler software (version 0.4.2, bold mode) [70], relying on the programs SPAdes [4] for prior short read assembly, miniasm [35] and Racon [66] for prior long-read assembly and Pilon [67] for polishing of the consensus sequence.

Unless specified otherwise, the following bioinformatic analyses were conducted using the Pantagruel pipeline under the default settings as described previously [31] and on the program webpage http://github.com/flass/pantagruel/. This pipeline is designed for the analysis of bacterial pangenomes, including the inference of a species tree, gene trees, and the detection of horizontal gene transfers through species tree/gene tree reconciliations [61]. A more detailed description of the bioinformatic analyses is given in the Supplementary Methods.

Genome annotation

We used Prokka [53] as part of *Pantagruel* (task 0) to annotate the new genomes sequences, using a reference database of annotated proteins derived from complete proteomes predicted from all available complete genomes from the *Rhizobium*/*Agrobacterium* group. These were downloaded from the NCBI RefSeq Assembly database on the 15 Dec 2017 using the query 'txid227290[Organism:exp] AND ("latest refseq"[filter] AND ("scaffold level"[filter] OR "chromosome level"[filter] OR "complete genome"[filter]) AND all[filter] NOT anomalous[filter])'.

Genomic dataset and gene family classification

We assembled a complete bacterial genome dataset covering all known representative of the subgroup in the alphaproteobacterial families *Rhizobiaceae* and (sister group) *Aurantimonadaceae*. This dataset comprises all 564 genomes available from the NCBI RefSeq Assembly database on the 23 Apr 2018, filtering anomalous genomes and those with a contig N50 < 98kb using query: '(txid82115[Organism:exp] OR txid255475[Organism:exp]) AND ("latest refseq"[filter] NOT anomalous[filter]) AND ("98000"[ContigN50] : "20000000"[ContigN50])'. From this dataset we removed assembly GCF_002000045.1 (*R. flavum* YW14) for which we had a newer, higher quality assembly. To these we added the new genome sequences from the eight strains mentioned above, for a total of 571 genomes (dataset '571Rhizob').

Reference species trees

From the 571Rhizob genome dataset, we define the pseudo-core genome as genes occurring only in a single copy and present in at least 561 out of the 571 genomes (98%). The resulting pseudo-core genome gene set (thereafter referred as pCG_{571}) includes 155 loci, which protein alignments were concatenated. This concatenated protein alignment was used to compute a reference species tree (S_{ML571}) with RAxML [56] under the model PROTCATLGX; branch supports were estimated by generating 200 rapid bootstraps under the same parameters. From the S_{ML571} tree, we identified the well-supported clade grouping 41 genomes including all representative of *Neorhizobium* spp. and *Pseudorhizobium* spp. and our new isolates (dataset '41NeoPseudo'). To gain further phylogenetic resolution in this clade of interest, we restricted the pCG_{571} concatenated alignment to the 41 genomes of this smaller genomic dataset, which we used as input to the Phylobayes program for a more accurate (but computationally more expensive) Bayesian phylogenetic inference under the CAT-GTR+G4 model [30]. This provided us with a robust non-ultrametric tree for the 41 genomes (S_{BA41}). We finally used this S_{BA41} tree as a fixed input topology for Phylobayes to infer an ultrametric tree (unitless 'time' tree) under the CIR clock model [34], further referred to as T_{BA41}.

Gene trees, reconciliations and orthologous group classification

Gene trees were computed for each of the 6,714 homologous gene family of the 41-species pangenome with at least four sequences using MrBayes [45] under the GTR+4G+I model, running the Metropolis-coupled Markov chain Monte-Carlo (MCMCMC) for 2,000,000 generations, sampling a tree every 500 generations. Convergence of bipartition distribution in independent pairs of 4,000 sampled gene tree sets ('chains') was achieved for all gene families under these conditions. We combined the pairs of gene tree samples (resulting in 8,000 trees/gene family) and discarded the first 25% trees as burn-in; we used the remainder as input for the ALEml program [60,62], to reconcile these gene trees with the dated tree T_{BA41} and estimate evolutionary scenarios for each gene family, featuring events of gene duplication, transfer and loss (DTL).

Based on the estimated gene family evolutionary scenarios, we could define gene groups based on a true criterion of orthology, i.e. common descent from an ancestor by means of speciation only [19], rather than a proxy criterion such as bidirectional best hits (BBH) in a similarity search. This has the advantage of explicitly detecting the gain of an orthologous group (OG) in a genome lineage by means of horizontal gene transfer (HGT) or gene duplication. To differentiate additive HGT and replacing HGT events (i.e. gene conversion by homologous recombination events), we used a heuristic based on the unicopy criterion [7] as described in a previous study [33], where transfer events that do not increase the gene copy number in a genome are not interpreted as the emergence of a new OG, i.e. considering homologous recombination can occur within an OG without breaking it. We then used this classification to build a matrix of OG presence/absence in the '41NeoPseudo' dataset, and computed the clades-specific core genome gene set for each clade of the species tree S_{BA41}, under the following criteria: 1) every OG present in all members of the focal clade and absent in all members of its sister clade, and 2) for chosen clades matching *Pseudorhizobium* constituent species, every OG present in all members of the species, and absent in the rest of the *Pseudorhizobium* genus.

Hierarchical clustering was performed based on the OG presence/absence matrix using the pvclust function form the pvclust R package version 2.0-0 [58] with default settings, to obtain bootstrap-derived p-values (BP) and approximately unbiased (AU) branch support estimates.

Gene Ontology enrichment tests

GO terms frequency of annotation (as derived from the automatic InterProScan annotation) were compared between test ('study') gene sets and respective reference ('population') gene sets, where the former is included in the latter. Gene set pairs that were compared comprised: 1) clade-specific core genome of a clade (a single representative sequence per OG) vs. whole core genome of this clade (a single representative sequence per OG); and 2) clade-specific core genome of a clade (all sequences) vs. pangenome of this clade (all sequences). All tests were performed using the topGO R package version 2.22.0 [1], using the 'weight01' algorithm and 'Fisher' statistic.

Marker gene phylogenies

CDS alignments and gene trees were extracted from the '41NeoPseudo' Pantagruel database described below for the following genes *atpD* (gene family id: RHIZOC034665, OG id: 1), *glnII* (RHIZOC002590, 0), *gyrB* (RHIZOC002739, 0), *recA* (RHIZOC040013, 0) and *rpoB* (RHIZOC004093, 0). Note that *R. marinum* MGL06^{T} sequence for *glnII* gene was missing in gene family RHIZOC002590, probably due to incomplete genome assembly. CDS alignments were concatenated and a tree was computed with MrBayes as described above.

Overall genome relatedness measurement

We used the GGDC tool (version 2.1) for digital DNA-DNA hybridization (dDDH) to compare the genomes of the closest relatives of the NT-25/26 clone, using the formula d_4 (BLASTN identities / HSP length) [39].

Biochemical tests

A range of phenotypic assays were performed on a set of ten strains, including the five newly PacBio-sequenced strains as well as the relevant type strains. Salt tolerance: growth of strains was assayed at 28°C in liquid R2A medium (DSMZ medium 830) supplemented with increasing concentrations of NaCl (0–9% range was tested) to determine their minimum inhibitory NaCl concentration (MIC_{NaCl}). Congo Red assay: strains were plated on yeast extract – mannitol agar medium (YEM) [57] with 0.1g/L Congo Red dye for seven days. The commercial biochemical identification system for Gram-negative bacteria 'Api 20 NE' (BioMérieux) was used for an initial analysis of the biochemical capacities. High-throughput phenotyping was conducted using the GenIII microplates (Biolog, Inc., Hayward, California) for testing for growth with 94 single carbon or nitrogen nutrient sources or with inhibitors (antibiotics, salt, etc.). The GenIII phenotype data were analysed using the 'opm' R package [65].

MALDI-TOF typing

Sample preparation for MALDI-TOF mass spectrometry was carried out according to Protocol 3 in [52]. Instrumental conditions for the measurement were used as described by [64]. The dendrogram was created by using the MALDI Biotyper Compass Explorer software (Bruker, Version 4.1.90).

Fatty acid profiles

Fatty acid methyl esters were obtained as previously described [25] and separated by using a gas chromatograph (model 6890 N; Agilent Technologies). Peaks were automatically computed and assigned using the Microbial Identification software package (MIDI), TSBA40 method, Sherlock version 6.1. The dendrogram was created with Sherlock Version 6.1.

Polar lipids and respiratory lipoquinones

Polar lipids and respiratory lipoquinones were extracted from 100 mg freeze-dried cells and separated by two-dimensional silica gel thin layer chromatography by the identification service of the DSMZ as previously described [23].

Phenotype association testing

The association of accessory gene occurrence with phenotypic profiles obtained with the Biolog GenIII (continuous values) was tested using the phylogenetic framework implemented in the 'treeWAS' R package [14]. The test was performed on 9,621 non-trivial OG presence/absence profiles using the ML method for ancestral state reconstruction. Significance was assessed by comparing test scores to those obtained by simulating random genotype profiles along the species tree (10-taxon tree derived from S_{BA41}) for 100 times more loci than the tested dataset (ca. 10^5 simulations) and scaling the *p*-values using FDR correction. Phenotypes with significant hits were then tested again with more stringent parameters, using 1,000 times more loci than the tested dataset and Bonferroni *p*-value correction.

Results

Genome sequencing of eight new *Rhizobiaceae* genomes

We determined the genome sequences of strains *Rhizobium* sp. NT-25, *R. flavum* YW14^T, *R.* sp. Q54, *R.* sp. TCK and Khangiran2 using hybrid assembly of Illumina short sequencing reads and PacBio long reads, both at high coverage (Sup. Table S1). Hybrid assembly yielded high-quality complete genomes with all circularized replicons (chromosomes and plasmids) for all strains except Q54. In the genome assembly of strain Q54, only one 463-kb plasmid is circularized, leaving eleven fragments, of which one is chromosomal (size 3.79 Mb) and ten (size range: 1–216 kb) that could not be assigned to a replicon type. All these genomes carry plasmids, with one to four confirmed circular plasmids in strains NT-25, TCK, YW14 and Khangiran2 (plasmid size range: 15–462 kb) and possibly more for strains Q54, AB21, JC140 and P007.

In addition, strains *R. halotolerans* AB21^T, *R. endolithicum* JC140^T and *R.* sp. P007 were sequenced using Illumina short sequencing reads only. Assembly produced high-quality draft genomes, with 20 to 84 contigs, with N50 statistics ranging 336–778 kb and average coverage 43x–75x (Sup. Table S1).

Comparison of genomes with digital DNA-DNA hybridization

To direct the assignment of strain NT-26 and the newly sequenced strains to existing or new species, we proceeded to pairwise comparisons of whole genome sequences. The dataset included the new whole genome sequences and already published reference genomes of strain *R.* sp. NT-26 and type

strains of related species *P. pelagicum* R1-200B4^T and *R. marinum* MGL06^T using dDDH (Table 1). As expected, strains NT-25 and NT-26 are highly related (98% dDDH) and are thus considered to belong to the same clone (thereafter referred to as the 'NT-25/26 clone'). They are also closely related to strain TCK (71%). Given the gap of relatedness observed with all other tested strains, these three strains are considered to form a new species.

Strains *R. flavum* YW14^T, *R. halotolerans* AB21^T, and *R.* sp. Khangiran2 are closely related, with their dDDH score almost reaching the classic 70% threshold [55], thus questioning whether these strains should be considered as part of separate or the same species. Strains *R.* sp. Q54 groups clearly with *R. endolithicum* JC140^T (81%) and thus is assigned to the species *R. endolithicum*. The dDDH score of 76.30% between the *P. pelagicum* and *R. marinum* type strain genomes (R1-200B4, MGL06) indicates that both strains belong to the same species, with *R. marinum [37]* having priority over *P. pelagicum* [26,41], warranting that *R. marinum* be kept as the valid species name and *P. pelagicum* as a synonym.

Phylogeny of *Neorhizobium* and *Pseudorhizobium*

We produced a large phylogeny based on the concatenated core-genomes of 571 Rhizobiaceae and Aurantimonadaceae complete genomes and using a fast ML method of inference (S_{ML571}) (Sup. Fig. S1). In addition, we generated a phylogeny focused on the group of interest encompassing the *Neorhizobium* and *Pseudorhizobium* genera ('41NeoPseudo' dataset) using a Bayesian inference and more realistic molecular evolution model (S_{BA41}) (Figure 1), which confirmed the groupings described above based on dDDH. Almost all branches in S_{BA41} are well supported with Bayesian posterior probability (PP) support > 0.97, apart from some internal branches in the *Neorhizobium* clade and the branch grouping of strains *R.* sp. Khangiran2, *R. halotolerans* AB21^T and *R. flavum* YW14^T. Both trees S_{ML571} and S_{BA41} place strain Khangiran2 closest to AB21 and then to YW14, but the relatively low PP support of 0.79 on the stem branch of this group suggests that other configurations are supported by a subset of genes.

The clade containing strains NT-25, NT-26 and TCK groups with the *R. halotolerans/R. flavum* clade, and further with *R. endolithicum* and then *R. marinum*/*Pseudorhizobium pelagicum*, as a well-separated clade from *Neorhizobium*. The clustering of all these strains with R1-200B4^T (=LMG 28314^T = CECT 8629^T), the type strain of the *Pseudorhizobium* genus, warrants the reclassification of all considered species, i.e. *R. halotolerans*, *R. flavum*, *R. endolithicum* and *R. marinum* into the *Pseudorhizobium* genus, thus becoming *P. halotolerans* (*Phalo*), *P. flavum* (*Pfla*), *P. endolithicum* (*Pendo*) and *P. marinum* (*Pmari*). Accordingly, the proposed new species including strains NT-25, NT-26 and TCK should belong to this genus and we propose to name it "*Pseudorhizobium banfieldii*" (*Pban*).

The positions of strains "*R. oryzae*" B4P and "*R. vignae*" CCBAU 05176 in the S_{ML571} and S_{BA41} phylogenetic trees makes it clear that they were incorrectly named (Sup. Fig. S1, S2) and should be designated as *R.* sp. B4P and *N. galegae* CCBAU 05176.

Another source of information on the evolutionary relationships between strains is the distribution of pangenome genes. A hierarchical clustering based on the distribution of groups of orthologous accessory genes present in the '41NeoPseudo' genomes (S_{CL41}) shows a very similar picture, with good support for most branches leading to major clades and species (Figure 2B; Sup. Fig. S3). The

branch separating *Pfla* YW14^T from strains Khangiran2 and *Phalo* AB21^T has again low support (BP: 0.51, AU: 0.52) indicating that almost half the accessory genes support an alternative topology. A major difference with the core genome tree S_{BA41} is that deep-branching strains in S_{BA41} all cluster as a sister group of the *Pseudorhizobium* clade in the pangenome tree S_{CL41} (Sup. Fig. S3). This illustrates that while gene presence/absence provides an interesting complement of data possibly allowing a better resolution at the genus level and below, its interpretation in terms of evolutionary history should be only complementary to (core genome) sequence-based data.

Finally, we compared these phylogenies based on core genome-wide and pangenome-wide data to those obtained using only a restricted set of classic marker genes, either separately or in concatenation, i.e. in a multi-locus sequence analysis (MLSA) (Sup. Dataset S1). The monophyly of *Pmari* and the monophyly of *Pendo* are consistently recovered for each of the five tested marker genes *atpD*, *recA*, *rpoB*, *glnII* and *gyrB*. However, only the *atpD*, *rpoB*, *glnII* and *gyrB* trees, but not *recA*, support *P. banfieldii* to be monophyletic. Also, none of these genes support the monophyly of the group *P. halotolerans*/*P. flavum* (*Phalo*+*Pfla*), specifically due to the inclusion of the *Pban* clade in the former. Combining these markers, the MLSA finds *Pmari*, *Pendo* and *Pban* as being monophyletic, but again not *Phalo*+*Pfla*, with the inclusion of *Pban* being moderately supported (PP support < 0.86). These results indicate that while marker gene-based analyses are mostly consistent with the information obtained from whole genomes, they do not allow for a systematically correct classification of strains into species. Marker genes should therefore only be used for preliminary analysis, but not definitive classification, which should be based on genome-wide data, in accordance to recent guidelines [29].

Phenotype-based classification

We tested the ability of high-throughput phenotyping methods to generate taxonomic information. We found that clustering of strains based on data generated from lipid, protein or metabolic screens yielded a classification broadly similar to that obtained with core-genome alignments. Fatty acid profiling shows the poorest resolution as it could not discriminate between species (Figure 2C). However, it distinctly clustered *Pban*+*Phalo*+*Pfla*+*Pendo* to the exclusion of its outgroups *Neorhizobium* and *P. marinum*, indicating a synapomorphic change of lipid composition at the common origin of these four species. A proteome screen (using MALTI-TOF mass spectrometry) provided a better taxonomic resolution as it allowed the differentiation of *Pendo* from the group *P. flavum*/*P. halotolerans*/*P. banfieldii* (*Phalo*+*Pfla*+*Pban*) (Figure 2D). Metabolism profiles (based on growth curves in 95 different conditions) proved the most accurate sequence-independent support for the phylogenomic tree as it also distinguished the group *Phalo*+*Pfla* from *Pban* and thus almost completely mirrored the internal branching pattern of the genus *Pseudorhizobium* (Figure 2E). The main difference is that all phenotype screens led to cluster *Pmari* strain MGL06^T with outgroups *N. galegae* and *R. leguminosarum* (Figure 2C-E), showing the limited ability of chemotaxonomic and phenotypic analyses to resolve taxonomy at deeper evolutionary scales, likely due to convergence of adaptive traits.

Clade-specific gene sets reveal specific functions and ecologies

We inferred gene family evolution scenarios accounting for HGT history by reconciling gene tree topologies with that of the species tree S_{BA41}. Based on these scenarios, we delineated groups of orthologous genes that reflect the history of gene acquisition in genome lineages – every gain of a

new gene copy in a genome lineage creating a new orthologous group (OG). We looked for groups of OGs with contrasting occurrence patterns between a focal clade and its relatives, to identify specific events of gene gain or loss that led to the genomic differentiation of the clade. Data for all clade comparisons in our '41NeoPseudo' dataset are presented in Sup. Table S2 and are summarized below for the clades on the lineage of strain NT-26; more detailed information and description of gene sets specific to other groups are listed in the Supplementary Text. Major gene sets that have contrasting pattern of occurrence in *Pseudorhizobium* are depicted in Figure 3.

'NT-25/NT-26 clone'

Most of the genome gene content specific to this group is composed of mobile or selfish elements. As expected, this includes the plasmids carrying the arsenite oxidation *aio* locus, and extra copies of the arsenic resistance *ars* operon and the phosphate-specific transporter *pst*/*phn* locus, found on a 322-kb plasmid in NT-26 and a 119-kb plasmid in NT-25. In addition, the chromosome is laden with specific mobile elements. This includes a prophage located between two tRNA genes (positions 1,347–1,419 kb; length 71 kb) characterized by an entire set of phage structural genes and an integrase gene at the end of the locus, as well as a putative integron (positions 352–394kb; length 41 kb) characterized by an integrase gene at the end of the locus, next to a tRNA gene. The prophage carries no gene with identified function other than phage-related ones, whereas the putative integron carries several genes involved in transport and metabolism of a putative branched-chain carbohydrate substrate.

'Pseudorhizobium banfieldii' (*Pban*)

There are exactly 100 genes exclusively present in *Pban* compared to other *Pseudorhizobium* species (see 'cladeB' in Sup. Table S2). Among them, the main feature is a locus grouping genes encoding the RuBisCO and other functions of the Calvin cycle, as well as respiratory chain cytochromes. These systems allow the provision of electrons and the fixation of carbon dioxide and are likely to be the main determinants of chemolithoautotrophy in this species. This locus belongs to a larger *Pban*-specific region composed of two closely located 27-kb and 80-kb fragments, which suggests it results from the recent insertion and domestication of a mobile element (likely interrupted by an even more recent insertion/rearrangement). Among the 37 *Pban*-specific genes in this extended region, several code for enzymes of the classes oxidoreductase, monooxygenase, decarboxylase and glutathione-S transferase, which all use reduced electron acceptors and/or protons, and with their putative substrates including aromatic cyclic and halogenated organic compounds. This suggests a functional link between chemoautotrophy and detoxification pathways. Based on the reconstruction of horizontal gene transfer scenarios, these genes are inferred to originate from a deep-branching lineage of *Neorhizobium*, or possibly from an ancient relative of *R. oryzae* (Sup. Fig. S4). The reconstructed scenarios of HGT (Sup. Table S3) indicate that *Pban* was the final recipient of a series of transfer events, dated as early as the diversification of the *Neorhizobium/Pseudorhizobium* group. This suggests genes coding for autotrophy have been circulating for a long time in this wider taxon, and were later fixed in the *Pban* lineage.

Other *Pban*-specific genes include a locus putatively encoding the biosynthesis of a lipopolysaccharide O-antigen with an N-acetylneuraminic acid function, and a 26-kb region

encoding putative enzymes and transporters related to taurine utilization pathways and to the degradation of possibly halogenated aromatic compounds (Sup. Table S4).

In contrast, there are several genomic regions that have been lost in the *Pban* lineage, including 17 genes specifically absent with respect to cognate *Pseudorhizobium* species. These include an operon present in sister species *Phalo, Pfla* and *Pendo* that encodes a multimeric Na^+/H^+ cation antiporter (also present in *Pmari* strain MGL06^T, with the gene trees indicating it is a HGT recipient from *Phalo*). Another notable *Pban*-specific loss is an operon encoding a cellulose synthase, indicating the likely presence of a cellulose-like polymer in the capsular exopolysaccharide of all other *Pseudorhizobium* species.

Finally, *Pban* genomes specifically lack genes coding for a respiratory complex including several cytochrome *c* oxidases, in linkage with a gene coding the EutK carboxysome-like microcompartment protein, whose known homologues are involved in the degradation of ethanolamine. This locus is present in *Pfla*, *Phalo*, *Pendo* and *Pmari*, and estimated HGT scenarios support a gain by the *Pseudorhizobium* common ancestor and a subsequent loss by the *Pban* ancestor (see Supplementary text). This locus often includes genes that encode redox enzymes that may be the terminal electron acceptor of that respiratory complex; interestingly, these genes vary with the species (Sup. Fig. S5): *Pfla* YW14 carries a copper-containing nitrite reductase, while *Phalo* strains AB21 and Khangiran2 carry a (non-homologous) TAT-dependent nitrous-oxide reductase; the locus in *Pendo* strains harbours no gene encoding such terminal electron acceptor, but other genes encoding metabolic enzymes that differ among strains. This suggests that this respiratory chain and associated putative micro-compartment are used as an evolutionary flexible platform for the reductive activities of these organisms.

‘*Pseudorhizobium sub-clade Pban+Phalo+Pfla*’

The largest specific features of the *Pban+Phalo+Pfla* clade are the presence of the 20-kb super-operon *paa* coding for the uptake and degradation of phenylacetate, and of a 13-kb locus including the *soxXYZABCD* operon that encodes the sulphur oxidation pathway, allowing the lithotrophic oxidation of thiosulphate.

‘*Pseudorhizobium sub-clade Pban+Phalo+Pfla+Pendo*’

Cellular process enriched in the functional annotation of genes specific to the *Pban+Phalo+Pfla+Pendo* clade include NAD cofactor biosynthesis (GO:0009435), tryptophan catabolism (GO:0019441) and phosphatidic acid biosynthesis (GO:0006654). It is also worth reporting the clade-specific presence of an operon coding a thiosulphate sulphurtransferase with a pyrroloquinoline quinone (PQQ)-binding motif, a SoxYZ-like thiosulphate carrier, a SoxH-like metallo-protease related to beta-lactamases, an inositol monophosphatase and an ABC-type transporter of the ferric iron ion. In addition, another clade-specific gene in this locus codes for a membrane-bound PQQ-dependent dehydrogenase with glucose, quinate or shikimate as predicted substrates. A 17-kb locus including *pqqBCDE* operon for the biosynthesis of cofactor PQQ and PQQ-dependent methanol metabolism enzymes was also specifically gained in this clade, but later lost by *Pfla* strain YW14. These genes seem to collectively code for a pathway where the periplasmic oxidation of thiosulphate provides electrons that are carried by SoxYZ and PQQ and used by the metallo-protease and membrane-bound dehydrogenase, respectively, to oxidise targeted

compounds, and doing so possibly lead to their degradation. The role of the monophosphatase and ferric iron transporter in this process is unclear.

Pseudorhizobium (*Pban+Phalo+Pfla+Pendo+Pmari*)

In comparison with the closely related genus *Neorhizobium*, *Pseudorhizobium*-specific genes are over-represented in genes involved in cellular processes related to energy metabolism: 'aerobic respiration' (GO:0009060) and 'electron transport coupled proton transport' (GO:0015990), and to anabolic processes, including the biosynthesis of cofactor NAD (GO:0009435), lipid precursor acetyl-CoA from acetate (GO:0019427) and amino-acid asparagine (GO:0006529). In addition, a *Pseudorhizobium*-specific operon encodes the biosynthesis of osmoprotectant N-acetylglutaminylglutamine (NAGGN) and its attachment to the lipopolysaccharide.

Other Pseudorhizobium species

Specific traits of other clades, including the species *Pendo*, *Pmari* and *Phalo*, are discussed in the Supplementary Text. Among the many species-specific traits found, we can highlight the following predictions: *Phalo* features a specific pathway involved in the biosynthesis of carotenoids; *Pendo* has specific accessory components of its flagellum, and misses many genes that are otherwise conserved in the genus, including a cyanase gene and a four-gene operon involved in degradation of aromatic compounds; *Pmari*, as the most diverged species in the genus, has several hundred species-specific genes, including a 27-kb locus coding for a potassium-transporting ATPase, extrusion transporters and degradation enzymes with putative phenolic compound substrates, and a poly(3-hydroxybutyrate) (PHB) depolymerase.

Verification of bioinformatic predictions of phenotypes

We aimed to validate the predictions of clade-specific phenotypes that would allow us to distinguish taxa and to also confirm the bioinformatically predicted functions of the identified genes. This is important as only experimental validation can clearly link a bacterial genotype with its phenotype, and ultimately with its relevance in the ecology of the taxon. We thus implement a new version of the polyphasic approach to taxonomy [55], where genome-based discovery of phenotypes complements genome relatedness-based delineation of taxa. We focus on *P. banfieldii*, for which our dataset provides the best phylogenetic contrast, with three genomes sampled within the taxon and eight genomes sampled in close relatives, ensuring the robust identification of species-specific genes (Figure 3).

Pban-*specific chemolithoautotrophy.* This is a known trait of all *Pban* strains, which were indeed isolated for that particular property [17,50]. All strains can grow with thiosulphate as a sole electron source and by fixing carbon dioxide as a C source; the use of arsenite as an electron donor is unique to the NT-25/26 clone, due to the presence of the *aio* operon on the clone's specific plasmid.

Pban-*specific use of taurine.* Contrary to the *in silico* prediction, *Pban* strains did not grow in a minimum salts medium with taurine as the sole electron donor under the tested conditions. This does not rule out that the strains can use some (possibly other) sulphonate compound, notably under different inducing conditions; further investigation of this trait needs to be undertaken.

Pban-*specific salt sensitivity.* Tolerance of salt is a known trait of all previously reported strains of *Phalo* (up to 4% NaCl), *Pfla* (up to 4% NaCl), *Pendo* (up to 5% NaCl), and *Pmari* (up to 7% NaCl for former *P. pelagicum* strains and up to 9% NaCl for strain MGL06^T), having indeed inspired the choice of the species epithet of *R. halotolerans* [18,22,26,37,44]. This phenotype could be conferred, at least in part, by the expression of a Na^+/H^+ antiporter, a function that was identified as a marine niche-associated trait in Rhodobacteraceae [54]. The Na^+/H^+ antiporter genes are missing in *Pban*, leading us to predict a lower salt tolerance in this species (Figure 3). The ability to grow in NaCl concentrations ranging from 0 to 9% was tested for 11 strains of *Pseudorhizobium* and related organisms (Sup. Table S5). The results showed no significant difference between *Pban* and other species in the genus with all tolerating up to 3–6% NaCl under our test conditions, apart from *Pmari* strain MGL06, which still grew in the presence of 7% NaCl. This common baseline of salt tolerance suggests that *Pseudorhizobium* core genes encode salt tolerance factors or, less parsimoniously, that all lineages have convergently evolved such traits. The levels of salt tolerance we measured are lower than previously reported for *Pendo* and *Pmari*, and higher for *Phalo*, suggesting that other factors that contribute to salt tolerance in other conditions were not expressed in our experiment. The Na^+/H^+ antiporter, in particular, could provide additional tolerance to those strains which carry it under the appropriate conditions, but may not be expressed under our test conditions. Alternatively, the Na^+/H^+ antiporter gene might have been mis-annotated and could function as a sodium ion-driven proton pump, or could even transport another type of cation than Na^+.

Pban-*specific lack of production of a cellulose polymer. Pban* and *Phalo* strains were plated on yeast extract–mannitol agar medium (YEM) supplemented with 0.1g/l Congo Red dye, a characteristic marker of beta-glucan polymers, to test for the presence of cellulose or a related polymer such as curdlan in their capsular polysaccharide [27]. Contrary to expectations, *P. banfieldii* strains were coloured by the dye, with an orange-red hue, which was observed for other tested *Pseudorhizobium* strains, including *Pendo* strains JC140^T and Q54, and *Pmari* strain MGL06^T. However, growth of *P. banfieldii* strains was inhibited in the presence of the Congo Red dye, resulting in small, dry colonies on YEM plates (Sup. Fig S6; Sup. Table S6) or no growth on MSM + 0.08% YE (data not shown). This growth phenotype was not found among the tested strains from other *Pseudorhizobium* species except for *Phalo* strain Khangiran2 (which was coloured salmon-orange). The inhibitory effect of the dye has been previously observed for other bacteria deficient in the production of beta-glucan polymers [59], and the inhibition observed on *Pban* strains could thus reflect the absence of protection that the cellulose-like polysaccharide provides to other *Pseudorhizobium* isolates. The case of strain Khangiran2 remains unclear as it seems to strongly bind the dye (thus suggesting the polymer is expressed), but it nonetheless suffers from exposure to it.

Pban+Phalo+Pfla-*specific degradation of phenylacetate*. A screen using the API 20 NE identification system showed *Phalo* and *Pfla* strains were positive for phenylacetate assimilation (Sup. Table S7), in line with *in silico* predictions. However, *Pendo* strain JC140 is also positive, and *Pban* strains are negative, contrary to expectations based on the presence of the *paa* operon. The fact that all positive strains but *Phalo* AB21 had a weak response suggests that induction of this function may not be optimal, or that enzymes encoded by this operon may have low affinity for phenylacetate, despite their automatic functional annotation. The activity observed in *Pendo* JC140

can be explained by the presence of an isolated gene coding for a 4-hydroxyphenylacetate 3-monooxygenase, the first step in the phenylacetate degradation pathway, which is not homologous to the one present in the *Pban+Phalo+Pfla*-specific *paa* operon. This gene however also occurs in the *Pendo* strain Q54, which is negative in the assimilation test. Based on our limited data, we can only speculate that, in strain Q54 and in *Pban* strains, the identified phenylacetate degradation genes could have lost their function or not be induced in the culture conditions used in the API identification system.

Search for genotype-phenotype associations

We took advantage of our well-defined phylogenomic framework and of the compendium of phenotypes that we had tested (Sup. Table S8) to search for potential associations between the distribution of accessory genes and the distribution of phenotypes, with the expectation of revealing new gene functions. Using a genome-wide association (GWAS) testing framework, we looked for the basis of significant phenotypes that are not necessarily distributed following the taxonomical structure of species. Specifically, we explored the association between metabolic traits and the distribution of OGs in the accessory genome, using the species tree to account for potential spurious associations linked to oversampling of closely related strains. The GWAS reported numerous significant associations, listed in Sup. Table S9. Manual exploration of results singled out only one association with a clear association pattern that we believe to be of biological relevance: the utilization of beta-methyl-D-glucoside, observed most strongly in strains *Pban* NT-26, *Phalo* AB21 and *N. galegae* HAMBI 540, was associated with the presence of several chromosomal clusters of genes. These include two operons, one encoding an allophanate hydrolase, and another encoding a transporter and a dienelactone hydrolase-related enzyme.

Discussion

The traditional polyphasic approach in bacterial taxonomy combined several criteria, in particular marker gene phylogeny and biochemical phenotypes, to determine the boundaries of taxa [55]. The validity of this approach has recently come into question, due to the growing evidence that most phenotypes are encoded by genes that may be accessory within a species or be shared promiscuously among species, making them poor diagnostic characters [28]. Instead, the growing practice in the field is to use whole genome sequences to estimate similarity between bacterial isolates [12] and to compute phylogenetic trees based on genome-wide data [42]. A tree provides the hierarchical relationships between organisms, while the level of overall genome relatedness provides an objective criterion for the delineation of species. This criterion requires a threshold, and 70%, which is equivalent to the classical score of DNA-DNA hybridization (DDH), has been proposed for dDDH [11], as obtained using the GGDC tool [39]. A 70% threshold is widely considered as adequate for species delineation – but its arbitrary value often calls into question the relevance of the proposed species boundaries.

In this study, the levels of dDDH for the three possible strain pairs among *R.* sp. Khangiran2, *R. halotolerans* AB21^T and *R. flavum* YW14^T are all below, but close to the 70% threshold. According to recent taxonomical guidelines [29], other criteria should be considered to decide on species boundaries. We therefore complemented this threshold-based taxonomic approach

with additional investigations on the genomic, chemotaxonomical and ecological differentiation between strains, with the view of proposing more meaningful species descriptions. These three strains form a clade (‘*Phalo+Pfla*’) that is well supported in the S_{ML571} ML tree, and they share 102 clade-specific genes and 47 genes unique to this group within the *Pseudorhizobium* genus (Sup. Table S2, ‘clade10’ and ‘cladeC’, respectively), which are enriched in functions involved in the biogenesis and modification of membrane lipids. This important cellular pathway could be the means of adaptation to a shared ecological niche, and this group could thus constitute an ecological species [32]. However, core genes specific to *Phalo* (32 genes when compared to all other *Pseudorhizobium* species, including *Pfla*; 86 compared to *Pfla* and closest species *Pban*; 234 compared to only sister strain *Pfla* YW14) are also significantly enriched with coherent cellular functions. They notably encode the biosynthesis of carotenoid pigments and related isoprenoid metabolism, a key metabolic pathway, suggesting that the core genome of *Phalo* may also be involved in the adaptation to its own specific niche. Therefore, the ecological arguments rather support *Phalo* and *Pfla* to be two distinct ecological species. In addition, the relatively lower support in S_{BA571} Bayesian tree for the *Phalo+Pfla* clade is a strong argument against its election to species status. For this reason, we recommend that *R. halotolerans* and *R. flavum* shall remain two distinct species until further evidence to the contrary. These species will however be transferred to the genus *Pseudorhizobium*, according to the presented phylogenetic evidence (Figure 1; Sup. Fig. S1, S2). *R.* sp. Khangiran2 is to be classified into the species of its closest relative type strain, *P. halotolerans* AB21^T.

Through the phylogeny-aware comparison of genomes, we explored the functional specificities of lineages within the bacterial genus *Pseudorhizobium* (Figure 3). From the functional annotation of genomes, we identified the genomic basis of known traits of the strains, such as chemolithoautotrophy or salt tolerance, and predicted others such as a cellulose component of the capsule or a lipopolysaccharide O-antigen. We mapped the distribution of these traits within a phylogenetic framework, identifying those traits which presence or absence was exclusive to a group. This led to the prediction of phenotypes, which we aimed to validate experimentally. The only prediction of contrasting phenotypes that could be verified in the lab is related to the absence of a cellulose-like polymer in *Pban*; this phenotype seems overly dependent on test conditions and should not be used to establish a diagnostic test. In addition, shared phenotypic features of the group were documented, including the general tolerance of members of the *Pseudorhizobium* genus to NaCl (Sup. Table S5). This shared feature suggests that the ancestor of the group might have been itself salt tolerant, and thus possibly a marine organism – a hypothesis consistent with the basal position in the genus tree of seaborne *P. marinum*.

In addition, we identified genes encoding cellular processes and pathways that were over-represented in the specific core genome of clades of the *Pseudorhizobium* genus. This pattern results from the synchronous gain of genes with related functions in a lineage and their subsequent conservation in all descendants – a pattern indicative of positive selection having shaped the gene repertoire of that taxon [33]. Indeed, under an ecotype diversification model, the acquisition of genes enabling the adaptation to a different ecological niche can trigger the emergence of a new ecotype lineage [13]. Ecological isolation may in turn drive the differentiation of the core genome between sister ecotype lineages, with additional adaptive mutations (including new gene gains and losses) producing a knock-on effect leading to ecological specialisation [32]. By exploring the

specific gene repertoire of each clade of the *Pseudorhizobium* genus, we thus shed light on putative ways of adaptations of these groups to their respective ecological niches.

The emergence of the highly specialised NT-25/NT-26 clone could be explained by a scenario involving a sequence of such ecotype diversification events: a first key event was the acquisition of multiple new cytochromes and interacting redox enzymes in the *Pseudorhizobium* genus ancestor, enhancing its capacity to exploit the redox gradients between available environmental compounds. This was followed by the acquisition of a first set of sulphur oxidation enzymes, with electrons from thiosulphate oxidation transferred to the carrier molecules SoxYZ and PQQ, likely to fuel enzymes such as a jointly acquired toxic carbohydrate-degrading metallo-hydrolase. Then, the *sox* gene cluster was gained by the ancestor of *Pban* and *Phalo*, allowing it to use thiosulphate as a source of electrons to fuel respiration and therefore to convert them into proton motive force and to recycle the cellular pool of redox cofactors. The joint acquisition of the phenylacetate degradation pathway – many reactions of which require reduced or oxidized cofactors [63] – allowed this organism to use this aromatic compound and its breakdown products as carbon and electron source. This set of new abilities must have allowed this lineage to colonize new habitats either depleted in organic nutrients or contaminated with toxic organic compounds. This was followed by the acquisition of RuBisCO and other Calvin cycle genes by the *Pban* ancestor, which allowed that strain to live chemolithoautotrophically using sulphur oxidation – again this has likely let this lineage colonize environments yet uncharted by most rhizobia, such as rock surfaces. Finally, the acquisition of a plasmid carrying the arsenite oxidation genes and other factors of resistance to arsenic and heavy metals, allowed the NT-25/NT-26 clone ancestor to successfully colonize the extremely toxic and organic nutrient-poor environment of a gold mine.

Aside from this scenario of extreme specialisation towards chemolithoautotrophy and resistance against toxic heavy metals, all species in the genus *Pseudorhizobium* have achieved significant ecological differentiation from the *bona fide* rhizobial lifestyle, which is characteristic for members of the most closely related genus *Neorhizobium* typically isolated from soil and the plant rhizosphere (Linström 1989; Wang et al. 1998; Österman et al. 2014; Mousavi et al. 2015; Haryono et al. 2018). *P. endolithicum* has only been found inside the mineral matrix of sand grains and its high salt and temperature tolerance phenotypes indicate it is likely adapted to this peculiar lifestyle, even though its capacity to nodulate soybean indicates its ecological niche encompasses various lifestyles [44]. The large core genome turnover that occurred over the long branch leading to this clade (Sup. Table S2) have provided it with a large scope for ecological adaptation, even though the phenotypic consequences of these changes are too complex to be predicted. One notable change occurred in the structural and biosynthetic genes of the flagellum, possibly leading to a deviant morphology of this bacterial motor in this species. This might be linked to its ability to colonize the interior of sand rock particles necessitating a particular type of motility.

The *P. marinum* core genome is largely differentiated from the rest of the genus, owing to its early divergence. Among its species-specific components, some genes are involved in functions that are key for survival in a typically marine lifestyle: transport of K^+ and Cl^- ions, urea and various sugars and organic acids and amino-acids and the degradation of toxic phenolic compounds, in combination with many signal transduction systems, must allow the rapid scavenging/extrusion of rare/excess ions or toxins in response to changing availability of mineral and organic nutrients and the rise in toxicity of the environment. In addition, the (de)polymerisation of storage compound

PHB may allow the cell to survive long-term starvation during nutrient-depleted phases. Finally, the presence of the osmoprotectant NAGGN O-antigen in the lipopolysaccharide – a trait common to all sequenced members of the genus *Pseudorhizobium* – makes this species particularly adapted to life in habitats where salinity can vary strongly.

Conclusion

In summary, we have used a comparative genomics approach within a phylogenomic framework to identify the unique characters of the five species of the genus *Pseudorhizobium*, shedding light on the genetic makeup allowing them to adapt to their respective ecological niche. In accordance with prior observation of phenotypes, our analysis highlighted how this clade of *Rhizobiaceae* evolved towards dramatically different ecological strategies, with marked innovations in tropism and resistance to environmental toxins, thus allowing each species to colonize its own peculiar niche.

Description of *Pseudorhizobium halotolerans* comb. nov.:

The description of the species is the same as the descriptions given by [18], except that it is tolerant to NaCl up to 5 % (w/v), instead of 4 %.

Basionym: *Rhizobium halotolerans* Diange and Lee, 2013.

The type strain, AB21^T (= DSM 105041^T = KEMC 224-056^T = JCM 17536^T), was isolated from chloroethylene-contaminated soil from Suwon, South Korea.

Description of *Pseudorhizobium flavum* comb. nov.:

The description of the species is the same as the descriptions given by [22]. Notably, it has a tolerance of 0–4 % NaCl (w/v).

Basionym: *Rhizobium flavum* Gu *et al.* 2014.

The type strain, YW14^T (= DSM 102134^T = CCTCC AB2013042^T = KACC 17222^T) was isolated from organophosphorus (OP) insecticide-contaminated soil.

Description of *Pseudorhizobium endolithicum* comb. nov.:

The description of the species is the same as the descriptions given by [44].

Basionym: *Rhizobium endolithicum* Parag *et al.* 2013.

The type strain is JC140^T (= DSM 104972^T = KCTC32077^T = CCUG64352^T = MTCC11723^T = HAMBI 2447^T), isolated from sand rock matrix.

Description of *Pseudorhizobium marinum* comb. nov.:

The species name *P. pelagicum* Kimes 2015 is a heterotypic synonym of *R. marinum* Liu 2015. Because the description of the genus *Pseudorhizobium* remains valid (with type strain R1-200B4^T = LMG 28314^T = CECT 8629^T), the species epithet *marinum* is now to be preceded by the *Pseudorhizobium* genus prefix.

The description of the species is the same as the descriptions given by (Liu et al. 2015).

The type strain is MGL06^T (= DSM 106576^T = MCCC 1A00836^T = JCM 30155^T), isolated from seawater that was collected from the surface of the South China Sea (118º 23' E 21º 03' N).

Description of *Pseudorhizobium banfieldii* sp. nov.:

See protologue (Table 2) generated on the Digital Protologue Database [46] under taxon number TA00814.

P. banfieldii ['bæn · 'fild · /ii/] is named in honour of Prof Jillian Banfield, environmental microbiologist whose research revolutionised the view of bacterial and archeal diversity.

P. banfieldii strains are salt tolerant up to 4% NaCl. The following phenotypes distinguish them from other members of *Pseudorhizobium*: they are sulphur oxidizers, and can harvest electrons from sulphur compounds including thiosulphate; they are autotrophic and can assimilate carbon from CO_2 in the presence of an electron source, such as the reduced inorganic sulphur compound thiosulphate. Note that arsenite oxidation and autotrophy in the presence of arsenite are accessory traits borne by a plasmid and are not diagnostic of the species.

In addition, growth of *P. banfieldii* strains is inhibited on yeast extract – mannitol agar medium (YEM) supplemented with 0.1g/l Congo Red dye, resulting on small, dry colonies, and are coloured orange-red by the dye.

The type strain is NT-26^T (= DSM 106348^T = CFBP 8663^T), isolated from arsenopyrite-containing rock in a sub-surface goldmine in the Northern Territory, Australia.

Data availability

All genomic data were submitted to the EBI-ENA under the BioProject accession PRJEB21840/ERP024139, in relation to BioSample accessions ERS1921026–ERS1921030. PacBio runs were submitted under the experiment accessions ERX2989729–ERX2989733. Illumina runs were submitted under the experiment accession ERX3427879, ERX3427880, ERX3427882–ERX3427884, ERX3431115 and ERX3431116. Annotated assemblies were submitted under the analysis accessions ERZ1027023–ERZ1027030.

Intermediary data and results from phenotypic and evolutionary analyses are available on the Figshare data repository under project 65498, available at: https://figshare.com/projects/Taxonomy_of_the_bacterial_genus_Pseudorhizobium/65498. It contains file sets relating to:

- the core genome gene alignment concatenate and the species trees S_{ML571}, S_{BA41} and T_{BA41} (doi: 10.6084/m9.figshare.8316827);
- individual marker gene and MLSA phylogenies (doi: 10.6084/m9.figshare.8332706);
- the pangenome gene alignments for the '571Rhizob' and '41NeoPseudo' datasets (doi :10.6084/m9.figshare.8343473 and doi: 10.6084/m9.figshare.8335265)
- the pangenome gene trees for the '41NeoPseudo' dataset (doi: 10.6084/m9.figshare.8320199);

- the *Pantagruel* phylogenomic database summarizing the pangenome analysis of the '571Rhizob' datasets (with focus on the included '41NeoPseudo' dataset) (doi: 10.6084/m9.figshare.8320142);
- the fatty acid profiling of *Pseudorhizobium* strains (doi: 10.6084/m9.figshare.8316383.v1);
- the API 20 NE biochemical profiling of *Pseudorhizobium* strains (doi: 10.6084/m9.figshare.8316770);
- the NaCl Plate phenotype of *Pseudorhizobium* strains (doi: 10.6084/m9.figshare.8316803);
- the Biolog Gen III metabolism profiling of *Pseudorhizobium* strains (doi: 10.6084/m9.figshare.8316746);
- the genome-wide association testing of Biolog GenIII phenotypes vs. accessory genome presence/absence (doi: 10.6084/m9.figshare.8316818).

Authors Contribution

FL, JMS, JP and FB designed the study. SMMD, FJZ, JZ, JMS, THO and JP isolated, provided and cultivated the bacterial strains. SV and AF performed phenotypic analyses. JS and FL analysed the phenotypic data. FL and HB conducted genome assemblies. FL and XD wrote the phylogenetic analysis software. FL conducted the bioinformatic and evolutionary analyses. FL, JMS, JP, FB and XD wrote the manuscript. All authors read and approve the content of the manuscript.

Acknowledgement

We would like to thank Pascal Bartling, Brian Tindall, Sabine Gronow, Uli Nübel, Gabi Pötter and Peter Schumann for bioinformatic, taxonomic and analytic support as well as very helpful discussions.

Funding

This work was supported by the European Research Council (ERC) (grant ERC260801—BIG_IDEA to FB). FL was supported by a Medical Research Council (MRC) grant (MR/N010760/1) to XD. Computational calculations were performed on Imperial College high-performance computing (HPC) cluster and on MRC Cloud Infrastructure for Microbial Bioinformatics (MRC CLIMB) cloud-based computing servers [15]. THO was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) grant (BB/N012674/1) to JMS.

[1] Alexa, A., Rahnenführer, J., Lengauer, T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13), 1600–7, Doi: 10.1093/bioinformatics/btl140.

[2] Andres, J., Arsène-Ploetze, F., Barbe, V., Brochier-Armanet, C., Cleiss-Arnold, J., Coppée, J.Y., Dillies, M.A., Geist, L., Joublin, A., Koechler, S., Lassalle, F., Marchal, M., Médigue, C., Muller, D., Nesme, X., Plewniak, F., Proux, C., Ramirez-Bahena, M.H., Schenowitz, C., Sismeiro, O., Vallenet, D., Santini, J.M., Bertin, P.N. (2013) Life in an arsenic-containing gold mine: genome and physiology of the autotrophic arsenite-oxidizing bacterium Rhizobium sp. NT-26. Genome Biol. Evol. 5(5), 934–53, Doi: 10.1093/gbe/evt061.

[3] Andres, J., Bertin, P.N. (2016) The microbial genomics of arsenic. FEMS Microbiol. Rev. 40(2), 299–322, Doi: 10.1093/femsre/fuv050.

[4] Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., Pevzner, P.A. (2016) plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 32(22), 3380–7, Doi: 10.1093/bioinformatics/btw493.

[5] Badilla, C., Osborne, T.H., Cole, A., Watson, C., Djordjevic, S., Santini, J.M. (2018) A new family of periplasmic-binding proteins that sense arsenic oxyanions. Sci. Rep. 8(1), 6282, Doi: 10.1038/s41598-018-24591-w.

[6] Bernhardt, P.V., Santini, J.M. (2006) Protein Film Voltammetry of Arsenite Oxidase from the Chemolithoautotrophic Arsenite-Oxidizing Bacterium NT-26. Biochemistry 45(9), 2804–9, Doi: 10.1021/bi0522448.

[7] Bigot, T., Daubin, V., Lassalle, F., Perrière, G. (2013) TPMS: a set of utilities for querying collections of gene trees. BMC Bioinformatics 14, 109, Doi: 10.1186/1471-2105-14-109.

[8] Bottomley, P.J., Maggard, S.P., Leung, K., Busse, M.D. (1991) Importance of saprophytic competence for introduced rhizobia. In: Keister, D.L., Cregan, P.B., (Eds.), The Rhizosphere and Plant Growth: Papers presented at a Symposium held May 8–11, 1989, at the Beltsville Agricultural Research Center (BARC), Beltsville, Maryland, Springer Netherlands, Dordrecht, pp. 135–40.

[9] Brunel, B., Cleyet-Marel, J.-C., Normand, P., Bardin, R. (1988) Stability of Bradyrhizobium japonicum Inoculants after Introduction into Soil. Appl Env. Microbiol 54(11), 2636–42.

[10] Carrascal, O.M.P., VanInsberghe, D., Juárez, S., Polz, M.F., Vinuesa, P., González, V. (2016) Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing Rhizobium species associated with Phaseolus vulgaris. Environ. Microbiol. 18(8), 2660–76, Doi: 10.1111/1462-2920.13415.

[11] Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., Rooney, A.P., Yi, H., Xu, X.-W., De Meyer, S., Trujillo, M.E. (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int. J. Syst. Evol. Microbiol. 68(1), 461–6, Doi: 10.1099/ijsem.0.002516.

[12] Chun, J., Rainey, F.A. (2014) Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. Int. J. Syst. Evol. Microbiol. 64(2), 316–24, Doi: 10.1099/ijs.0.054171-0.

[13] Cohan, F.M. (2017) Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla. Microbiol. Spectr. 5(5), Doi: 10.1128/microbiolspec.MTBP-0014-2016.

[14] Collins, C., Didelot, X. (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLOS Comput. Biol. 14(2), e1005958, Doi: 10.1371/journal.pcbi.1005958.

[15] Connor, T.R., Loman, N.J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M.J., Richardson, E., Ismail, M., Thompson, S.E.-., Kitchen, C., Guest, M., Bakke, M., Sheppard, S.K., Pallen, M.J. (2016) CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. Microb. Genomics 2(9), Doi: 10.1099/mgen.0.000086.

[16] Corsini, P.M., Walker, K.T., Santini, J.M. (2018) Expression of the arsenite oxidation regulatory operon in Rhizobium sp. str. NT-26 is under the control of two promoters that

respond to different environmental cues. MicrobiologyOpen 7(3), e00567, Doi: 10.1002/mbo3.567.

[17] Deb, C., Stackebrandt, E., Pradella, S., Saha, A., Roy, P. (2004) Phylogenetically Diverse New Sulfur Chemolithotrophs of α-Proteobacteria Isolated from Indian Soils. Curr. Microbiol. 48(6), 452–8, Doi: 10.1007/s00284-003-4250-y.

[18] Diange, E.A., Lee, S.-S. (2013) Rhizobium halotolerans sp. nov., Isolated from Chloroethylenes Contaminated Soil. Curr. Microbiol. 66(6), 599–605, Doi: 10.1007/s00284-013-0313-x.

[19] Doyon, J.-P., Ranwez, V., Daubin, V., Berry, V. (2011) Models, Algorithms and Programs for Phylogeny Reconciliation. Brief. Bioinform. 12(5), 392–400, Doi: 10.1093/bib/bbr045.

[20] Engelhardt, T., Sahlberg, M., Cypionka, H., Engelen, B. (2011) Induction of prophages from deep-subseafloor bacteria. Environ. Microbiol. Rep. 3(4), 459–65, Doi: 10.1111/j.1758-2229.2010.00232.x.

[21] Gonzalez, V., Acosta, J.L., Santamaria, R.I., Bustos, P., Fernandez, J.L., Hernandez Gonzalez, I.L., Diaz, R., Flores, M., Palacios, R., Mora, J., Davila, G. (2010) Conserved Symbiotic Plasmid DNA Sequences in the Multireplicon Pangenomic Structure of Rhizobium etli. Appl Env. Microbiol 76(5), 1604–14, Doi: 10.1128/AEM.02039-09.

[22] Gu, T., Sun, L.N., Zhang, J., Sui, X.H., Li, S.P. (2014) Rhizobium flavum sp. nov., a triazophos-degrading bacterium isolated from soil under the long-term application of triazophos. Int. J. Syst. Evol. Microbiol. 64(6), 2017–22, Doi: 10.1099/ijs.0.061523-0.

[23] Hahnke, S., Tindall, B.J., Schumann, P., Sperling, M., Brinkhoff, T., Simon, M. (2012) Planktotalea frisia gen. nov., sp. nov., isolated from the southern North Sea. Int. J. Syst. Evol. Microbiol. 62(7), 1619–24, Doi: 10.1099/ijs.0.033563-0.

[24] Haryono, M., Tsai, Y.-M., Lin, C.-T., Huang, F.-C., Ye, Y.-C., Deng, W.-L., Hwang, H.-H., Kuo, C.-H. (2018) Presence of an Agrobacterium-Type Tumor-Inducing Plasmid in Neorhizobium sp. NCHU2750 and the Link to Phytopathogenicity. Genome Biol. Evol. 10(12), 3188–95, Doi: 10.1093/gbe/evy249.

[25] Kämpfer, P., Kroppenstedt, R.M. (1996) Numerical analysis of fatty acid patterns of coryneform bacteria and related taxa. Can. J. Microbiol. 42(10), 989–1005, Doi: 10.1139/m96-128.

[26] Kimes, N.E., López-Pérez, M., Flores-Félix, J.D., Ramírez-Bahena, M.-H., Igual, J.M., Peix, A., Rodriguez-Valera, F., Velázquez, E. (2015) Pseudorhizobium pelagicum gen. nov., sp. nov. isolated from a pelagic Mediterranean zone. Syst. Appl. Microbiol. 38(5), 293–9, Doi: 10.1016/j.syapm.2015.05.003.

[27] Kneen, B.E., Larue, T.A. (1983) Congo Red Absorption by Rhizobium leguminosarum. Appl. Environ. Microbiol. 45(1), 340–2.

[28] Kumar, N., Lad, G., Giuntini, E., Kaye, M.E., Udomwong, P., Shamsani, N.J., Young, J.P.W., Bailly, X. (2015) Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. Open Biol. 5(1), 140133, Doi: 10.1098/rsob.140133.

[29] de Lajudie, P.M., Andrews, M., Ardley, J., Eardly, B., Jumas-Bilak, E., Kuzmanović, N., Lassalle, F., Lindström, K., Mhamdi, R., Martínez-Romero, E., Moulin, L., Mousavi, S.A., Nesme, X., Peix, A., Puławska, J., Steenkamp, E., Stępkowski, T., Tian, C.-F., Vinuesa, P., Wei, G., Willems, A., Zilli, J., Young, P. (2019) Minimal standards for the description of new genera and species of rhizobia and agrobacteria. Int. J. Syst. Evol. Microbiol., Doi: 10.1099/ijsem.0.003426.

[30] Lartillot, N., Brinkmann, H., Philippe, H. (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(1), S4, Doi: 10.1186/1471-2148-7-S1-S4.

[31] Lassalle, F., Jauneikaite, E., Veber, P., Didelot, X. (2019) Automated reconstruction of all gene histories in large bacterial pangenome datasets and search for co-evolved gene modules with Pantagruel. BioRxiv, 586495, Doi: 10.1101/586495.

[32] Lassalle, F., Muller, D., Nesme, X. (2015) Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. Res. Microbiol. 166(10), 729–41, Doi: 10.1016/j.resmic.2015.06.008.
[33] Lassalle, F., Planel, R., Penel, S., Chapulliot, D., Barbe, V., Dubost, A., Calteau, A., Vallenet, D., Mornico, D., Bigot, T., Guéguen, L., Vial, L., Muller, D., Daubin, V., Nesme, X. (2017) Ancestral Genome Estimation Reveals the History of Ecological Diversification in Agrobacterium. Genome Biol. Evol. 9(12), 3413–31, Doi: 10.1093/gbe/evx255.
[34] Lepage, T., Bryant, D., Philippe, H., Lartillot, N. (2007) A General Comparison of Relaxed Molecular Clock Models. Mol. Biol. Evol. 24(12), 2669–80, Doi: 10.1093/molbev/msm193.
[35] Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32(14), 2103–10, Doi: 10.1093/bioinformatics/btw152.
[36] Linström, K. (1989) Rhizobium galegae, a New Species of Legume Root Nodule Bacteria. Int. J. Syst. Evol. Microbiol. 39(3), 365–7, Doi: 10.1099/00207713-39-3-365.
[37] Liu, Y., Wang, R.-P., Ren, C., Lai, Q.-L., Zeng, R.-Y. (2015) Rhizobium marinum sp. nov., a malachite-green-tolerant bacterium isolated from seawater. Int. J. Syst. Evol. Microbiol. 65(12), 4449–54, Doi: 10.1099/ijsem.0.000593.
[38] Mauchline, T.H., Fowler, J.E., East, A.K., Sartor, A.L., Zaheer, R., Hosie, A.H.F., Poole, P.S., Finan, T.M. (2006) Mapping the Sinorhizobium meliloti 1021 solute-binding protein-dependent transportome. Proc. Natl. Acad. Sci. U. S. A. 103(47), 17933–8, Doi: 10.1073/pnas.0606673103.
[39] Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14(1), 60, Doi: 10.1186/1471-2105-14-60.
[40] Mousavi, S.A., Willems, A., Nesme, X., de Lajudie, P., Lindström, K. (2015) Revised phylogeny of Rhizobiaceae: Proposal of the delineation of Pararhizobium gen. nov., and 13 new species combinations. Syst. Appl. Microbiol. 38(2), 84–90, Doi: 10.1016/j.syapm.2014.12.003.
[41] Oren, A., Garrity, G.M. (2017) List of new names and new combinations previously effectively, but not validly, published. Int. J. Syst. Evol. Microbiol. 67(9), 3140–3, Doi: 10.1099/ijsem.0.002278.
[42] Ormeño-Orrillo, E., Servín-Garcidueñas, L.E., Rogel, M.A., González, V., Peralta, H., Mora, J., Martínez-Romero, J., Martínez-Romero, E. (2015) Taxonomy of rhizobia and agrobacteria from the Rhizobiaceae family in light of genomics. Syst. Appl. Microbiol. 38(4), 287–91, Doi: 10.1016/j.syapm.2014.12.002.
[43] Österman, J., Marsh, J., Laine, P.K., Zeng, Z., Alatalo, E., Sullivan, J.T., Young, J.P.W., Thomas-Oates, J., Paulin, L., Lindström, K. (2014) Genome sequencing of two Neorhizobium galegae strains reveals a noeT gene responsible for the unusual acetylation of the nodulation factors. BMC Genomics 15(1), 500, Doi: 10.1186/1471-2164-15-500.
[44] Parag, B., Sasikala, C., Ramana, C.V. (2013) Molecular and culture dependent characterization of endolithic bacteria in two beach sand samples and description of Rhizobium endolithicum sp. nov. Antonie Van Leeuwenhoek 104(6), 1235–44, Doi: 10.1007/s10482-013-0046-7.
[45] Ronquist, F., Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma. Oxf. Engl. 19(12), 1572–4.
[46] Rosselló-Móra, R., Trujillo, M.E., Sutcliffe, I.C. (2017) Introducing a digital protologue: a timely move towards a database-driven systematics of archaea and bacteria. Antonie Van Leeuwenhoek 110(4), 455–6, Doi: 10.1007/s10482-017-0841-7.
[47] Sadowsky, M.J., Graham, P.H. (1998) Soil Biology of the Rhizobiaceae. In: Spaink, H.P., Kondorosi, A., Hooykaas, P.J.J., (Eds.), The Rhizobiaceae: Molecular Biology of Model Plant-Associated Bacteria, Springer Netherlands, Dordrecht, pp. 155–72.
[48] Santini, J.M., vanden Hoven, R.N. (2004) Molybdenum-Containing Arsenite Oxidase of the Chemolithoautotrophic Arsenite Oxidizer NT-26. J. Bacteriol. 186(6), 1614–9, Doi: 10.1128/JB.186.6.1614-1619.2004.

[49] Santini, J.M., Kappler, U., Ward, S.A., Honeychurch, M.J., vanden Hoven, R.N., Bernhardt, P.V. (2007) The NT-26 cytochrome c552 and its role in arsenite oxidation. Biochim. Biophys. Acta BBA - Bioenerg. 1767(2), 189–96, Doi: 10.1016/j.bbabio.2007.01.009.

[50] Santini, J.M., Sly, L.I., Schnagl, R.D., Macy, J.M. (2000) A New Chemolithoautotrophic Arsenite-Oxidizing Bacterium Isolated from a Gold Mine: Phylogenetic, Physiological, and Preliminary Biochemical Studies. Appl. Environ. Microbiol. 66(1), 92–7, Doi: 10.1128/AEM.66.1.92-97.2000.

[51] Sardiwal, S., Santini, J.M., Osborne, T.H., Djordjevic, S. (2010) Characterization of a two-component signal transduction system that controls arsenite oxidation in the chemolithoautotroph NT-26. FEMS Microbiol. Lett. 313(1), 20–8, Doi: 10.1111/j.1574-6968.2010.02121.x.

[52] Schumann, P., Maier, T. (2014) Chapter 13 - MALDI-TOF Mass Spectrometry Applied to Classification and Identification of Bacteria. In: Goodfellow, M., Sutcliffe, I., Chun, J., (Eds.), Methods in Microbiology, vol. 41, Academic Press, pp. 275–306.

[53] Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14), 2068–9, Doi: 10.1093/bioinformatics/btu153.

[54] Simon, M., Scheuner, C., Meier-Kolthoff, J.P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M., Klenk, H.-P., Schomburg, D., Petersen, J., Göker, M. (2017) Phylogenomics of *Rhodobacteraceae* reveals evolutionary adaptation to marine and non-marine habitats. ISME J. 11(6), 1483–99, Doi: 10.1038/ismej.2016.198.

[55] Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., Whitman, W.B. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52(Pt 3), 1043–7.

[56] Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, btu033, Doi: 10.1093/bioinformatics/btu033.

[57] Surange, S., Wollum II, A.G., Kumar, N., Nautiyal, C.S. (1997) Characterization of Rhizobium from root nodules of leguminous trees growing in alkaline soils. Can. J. Microbiol. 43(9), 891–4, Doi: 10.1139/m97-130.

[58] Suzuki, R., Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12), 1540–2, Doi: 10.1093/bioinformatics/btl117.

[59] Suzuki, T., Campbell, J., Kim, Y., Swoboda, J.G., Mylonakis, E., Walker, S., Gilmore, M.S. (2012) Wall teichoic acid protects Staphylococcus aureus from inhibition by Congo red and other dyes. J. Antimicrob. Chemother. 67(9), 2143–51, Doi: 10.1093/jac/dks184.

[60] Szöllősi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E., Daubin, V. (2013) Efficient Exploration of the Space of Reconciled Gene Trees. Syst. Biol. 62(6), 901–12, Doi: 10.1093/sysbio/syt054.

[61] Szöllősi, G.J., Tannier, E., Daubin, V., Boussau, B. (2015) The inference of gene trees with species trees. Syst. Biol. 64(1), e42–e62, Doi: 10.1093/sysbio/syu048.

[62] Szöllősi, G.J., Tannier, E., Lartillot, N., Daubin, V. (2013) Lateral Gene Transfer from the Dead. Syst. Biol. 62(3), 386–97, Doi: 10.1093/sysbio/syt003.

[63] Teufel, R., Mascaraque, V., Ismail, W., Voss, M., Perera, J., Eisenreich, W., Haehnel, W., Fuchs, G. (2010) Bacterial phenylalanine and phenylacetate catabolic pathway revealed. Proc. Natl. Acad. Sci. 107(32), 14390–5, Doi: 10.1073/pnas.1005399107.

[64] Tóth, E.M., Schumann, P., Borsodi, A.K., Kéki, Z., Kovács, A.L., Márialigeti, K. (2008) Wohlfahrtiimonas chitiniclastica gen. nov., sp. nov., a new gammaproteobacterium isolated from Wohlfahrtia magnifica (Diptera: Sarcophagidae). Int. J. Syst. Evol. Microbiol. 58(4), 976–81, Doi: 10.1099/ijs.0.65324-0.

[65] Vaas, L.A.I., Sikorski, J., Hofner, B., Fiebig, A., Buddruhs, N., Klenk, H.-P., Göker, M. (2013) opm: an R package for analysing OmniLog(R) phenotype microarray data. Bioinforma. Oxf. Engl. 29(14), 1823–4, Doi: 10.1093/bioinformatics/btt291.

[66] Vaser, R., Sović, I., Nagarajan, N., Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5), 737–46, Doi: 10.1101/gr.214270.116.

[67] Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9(11), e112963, Doi: 10.1371/journal.pone.0112963.

[68] Wang, E.T., van Berkum, P., Beyene, D., Sui, X.H., Dorado, O., Chen, W.X., Martínez-Romero, E. (1998) Rhizobium huautlense sp. nov., a symbiont of Sesbania herbacea that has a close phylogenetic relationship with Rhizobium galegae. Int. J. Syst. Evol. Microbiol. 48(3), 687–99, Doi: 10.1099/00207713-48-3-687.

[69] Warelow, T.P., Pushie, M.J., Cotelesage, J.J.H., Santini, J.M., George, G.N. (2017) The active site structure and catalytic mechanism of arsenite oxidase. Sci. Rep. 7(1), 1757, Doi: 10.1038/s41598-017-01840-y.

[70] Wick, R.R., Judd, L.M., Gorrie, C.L., Holt, K.E. (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput. Biol. 13(6), e1005595, Doi: 10.1371/journal.pcbi.1005595.

[71] Will, S.E., Henke, P., Boedeker, C., Huang, S., Brinkmann, H., Rohde, M., Jarek, M., Friedl, T., Seufert, S., Schumacher, M., Overmann, J., Neumann-Schaal, M., Petersen, J. (2019) Day and Night: Metabolic Profiles and Evolutionary Relationships of Six Axenic Non-Marine Cyanobacteria. Genome Biol. Evol. 11(1), 270–94, Doi: 10.1093/gbe/evy275.

[72] 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., Mauchline, T.H., East, A.K., Quail, M.A., Churcher, C., Arrowsmith, C., Cherevach, I., Chillingworth, T., Clarke, K., Cronin, A., Davis, P., Fraser, A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Sanders, M., Simmonds, M., Whitehead, S., Parkhill, J. (2006) The genome of Rhizobium leguminosarum has recognizable core and accessory components. Genome Biol. 7(4), R34, Doi: 10.1186/gb-2006-7-4-r34.

Tables

Table 1. Genome-to-Genome Distance Calculation (GGDC) of similarity between *Pseudorhizobium* strains.

Genome similarity estimated using the GGDC tool (version 2.1) with the formula d_4 (BLASTN identities / HSP length). Similarity values are in percent (%). Green shading indicate values over the 70% threshold recommended for assignment to the same species; orange shading indicate contentious values close to the threshold.

			NT-26	NT-25	TCK	**AB21**	Khangiran2	**YW14**	**JC140**	Q54	**MGL06**
1	***Pban* NT-26^T**	DSM 106348	x	x	x	x	x	x	x	x	x
2	*Pban* NT-25	DSM 106347	**98.1**	x	x	x	x	x	x	x	x
3	*Pban* TCK	DSM 13828	**71.5**	**71.7**	x	x	x	x	x	x	x
4	***Phalo* AB21^T**	DSM 105041	61.9	61.7	53.2	x	x	x	x	x	x
5	*Phalo* Khangiran2	DSM 106339	60.5	60.6	52.4	66.9	x	x	x	x	x
6	***Pfla* YW14^T**	DSM 102134	62.1	61.9	53.9	67.4	65.6	x	x	x	x
7	***Pendo* JC140^T**	DSM 104972	24.4	24.4	23.2	24.3	24.2	24.6	x	x	x
8	*Pendo* Q54	DSM 106353	24.9	24.9	23.2	24.9	24.4	25.1	**81.1**	x	x
9	***Pmari* MGL06^T**	DSM 106576	21.7	21.7	21.1	21.7	21.7	21.8	22.1	22.2	x
10	***Pmari* R1-200B4^T**	-	21.7	22.3	21.8	21.6	21.7	21.8	22.2	22.3	**76.3**

Table 2. Digital Protologue description *Pseudorhizobium banfieldii* sp. nov. (TA00814)

TAXONUMBER	TA00814
TYPE OF DESCRIPTION	New Description
SPECIES NAME	*Pseudorhizobium banfieldii*
GENUS NAME	*Pseudorhizobium*
SPECIFIC EPITHET	*banfieldii*
SPECIES STATUS	sp. nov.
SPECIES ETYMOLOGY	in honour of Prof Jillian Banfield, environmental microbiologist whose research revolutionised the view of bacterial and archeal diversity.
DESIGNATION OF THE TYPE STRAIN	NT-26
STRAIN COLLECTION NUMBERS	DSM 106348, CFBP 8663
16S rRNA GENE ACCESSION NUMBER	AF159453
GENOME ACCESSION NUMBER [RefSeq]	GCF_000967425.1
GENOME STATUS	complete
GENOME SIZE	4577425
GC mol %	61.84
COUNTRY OF ORIGIN	Australia
REGION OF ORIGIN	Northern Territory
DATE OF ISOLATION UNKNOWN (< yyyy)	< 1999
SOURCE OF ISOLATION	moist arsenopyrite-containing rock
SAMPLING DATE	1999-01-01
GEOGRAPHIC LOCATION	Granites Gold Mine
LATITUDE	20°32′18.4″S
LONGITUDE	130°18′37.7″E
DEPTH	60
NUMBER OF STRAINS IN STUDY	3
SOURCE OF ISOLATION OF NON-TYPE STRAINS	moist arsenopyrite-containing rock, soil
GROWTH MEDIUM, INCUBATION CONDITIONS [Temperature, pH, and further information] USED FOR STANDARD CULTIVATION	Minimal salts medium (MSM) as per (Santini et al., 2000. Applied & Environmental Microbiology 66(1):92-97). pH 8.0 28°C
IS A DEFINED MEDIUM AVAILABLE	yes (Santini et al., 2000. Applied & Environmental Microbiology 66(1):92-97).
ALTERNATIVE MEDIUM 1	Luria Bertani
GRAM STAIN	NEGATIVE
CELL SHAPE	rod
CELL SIZE (length or diameter)	1
MOTILITY	motile
IF MOTILE	flagellar
IF FLAGELLATED	2 sub-terminal flagella
SPORULATION (resting cells)	none
COLONY MORPHOLOGY	produces EPS
LOWEST pH FOR GROWTH	4 in MSM + arsenite
HIGHEST pH FOR GROWTH	9 in MSM + arsenite
pH OPTIMUM	8.0 in MSM + arsenite
pH CATEGORY	neutrophile
RELATIONSHIP TO O2	facultative aerobe
O2 CONDITIONS FOR STRAIN TESTING	aerobiosis
CARBON SOURCE USED [class of compounds]	sugars, organic acids, carbon dioxide
CARBON SOURCE USED [specific compounds]	acetate, arabinose, galactose, fructose, fumarate, glucose, glycerol, inositol, lactate, lactose, malate, maltose, mannitol, pyruvate, trehalose, raffinose, salicin, succinate, sucrose, xylose
CARBON SOURCE NOT USED [specific compounds]	citrate, rhamnose, sorbitol.
NITROGEN SOURCE	NO3, NH4+
TERMINAL ELECTRON ACCEPTOR	oxygen, nitrate, nitrite
ENERGY METABOLISM	mixotroph
BIOSAFETY LEVEL	1
HABITAT	ENVO:00001995, ENVO:00001998, ENVO:00005801
BIOTIC RELATIONSHIP	free-living
KNOWN PATHOGENICITY	none
MISCELLANEOUS, EXTRAORDINARY FEATURES RELEVANT FOR THE DESCRIPTION	facultative chemolithoautotrophe on thiosulfate as electron source and carbon dioxide as C source

Figures

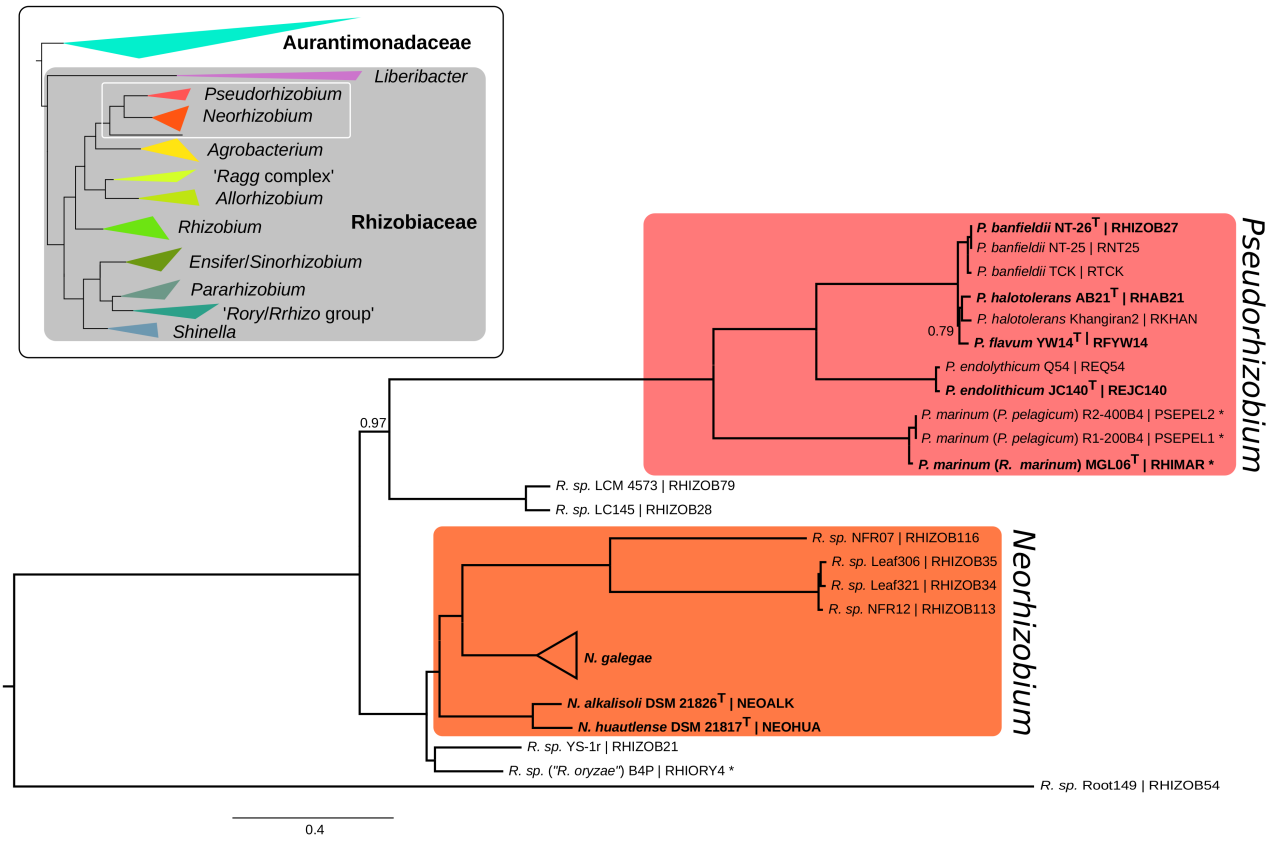


Figure 1: Bayesian phylogenetic tree of 41 organisms from the *Neorhizobium* and *Pseudorhizobium* genera and close relatives (S_{BA41}). Tree obtained with Phylobayes under the GTR-CAT protein evolution model, based on a concatenated alignment of 155 pseudo-core protein loci. All posterior probability branch supports are 1.00 unless indicated. The organism name is followed (after the pipe symbol |) by the identifier in the Pantagruel pangenome database of this study. Strains whose species affiliation are corrected in this study are marked with an asterisk *. Species type strains are in bold. The clade *N. galegae* was collapsed; it includes the type strain *N. galegae bv. orientalis* HAMBI 540^T (organism id: NEOGAL2). The full tree is presented in Sup. Fig. S2. A schematic view of the phylogenetic context of this group is indicated in inset (based on tree S_{ML571}, which presented in full in Sup. Fig. S1). *Raag*: *R. aggregatum*; *Rory*: *R. oryzae*; *Rrhizo*: *R. rhizosphereae*. The alignment and tree files are available on Figshare (doi: 10.6084/m9.figshare.8316827).

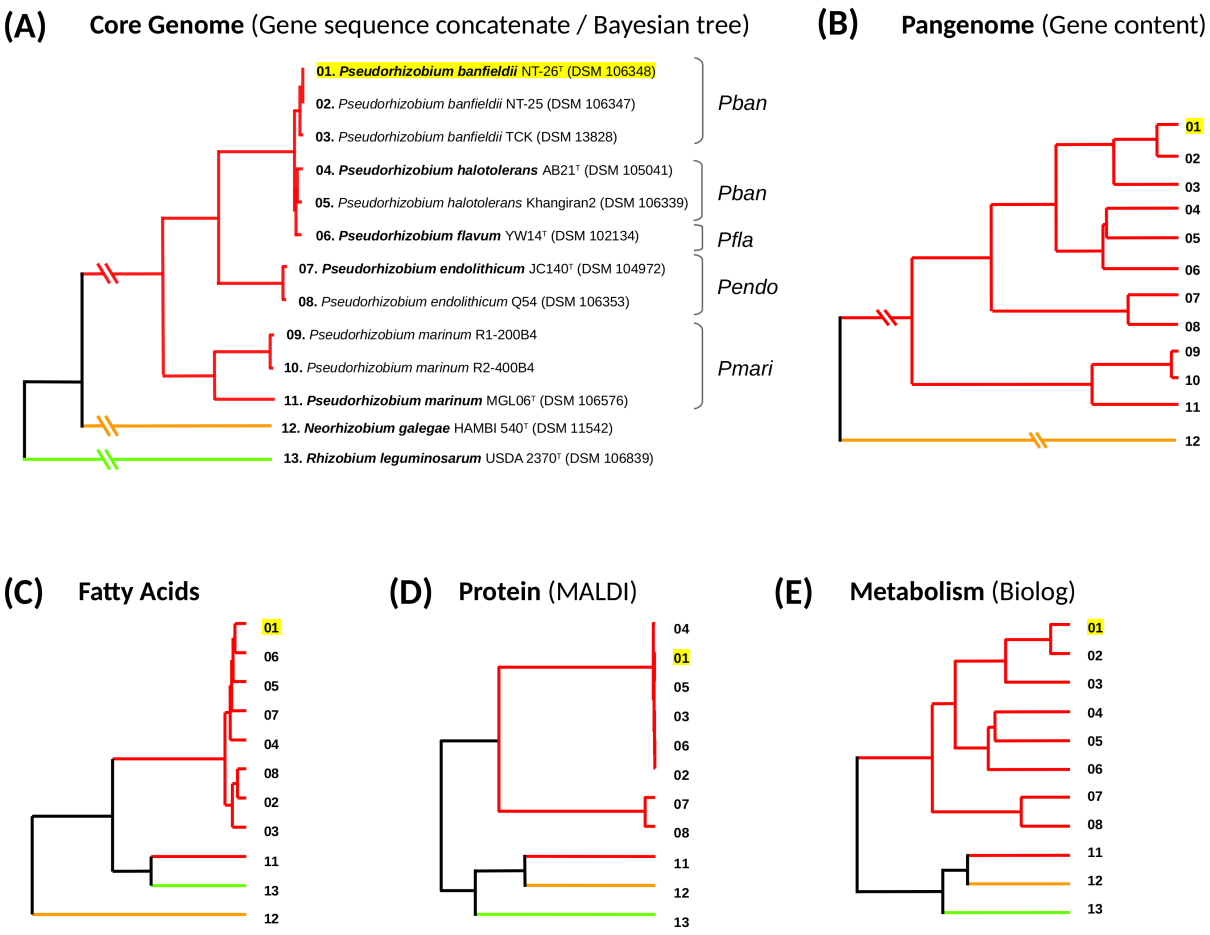


Figure 2: Phylogenetic clustering of Pseudorhizobium strains based on genomic and phenotypic characters.

A numeric code corresponding to strains as indicated in panel (A) is used in the other panels. **A.** Bayesian tree S_{BA41} based on core genome gene concatenate, adapted from Figure 1; *R. leguminosarum* is indicated as an outgroup, in accordance with the maximum-likelihood tree S_{ML571} based on the same core genome gene set and wider taxon sample. **B**. Hierarchical clustering dendrogram based on the accessory gene content of strains defined at the orthologous group level. **C-E**. Hierarchical clustering dendrogram based on phenotypic data relating to fatty acid content, protein content and metabolic abilities. Underlying data are available on Figshare:
doi: 10.6084/m9.figshare.8316827; doi: 10.6084/m9.figshare.8316383;
doi: 10.6084/m9.figshare.8316746.

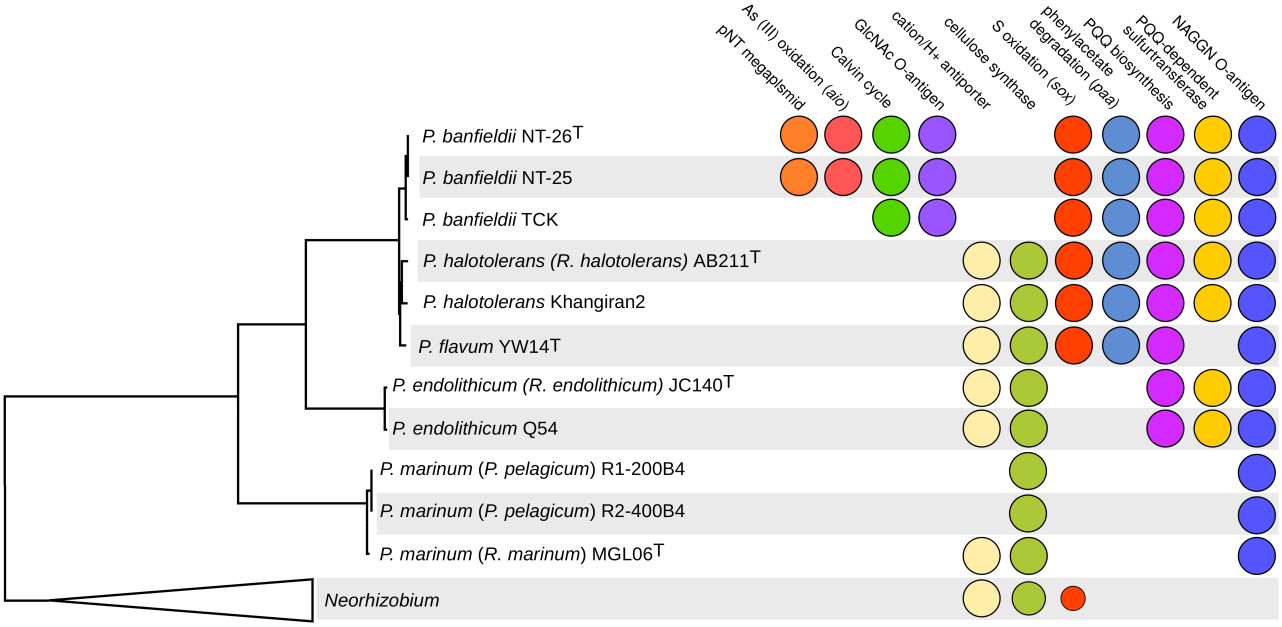
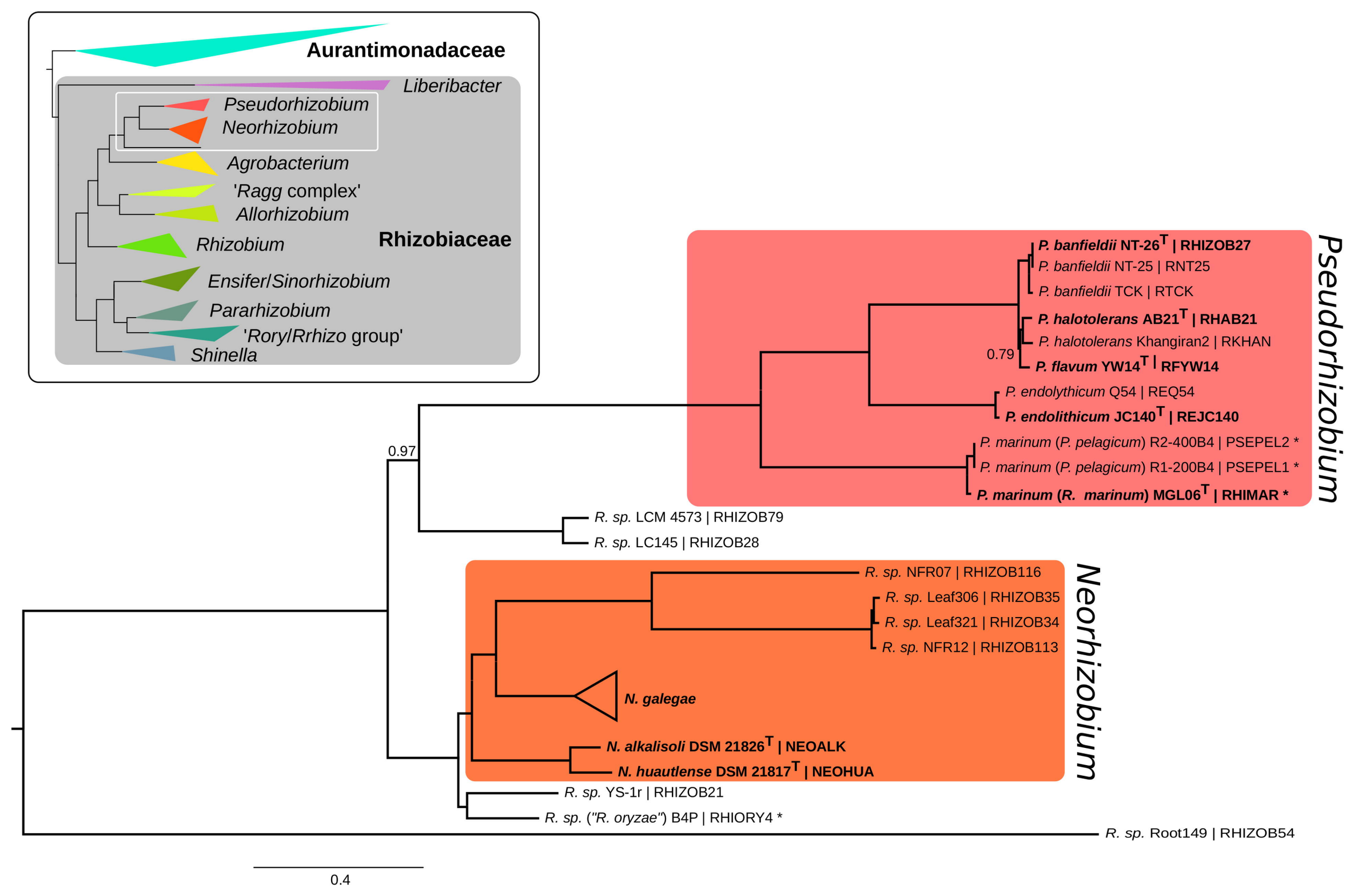
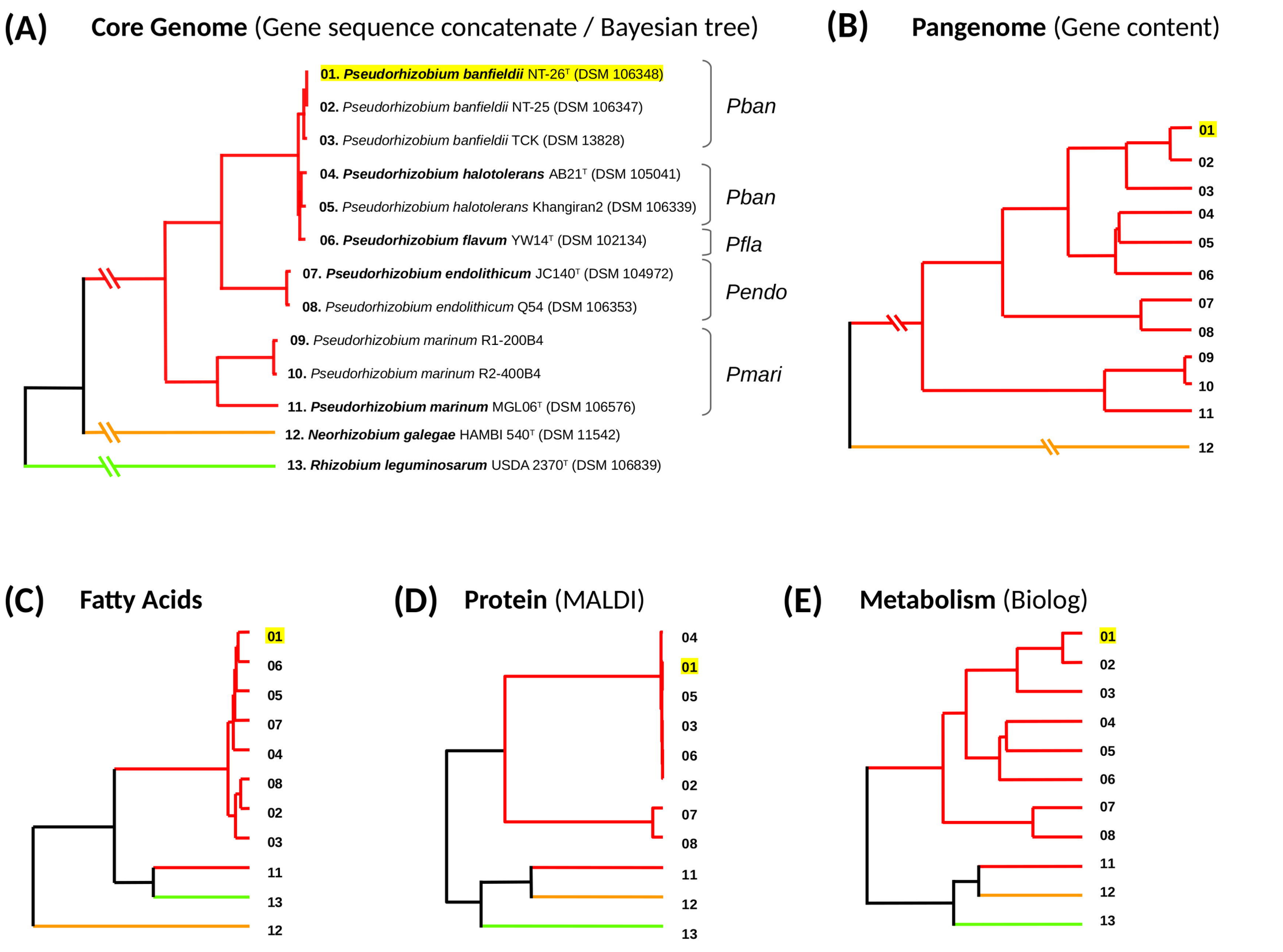
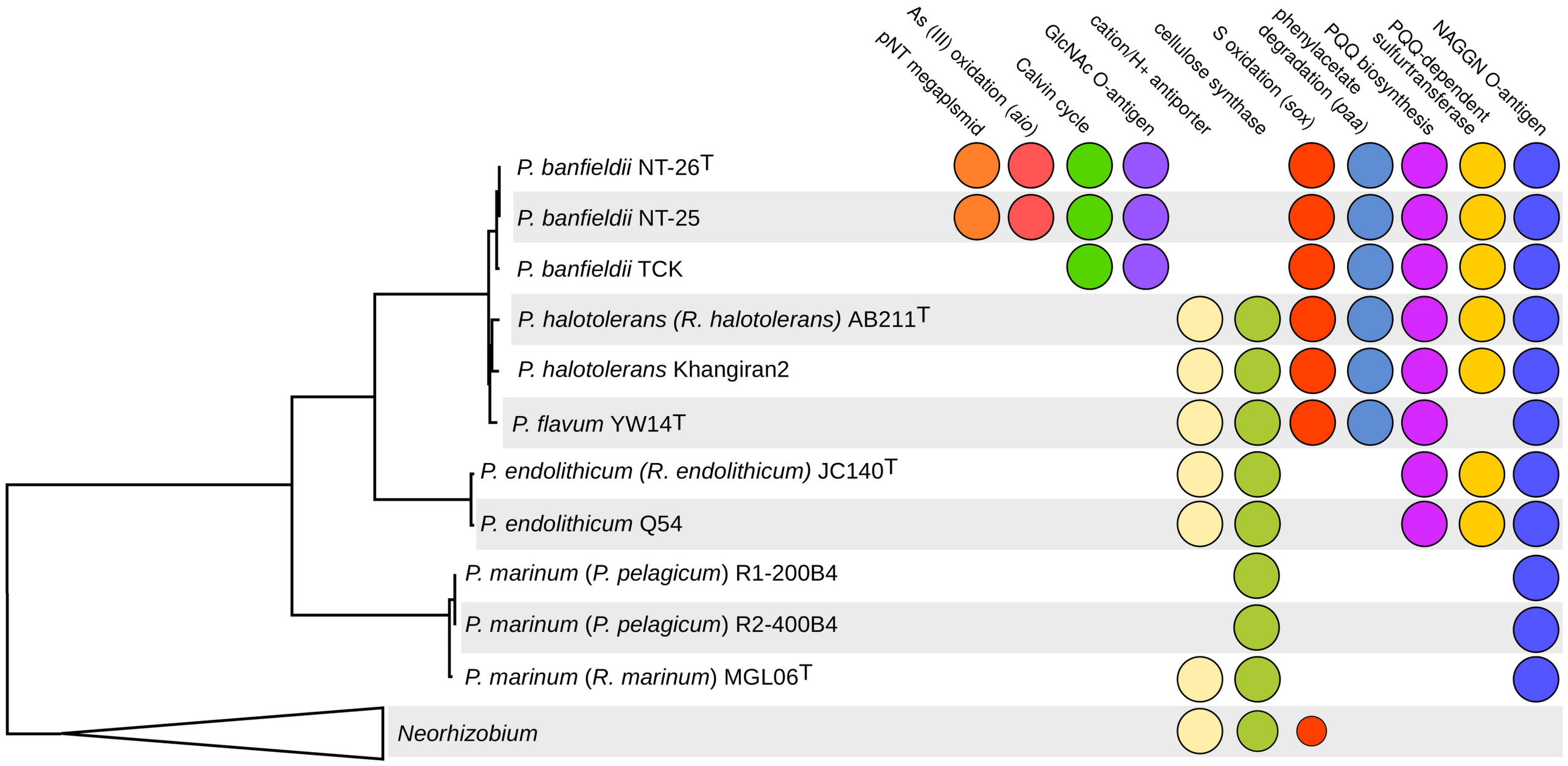


Figure 3: phylogenetic distribution of selected pathways in *Pseudorhizobium*
A circle indicates the presence of the genes or pathway in the genome. For the collapsed *Neorhizobium* clade, the frequency of presence of the genes is indicated by the area of the circle.


Aurantimonadaceae
Liberibacter
Pseudorhizobium
Neorhizobium
Agrobacterium
'Ragg complex'
Allorhizobium
Rhizobium
Rhizobiaceae
Ensifer/Sinorhizobium
Pararhizobium
'Rory/Rrhizo group'
Shinella
P. banfieldii NT-26T | RHIZOB27
P. banfieldii NT-25 | RNT25
P. banfieldii TCK | RTCK
P. halotolerans AB21T | RHAB21
P. halotolerans Khangiran2 | RKHAN
0.79
P. flavum YW14T | RFYW14
P. endolythicum Q54 | REQ54
P. endolithicum JC140T | REJC140
P. marinum (P. pelagicum) R2-400B4 | PSEPEL2 *
P. marinum (P. pelagicum) R1-200B4 | PSEPEL1 *
P. marinum (R. marinum) MGL06T | RHIMAR *
Pseudorhizobium
0.97
R. sp. LCM 4573 | RHIZOB79
R. sp. LC145 | RHIZOB28
R. sp. NFR07 | RHIZOB116
R. sp. Leaf306 | RHIZOB35
R. sp. Leaf321 | RHIZOB34
R. sp. NFR12 | RHIZOB113
N. galegae
N. alkalisoli DSM 21826T | NEOALK
N. huautlense DSM 21817T | NEOHUA
Neorhizobium
R. sp. YS-1r | RHIZOB21
R. sp. ("R. oryzae") B4P | RHIORY4 *
R. sp. Root149 | RHIZOB54
0.4


(A) Core Genome (Gene sequence concatenate / Bayesian tree)
01. Pseudorhizobium banfieldii NT-26T (DSM 106348)
02. Pseudorhizobium banfieldii NT-25 (DSM 106347)
03. Pseudorhizobium banfieldii TCK (DSM 13828)
Pban
04. Pseudorhizobium halotolerans AB21T (DSM 105041)
05. Pseudorhizobium halotolerans Khangiran2 (DSM 106339)
Pban
06. Pseudorhizobium flavum YW14T (DSM 102134)
Pfla
07. Pseudorhizobium endolithicum JC140T (DSM 104972)
08. Pseudorhizobium endolithicum Q54 (DSM 106353)
Pendo
09. Pseudorhizobium marinum R1-200B4
10. Pseudorhizobium marinum R2-400B4
11. Pseudorhizobium marinum MGL06T (DSM 106576)
Pmari
12. Neorhizobium galegae HAMBI 540T (DSM 11542)
13. Rhizobium leguminosarum USDA 2370T (DSM 106839)
(B) Pangenome (Gene content)
01
02
03
04
05
06
07
08
09
10
11
12
(C) Fatty Acids
01
06
05
07
04
08
02
03
11
13
12
(D) Protein (MALDI)
04
01
05
03
06
02
07
08
11
12
13
(E) Metabolism (Biolog)
01
02
03
04
05
06
07
08
11
12
13


pNT megaplsmid
As (III) oxidation (aio)
Calvin cycle
GlcNAc O-antigen
cation/H+ antiporter
cellulose synthase
S oxidation (sox)
phenylacetate degradation (paa)
PQQ biosynthesis
PQQ-dependent sulfurtransferase
NAGGN O-antigen
P. banfieldii NT-26T
P. banfieldii NT-25
P. banfieldii TCK
P. halotolerans (R. halotolerans) AB211T
P. halotolerans Khangiran2
P. flavum YW14T
P. endolithicum (R. endolithicum) JC140T
P. endolithicum Q54
P. marinum (P. pelagicum) R1-200B4
P. marinum (P. pelagicum) R2-400B4
P. marinum (R. marinum) MGL06T
Neorhizobium