Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid

Melanie J. Barnett[a], Robert F. Fisher[a], Ted Jones[b], Caridad Komp[b], A. Pia Abola[b], Frédérique Barloy-Hubler[c], Leah Bowser[b], Delphine Capela[c,d,e], Francis Galibert[c], Jérôme Gouzy[d], Mani Gurjal[b], Andrea Hong[a], Lucas Huizar[b], Richard W. Hyman[b], Daniel Kahn[d], Michael L. Kahn[f], Sue Kalman[b,g], David H. Keating[a,h], Curtis Palm[b], Melicent C. Peck[a], Raymond Surzycki[b,i], Derek H. Wells[a], Kuo-Chen Yeh[a,h,j], Ronald W. Davis[b], Nancy A. Federspiel[b,k], and Sharon R. Long[a,h,l]

[a]Department of Biological Sciences, and [h]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305; [b]Stanford Center for DNA Sequencing and Technology, 855 California Avenue, Palo Alto, CA 94304; [c]Laboratoire de Génétique et Développement, Faculté de Médecine, 2 Avenue du Pr. Léon Bernard, F-35043 Rennes Cedex, France; [d]Laboratoire de Biologie Moléculaire de Relations Plantes–Microorganisms, Unité Mixte de Recherche, 215 Institut National de la Recherche Agronomique–Centre National de la Recherche Scientifique, F-31326 Castanet Tolosan, France; and [f]Institute of Biological Chemistry, Washington State University, Pullman, WA 99164

Contributed by Sharon R. Long, June 12, 2001

The symbiotic nitrogen-fixing soil bacterium *Sinorhizobium meliloti* contains three replicons: pSymA, pSymB, and the chromosome. We report here the complete 1,354,226-nt sequence of pSymA. In addition to a large fraction of the genes known to be specifically involved in symbiosis, pSymA contains genes likely to be involved in nitrogen and carbon metabolism, transport, stress, and resistance responses, and other functions that give *S. meliloti* an advantage in its specialized niche.

Genome structure in the *Rhizobiaceae* is quite diverse: members generally possess large, often multipartite genomes (1). For example, some *Agrobacterium* strains have both circular and linear replicons; *Rhizobium* sp. NGR234 has a 3.5-megabase (Mb) chromosome, a >2-Mb megaplasmid, and a smaller 536-kb plasmid that carries most symbiotic functions (2); *Bradyrhizobium japonicum* has a single 8.7-Mb chromosome (3); *Mesorhizobium loti* has two plasmids (352 and 208 kb) and a 7-Mb chromosome (4) that contains a 610-kb "symbiosis island" that is transmissible to other nonsymbiotic mesorhizobia (5). The symbiotic soil bacterium *Sinorhizobium meliloti* strain 1021 has three replicons (3.65, 1.68, and 1.35 Mb), of which SymA is the smallest (6).

Previous size predictions for pSymA ranged from 1.325 to 1.42 Mb, which is comparable to the size of some entire bacterial genomes (7, 8). Most of the previously characterized genes on pSymA had been identified by using classical bacterial genetics to search for genes required for formation of nitrogen-fixing nodules on alfalfa. Clustered within a 275-kb region, these include *nod* genes required for synthesis of Nod factor as well as the *nol* and *noe* genes, which are encoded in six operons on pSymA (9). *nodD1*, *nodD2*, and *nodD3* encode LysR-type transcriptional regulators that activate expression of these operons in response to plant signals or as part of signal-independent regulatory circuitry (10). SyrM and SyrB are pSymA-encoded regulators that also operate within this regulatory circuit (11). Previously discovered *nif* and *fix* genes for symbiotic nitrogen fixation also lie within the 275-kb region (12), as do genes encoding nitrous oxide reductase (*nos*; ref. 13), a functional copy of the *groESL* chaperonin operon (14), and genes needed for catabolism of betaines (15). However, except for *syrB* and a locus that influences symbiotic effectiveness (16), little was known outside this 275-kb region. Recently, an alcohol dehydrogenase (*adhA*; ref. 17) and the rhizobactin regulon (*rhbF*) were physically mapped to pSymA (7).

pSymA of the closely related strain Rm2011 can be cured without affecting growth in either rich or minimal-succinate media, but this strain is defective in the utilization of certain carbon sources (18). Our analysis shows that many genes on pSymA provide versatility to *S. meliloti* and may be adaptive in both the free-living and symbiotic states.

Materials and Methods

Library Construction and Sequencing. Three *S. meliloti* strain 1021 (19) genomic libraries were constructed: one from *Swa*I-digested DNA enriched by pulsed-field gel electrophoresis for the 1.4-Mb linearized pSymA, and two from total genomic DNA (see *Supplemental Text* and Fig. 2, which are published as supplemental data on the PNAS web site, www.pnas.org). Randomly sheared DNA (1–2 kb) (20) was purified by HPLC, cloned into our linker/adaptor version of M13mp18 to minimize chimera formation (R.W.H., unpublished data), and sequenced by using BigDye terminator technology on ABI377-XL sequencers (Applied Biosystems). Base calling used PHRED software (21, 22).

Assembly and Gap Closure. Sequence was assembled with PHRAP (Phil Green, http://www.phrap.org). The final assembly included 32,325 sequence reads from the pSymA-enriched library and 3,441 sequences from the total genomic libraries. Assembly data were viewed in CONSED (23). Sequences from the high-resolution physical map of ordered bacterial artificial chromosomes (BACs) (7) served as a scaffold to order contigs; ordering these known markers confirmed correct assembly. The final average high-quality base coverage was ≈10×. Sequence across the *Swa*I site in pSymA, the restriction site used to purify pSymA via pulsed-field gel electrophoresis, was obtained with reads from the total genomic libraries prepared from undigested DNA. Gaps, single-stranded, single-subclone, and low-quality regions were covered by sequence from PCR products obtained from pSymA BACs (7) or genomic DNA. An error rate of <0.5 per 10,000 bases was computed by using base qualities determined by the PHRAP assembler.

Abbreviations: GABA, 4-aminobutyrate; FDH, formate dehydrogenase; Mb, megabase.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AE006469).

[e]Present address: Institut Curie, 26 Rue d'Ulm, 75005 Paris, France.

[g]Present address: Incyte Genomics, 3160 Porter Drive, Palo Alto, CA 94304.

[i]Present address: Département de Biologie Moléculaire Sciences 2, Université de Genève, Geneva 1211, Switzerland.

[j]Present address: Institute of BioAgricultural Sciences, Academia Sinica, Nankang, Taipei 11529, Taiwan.

[k]Present address: Exelixis, Inc., 170 Harbor Way, P.O. Box 511, South San Francisco, CA 94083-0511.

[l]To whom reprint requests should be addressed. E-mail: srl@leland.stanford.edu.

www.pnas.org/cgi/doi/10.1073/pnas.161294798

Annotation and Analysis. After training GLIMMER 2.0 (24) on a set of 180 known genes from *S. meliloti* strains 1021 and 2011, we used it to predict ORFs in the pSymA sequence. We checked GLIMMER predictions with CODONPREFERENCE (25) and FRAMED (26). We conducted similarity searches by using BLASTP with the National Center for Biotechnology Information/GenBank protein database and HMMER (S. Eddy) with the PFAM Ver. 5.4 database (27). Further analyses used tools available on the *S. meliloti* consortium (6) and EcoCyc web sites (28). We categorized predicted proteins by using a modified Riley classification (29). We used tRNA SCAN Ver. 1.11 to identify potential tRNAs (30). We assigned gene names to predicted ORFs when the analysis supported such assignment. However, these are predicted functions only; proof of function awaits functional tests. Otherwise, predicted genes were designated SMa and predicted proteins, SMA.

Results and Discussion

A concurrent publication on the comparative analysis of the entire genome describes the general structure of pSymA; it reports such features as GC content, codon usage, repeated sequences and putative replication, and transfer functions (6). As global comparisons of the *S. meliloti* genome to the *Rhizobium* sp. NGR234a sym plasmid (31) and the complete genome of *Mesorhizobium loti* (4) are presented in the aforementioned overview paper, we will comment only on specific examples here. The sequence of a 410-kb region of the *B. japonicum* chromosome was recently reported (32), but except for the symbiotic genes (*nod*, *nif*, and *fix*), most genes in the region are not conserved in pSymA. More detailed analyses are available at http://sequence.toulouse.inra.fr/meliloti.html. pSymA more closely resembles plasmids of related bacteria than a true bacterial chromosome (6). We identified 1,293 putative genes on pSymA, yielding a coding capacity of 83.6% (6). Our analysis of pSymA in the context of the total genome failed to find evidence that any of these genes might be absolutely required for free-living growth. However, many genes on pSymA are necessary for nodulation and nitrogen fixation by *S. meliloti*. In addition to the already known genes, the sequence revealed more genes that may be important for symbiosis, utilization of diverse nitrogen and carbon sources, and response to environmental stresses.

Nodulation and Nitrogen Fixation. Nodulation (*nod*) genes encoding Nod factor biosynthetic enzymes and transcription activators are well characterized in *S. meliloti* (Fig. 1; ref. 33). We found no obvious examples of other *nod* genes on the basis of sequence analysis.

Sequence data led us to revise some previously published gene annotations. The *nodM nod* box regulatory sequence presumably controls expression of an operon that was previously reported to consist of six genes (*nodM nolFGHI nodN*; ref. 34). Our analysis indicates that the previously described NolG, NolH, and NolI instead encode a single 1,065-aa protein, NolG, which shares global homology with proteins such as EnvD, CnrA, and CzcA (35). Our analysis also differs from published results for the identification of *nolS*, *nolQa*, and *nolQb* (36). Instead, upstream of *nodD2*, we annotated a 108-aa hypothetical protein (SMA0754).

For nitrogen fixation genes, comparisons with other systems allowed us to search for new loci on the basis of sequence similarity. We confirmed the presence of *nifE* and *nifX*, genes likely to be needed for synthesis of the iron–molybdenum cofactor of nitrogenase, and of an *fdxB* ferredoxin downstream from the *nifHDK* nitrogenase. FixU, a hypothetical protein of unknown function identified in the *nif* regions of *Rhizobium* sp. NGR234, *Rhizobium leguminosarum* bv. *trifolii, B. japonicum*, and *M. loti*, is downstream of the *fixABCX nifAB fdxN* operon on pSymA (SMA0810). SMA1142 is 64% identical to the FixL response-regulator/kinase of *R. leguminosarum* and, like the only FixL reported for *R. leguminosarum*, is adjacent to a Fnr/Crp-like regulator (SMA1141). SMA1142 is not conserved in *M. loti. R. leguminosarum* FixL has a heme-binding domain, a transmitter domain, and a C-terminal receiver domain (37). In *R. leguminosarum*, *fixL* is induced by microaerobic conditions, but mutants are not impaired in *nifA* expression and form nitrogen-fixing nodules (37), unlike a *S. meliloti fixL* mutant (38). The presence of a heme-binding domain in SMA1142 suggests that this protein responds to oxygen concentrations; therefore, it will be interesting to determine what role, if any, this protein plays in symbiosis or other signal transduction pathways. We identified an ORF 58% identical to *B. japonicum* FixR (SMA1757). In *B. japonicum*, FixR function remains elusive, but its gene is cotranscribed with *nifA*, is activated during free-living anaerobic growth and during symbiosis, and may encode a 3-oxoacyl-(acyl carrier protein) reductase (39, 40).

In *B. japonicum*, *fixNOQP* encodes the high-affinity terminal oxidase required for microaerobic respiration and nitrogen

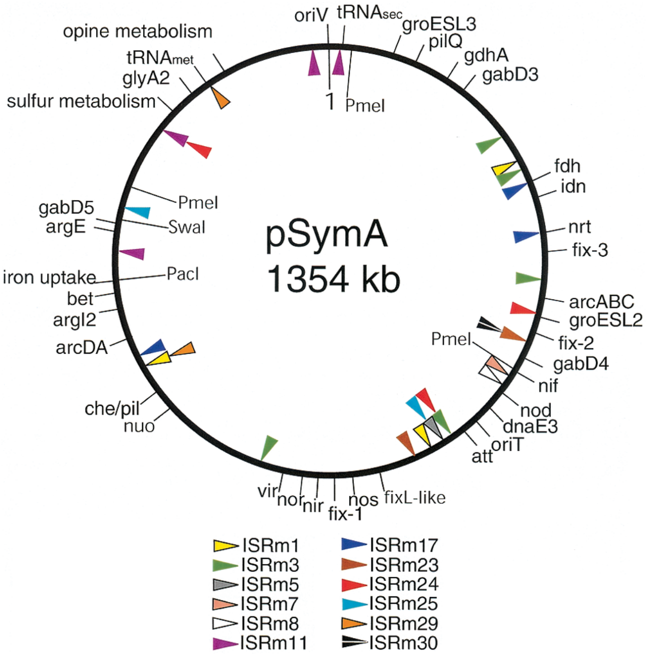

Fig. 1. Map of pSymA. The position of the first nucleotide (denoted 1) was assigned adjacent to and clockwise from the *repABC* genes and putative origin of replication (6). Selected genes and regions mentioned in text are labeled. Locations of insertion sequence (IS) elements and IS fragments are marked with colored triangles. Additional information on IS elements is available in ref. 6 and at the genome web site (http://sequence.toulouse.inra.fr/meliloti.html). Putative chemotaxis genes (*che*), pilus assembly genes (*pil*), and NADH–ubiquinone dehydrogenase genes (*nuo*) are discussed in a concurrent publication (6). Sites of infrequently cutting restriction enzymes (*Swa*I, *Pac*I, and *Pme*I) are marked. SMA designations for regions shown on the map are: tRNA-sec, SMa0011; *groESL3*, SMA0124–125; *pilQ*, SMA0163; *gdhA*, SMA0228; *gabD3*, SMA0260; *fdh*, SMA0478; *idn*, SMA0512–514; *nrt*, SMA0581–585; *fix*-3, SMA0612–622; *arcABC*, SMA0693–697; *groESL2*, SMA0744–745; *fix*-2, SMA0760–769; *gabD4*, SMA0805; *nif*, SMA0825–831; *nod*, SMA0840–878; *dnaE3*, SMA0892; *oriT*, SMa4018; *att*, SMA0950–958; *fixL*-like, SMA1142; *nos*, SMA1179–1188; *fix*-1, SMA107–1229; *nir*, SMA1247–1250; *nor*, SMA1269–1279; *vir*, SMA1302–1321; *nuo*, SMA1516–1536; *che*/*pil*, SMA1552–1578; *arcDA*, SMA1667–1670; *argI2*, SMA1711; *bet*, SMA1726–1731; iron uptake, SMA1740–1747; *argE*, SMA1836; *gabD5*, SMA1848; sulfur metabolism, SMA2067–2103; *glyA2*, SMA2135; tRNA-met, SMa2168; opine metabolism, SMA2223–2225; and *oriV*, SMa4019.

fixation (41, 42). pSymA contains a previously known reiteration of *fixTKNOQP* (fix-2) located 250 kb from the complete and well-studied *fix* cluster (Fig. 1; ref. 43). We found a third reiteration of *fixNOQPIS* (fix-3) about 335 kb from the complete fix-1 cluster. Operon structure diverges in this third set of genes: a gene with no database ortholog, SMa0620, is located between *fixP* and *fixI*, where *fixG* and *fixH* are found in fix-1. The fix-2 reiteration also lacks *fixG* and *fixH* and is flanked by fragments of *fixJ* and *fixI* and by novel ORFs. In *M. loti*, the two reiterations of the *fixNOQPGHIS* clusters have the same gene organization.

S. meliloti FixK1 and FixK2 are Fnr/Crp-like transcription factors that positively control expression of *fixNOQP* and negatively control expression of *nifA* (44, 45). We identified a third FixK-like protein (SMA0662). *M. loti* also has a small family of FixK-like proteins, two of which are present within the symbiotic island.

S. meliloti chromosomal *ntrR* is thought to be responsible for repression of *nod* gene expression in the presence of nitrogen, but mutants display a weak phenotype (46). We identified a putative second copy of this NtrR regulator (SMA0981); it will be interesting to discover whether double mutants exhibit a more severe phenotype.

In addition to searching for structural genes, we examined pSymA for regulatory motifs related to symbiosis. A search for additional *nod* gene promoters (*nod* boxes) yielded no significant matches. A search for NtrA −26 to −10 promoter consensus sequences, NifA upstream activating sequences (UAS), and NtrC UAS (47) identified the previously known consensus sites as well as many that were discounted because of gene organization. One perfect match to the NifA-activated −26 to −10 sequence, located in a region previously characterized (48), lies 326 bp upstream of the translational start of SMa0824, a putative gene that lies immediately upstream of *nifHDKE*; the encoded protein has N-terminal similarity to the C-terminal region of NtrA proteins. We also identified a sequence 173 bp upstream of the predicted start codon for the *nrtAB* nitrate transport genes that has one mismatch to the −26 to −10 NifA-dependent consensus.

We failed to identify *nifQ*, *nifV*, or *nifW* genes and, except for *nifS* and a possible *nifV* on the chromosome (49), the *nif* and *fix* genes listed above appear to be the only newly identified genes similar to known *nif* and *fix* genes from other organisms.

Nitrogen Metabolism. One in twelve of the genes we annotated encodes proteins related to nitrogen metabolism. A 53-kb segment of pSymA is particularly rich in such genes, including a complete pathway for denitrification that surrounds the *fix* gene cluster (fix-1; Fig. 1). We identified a NapAB-type (SMA1236) periplasmic dissimilatory nitrate reductase in this region, but *S. meliloti* apparently lacks a NarGHJI-type membrane-bound nitrate reductase. SMa1250 encodes a nitrite reductase that is 79% identical to *nirK* from a *Pseudomonas* and is associated with a NirV-type (SMA1247) protein (50). A nitric oxide reductase is encoded by the *nor* operon (position 693,908–699,534). Other proteins that are potentially important in denitrification are also here: NnrU (SMA1283), required for expression of *nir* and *nor* genes; Azu1, a blue copper protein associated with periplasmic nitrite reductase; HemN (SMA1266), which is involved in heme maturation; the Crp/Fnr-like regulatory protein NnrR (SMA1245); and the regulator NnrS (SMA1252). The previously identified genes encoding nitrous oxide reductase (*nos*) are also located here (13). From similarity searches, *M. loti* apparently lacks genes encoding denitrification enzymes. Putative genes encoding nitrate transport, *nrtAB*, are located elsewhere on pSymA (SMa0583 and SMa0585; Fig. 1).

The contribution of pSymA to versatile nitrogen metabolism is evident in the variety of genes predicted to be involved in amino acid catabolism, transport, and interconversions. Besides the NodM (Nod factor synthesis) and RhbA (siderophore synthesis) aminotransferases, pSymA encodes six proteins with similarity to aminotransferases whose substrates are unknown (SMA0093, 0387, 1495, 1761, 1855, and 2139).

Many putative genes involved in arginine metabolism are found on pSymA: these may allow *S. meliloti* to use arginine as a sole carbon and nitrogen source. We identified a complete pathway for fermentation of arginine. There are two copies of *arcA* (SMa0693 and SMa1670), which encode arginine deiminases; *arcA1* is in an operon with *arcB* (catabolic ornithine carbamoyl transferase) and *arcC* (carbamate kinase; Fig. 1). Two putative *arcD* genes (SMa1667 and 1668), which encode arginine–ornithine antiporters, are adjacent to *arcA2* (Fig. 1). This pathway may be important for generating ATP and membrane potential in the symbiotic state where oxygen concentrations are low. The arginine deiminase pathway is also found in *Rhizobium etli*, where an *arcA*-reporter is strongly induced under anaerobic conditions. An *arcA* mutant could still fix nitrogen, albeit at a lower efficiency (51), but it is unknown whether a second copy of *arcA* exists in *R. etli*. *M. loti* encodes proteins 90% identical to *S. meliloti* ArcA2 and 65–70% identical to ArcD1 and ArcD2, but the *arcABC* operon is absent. pSymA also encodes a putative second copy of another ornithine carbamoyltransferase, *argI2* (SMa1711), as does *M. loti*.

We also identified a putative *argE* (acetylornithine deacetylase, SMa1836; Fig. 1) that is required for arginine biosynthesis in *Myxococcus* and enteric bacteria; ArgJ (ornithine acetyltransferase) substitutes for ArgE in most other prokaryotes. Because the *S. meliloti* chromosome encodes ArgJ, this ArgE-like protein may provide an alternate means for biosynthesis or possibly a novel catabolic function. *M. loti* also appears to have an *argE* ortholog. SMA0680 and 0682 are more than 50% identical to amino acid decarboxylases that convert arginine, glutamate, ornithine, and lysine to agmatine, GABA, putrescine, and cadaverine, respectively, and are found neither on the other replicons nor in the *M. loti* genome. SMA2203, 2205, and 2209 are putative PotB, PotC, and PotD putrescine/ornithine transport proteins.

pSymA contains a putative HutH histidine ammonia lyase that cleaves histidine to ammonia and urocanate (SMA0306). An ORF with similarity to the histidine utilization repressor is also present, but urocanase, necessary for catabolism of urocanate, is encoded on pSymB (SMb21163; ref. 52).

Free-living rhizobia use the glutamine synthase (GS)-GOGAT pathway instead of the glutamate dehydrogenase (*gdhA*) pathway for assimilation of ammonium. Although *S. meliloti* possesses multiple copies of GS, none are on pSymA. The only copy of *gdhA*, to our knowledge the first one identified in rhizobia, is on pSymA (SMa0228; Fig. 1). Similarity searches failed to find GdhA encoded by *M. loti*. GdhA is probably not required for glutamate catabolism in free-living *S. meliloti* because a pSymA-cured strain can still use glutamate for growth (18). Interestingly, the pSymA-cured strain will not grow on minimal medium with ammonium as the sole nitrogen source, suggesting that pSymA plays a role in ammonium assimilation (M.J.B., unpublished work). pSymA also carries a putative glutamate/aspartate transporter gene (SMa0677).

Opine Metabolism. *S. meliloti* is closely related to *Agrobacterium* spp., pathogens that form crown gall or hairy root tumors on host plants. In these tumors, *Agrobacterium* genes transferred to the plant direct the synthesis of species-specific amino acid derivatives (opines), which can be used as sole carbon and nitrogen sources by the infecting *Agrobacterium*. pSymA contains genes similar to *Agrobacterium tumefaciens* opine catabolic genes: *ooxA* (SMa2223; Fig. 1), *ooxB* (SMa2225; Fig. 1), two cyclodeaminases (SMA0486, 1871), and the AgaE-like deaminase (SMA1869), which converts the opine, mannopinic acid, to mannose and glutamate (53, 54). If sequence similarity is predictive of function, these enzymes appear sufficient to convert opines to proline. *S. meliloti* may transport opines as well: a cluster of genes encode putative ABC transporter proteins 31–48% identical to the OccM (SMA0492), OccQ (SMA0493), and OccT (SMA0495) octopine transporter of *A. tumefaciens*. We identified no ORFs in the *S. meliloti* genome with strong similarity to *A. tumefaciens* opine synthesis enzymes; however, SMB20286 on pSymB is 23% identical to octopine synthase. None of these proteins appear conserved in *M. loti* except SMA1869, which is 70% identical to mll7029. If *S. meliloti* 1021 catabolizes but does not synthesize opines, it raises the possibility that this strain is capable of parasitizing *Agrobacterium* tumors.

Correlation of Carbon Utilization Phenotypes with pSymA Sequence. A *S. meliloti* 2011 mutant lacking pSymA was tested for its ability to use various carbon sources and failed to use inosine, 4-aminobutyrate (GABA), serine, glycine, or gluconate as sole carbon sources (18). We examined whether pSymA contained genes likely to be involved in utilization of these compounds.

None of the predicted enzymes for salvage of inosine are encoded on pSymA; however, it is possible either that inosine catabolism is different in *S. meliloti* or that an unidentified regulatory or transport function is localized on pSymA.

In *Escherichia coli*, GABA utilization requires GABA transaminase, encoded by *gabT,* and succinate semialdehyde dehydrogenase, encoded by *gabD.* A putative *gabT* is located on pSymB. Three aldehyde dehydrogenases on pSymA (SMA0260, 0805, and 1848) are each at least 45% identical to the *E. coli* GabD (Fig. 1). Although the presence of three pSymA-encoded GabD proteins may explain the GABA phenotype of the cured strain, we note that pSymB and the chromosome also each contain a putative *gabD*. Interestingly, a *gabD* homolog from *A. tumefaciens*, *attK*, and an adjacent gene, *attL*, are required for attachment of bacteria to plant cells (55). We found an *attL*-like gene adjacent to the pSymA *gabD3* (SMa0260). *attK* and *attL* also appear conserved in *M. loti*. Experiments are required to determine which of these putative proteins is necessary for GABA utilization and whether they are important for symbiosis.

E. coli cannot use serine and glycine as sole carbon sources, but *S. meliloti* can. Many orthologs of the serine cycle genes needed to catabolize serine and glycine are encoded on pSymA; absence of these genes in the pSymA deletion strain could account for its inability to use serine or glycine as sole carbon sources. Serine cycle enzymes encoded on pSymA include serine glyoxylate aminotransferase (SgaA; SMA2139), serine hydroxymethyltransferase (GlyA2; SMA2135; Fig. 1), glycerate dehydrogenase (SMA2137), hydroxypyruvate reductase (SMA1406), a serine/glycine transaminase similar to eukaryotic enzymes (SMA1495), and a possible serine deaminase (SMA1872). pSymA also encodes a PurU homolog (formyltetrahydrofolate deformylase; SMA2141) and a 5,10-methylene tetrahydrofolate reductase (SMA2143). These two enzymes may be important for interconversion of one-carbon donors of the serine cycle. All of these except SMA2141 and 2143 detect proteins in *M. loti* by BLASTP, but there is no clustering of genes in the same location as for *S. meliloti*. The *S. meliloti* chromosomal copy of *glyA* is most likely involved in glycine biosynthesis. The presence of two copies of *glyA* is interesting, because a *glyA* mutant of *B. japonicum* forms nonfixing nodules (56). The authors of that study proposed that the fixation phenotype resulted from an inability to synthesize glycine, yet the mutant was only a leaky auxotroph for glycine. These results, combined with our data, suggest that the symbiotic phenotype arises from a defect in catabolism.

The pSymA deletion strain cannot use gluconate as a sole carbon source, and our sequence data suggest that the absence of pSymA-encoded thermosensitive-type gluconate kinase activity (IdnK) is responsible for this phenotype; however, pSymB encodes an enzyme similar to *gntK*, the thermoresistant gluconate kinase of *E. coli* (SMB2119). In *E. coli*, either IdnK or GntK suffices to catabolize gluconate; we do not know whether this is the case in *S. meliloti*. Located adjacent to *idnK* are *idnD* and *idnO* (Fig. 1), genes involved in catabolism of the sugar L-idonate in *E. coli* (57). The Idn pathway may enable *S. meliloti* to grow on a variety of sugar acids, with the end product 6-phosphogluconate metabolized via either the pentose phosphate or Entner–Doudoroff pathway. *S. meliloti* may possess a unique mechanism for transport of gluconate because neither of the *E. coli* permeases is present, as judged by sequence similarity.

Transport. We predict that about one of every seven genes on pSymA is involved in transport. The 34 clusters of putative ABC transporter genes are fairly well distributed on pSymA, except a 100-kb region (732,000–833,000) that contains 8 clusters. The high conservation between ABC transporters makes it difficult to predict the transported solute for the majority of these. In addition to the previously known regulon encoding the high-affinity siderophore iron transport system (*rhbABCDEFrhrArhtA*; Fig. 1), there are two clusters of putative iron transport genes at positions 283,000 and 985,000. Three different potassium transporters are present on pSymA: the TrkH- and KUP-types are present on the chromosome, whereas the KdpABC type is unique to pSymA. Potassium transport may be important for pH adaptation during symbiosis, because a Fix$^-$ mutant was shown to be defective in potassium efflux (58).

pSymA contains a large operon that is highly conserved with the *virB1–11* operon of *A. tumefaciens* except that *virB7* is replaced by a hypothetical ORF (Fig. 1). In *A. tumefaciens*, the VirB transport system transfers bacterial tDNA into the plant cell (59). Other bacteria contain Vir homologues, but their function is unknown. *S. meliloti* strains containing deletions of the *virB* operon appear to be normal for nodulation and nitrogen fixation (D.H.W., unpublished data). *M. loti* has orthologs of some of these genes on its chromosome (*virB10*) and pMLa plasmid (*virB4*, *virB6*, *virB8*, *virB9*, and *virB11*).

We found a cluster of genes (529,782–537,888) whose products are 58–70% identical to *A. tumefaciens* proteins required for attachment to plant cells (60). *attABC* encode a putative ABC transporter, and *atrABC* encode proteins similar to a transcriptional regulator, glutamate-1-semialdehyde 2,1-aminomutase, and acetolactate synthase, respectively (55).

Regulatory Proteins. The largest family of transcriptional regulators in *S. meliloti* is the LysR/NodD type; this group is overrepresented on pSymA, with 36 of the 85 total *S. meliloti* members. The phylogenetic subfamily containing the NodD activators of *nod* gene expression has 11 members, 9 of which are on pSymA. We found no SorC or DeoR-type regulators on pSymA, consistent with other data suggesting that pSymA is not specialized for sugar metabolism (6). Nor did we identify any LuxR- or NtrC-like activators on pSymA. We discovered a third SyrB-like regulator on pSymA; none of these are found on the other replicons. Three Crp/Fnr-like regulators are present on pSymA (SMA1067, 1141, and 1245); these constitute a family distinct from the FixK family. We failed to discern any particular bias in the locations of the regulators, as they are fairly evenly distributed about the replicon.

Housekeeping Functions. Two intact *groESL* operons are on pSymA (SMa0744, 0745 and 0124, 0125; Fig. 1) (14). In addition, pSymA encodes a putative DnaJ/CpbA-like chaperonin. Other putative pSymA housekeeping functions that are also encoded on other replicons include: UvrD2, DnaE3, RpoE6, DNA ligase, and DNA-damage inducible protein.

pSymA contains two tRNA genes (Fig. 1). One specifies methionine, is redundant with the one on the chromosome, and may be nonfunctional on the basis of predictions of its secondary structure. The other, with a UCA anticodon specifying selenocysteine, appears unique in the genome. *selC*, encoding the seryl tRNA, is adjacent to genes encoding SelA (selenocysteine synthase), SelD (required to modify seryl tRNA to selenocysteine tRNA), and SelB (selenocysteine-specific elongation factor). A transposon separates *selA* and *selB* from *selC* and *selD*.

Upstream of the *sel* genes is an operon encoding the three subunits of formate dehydrogenase (*fdoGHI*). This formate dehydrogenase (FDH) is 45–61% identical to the O and N isozymes of *E. coli* that contain both molybdenum cofactors and iron–sulfur centers. As expected, the α subunit of this FDH contains a selenocysteine residue, the only predicted selenocysteine codon we were able to identify in the entire *S. meliloti* genome. pSymA has a second FDH, of the homodimeric NAD-dependent type found in methylotrophic bacteria, fungi, and plants (SMA0478; Fig. 1) (61). The chromosome encodes a FdsGBACD-type FDH (49, 62). That pSymA encodes two FDHs may be an indication that formate respiration is important in the oxygen-limited environment of the nodule. *M. loti* lacks *sel* genes as well as the selenocysteine-containing and homodimeric FDHs.

Stress Responses. Little is known about the response of symbiotic soil bacteria to environmental stresses. A number of ORFs specified by pSymA may be involved in stress responses, including three for cold shock (SMA0126, 0181, and 0738) and one for heat shock (SMA1118). A hydroperoxidase (SMA2379) and two haloperoxidases (SMA1809 and 2031) may be part of a protective mechanism in symbiotic or environmental oxidative stresses. SMA2389 is 57–63% identical to the Ohr stress-induced proteins from *M. loti* and *B. japonicum*, respectively (63).

SMA1896 is 43% identical to domains of methionine sulfoxide reductase, an enzyme that protects against oxidative damage by repairing both free and incorporated methionine (64). SMA1547 is the best match in the entire genome to *E. coli* PimT, L-isoaspartate protein carboxymethyl transferase, and may play a role in repair or degradation of damaged proteins.

Other putative pSymA proteins may confer resistance to toxins via export such as the AcrB-like cation efflux pumps: SMA1664 and SMA1662 are 32 and 44% identical to AcrA and AcrB, respectively. SMA1884 is also similar to AcrB and is adjacent to an AcrR-like regulator (SMA1882). The CopC copper export protein (SMA1198) may help *S. meliloti* withstand toxic levels of copper and thus aid survival in certain soils.

It is important for a symbiotic soil bacterium to be able to withstand osmotic stress. Betaine aldehyde dehydrogenase (BetB2, SMA1731), which catalyzes the second step in betaine synthesis, is encoded near genes for a putative glycine–betaine-binding lipoprotein (SMA1729), a regulator of betaine synthesis, BetI (SMA1726), and a lipase (SMA1727; Fig. 1). The *S. meliloti* chromosome also encodes a betaine aldehyde dehydrogenase, as well as choline dehydrogenase, BetA (65). The pSymA BetI may be redundant with chromosomal SMC00095 (49). SMA1466 and 1467 are 44% identical to ABC transport proteins for glycine, betaine, carnitine, and choline.

OtsA, trehalose synthase (SMA0233), presumably is required for synthesis of trehalose, an endogenous osmolyte in *S. meliloti* (66). *otsAB* are cotranscribed in *Rhizobium* NGR234, and mutants form small less effective nodules (67). *M. loti* appears to contain an *otsAB* operon as well. Surprisingly, *S. meliloti* lacks OtsB, trehalose phosphatase, leading us to speculate that trehalose synthesis occurs via a different pathway.

Sulfur Metabolism. We showed earlier that pSymA encodes a NodH sulfotransferase, a NodP1 PAPS synthase, and a NodQ1 APS kinase (68, 69), enzymes responsible for the synthesis and transfer of activated sulfate to Nod factor. Our analysis identifies additional putative enzymes involved in sulfur metabolism on pSymA. Seven lie within a 21-kb region (Fig. 1): a sulfite oxidase (SMA2103), sulfate/thiosulfate transport proteins (SMA2067 and 2069), and proteins similar to desulfurization enzymes (SMA2073, 2087, 2093, and 2101). The last are similar to the dibenzothiophene desulfurization proteins A, B, and C and to sulfonate-binding proteins and may be important for scavenging sulfur during sulfate or cysteine starvation. The transport proteins may be present in *M. loti*, whereas BLAST analysis of the putative sulfite oxidase and desulfurization enzymes returned only weak or no matches. Two arylsulfatases may be important for sulfur scavenging: one (SMA0943) is 51 and 79% identical to a putative arylsulfatase from *B. japonicum* and mll5471 from *M. loti*, respectively. The other is similar to the human ArsA-type (SMA1683). Also, we identified a sulfate uptake protein, SMA1916, which is 56% identical to the chromosomal protein SMC04179.

Putative Calcium-Binding Proteins. SMA0060 is 31% identical over most of its length to the eukaryotic calcium-binding protein regucalcin/senescence marker protein-30, and a similar protein is present in *M. loti*. SMA0717, not conserved in *M. loti*, is 33% identical to regucalcin, but only in the C-terminal domain; the N-terminal domain is similar to IclR-type DNA-binding domains. Regucalcin is proposed to be a calcium-binding protein that is stimulated by calcium, calcitonin, insulin, and estrogen, but contains no EF-hand motif (70). And last, SMA2111 has a motif similar to the calcium-binding protein, hemolysin, and to the *R. leguminosarum* NodO; an ortholog was not found in *M. loti*.

Conclusions

Many genes found on pSymA are similar to interesting genes of other bacteria but have no obvious symbiotic function. pSymA is clearly specialized for nodulation and nitrogen fixation, and our analysis suggests some additional symbiotic loci. In addition, the many pSymA genes involved in nitrogen metabolism reveal their significant role in providing versatility for dealing with nitrogen in many oxidation states and chemical combinations. Physiologically, low oxygen conditions characterize the nodule environment and may be encountered by rhizobia in soil. Thus it is interesting to find that symbiotic genes are linked to other genes likely to be useful under low oxygen conditions.

Last, publication of the pSymA DNA sequence should be viewed as the starting point for an era of intensive research. This invaluable resource will facilitate all future molecular genetic studies on this important model microorganism. The study of individual genes can now be viewed in the context of the organism's potential, and the often alien world of symbiosis can now be viewed as the result of the action of a large, but known, set of possibilities. We now have the opportunity to use this information to attempt to fully understand the metabolic and symbiotic complexities that allow *S. meliloti* to successfully occupy its niches, surviving in the soil, infecting plants, and fixing atmospheric dinitrogen into ammonia for a world where fixed nitrogen is frequently in short supply.

We are grateful to fellow members of the *S. meliloti* consortium for useful discussions. We thank Lee Kozar for help with computers and software and Glen Barnett for help with software. This project was supported by Department of Energy (DOE) Energy Biosciences Grant DE-FG03-90ER20010 to S.R.L. and by the Howard Hughes Medical Institute. Additional funds were provided by National Institutes of Health Grant GM30962 to S.R.L. and DOE Grant DE-FG03-99ER20225 to M.K.

1. Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., Ramuz, M. & Allardet-Servent, A. (1998) *J. Bacteriol.* **180,** 2749–2755.
2. Viprey, V., Rosenthal, A., Broughton, W. J. & Perret, X. (2000) *Genome Biol.* **1,** 1–17.
3. Kündig, C. H., Hennecke, H. & Göttfert, M. (1993) *J. Bacteriol.* **175,** 613–622.
4. 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.
5. Sullivan, J. T. & Ronson, C. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5145–5149.
6. Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, A. P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., *et al.* (2001) *Science* **293,** 668–672.
7. Barloy-Hubler, F., Capela, D., Barnett, M. J., Kalman, S., Federspiel, N. A., Long, S. R. & Galibert, F. (2000) *J. Bacteriol.* **182,** 1185–1189.
8. Honeycutt, R. J., McClelland, M. & Sobral, B. W. S. (1993) *J. Bacteriol.* **175,** 6945–6952.
9. Schlaman, H. L., Phillips, D. A. & Kondorosi, E. (1998) in *The Rhizobiaceae*, eds. Spaink, H. P., Kondorosi, A. & Hooykaas, P. J. J. (Kluwer, Dordrecht, The Netherlands), pp. 361–386.
10. Swanson, J. A., Mulligan, J. T. & Long, S. R. (1993) *Genetics* **134,** 435–444.
11. Barnett, M. J. & Long, S. R. (1997) *Mol. Plant–Microbe Interact.* **5,** 550–559.
12. Kaminski, P. A., Batut, J. & Boistard, P. (1998) in *The Rhizobiaceae*, eds. Spaink, H. P., Kondorosi, A. & Hooykaas, P. J. J. (Kluwer, Dordrecht, The Netherlands), pp. 431–460.
13. Holloway, P., McCormick, W., Watson, R. J. & Chan, Y.-K. (1996) *J. Bacteriol.* **178,** 1505–1514.
14. Ogawa, J. & Long, S. R. (1995) *Genes Dev.* **9,** 714–729.
15. Goldmann, A., Boivin, C., Fluery, V., Message, B., Lecoeur, L., Maille, M. & Tepfer, D. (1991) *Mol. Plant–Microbe Interact.* **4,** 571–578.
16. Sharypova, L. A., Yurgel, S. N., Keller, M., Simarov, B. V., Pühler, A. & Becker, A. (1999) *Mol. Gen. Genet.* **261,** 1032–1044.
17. Willis, L. B. & Walker, G. C. (1998) *Biochim. Biophys. Acta* **1384,** 197–203.
18. Oresnik, I. J., Liu, L.-L., Yost, C. K. & Hynes, M. F. (2000) *J. Bacteriol.* **182,** 3582–3586.
19. Meade, H. M., Long, S. R., Ruvkun, G. B., Brown, S. E. & Ausubel, F. M. (1982) *J. Bacteriol.* **149,** 114–122.
20. Thorstenson, Y. R., Hunicke-Smith, S. P., Oefner, P. J. & Davis, R. W. (1998) *Genome Res.* **8,** 848–855.
21. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8,** 175–185.
22. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
23. Gordon, D., Abajian, C. & Green, P. (1998) *Genome Res.* **8,** 195–202.
24. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998) *Nucleic Acids Res.* **26,** 544–548.
25. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
26. Schiex, T., Thébault, P. & Kahn, D. (2000) in *JOBIM Conference Proceedings* (*Montpellier, France*), pp. 321–328.
27. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. L. (2000) *Nucleic Acids Res.* **28,** 263–266.
28. Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. & Krummenacker, M. (1999) *Nucleic Acids Res.* **27,** 55–58.
29. Riley, M. (1993) *Microbiol. Rev* **57,** 862–952.
30. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25,** 955–964.
31. Freiberg, C., Fellay, R., Bairoch, A., Broughton, W. J., Rosenthal, A. & Perret, X. (1997) *Nature (London)* **387,** 394–401.
32. Göttfert, M., Röthlisberger, S., Kündig, C., Beck, C., Marty, R. & Hennecke, H. (2001) *J. Bacteriol.* **183,** 405–1412.
33. Downie, J. A. (1998) in *The Rhizobiaceae*, eds. Spaink, H. P., Kondorosi, A. & Hooykaas, P. J. J. (Kluwer, Dordrecht, The Netherlands), pp. 387–402.
34. Baev, N., Endre, G., Petrovics, G., Banfalvi, Z. & Kondorosi, A. (1991) *Mol. Gen. Genet.* **228,** 113–124.
35. Saier, M. H., Tam, R., Reizer, A. & Reizer, J. (1994) *Mol. Microbiol.* **11,** 841–847.
36. Plazanet, C., Réfrégier, G., Demont, N., Truchet, G. & Rosenberg, C. (1995) *FEMS Microbiol. Lett.* **133**.
37. Patschkowski, T., Schlüter, A. & Priefer, U. B. (1996) *Mol. Microbiol.* **21,** 267–280.
38. David, M., Daveran, M.-L., Batut, J., Dedieu, A., Domergue, O., Ghai, J., Hertig, C., Boistard, P. & Kahn, D. (1988) *Cell* **54,** 671–683.
39. Barrios, H., Fischer, H.-M., Hennecke, H. & Morett, E. (1995) *J. Bacteriol.* **177,** 1760–1765.
40. Fischer, T. B., Anthamatten, D., Bruderer, T. & Hennecke, H. (1987) *Nucleic Acids Res.* **15,** 8479–8499.
41. Preisig, O., Anthamatten, D. & Hennecke, H. (1993) *Proc. Natl. Acad. Sci.* **90,** 3309–3313.
42. Preisig, O., Zufferey, R., Thony-Meyer, L., Appleby, C. A. & Hennecke, H. (1996) *J. Bacteriol.* **178,** 1532–1538.
43. Renalier, M.-H., Batut, J., Ghai, J., Terzaghi, B., Gheerardi, M., David, M., Garnerone, A.-M., Vasse, J., Truchet, G., Huguet, T., *et al.* (1987) *J. Bacteriol.* **169,** 2231–2238.
44. Batut, J., Daveran-Mingot, M.-L., David, M., Jacobs, J., Garnerone, A. M. & Kahn, D. (1989) *EMBO J.* **8,** 1279–1286.
45. Foussard, M., Garnerone, A.-M., Ni, F., Soupene, E., Boistard, P. & Batut, J. (1997) *Mol. Microbiol.* **25,** 27–37.
46. Dusha, I. & Kondorosi, A. (1993) *Mol. Gen. Genet.* **240,** 435–444.
47. Gussin, G. N., Ronson, C. W. & Ausubel, F. M. (1986) *Annu. Rev. Genet.* **20,** 567–591.
48. Better, M., Lewis, B., Corbin, D., Ditta, G. & Helinski, D. R. (1983) *Cell* **35,** 479–485.
49. Capela, D., Barloy-Hubler, F., Gouzy, J., Bothe, G., Ampe, F., Batut, J., Boistard, P., Becker, A., Boutry, M., Cadieu, E., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 9877–9882. (First Published July 31, 2001; 10.1073/pnas.161294398)
50. Bedzyk, L., Wang, T. & Ye, R. W. (1999) *J. Bacteriol.* **181,** 2802–2806.
51. D'Hooghe, I., Vander Wauven, C., Michiels, J., Tricot, C., de Wilde, P., Vanderleyden, J. & Stalon, V. (1997) *J. Bacteriol.* **179,** 7403–7409.
52. 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)
53. Cho, K., Fuqua, C. & Winans, S. C. (1997) *J. Bacteriol.* **179,** 1–8.
54. Lyi, S. M., Jafri, S. & Winans, S. C. (1999) *Mol. Microbiol.* **31,** 339–347.
55. Matthysse, A., Yarnall, H., Boles, S. B. & McMahan, S. (2000) *Biochim. Biophys. Acta* **1490,** 208–212.
56. Rossbach, S. & Hennecke, H. (1991) *Mol. Microbiol.* **5,** 39–47.
57. Bausch, C., Peekhaus, N., Utz, C., Blais, T., Murray, E., Lowary, T. & Conway, T. (1998) *J. Bacteriol.* **180,** 3704–3710.
58. Putnoky, P., Kereszt, A., Nakamura, T., Endre, G., Grosskopf, E., Kiss, P. & Kondorosi, A. (1998) *Mol. Microbiol.* **28,** 1091–1101.
59. Kado, C. I. (2000) *Curr. Opin. Microbiol.* **3,** 643–648.
60. Matthyse, A., Yarnall, H. A. & Young, N. (1996) *J. Bacteriol.* **178,** 5302–5308.
61. Lamzin, V. S., Aleshin, A. E., Strokopytov, B. V., Yukhnevich, M. G. & Popov, V. O. (1992) *Eur. J. Biochem.* **206,** 441–452.
62. Oh, J.-I. & Bowien, B. (1998) *J. Biol. Chem.* **273,** 26349–26360.
63. Muller, P., Ahrens, K., Keller, T. & Klaucke, A. (1995) *Mol. Microbiol.* **18,** 831–840.
64. Lowther, W. T., Brot, N., Weissbach, H. & Matthews, B. W. (2000) *Biochemistry* **39,** 13307–13312.
65. Pocard, J.-A., Vincent, N., Boncompagni, E., Tombras-Smith, L., Poggi, M.-C. & Le Rudulier, D. (1997) *Microbiology* **143,** 1369–1379.
66. Gouffi, K., Pica, N., Pichereau, V. & Blanco, C. (1999) *Appl. Environ. Microbiol.* **65,** 1491–1500.
67. Englehard, M., Viprey, V., Perret, X., Broughton, W., Wiemken, A., Boller, T. & Müller, J. (1999) in *Molecular Plant Microbe Interactions* (International Society for Molecular Plant–Microbe Interactions, St. Paul, MN), pp. 177–178.
68. Ehrhardt, D. W., Atkinson, E. M., Faull, K. F., Freedberg, D. I., Sutherlin, D. P., Armstrong, R. & Long, S. R. (1995) *J. Bacteriol.* **177,** 6237–6245.
69. Schwedock, J. & Long, S. R. (1990) *Nature (London)* **348,** 644–647.
70. Yamaguchi, M. (2000) *Life Sci.* **66,** 1769–1780.