Letter

# Genome sequence of the β-rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia

Claire Amadou,[1,6] Géraldine Pascal,[1,6] Sophie Mangenot,[2,6] Michelle Glew,[1] Cyril Bontemps,[1] Delphine Capela,[1] Sébastien Carrère,[1] Stéphane Cruveiller,[3] Carole Dossat,[2] Aurélie Lajus,[3] Marta Marchetti,[1] Véréna Poinsot,[4] Zoé Rouy,[3] Bertrand Servin,[5] Maged Saad,[1] Chantal Schenowitz,[2] Valérie Barbe,[2] Jacques Batut,[1] Claudine Médigue,[3] and Catherine Masson-Boivin[1,7]

[1]*Laboratoire des Interactions Plantes Micro-organismes (LIPM), UMR CNRS-INRA 2594/441, 31326 Castanet-Tolosan Cedex, France;* [2]*C.E.A/IG/Génoscope, 91057 Evry Cedex, France;* [3]*CNRS-UMR 8030, 91057 Evry Cedex, France;* [4]*Laboratoire des IMRCP, 31062 Toulouse-Cedex, France;* [5]*Laboratoire de Génétique Cellulaire, UMR INRA-ENVT, 31326 Castanet-Tolosan Cedex, France*

We report the first complete genome sequence of a β-proteobacterial nitrogen-fixing symbiont of legumes, *Cupriavidus taiwanensis* LMG19424. The genome consists of two chromosomes of size 3.42 Mb and 2.50 Mb, and a large symbiotic plasmid of 0.56 Mb. The *C. taiwanensis* genome displays an unexpected high similarity with the genome of the saprophytic bacterium *C. eutrophus* H16, despite being 0.94 Mb smaller. Both organisms harbor two chromosomes with large regions of synteny interspersed by specific regions. In contrast, the two species host highly divergent plasmids, with the consequence that *C. taiwanensis* is symbiotically proficient and less metabolically versatile. Altogether, specific regions in *C. taiwanensis* compared with *C. eutrophus* cover 1.02 Mb and are enriched in genes associated with symbiosis or virulence in other bacteria. *C. taiwanensis* reveals characteristics of a minimal rhizobium, including the most compact (35-kb) symbiotic island (*nod* and *nif*) identified so far in any rhizobium. The atypical phylogenetic position of *C. taiwanensis* allowed insightful comparative genomics of all available rhizobium genomes. We did not find any gene that was both common and specific to all rhizobia, thus suggesting that a unique shared genetic strategy does not support symbiosis of rhizobia with legumes. Instead, phylodistribution analysis of more than 200 *Sinorhizobium meliloti* known symbiotic genes indicated large and complex variations of their occurrence in rhizobia and non-rhizobia. This led us to devise an in silico method to extract genes preferentially associated with rhizobia. We discuss how the novel genes we have identified may contribute to symbiotic adaptation.

[Supplemental material is available online at www.genome.org. The sequence data from this study for *C. taiwanensis* LMG19424 have been submitted to EMBL under accession nos. CU633749 (chromosome 1), CU633750 (chromosome 2), and CU633751 (pRalta).]

Legumes and bacteria, collectively known as rhizobia, cooperate in a symbiosis of major ecological importance, manifested by the development of nodules on the plant roots inside which rhizobia fix atmospheric nitrogen ($N_2$). As a benefit, nodulated legumes grow needless of N-fertilizers whose synthesis and agricultural abuse waste fossil energy and heavily contribute to groundwater pollution. Symbiotic nitrogen fixation instead is environmentally friendly and key to nitrogen cycling on earth.

Rhizobia provide a rare example of bacteria spread over very large phylogenetic distances yet sharing a seemingly very specific biological function, i.e., symbiotic proficiency with legume plants. This raises the questions of whether rhizobia share a common genetic equipment for entering symbiosis with legumes, when and how many times symbiotic adaptation has emerged during evolution, and why should it be restricted to proteobacteria. Genome scrutinizing now helps to answer these questions.

**[6]These authors equally contributed to this work.**
**[7]Corresponding author.**
**E-mail catherine.masson@toulouse.inra.fr; fax 33-5-61-28-50-61.**
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.076448.108.

Most rhizobia belong to α-proteobacteria (abbreviated to α-rhizobia), where they are distributed so far in nine genera and more than 50 species. The complete genome sequences of seven α-rhizobia were available at the time this work was initiated (*Bradyrhizobium japonicum*, *Bradyrhizobium* sp. ORS278, *Bradyrhizobium* sp. BTAi1, *Mesorhizobium loti*, *Rhizobium leguminosarum*, *Rhizobium etli*, *Sinorhizobium meliloti*) (Fig. 1). Rhizobia have also been identified among β-proteobacteria (abbreviated to β-rhizobia), in genera *Burkholderia* and *Cupriavidus* (formerly *Ralstonia*) (Moulin et al. 2001), and are the predominant *Mimosa* symbionts in Asia. The symbiotic ability is so far restricted to a few species within these two genera, whose other members are often plant or human/animal pathogens as well as saprophytes with sometimes exceptional metabolic properties and great potential for bioremediation. Four strains have been completely sequenced within the *Ralstonia–Cupriavidus* branch (Vandamme and Coenye 2004): the plant pathogen *R. solanacearum* GMI1000 (Salanoubat et al. 2002), the metal-resistant *C. metallidurans* CH34 (http://genome.jgi-psf.org/finished_microbes/ralme/ralme.home.html), the pollutant-degrading *C. pinatubonensis* (formerly *R. eutropha* [Sato et al. 2006]) JMP134 (http://genome.jgi-psf.org/finished_microbes/raleu/raleu.home.html), and the


18:1472–1483 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

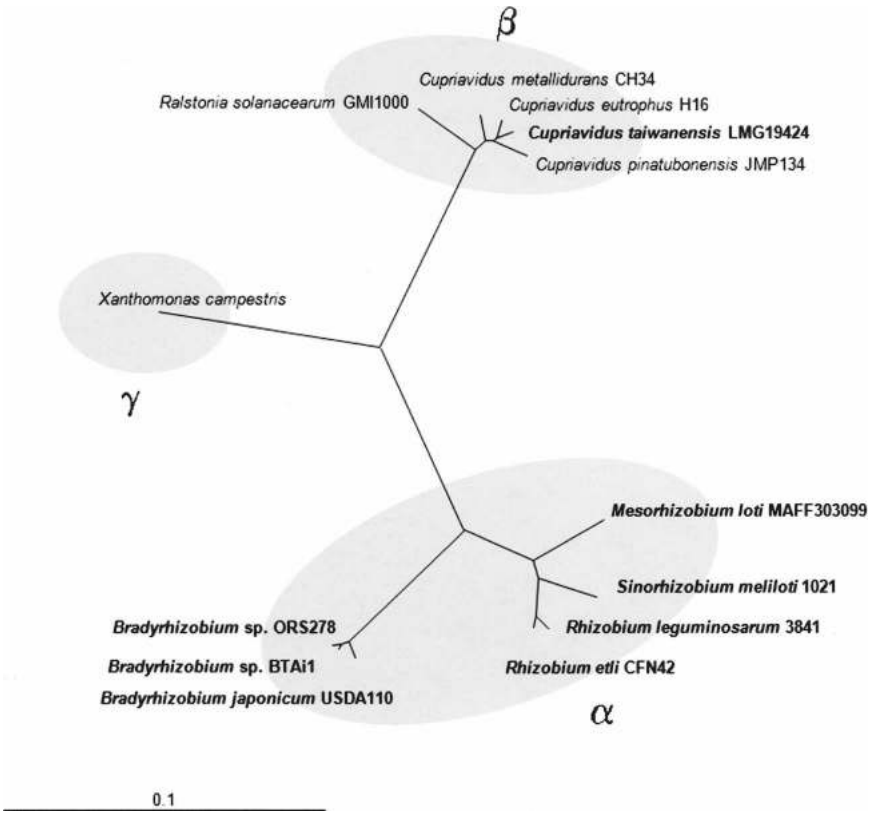


**Figure 1.** Unrooted 16S rDNA tree of completely sequenced rhizobia and *Ralstonia/Cupriavidus* strains. Rhizobia are in bold. α, β, and γ represent different subdivisions of proteobacteria. The tree was constructed by using the neighbor-joining method. 16S rDNA sequences are available in GenBank. For *Ralstonia/Cupriavidus* species, which possess nonidentical 16S rDNA, a consensus sequence was determined with DNAstar.

chemolithoautotrophic *C. eutrophus* (formerly *R. eutropha*) H16 (Pohlmann et al. 2006) (Fig. 1).

Here, we first report on the genome sequence determination and analysis of the β-proteobacterial rhizobium *Cupriavidus taiwanensis* LMG19424 (Chen et al. 2001) with emphasis on its symbiotic properties. Second, this phylogenetically distant rhizobial genome prompted us to initiate an original comparative genomics analysis of all available rhizobia. We discuss implications of our findings for understanding the rhizobium–legume symbiosis.

## Results

### The *Cupriavidus taiwanensis* genome

#### *General features*

*C. taiwanensis* LMG19424 has a genome of 6,476,523 bp, organized in three circular replicons of 3,416,912 bp (chromosome 1), 2,502,411 bp (chromosome 2), and 557,200 bp (pRalta plasmid). This is consistent with earlier electrophoretic data on this strain (Chen et al. 2003). The overall genomic organization is thus comparable to that of other *Cupriavidus* species, with two main circular replicons and additional plasmid(s) (Supplemental Table S1). The *C. taiwanensis* genome is significantly smaller than the three other *Cupriavidus* genomes available. General features of the replicons are listed in Table 1, highlighting strong differences between pRalta and the two chromosomes in terms of coding regions, G+C content, and abundance of transposases. Five complete and different rRNA operons were identified, three on chromosome 1 and two on chromosome 2.

The putative chromosome 1 and 2 replication origins have the same characteristics as those of *C. eutrophus* H16 (Pohlmann et al. 2006). Their location was confirmed by GGGNAGGG motif (KOPS) distribution analysis (Supplemental Fig. S1; Bigot et al. 2005; Young et al. 2006). The pRalta replication/maintenance cluster includes *parABrepAB* genes highly homologous to those of *Cupriavidus*, *Ralstonia*, and *Burkholderia* plasmids, indicating a common mode of replication. However, the *parABrepAB* gene cluster is framed by transposases, thus suggesting that it may have recombined on a plasmid of foreign origin. In addition, the origin of replication of pRalta could not be located precisely by cumulative GC skew analysis, nor by KOPS distribution analysis.

The distribution of genes in product type and major functional categories is listed in Supplemental Table S2.

**Table 1.** General features of the *C. taiwanensis* LMG19424 and *C. eutrophus* H16 genomes

| | LMG19424 | | | H16 | | |
|---|---|---|---|---|---|---|
| **Feature** | **C 1** | **C 2** | **pRalta** | **C 1** | **C 2** | **pHG1** |
| Size (bp) | 3,416,912 | 2,502,411 | 557,200 | 4,052,032 | 2,912,490 | 452,156 |
| G+C content (%) | 67.51 | 67.91 | 59.66 | 66.4 | 66.7 | 62.3 |
| Percentage coding | 89.2 | 89.4 | 83.9 | 88.1 | 88.6 | 79.7 |
| tRNA genes | 57 | 6 | 0 | 51 | 7 | 1 |
| rRNA operons | 3 | 2 | 0 | 3 | 2 | 0 |
| Transposable elements | 5 | 10 | 207 | 4 | 3 | 7 |
| Total no. of CDS | 3145 | 2254 | 583 | 3651 | 2555 | 420 |
| Nb (%) CDS with assigned functions | 2385 (75.8%) | 1508 (66.9%) | 399 (68.4%) | 2382 | 1618 | 225 |
| Nb (%) CDS with unknown functions | 734 (23.3%) | 704 (31.2%) | 183 (31.4%) | 841 | 680 | 157 |
| Average length of genes (kb) | 969 | 993 | 802 | 916 | 938 | 725 |

CDS, coding sequences.

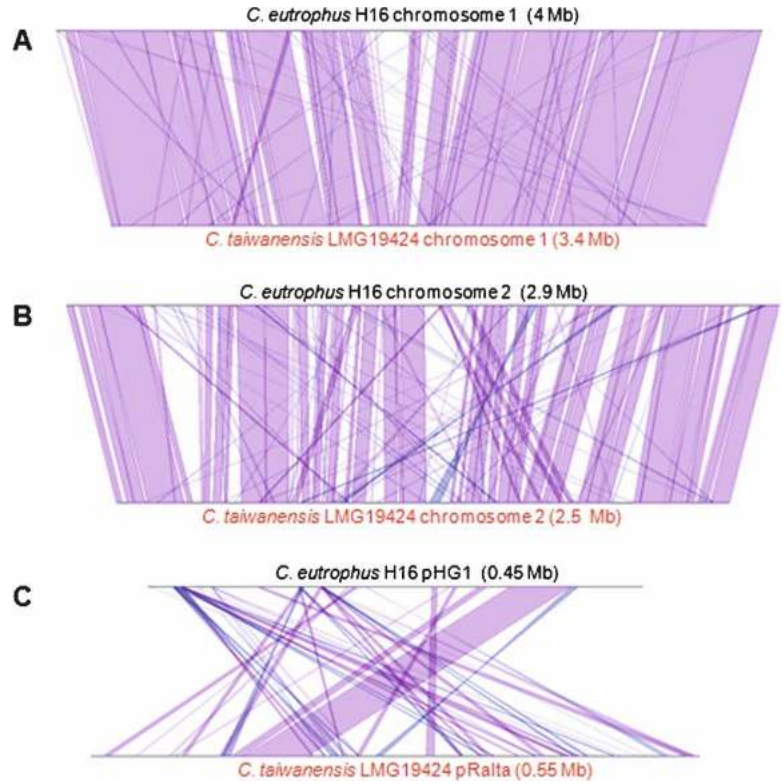


**Figure 2.** Synteny plots between the *C. taiwanensis* and *C. eutrophus* genomes. Conserved gene clusters, i.e., synteny groups, were computed according to Vallenet et al. (2006). Synteny groups containing a minimum of five (*A*,*B*) or three (*C*) genes are shown for colinear regions (purple), and for inverted regions (blue). The display has been obtained using the MaGe graphical interface of the CupriaviduScope project (https://www.genoscope.cns.fr/agc/mage).

### *The* C. taiwanensis *genome closely resembles that of* C. eutrophus *H16*

BLASTP comparisons of *C. taiwanensis* LMG19424 with the genomes of the three other sequenced *Cupriavidus* spp. indicated that the genome of *C. eutrophus* H16 was closest to that of *C. taiwanensis* (Fig. 2; Supplemental Fig. S2), in agreement with 16S rRNA phylogeny (Fig. 1). Seventy-two percent of *C. taiwanensis* genes have likely orthologs in the *C. eutrophus* genome, 71% of which are organized in groups of neighbor orthologs (i.e., synteny groups), although the level of synteny is dependent on the replicon. *C. taiwanensis* and *C. eutrophus* share common core chromosomes 1 and 2 of 3.09 Mb (2751 genes) and 1.8 Mb (1507 genes), respectively, interspersed with specific regions (SR) in each chromosome of each genome (Fig. 3; Supplemental Table S3). Few rearrangements of the core genome have occurred between the two species for chromosomes 1, while more were observed for chromosomes 2. On *C. taiwanensis* chromosome 1, SR total ~323 kb, distributed in 43 regions from <1 kb up to 47 kb. For chromosome 2, *C. taiwanensis* SR total ~704 kb, distributed in 97 regions from <1 kb up to 76 kb.

Most of the *C. taiwanensis* SR display characteristics of laterally acquired DNA including GC deviation and presence of a mobile element or tRNA gene nearby, suggestive of exogenous DNA insertion in the core backbone (Supplemental Table S3). Notably, 25% of the chromosome 1 SR are located adjacent to a tRNA, suggesting that tRNA insertion has extensively contributed to genetic remodeling of the main chromosome in this bacterium. The paucity of integrase genes, repeat structures, and mobile characters suggests that SR have become permanently integrated into the host genome. Interestingly, SR often occur at the same positions in *C. eutrophus* and *C. taiwanensis* (see Fig. 3), indicating that both chromosomes have expanded from a common backbone. *C. taiwanensis* SR total 800 kb less than *C. eutrophus* SR, suggesting that *C. taiwanensis* has undergone less genome expansion than *C. eutrophus*.

In contrast to the similarities between the chromosomes, the pRalta plasmid appears very different to its pHG1 counterpart in H16 (Schwartz et al. 2003). pRalta and pHG1 display very low levels of synteny, except a 57.9-kb region encompassing a *pil*, *trb*, and *tra* locus related to pilus biogenesis and conjugative plasmid transfer (Monchy et al. 2007; Fig. 3). This suggests that, like pHG1 (Friedrich et al. 1981), pRalta is a self-transferable plasmid. Absence of the pHG1 plasmid that encodes key enzymes of $H_2$-based lithoautotrophy and anaerobiosis in *C. eutrophus* implies that *C. taiwanensis* has reduced metabolic properties compared with *C. eutrophus* (for a detailed account of *C. taiwanensis* metabolic properties, see Supplemental Data S1).

In the context of this publication, emphasis will be put on the chromosomal SR and on the symbiotic cluster on pRalta, as these regions most likely account for differences in lifestyle between the symbiotic *C. taiwanensis* LMG19424 and the free-living *C. eutrophus* H16.

### *Nodulation and nitrogen fixation*

In most rhizobia, the biosynthesis and transport of Nod Factors (NF), the signaling molecules that induce nodule organogenesis, are encoded by nodulation genes (*nod*, *nol*, *noe*). *C. taiwanensis* carries 10 nodulation genes, *nodBCIJHASUQ*, and one regulatory gene, *nodD*, on pRalta (Supplemental Table S4) that are tightly clustered in a single 10-kb region (Fig. 4). The *nodBCIJHASUQ* genes are probably arranged in a single operon preceded by a *nod*-box and transcribed divergently from the unique *nodD* gene, the product of which binds *nod*-boxes and activates *nod* gene expression in rhizobia. In contrast to other rhizobial NodQ that are bifunctional enzymes appearing as a fusion of CysN (large subunit of ATP sulfurylase) and CysC (APS kinase/adenylyl sulfate kinase), the *C. taiwanensis nodQ*-like gene consists only of a *cysC* homolog. A bifunctional CysNCysC enzyme (RALTA_B0475), however, is present on chromosome 2 in addition to the chromosome 1 *cysN* gene (RALTA_A2469). No extrachromosomal copies of *glmS* (glucosamine:fructose-6-phosphate aminotransferase) and *cysD* (ATP-sulfurylase small subunit), referred to as *nodM* and *nodP*, respectively, in rhizobia, were found, suggesting that the chromosomal genes participate in NF biosynthesis. *nolFG* homologs (secretion protein and efflux transporter) are present on the plasmid 12 kb away from the *nif* cluster (see

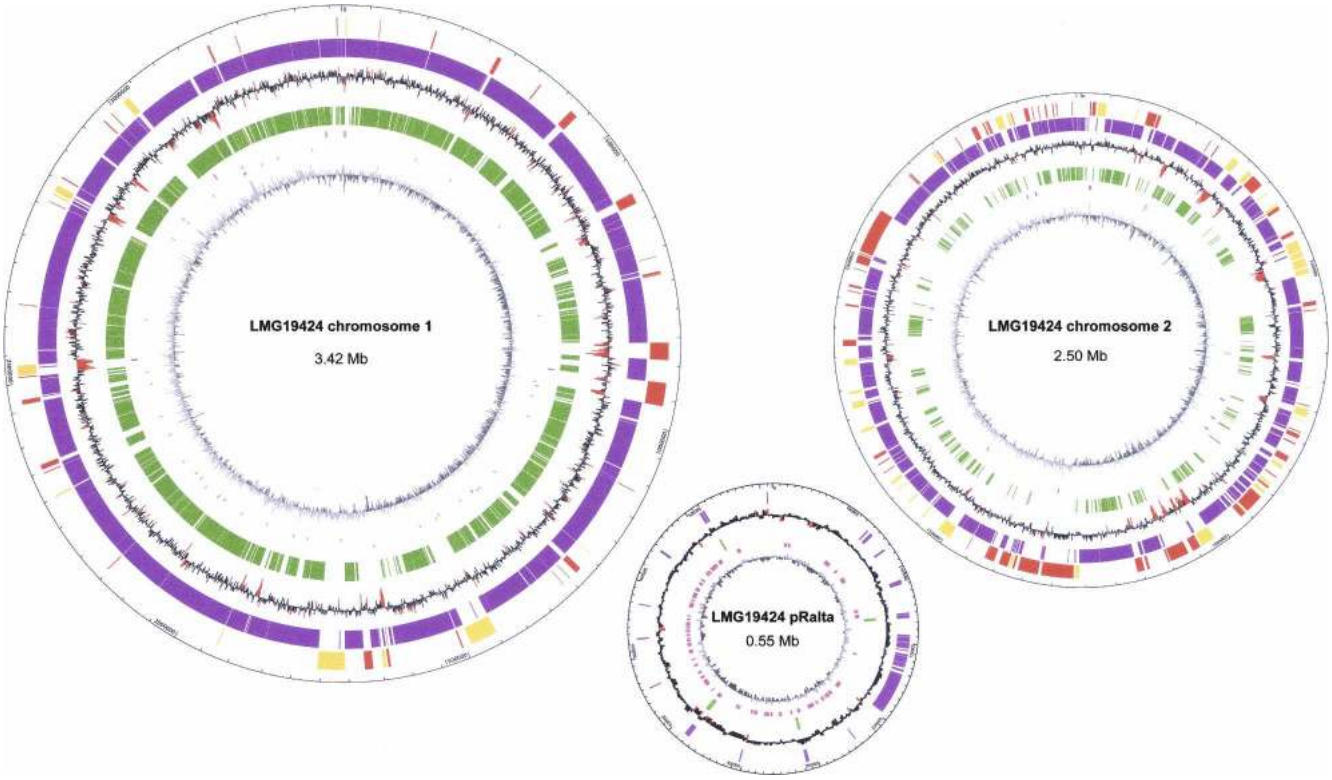


**Figure 3.** Circular representation of the three replicons of *C. taiwanensis* LMG19424. Circles, from the *inside out*, show: (1) GC skew; (2) tRNA and rRNA (green) and TE (pink); (3) CDS common to all *Cupriavidus* (green); (4) GC deviation (red when significantly different from the mean); (5) common CDS to *C. taiwanensis* LMG19424 and *C. eutrophus* H16 (purple); (6) for chromosomes 1 and 2 only, SR of *C. taiwanensis* versus *C. eutrophus* (yellow when a SR occurs at the same location in H16, red otherwise). Circles were drawn using GenVision software (http://www.dnastar.com/products/genvision.php).

below) but are not preceded by a *nod*-box. We predicted only two *nod*-boxes, one located in the *nodB–nodD* intergenic region, the other one located upstream of a putative sterol desaturase (pRALTA_0471) of unknown function (Fig. 4). The nodulation genes of *C. taiwanensis* are induced by the flavonoids luteolin and apigenin (M. Glew and C. Masson-Boivin, unpubl.), which are classical inducers of rhizobial *nod* genes. *nodBCIJHASUQ* are responsible for the synthesis of the core Nod factor and the adjunction of *O*-carbamoyl (*nodU*), *N*-methyl (*nodS*), and 6-*O*-sulfate (*nodH*) groups. Consistent with in silico prediction, Nod factors produced by *C. taiwanensis* LMG19424 were found to be pentameric chito-oligomers sulfated at the reducing end and *N*-acylated by vaccenic acid ($C_{18:1}$) or palmitic acid ($C_{16:0}$) on their non-reducing end (Fig. 4). Most of the molecules are substituted by a *N*-methyl and a carbamoyl group at the terminal non-reducing sugar. Although these structures are very similar to that produced by *Acacia*-nodulating α-rhizobia (Lorquin et al. 1997; Ba et al. 2002), *C. taiwanensis* has a restricted host range as compared with *Acacia* rhizobia (C. Masson-Boivin, unpubl.), suggesting that other determinants may limit its host range.

Next to *nod* genes, we identified 19 genes, presumably arranged in five operons and covering 25 kb, that are involved in nitrogenase synthesis and functioning (Supplemental Table S4). These operons are *nifA*, encoding a sigma 54-dependent regulator, *nifENfdxBnifQ* and a modified *nifX*, *nifVWfixABCX*, *nifBfdxNnifZfixU*, and the nitrogenase structural genes *nifHDK*. In all rhizobia studied so far, energy generation under the oxygen-restricted condition of the nodule involves a cytochrome oxidase with a high affinity for oxygen (*cbb3* type) that is also widespread in proteobacteria. A single copy of the *ccoNOQP* operon encoding this oxidase and of the associated copper-transport complex *ccoGIS* were found on the *C. taiwanensis* main chromosome, whereas it is often located in symbiotic regions in other rhizobia. Synthesis of this *ccb3* oxidase under low oxygen probably depends on the nearby *fnr* gene (RALTA_A1857) since no *fixLJ*-like genes, which drive *ccoNOQP* expression in many α-rhizobia, were found in *C. taiwanensis*. This is consistent with previous evidence that the FixLJ two-component system is restricted to α-proteobacteria (Cosseau and Batut 2004).

pRalta, which carries the *nod–nif* symbiotic cluster, has a typical β-proteobacterial gene replication cluster (although framed by transposases), but contains a huge number of transposable elements (TE) compared with pHG1 and other *Cupriavidus* plasmids. Indeed, 207 TE (36% of the plasmid CDS, 93% of total TE) distributed in 12 IS (insertion sequence) families are spread all along the replicon. *nod–nifA* and *nif–fix* clusters are separated and surrounded by transposable elements, indicating their acquisition via lateral transfer. Both clusters have GC values (52%) much lower than the plasmid and genome means (59.60% and 67.3%, respectively). Altogether, these data suggest that the pRalta plasmid has been the theater of numerous lateral transfer events or rearrangements, remodeling an ancestral replicon into a symbiotic plasmid. Accumulation of IS and IS-related elements has already been reported for symbiotic regions or symbiotic plasmids of other rhizobia (Viprey et al. 2000; Capela et al. 2001; Gonzalez et al. 2006).

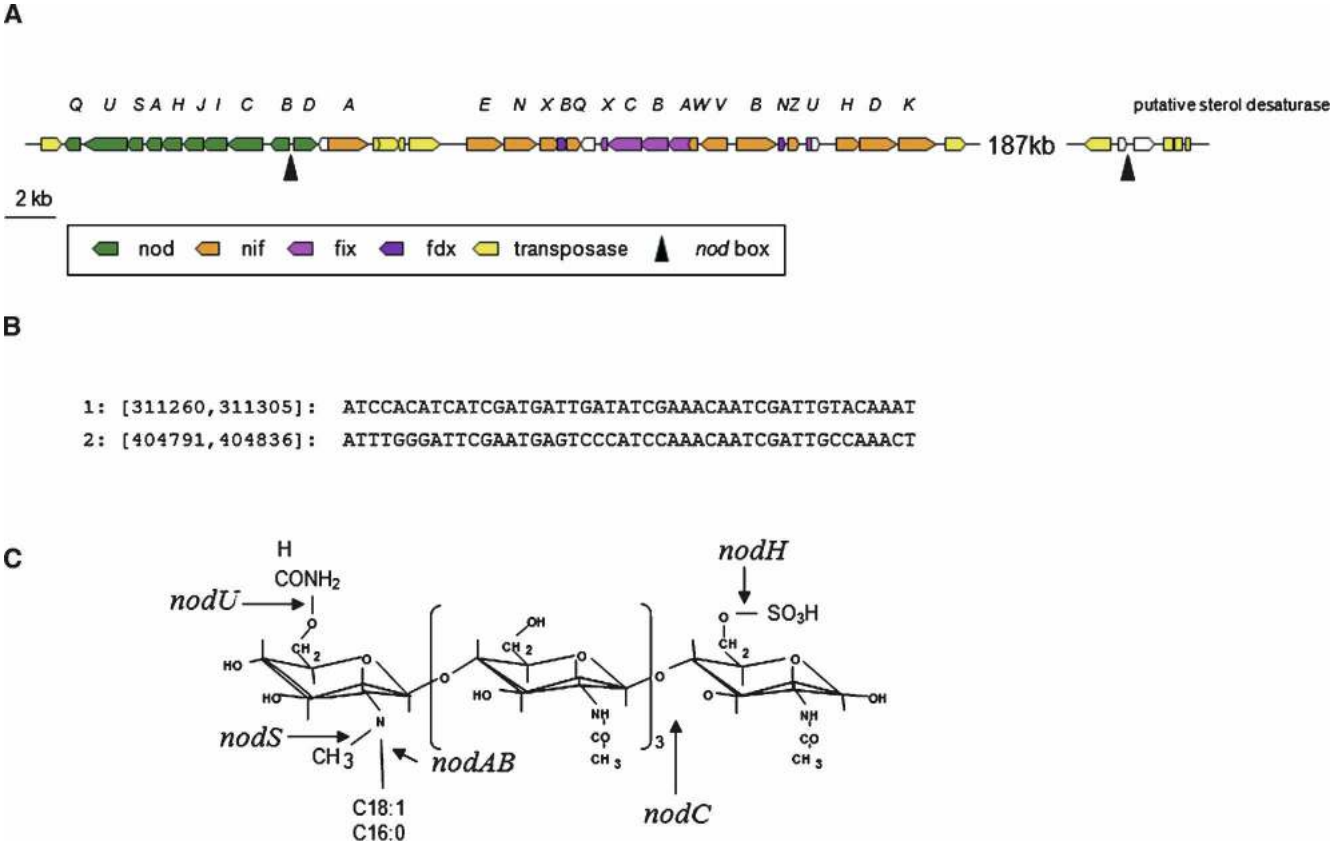


**Figure 4.** Symbiotic features of *C. taiwanensis*. (*A*) Map of symbiotic regions. Genes are colored according to their name. (Arrows) The location of potential *nod*-boxes. (*B*) Sequence and position on pRalta of predicted *nod*-boxes. (*C*) Chemical structure of Nod factors produced by *C. taiwanensis*. (Arrows) Contribution of Nod proteins to NF biosynthesis based on the literature.

### *Additional functions related to interactions with eukaryotes*

In addition to the symbiotic cluster on pRalta, our analysis indicated that SR may contribute to the interaction with the plant for many of them host genes associated with bacteria–eukaryote interactions (Supplemental Tables S3, S5). Several LPS/EPS loci for polysaccharide biosynthesis and transport (SR24, SR32, SR46, SR112, SR119) as well as a *kps* locus (SR88) map on chromosomes 1 and 2. Exopolysaccharide and polysaccharide genes have been shown to play a major role in rhizobial infection, probably by preventing plant defense responses (Jones et al. 2007). A non-flagellar type three secretion system (T3SS), which is the principal arsenal of pathogens (Hueck 1998), is present on chromosome 2 (SR92). Such a system is present in some rhizobia, where it perturbs host defenses and modulates host specificity (Bartsev et al. 2004). Surprisingly, its organization in *C. taiwanensis* differs from that of both *R. solanacearum* (Genin and Boucher 2004) and α-rhizobia (Marie et al. 2001) (Supplemental Fig. S4). Furthermore, none of the effector proteins identified in *R. solanacearum* or in rhizobia have been detected in *C. taiwanensis* by BLASTP analysis. The *C. taiwanensis* T3SS most closely resembles that of a human opportunist, *Burkholderia cenocepacia* (Glendinning et al. 2004), suggesting a common origin for these two secretion systems (Supplemental Fig. S4). *C. taiwanensis* T3SS structural genes are expressed in nodules, indicating that this system may contribute to symbiosis (M. Saad and C. Masson-Boivin, unpubl.). In addition, a 38-kb prophage of the P2 family is present on chromosome 1 (SR13) that closely resembles the virulence-associated Phi CTX phage of *Pseudomonas aeruginosa* (Nakayama et al. 1999). The *C. taiwanensis* prophage contains several unique CDS, suggesting that the phage may have acted as a vehicle for foreign DNA. Chromosome 2 carries a phospholipase D gene (*pld2*, SR85) whose homolog in *P. aeruginosa* (*pldA*) has a role for persistence in a chronic infection model (Wilderman et al. 2001). In *P. aeruginosa*, *pldA* is associated with a *vgr* gene and an ORF4 of unknown function, which together constitute a putative mobile module (Wilderman et al. 2001). A similar organization was found in *C. taiwanensis*, except that three copies of the ORF4 homolog (RALTA_B1036, RALTA_B 1040, RALTA_B 1041) were identified downstream of *pld2*. Two non-ribosomal peptide synthases were also found to be specific to *C. taiwanensis*, compared with H16 (SR91).

Besides the *C. taiwanensis*-specific elements described above, additional virulence-associated determinants were shared by the *C. taiwanensis* and *C. eutrophus* H16 genomes (Supplemental Table S5). The biological significance of this is unclear, as neither *C. taiwanensis* nor *C. eutrophus* have known pathogenic capabilities, even though they are phylogenetically close to pathogens (Chen et al. 2001). This set of genes includes a type six secretion system (T6SS) encoded by a 27-kb locus on chromosome 1 with high similarity to the T6SS of *P. aeruginosa* and *Salmonella enterica* (Folkesson et al. 2002; Mougous et al. 2006). T6SS are conserved in numerous Gram-negative pathogens and symbionts (including *R. leguminosarum*) where they are involved in the interaction

with eukaryotic host cells (Bladergroen et al. 2003; Parsons and Heffron 2005; Pukatzki et al. 2006). Eight Mip-like (macrophage infectivity potentiators) genes have been found on chromosomes 1 (seven) and 2 (one). Mip proteins, which display peptidyl-prolyl *cis-trans* isomerase activity, are common virulence factors of many (human) pathogens (Zang et al. 2007). Interestingly, *C. taiwanensis* possesses most of the genes of the complex network that regulates virulence and pathogenicity in *R. solanacearum*: the *phcAphcBRSQ* system, the *pehRS*, *vsrABCD*, and *xpsR* regulators (Schell 2000). However, most virulence genes controlled by these regulators in *R. solanacearum*, such as *pglA*, *pme*, *pehBC*, encoding pectinolytic enzymes, *egl* and *cbhA*, encoding cellulolytic enzymes, and most *eps* genes, encoding the exopolysaccharide I production, are absent in *C. taiwanensis*, except for the *tek* gene, encoding an exoprotein of unclear function. *C. taiwanensis*, however, possesses several other hydrolytic enzymes.

## Comparative genomics of rhizobia

The availability of the *C. taiwanensis* genome sequence opened new perspectives for the overall comparative genomics of rhizobia. Indeed, because *C. taiwanensis* was the first β-rhizobium genome to be known, it provided the long-needed output reference to fruitfully compare α-rhizobia genomes. Comparative genomics of rhizobia was achieved in three steps: (1) We probed for the existence of a common and specific core genome of rhizobia, (2) we took advantage of the model rhizobium *Sinorhizobium meliloti* to assess the phylodistribution pattern(s) of known symbiotic genes in the prokaryotic world, and (3) we devised an in silico method to identify genes preferentially associated with rhizobia.

### *Core genome of rhizobia*

By Bi-directional Best Hit (BBH) analysis, 1258 genes were found to be common to α-rhizobia, and 647 were common to both α- and β-rhizobia (see Methods). After removing from these sets of genes those that were present in any other α- or β-proteobacterium, we found that no gene was both common and specific to either α-rhizobia or α- and β-rhizobia, showing that there is no common and specific genetic determinism for legume symbiosis. This suggested first that rhizobia may use different strategies for symbiosis, and second that some rhizobial symbiotic genes may be shared by non-rhizobia. This prompted us to analyze the phylodistribution of known symbiotic genes in a model rhizobium.

### *Phylodistribution of* S. meliloti *symbiotic genes*

*Sinorhizobium meliloti* is the rhizobium in which the most symbiotic genes have been identified to date over two decades of genetic analysis. We selected from the NodMutDB database (http://nodmutdb.vbi.vt.edu/; Mao et al. 2005) 214 individual genes experimentally demonstrated as serving a symbiotic function in this model rhizobium (Supplemental Table S6). We searched for homologs of each of these genes in the 400 bacterial and archeal genomes available in July 2007, taking into account the large phylogenetic distance between them (see Methods). Hierarchical clustering of the genes grouped them into seven clusters of size eight to 73 that highlighted three very distinct patterns of phylodistribution (Supplemental Figs. S5, S6).

Forty-one percent of *S. meliloti* symbiotic genes, corresponding to clusters V (14 genes), VI (32 genes), and VII (43 genes), are widespread in the prokaryotic world except in obligatory parasites/symbionts such as *Mycoplasma*, *Rickettsia*, *Wolbachia*, and *Buchnera*. These three clusters include metabolic and housekeeping genes (e.g., *ilvC*, *pdhAa*, *pdhAb*, *mdh*, *pckA*, *ppiB*, *hemA*, *ccmAC*, *cycLJK*, *relA*, *tig*, *greA*, *purl*, *ntrC*, *glnB*, *glgA1*, *pgm*, *queA*, *acpxL*, *lpsL*), transporters (*phoUBC*, *sitABCD*, *dctA*, *exsA*), and stress resistance genes (*typA*, *rpoH2*, *catC*, *katA*, *sodB*, *lon*). A number of nodulation genes appear in these clusters as paralogs of housekeeping genes, e.g., *nodM* (*glmS*), *nodE* (*fabF*), *nodG* (*fabG*), *nodP* (*cysD*), *nodQ* (*cysNcysC*). A few nodulation genes (*nodI*, *nodL*) and several genes involved in polysaccharide synthesis (*exoY*, *exoB*, *exoN*, *expD1*, *expA7A8A9A10*, *chvI*, *ndvA*, *rkpH*) have widespread homologs.

At the other extreme, and unexpectedly, the 73 genes from cluster I are extremely scarce in the prokaryotic world, and, furthermore, 25 of them (34%) were even specific to *S. meliloti*. The majority of the genes of cluster I encode either nodulation functions (*nodA*, *nodC*, *nodF*, *nodH*, *noeAB*, *nolF*) or the synthesis of succinoglycan and galactoglucan exopolysaccharides (*syrA*, *syrB*, *syrB2*, *exoHK*, *exoLAMO*, *exoP*, *exoI*, *exoX*, *exoF1QZ*, *exoT*, *exoI*, *exoW*, *exoX*, *exoU*, *exoD*, *exoR*, *expC*, *expA1*, *exp4–A6*, *expA23*, *expE2–E8*, *expG*) that are crucial for *Medicago* infection by *S. meliloti*. The key nodulation gene *nodA* belongs to this cluster; *nodA* and *nodH* are the two genes that have been exclusively found in rhizobia so far. *nodA* is specific to rhizobia but is not present in all of them, for it is missing in photosynthetic bradyrhizobia as reported recently (Giraud et al. 2007). *nodH* is present in only two instances in available complete genomes so far, i.e., in *S. meliloti* and *C. taiwanensis*. Indeed, both of these rhizobia synthesize sulfated nodulation factors (see Fig. 4).

The remaining genes corresponding to clusters II, III, and IV are of intermediate abundance in prokaryotes, with different patterns of distribution. Cluster IV gathers seven *nif* genes involved in nitrogenase synthesis and functioning as well as related *fixU* and *fdxB* genes. Genes of this cluster appear sporadically in the prokaryotic world, in both free-living and symbiotic $N_2$-fixing organisms. Cluster III (eight genes) that gathers *fix(cco)NOP* and *fix(cco)GS* driving the synthesis of a high-affinity cytochrome oxidase is almost specific for proteobacteria. Lastly, Cluster II gathers 35 genes that are poorly represented in the prokaryotic world with a variety of phylodistribution patterns. This cluster includes *fixABC* genes that together with *fix(cco)NOP* and *fix(cco)GS* are present in all rhizobia.

Eighty-nine *S. meliloti* symbiotic proteins distributed across all seven clusters have orthologs (BBH) in the genome of *C. taiwanensis* (Supplemental Table S6); thus, they are good candidates for serving symbiotic functions, even though most of them (except the nodulation and nitrogen fixation genes and the *cycH*, *rkpJ*, *lpsC*, *catC*, *exoB*, *expD1*, *nolF* genes) were also predicted in the genome of the nonsymbiotic organism *C. eutrophus* H16 by BLAST analysis (Supplemental Table S6).

The above analysis had two implications. First, it shed new light on the evolution of symbiotic functions in rhizobia, as discussed below; second, it had practical consequences for the in silico comparative genomics of rhizobia. Clearly, genes that are either ubiquitous or unique to one rhizobium are not prone to predictive comparative genomics, contrary to genes showing intermediate patterns of distribution (e.g., clusters II, III, and IV). Specifically, we noticed that some genes belonging to this last category (e.g., *nodBC*) appeared overrepresented in rhizobia. This prompted us to make the analysis described below.

### *Genes preferentially associated with rhizobia*

We designed a protocol to identify genes preferentially associated with rhizobia, on a statistically sound basis, using either *S. meliloti* or *C. taiwanensis* as a reference genome.

Briefly, we identified in all prokaryotic genomes gene products having a significant match (see Methods) to the complete proteome of either *S. meliloti* or *C. taiwanensis* taken as a reference genome. Then the relative abundance of these genes in rhizobia versus all non-rhizobia was computed. As the number of non-rhizobial genomes (392) far exceeded that of rhizobial genomes (eight), appropriate terms of correction were included in the statistical treatment (see Methods). As an output, we selected genes that were preferentially associated with rhizobia at a confidence level of 99% ($P < 0.01$). Hierarchical clustering of the selected genes was then performed so as to determine their phylodistribution.

With *S. meliloti* being taken as the reference genome, we ended up with 133 genes overrepresented in rhizobia (Supplemental Table S7; Table 2). These genes grouped into four clusters featuring different phylodistribution (Supplemental Fig. S7). The overall picture shows that the applied in silico procedure was successful in selecting genes that were overrepresented in rhizobia, compared with other bacteria. The majority (96) of these genes were of unknown function. As expected, 12 key symbiotic genes involved in nodulation (*nodA*, *nodC*, and *nodD*) and $N_2$ fixation (*nifDKE*, *nifN*, *nifB*, *fixA*, *fixC*, *fixX*, and *fdxB*) were selected by this procedure. In addition to *nodA*, three genes were found to be strictly specific for rhizobia: SMb21483, SMc03842, and SMa1329 (Supplemental Table S7). BLAST analysis against incomplete genomes at NCBI confirmed SMa1329, a predicted proline peptidase, to be specific for rhizobia. In addition, neighbor genes (SMa1332, SMa1334, SMa1337) were also identified as being rhizobia-associated. Distant homologs of the small proteins of unknown function SMb21483 and SMc03842 were found in incomplete non-rhizobial genomes, thus indicating that these proteins might not be strictly rhizobium-specific, but still overrepresented in rhizobia. Among other genes significantly overrepresented in rhizobia, we found three adenylate cyclase genes, *cyaH*, *cyaG1*, *cyaG2*, which have a similar domain organization. Transcriptome studies have shown that *cyaH* expression was enhanced in nodules as compared with free-living conditions (Capela et al. 2006). Among other genes of interest were *nthA* and *nthB*, which encode enzymes contributing to the synthesis of the

**Table 2. Genes common to *S. meliloti* and *C. taiwanensis* and overrepresented in rhizobia**

| *S. meliloti* gene | Product | Non- rhizobia | α-Rhizobia | *C. taiwanensis* gene |
|---|---|---|---|---|
| *fdsD* | Putative NAD-dependent formate dehydrogenase delta subunit protein | 36 | 7 | *fdsD* |
| *fdxB* | Ferredoxin III | 53 | 7 | *fdxB* |
| *fixA* | Electron transfer flavoprotein beta chain | 42 | 7 | *fixA* |
| *fixC* | Oxidoreductase | 41 | 7 | *fixC* |
| *fixX* | Ferredoxin-like protein | 43 | 7 | *fixX* |
| *gst10* | Putative glutathione s-transferase protein | 23 | 6 | *gst*-like |
| *nifB* | Fe–Mo cofactor biosynthesis protein | 43 | 7 | *nifB* |
| *nifD* | Nitrogenase Fe–Mo alpha chain | 39 | 7 | *nifD* |
| *nifE* | Oxidoreductase | 38 | 7 | *nifE* |
| *nifK* | Nitrogenase Fe–Mo beta chain | 38 | 7 | *nifK* |
| *nifN* | Nitrogenase Fe–Mo cofactor biosynthesis protein | 33 | 7 | *nifN* |
| *nodA* | N-acyltransferase | 0 | 5 | *nodA* |
| *nodC* | N-acetylglucosaminyltransferase | 10 | 5 | *nodC* |
| *nodD2* | *nod*-box-dependent transcription activator | 5 | 5 | *nodD* |
| *phnG* | Putative C–P (carbon–phosphorus) lyase component protein | 44 | 7 | *phnG* |
| *phnJ* | Putative C–P (carbon–phosphorus) lyase component protein | 51 | 7 | *phnJ* |
| *pilA1* | Putative pilin subunit protein | 50 | 7 | *flp-1* |
| SMa0241 | Hypothetical protein | 44 | 7 | RALTA_A1483 |
| SMa0247 | Hypothetical protein | 25 | 7 | RALTA_A1656 |
| SMa0585 | nrtA-type periplasmic nitrate transport binding protein, probable | 46 | 7 | *nasF* |
| SMa1927 | Hypothetical protein | 24 | 6 | RALTA_B0210 |
| SMb20025 | Hypothetical protein | 24 | 6 | RALTA_A0141 |
| SMb20039 | Putative transcriptional regulator protein | 37 | 7 | RALTA_A2147 |
| SMb20040 | Hypothetical protein transmembrane | 38 | 7 | RALTA_A2151 |
| SMb20114 | Hypothetical protein | 10 | 6 | RALTA_A1236 |
| SMb20216 | Putative epoxide hydrolase protein | 20 | 7 | RALTA_A1627 |
| SMb20865 | Hypothetical protein | 20 | 6 | RALTA_B0852 |
| SMb21004 | Hypothetical protein | 27 | 6 | RALTA_A1752 |
| SMb21292 | Conserved hypothetical membrane protein, paralog of y20848 | 45 | 7 | RALTA_A0994 |
| SMb21379 | Hypothetical protein | 54 | 7 | RALTA_B2206 |
| SMb21691 | Putative nitrilotriacetate monooxygenase component a protein | 54 | 7 | RALTA_B2001 |
| SMc00080 | Hypothetical protein | 52 | 7 | RALTA_B2187 |
| SMc00148 | Hypothetical protein | 38 | 7 | RALTA_A2272 |
| SMc00520 | Hypothetical protein | 41 | 7 | RALTA_A0191 |
| SMc01162 | Hypothetical protein | 44 | 7 | RALTA_A1535 |
| SMc01236 | Hypothetical protein | 49 | 7 | RALTA_A0022 |
| SMc01986 | Hypothetical transmembrane protein | 9 | 5 | RALTA_B1778 |
| SMc02514 | Putative periplasmic binding abc transporter protein | 35 | 7 | RALTA_A2001 |
| SMc03176 | Hypothetical protein | 52 | 7 | RALTA_B1593 |
| SMc04314 | Hypothetical protein | 7 | 5 | RALTA_A2657 |

Genes overrepresented in rhizobia when both *S. meliloti* and *C. taiwanensis* were taken as reference genomes are indicated. The number of non-rhizobial or α-rhizobial genomes that contain a homolog is indicated. Three-hundred-ninety-two non-rhizobial and seven α-rhizobial genomes were considered.

phytohormone indole-3-acetic acid (IAA) in some rhizobium and *Agrobacterium* strains (Kobayashi et al. 1995). It has been suggested that plant-associated bacteria may use this phytohormone to interact with plants as part of their colonization strategy; PhnGJ are C–P lyase components of the phosphonate uptake and degradation pathways that allow *S. meliloti* and other rhizobia to use phosphonates as a phosphate source (Parker et al. 1999). *S. meliloti* requires phosphate for successfully entering symbiosis with legumes (Bardin et al. 1996), and *phn* gene expression is induced under inorganic phosphate (Pi) starvation (Krol and Becker 2004). Although its is highly unlikely that *phn* genes are required for symbiosis under usual laboratory assays conditions (Fahreus medium) where free Pi is plenty, phosphate solubilization from phosphonates might be essential in Pi-depleted soils or rhizosphere environments. A predicted nitrate transporter (*nrtA*, SMa0585) was also selected that could be of interest given the inhibitory effect of nitrates on nodulation and nitrogen fixation in all rhizobia. We also found two glutathione-S-transferase (GST) genes among the genes overrepresented in rhizobia. Protection from oxidative-generated damage is known to be crucial at different stages of the symbiotic interaction, and *S. meliloti* makes wide use of glutathione for detoxification purposes since it has a panoply of 15 GSTs. Gst3 and Gst10 are two members of the family that are overrepresented in rhizobia in which they may specifically detoxify compounds generated during the symbiotic interaction.

When *C. taiwanensis* was taken as a reference genome, 169 genes were preferentially associated with rhizobia (Table 2; Supplemental Table S8; Supplemental Fig. S8). Forty genes were common to the list of *S. meliloti* rhizobia-associated genes described above. These were *nodA*, *nodC*, *nodD*, *nifB*, *nifDKE*, *nifN*, *fixA*, *fixC*, *fixX*, *fdxB*, *phnG*, *phnJ*, *fdsD*, *gst10*, *pilA1*, and 23 additional genes of unknown function. Some of the genes of this common core were highly significantly overrepresented in rhizobia, as high as *nodAC* genes, for example (Table 2). These include genes of unknown function RALTA_A2657/SMc04314, RALTA_B1778/SMc01986, RALTA_A1236/SMb20114, and RALTA_A1627/SMb20216 and genes encoding proteins of the Bug family (UPF0065) (SMa1927/RALTA_B0210, SMb20025/RALTA_A0141), which are very abundant in β-proteobacteria. *C. taiwanensis* has a total of 100 proteins of this family (corresponding to cluster I in Supplemental Fig. S8), whereas *S. meliloti* has only six. Bug proteins are periplasmic solute receptors likely involved in import of carboxylates (Antoine et al. 2003), a common C-source for endosymbiotic rhizobia.

Lastly, 129 genes (169 − 40) were coined as "rhizobia-associated" only when taking *C. taiwanensis* as a reference genome. Forty-nine of them belong to the Bug family, which is indeed overrepresented in some β-proteobacteria as compared with α-proteobacteria. Hence, this high feature of 49 is largely artifactual as it results from many paralogous *C. taiwanensis* query proteins matching a few targets. This is a limitation of the Phydbac-like approach used here in which sequence similarities are assessed only unidirectionally. More interestingly, eight genes (*cobS*, *coxG*, *phnF*, *hyuA*, *hss*, RALTA_A0880, RALTA_A1094, RALTA_A0495) of the *C. taiwanensis* set had been excluded from the *S. meliloti* set as they were also present in at least 70% of the Rhizobiales order to which all α-rhizobia belong (see Methods). Their presence in the *C. taiwanensis*-based set nevertheless validates them and indicates they are very likely bona fide rhizobia-associated genes that may be worth analyzing experimentally in addition to the set of 40 genes described above. It is noteworthy that seven genes among the 129 (*coxO*, *phnI*, RALTA_A1237, RALTA_A1655, RALTA_A1753, RALTA_A2002, RALTA_B2207) are immediate neighbors of some of the 40 "common" genes. Finally, half (32 out of 63) of the remaining genes in the 129 set encoded proteins of unknown function.

## Discussion

The ability to enter into a $N_2$-fixing symbiosis with legumes, which defines rhizobia, is shared by phylogenetically distant bacteria spread over two subclasses of the proteobacteria. *C. taiwanensis* is the first rhizobium of the β-subclass to be sequenced, thus allowing comparison with α-rhizobia. Furthermore, the sequences of *C. taiwanensis* and *C. eutrophus* H16 provide the first instance of a reasonably close symbiotic/nonsymbiotic couple available for genome comparison. The sequence of a saprophytic strain called *Mesorhizobium* BNC1 is also available but actually shows little similarity to the *Lotus* symbiont *Mesorhizobium loti*. It is also noteworthy that, despite the greater divergence of their 16S rRNA sequences (Fig. 1), the genomes of *C. taiwanensis* and *C. eutrophus* are more similar to each other than are the two recently sequenced genomes of the photosynthetic *Bradyrhizobium* strains BTAi1 and ORS278 (Giraud et al. 2007; data not shown).

The *C. taiwanensis* LMG19424 genome is significantly smaller than that of *C. eutrophus* H16 and other available *Cupriavidus* genomes, which have a wide range of metabolic capacities. Specifically, *C. taiwanensis* has reduced metabolic properties compared with *C. eutrophus*, mainly due to the absence of the pHG1 plasmid that encodes key enzymes of $H_2$-based lithoautotrophy and anaerobiosis in *C. eutrophus*. *C. taiwanensis* also shows higher accumulation of IS than other *Cupriavidus*. These two features, i.e., genome size reduction and IS accumulation, often characterize bacteria that have recently passed an evolutionary bottleneck and adapted to a stable environment (Stinear et al. 2007), as we may expect for a plant symbiont.

Although very distant phylogenetically from other rhizobia, *C. taiwanensis* uses the same strategy for nodulation as classical rhizobia. The chemical structure of its Nod Factors (NF) reported here is in full agreement with the genes present in the symbiotic cluster, thus indicating that no structural *nod* gene has been missed during genome annotation. The *nod* cluster is bracketed by transposases, which is consistent with *nod* genes of α- and β-rhizobia being monophyletic and laterally transferred. Strikingly, *C. taiwanensis* exhibits the most minimal symbiotic genome structure and traits among rhizobia: (1) It shows the tightest packing (35 kb) of nodulation and nitrogen-fixation genes described so far for any rhizobium, (2) it likely uses the housekeeping genes *glmS* and *cysD* genes for NF biosynthesis since, for the first time in rhizobia, no paralogous *nod* copy of these genes has been found (*nodM* and *nodP*, respectively), (3) only two *nod*-boxes were identified in the whole genome, (4) it produces very few different NFs, and (5) the *ccoNOP* genes for microaerobic respiration have a chromosomal instead of symbiotic island location. Altogether, this suggests that this symbiont has evolved recently.

Another feature of the *C. taiwanensis* genome was the unexpected occurrence of genes usually associated with virulence, most of them being also present in *C. eutrophus* H16. It should be recalled in this context that a *C. taiwanensis* strain was also retrieved from a clinical human isolate (Chen et al. 2001) and that the *Cupriavidus* species are phylogenetically close to pathogens. This may suggest that this bacterial genus is genetically adapted

to ecological transitions between mutualism and parasitism in either sense.

A long-standing question in the field was whether a core genome for rhizobia existed. Our work provides a pretty clear answer to this question by demonstrating that no gene is both common to all and specific to rhizobia. Thus, a unique shared genetic strategy does not support symbiosis of rhizobia with legumes. For instance, two *nod* genes, *nodA* and *nodH*, are specific for rhizobia, yet they are not present in photosynthetic *Bradyrhizobium*, as demonstrated before (Giraud et al. 2007). *nif* genes, *ccoNOP*, *ccoGIS*, and *fixABC* genes are common to all rhizobia but are found in some non-rhizobia as well.

This finding prompted us to perform a phylogenomic analysis of known symbiotic rhizobial genes. Specifically, we analyzed the distribution of the known symbiotic genes from the model rhizobium *S. meliloti* in the prokaryotic world. Among the many tools available for phylogenomics (de Crecy-Lagard and Hanson 2007), we used a Phydbac-based method of phylogenomic profiling using normalized BLAST scores (Enault et al. 2004) combined with a two-dimensional representation of gene conservation profiles, similar in its principle to that described by Martin et al. (2003). We established for the first time to our knowledge that different symbiotic genes display from a very narrow to a very large phylodistribution in the prokaryotic world. Occurrence of widely distributed genes can be easily rationalized by the fact that symbiosis is life in a specific ecological niche and thus requires appropriate housekeeping genes. In rhizobia, some of these widely distributed genes are paralogs of housekeeping genes, thus indicating that local recruitment of indigenous genes has played a major role during evolution of symbiotic properties. At the other end, we have found that some symbiotic genes, especially those involved in plant infection, have a very narrow host range, some of them being (so far) specific to *S. meliloti*. It is possible that such genes are related to plant host specificity and may have been recruited among genes present in the phylum. Together with the already documented role of lateral transfer in spreading *nod* and *nif* genes, this analysis suggests that evolution of symbiotic properties may result from a complex scenario and that different rhizobia have adopted partially different genetic strategies to become legume symbionts.

We realized as another output of this analysis that some symbiotic genes besides *nod* and *nif* genes tend to be overrepresented in rhizobia compared with non-rhizobia. This led us to a systematic and statistically sound search of overrepresented genes in rhizobial complete genomes. Identification of such genes was not trivial as (1) it required identifying potential orthologs over large phylogenetic distances, (2) genes of the core genome of α-proteobacteria had to be filtered out, and (3) the number of non-rhizobia far exceeded the number of rhizobia, so appropriate statistics had to be applied. An important result of this study is the identification of a class of genes overrepresented in rhizobia whose biological function can now be addressed. This class contains 133 *S. meliloti* and 169 *C. taiwanensis* genes, at least 40 of them being common to both rhizobia. The fact that the key symbiotic *nod* and *nif* genes that we know to be frequently associated with rhizobia had indeed been selected by our in silico procedure attests the success of our analysis. *nifH* was not found from the selection, for we were deliberately very selective in the selection of these genes (both at the 0.9 threshold level and $P < 0.01$; see Methods). Hence, *nifH* and probably additional genes could be faithfully identified as being rhizobia-associated upon relaxing the selection criteria. Besides *nod* and *nif/fix* genes, the overall biological significance of genes preferentially associated with rhizobia must now be addressed experimentally. Data mining of some of the selected genes, however, suggests some clues. Some of them are involved in (1) plant hormone (IAA) synthesis (*nthAB*), which was previously suggested to be key in several plant–microbe interactions; (2) solubilization of phosphate (*phnJG*), a major limiting factor in soils; or (3) the import of nitrate (*nrtA*), a key regulatory metabolite for both nodulation and $N_2$ fixation. These three examples suggest that we may have identified genes favoring symbiotic adaptation. It is worth recalling in this context that the laboratory conditions under which symbiotic proficiency of rhizobia is assayed have little in common with the conditions rhizobia actually face in nature. Thus, screening of mutant libraries in standard laboratory conditions may fail to identify genes that are only beneficial or even crucial in natural conditions. Hopefully, the in silico procedure we have developed may help fill this gap in rhizobial ecology. Finally, some rhizobia-associated genes belong to large paralogous gene families such as glutathione-S-transferases (*gst*) and adenylate cyclases (*cya*). These enzymes are very numerous in at least some rhizobia (*S. meliloti* has 10 *gst*, 26 *cya*). Conceivably, the ones that are rhizobia-associated are the most promising for playing a role in the interaction with the plant.

It is noteworthy that different sets of genes were selected upon using *S. meliloti* or *C. taiwanensis* as a reference genome, besides an overlap of 40 genes, i.e., 25%–30% of each individual set. Such a limited overlap can be explained in different ways. First, the overlap can be extended by relaxing our selection criteria. For example some *phn* genes were exclusively found with *S. meliloti* as a reference genome, others only with *C. taiwanensis*, and some others were common. This suggests that the whole *phn* cluster is actually rhizobia-associated, whatever the rhizobium genome taken as a reference. Second, the two selection procedures were not symmetrical; i.e., *S. meliloti*-associated genes were picked up after excluding genes present in at least 70% of the bacteria of the order "Rhizobiales" to which all α-rhizobia belong, whereas when *C. taiwanensis* was taken as a reference genome, "Rhizobiales" genes were not excluded (see Methods). Third, *S. meliloti* and *C. taiwanensis* may have partly different genetic adaptation to symbiosis as their host plants; i.e., *Medicago* and *Mimosa*, respectively, may provide different chemical and physical environments to which their cognate symbionts have adapted. Altogether, we see the 40 common genes as points of entry for loci or pathways of symbiotic interest. Detailed analysis of each locus or pathway is required to identify the full set of genes involved. Some of the genes might be common to several rhizobia whereas others might be species-specific, as already described for *nod* genes, for example.

We have conducted this search of rhizobia-associated genes on the eight rhizobium genomes presently available. Two others became available during completion of this work (*Azorhizobium caulinodans*, *Sinorhizobium medicae*), while others should be available soon (*Sinorhizobium* sp. NGR234, *Burkholderia phymatum*). Thus, the comparison presented here can be updated and refined as more rhizobial genomes become available. The same type of approach could be conducted on any group of phylogenetically distant bacteria sharing the same biological function (e.g., bioremediation of a given compound) or habitat (e.g., a given plant rhizosphere), especially when the microorganisms of interest are recalcitrant to genetic approaches, or the phenotype is either difficult to screen for or under complex genetic determinism (functional redundancy or QTL). This approach could prove es-

pecially useful for microbial ecology studies and for genome data mining in general as more genome sequences become available.

## Methods

### Bacterial strains

*C. taiwanensis* LMG19424 (type strain) was isolated from root nodules of *Mimosa pudica* in Taiwan (Chen et al. 2001). The sequenced strain (referred to as CBM777 in our culture collection) is available upon request to the corresponding author.

### Shotgun DNA sequencing and genome assembly

The complete genome sequence of *C. taiwanensis* was determined using the whole-genome shotgun method. Three libraries (A, B, and C) were constructed: two of them were obtained after mechanical shearing of genomic DNA and cloning of generated 3-kb and 10-kb inserts into the plasmids pNav (A) and pCNS (B) (pcdna2.1- and pSU18-derived, respectively). DNA fragments of ~30 kb (generated after partial digestion with *Sau*3A) were introduced into the plasmid pBeloBac11 to generate a BAC library (C). Plasmid DNAs were purified and end-sequenced (68,448 clones for A, 23,232 for B, and 8448 for C) by dye-terminator chemistry on ABI3730 sequencers (Applied Biosystems) leading to an average 10-fold coverage. The *phred/phrap/consed* software package (www.phrap.com) was used for sequence assembly and quality assessment. 26,200 additional sequence reactions were necessary for gap closure and sequence polishing that consisted of random sequencing of subclones (for 23,200 sequence reactions) supplemented with 1500 sequences of PCR products and 1300 sequences of oligonucleotide-targeted regions. Final error estimation rate as computed by *phred/phrap/consed* was <0.04 errors per 10 kb.

### Prediction and annotation of CDS

Gene prediction was conducted using the AMIGene software (Bocs et al. 2003) A total of 5986 coding sequences (CDSs) were predicted and assigned a unique identifier prefixed with "RALTA_A" for chromosome 1 (3147 genes), "RALTA_B" for chromosome 2 (2254 genes), and "pRALTA_" for the plasmid (585 genes). This set of CDSs was submitted to automatic functional annotation (Vallenet et al. 2006). Sequence data for comparative analyses were obtained from the NCBI databank (RefSeq and WGS [whole genome shotgun] sections). Putative orthologs and groups of neighbor orthologs (i.e., synteny groups) were computed between *C. taiwanensis* and all other complete genomes, as previously described (Vallenet et al. 2006). All the data were stored in a relational database, called CupriaviduScope, and manual validation of the automatic annotation was performed using the web interface MaGe (Magnifying Genomes) that allows graphic visualization of the *C. taiwanensis* annotations enhanced by a synchronized representation of synteny groups in other genomes chosen for comparisons. Complete sets of automatic and expert annotations are publicly available at http://www.genoscope.cns.fr/agc/mage ("CupriaviduScope" project).

For *nod*-box prediction, the AT(N11)AT(N7)AT(N11)AT motif was searched using PatScan (Dsouza et al. 1997) and confirmed by expert examination.

### Rhizobium core genome identification

Identification of putative orthologs between rhizobial genomes was based on bidirectional best-hit (BBH) identification (Koski and Golding 2001). We used NCBI BLAST version 2.2.17 and applied the following thresholds: *E*-value < $10^{-3}$, identity > 30%, %Q > 60, and %S > 60. The intersection between all possible BBH sets of genes among α-rhizobia or among α- and β-rhizobia determined the α- or α–β rhizobial common core genome. The specific α- or α–β rhizobial core genome was obtained after removing from these sets of genes all proteins that have a BBH-based putative ortholog in any other α- or α- and β-proteobacterial genome.

### Phylogenomic profiling

The complete *S. meliloti* proteome (6199 proteins) or the complete *C. taiwanensis* proteome (5982 proteins) was compared to all open reading frames (ORFs) of the 400 bacterial and archaeal complete genomes available as July 2007, using BLASTP (Altschul et al. 1997). We used the Phydbac-based method of phylogenomic profiling using normalized BLAST scores (Enault et al. 2004). Phydbac scores reflect both the level of similarity between two proteins and the phylogenetic distance between the organisms they come from. For the same level of sequence similarity, the score increases with the phylogenetic distance. As for *S. meliloti*, the conservation scores of proteins in the prokaryotic world ranged from 0 to 4. Low scores often resulted in a partial alignment between proteins. Manual expert examination of several well-known *S. meliloti* proteins indicated that a score greater than 0.9 reflected a likely orthology.

### Genes preferentially associated with rhizobia

We investigated whether some proteins were statistically more present in rhizobia than in other bacteria. We selected proteins from any genomes scoring greater than 0.9 (see above) (thus corresponding to a strong similarity between sequences) using either *S. meliloti* or *C. taiwanensis* as a reference genome, and tested the relative abundance of these proteins in rhizobia versus non-rhizobia. For each protein, we performed a $\chi^2$ test contrasting the number of hits in the eight rhizobia and in the 392 non-rhizobia. The test *P*-values were estimated empirically using 10 million permutations. We used a Bonferroni correction for 6199 tests to assess the significance of the *P*-values. When *S. meliloti* was taken as a reference genome, we finally excluded genes that were present in at least 70% of the members of the "Rhizobiales" order, since all α-rhizobia belong to this order. The aim was to filter out widespread genes in the "Rhizobiales" order that might not be meaningful for our analysis.

### Nod factor purification and characterization

Five-liter cultures were grown in minimal medium (MM) supplemented with 10 mM succinate and 10 µM luteolin overnight ($OD_{600}$ = 1.0). Supernatants were extracted with XAD4 beads (Fluka) or puriss 2-butanol (Fluka) as previously described (Roche et al. 1991). Dry residues were dissolved in ACN/$H_2O$ (20:80) and injected in HPLC (10A chain, Shimazu). The UV trace was recorded at 203 nm. HPLC separations were performed on a semi-preparative C18 column (equisorb ODS2 5 µm, 250 × 8 mm; C.I.L.) using an isocratic step for 10 min with 20% AcCN, followed by a linear gradient running from this solvent to pure acetonitrile for 40 min, at a 2 mL/min flow rate (Poinsot et al. 2001). Peaks were collected and three runs were pooled.

The HPLC fractions were analyzed on a ESI-Q-ToF Ultima (Waters) using direct infusion (solvent AcCN/$H_2O$ 1:1, 1% acetic acid, rate: 10 µL/min). MS settings: E probe, 3 kV; E cone, 100 V; E Rf, 70 V; E coll, 15 V for MS, 45 V for MS/MS; collision gas, Argon. Spectra were recorded in both the positive and the negative mode. Peaks detected in the awaited range (*m/z* 1000–1500 for the simple charged species or 600–750 for the double charged

ones) were submitted to MS/MS analysis to confirm their lipochitooligosaccharidic nature.

## Acknowledgments

We thank Michael Chandler and Patricia Siguier for sharing expertise on insertion sequence elements, and Thomas Faraut, François Enault, and Jérome Gouzy for helpful discussions. We thank Julie Cullimore for critical reading of and suggestions on the manuscript. Sequencing was supported by the Genoscope "Séquençage" program and by INRA, and annotation was partly supported by a grant from ACI IMPBio 2004, MicroScope project. G.P., M.G., and M.S. were supported by a post-doctoral fellowship from INRA, and C.A. was supported by an INRA SPE department short-term contract.

## References

Altschul, S., Madden, T., Schaffer, A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Antoine, R., Jacob-Dubuisson, F., Drobecq, H., Willery, E., Lesjean, S., and Locht, C. 2003. Overrepresentation of a gene family encoding extracytoplasmic solute receptors in *Bordetella*. *J. Bacteriol.* **185:** 1470–1474.

Ba, S., Willems, A., De Lajudie, P., Roche, P., Jeder, H., Quatrini, P., Neyra, M., Ferro, M., Prome, J.C., Gillis, M., et al. 2002. Symbiotic and taxonomic diversity of rhizobia isolated from *Acacia tortilis* subsp. raddiana in Africa. *Syst. Appl. Microbiol.* **25:** 130–145.

Bardin, S., Dan, S., Osteras, M., and Finan, T.M. 1996. A phosphate transport system is required for symbiotic nitrogen fixation by *Rhizobium meliloti*. *J. Bacteriol.* **178:** 4540–4547.

Bartsev, A.V., Deakin, W.J., Boukli, N.M., McAlvin, C.B., Stacey, G., Malnoe, P., Broughton, W.J., and Staehelin, C. 2004. NopL, an effector protein of *Rhizobium* sp. NGR234, thwarts activation of plant defense reactions. *Plant Physiol.* **134:** 871–879.

Bigot, S., Saleh, O.A., Lesterlin, C., Pages, C., El Karoui, M., Dennis, C., Grigoriev, M., Allemand, J.F., Barre, F.X., and Cornet, F. 2005. KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *EMBO J.* **24:** 3770–3780.

Bladergroen, M.R., Badelt, K., and Spaink, H.P. 2003. Infection-blocking genes of a symbiotic *Rhizobium leguminosarum* strain that are involved in temperature-dependent protein secretion. *Mol. Plant Microbe Interact.* **16:** 53–64.

Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Medigue, C. 2003. AMIGene: Annotation of MIcrobial genes. *Nucleic Acids Res.* **31:** 3723–3726.

Capela, D., Barloy-Hubler, F., Gouzy, J., Bothe, G., Ampe, F., Batut, J., Boistard, P., Becker, A., Boutry, M., Cadieu, E., et al. 2001. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc. Natl. Acad. Sci.* **98:** 9877–9882.

Capela, D., Filipe, C., Bobik, C., Batut, J., and Bruand, C. 2006. *Sinorhizobium meliloti* differentiation during symbiosis with alfalfa: A transcriptomic dissection. *Mol. Plant Microbe Interact.* **19:** 363–372.

Chen, W.M., Laevens, S., Lee, T.M., Coenye, T., De Vos, P., Mergeay, M., and Vandamme, P. 2001. *Ralstonia taiwanensis* sp. nov., isolated from root nodules of *Mimosa* species and sputum of a cystic fibrosis patient. *Int. J. Syst. Evol. Microbiol.* **51:** 1729–1735.

Chen, W.M., Moulin, L., Bontemps, C., Vandamme, P., Béna, G., and Boivin-Masson, C. 2003. Legume symbiotic nitrogen fixation by beta-proteobacteria is widespread in nature. *J. Bacteriol.* **185:** 7266–7272.

Cosseau, C. and Batut, J. 2004. Genomics of the *ccoNOQP*-encoded cbb(3) oxidase complex in bacteria. *Arch. Microbiol.* **181:** 89–96.

de Crecy-Lagard, V. and Hanson, A.D. 2007. Finding novel metabolic genes through plant–prokaryote phylogenomics. *Trends Microbiol.* **15:** 563–570.

Dsouza, M., Larsen, N., and Overbeek, R. 1997. Searching for patterns in genomic data. *Trends Genet.* **13:** 497–498.

Enault, F., Suhre, K., Poirot, O., Abergel, C., and Claverie, J.M. 2004. Phydbac2: Improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.* **32:** W336–W339.

Folkesson, A., Lofdahl, S., and Normark, S. 2002. The *Salmonella enterica* subspecies I specific centisome 7 genomic island encodes novel protein families present in bacteria living in close contact with eukaryotic cells. *Res. Microbiol.* **153:** 537–545.

Friedrich, B., Hogrefe, C., and Schlegel, H.G. 1981. Naturally ocurring genetic transfer of hydrogen-oxidizing ability between strains of *Alcanigenes eutrophus*. *J. Bacteriol.* **147:** 198–205.

Genin, S. and Boucher, C. 2004. Lessons learned from the genome analysis of *Ralstonia solanacearum*. *Annu. Rev. Phytopathol.* **42:** 107–134.

Giraud, E., Moulin, L., Vallenet, D., Barbe, V., Cytryn, E., Avarre, J.C., Jaubert, M., Simon, D., Cartieaux, F., Prin, Y., et al. 2007. Legumes symbioses: Absence of Nod genes in photosynthetic bradyrhizobia. *Science* **316:** 1307–1312.

Glendinning, K.J., Parsons, Y.N., Duangsonk, K., Hales, A., Humphreys, D., Hart, C.A., and Winstanley, C. 2004. Sequence divergence in type III secretion gene clusters of the *Burkholderia cepacia* complex. *FEMS Microbiol. Lett.* **235:** 229–235.

Gonzalez, V., Santamaria, R.I., Bustos, P., Hernandez-Gonzalez, I., Medrano-Soto, A., Moreno-Hagelsieb, G., Janga, S.C., Ramirez, M.A., Jimenez-Jacinto, V., Collado-Vides, J., et al. 2006. The partitioned *Rhizobium etli* genome: Genetic and metabolic redundancy in seven interacting replicons. *Proc. Natl. Acad. Sci.* **103:** 3834–3839.

Hueck, C.J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol. Mol. Biol. Rev.* **62:** 379–433.

Jones, K.M., Kobayashi, H., Davies, B.W., Taga, M.E., and Walker, G.C. 2007. How rhizobial symbionts invade plants: The *Sinorhizobium–Medicago* model. *Nat. Rev. Microbiol.* **5:** 619–633.

Kobayashi, M., Suzuki, T., Fujita, T., Masuda, M., and Shimizu, S. 1995. Occurrence of enzymes involved in biosynthesis of indole-3-acetic-acid from indole-3-acetonitrile in plant-associated bacteria, *Agrobacterium* and *Rhizobium*. *Proc. Natl. Acad. Sci.* **92:** 714–718.

Koski, L.B. and Golding, G.B. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52:** 540–542.

Krol, E. and Becker, A. 2004. Global transcriptional analysis of the phosphate starvation response in *Sinorhizobium meliloti* strains 1021 and 2011. *Mol. Genet. Genomics* **272:** 1–17.

Lorquin, J., Lortet, G., Ferro, M., Mear, N., Prome, J.C., and Boivin, C. 1997. *Sinorhizobium teranga* bv. acaciae ORS1073 and *Rhizobium* sp. strain ORS1001, two distantly related Acacia-nodulating strains, produce similar Nod factors that are O carbamoylated, N methylated, and mainly sulfated. *J. Bacteriol.* **179:** 3079–3083.

Mao, C.H., Qiu, J., Wang, C.X., Charles, T.C., and Sobral, B.W.S. 2005. NodMutDB: A database for genes and mutants involved in symbiosis. *Bioinformatics* **21:** 2927–2929.

Marie, C., Broughton, W.J., and Deakin, W.J. 2001. Rhizobium type III secretion systems: Legume charmers or alarmers? *Curr. Opin. Plant Biol.* **4:** 336–342.

Martin, M.J., Herrero, J., Mateos, A., and Dopazo, J. 2003. Comparing bacterial genomes through conservation profiles. *Genome Res.* **13:** 991–998.

Monchy, S., Benotmane, M.A., Janssen, P., Vallaeys, T., Taghavi, S., van der Lelie, D., and Mergeay, M. 2007. Plasmids pMOL28 and pMOL30 of *Cupriavidus metallidurans* are specialized in the maximal viable response to heavy metals. *J. Bacteriol.* **189:** 7417–7425.

Mougous, J.D., Cuff, M.E., Raunser, S., Shen, A., Zhou, M., Gifford, C.A., Goodman, A.L., Joachimiak, G., Ordonez, C.L., Lory, S., et al. 2006. A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* **312:** 1526–1530.

Moulin, L., Munive, A., Dreyfus, B., and Boivin-Masson, C. 2001. Nodulation of legumes by members of the beta-subclass of Proteobacteria. *Nature* **411:** 948–950.

Nakayama, K., Kanaya, S., Ohnishi, M., Terawaki, Y., and Hayashi, T. 1999. The complete nucleotide sequence of phi CTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa*: Implications for phage evolution and horizontal gene transfer via bacteriophages. *Mol. Microbiol.* **31:** 399–419.

Parker, G.F., Higgins, T.P., Hawkes, T., and Robson, R.L. 1999. *Rhizobium* (*Sinorhizobium*) *meliloti phn* genes: Characterization and identification of their protein products. *J. Bacteriol.* **181:** 389–395.

Parsons, D.A. and Heffron, F. 2005. *sciS*, an *icmF* homolog in *Salmonella enterica* serovar typhimurium, limits intracellular replication and decreases virulence. *Infect. Immun.* **73:** 4338–4345.

Pohlmann, A., Fricke, W.F., Reinecke, F., Kusian, B., Liesegang, H., Cramm, R., Eitinger, T., Ewering, C., Potter, M., Schwartz, E., et al. 2006. Genome sequence of the bioplastic-producing "Knallgas" bacterium *Ralstonia eutropha* H16. *Nat. Biotechnol.* **24:** 1257–1262.

Poinsot, V., Belanger, E., Laberge, S., Yang, G.P., Antoun, H., Cloutier, J., Treilhou, M., Dénarié, J., Prome, J.C., and Debellé, F. 2001. Unusual methyl-branched alpha,beta-unsaturated acyl chain substitutions in the Nod factors of an arctic rhizobium, *Mesorhizobium* sp. strain N33

(*Oxytropis arctobia*). *J. Bacteriol.* **183:** 3721–3728.
Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. 2006. Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc. Natl. Acad. Sci.* **103:** 1528–1533.
Roche, P., Debellé, F., Maillet, F., Lerouge, P., Faucher, C., Truchet, G., Dénarié, J., and Prome, J.C. 1991. Molecular basis of symbiotic host specificity in *Rhizobium meliloti*: *nodH* and *nodPQ* genes encode the sulfatation of lipo-oligosaccharide signals. *Cell* **67:** 1131–1143.
Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J.C., Cattolico, L., et al. 2002. Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415:** 497–502.
Sato, Y., Nishihara, H., Yoshida, M., Watanabe, M., Rondal, J.D., Concepcion, R.N., and Ohta, H. 2006. *Cupriavidus pinatubonensis* sp. nov and *Cupriavidus laharis* sp. nov., novel hydrogen-oxidizing, facultatively chemolithotrophic bacteria isolated from volcanic mudflow deposits from Mt. Pinatubo in the Philippines. *Int. J. Syst. Evol. Microbiol.* **56:** 973–978.
Schell, M.A. 2000. Control of virulence and pathogenicity genes of *Ralstonia solanacearum* by an elaborate sensory network. *Annu. Rev. Phytopathol.* **38:** 263–292.
Schwartz, E., Henne, A., Cramm, R., Eitinger, T., Friedrich, B., and Gottschalk, G. 2003. Complete nucleotide sequence of pHG1: A R*alstonia eutropha* H16 megaplasmid encoding key enzymes of H-2-based lithoautotrophy and anaerobiosis. *J. Mol. Biol.* **332:** 369–383.
Stinear, T.P., Seemann, T., Pidot, S., Frigui, W., Reysset, G., Garnier, T., Meurice, G., Simon, D., Bouchier, C., Ma, L., et al. 2007. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Res.* **17:** 192–200.
Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Medigue, C. 2006. MaGe: A microbial genome annotation system supported by synteny results. *Nucleic Acids Res.* **34:** 53–65.
Vandamme, P. and Coenye, T. 2004. Taxonomy of the genus *Cupriavidus*: A tale of lost and found. *Int. J. Syst. Evol. Microbiol.* **54:** 2285–2289.
Viprey, V., Rosenthal, A., Broughton, W.J., and Perret, X. 2000. Genetic snapshots of the *Rhizobium* species NGR234 genome. *Genome Biol.* **1:** research0014.1–14.17. doi: 10.1186/gb-2000-1-6-research0014.
Wilderman, P.J., Vasil, A.I., Johnson, Z., and Vasil, M.L. 2001. Genetic and biochemical analyses of a eukaryotic-like phospholipase D of *Pseudomonas aeruginosa* suggest horizontal acquisition and a role for persistence in a chronic pulmonary infection model. *Mol. Microbiol.* **39:** 291–303.
Young, J.P., Johnston, A.W., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Wexler, M., Curson, A.R., Todd, J.D., and Poole, P.S. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **7:** R34. doi: 10.1186/gb-2006-7-4-r34.
Zang, N., Tang, D.J., Wei, M.L., He, Y.Q., Chen, B.S., Feng, J.X., Xu, J., Gan, Y.Q., Jiang, B.L., and Tang, J.L. 2007. Requirement of a *mip*-like gene for virulence in the phytopathogenic bacterium *Xanthomonas campestris* pv. campestris. *Mol. Plant Microbe Interact.* **20:** 21–30.

*Received January 28, 2008; accepted in revised form May 19, 2008.*

# Genome sequence of the β-rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia

Claire Amadou, Géraldine Pascal, Sophie Mangenot, et al.

*Genome Res.* 2008 18: 1472-1483 originally published online May 19, 2008
Access the most recent version at doi:10.1101/gr.076448.108