



**HAL**  
open science

## Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021

D Capela, F Barloy-Hubler, J Gouzy, G Bothe, F Ampe, J Batut, P Boistard,  
A Becker, M Boutry, E Cadieu, et al.

► **To cite this version:**

D Capela, F Barloy-Hubler, J Gouzy, G Bothe, F Ampe, et al.. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, 98 (17), pp.9877-9882. 10.1073/pnas.161294398 . hal-01213727

**HAL Id: hal-01213727**

**<https://hal.science/hal-01213727>**

Submitted on 9 Oct 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021

Delphine Capela<sup>†††</sup>, Frédérique Barloy-Hubler<sup>\*\*</sup>, Jérôme Gouzy<sup>\*†</sup>, Gordana Bothe<sup>§</sup>, Frédéric Ampe<sup>†</sup>, Jacques Batut<sup>†</sup>, Pierre Boistard<sup>†</sup>, Anke Becker<sup>¶</sup>, Marc Boutry<sup>¶</sup>, Edouard Cadieu<sup>‡</sup>, Stéphane Dréano<sup>‡</sup>, Stéphanie Gloux<sup>‡</sup>, Thérèse Godrie<sup>\*\*</sup>, André Goffeau<sup>¶</sup>, Daniel Kahn<sup>†</sup>, Ernö Kiss<sup>†</sup>, Valérie Lelaure<sup>‡</sup>, David Masuy<sup>¶</sup>, Thomas Pohl<sup>§</sup>, Daniel Portetelle<sup>\*\*</sup>, Alfred Pühler<sup>¶</sup>, Bénédicte Purnelle<sup>¶</sup>, Ulf Ramsperger<sup>§</sup>, Clotilde Renard<sup>†</sup>, Patricia Thébault<sup>†</sup>, Micheline Vandenberg<sup>\*\*</sup>, Stefan Weidner<sup>¶</sup>, and Francis Galibert<sup>†††</sup>

<sup>†</sup>Laboratoire de Biologie Moléculaire des Relations Plantes–Microorganismes, Unité Mixte de Recherche (UMR) 215 Centre National de la Recherche Scientifique (CNRS), Institut National de la Recherche Agronomique, Chemin de Borde Rouge, BP27, F-31326 Castanet Tolosan Cedex, France;

<sup>‡</sup>Laboratoire de Génétique et Développement UMR 6061 CNRS, Faculté de Médecine, 2 Avenue du Pr. Léon Bernard, F-35043 Rennes Cedex, France; <sup>§</sup>GATC GmbH, Fritz-Arnold-Strasse 23, D-78467 Konstanz, Germany; <sup>¶</sup>Universität Bielefeld, Biologie IV (Genetik), Universitätsstrasse 25, D-33615 Bielefeld, Germany; <sup>¶</sup>Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud 2, Bte. 20, B-1348 Louvain-la-Neuve, Belgium; and <sup>\*\*</sup>Unité de Microbiologie, Faculté des Sciences Agronomiques de Gembloux, Avenue Maréchal Juin 6, B-5030 Gembloux, Belgium

Communicated by Sharon R. Long, Stanford University, Stanford, CA, June 12, 2001 (received for review March 12, 2001)

*Sinorhizobium meliloti* is an  $\alpha$ -proteobacterium that forms agronomically important  $N_2$ -fixing root nodules in legumes. We report here the complete sequence of the largest constituent of its genome, a 62.7% GC-rich 3,654,135-bp circular chromosome. Annotation allowed assignment of a function to 59% of the 3,341 predicted protein-coding ORFs, the rest exhibiting partial, weak, or no similarity with any known sequence. Unexpectedly, the level of reiteration within this replicon is low, with only two genes duplicated with more than 90% nucleotide sequence identity, transposon elements accounting for 2.2% of the sequence, and a few hundred short repeated palindromic motifs (RIME1, RIME2, and C) widespread over the chromosome. Three regions with a significantly lower GC content are most likely of external origin. Detailed annotation revealed that this replicon contains all housekeeping genes except two essential genes that are located on pSymB. Amino acid/peptide transport and degradation and sugar metabolism appear as two major features of the *S. meliloti* chromosome. The presence in this replicon of a large number of nucleotide cyclases with a peculiar structure, as well as of genes homologous to virulence determinants of animal and plant pathogens, opens perspectives in the study of this bacterium both as a free-living soil microorganism and as a plant symbiont.

Symbiotic bacteria, together with phytopathogenic microbes, are plant-interacting microorganisms of major agronomic importance. Among these, rhizobia are root-nodule-forming nitrogen-fixing legume symbionts. These microorganisms, in addition to providing an ideal model for studying plant–microbe interactions, play an important role in global nutrient cycling, because symbiotic fixation of nitrogen by rhizobia accounts for over 50% of the nitrogen fixed annually by living organisms (80 million tons). Rhizobia, by diminishing input costs and positively influencing plant crop growth and health, also contribute to the development of a sustainable agriculture.

One of the most studied and best characterized rhizobia is *Sinorhizobium meliloti* (strain 1021), the symbiont of alfalfa (*Medicago sativa*), and its close relative *Medicago truncatula*. As *M. truncatula* and *S. meliloti* have been chosen by many international groups as model organisms, the association between these two organisms is emerging as a worldwide leading model system for symbiosis and nitrogen fixation studies along with *Lotus japonicus* and *Mesorhizobium loti*. The genome sequence of the latter was recently published (1).

*S. meliloti* strain 1021 is a free-living Gram-negative soil bacterium whose genome consists of three large replicons: a chromosome and two megaplasmids (2, 3). Meade and Signer reported a circular genetic map of the *S. meliloti* chromosome (4) in 1977, which was recently updated by the inclusion of 447 more

markers (5). The three replicons have been entirely sequenced by an international consortium (6).

We describe here the general characteristics and main biological functions encoded by the 3.6-Mb chromosome.

## Chromosome Sequence and Annotation

The entire nucleotide sequence of the *S. meliloti* strain 1021 chromosome has been determined by a European consortium composed of six laboratories. The project involved shotgun sequencing of 50 ordered recombinant bacterial artificial chromosome (BAC) clones that cover the whole replicon (5). This strategy was preferred to a whole-chromosome shotgun sequencing strategy, mainly to avoid contamination by megaplasmid DNA and to limit potential problems arising from repetitive sequences. Data were processed by using the PHRED/PHRAP/CONSED package (7). Finishing was done after careful inspection and resequencing by using specific primers to resolve ambiguities, achieve sufficient redundancy (>4-fold), and ensure complete double-strand sequence. Up to 3.5% (130 kb) was sequenced twice independently as overlaps between BAC clones sequenced by different teams of the consortium, and no sequence discrepancies were found. In addition, potential frameshifts were searched for by using NCBI-BLASTX comparison and the FRAMED program (8). Our sequence quality estimates suggest less than 1 error per 100,000 nucleotides; further details on methods are described in detail on our web site (<http://sequence.toulouse.inra.fr/meliloti.html>).

Sequence analysis and annotation were managed with the IANT (integrated Annotation Tool) web-based semiautomated annotation environment (9). Protein-coding ORFs were predicted by using the FRAMED program (8) trained on known *S. meliloti* genes and the NCBI-BLASTX program. tRNA genes were detected with TRNASCAN-SE (10). WU-BLASTN comparisons were also used for the detection of *rrn* operons, insertion sequences, and other DNA repeats (RIME, motifs. . .). Proteins were classified according to the Riley rules (11), modified to account for nodulation and nitrogen-fixation genes.

Abbreviations: ABC, ATP-binding cassette; LPS, lipopolysaccharide; CPS, capsular polysaccharide.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AL591688).

\*D.C., F.B.-H., and J.G. contributed equally to this work.

††To whom reprint requests should be addressed. E-mail: francis.galibert@univ-rennes1.fr.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

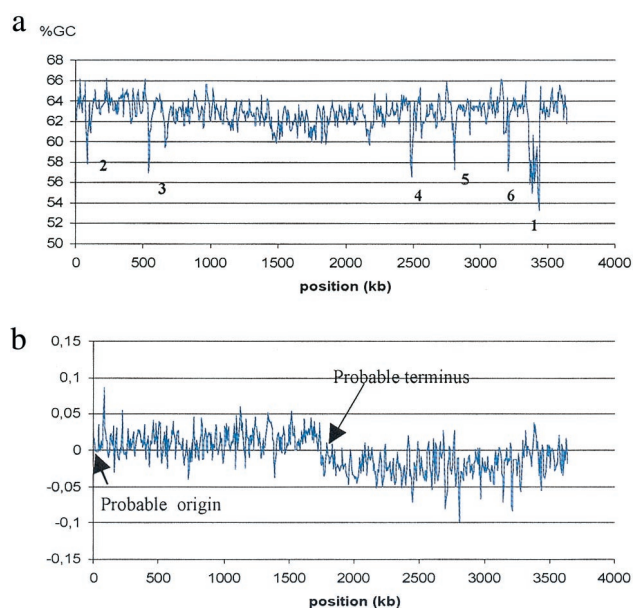
**Table 1. General features of the *S. meliloti* strain 1021 chromosome**

Length, bp	3,654,135
G + C ratio	62.7%
Protein and RNA coding regions	86.4%
tRNAs	51
tmRNA	1
Ribosomal RNA operons	3
Protein-coding genes	3341
Average length of protein-coding genes, bp	938
Start codons	
ATG	86%
GTG	7.8%
TTG	6.2%
Stop codons	
TAA	16.7%
TAG	17.3%
TGA	66%
Genes with functional assignment	59%
Orphan genes (% of total protein-coding genes)	5%
Regulatory genes (% of total protein-coding genes)	7.2%
Insertion sequence and phage-related sequences (% of replicon size)	2.2%
RIME elements	185
Palindromes A, B, C	253

The complete annotated genetic map, search tools (SRS, BLAST), annotation procedure, and classification are available on our web site (<http://sequence.toulouse.inra.fr/meliloti.html>).

### General Features

General features of the chromosome sequence are shown in Table 1. The average GC content is 62.7%, as in the *M. loti* chromosome (1). However, the study of GC percentage variations over the replicon revealed six large regions with lower GC content (Fig. 1a). Three of these regions (regions 2, 5, and 6) correspond to the *rrn* operons (54.4% GC). The largest and most intriguing region



**Fig. 1.** (a) Distribution of GC content variation (sliding windows of 10,000 bp and a step of 5,000 bp). Six regions with a lower GC percentage are discernible. (b) GC skew  $[(G - C)/(G + C)]$  calculated with a sliding window of 10,000 bp and a step of 5,000 bp.

(region 1), 80 kb in size, contains a number of genes from mobile elements, i.e., two integrase genes and 10 transposon-associated genes. In addition, we identified nine gene fragments in this region, six of which are remnant transposase fragments, indicating that this DNA region has undergone many rearrangements. Region 1 also contains a second copy of *proB* (49% amino acids identical to *proBI*), which encodes a glutamate-5-kinase involved in the proline biosynthesis pathway. The presence of two *proB* genes has not been encountered so far in any other bacterial genome. This observation suggests that chromosomal region 1 results from horizontal gene transfer. Another interesting gene in this region is *fixT3*, a third copy of the *fixT* gene (93% nucleotide identity), which likely results from DNA transfer from pSymA. Region 3, 20 kb long, is flanked by a tRNA<sup>Thr</sup> on one side and a phage integrase on the other side, which suggests that this DNA fragment could have been transduced by a phage. Ten of the twelve genes predicted in this region encode hypothetical proteins of unknown function. None of these shows similarity to phage protein sequences. The last region (region 4), 15 kb in size, is limited on one side by a tRNA<sup>Met</sup> and contains 17 short hypothetical genes.

The replication origin of the *S. meliloti* chromosome has been tentatively placed from the following observations: (i) one of the two major inversions revealed by GC skew analysis lies at this position (Fig. 1b); (ii) all three *rrn* operons are in the neighborhood of the putative origin, on the leading strand; (iii) as is often observed in bacterial genomes, genes such as *parAB*, involved in chromosome partition, and *gidAB*, the glucose-inhibited division genes, are located in the vicinity of the origin. Very recently, Brassinga *et al.* (12) identified a conserved gene cluster around the replication origins of the  $\alpha$ -proteobacteria *Caulobacter crescentus* and *Rickettsia prowazekii*. The same genetic organization, with the exception of the *hemH* gene, was found around the putative origin of replication of the *S. meliloti* chromosome, thus validating our prediction and placing the *S. meliloti* chromosome origin between *smc02794/hemE* and *smc02793*.

The plasticity of the *S. meliloti* chromosome is less extensive than expected. Indeed, transposon-related functions account for only  $\approx 2.2\%$  of the total chromosome sequence, with 0.5% located in region 2 described above (Fig. 1a). In addition, except for the three *rrn* operons, only two recently duplicated genes were found on this replicon with more than 90% nucleotide sequence identity *tuf* and *purU*. Other chromosomal genes have copies on the pSym plasmids (1). The intergenic elements (RIME1, RIME2), composed of two large palindromic sequences and motifs C described by Osteras *et al.* (13, 14), are widespread over the chromosome. Both half-RIME elements and motifs C have been found organized in diverse unpublished combinations. With the exception of one motif C in the *uvrD* gene, these repeated elements have been exclusively detected in intergenic regions. Interestingly, it appears that motifs C are preferentially located downstream of convergent genes, their orientation being correlated with gene transcription direction (see additional data on our web site, <http://sequence.toulouse.inra.fr/meliloti.html>). These elements could play a role in gene transcription termination, as suggested before (15).

Coding regions, including protein- and RNA-coding genes, represent 86.4% of the total chromosome sequence. An interesting feature of gene organization is the frequent changes in polarity. Although 51.5% of genes are transcribed on the leading strand, the coding strand switches every three to four genes. Data on the relative orientation of neighboring chromosomal genes are given in Table 2. These data show that the distribution is strongly biased, as in most cases the head-to-head configuration is observed only when the distance between the two genes exceeds 150 bp. This bias likely reflects the minimal length for two head-to-head promoters.

A total of 3,341 protein-encoding genes have been predicted, with a mean length of 938 bp. The longest identified chromo-

**Table 2. Number of gene pairs in tail-to-head, tail-to-tail, and head-to-head configuration**

Intergenic distance	Tail-head ⇒ ⇒	Tail-tail ⇒ ⇐	Head-head ⇐ ⇒
0–50 bp	985	155	16
51–100 bp	323	109	30
101–150 bp	302	64	87
≥151 bp	683	195	390

somal gene is 8,496 nt long and encodes the cyclic-(1, 2)-glucan biosynthesis NdvB protein (16). The proportion of start and stop codons used in the *S. meliloti* chromosome is indicated in Table 1. The TGA stop codon is over-represented (66%), as in other genomes with a high GC content (17).

Putative functions were assigned, on the basis of homology, to 59% of the chromosomal genes, a figure slightly higher than with recently sequenced bacterial genomes such as *Pseudomonas aeruginosa* (54.2%) (18), *M. loti* (54%) (1), *Vibrio cholerae* (53.5%) (19), or *Xylella fastidiosa* (47%) (20). The 41% of proteins with unknown function are comprised of three categories: 5% present no similarity with any protein in databases, 10.3% show only partial or weak similarity with database proteins, and 25.7% are conserved hypothetical proteins. Genes encoding similar biological functions or pathways are evenly distributed across the chromosome, except for genes encoding ribosomal proteins and chemotaxis proteins that are organized in large clusters.

### DNA Replication, Repair, and Recombination

*S. meliloti* contains at least 57 genes involved in DNA metabolism and replication, including the DNA polymerase I gene *polA* and genes encoding  $\alpha$ ,  $\beta$ ,  $\delta'$ ,  $\epsilon$ ,  $\gamma$ , and  $\tau$  subunits of DNA polymerase III. The number of identified genes involved in DNA replication is smaller than in *Escherichia coli*, as genes encoding DNA polymerase II (*polB*) and other DNA polymerase III accessory subunits ( $\delta$ ,  $\chi$ ,  $\Psi$ , and  $\theta$ ) have not been identified in *S. meliloti*. These genes are also missing in several other bacterial genomes like *Bacillus subtilis* and *Mycobacterium tuberculosis*, suggesting that these might not be essential for DNA replication, or alternatively that other genes, yet to be identified, substitute for them. The DnaG primase, the PriA protein required for primosome assembly, and the DnaA replication initiator protein are also encoded by the chromosome. However, other proteins involved in primosome assembly in *E. coli*, such as DnaT, DnaC, PriB, and PriC, are missing in *S. meliloti*.

In addition to replication-associated proteins, we identified 30 chromosomal genes involved in DNA mismatch repair (*mut-LSY*), UV-damaged DNA repair (*uvrA*, *uvrB*, *uvrC*, and *uvrD*) and homologous recombination (*recAFGNRQJ*). As was found in *R. prowazekii*, *Mycoplasma genitalium*, and *Helicobacter pylori*, the RecBCD complex is absent from *S. meliloti*. Interestingly, a restriction-modification system (*hsdMSR*) flanked by the two insertion sequences, *ISRM11* and *ISRM1*, was also identified. This composite transposon-like structure suggests that this DNA region could have been acquired by horizontal transfer.

### Cell Division

The cell division and cell cycle control processes in *S. meliloti* are poorly understood. These mechanisms are likely to be particularly important in symbiosis, because after a phase of active cell divisions during infection, cell division and DNA replication stop in differentiated bacteroids (21, 22). Moreover, the mechanisms by which coordination of cell division with DNA replication is achieved, as well as the partition of the three replicons, remain to be elucidated.

Six of the nine genes required for the cytokinesis process in *E. coli* (*ftsA*, *ftsI*, *ftsK*, *ftsQ*, *ftsW*, and two *ftsZ* genes) have been found on the *S. meliloti* chromosome. Some of these have previously been identified by Margolin *et al.* (23, 24), whereas *ftsL*, *ftsN*, and *zipA* are apparently missing. We also identified two additional genes (*smc02311* and *smc02792*) coding for proteins sharing 54 and 50% amino acid sequence identity, respectively, with the *B. subtilis* Maf protein, involved in septum formation. Chromosome partitioning in *S. meliloti* may require *smc* and *parAB* as in *C. crescentus*. Another interesting orthologous gene of *C. crescentus* identified in *S. meliloti* is *ctrA*, a member of the two-component signal transduction family involved in the control of a number of cell cycle-regulated genes, including the cell division genes *ftsQA* and *ftsZ* and DNA methyltransferase encoding gene *ccrM* (25). CtrA shares similar features to *C. crescentus* and, by analogy, may play a role in cell division (26). Interestingly, none of the regulators (*cckA*, *divF*, and *divJ*) that control *ctrA* expression in *C. crescentus* (27), except *divK*, was retrieved in the entire *S. meliloti* genome.

### Transcription and Translation

All of the genes, but one tRNA<sup>Arg</sup>, required for transcription and translation have been detected on the chromosome. The three ribosomal *rnm* operons, as well as all genes coding for ribosomal protein, are located on this replicon. Fifty-five genes corresponding to all of the *E. coli* ribosomal proteins except the S22 protein of the 30S subunit have been found. An original finding is the presence of two copies of *rpsU*, coding for the S21 ribosomal-protein encoding gene, which has never been described so far in bacterial genomes. The respective role of these two subunits is still unknown. Fifty-one tRNA genes, evenly distributed on the chromosome, have been detected. These tRNAs correspond to 43 different tRNA acceptors and may recognize, by virtue of the base wobble, all possible codons with the exception of CGG used for arginine. The missing essential tRNA is encoded by the pSymB megaplasmid (28). In addition, neither glutamyl- nor asparaginyl-tRNA synthetase encoding genes (*glnS* and *asnS*) has been found in *S. meliloti*. As previously suggested in *Deinococcus radiodurans* and *B. subtilis* (29, 30), both enzymatic activities might be replaced by a transamidation process of misacylated Glu-tRNA<sup>Gln</sup> and Asp-tRNA<sup>Asn</sup>, which probably involve the chromosomal *gatABC* gene products.

### Biosynthesis

The complete pathways for amino acid biosynthesis have been found, with the noticeable exception of asparagine synthase, for which the two *asn* genes are on pSymB. Interestingly, *S. meliloti* exhibits two different pathways for methionine biosynthesis, the classical *metABC* as in *E. coli* and *metZ* essential for symbiosis in *Rhizobium etli* (31). All essential genes for *de novo* synthesis of purines and pyrimidines as well as salvage pathways for most nucleotides have been found. Biosynthesis genes are not organized in operons, as is the case in *E. coli*, and no global regulator such as PurR or PyrR has been identified. The chromosome of *S. meliloti* also bears most of the essential genes for cofactor and vitamin biosynthesis. However, consistent with the biotin auxotrophy of strain 1021, biotin synthetase (EC 2.8.1.6) and dethiobiotin synthetase (EC 6.3.3.3), although found in *M. loti*, were not detected in the *S. meliloti* 1021 genome. The chromosome also contains genes for fatty acid, lipid, phospholipid, peptidoglycan, lipid A, and cyclic  $\beta$ -glucan synthesis.

### Central Metabolism and Energy Formation

All of the enzymes involved in glycolysis and gluconeogenesis are encoded by chromosomal genes, with the exception of the gluconeogenic fructose-1, 6-biphosphatase (*cbbF*), found on pSymB. For glycolysis, *S. meliloti* lacks the classical ATP-dependent phosphofructokinase (*pfk*) but possesses a complete Entner–Doudoroff pathway that makes it the main route for

glucose utilization (32). In addition, the chromosome bears a pyrophosphate-dependent PFK (SMc01852), an enzyme not found before in proteobacteria, the role of which remains to be investigated. Together with a complete Krebs cycle and glyoxylate shunt, a large number of enzymes involved in pyruvate and acetylCoA metabolism have been identified (PykA, PpdK, LldD, Dld, Pyc, PckA, PdhA, LpdA, Mdh, Tme, Mde, and AcsA). However, oxaloacetate decarboxylase and classical anaerobic enzymes are notably missing. Interestingly, a complete methylglyoxal bypass has been identified. This rather rare pathway enables the conversion of dihydroxyacetone-phosphate to pyruvate in a low-phosphate environment (33). The chromosome bears the necessary genes for ammonium transport (*amtB*) and assimilation, including two of the three glutamine synthetases, *glnA*, *glnT* (the third copy, *glnII*, is located on pSymB), and the two subunits of glutamate synthase (*gltBD*). The corresponding regulatory system, including P<sub>II</sub>, the P<sub>II</sub> uridylyl-transferase (GlnD), and the GS-adenylyl-transferase (GlnE), have also been found. In addition, and next to the NtrBC system regulating nitrogen assimilation in free-living bacteria, we have identified homologs of NrfA and of the two-component system NtrYX that both regulate *nifA* expression in *Azorhizobium caulinodans* (34, 35). The only potential nitrogen fixation gene found on the chromosome encodes a protein with 49% amino acid identity to the NifS cysteine desulfurase required for assembly of iron-sulfur clusters (36). Three lines of evidence support the contention that this gene is indeed a *nifS* gene: (i) a *nifS* ortholog exists in *Rhizobium* sp. NGR234; (ii) no better ortholog candidate was found on the megaplasmids; and (iii) the sequence corresponding to this gene is specifically expressed in the nodule (22).

*S. meliloti* generates energy through aerobic respiration coupled with proton translocation. A new cytochrome *c* oxidase complex (*coxMNOP*) and terminal quinol oxidases (*qtxA qtxB*), adding two new branches to the *S. meliloti* respiratory chain, have been found. Proton translocation can be performed by the ATPase complex (*atpHAGDC*, *atpIBEF2F1*) and by an H<sup>+</sup>-pyrophosphatase (RrpP), an enzyme rarely found in prokaryotes that uses P<sub>Pi</sub> instead of ATP (37).

### Transport and Catabolism

As many as 10.8% of the chromosomal proteins are involved in transport, 33% of which belong to the ATP-binding cassette (ABC) and 18% to Major Facilitator superfamilies. A total of 78 ABC transporters have been detected, of which 50 are predicted to be import and 20 export systems, based on the presence of a periplasmic substrate-binding protein. Among these, the proportion of amino acid, amine, and peptide transport systems is strikingly high in the *S. meliloti* chromosome (22% of the transport proteins). In particular, 10 proteins belong to the Resistance to homoserine/threonine (Rht) family, whereas these proteins are rare in genomes (only ≈25 in databases). Genes have been found for the catabolism of a large number of amino acids, among which *argII* (arginase) and *ald* (alanine dehydrogenase) are noteworthy. Moreover, and in addition to the known *degP* and *clpP*, 33 genes involved in protease/peptidase activity have been identified, thus highlighting the importance of protein and amino acid degradation as a main feature of this bacterium. The presence of 15 sugar kinases on the *S. meliloti* chromosome is in good agreement with the known ability of this bacterium to degrade a large number of simple sugars (38) and with the absence of an effective phosphotransferase system (PTS). The degradation of di- and oligosaccharides can be predicted from the presence of several new glycosidases (SMc01846, SMc02071, SMc03064, SMc03160, SMc04247, and SMc04255), in addition to the *aglEFGAK* cluster and three transglycosylases. Other carbon compounds that can be degraded by means of chromosomal genes include organic and fatty acids (21 genes including acyl-CoA synthases, enoylCoA hydratases, ligases, and transferases) and 4-hydroxyphenylpyruvate. Interestingly, eight oxygenases have

been detected, three of which, most probably cytochrome P450 oxygenases, could be involved in nonspecific biodegradation of recalcitrant compounds. Finally, the three subunits of a CO dehydrogenase have been identified on the *S. meliloti* chromosome (SMc03101–3103).

### Adaptation to Atypical Conditions

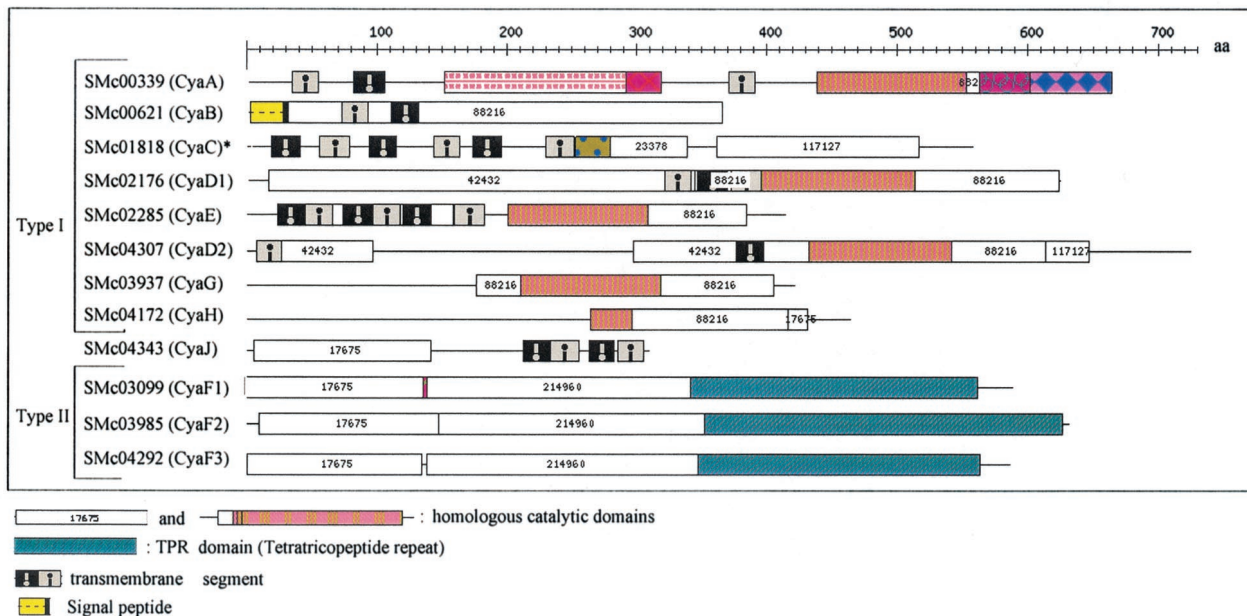
*S. meliloti* seems to be well equipped to face a large variety of stress conditions, including osmotic shocks—in particular by means of the glycine betaine biosynthesis genes *betICBA* (39)—and temperature variations. Temperature upshift induces the synthesis of heat-shock proteins (HSPs). Major bacterial HSPs are molecular chaperones, such as DnaJ, DnaK, GrpE, GroESL, and proteases. They are particularly numerous in *S. meliloti*. In addition to the known complete chromosomal copy of *groESL* (40), we have identified a second chromosomal copy of the *groEL* gene (79% amino acid identity to GroEL1). Moreover, besides the *E. coli* *grpE* and *dnaK-dnaJ* orthologous genes, we have found in *S. meliloti* another *dnaK* homologue (*smc02376*) and four additional genes (*smc00003*, *smc00699*, *smc01853*, and *smc04233*) containing a specific DnaJ domain (41–56% amino acid identity to the *E. coli* DnaJ domain). As shown for DnaJ in *Rhizobium leguminosarum* (41) and GroELc in *S. meliloti* (42), these new chaperones might play a crucial role in symbiosis. The *S. meliloti* chromosome also encodes: (i) three putative small HSPs that belong to the Hsp20 (SMc01106, SMc04040) and Hsp15 (SMc03876) families; (ii) a series of ATP-dependent heat-shock proteases including ClpA and ClpB; (iii) three copies of ClpP and ClpX that can perform chaperone functions in the absence of ClpP; and (iv) the two members of the Hsp100 ATPase family, HslV and HslU. The *S. meliloti* chromosome also carries five cold-shock homologous genes.

### Oxidative Stress Protection

*S. meliloti* grows optimally in aerobic (or microaerobic) conditions and possesses protective enzymes against oxidative stresses. Oxygen protection might be essential for efficient infection if rhizobia, like many pathogens, induce an oxidative burst on plant cell infection. One of the three known catalase genes (43) of *S. meliloti*, *katA*, lies on the chromosome next to *oxyR*, whereas the other two are located on the two other replicons. The chromosome contains two superoxide dismutase genes, one belonging to the iron/manganese (*sodB*) and the other to the copper/zinc dismutase (*sodC*) families. Eleven glutathione *S*-transferase (*gst*) and three *rpoE* ( $\sigma$ -factor 24) were identified on the chromosome and might contribute to protection against oxygen or other reactive molecular species. Six additional *gst* and five additional *rpoE* have been detected on the megaplasmids. The same number of *gst* (17 genes) have been described in the sequence of *M. loti*, but in this organism, they are all carried by the chromosome (1).

### Iron Transport and Heme Biosynthesis

Iron and heme are of central importance in symbiosis, because many key enzymes of the nitrogen fixation process are redox proteins. The chromosome carries multiple systems for iron or heme transport: (i) *afuABC*, an ABC transporter likely involved in iron transport; (ii) *sitABCD*, closely homologous to *Salmonella* and *Yersinia* iron transport genes, lying next to *fur*, a regulatory gene that might operate in a siderophore-independent manner (44); (iii) *fbpB* and *smc00784* encoding a permease and a periplasmic solute-binding protein member of an iron transport system; and (iv) at least six TonB-dependent receptor genes, four of which are involved in siderophore-iron transport (*fhuA1*, *fhuA2*, *smc02721*, and *smc02890*), and two in heme transport (*smc02726* and *smc04205*). The latter lies next to the *fecIR* regulatory system (45). As in *R. leguminosarum*, high-affinity iron acquisition may also require the *feuPQ* genes encoding a two-component regulatory system. We have also identified an ortholog of the *Bradyrhizobium japonicum* *irr* gene that coordi-



**Fig. 2.** Structure of the chromosomal nucleotide cyclases of *S. meliloti*. \*, the CyaC catalytic domain is masked by ProDom PD117127 domain (52). CyaJ is an intermediate between types I and II.

nates heme biosynthesis with iron uptake (46). The *hemABCDEFK* genes responsible for heme biosynthesis, with the exception of *hemG*, have been found scattered over the chromosome. Finally, a bacterioferritin gene, *bfr*, possibly involved in iron storage and detoxification, was found next to *ialA*.

### Cell Surface Components

*S. meliloti*, as other Gram-negative bacteria, produces a variety of surface polysaccharides, including exopolysaccharides (EPS), lipopolysaccharides (LPS), and capsular polysaccharides (CPS), in addition to periplasmic  $\beta$ -glucans. These compounds probably contribute to the adaptation of *S. meliloti* to different environmental conditions. The chromosome contributes very significantly to polysaccharide synthesis. Whereas the structural genes for EPS biosynthesis are located on pSymB (28), most of the corresponding regulators, including MucR, ExoR, ChvIG, and PhoB, are encoded by the chromosome. Two (*rkpAGHIJ* and *rkpK-lpsL*) of the three loci for CPS production described in *S. meliloti* strain 41 (47) have been retrieved from the 1021 chromosome, albeit with significant differences. Indeed, the *rkpA* gene from strain 1021 corresponds to a fusion of the *rkpABCDE* and *F* genes from strain 41. Moreover, a part of the third locus, corresponding to *rkpLMNOPQ* and *rkpZ*, is missing from the *S. meliloti* 1021 genome. This finding gives support to the fact that CPS structure is strain-specific in *Sinorhizobium* sp (48).

The chromosome also encodes enzymes needed to synthesize lipid A (LpxABCDK, FabZ, and MsbB) and the core (LpsABCDEL, KdsAB, KdtA) for LPS synthesis. Two new loci involved in polysaccharide synthesis were predicted from the sequence. One of them (SMc04270 to SMc04278) is possibly involved in the biosynthesis of fatty acids, precursors for lipid A. The other locus (SMc01790 to SMc01796) is supposed to be involved in polymerization of sugar precursors for either LPS or CPS. Genes necessary for the biosynthesis, modification, and export of cyclic  $\beta$ -glucans (*ndvAB* and *cgmAB*) are chromosomal.

Sequence inspection suggests that *S. meliloti* may synthesize a pilus of a novel type recently described in *C. crescentus*. The *S. meliloti* pilus is encoded by a set of nine chromosomal genes, in the same order as in *C. crescentus* (49). The *S. meliloti pil/cpa*

cluster is bracketed by two insertion sequence elements, raising the possibility of the *pil* cluster being genetically mobile. Accordingly, most of the *pil/cpa* gene cluster is also present on pSymA. Last, we have identified on the chromosome at least 19 new genes coding for outer membrane proteins. Many of these proteins have been previously characterized in pathogenic bacteria such as *Brucella* (*omp1*, *omp10*, *omp16*, *omp19*, *omp25*), *Coxiella* (*com1*), *Yersinia* (*fyuA*), or in other rhizobia such as *Rhizobium* sp. NGR234 (Y4fJ) and *R. leguminosarum* (*ropB*).

### Virulence-Like Genes

In addition to the cell surface components described above, the sequence has revealed other candidates for interaction with the plant host. Namely, we have identified a number of genes homologous to genes involved in virulence in animal and plant pathogens, including the recently sequenced *X. fastidiosa* (20). The most significant examples include: (i) an ortholog of the *acvB* virulence gene of *Agrobacterium tumefaciens* that is also present in *Xylella*; (ii) *ialA* and *ialB*, which are required for invasion of human erythrocytes by *Bartonella bacilliformis*, the Oroya fever causative agent; (iii) *rnr*, encoding a ribonuclease required for expression of virulence genes in *Shigella flexneri* and possibly *Xylella*; (iv) *typA*, a GTPase that mediates interactions between enteropathogenic *E. coli* and epithelial cells. SMc00973 closely resembles the HlyA hemolysin of *Treponema hyodysenteriae*, whereas SMc00286, SMc04171, and SMc04206 may belong to the RTX-toxin family. We have also identified several pairs of homologs of virulence-associated proteins (Vap) of *Bacteroides nodosus*, the ovine foot-rot agent. The biological function of the VapBC proteins remains elusive. However, a VapC homolog, encoded by the *ntrR* gene, has a dramatic effect on the symbiotic properties of *S. meliloti* (50). The *vapAD* genes, like the homologous *higAB* genes of *E. coli*, encode a poison-antidote system that often ensures the genetic stability of virulence plasmids or pathogenicity islands. The VapAD genes on the *S. meliloti* chromosome might contribute to the genetic stability of the chromosome or, alternatively, might be relics of the integration of ancient plasmids.

### Regulation and Signal Transduction

Regulatory functions have been assigned to 7.2% of total chromosomal genes, i.e., nearly half of the regulatory genes

identified on the whole *S. meliloti* genome (6). Chromosomal regulators include: (i) nine  $\sigma$  factors; (ii) 166 transcription regulators with a large number (31 members) belonging to the LysR family; (iii) 11 complete two-component regulatory systems plus eight additional isolated histidine kinases; and (iv) one LuxIR-type quorum-sensing system. In addition to these common regulator families, we have found on the chromosome 12 of the 26 identified *S. meliloti* nucleotide cyclases, an unusually high number. Domain analysis reveals the presence of two different types of cyclase, both encountered on the three replicons (Fig. 2). Both types contain a homologous catalytic domain that is found in class III adenylate/guanylate cyclases (51). The first type of *S. meliloti* cyclases harbors the catalytic domain in the carboxyl-terminal part of the protein and a variable amino-terminal domain that often harbors transmembrane segments. The diversity of these amino-terminal domains suggests that these cyclases could act in response to specific extracellular signals. The second type of cyclase has been exclusively observed so far in *S. meliloti*. It consists of the catalytic domain located in the amino-terminal part of the protein associated with carboxyl-terminal tetratricopeptide repeats (TPR). TPR domains usually mediate protein-protein interactions. The role of these cyclases and the signal transduction pathways in which they participate are unknown in *S. meliloti*.

## Conclusions

The *S. meliloti* chromosome sequence has revealed that, besides housekeeping genes (i.e., genes involved in nucleic acid and protein metabolism as well as universal biosynthetic pathways), this replicon also carries genetic information for mobility and chemotaxis processes, plant interaction (putative virulence genes), as well as stress responses. Strikingly, despite the availability of more than 37 complete and annotated bacterial and several complete or partial eukaryote genome sequences, functions could not be assigned for up to 41% of the *S. meliloti* chromosomal genes—12% of them (5% of total protein coding ORFs) are unique, as they have no sequence analogy. This subset might well be characteristic of *Sinorhizobium*, because they were not found in other  $\alpha$ -proteobacteria like *M. loti* (1) or *C. crescentus* (27). Functional analysis will determine whether the existence of these genes reflects the phylogenetic position of *S. meliloti* or its adaptation to soil, rhizospheric, or symbiotic life.

We express our gratitude to Dr. Guido Volckaert, who helped resolve a difficult sequencing problem, Dr. Thomas Schiex for the FRAMED program, and IIT Biotech GmbH, Bielefeld. This work has been supported by grants from the European Union (BIO4-CT98-0109), Institut National de la Recherche Agronomique (AIP 98/P00206), and Centre National de la Recherche Scientifique (Genome Research program) to J.B. and F.G.

- Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., et al. (2000) *DNA Res.* **7**, 331–338.
- Honeycutt, R. J., McClelland, M. & Sobral, B. W. (1993) *J. Bacteriol.* **175**, 6945–6952.
- Banfalvi, Z., Kondorosi, E. & Kondorosi, A. (1985) *Plasmid* **13**, 129–138.
- Meade, H. M. & Signer, E. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 2076–2078.
- Capela, D., Barloy-Hubler, F., Gatius, M. T., Gouzy, J. & Galibert, F. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 9357–9362.
- Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., et al. (2001) *Science* **293**, 668–672.
- Nickerson, D. A., Tobe, V. O. & Taylor, S. L. (1997) *Nucleic Acids Res.* **25**, 2745–2751.
- Schiex, T., Thébault, P. & Kahn, D., eds. (2000) in *JOBIM Conference Proceedings* (ENSA & LIRM, Montpellier, France), pp. 321–328.
- Thébault, P., Servant, F., Schiex, T., Kahn, D. & Gouzy, J., eds. (2000) in *JOBIM Conference Proceedings* (ENSA & LIRM, Montpellier, France), pp. 361–365.
- Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25**, 955–964.
- Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1999) *Nucleic Acids Res.* **27**, 55–58.
- Brassinga, A. K., Siam, R. & Marczyński, G. T. (2001) *J. Bacteriol.* **183**, 1824–1829.
- Osteras, M., Stanley, J. & Finan, T. M. (1995) *J. Bacteriol.* **177**, 5485–5494.
- Osteras, M., Boncompagni, E., Vincent, N., Poggi, M. C. & Le Rudulier, D. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 11394–11399.
- Gilson, E., Rousset, J. P., Clement, J. M. & Hofnung, M. (1986) *Ann. Inst. Pasteur Microbiol.* **137B**, 259–270.
- Ielpi, L., Dylan, T., Ditta, G. S., Helinski, D. R. & Stanfield, S. W. (1990) *J. Biol. Chem.* **265**, 2843–2851.
- Rocha, E. P., Danchin, A. & Viari, A. (1999) *Nucleic Acids Res.* **27**, 3567–3576.
- Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrenner, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., et al. (2000) *Nature (London)* **406**, 959–964.
- Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Clayton, R. A., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Umayam, L., et al. (2000) *Nature (London)* **406**, 477–483.
- Simpson, A. J., Reinach, F. C., Arruda, P., Abreu, F. A., Acencio, M., Alvarenga, R., Alves, L. M., Araya, J. E., Baía, G. S., Baptista, C. S., et al. (2000) *Nature (London)* **406**, 151–157.
- Oke, V. & Long, S. R. (1999) *Curr. Opin. Microbiol.* **2**, 641–646.
- Oke, V. & Long, S. R. (1999) *Mol. Microbiol.* **32**, 837–849.
- Margolin, W., Corbo, J. C. & Long, S. R. (1991) *J. Bacteriol.* **173**, 5822–5830.
- Margolin, W. & Long, S. R. (1994) *J. Bacteriol.* **176**, 2033–2043.
- Wright, R., Stephens, C. & Shapiro, L. (1997) *J. Bacteriol.* **179**, 5869–5877.
- Barnett, M. J., Hung, D. Y., Reisenauer, A., Shapiro, L. & Long, S. (2001) *J. Bacteriol.* **183**, 3204–3210.
- Jenal, U. (2000) *FEMS Microbiol. Rev.* **24**, 177–191.
- Finan, T. M., Weidner, S., Wong, K., Buhrmester, J., Chain, P., Vorhölter, F. J., Hernandez-Lucas, I., Becker, A., Cowie, A., Gouzy, J., et al. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 9889–9894. (First Published July 31, 2001; 10.1073/pnas.161294698)
- Curnow, A. W., Hong, K. W., Yuan, R. & Soll, D. (1997) *Nucleic Acids Symp. Ser.* **36**, 2–4.
- Curnow, A. W., Tumbula, D. L., Pelaschier, J. T., Min, B. & Soll, D. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 12838–12843.
- Tate, R., Riccio, A., Caputo, E., Iaccarino, M. & Patriarca, E. J. (1999) *Mol. Plant-Microbe Interact.* **12**, 24–34.
- Finan, T. M., Oresnik, I. & Bottacin, A. (1988) *J. Bacteriol.* **170**, 3396–3403.
- Ferguson, G. P., Totemeyer, S., MacLean, M. J. & Booth, I. R. (1998) *Arch. Microbiol.* **170**, 209–218.
- Pawlowski, K., Klosse, U. & de Bruijn, F. J. (1991) *Mol. Gen. Genet.* **231**, 124–138.
- Kaminski, P. A., Desnoues, N. & Elmerich, C. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 4663–4667.
- Zheng, L., White, R. H., Cash, V. L., Jack, R. F. & Dean, D. R. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 2754–2758.
- Baltscheffsky, M., Schultz, A. & Baltscheffsky, H. (1999) *FEBS Lett.* **457**, 527–533.
- Kahn, M. L., McDermott, T. R. & Udvardi, M. K. (1998) in *The Rhizobiaceae*, eds. Spaink, H. P., Kondorosi, A. & Hooykaas, P. J. J. (Kluwer, Dordrecht, The Netherlands), pp. 461–485.
- Pocard, J. A., Vincent, N., Boncompagni, E., Smith, L. T., Poggi, M. C. & Le Rudulier, D. (1997) *Microbiology* **143**, 1369–1379.
- Rusanganwa, E. & Gupta, R. S. (1993) *Gene* **126**, 67–75.
- Labidi, M., Laberge, S., Vezina, L. P. & Antoun, H. (2000) *Mol. Plant-Microbe Interact.* **13**, 1271–1274.
- Ogawa, J. L. & Long, S. R. (1995) *Genes Dev.* **9**, 714–729.
- Sigaud, S., Becquet, V., Frendo, P., Puppo, A. & Herouart, D. (1999) *J. Bacteriol.* **181**, 2634–2639.
- Zhou, D., Hardt, W. D. & Galan, J. E. (1999) *Infect. Immun.* **67**, 1974–1981.
- Van Hove, B., Staudenmaier, H. & Braun, V. (1990) *J. Bacteriol.* **172**, 6749–6758.
- Hamza, I., Chauhan, S., Hassett, R. & O'Brian, M. R. (1998) *J. Biol. Chem.* **273**, 21669–21674.
- Kereszt, A., Kiss, E., Reuhs, B. L., Carlson, R. W., Kondorosi, A. & Putnok, P. (1998) *J. Bacteriol.* **180**, 5426–5431.
- Reuhs, B. L., Stephens, S. B., Geller, D. P., Kim, J. S., Glenn, J., Przytycki, J. & Ojanen-Reuhs, T. (1999) *Appl. Environ. Microbiol.* **65**, 5186–5191.
- Skerker, J. M. & Shapiro, L. (2000) *EMBO J.* **19**, 3223–3234.
- Oláh, B., Kiss, E., Györgypál, Z., Borzi, J., Cinege, G., Csanádi, G., Batut, J., Kondorosi, A. & Dusha, I. (2001) *Mol. Plant-Microbe Interact.* **14**, 887–894.
- Danchin, A., Pidoux, J., Krin, E., Thompson, C. J. & Ullmann, A. (1993) *FEMS Microbiol. Lett.* **114**, 145–151.
- Corpet, F., Servant, F., Gouzy, J. & Kahn, D. (2000) *Nucleic Acids Res.* **28**, 267–269.