# Genome sequence of *Ensifer arboris* strain LMG 14919$^T$; a microsymbiont of the legume *Prosopis chilensis* growing in Kosti, Sudan

**Wayne Reeve[1*], Rui Tian[1], Lambert Bräu[2], Lynne Goodwin[3], Christine Munk[3], Chris Detter[3], Roxanne Tapia[3], Cliff Han[3], Konstantinos Liolios[4], Marcel Huntemann[4], Amrita Pati[4], Tanja Woyke[4], Konstantinos Mavrommatis[5], Victor Markowitz[5], Natalia Ivanova[4], Nikos Kyrpides[4] & Anne Willems[6].**

**[1] Centre for Rhizobium Studies, Murdoch University, Western Australia, Australia**

**[2] School of Life and Environmental Sciences, Deakin University, Victoria, Australia**

**[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA**

**[4] DOE Joint Genome Institute, Walnut Creek, California, USA**

**[5] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA**

**[6] Laboratory of Microbiology, Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Belgium**

***Correspondence:** Wayne Reeve (W.Reeve@murdoch.edu.au)

**Keywords:** root-nodule bacteria, nitrogen fixation, rhizobia, *Alphaproteobacteria*

*Ensifer arboris* LMG 14919$^T$ is an aerobic, motile, Gram-negative, non-spore-forming rod that can exist as a soil saprophyte or as a legume microsymbiont of several species of legume trees. LMG 14919$^T$ was isolated in 1987 from a nodule recovered from the roots of the tree *Prosopis chilensis* growing in Kosti, Sudan. LMG 14919$^T$ is highly effective at fixing nitrogen with *P. chilensis* (Chilean mesquite) and *Acacia senegal* (gum Arabic tree or gum acacia). LMG 14919$^T$ does not nodulate the tree *Leucena leucocephala*, nor the herbaceous species *Macroptilium atropurpureum*, *Trifolium pratense*, *Medicago sativa*, *Lotus corniculatus* and *Galega orientalis*. Here we describe the features of *E. arboris* LMG 14919$^T$, together with genome sequence information and its annotation. The 6,850,303 bp high-quality-draft genome is arranged into 7 scaffolds of 12 contigs containing 6,461 protein-coding genes and 84 RNA-only encoding genes, and is one of 100 rhizobial genomes sequenced as part of the DOE Joint Genome Institute 2010 Genomic Encyclopedia for Bacteria and Archaea-Root Nodule Bacteria (GEBA-RNB) project.

## Introduction

Legume plants form nitrogen fixing symbiosis with root nodule bacteria, collectively called rhizobia. These legumes are particularly useful crop plants that do not require exogenous nitrogenous fertilizer to support growth in less fertile, nitrogen-deficient conditions. They include some of our staple food and feed plants such as beans, peas, soybeans, lentils, clover, peanuts and alfalfa and are mostly annual crops. In many arid and savannah regions, leguminous trees represent a particularly valuable resource as they are often deep-rooted and drought resistant. They have been used traditionally in the Sahel region as sources of timber, fodder and for soil improvement [1]. *Prosopis chilensis*, also known as Chilean mesquite, is a native tree from South America that has many uses: its nutritious pods can be ground to produce flour and are also eaten by livestock; its wood is used for construction and furniture. Chilean mesquite is also used for intercropping

with other plants, for which it provides shelter and nutrients (leaf compost, nitrogen). *Acacia senegal* (recently renamed as *Senegalia senegal*) is a plant of particular importance in the production of gum arabic in the Sahel region and the Middle East. Its seeds are dried for human consumption, and its leaves and pods serve as feed for sheep, goats and camels. The plant is also used in agroforestry in intercropping with watermelon and grasses, and in rotation systems with other crops (Agroforestree Database [2]).

The microsymbiont of these legume trees from Sudan and Kenya [3] has been renamed as *Ensifer arboris* [4], of which LMG 14919$^T$ (= HAMBI 1552, ORS 1755, TTR38) is the type strain. This strain was isolated from root nodules of *Prosopis chilensis* from Kosti, Sudan, and shown to effectively nodulate its original host as well as *Acacia senegal* [5].

Given the drought tolerance of the host trees, it seems fitting that their symbionts are also stress resistant: *Ensifer arboris* was described as tolerant to temperatures up to 41-43 °C, 3% NaCl, several heavy metals (including Pb, Cd, Hg, Cu) and a wide range of antibiotics [3,5], characteristics that contribute to the success of the rhizobial-legume tree association in challenging environmental conditions [6]. Here we present a summary classification and a set of features for *E. arboris* strain LMG 14919$^T$ (Table 1), together with the description of the complete genome sequence and its annotation.

## Classification and features

*E. arboris* LMG 14919$^T$ is a motile, non-sporulating, non-encapsulated, Gram-negative rod in the order *Rhizobiales* of the class *Alphaproteobacteria*. The rod-shaped form varies in size with dimensions of approximately 0.25 µm in width and 1.0-1.5 µm in length (Figure 1, Left and Center). The strain is fast-growing, forming colonies within 3-4 days when grown on half strength Lupin Agar (½LA) [19], tryptone-yeast extract agar (TY) [20] or a modified yeast-mannitol agar (YMA) [21] at 28°C. Colonies on ½LA are white-opaque, slightly domed and moderately mucoid with smooth margins (Figure 1 Right).

*E. arboris* LMG 14919$^T$ is capable of using several amino acids, including L-proline, L-arginine, sodium glutamate and L-histidine as sole nitrogen sources and can use a wide range of different carbon sources including L-arabinose, D-galactose, raffinose, L-rhamnose, maltose, lactose, D-fructose, D-mannose, trehalose, D-ribose, xylene, methyl-D-mannoside, sorbitol, dulcitol, meso-inositol, inulin, dextrin, amygdalin, arbutin, sodium citrate, itaconate, α-ketoglutarate, sodium maltose, 1,2-propylene glycol, and 1,2-butylene glycol [5].

Minimum Information about the Genome Sequence (MIGS) is provided in Table 1. Figure 2 shows the phylogenetic neighborhood of *E. arboris* LMG 14919$^T$ in a 16S rRNA sequence based tree. This strain shares 99% (1361/1366 bp) and 99% (1361/1366 bp) sequence identity to the 16S rRNA of the fully sequenced *E. meliloti* Sm1021 [26] and *E. medicae* WSM419 [27] strains, respectively.

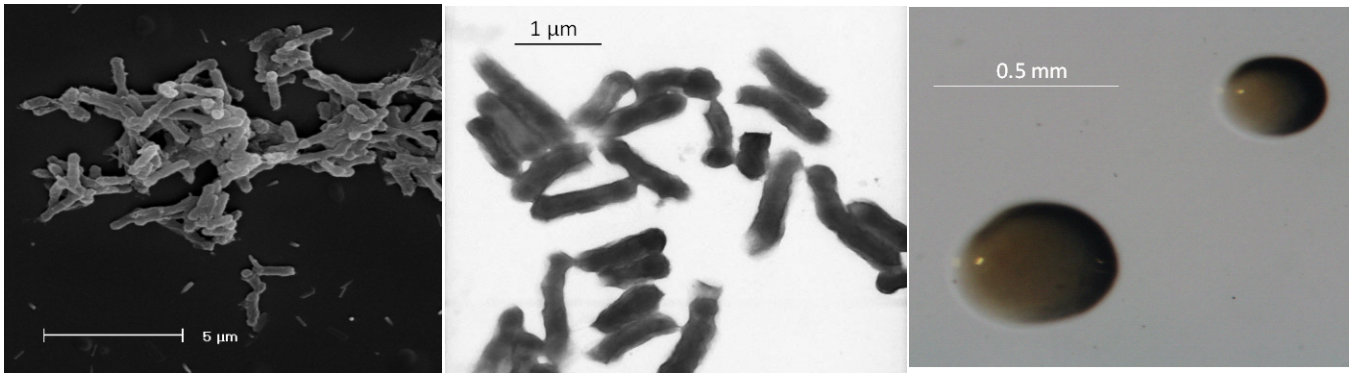


**Figure 1**. Images of *Ensifer arboris* LMG 14919$^T$ using scanning (Left) and transmission (Center) electron microscopy and the appearance of colony morphology on a solid medium (Right).

**Table 1**. Classification and general features of *Ensifer arboris* LMG 14919$^T$ according to the MIGS recommendations [7]

| MIGS ID | Property | Term | Evidence code |
|---|---|---|---|
| | | Domain *Bacteria* | TAS [8] |
| | | Phylum *Proteobacteria* | TAS [9] |
| | | Class *Alphaproteobacteria* | TAS [10,11] |
| | Current classification | Order *Rhizobiales* | TAS [11,12] |
| | | Family *Rhizobiaceae* | TAS [13,14] |
| | | Genus *Ensifer* | TAS [4,15,16] |
| | | Species *Ensifer arboris* | TAS [4] |
| | | Strain LMG 14919$^T$ | |
| | Gram stain | Negative | IDA |
| | Cell shape | Rod | IDA |
| | Motility | Motile | IDA |
| | Sporulation | Non-sporulating | NAS |
| | Temperature range | Mesophile | NAS |
| | Optimum temperature | 28°C | NAS |
| | Salinity | Non-halophile | NAS |
| MIGS-22 | Oxygen requirement | Aerobic | TAS [3] |
| | Carbon source | Varied | TAS [5] |
| | Energy source | Chemoorganotroph | NAS |
| MIGS-6 | Habitat | Soil, root nodule, on host | TAS [3,5] |
| MIGS-15 | Biotic relationship | Free living, symbiotic | TAS [3,5] |
| MIGS-14 | Pathogenicity | Non-pathogenic | NAS |
| | Biosafety level | 1 | TAS [17] |
| | Isolation | Root nodule | TAS [5] |
| MIGS-4 | Geographic location | Kosti, Sudan | TAS [5] |
| MIGS-5 | Soil collection date | 1987 | IDA |
| MIGS-4.1 | Longitude | 32.66342 | TAS [5] |
| MIGS-4.2 | Latitude | 13.16125 | TAS [5] |
| MIGS-4.3 | Depth | Not reported | NAS |
| MIGS-4.4 | Altitude | Not reported | NAS |

Evidence codes – IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [18].

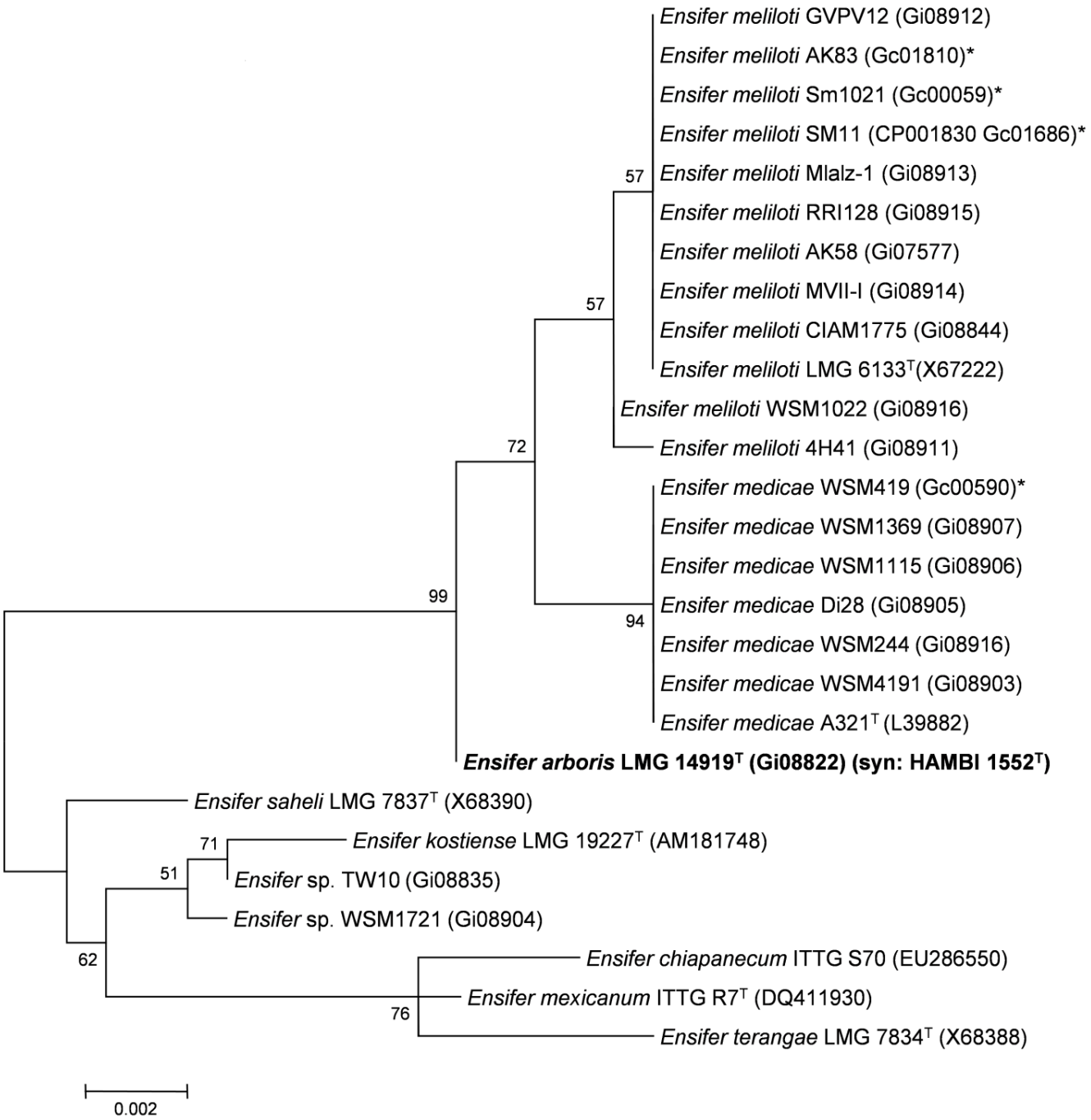


**Figure 2.** Phylogenetic tree showing the relationship of *Ensifer arboris* LMG 14919$^T$ (shown in bold print) to other *Ensifer* spp. in the order *Rhizobiales* based on aligned sequences of the 16S rRNA gene (1,290 bp internal region). All sites were informative and there were no gap-containing sites. Phylogenetic analyses were performed using MEGA, version 5 [22]. The tree was built using the Maximum-Likelihood method with the General Time Reversible model [23]. Bootstrap analysis [24] with 500 replicates was performed to assess the support of the clusters. Type strains are indicated with a superscript T. Brackets after the strain name contain a DNA database accession number and/or a GOLD ID (beginning with the prefix G) for a sequencing project registered in GOLD [25]. Published genomes are indicated with an asterisk.

### Symbiotaxonomy

*E. arboris* LMG 14919$^T$ was initially shown to form nodules (Nod$^+$) and fix nitrogen (Fix$^+$) with two leguminous tree species, *P. chilensis* and *A. senegal.* It was unable to elicit nodules on the herbaceous perennials *Macroptilium atropurpureum*, *Trifolium pratense*, *Medicago sativa*, *Lotus corniculatus* and *Galega orientalis* [5]. The symbiotic properties of this strain in seedlings of *Acacia* and *Prosopis* spp. in Sudan and Senegal have been reported in detail [6]. Indeterminate nodules are induced, mainly on the lateral roots either in clusters or individually. Young nodules are spherical and later become elongated and are commonly branched. LMG 14919$^T$ (=HAMBI 1552) was shown to nodulate and fix nitrogen in seedlings of African *A. mellifera*, *A. nilotica*, *A. oerfota* (synonym *A. nubica*), *A. senegal*, *A. seyal*, *A. sieberiana*, *A. tortilis* subsp. *raddiana*, Latin American *A. angustissima*, *P. chilensis* and *P. pallida*, and Afro-Asian *P. cineraria*. It also effectively nodulates with Latin-American introductions of *P. chilensis* and *P. juliflora* in Africa [6]. It induced small ineffective nodules on Australian *A. holosericea* and African *P. africana* [6].

## Genome sequencing and annotation

### Genome project history

This organism was selected for sequencing on the basis of its environmental and agricultural relevance to issues in global carbon cycling, alternative energy production, and biogeochemical importance, and is part of the Community Sequencing Program at the U.S. Department of Energy, Joint Genome Institute (JGI) for projects of relevance to agency missions. The genome project is deposited in the Genomes OnLine Database [25] and an improved-high-quality-draft genome sequence in IMG. Sequencing, finishing and annotation were performed by the JGI. A summary of the project information is shown in Table 2.

### Growth conditions and DNA isolation

*E. arboris* LMG 14919$^T$ was cultured to mid logarithmic phase in 60 ml of TY rich medium on a gyratory shaker at 28°C [28]. DNA was isolated from the cells using a CTAB (Cetyl trimethyl ammonium bromide) bacterial genomic DNA isolation method [29].

### Genome sequencing and assembly

The genome of *Ensifer arboris* LMG 14919$^T$ was sequenced at the Joint Genome Institute (JGI) using Illumina technology [30]. An Illumina short-insert paired-end library with an average insert size of 270 bp generated 19,256,666 reads and an Illumina long-insert paired-end library with an average insert size of 9,232.94 +/- 2,530.88 bp generated 1,365,298 reads totaling 3,093.3 Mbp of Illumina data. All general aspects of library construction and sequencing performed at the JGI can be found at the JGI user home.

**Table 2**. Genome sequencing project information for *E. arboris* LMG 14919$^T$.

| MIGS ID | Property | Term |
|---|---|---|
| MIGS-31 | Finishing quality | Improved high-quality draft |
| MIGS-28 | Libraries used | Illumina Standard (short PE) and Illumina CLIP (long PE) library |
| MIGS-29 | Sequencing platforms | Illumina HiSeq 2000 |
| MIGS-31.2 | Sequencing coverage | Illumina: 448× |
| MIGS-30 | Assemblers | Velvet version 1.1.05; Allpaths-LG version r38445 |
| MIGS-32 | Gene calling methods | Prodigal 1.4, GenePRIMP |
| | GenBank | ATYB00000000 |
| | GenBank release date | July 15, 2013 |
| | GOLD ID | Gi08822 |
| | NCBI project ID | 74465 |
| | Database: IMG | 2512047086 |
| | Project relevance | Symbiotic $N_2$ fixation, agriculture |

The initial draft assembly contained 27 contigs in 9 scaffolds. The initial draft data was assembled with Allpaths, version r38445, and the consensus was computationally shredded into 10 Kbp overlapping fake reads (shreds). The Illumina draft data was also assembled with Velvet, version 1.1.05 [31], and the consensus sequences were computationally shredded into 1.5 Kbp overlapping fake reads (shreds). The Illumina draft data was assembled again with Velvet using the shreds from the first Velvet assembly to guide the next assembly. The consensus from the second VELVET assembly was shredded into 1.5 Kbp overlapping fake reads. The fake reads from the Allpaths assembly and both Velvet assemblies and a subset of the Illumina CLIP paired-end reads were assembled using parallel phrap, version SPS 4.24 (High Performance Software, LLC). Possible misassemblies were corrected with manual editing in Consed [32-34]. Gap closure was accomplished using repeat resolution software (Wei Gu, unpublished), and sequencing of bridging PCR fragments using Sanger (unpublished, Cliff Han) technology. For the improved high quality draft, one round of manual/wet lab finishing was completed. A total of 46 additional sequencing reactions, were completed to close gaps and to raise the quality of the final sequence. The estimated total size of the genome is 6.9 Mbp and the final assembly is based on 3,093.3 Mbp of Illumina draft data, which provides an average of 448× coverage of the genome.

### Genome annotation

Genes were identified using Prodigal [35] as part of the DOE-JGI annotation pipeline [36] followed by a round of manual curation using the JGI GenePRIMP pipeline [37]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) non-redundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. These data sources were combined to assert a product description for each predicted protein. Non-protein coding genes and miscellaneous features were predicted using tRNAscan-SE [38], RNAMMer [39], searches against models of the ribosomal RNA genes built from SILVA [40], Rfam [41], TMHMM [42], and SignalP [43]. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG-ER) platform [44].

## Genome properties

The genome is 6,850,303 nucleotides with 62.02% GC content (Table 3) and comprised of 7 scaffolds (Figure 3) of 12 contigs. From a total of 6,545 genes, 6,461 were protein encoding and 84 RNA only encoding genes. The majority of genes (80.78%) were assigned a putative function whilst the remaining genes were annotated as hypothetical. The distribution of genes into COGs functional categories is presented in Table 4.

**Table 3.** Genome Statistics for *Ensifer arboris* LMG 14919$^T$

| Attribute | Value | % of Total |
|---|---|---|
| Genome size (bp) | 6,850,303 | 100.00 |
| DNA coding region (bp) | 5,921,899 | 86.45 |
| DNA G+C content (bp) | 4,248,771 | 62.02 |
| Number of scaffolds | 7 | |
| Number of contigs | 12 | |
| Total gene | 6,545 | 100.00 |
| RNA genes | 84 | 1.28 |
| rRNA operons | 3 | 0.05 |
| Protein-coding genes | 6,461 | 98.72 |
| Genes with function prediction | 5,287 | 80.78 |
| Genes assigned to COGs | 5,233 | 79.95 |
| Genes assigned Pfam domains | 5,438 | 83.09 |
| Genes with signal peptides | 588 | 8.98 |
| Genes with transmembrane helices | 1,456 | 22.25 |
| CRISPR repeats | 0 | |

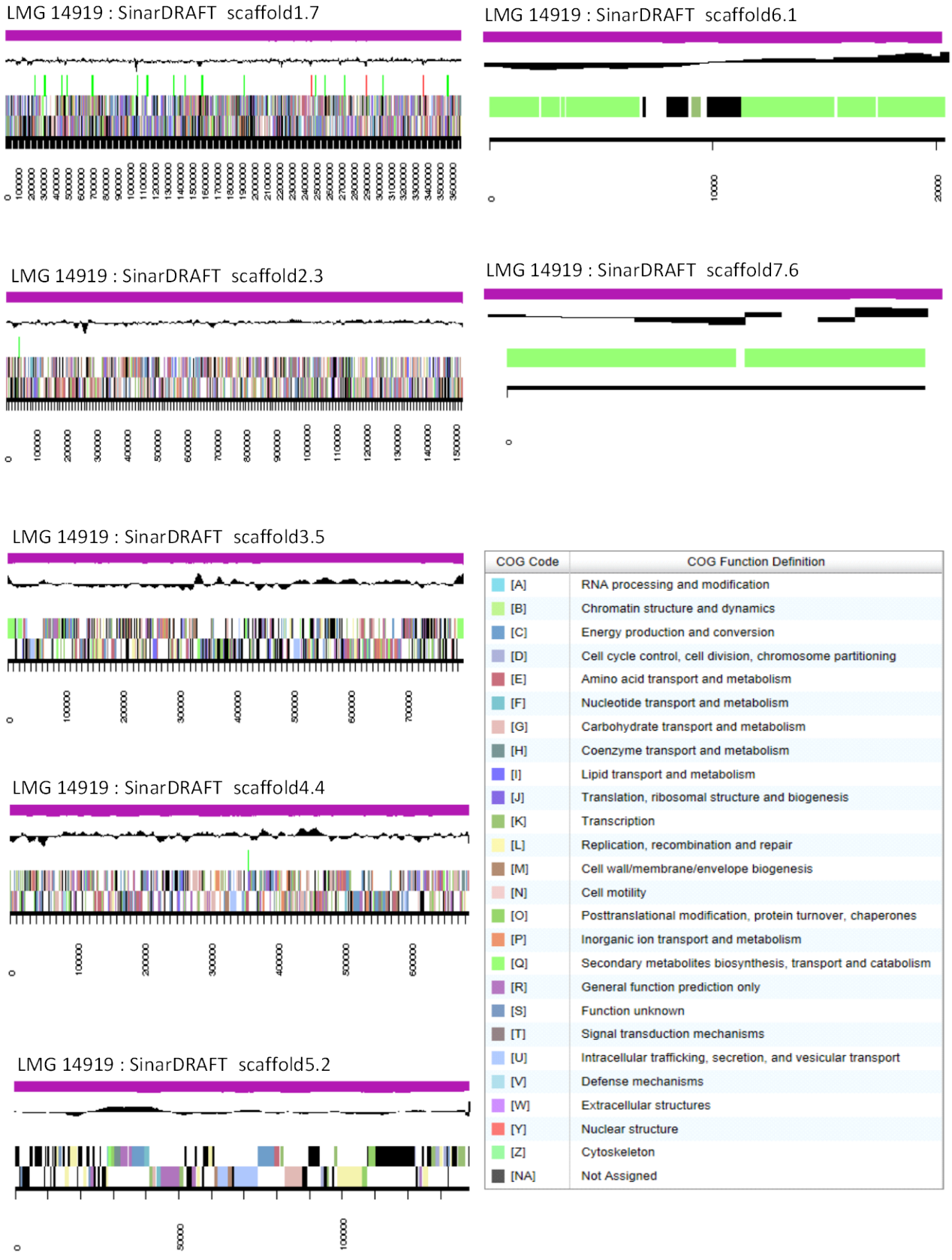


**Figure 3**. Graphical map of the genome of *Ensifer arboris* LMG 14919$^T$ showing the seven largest scaffolds. From bottom to the top of each scaffold: Genes on forward strand (color by COG categories as denoted by the IMG platform), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, sRNAs red, other RNAs black), GC content, GC skew.

**Table 4.** Number of protein coding genes of *Ensifer arboris* LMG 14919$^T$ associated with the general COG functional categories.

| Code | Value | % age | Description |
|---|---|---|---|
| J | 195 | 3.35 | Translation, ribosomal structure and biogenesis |
| A | 0 | 0.00 | RNA processing and modification |
| K | 510 | 8.76 | Transcription |
| L | 212 | 3.64 | Replication, recombination and repair |
| B | 1 | 0.02 | Chromatin structure and dynamics |
| D | 49 | 0.84 | Cell cycle control, mitosis and meiosis |
| Y | 0 | 0.00 | Nuclear structure |
| V | 60 | 1.03 | Defense mechanisms |
| T | 248 | 4.26 | Signal transduction mechanisms |
| M | 274 | 4.71 | Cell wall/membrane biogenesis |
| N | 77 | 1.32 | Cell motility |
| Z | 0 | 0.00 | Cytoskeleton |
| W | 0 | 0.00 | Extracellular structures |
| U | 122 | 2.10 | Intracellular trafficking and secretion |
| O | 185 | 3.18 | Posttranslational modification, protein turnover, chaperones |
| C | 349 | 6.00 | Energy production conversion |
| G | 598 | 10.27 | Carbohydrate transport and metabolism |
| E | 653 | 11.22 | Amino acid transport metabolism |
| F | 104 | 1.79 | Nucleotide transport and metabolism |
| H | 201 | 3.45 | Coenzyme transport and metabolism |
| I | 205 | 3.52 | Lipid transport and metabolism |
| P | 292 | 5.02 | Inorganic ion transport and metabolism |
| Q | 182 | 3.13 | Secondary metabolite biosynthesis, transport and catabolism |
| R | 721 | 12.39 | General function prediction only |
| S | 582 | 10.00 | Function unknown |
| - | 1,312 | 20.05 | Not in COGS |

## Acknowledgements

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

## References


1. Deans JD, Diagne O, Nizinski J, Lindley DK, Seck M, Ingleby K, Munro RC. Comparative growth, biomass production, nutrient use and soil amelioration by nitrogen-fixing tree species in semi-arid Senegal. For Ecol Manage 2003; 176:253-264. http://dx.doi.org/10.1016/S0378-1127(02)00296-7
2. Agroforestree Database. http://www.worldagroforestrycentre.org/resources/databases/agroforestree
3. Nick G, de Lajudie P, Eardly BD, Suomalainen S, Paulin L, Zhang X, Gillis M, Lindstrom K. *Sinorhizobium arboris* sp. nov. and *Sinorhizobium kostiense* sp. nov., isolated from leguminous trees in Sudan and Kenya. *Int J Syst Bacteriol* 1999; **49**:1359-1368. PubMed http://dx.doi.org/10.1099/00207713-49-4-1359
4. Young JM. 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* 2003; **53**:2107-2110. PubMed http://dx.doi.org/10.1099/ijs.0.02665-0
5. Zhang X, Harper R, Karsisto M, Lindstrom K. Diversity of *Rhizobium* bacteria isolated from the root nodules of leguminous trees. *Int J Syst Evol Microbiol* 1991; **41**:104-113.
6. Räsänen LA, Lindström K. Effects of biotic and abiotic constraints on the symbiosis between rhizobia and the tropical leguminous trees *Acacia* and *Prosopis*. *Indian J Exp Biol* 2003; **41**:1142-1159. PubMed
7. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen M, Angiuoli SV, *et al*. Towards a richer description of our complete collection of genomes and metagenomes "Minimum Information about a Genome Sequence " (MIGS) specification. *Nat Biotechnol* 2008; **26**:541-547. PubMed http://dx.doi.org/10.1038/nbt1360
8. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains *Archaea, Bacteria*, and *Eucarya*. *Proc Natl Acad Sci USA* 1990; **87**:4576-4579. PubMed http://dx.doi.org/10.1073/pnas.87.12.4576
9. Garrity GM, Bell JA, Lilburn T. Phylum XIV. *Proteobacteria* phyl. nov. In: Garrity GM, Brenner DJ, Krieg NR, Staley JT (eds), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 2, Part B, Springer, New York, 2005, p. 1.
10. Garrity GM, Bell JA, Lilburn T. Class I. *Alphaproteobacteria* class. nov. In: Garrity GM, Brenner DJ, Krieg NR, Staley JT (eds), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 2, Part C, Springer, New York, 2005, p. 1.
11. Validation List No. 107. List of new names and new combinations previously effectively, but not validly, published. *Int J Syst Evol Microbiol* 2006; **56**:1-6. PubMed http://dx.doi.org/10.1099/ijs.0.64188-0
12. Kuykendall LD. Order VI. *Rhizobiales* ord. nov. In: Garrity GM, Brenner DJ, Kreig NR, Staley JT, editors. Bergey's Manual of Systematic Bacteriology. Second ed: New York: Springer - Verlag; 2005. p 324.
13. Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. *Int J Syst Bacteriol* 1980; **30**:225-420. http://dx.doi.org/10.1099/00207713-30-1-225
14. Conn HJ. Taxonomic relationships of certain nonsporeforming rods in soil. *J Bacteriol* 1938; **36**:320-321.
15. Casida LE. *Ensifer adhaerens* gen. nov., sp. nov.: a bacterial predator of bacteria in soil. *Int J Syst Bacteriol* 1982; **32**:339-345. http://dx.doi.org/10.1099/00207713-32-3-339
16. Judicial Commission of the International Committee on Systematics of Prokaryotes. The genus name *Sinorhizobium* Chen et al. 1988 is a later synonym of *Ensifer* Casida 1982 and is not conserved over the latter genus name, and the species name '*Sinorhizobium adhaerens*' is not validly published. Opinion 84. *Int J Syst Evol Microbiol* 2008; **58**:1973. PubMed http://dx.doi.org/10.1099/ijs.0.2008/005991-0
17. Agents B. Technical rules for biological agents. TRBA (http://www.baua.de):466.
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**:25-29. PubMed http://dx.doi.org/10.1038/75556
19. Howieson JG, Ewing MA, D'antuono MF. Selection for acid tolerance in *Rhizobium meliloti*. *Plant Soil* 1988; **105**:179-188. http://dx.doi.org/10.1007/BF02376781

20. Beringer JE. R factor transfer in *Rhizobium leguminosarum*. *J Gen Microbiol* 1974; **84**:188-198. PubMed http://dx.doi.org/10.1099/00221287-84-1-188

21. Terpolilli JJ. Why are the symbioses between some genotypes of *Sinorhizobium* and *Medicago* suboptimal for $N_2$ fixation? Perth: Murdoch University; 2009. 223 p.

22. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 2011; **28**:2731-2739. PubMed http://dx.doi.org/10.1093/molbev/msr121

23. Nei M, Kumar S. Molecular Evolution and Phylogenetics. New York: Oxford University Press; 2000.

24. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985; **39**:783-791. http://dx.doi.org/10.2307/2408678

25. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2008; **36**:D475-D479. PubMed http://dx.doi.org/10.1093/nar/gkm884

26. Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, *et al*. The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* 2001; **293**:668-672. PubMed http://dx.doi.org/10.1126/science.1060966

27. Reeve W, Chain P, O'Hara G, Ardley J, Nandesena K, Brau L, Tiwari R, Malfatti S, Kiss H, Lapidus A, *et al*. Complete genome sequence of the *Medicago* microsymbiont *Ensifer* (*Sinorhizobium*) *medicae* strain WSM419. *Stand Genomic Sci* 2010; **2**:77-86. PubMed http://dx.doi.org/10.4056/sigs.43526

28. Reeve WG, Tiwari RP, Worsley PS, Dilworth MJ, Glenn AR, Howieson JG. Constructs for insertional mutagenesis, transcriptional signal localization and gene regulation studies in root nodule and other bacteria. *Microbiology* 1999; **145**:1307-1316. PubMed http://dx.doi.org/10.1099/13500872-145-6-1307

29. DOE Joint Genome Institute. http://my.jgi.doe.gov/general/index.html

30. Bennett S. Solexa Ltd. *Pharmacogenomics* 2004; **5**:433-438. PubMed http://dx.doi.org/10.1517/14622416.5.4.433

31. Zerbino DR. Using the Velvet *de novo* assembler for short-read sequencing technologies. Current Protocols in Bioinformatics 2010;Chapter 11:Unit 11 5.

32. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; **8**:186-194. PubMed http://dx.doi.org/10.1101/gr.8.3.175

33. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998; **8**:175-185. PubMed http://dx.doi.org/10.1101/gr.8.3.175

34. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Res* 1998; **8**:195-202. PubMed http://dx.doi.org/10.1101/gr.8.3.195

35. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**:119. PubMed http://dx.doi.org/10.1186/1471-2105-11-119

36. Mavromatis K, Ivanova NN, Chen IM, Szeto E, Markowitz VM, Kyrpides NC. The DOE-JGI Standard operating procedure for the annotations of microbial genomes. *Stand Genomic Sci* 2009; **1**:63-67. PubMed http://dx.doi.org/10.4056/sigs.632

37. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 2010; **7**:455-457. PubMed http://dx.doi.org/10.1038/nmeth.1457

38. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997; **25**:955-964. PubMed

39. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007; **35**:3100-3108. PubMed http://dx.doi.org/10.1093/nar/gkm160

40. Pruesse E, Quast C, Knittel K. Fuchs BdM, Ludwig W, Peplies J, Glöckner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007; **35**:7188-

7196. PubMed
http://dx.doi.org/10.1093/nar/gkm864

41. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res* 2003; **31**:439-441. PubMed
http://dx.doi.org/10.1093/nar/gkg006

42. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001; **305**:567-580. PubMed
http://dx.doi.org/10.1006/jmbi.2000.4315

43. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004; **340**:783-795. PubMed
http://dx.doi.org/10.1016/j.jmb.2004.05.028

44. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 2009; **25**:2271-2278. PubMed
http://dx.doi.org/10.1093/bioinformatics/btp393