Title: Symbiotic and non-symbiotic members of the genus *Ensifer* (syn. *Sinorhizobium*) are separated into two clades based on comparative genomics and high-throughput phenotyping

Authors: Camilla Fagorzi[1], Alexandru Ilie[1], Francesca Decorosi[2], Lisa Cangioli[1], Carlo Viti[2], Alessio Mengoni[1*], George C diCenzo[1,3*]

Affiliations: [1] Department of Biology, University of Florence, Sesto Fiorentino, Italy

[2] Genexpress Laboratory, Department of Agriculture, Food, Environment and Forestry, University of Florence, Sesto Fiorentino, Italy

[3] Department of Biology, Queen's University, Kingston, Ontario, Canada

*** Corresponding Authors:** Alessio Mengoni (alessio.mengoni@unifi.it) and George diCenzo (george.dicenzo@queensu.ca).

Data Deposition: Genome sequences were deposited to the NCBI under the BioProject accession PRJNA622509.

ABSTRACT

Rhizobium – legume symbioses serve as a paradigmatic example for the study of mutualism evolution. The genus *Ensifer* (syn. *Sinorhizobium*) contains diverse plant-associated bacteria, a subset of which can fix nitrogen in symbiosis with legumes. To gain insights into the evolution of symbiotic nitrogen fixation (SNF), and inter-kingdom mutualisms more generally, we performed extensive phenotypic, genomic, and phylogenetic analyses of the genus *Ensifer*. The data suggest that SNF likely emerged several times within the genus *Ensifer* through independent horizontal gene transfer events. Yet, the majority (105 of 106) of the *Ensifer* strains with the *nodABC* and *nifHDK* nodulation and nitrogen fixation genes were found within a single, monophyletic clade. Comparative genomics highlighted several differences between the "symbiotic" and "non-symbiotic" clades, including divergences in their pangenome content. Additionally, strains of the symbiotic clade carried 325 fewer genes, on average, and appeared to have fewer rRNA operons than strains of the non-symbiotic clade. Initial characterization of a subset of ten *Ensifer* strains identified several putative phenotypic differences between the clades. Tested strains of the non-symbiotic clade could catabolize 25% more carbon sources, on average, than strains of the symbiotic clade, and they were better able to grow in LB medium and tolerate alkaline conditions. On the other hand, the tested strains of the symbiotic clade were better able to tolerate heat stress and acidic conditions. We suggest that these data support the division of the genus *Ensifer* into two main subgroups, as well as the hypothesis that pre-existing genetic features are required to facilitate the evolution of SNF in bacteria.

Keywords: Mutualism, evolutionary biology, phenomics, comparative genomics, rhizobia, Proteobacteria

SIGNIFICANCE STATEMENT

The bacterial genus *Ensifer* contains ecologically important N_2-fixing symbionts of leguminous plants, as well as non-symbiotic species. However, the evolutionary dynamics of symbiotic nitrogen fixation within this genus are unclear, and it remains an open question of whether the gain of classical symbiotic N_2-fixation genes is sufficient to allow a bacterium to fix nitrogen. Our results suggest that the symbiotic species of the genus *Ensifer* predominately group separately from the non-symbiotic species, but that symbiotic abilities were likely acquired multiple times within this group. This study provides new insight into the evolution of symbiotic N_2-fixation in a bacterial genus, while supporting the hypothesis that genetic features aside from the classical symbiotic N_2-fixation genes contribute to the evolution of symbiotic potential.

INTRODUCTION

Symbioses are pervasive phenomena present in all Eukaryotic forms of life (López-García et al. 2017). These includes the facultative symbiotic interactions, obligate symbioses, and the evolution of organelles (Douglas 2014), with symbiotic nitrogen fixation (SNF) being a paradigmatic example of the latter (Masson-Boivin & Sachs 2018). SNF (the conversion of N_2 to NH_3) is performed by a polyphyletic group of bacteria from the Alphaproteobacteria and Betaproteobacteria (whose nitrogen-fixing members are collectively called rhizobia) and members of the genus *Frankia* (Wang & Young 2019; Masson-Boivin et al. 2009) while intra-cellularly housed within specialized organs (nodules) of specific plants in the family *Fabaceae* and the genus *Parasponia*, as well as the actinorhizal plants (Werner et al. 2014; Velzen et al. 2018; Griesmann et al. 2018). The advantages and evolutionary constraints to SNF have long been investigated in the conceptual framework of mutualistic interactions and the exchange of goods (see for instance (Sørensen et al. 2019; Werner et al. 2015; Heath & Tiffin 2007)), and quantitative estimations with metabolic reconstructions have also been performed (Pfau et al. 2018; diCenzo, Tesi, et al. 2019).

The establishment of a symbiotic nitrogen-fixing interaction requires that the bacterium encode several diverse molecular functions, including those related to signalling and metabolic exchange with the host plant, nitrogenase and nitrogenase-related functions, and escaping or resisting the plant immune system (Oldroyd et al. 2011; Haag et al. 2013; Poole et al. 2018). In general, the primary genes required for SNF (i.e., the *nod*, *nif*, and *fix* genes) are located within mobile genetic elements that include symbiotic islands and symbiotic (mega)-plasmids (Tian & Young 2019; Checcucci et al. 2019; Geddes et al. 2020), facilitating their spread through horizontal gene transfer (HGT) (Sullivan et al. 1995; Barcellos et al. 2007; Pérez Carrascal et al. 2016). Emphasizing the role of HGT in the evolution of rhizobia, rhizobia are dispersed across seven families of the Alphaproteobacteria and one family of the

Betaproteobacteria, and most genera with rhizobia also contain non-rhizobia (Garrido-Oter et al. 2018; Wang 2019).

An interesting area of investigation is whether the evolution of mutualistic symbioses, such as SNF, depends on metabolic/genetic requirements ("facilitators", as in (Gerhart & Kirschner 2007)) aside from the strict symbiotic genes (Zhao et al. 2017; Long 2001; Sanjuán 2016). In other words, i) is the acquisition of symbiotic genes present in genomic islands or plasmids sufficient to become a symbiont or, ii) are metabolic pre-requirements or adaptation successive to HGT required? A comparative genomics study of 1,314 Rhizobiales genomes identified no functional difference between rhizobia and non-rhizobia based on Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations (Garrido-Oter et al. 2018), suggestive of an absence of obvious facilitators. In contrast, experimental studies are generally consistent with an important role of non-symbiotic genes in establishing or optimizing rhizobium – legume symbioses. Several studies have shown that effective symbionts are not produced following the transfer of symbiotic plasmids from rhizobia of the genera *Rhizobium* or *Ensifer* (syn. *Sinorhizobium*) to closely related non-rhizobia from the genera *Agrobacterium* or *Ensifer* (see for instance (Hooykaas et al. 1982; Finan et al. 1986; Rogel et al. 2001); reviewed in (diCenzo, Zamani, et al. 2019)). Similarly, the same symbiotic island is associated with vastly different symbiotic phenotypes depending on the *Mesorhizobium* genotype (Nandasena et al. 2007; Haskett et al. 2016). Further supporting the need for additional adaptations to support SNF, symbiosis plasmid transfer coupled to experimental evolution can lead to the gain of more advanced symbiotic phenotypes (Doin de Moura et al. 2020).

The genus *Ensifer* provides an ideal model to further explore the differentiation, or lack thereof, of symbiotic bacteria from non-symbionts. This genus comprises rhizobia such as *E. meliloti* and *E. fredii*, as well as non-rhizobia like *E. morelense* and *E. adhaerens*, and many members have been extensively studied producing an abundant set of experimental and

genomic data (for a recent review, see (diCenzo, Zamani, et al. 2019)). The genus *Ensifer*, as currently defined, resulted from the combination of the genera *Sinorhizobium* and *Ensifer* based on similarities in the 16S rRNA and *recA* sequences of the type strains and the priority of the name *Ensifer* (Young 2003; Willems et al. 2003). Multilocus sequence analysis supported the amalgamation of these genera (Martens et al. 2007), although it was subsequently noted that *E. adhaerens* (the type strain) is an outgroup of this taxon based on whole genome phylogenomics (Ormeño-Orrillo et al. 2015). A more recent taxonomy approach based on genome phylogeny suggests that the genus *Ensifer* should again be split, with the initial type strains of *Ensifer* and *Sinorhizobium* belonging to separate genera (Parks et al. 2018).

In this paper we report an extensive comparative genomic and initial phenotypic characterization of legume symbionts and non-symbionts of the genus *Ensifer*. We identified that SNF likely evolved multiple times through independent HGT events; even so, most symbionts were found in a single clade, consistent with a requirement for pre-existing genetic features to facilitate the evolution of SNF. Moreover, the symbiotic and non-symbiotic clades differed in their pangenome composition, and tests with a subset of strains suggested they also differed in their substrate utilization and resistance phenotypes as measured by the Phenotype MicroArray™ platform. We suggest that the data support the division of the genus *Ensifer* into two subgroups, corresponding to the genera *Ensifer* and *Sinorhizobium* of the Genome Taxonomy Database (Parks et al. 2018).

MATERIALS AND METHODS

Genome Sequencing, Assembly, and Annotation

Prior to short-read sequencing, all strains were grown to stationary phase at 30˚C in TY medium (5 g L^{-1} tryptone, 3 g L^{-1} yeast extract, and 0.4 g L^{-1} $CaCl_2$). Total genomic DNA was isolated using a standard cetyltrimethylammonium bromide (CTAB) method (Perrin et al.

2015). Short-read sequencing was performed at IGATech (Udine, Italy) using an Illumina HiSeq2500 instrument with 125 bp paired-end reads. Two independent sequencing runs were performed for *E. morelense* Lc04 and *E. psoraleae* CCBAU 65732, whereas *E. morelense* Lc18 and *E. sesbaniae* CCBAU 65729 were sequenced once. For the long-read sequencing, *E. sesbaniae* was grown to mid-exponential phase at 30˚C in MM9 minimal medium (MOPS buffer [40 mM MOPS, 20 mM KOH], 19.2 mM NH_4Cl, 85.6 mM NaCl, 2 mM KH_2PO_4, 1 mM $MgSO_4$, 0.25 mM $CaCl_2$, 1 μg ml^{-1} biotin, 42 nM $CoCl_2$, 38 μM $FeCl_3$, 10 μM thiamine-HCl, and 10 mM sucrose). Total genomic DNA was isolated as described elsewhere (Cowie et al. 2006). Long-read sequencing was performed in-house with a Pacific Biosciences Sequel instrument.

Reads were assembled into scaffolds using SPAdes 3.9.0 (Bankevich et al. 2012; Vasilinetc et al. 2015); in the case of *E. sesbaniae*, long reads were corrected and trimmed using Canu 1.7.1 (Koren et al. 2017) prior to assembly. Scaffolds returned by SPAdes were parsed to remove those with less than 20x coverage or with a length below 200 nucleotides. Using FastANI (Jain et al. 2018), one-way average nucleotide identity (ANI) values of each assembly were calculated against 887 alpha-proteobacterial genomes available through the National Center for Biotechnological Information (NCBI) with an assembly level of complete or chromosome. Based on the FastANI output, each draft genome assembly was further scaffolded using MeDuSa (Bosi et al. 2015) and the reference genomes listed in Supplementary Table S1. For most assemblies, scaffolds under 1 kb in length were discarded. The exception was for *S. sesbaniae*, for which case scaffolds under 10 kb were discarded. Genome assemblies were annotated using Prokka 1.12-beta (Seemann 2014), annotating coding regions with Prodigal (Hyatt et al. 2010), tRNA with Aragon (Laslett & Canbäck 2004), rRNA with Barrnap (github.com/tseemann/barrnap), and ncRNA with Infernal (Kolbe & Eddy 2011) and Rfam (Kalvari et al. 2018).

Species Phylogenetic Analyses

All *Ensifer* (and *Sinorhizobium*) genomes were downloaded from the NCBI Genome Database regardless of assembly level. Strains that either i) lacked a RefSeq assembly, ii) had genome sizes < 1 Mb, or iii) appeared to not belong to the *Ensifer* clade based on preliminary phylogenetic analyses were discarded, leaving a final set of 157 strains (Supplementary Dataset S1). Eight complete *Rhizobium* genomes (Supplementary Dataset S2) were downloaded to serve as an outgroup. Genomes were reannotated with prokka to ensure consistent annotation. All genomes were downloaded on 12 November 2018, and associated metadata is available as Supplementary Dataset S3.

To construct an unrooted, core gene phylogeny, the pangenome of the 157 *Ensifer* strains was calculated using Roary 3.11.3 (Page et al. 2015) with a percent identify threshold of 70%. As part of the running of Roary, the nucleotide sequences of the 1,049 core genes (identified as those found in at least 99% of the genomes; Supplementary Dataset S3) were individually aligned with PRANK (Löytynoja 2014) and the alignments concatenated. The concatenated alignment was trimmed using TRIMAL 1.2rev59 (Capella-Gutiérrez et al. 2009) with the automated1 option, and used to construct a maximum likelihood phylogeny (the bootstrap best tree following 100 bootstrap replicates, as determined by the extended majority-rule consensus tree criterion) using RAxML 8.2.9 (Stamatakis 2014) with the GTRCAT model as recommended (cme.h-its.org/exelixis/resource/download/NewManual.pdf). All phylogenies prepared in this study were visualized with the online iTOL webserver (Letunic & Bork 2016).

To construct a rooted phylogeny, the AMPHORA2 pipeline (Wu & Scott 2012) was used to identify 31 highly conserved bacterial proteins in each *Ensifer* and *Rhizobium* proteome, based on the 31 hidden Markov models (HMMs) that come with AMPHORA2 and HMMER 3.1b2 (Eddy 2009). Customs Perl scripts were then used to remove proteins that were

either found in less than 95% of genomes or was found in multi-copy in at least one genome, leaving a set of 30 proteins (Frr, InfC, NusA, Pgk, PyrG, RplA, RplB, RplC, RplD, RplE, RplF, RplK, RplL, RplM, RplN, RplP, RplS, RplT, RpmA, RpoB, RpsB, RpsC, RpsE, RpsI, RpsJ, RpsK, RpsM, RpsS, SmpB, Tsf). Orthologous groups were aligned using MAFFT 7.310 (Katoh & Standley 2013) with the localpair option, following which the alignments were trimmed using TRIMAL 1.2rev59 with the automated1 option. Alignments were concatenated and used to construct a maximum likelihood phylogeny (the bootstrap best tree following 304 bootstrap replicates, as determined by the extended majority-rule consensus tree criterion) using RAxML with the PROTGAMMAJTTDCMUT model. This model was chosen as a preliminary run using RAxML with the automatic model selection indicated that the best scoring tree was obtained with the selected model.

ANI and AAI Calculations

Pairwise ANI values were calculated for all *Ensifer* strains using FastANI (Jain et al. 2018) with default parameters; a value of 78% was used in cases where no value was returned by FastANI. Pairwise average amino acid identity (AAI) values were calculated with the compareM workflow (github.com/dparks1134/CompareM). Results were visualized and clustered using the heatmap.2 function of the gplots package in R (Warnes et al. 2016), with average linkage and Pearson correlation distances.

Pangenome Calculation

All proteins of the reannotated *Ensifer* strains were clustered into orthologous groups using CD-HIT 4.6 (Li & Godzik 2006) with a percent identity threshold of 70% and an alignment length of 80% of the longer protein. The output was used to determine core and accessory genomes using a prevalence threshold of 90% as many of the genomes were draft genomes.

Gene accumulation curves were produced using the specaccum function of the vegan package of R (Oksanen et al. 2018), with the random method and 500 permutations. Principal component analysis (PCA) was performed with the prcomp function of R, and was visualized with the autoplot function the ggplot2 package (Wickham 2016).

Identification and Phylogenetic Analysis of Common Nod, Nif, and Rep Proteins

The proteomes were collected for the 157 *Ensifer* strains, as well as all strains from the genera *Rhizobium*, *Neorhizobium*, *Agrobacterium*, *Mesorhizobium*, and *Ochrobactrum* with an assembly status of Complete or Chromosome (Supplementary Dataset S4). Additionally, the seed alignments for the HMMs of the nodulation proteins NodA (TIGR04245), NodB (TIGR04243), and NodC (TIGR04242), the nitrogenase proteins NifH (TIGR01287), NifD (TIGR01282), NifK (TIGR01286), and the replicon partitioning proteins RepA (TIGR03453), and RepB (TIGR03454) were downloaded from TIGRFAM (Haft et al. 2013). Seed alignments were converted into HMMs with the HMMBUILD function of HMMER 3.1b2 (Eddy 2009). Each HMM was searched against the complete set of proteins from all 157 reannotated *Sinorhizobium* and *Ensifer* strains using the HMMSEARCH function of HMMER. The amino acid sequences for each hit (regardless of e-value) were collected. Each set of sequences was searched against a HMM database containing all 21,200 HMMs from the Pfam (Finn et al. 2016) and TIGRFAM databases using the HMMSCAN function of HMMER, and the top scoring HMM hit for each query protein was identified. Proteins were annotated as NodA, NodB, NodC, NifH, NifD, NifK, RepA, or RepB according to Supplementary Table S2.

The *nodA*, *nodB*, and *nodC* genes are generally found as an operon. Thus, the NodA, NodB, and NodC proteins were putatively associated to operons based on identifying proteins that are encoded by adjacent genes in their respective genomes; orphan proteins not encoded by adjacent genes were discarded as the subsequent phylogenetic analysis was based on

concatenated NodA, NodB, and NodC alignments. Each set of orthologs were aligned using MAFFT with the localpair option, and alignments trimmed using TRIMAL and the automated1 algorithm. Alignments were concatenated so as to combine alignments for proteins encoded by adjacent genes, producing a NodABC alignment. The same procedure was followed to produce NifHDK and RepAB alignments. Maximum likelihood phylogenies were built on the basis of each combined alignment using RAxML with the PROTGAMMAJTT (NodABC, NifHDK) or the PROTGAMMALG (RepAB) models. These models were chosen as preliminary runs using RAxML with the automatic model selection indicated that the best scoring trees were obtained with the selected models. The final phylogenies are the bootstrap best trees following 352 bootstrap replicates, as determined by the extended majority-rule consensus tree criterion.

Plant Assays

Phaseolus vulgaris (var. TopCrop, Mangani Sementi, Italy) seeds were surface sterilized in 2.5% $HgCl_2$ solution for two minutes and washed five times with sterile water. Seeds were germinated in the dark at 23°C, following which seedlings were placed in sterile polypropylene jars containing vermiculite:perlite (1:1) and nitrogen-free Fåhraeus medium, and grown at 23°C with a 12 hour photoperiod (100 µE m^{-2} s^{-1}). One-week old plantlets were inoculated with 100 µL of the appropriate rhizobium strain (suspended in 0.9% NaCl at an OD_{600} of 1); five plants were inoculated per strain and then grown for four weeks at 23°C with a 12 hour photoperiod (100 µE m^{-2} s^{-1}). Plant growth assays were repeated three independent times. Nodules were collected and surface sterilized as described elsewhere (Checcucci et al. 2016), crushed in sterile 0.9% NaCl solution, and serial dilutions were plated on TY agar plates and incubated at 30˚C for two days. PCR amplification of the 16S rRNA gene was performed using crude lysates from single colonies recovered from root nodules, as in (Barzanti et al. 2007). Sequencing of the PCR amplified 16S rRNA gene was performed from both the 27f and 1495r

primers using BrilliantDye™ Terminator Cycle Sequencing chemistry (Nimagen, Nijmegen, The Netherlands) on a 3730xl DNA Analyzer (ThermoFisher Scientific, Waltham, MA, USA).

Phenotype MicroArray™

Phenotype MicroArray™ experiments using Biolog plates PM1 and PM2A (carbon sources), PM9 (osmolytes), and PM10 (pH) were performed as described previously (Biondi et al. 2009). Data were collected over 96 hours with an OmniLog™ instrument. Data analysis was performed with DuctApe (Galardini et al. 2014). Activity index (AV) values were calculated following subtraction of the blank well from the experimental wells. Growth with each compound was evaluated with AV values from 0 (no growth) to 9 (maximal growth), following an elbow test calculation. Phenotype MicroArray™ experiments were performed once as results for these experiments are highly repeatable (Johnson et al. 2008; Bochner et al. 2010; Dunkley et al. 2019).

Biofilm Assays

Overnight cultures of strains grown in TY and LB (10 g L^{-1} tryptone, 5 g L^{-1} yeast extract, 5 g L^{-1} NaCl) media were diluted to an OD_{600} of 0.02 in fresh media, and six replicates of 100 µL aliquots were transferred to a 96-well microplate. Plates were incubated at 30˚C for 24 hours, after which the OD_{600} was measured with a Tecan Infinite 200 PRO (Switzerland). Each well was then stained with 30 µL of a filtered 0.1% (w/v) crystal violet solution for 10 minutes, and then the medium containing the planktonic cells was gently removed from the wells. Next, the wells were rinsed three times with 200 µL of phosphate-buffered saline (PBS; 0.1 M, pH 7.4) and allowed to dry for 15 min. One hundred µL of 95% (v/v) ethanol was added to each well and then incubated for 15 minutes at room temperature. The OD_{540} of each well was measured

(Rinaudi & González 2009), and biofilm production reported as the ratio of the OD_{540}/OD_{600} ratio. Biofilm assays consisted of six replicates, and were performed two independent times.

Growth Curves

Overnight cultures of each strain were grown in the same medium to be used for the growth curve. For minimal media, either 0.2% (w/v) of glucose or succinate was added as the carbon source. Cultures were diluted to an OD_{600} of 0.05 in the same media, and triplicate 150 µL aliquots were added to a 96-well microplate. Microplates were incubated without shaking at 30˚C or 37˚C in a Tecan Infinite 200 PRO, with OD_{600} readings taken every hour for 48 hours. Growth rates were evaluated over two-hour windows during the exponential growth phase. All growth curves were performed in triplicate and repeated two independent times.

To evaluate bacterial growth when provided root exudates as a nitrogen source, root exudates were produced from *Medicago sativa* cv. Maraviglia as described elsewhere (Checcucci et al. 2017). Single bacterial colonies from TY plates were resuspended in a 0.9% NaCl solution to an OD_{600} of 0.5. Then, each well of a 96-well microplate was inoculated with 5 µL of culture, 75 µL of nitrogen-free M9 with 0.2% (w/v) succinate as a carbon source, and 20 µL of root exudate as a nitrogen-source as done previously (Checcucci et al. 2017). Triplicates were performed for each strain. Microplates were incubated without shaking at 30˚C in a Tecan Infinite 200 PRO, with OD_{600} readings taken every hour for 48 hours. Growth rates were determined as described above.

Data Availability.

Genome assemblies were deposited to NCBI under the BioProject accession PRJNA622509. Scripts to repeat the computational analyses reported in this study are available at: github.com/diCenzo-GC/Ensifer_phylogenomics.

RESULTS

Genome sequencing of four *Rhizobiaceae* strains

Draft genomes of *E. morelense* Lc04, *E. morelense* Lc18, *E. sesbaniae* CCBAU 65729, and *E. psoraleae* CCBAU 65732 (Wang et al. 1999, 2013) were generated to increase the species diversity available for our analyses. Summary statistics of the assemblies are provided in Supplementary Table S3. The genome sequences confirmed the presence of nodulation and nitrogen-fixing genes in *E. morelense* Lc18, *E. sesbaniae* CCBAU 65729, and *E. psoraleae* CCBAU 65732, while these genes appeared absent in the *E. morelense* Lc04 assembly. Strains Lc04, CCBAU 65729, and CCBAU 65732 were confirmed to belong to the genus *Ensifer*, as one-way ANI comparisons revealed that the most similar alpha-proteobacterial genomes were from the genus *Ensifer*. However, the genome of strain Lc18 was most similar to genomes from the genera *Rhizobium* and *Agrobacterium*, consistent with an earlier 16S rRNA gene restriction fragment length polymorphism analysis (34). Thus, we propose renaming *E. morelense* Lc18 to *Rhizobium* sp. Lc18. As this strain does not belong to the genus *Ensifer*, it was excluded from further analyses.

Symbiotic and non-symbiotic *Ensifer* strains segregate phylogenetically

An unrooted, core gene phylogeny of 157 *Ensifer* strains was prepared to evaluate the phylogenetic relationships between the symbiotic and non-symbiotic strains (Figure 1). A rooted phylogeny based on a multi-locus sequence analysis was also prepared (Supplementary Figure S1). Each of the 157 strains were annotated as symbiotic or non-symbiotic based on the presence of the common *nodABC* nodulation genes and the *nifHDK* nitrogenase genes. Consistent with previous work (Garrido-Oter et al. 2018), both phylogenies revealed a clear division of the symbiotic and non-symbiotic strains into two well-defined clades. However, a few exceptions were noted. *E. sesbaniae* was found within the non-symbiotic clade; however,

E. sesbaniae was reported to be a symbiont of legumes such as *Phaseolus vulgaris* (Wang et al. 2013), and the ability of *E. sesbaniae* to nodulate *P. vulgaris* was confirmed in this study (Supplementary Figure S2). Similarly, at least one of the six symbiotic proteins were not detected in five strains of the symbiotic group, although we cannot rule out that these are false negatives due to incomplete genome assemblies or genome assembly errors. ANI (genospecies threshold: 95%) and AAI (genospecies threshold: 96%) calculations suggested the presence of 12 and 20 genospecies within the non-symbiotic and symbiotic groups, respectively (Figure 1, Supplementary Figures S3, S4), confirming that the non-symbiotic clade was not an artefact of low species diversity. Thus, we conclude that the genus *Ensifer* consists of two well-defined clusters, each consisting predominately of either symbiotic or non-symbiotic strains.

SNF likely arose multiple times within the genus *Ensifer*

A possible explanation for the phylogenetic segregation of SNF within the genus *Ensifer* was that the symbiotic genes were gained once through a single HGT event. To test this hypothesis, the phylogenetic relationships of the NodABC and NifHDK proteins of the order Rhizobiales were examined (Figures 2A, 2B). SNF genes are situated on megaplasmids in the genus *Ensifer*; thus, a phylogeny of RepAB partitioning proteins of the order Rhizobiales was prepared as a proxy of the evolutionary relationships among the symbiotic megaplasmids (Figure 2C). We predicted that the NodABC, NifHDK, and RepAB proteins of the genus *Ensifer* would form a single, monophyletic clade in each phylogeny, if the above hypothesis were true. This was not observed. Instead, all three phylogenies were inconsistent with a single origin of SNF within the genus *Ensifer* as the *Ensifer* strains were predominantly split into three clades: i) *E. meliloti* and *E. medicae*, ii) *E. fredii* and related strains, and iii) *E. americanum* and related strains. As the same clades are observed in the species tree (Figure 2D), this observation suggests SNF was independently acquired through HGT in each clade.

The relationships between the SNF genes of the remaining *Ensifer* species (e.g., *E. aridi* and *E. psoraleae*) was not clear; however, the most parsimonious solution was that there were three additional acquisitions of SNF (Figure 2D). Overall, the phylogenetic analyses of the NodABC, NifHDK, and RepAB proteins support the hypothesis that there were multiple, independent acquisition of symbiosis genes (hence SNF) by lineages within the genus *Ensifer*; however, the gain (and/or maintenance) of symbiosis genes preferentially occurred within one monophyletic group of species.

The genomic features of the symbiotic and non-symbiotic clades differ

The pangenome of the 157 *Ensifer* strains was calculated to evaluate if there were global genomic differences between the symbiotic and non-symbiotic clades. Both clades had open pangenomes (Supplementary Figure S5). A PCA based on gene presence/absence revealed a clear separation of the two clades (Figure 3A), suggesting a divergence of the pangenomes of these clades. The symbiotic clade was sub-divided into two groups along the second component of the PCA (Figure 3A), which may suggest further levels of genomic separation. 2,130 genes were found in the core genomes of both clades, while 20% (542 genes) and 40% (1,377 genes) of the core genomes of the symbiotic and non-symbiotic clades, respectively, were absent from the core genome of the other clade; of these, about a third were completely absent from the other clade's pangenome (Figure 3C). Of the 14,514 accessory genes (defined as genes found in at least 10% of at least one clade, excluding the 2,130 *Ensifer* core genes), only 2,352 (16%) were found in the pangenomes of both the symbiotic and non-symbiotic clades. Moreover, a statistically significant difference (Wilcoxon rank-sum test, $p < 0.0001$) in the genome sizes of the two clades was observed (Figure 3B); strains of the non-symbiotic clade carried 325 more genes, on average, than strains of the symbiotic clade (median difference of 470). Finally, based on the limited number of strains with finished genomes, strains of the symbiotic clade appear

to generally have three copies of the rRNA operon whereas strains of the non-symbiotic clade appear to have a norm of five copies of their rRNA operon. Together, these multiple lines of data are consistent with there being a broad genomic divergence of the symbiotic and non-symbiotic clades of the genus *Ensifer*.

Phenotypic features of the symbiotic and non-symbiotic clades differ

A subset of ten strains (Table 1), five each from the symbiotic and non-symbiotic clades, were subjected to a panel of assays to investigate how phenotypes vary across the genus *Ensifer*. These ten strains were chosen so as to provide broad phylogenetic coverage of the genus, while excluding strains for which extensive phenotypic characterizations have been previously published. No statistically significant differences were observed in the ability of members of the two clades to form biofilm (Supplementary Figure S6, Supplementary Table S4). However, the tested strains clearly differed in their ability to grow in LB media; whereas the tested strains of the non-symbiotic clade displayed robust growth in LB, the tested strains of the symbiotic clade largely failed to grow (Figure 4A). Tested strains of the non-symbiotic clade also displayed a slightly faster specific growth rate, on average, than the tested strains of the symbiotic clade in TY media (non-symbiotic clade: 0.54 ± 0.03 h^{-1}; symbiotic clade: 0.44 ± 0.08 h^{-1}; $p = 0.03$ from an ANOVA followed by Tukey's test; Supplementary Figure S7A, Supplementary Table S5). On the other hand, tested strains of the symbiotic clade were, on average, better able to withstand heat stress (37˚C) in TY media (Figure 4B). No statistically significant difference in the average ability of the tested strains of the two clades to grow in minimal media with succinate or glucose as a carbon source, or with *M. sativa* (a symbiotic partner of *E. meliloti* and *E. medicae*) root exudates as a nitrogen source, was detected (Supplementary Figure S7, Supplementary Table S5).

The phenotypic properties of the genus *Ensifer* were further examined through

evaluating the ability of the same ten strains to catabolize 190 carbon sources, and to grow in 96 osmolyte and 96 pH conditions, through the use of Biolog Phenotype MicroArrays™. Clustering the strains based on growth properties largely separated the tested strains of the symbiotic clade and non-symbiotic clade into distinct groups (Figure 5, Supplementary Figure S8). The exception was *E. sojae*, which formed its own intermediate group in the phenotype data. To aid in identifying which conditions best separate the tested strains of the symbiotic clade (including *E. sojae*) from the tested strains of the non-symbiotic clade, a linear discriminant analysis (LDA) was run over the AV values summarizing growth in each condition (Supplementary Dataset S5). In general, tested strains from the non-symbiotic clade better tolerated high pH (pH 9.0 to 9.5) than did the tested strains from the symbiotic cluster. In contrast, tested strains of the symbiotic clade had better tolerance to low pH conditions (pH 3.5 to 4.5). In addition, the tested strains from the two clades clearly differed in their overall metabolic abilities with tested strains of the non-symbiotic clade generally having a broader metabolic capacity than those of the symbiotic clade (Supplementary Table S6). Whereas tested strains of the symbiotic clade displayed robust growth on 65 carbon sources on average, the tested strains of the non-symbiotic clade grew on an average of 81 carbon sources ($p < 0.05$, Student's *t*-test). Overall, these initial experiments provide support for the hypothesis that a variety of phenotypes, not just the ability to nodulate legumes, differ between the symbiotic and non-symbiotic clades of the genus *Ensifer*.

DISCUSSION

We investigated the evolution of SNF within the genus *Ensifer*, which includes a mix of nitrogen-fixing and non-nitrogen-fixing bacteria, as a model for the evolution of inter-kingdom mutualisms. Our results indicate that, despite SNF having likely evolved multiple times within the genus *Ensifer*, the symbiotic and non-symbiotic strains are largely separated into two

phylogenetic clades reminiscent of the general division of pathogenic and environmental strains between the genera *Burkholderia* and *Paraburkholderia* (Sawana et al. 2014). While it is possible that this result will fail to remain true as more genome sequences are published, we believe the result to be robust as the current set of genome sequences are of strains isolated from diverse legumes and other diverse environments including the rhizosphere, pristine caves, and an abandoned mine. In addition to the prevalence of SNF, the two clades differed with respect to their genomic composition (pangenome content and genome size) and phenotypic properties (metabolic capacity and stress tolerance) based on initial tests of a subset of strains. There have been several revisions to the taxonomy of the genus *Ensifer*. Recently, a genome-based approach was proposed to standardize bacterial taxonomy (Parks et al. 2018) that splits the genus *Ensifer* into two genera: *Sinorhizobium* and *Ensifer*. The symbiotic and non-symbiotic clades identified here correspond with the genera *Sinorhizobium* and *Ensifer*, respectively, supporting the proposal to divide the genus *Ensifer* into two genera.

Our analyses revealed a complex evolutionary history of SNF within the genus *Ensifer*. In addition to SNF emerging a predicted six times or more, we detected possible losses of SNF and allele switches. Between one and six of the NodABC and NifHDK proteins were not detected in five of the strains in the symbiotic clade (Figure 1). While this may be indicative of multiple losses of symbiosis, we cannot rule out that these are false negatives due to incomplete genome assemblies or genome assembly errors; five of the six genomes were draft genomes, and the one strain with a complete genome (*E. meliloti* M162) can nodulate 10 of 27 tested *Medicago truncatula* genotypes suggesting it does contain *nod* and *nif* genes (Sugawara et al. 2013). Based on the RepAB phylogeny (Figure 2), the symbiotic megaplasmid of *E. arboris* likely shares common ancestry with the symbiotic megaplasmids of the sister species *E. meliloti* and *E. medicae*. Yet, the nodulation and nitrogen fixation genes appeared distinct. Thus, we hypothesize that there was a recent replacement of the symbiotic genes in *E. arboris*,

or alternatively, in *E. meliloti* and *E. medicae*. This hypothesis is supported by the observation that the host ranges of these rhizobia differ; unlike *E. meliloti* and *E. medicae*, *E. arboris* does not nodulate plants of the genus *Medicago* (Zhang et al. 1991).

The reason for the phylogenetic bias in the evolution of SNF within the genus *Ensifer* remains unclear, especially considering that strains from both clades are plant-associated (Bai et al. 2015). One hypothesis is that each clade occupies distinct niches within the soil and plant-associated environments. Indeed, analysis of a subset of strains suggested that species of the non-symbiotic clade have a broader metabolic capacity (Supplementary Table S6), which corresponded to a larger average genome size (Figure 3B), is consistent with these species being more capable of adapting to fluctuating nutritional environments. This is further supported by the apparently higher number of rRNA operons in strains of the non-symbiotic clade, which is generally thought to allow bacteria to more quickly respond to changing nutrient conditions (Stevenson & Schmidt 2004; Roller et al. 2016). Moreover, the non-symbiotic clade could be differentiated from the symbiotic clade based on its pangenome content (Figure 3A), which leads us to hypothesize that strains of these clades acquire genes from distinct gene pools, further supporting the hypothesis that they belong to distinct gene-cohesive groups and ecological niches. This hypothesis may then be interpreted in the framework of the stable ecotype model (*sensu* Cohan (Cohan 2006)), where the symbiotic and non-symbitic clades represent two, ecologically distinct and monophyletic groups and where periodic selection events (e.g. fitness for SNF) are recurrent.

An alternate, but not mutually exclusive, hypothesis is that the symbiotic *Ensifer* clade contains "facilitator" genes required to support SNF, similar to the theory that ancestral legumes contained a genetic "predisposition" necessary for the eventual evolution of rhizobium symbioses (Werner et al. 2014; Soltis et al. 1995; Doyle 2011). Conversely, evolution of SNF may also require the absence of "inhibitor" genes, such as the absence of virulence factors

(Marchetti et al. 2010). As we did not evaluate cause-and-effect relationships, our dataset does not definitely address these hypotheses. However, we observed numerous genotypic and likely phenotypic differences between the symbiotic and non-symbiotic clade, providing some support for these hypotheses. For example, the tested strains of the symbiotic clade appeared to have higher tolerance to low pH (Supplementary Figure S8, Supplementary Dataset S5), which is notable as the curled root hair is an acidic environment (Hawkins et al. 2017). At the genomic level, 231 of the core genes of the symbiotic clade were absent from the pangenome of the non-symbiotic clade and thus are good candidates as possible facilitators and follow-up studies. However, facilitators and inhibitors could also take the form of polymorphisms within highly conserved genes, as shown for *bacA* and the *Sinorhizobium – Medicago* symbiosis (diCenzo et al. 2017).

In summary, we show that the legume symbionts and non-symbionts of the genus *Ensifer* are largely segregated into two phylogenetically distinct clades that differ in their genomic and phenotypic properties. We suggest that these observations, which follow the guidelines recently reported for rhizobia and agrobacteria (de Lajudie et al. 2019), support the division of the genus into two genera: *Ensifer* for the non-symbiotic clade and *Sinorhizobium* for the symbiotic clade. However, formal descriptions and publication of the genera in the International Journal of Systematic and Evolutionary Microbiology (IJSEM) are still required. We also provide evidence that SNF genes were likely acquired several independent times within this genus, but predominately within one monophyletic clade. These observations suggest that the presence or absence of other genomic features ("facilitators" or "inhibitors") aside from the core symbiotic genes could be required for the establishment of an effective symbiosis. This suggestion is supported by the ability to differentiate the strains of the two clades based on their pangenome content and, at least for the tested subset, their phenotypic properties. However, as cause-and-effect relationships were not examined, follow-up study is

required to more directly test this facilitators hypothesis.

ACKNOWLEDGEMENTS

We are grateful to E. Martinez-Romero for providing strains *E. morelense* Lc04 and *E. morelense* Lc18, to E. Mullins (Teagasc, MTA2018233) for *E. adhaerens* OV14, to C.-F. Tian for *S. fredii* NGR 234, and to L. Dziewit for *Ensifer* sp. M14. AM was supported by the Fondazione Cassa di Risparmio di Firenze, grant n. 18204, 2017.0719, by the "MICRO4Legumes" grant (Italian Ministry of Agriculture), and by the grant "Dipartimento di Eccellenza 2018-2022" from the Italian Ministry of Education, University and Research (MIUR). LC was supported by the MICRO4Legumes grant (Italian Ministry of Agriculture). GCD was supported by a postdoctoral fellowship from the Natural Science and Engineering Research Council of Canada, and funding from Queen's University.

REFERENCES

Bai Y et al. 2015. Functional overlap of the Arabidopsis leaf and root microbiota. Nature. 528:364–369.

Bankevich A et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19:455–477.

Barcellos FG, Menna P, Batista JS da S, Hungria M. 2007. Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous diazotrophs *Sinorhizobium* (*Ensifer*) *fredii* and *Bradyrhizobium elkanii* in a Brazilian savannah soil. Appl Environ Microbiol. 73:2635–2643.

Barzanti R et al. 2007. Isolation and characterization of endophytic bacteria from the nickel hyperaccumulator plant *Alyssum bertolonii*. Microb Ecol. 53:306–316.

Biondi EG et al. 2009. Metabolic capacity of *Sinorhizobium* (*Ensifer*) *meliloti* strains as determined by Phenotype MicroArray analysis. Appl Environ Microbiol. 75:5396–5404.

Bochner B, Gomez V, Ziman M, Yang S, Brown SD. 2010. Phenotype microarray profiling

of *Zymomonas mobilis* ZM4. Appl Biochem Biotechnol. 161:116-123.

Bosi E et al. 2015. MeDuSa: a multi-draft based scaffolder. Bioinformatics. 31:2443–2451.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Casida LE. 1982. *Ensifer adhaerens* gen. nov., sp. nov.: a bacterial predator of bacteria in soil. Int J Syst Evol Microbiol. 32:339–345.

Checcucci A et al. 2016. Mixed nodule infection in *Sinorhizobium meliloti-Medicago sativa* symbiosis suggest the presence of cheating behavior. Front Plant Sci. 7:835.

Checcucci A et al. 2017. Role and regulation of ACC deaminase gene in *Sinorhizobium meliloti*: is it a symbiotic, rhizospheric or endophytic gene? Front. Genet. 8:6.

Checcucci A, diCenzo GC, Perrin E, Bazzicalupo M, Mengoni A. 2019. Genomic diversity and evolution of rhizobia. In: Microbial Diversity in the Genomic Era. Das, S & Dash, HR, editors. Academic Press pp. 37–46.

Cohan FM. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Phil Trans R Soc B. 361:1985–1996.

Cowie A et al. 2006. An integrated approach to functional genomics: construction of a novel reporter gene fusion library for *Sinorhizobium meliloti*. Appl Environ Microbiol. 72:7156–7167.

de Lajudie PM et al. 2019. Minimal standards for the description of new genera and species of rhizobia and agrobacteria. Int J Syst Evol Microbiol. 69:1852–1863.

diCenzo GC, Zamani M, et al. 2019. Multi-disciplinary approaches for studying rhizobium - legume symbioses. Can J Microbiol. 65:1–33.

diCenzo GC, Tesi M, Pfau T, Mengoni A, Fondi M. 2019. A Virtual Nodule Environment (ViNE) for modelling the inter-kingdom metabolic integration during symbiotic nitrogen fixation. bioRxiv. 1:765271.

diCenzo GC, Zamani M, Ludwig HN, Finan TM. 2017. Heterologous complementation reveals a specialized activity for BacA in the *Medicago-Sinorhizobium meliloti* symbiosis. Mol Plant Microbe Interact. 30:312–324.

Doin de Moura GG, Remigi P, Masson-Boivin C, Capela D. 2020. Experimental evolution of legume symbionts: what have we learnt? Genes. 11:339.

Douglas AE. 2014. Symbiosis as a general principle in eukaryotic evolution. Cold Spring Harb Perspect Biol. 6:a016113.

Doyle JJ. 2011. Phylogenetic perspectives on the origins of nodulation. Mol Plant Microbe Interact. 24:1289–1295.

Drewniak L, Matlakowska R, Sklodowska A. 2008. Arsenite and arsenate metabolism of *Sinorhizobium* sp. M14 living in the extreme environment of the Zloty Stok gold mine. Geomicrobiol J. 25:363–370.

Dunkley EJ, Chalmers JD, Cho S, Finn TJ, Patrick WM. 2019. Assessment of Phenotype Microarray plates for rapid and high-throughput analysis of collateral sensitivity networks. PLOS One. 14:e0219879.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23:205–211.

Finan TM, Kunkel B, De Vos GF, Signer ER. 1986. Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. J Bacteriol. 167:66–72.

Finn RD et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.

Galardini M et al. 2014. DuctApe: a suite for the analysis and correlation of genomic and OmniLog™ Phenotype Microarray data. Genomics. 103:1–10.

Garrido-Oter R et al. 2018. Modular traits of the Rhizobiales root microbiota and their

evolutionary relationship with symbiotic rhizobia. Cell Host Microbe. 24:155–167

Geddes BA, Kearsley J, Morton R, diCenzo GC, Finan TM. 2020. The genomes of rhizobia. In: Advances in Botanical Research. Frendo, P, Frugier, F, & Masson-Boivin, C, editors. Regulation of Nitrogen-Fixing Symbioses in Legumes Vol. 94 Academic Press pp. 213–249.

Gerhart J, Kirschner M. 2007. The theory of facilitated variation. Proc Natl Acad Sci USA. 104:8582–8589.

Gordon SA, Weber RP. 1951. Colorimetric estimation of indoacetic acid. Plant Physiol. 26:192–195.

Griesmann M et al. 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science. 361.

Haag AF et al. 2013. Molecular insights into bacteroid development during *Rhizobium*-legume symbiosis. FEMS Microbiol Rev. 37:364–383.

Haft DH et al. 2013. TIGRFAMs and genome properties in 2013. Nucleic Acid Res. 41:D387–D395.

Haskett TL et al. 2016. Assembly and transfer of tripartite integrative and conjugative genetic elements. Proc Natl Acad Sci USA. 113:12268–12273.

Hawkins JP, Geddes BA, Oresnik IJ. 2017. Succinoglycan production contributes to acidic pH tolerance in *Sinorhizobium meliloti* Rm1021. Mol Plant Microbe Interact. 30:1009–1019.

Heath KD, Tiffin P. 2007. Context dependence in the coevolution of plant and rhizobial mutualists. Proc Biol Sci. 274:1905–1912.

Hooykaas PJ, Snijdewint FG, Schilperoort RA. 1982. Identification of the Sym plasmid of *Rhizobium leguminosarum* strain 1001 and its transfer to and expression in other rhizobia and *Agrobacterium tumefaciens*. Plasmid. 8:73–82.

Howieson JG, Ewing MA. 1986. Acid tolerance in the *Rhizobium meliloti* - *Medicago*

symbiosis. Aust J Agric Res. 37:55–64.

Hyatt D et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11:119.

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 9:7200.

Johnson DA et al. 2008. High-throughput phenotypic characerization of *Pseudomonas aeruginosa* membrane transport genes. PLOS Genet. 4:e1000211.

Kalvari I et al. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46:D335–D342.

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

Kolbe DL, Eddy SR. 2011. Fast filtering for RNA homology search. Bioinformatics. 27:3102–3109.

Koren S et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736.

Laslett D, Canbäck B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11–16.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44:W242–5.

Li QQ et al. 2011. *Ensifer sojae* sp. nov., isolated from root nodules of *Glycine max* grown in saline-alkaline soils. Int J Syst Evol Microbiol. 61:1981–1988.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659.

Long SR. 2001. Genes and signals in the rhizobium-legume symbiosis. Plant Physiol. 125:69–

72.

López-García P, Eme L, Moreira D. 2017. Symbiosis in eukaryotic evolution. J Theor Biol. 434:20–33.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In: Methods in Molecular Biology (Methods and Protocols). Russell, D, editor. Vol. 1079 Humana Press: Totowa, NJ pp. 155–170.

Marchetti M et al. 2010. Experimental evolution of a plant pathogen into a legume symbiont. PLOS Biol. 12:e1000280.

Martens M et al. 2007. Multilocus sequence analysis of *Ensifer* and related taxa. Int J Syst Evol Microbiol. 57:489–503.

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

Masson-Boivin C, Sachs JL. 2018. Symbiotic nitrogen fixation by rhizobia—the roots of a success story. Curr Opin Plant Biol. 44:7–15.

Nandasena KG, O'Hara GW, Tiwari RP, Sezmiş E, Howieson JG. 2007. *In situ* lateral transfer of symbiosis islands results in rapid evolution of diverse competitive strains of mesorhizobia suboptimal in symbiotic nitrogen fixation on the pasture legume *Biserrula pelecinus* L. Environ Microbiol. 9:2496–2511.

Oksanen J et al. 2018. *vegan: Community Ecology Package. R package version 2.5-3. https://CRAN.R-project.org/package=vegan*.

Oldroyd GED, Murray JD, Poole PS, Downie JA. 2011. The rules of engagement in the legume-rhizobial symbiosis. Annu Rev Genet. 45:119–144.

Ormeño-Orrillo E et al. 2015. Taxonomy of rhizobia and agrobacteria from the *Rhizobiaceae* family in light of genomics. Syst Appl Microbiol. 38:287–291.

Page AJ et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics.

31:3691–3693.

Parks DH et al. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 36:996–1004.

Pérez Carrascal OM et al. 2016. Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing *Rhizobium* species associated with *Phaseolus vulgaris*. Environ Microbiol. 18:2660–2676.

Perrin E et al. 2015. Genomes analysis and bacteria identification: The use of overlapping genes as molecular markers. J Microbiol Methods. 117:108–112.

Pfau T et al. 2018. The intertwined metabolism during symbiotic nitrogen fixation elucidated by metabolic modelling. Sci Rep. 8:12504.

Poole P, Ramachandran V, Terpolilli J. 2018. Rhizobia: from saprophytes to endosymbionts. Nat Rev Microbiol. 180:1691.

Rinaudi LV, González JE. 2009. The low-molecular-weight fraction of exopolysaccharide II from *Sinorhizobium meliloti* is a crucial determinant of biofilm formation. J Bacteriol. 191:7216–7224.

Rogel MA, Hernández-Lucas I, Kuykendall LD, Balkwill DL, Martinez-Romero E. 2001. Nitrogen-fixing nodules with *Ensifer adhaerens* harboring *Rhizobium tropici* symbiotic plasmids. Appl Environ Microbiol. 67:3264–3268.

Roller BRK, Stoddard SF, Schmidt TM. 2016. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. Nat Microbiol. 1:16160.

Sanjuán J. 2016. Towards the minimal nitrogen-fixing symbiotic genome. Environ Microbiol. 18:2292–2294.

Sawana A, Adeolu M, Gupta RS. 2014. Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen.

nov. harboring environmental species. Front Genet. 5:429.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30:2068–2069.

Soltis DE et al. 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proc Natl Acad Sci USA. 92:2647–2651.

Sørensen MES et al. 2019. The role of exploitation in the establishment of mutualistic microbial symbioses. FEMS Microbiol Lett. 366.

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

Stevenson BS, Schmidt TM. 2004. Life history implications of rRNA gene copy number in *Escherichia coli*. Appl Environ Microbiol. 70:6670–6677.

Sugawara M et al. 2013. Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. Genome Biol. 14:R17.

Sullivan JT, Patrick HN, Lowther WL, Scott DB, Ronson CW. 1995. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. Proc Natl Acad Sci USA. 92:8985–8989.

Tian CF, Young JPW. 2019. Genomics and evolution of rhizobia. In: Ecology and Evolution of Rhizobia: Principles and Applications. Wang, ET, Tian, CF, Chen, WF, Young, JPW, & Chen, WX, editors. Springer: Singapore pp. 103–119.

Toledo I, Lloret L, Martínez-Romero E. 2003. *Sinorhizobium americanus* sp. nov., a new *Sinorhizobium* species nodulating native *Acacia* spp. in Mexico. Syst Appl Microbiol. 26:54–64.

Trinick MJ. 1980. Relationships amongst the fast-growing rhizobia of *Lablab purpureus*, *Leucaena leucocephala*, *Mimosa* spp., *Acacia farnesiana* and *Sesbania grandiflora* and

their affinities with other rhizobial groups. J Appl Bacteriol. 49:39–53.

Vasilinetc I, Prjibelski AD, Gurevich A, Korobeynikov A, Pevzner PA. 2015. Assembling short reads from jumping libraries with large insert sizes. Bioinformatics. 31:3262–3268.

Velzen R van et al. 2018. Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. Proc Natl Acad Sci USA. 115:E4700–E4709.

Wang ET. 2019. Current systematics of rhizobia. In: Ecology and Evolution of Rhizobia: Principles and Applications. Wang, ET, Tian, CF, Chen, WF, Young, JPW, & Chen, WX, editors. Springer: Singapore pp. 41–102.

Wang ET, Romero JM, Romero EM. 1999. Genetic diversity of rhizobia from *Leucaena leucocephala* nodules in Mexican soils. Mol Ecol. 8:711–724.

Wang ET, Young JPW. 2019. History of rhizobial taxonomy. In: Ecology and Evolution of Rhizobia: Principles and Applications. Wang, ET, Tian, CF, Chen, WF, Young, JPW, & Chen, WX, editors. Springer: Singapore pp. 23–39.

Wang YC et al. 2013. Proposal of *Ensifer psoraleae* sp. nov., *Ensifer sesbaniae* sp. nov., *Ensifer morelense* comb. nov. and *Ensifer americanum* comb. nov. Syst Appl Microbiol. 36:467–473.

Warnes GR et al. 2016. *gplots: Various R programming tools for plotting data. R package version 3.0.1. https://CRAN.R-project.org/package=gplots*.

Wendt T, Doohan F, Mullins E. 2012. Production of *Phytophthora infestans*-resistant potato (*Solanum tuberosum*) utilising *Ensifer adhaerens* OV14. Transgenic Res. 21:567–578.

Werner GDA, Cornwell WK, Cornelissen JHC, Kiers ET. 2015. Evolutionary signals of symbiotic persistence in the legume–rhizobia mutualism. Proc Natl Acad Sci USA. 112:10262–10269.

Werner GDA, Cornwell WK, Sprent JI, Kattge J, Kiers ET. 2014. A single evolutionary

innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. Nat Commun. 5:4087.

Wickham H. 2016. *ggplot2: elegant graphics for data analysis.* Springer-Verlag: New York, USA https://ggplot2.tidyverse.org.

Willems A et al. 2003. Description of new *Ensifer* strains from nodules and proposal to transfer *Ensifer adhaerens* Casida 1982 to *Sinorhizobium* as *Sinorhizobium adhaerens* comb. nov. Request for an opinion. Int J Syst Evol Microbiol. 43:1207–1217.

Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 28:1033–1034.

Young JM. 2003. The genus name *Ensifer* Casida 1982 takes priority over *Sinorhizobium* Chen et al. 1988, and *Sinorhizobium morelense* Wang et al. 2002 is a later synonym of *Ensifer adhaerens* Casida 1982. Is the combination '*Sinorhizobium adhaerens*' (Casida 1982) Willems et al. 2003 legitimate? Request for an opinion. Int J Syst Evol Microbiol. 53:2107–2110.

Zhang X, Harper R, Karisto M, Lindström K. 1991. Diversity of *Rhizobium* bacteria isolated from the root nodules of leguminous trees. Int J Syst Evol Microbiol. 41:104–113.

Zhao R et al. 2017. Adaptive evolution of rhizobial symbiotic compatibility mediated by co-evolved insertion sequences. ISME J. 12:101–111.

FIGURE LEGENDS

Figure 1. Unrooted phylogeny of the genus *Ensifer*. A maximum likelihood phylogeny of 157 strains was prepared from a concatenated alignment of 1,049 core genes. Nodes with a bootstrap value of 100 are indicated with the gray dots. The scale represents the mean number of nucleotide substitutions per site. The white and grey shading is used to group strains into genospecies on the basis of ANI and AAI results (Supplementary Figures S3, S4), based ANI and AAI genospecies threshold of 95% and 96%, respectively. From outside to inside, rings represent the genome assembly level (black – finished, white – draft), and the presence (black) or absence (white) of NodA, NodB, NodC, NifH, NifD, and NifK. Grey boxes indicate the presence of a truncated (as a result of incomplete genome assembly) version of the corresponding gene detected through inspection of the RefSeq annotations. Strains are named as recorded in NCBI at the time of collection.

Figure 2. Evolution of SNF within the genus *Ensifer*. Maximum likelihood phylogenies of concatenated alignments of (**A**) NodABC nodulation proteins, (**B**) NifHDK nitrogenase proteins, and (**C**) RepAB replicon partitioning proteins of the order Rhizobiales. Branches corresponding to proteins from the genus *Ensifer* are indicated with colour. (**D**) A subtree of the core gene species phylogeny of Figure 1. Colours denote taxa whose symbiotic proteins are predicted to have been vertically acquired from a common ancestor. The scale bars represent the mean number of amino acid (A-C) or nucleotide (D) substitutions per site.

Figure 3. Global genome properties of the genus *Ensifer*. (**A**) A PCA plot based on the presence and absence of all orthologous protein groups in each of the 157 *Ensifer* strains. (**B**) Box-and-whisker plots displaying the number of genes per genome in the symbiotic and non-symbiotic *Ensifer* clades. (**C**) A Venn Diagram displaying the overlap in the core genomes of the symbiotic and non-symbiotic *Ensifer* clades. (**D**) A Venn Diagram displaying the overlap in the accessory genomes of the symbiotic and non-symbiotic *Ensifer* clades.

Figure 4. Growth properties of phylogenetically diverse *Ensifer* strains. *Ensifer* strains were grown in microplates without shaking. Data points represent the average of triplicate samples, while the error bars indicate the standard deviation. Shades of pink are used for strains of the symbiotic clade, while shades of blue are used for strains of the non-symbiotic clade. (**A**) Growth in LB medium at 30˚C. (**B**) Growth in TY medium during heat stress (37˚C).

Figure 5. Phenotypic properties of phylogenetically diverse *Ensifer* strains. Ten *Ensifer* strains were screened for their ability to catabolize 190 carbon sources, and to grow in 96 osmolyte and 96 pH conditions using Biolog Phenotype MicroArray™ plates PM1, PM2, PM9, and PM10. Growth in each well was summarized on a scale of 0 (dark blue) through 9 (dark red), with higher numbers representing more robust growth. A larger version of this figure, in which each condition is labelled along the Y-axis, is provided as Supplementary Figure S8.

785 **Table 1.** *Ensifer* strains phenotypically characterized in this study.

Strain	Original source	SNF [a]	*Ensifer* clade [b]	Reference
Ensifer adhaerens Casida A	Isolated from a Pennsylvania (USA) soil sample	No	Non-symbiotic	(Casida 1982)
Ensifer adhaerens OV14	Isolated from the rhizosphere of *Brassica napus*	No	Non-symbiotic	(Wendt et al. 2012)
Ensifer sp. M14	Isolated from arsenic-rich sediments of a gold mine	No	Non-symbiotic	(Drewniak et al. 2008)
Ensifer morelense Lc04	Isolated from root nodules of *Leucaena leucocephala*	No	Non-symbiotic	(Wang et al. 1999)
Ensifer sesbaniae CCBAU 65729	Isolated from root nodules of *Sesbania cannabina*	Yes	Non-symbiotic	(Wang et al. 2013)
Ensifer fredii NGR234	Isolated from root nodules of *Lablab purpureus*	Yes	Symbiotic	(Trinick 1980)
Ensifer sojae CCBAU 05684	Isolated from root nodules of *Glycine max* grown in saline-alkaline soils	Yes	Symbiotic	(Li et al. 2011)
Ensifer americanum CFNEI 156	Isolated from root nodules of *Acacia acatlensis*	Yes	Symbiotic	(Toledo et al. 2003)
Ensifer psoraleae CCBAU 65732	Isolated from root nodules of *Psoralea corylifolia*	Yes	Symbiotic	(Wang et al. 2013)
Ensifer medicae WSM419	Isolated from root nodules of *Medicago murex*	Yes	Symbiotic	(Howieson & Ewing 1986)

786 [a] This column indicates if the strain can (Yes) or cannot (No) form nitrogen-fixing nodules on
787 legumes.
788 [b] This column indicates if the strain belongs to the symbiotic or non-symbiotic clade of the
789 genus *Ensifer* as defined in this study.
790
791

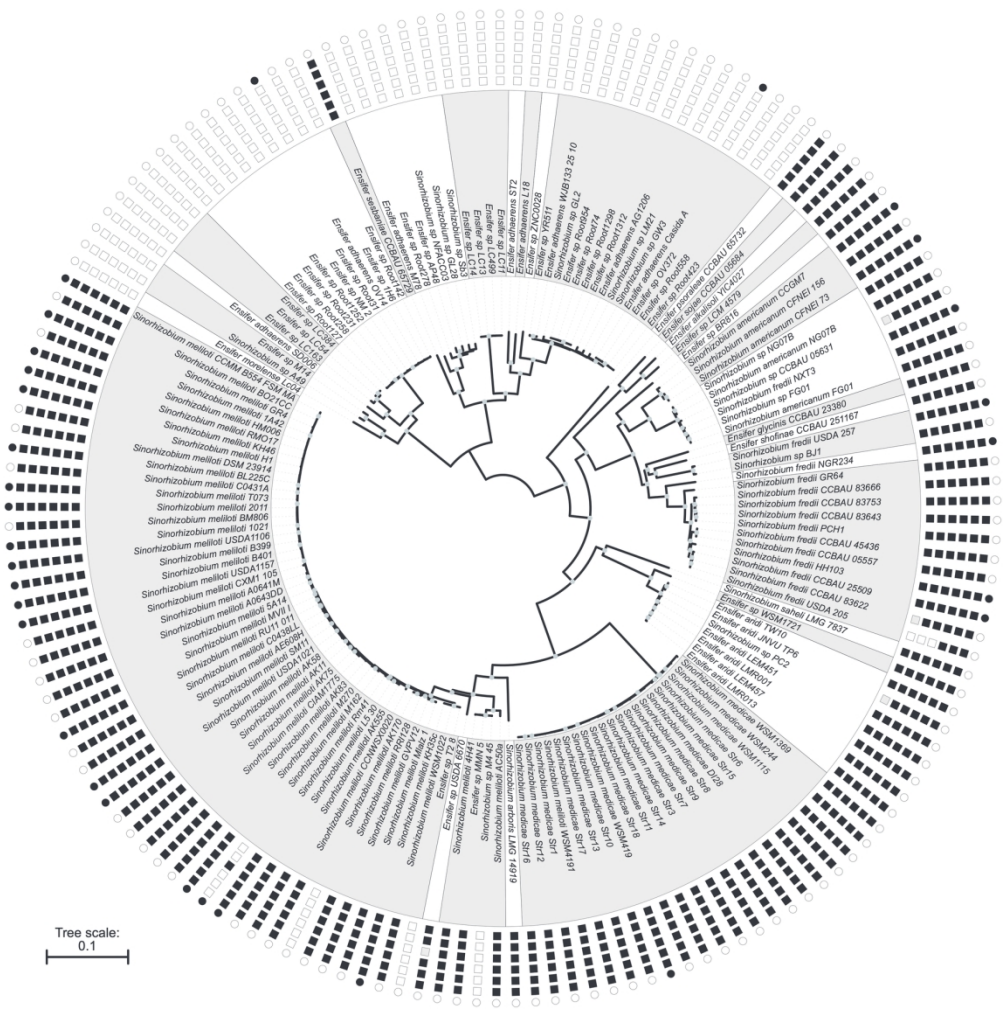

Figure 1. Unrooted phylogeny of the genus Ensifer. A maximum likelihood phylogeny of 157 strains was prepared from a concatenated alignment of 1,049 core genes. Nodes with a bootstrap value of 100 are indicated with the gray dots. The scale represents the mean number of nucleotide substitutions per site. The white and grey shading is used to group strains into genospecies on the basis of ANI and AAI results (Supplementary Figures S3, S4), based ANI and AAI genospecies threshold of 95% and 96%, respectively. From outside to inside, rings represent the genome assembly level (black – finished, white – draft), and the presence (black) or absence (white) of NodA, NodB, NodC, NifH, NifD, and NifK. Grey boxes indicate the presence of a truncated (as a result of incomplete genome assembly) version of the corresponding gene detected through inspection of the RefSeq annotations. Strains are named as recorded in NCBI at the time of collection.

160x160mm (300 x 300 DPI)

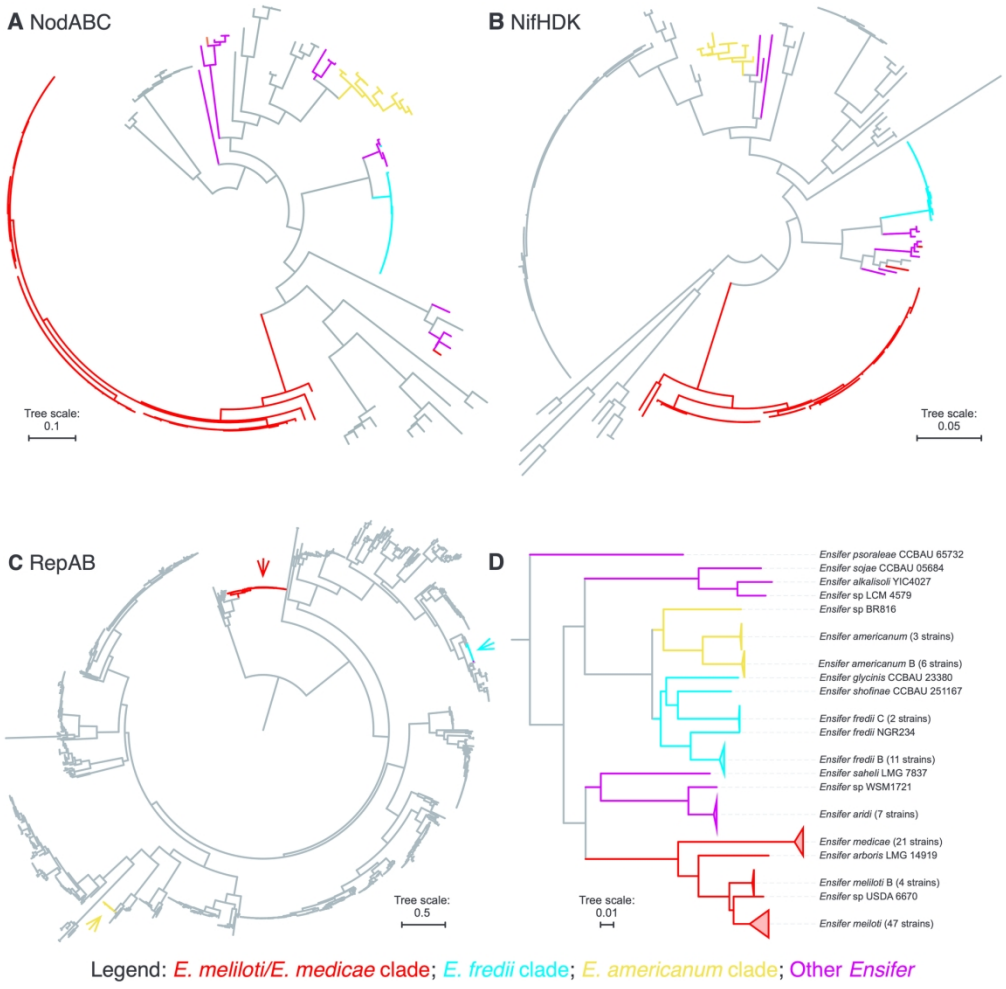

Figure 2. Evolution of SNF within the genus Ensifer. Maximum likelihood phylogenies of concatenated alignments of (A) NodABC nodulation proteins, (B) NifHDK nitrogenase proteins, and (C) RepAB replicon partitioning proteins of the order Rhizobiales. Branches corresponding to proteins from the genus Ensifer are indicated with colour. (D) A subtree of the core gene species phylogeny of Figure 1. Colours denote taxa whose symbiotic proteins are predicted to have been vertically acquired from a common ancestor. The scale bars represent the mean number of amino acid (A-C) or nucleotide (D) substitutions per site.

167x165mm (300 x 300 DPI)

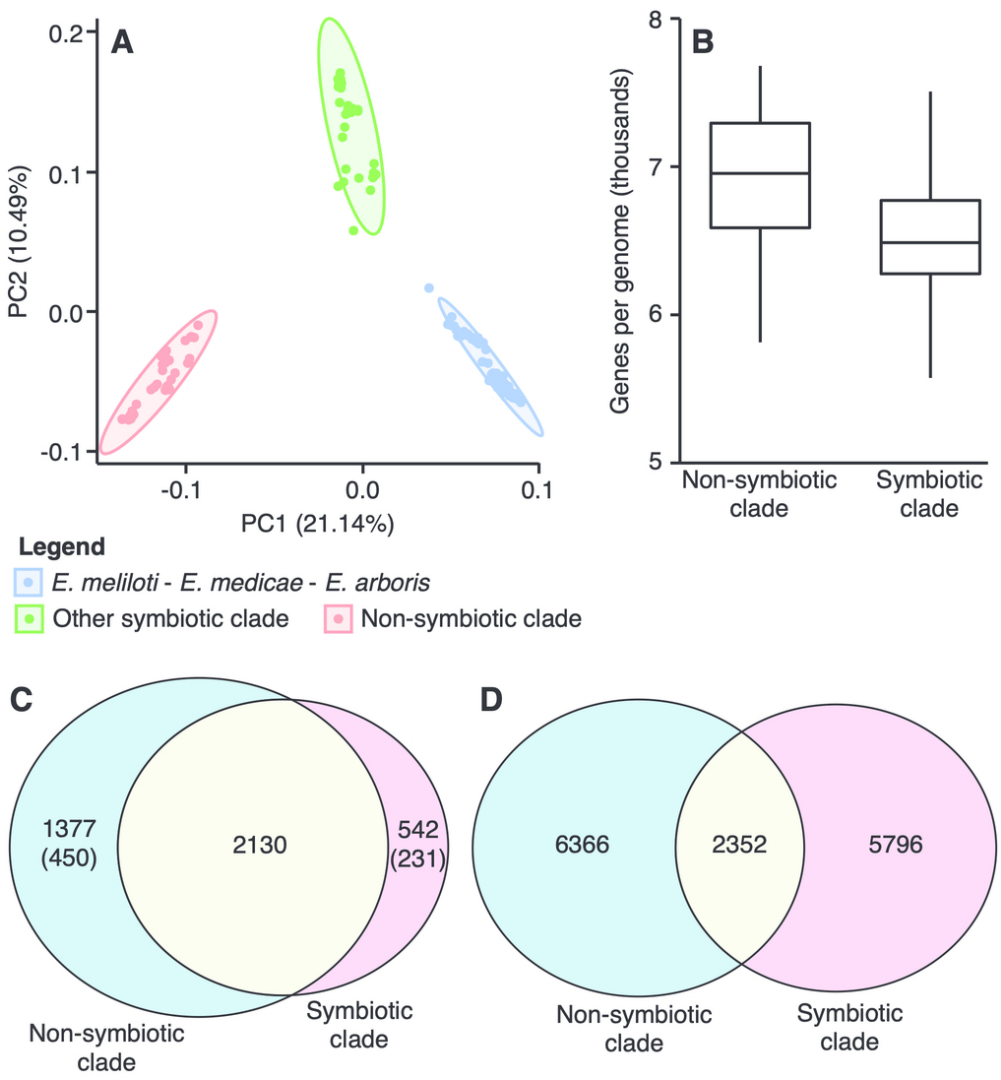

Figure 3. Global genome properties of the genus Ensifer. (A) A PCA plot based on the presence and absence of all orthologous protein groups in each of the 157 Ensifer strains. (B) Box-and-whisker plots displaying the number of genes per genome in the symbiotic and non-symbiotic Ensifer clades. (C) A Venn Diagram displaying the overlap in the core genomes of the symbiotic and non-symbiotic Ensifer clades. (D) A Venn Diagram displaying the overlap in the accessory genomes of the symbiotic and non-symbiotic Ensifer clades.

86x92mm (300 x 300 DPI)

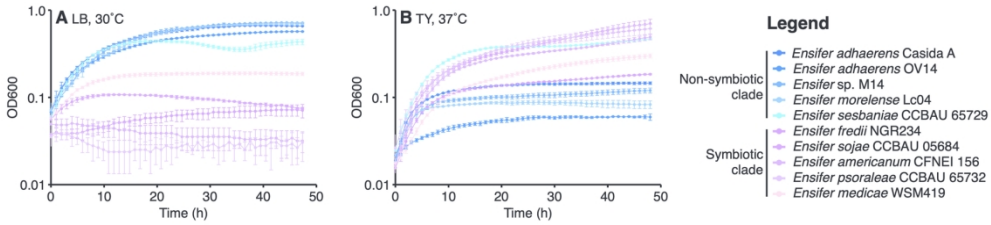

Figure 4. Growth properties of phylogenetically diverse Ensifer strains. Ensifer strains were grown in microplates without shaking. Data points represent the average of triplicate samples, while the error bars indicate the standard deviation. Shades of pink are used for strains of the symbiotic clade, while shades of blue are used for strains of the non-symbiotic clade. (A) Growth in LB medium at 30˚C. (B) Growth in TY medium during heat stress (37˚C).

159x35mm (300 x 300 DPI)

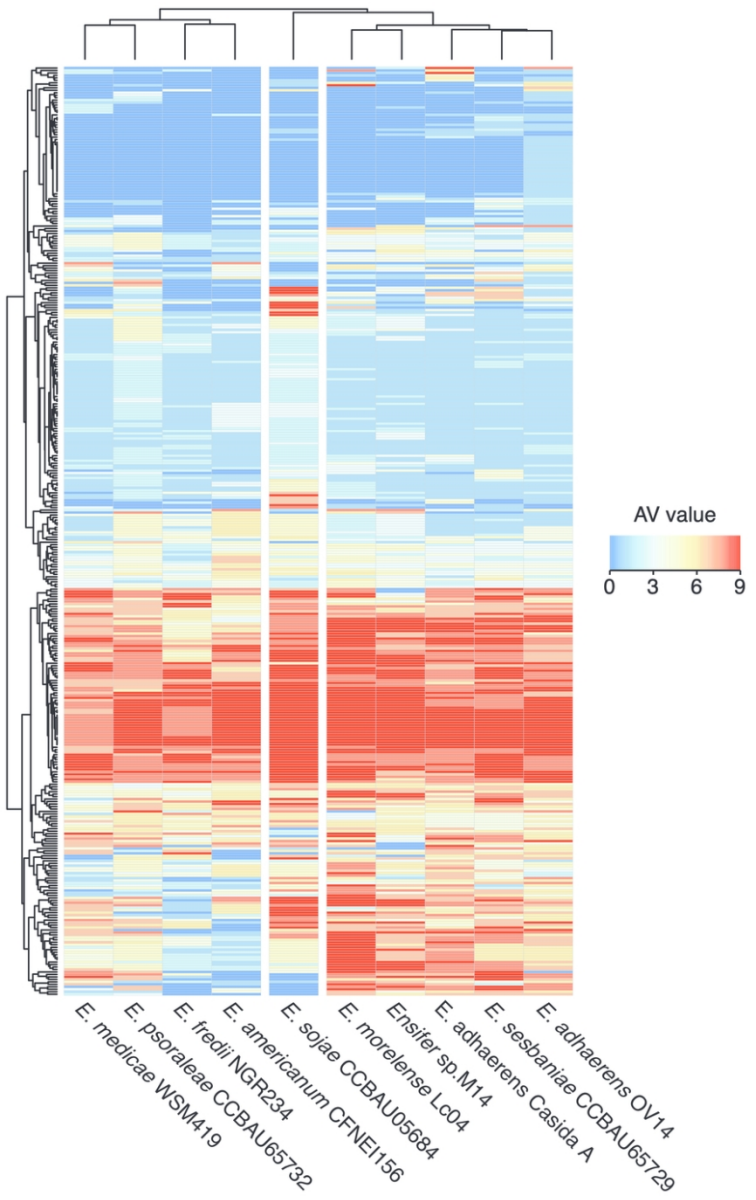

Figure 5. Phenotypic properties of phylogenetically diverse Ensifer strains. Ten Ensifer strains were screened for their ability to catabolize 190 carbon sources, and to grow in 96 osmolyte and 96 pH conditions using Biolog Phenotype MicroArrayTM plates PM1, PM2, PM9, and PM10. Growth in each well was summarized on a scale of 0 (dark blue) through 9 (dark red), with higher numbers representing more robust growth. A larger version of this figure, in which each condition is labelled along the Y-axis, is provided as Supplementary Figure S8.

76x119mm (300 x 300 DPI)