The complete sequence of the 1,683-kb pSymB megaplasmid from the N_2-fixing endosymbiont *Sinorhizobium meliloti*

Turlough M. Finan*†‡, Stefan Weidner*§, Kim Wong†, Jens Buhrmester§, Patrick Chain†, Frank J. Vorhölter§, Ismael Hernandez-Lucas†, Anke Becker§, Alison Cowie†, Jérôme Gouzy¶, Brian Golding†, and Alfred Pühler§

†Department of Biology, McMaster University, 1280 Main Street West, Hamilton, ON, Canada L8S 4K1; §Universität Bielefeld, Fakultät für Biologie, Lehrstuhl für Genetik, Universitätsstrasse 25, D-33615 Bielefeld, Germany; and ¶Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes, Unité Mixte de Recherche 215, Chemin de Borde Rouge, BP27, F-31326 Castanet Tolosan, France

Communicated by Sharon R. Long, Stanford University, Stanford, CA, June 12, 2001 (received for review March 13, 2001)

Analysis of the 1,683,333-nt sequence of the pSymB megaplasmid from the symbiotic N_2-fixing bacterium *Sinorhizobium meliloti* revealed that the replicon has a high gene density with a total of 1,570 protein-coding regions, with few insertion elements and regions duplicated elsewhere in the genome. The only copies of an essential arg-tRNA gene and the *minCDE* genes are located on pSymB. Almost 20% of the pSymB sequence carries genes encoding solute uptake systems, most of which were of the ATP-binding cassette family. Many previously unsuspected genes involved in polysaccharide biosynthesis were identified and these, together with the two known distinct exopolysaccharide synthesis gene clusters, show that 14% of the pSymB sequence is dedicated to polysaccharide synthesis. Other recognizable gene clusters include many involved in catabolic activities such as protocatechuate utilization and phosphonate degradation. The functions of these genes are consistent with the notion that pSymB plays a major role in the saprophytic competence of the bacteria in the soil environment.

Among the bacteria, the α-proteobacteria appear unusual because of the presence of multiple replicons within the same bacterial strain (1). In the case of *Agrobacterium tumefaciens*, the causative agent of crown gall disease, the genome contains both a linear and a circular chromosome (2). Many (but not all) of the bacteria that form N_2-fixing root nodules on leguminous plants are characterized by the presence of multiple plasmids greater than 400 kb in size. In the case of the N_2-fixing symbiont *Sinorhizobium meliloti*, there are three replicons, a 3,654-kb circular chromosome (3, 4) and two megaplasmids 1,354 and 1,683 kb in size (5–7). The smaller of the megaplasmids, variously called pSymA, pNod-Nif, or pRmeSU47a, is known to carry many of the genes involved in root nodule formation (*nod*) and nitrogen fixation (*nif*) (8, 9).

The 1,683-kb megaplasmid, referred to as pSymB, pExo, or pRmeSU47b, is known to carry various gene clusters involved in exopolysaccharide (EPS) synthesis, C_4-dicarboxylate transport, and lactose metabolism (10–12). Early studies focused on mutations that abolished synthesis of the succinoglycan EPS, EPS I, because these mutations resulted in a loss of the ability to form normal N_2-fixing root nodules. This symbiotic defect was rescued by second-site mutations that increased the synthesis of a second galactoglucan EPS (EPS II), whose biosynthetic genes were also located on the pSymB megaplasmid (13, 14). Other genes located on pSymB that are required for the formation of N_2-fixing root nodules include the C_4-dicarboxylate (*dctA*) and phosphate transport (*phoCDET*) genes and the *bacA* gene (15–18). The presence of large plasmids in bacteria that form associations with plants was described over 20 years ago (19). However, with the exception of the symbiotic genes in relatively small regions of these plasmids, the broader biological role of the plasmids in the biology of the organism has remained obscure. We constructed a genetic map of the pSymB megaplasmid and then generated strains in which over 1,200 kb of pSymB was removed (20, 21). Phenotypic analysis of these deletion derivatives revealed a small number of phenotypes; however, the functions of over 90% of the plasmid remain unknown.

Here we describe the complete nucleotide sequence and annotation of the pSymB megaplasmid. We find that the plasmid has a high gene density, similar to that of previously sequenced bacterial chromosomes. Annotation of the sequence revealed a previously unsuspected richness with respect to the number of solute transport systems, polysaccharide synthesis gene clusters, transcriptional regulators, cell protection, and other genes that appear to have catabolic roles. In addition, the megaplasmid contains genes that we assume are essential for viability of the bacteria. The question whether the pSymB replicon should be called a megaplasmid or chromosome is one of the issues that arises from the nucleotide sequence analysis described in this report. For the sake of clarity, and to be consistent with the previous literature, we refer to the 3.6-Mb replicon as the chromosome and the pSymB and pSymA replicons as megaplasmids.

Materials and Methods

Bacterial Strains. *S. meliloti* was isolated from New South Wales, Australia, in 1939 and was originally referred to as strain SU47 (22). DNA from the streptomycin-resistant SU47 derivative Rm1021 was used for sequencing.

Library Construction and Sequencing of Shotgun Clones. The large DNA inserts from pSymB region Ω5056 to Ω5102 were cloned into BAC vectors by using an *in vivo* cloning procedure (23), and the BAC DNA then was purified, nebulized, size fractionated to 1- to 2-kb fragments, and cloned into phosphatase-treated *Sma*I-restricted pUC118. A minimal set of 24 overlapping BAC clones (24) covering the complete pSymB megaplasmid was also used to construct shotgun libraries with different insert sizes (1.3–1.7 and 2.8–3.2 kb) by nebulization of BAC clone DNA and cloning into pTZ18R (Amersham Pharmacia).

Sequencing of shotgun clones for all of pSymB was performed by using M13 forward and reverse standard sequencing primers to achieve ≈7.5-fold coverage. Contigs were closed by PCR amplification from specific primers. Sequencing reactions were performed with dye-terminator and dye-primer chemistry on ABI Prism

MICROBIOLOGY

Abbreviations: ABC, ATP-binding cassette; EPS, exopolysaccharide; CPS, capsular polysaccharide.

Data deposition: The sequence reported in this paper has been deposited in the EMBL database (accession no. AL591985).

*T.M.F. and S.W. contributed equally to this work.

‡To whom reprint requests should be addressed. E-mail: finan@mcmaster.ca.

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.

377–96, ABI 373S (Applied Biosystems), and LI-COR LongRead IR 4200 (Li-Cor, Lincoln, NE) DNA sequencers.

Contig Assembly and Sequence Editing. The replicon was partly sequenced and annotated by a team based at McMaster University and partly by a team based at the University of Bielefeld. The final sequence and annotation were checked and corrected by both groups. Sequence data were processed and assembled by using the ACEDB assembly package [R. Durbin and J. Thierry Mieg (1998) http://htp.ncbi.nlm.nih.gov/repository/acedb] at McMaster University or the PHRED/PHRAP and STADEN (GAP4) packages of base-calling, sequence assembly, and finishing/editing software at the University of Bielefeld (25–27). The vector sequence was identified and removed from individual clones before assembly.

Regions of base-call uncertainty, low coverage, or potential frameshifts were resequenced from shotgun clones, BAC clones, or from sequencing of the PCR product from amplification of genomic DNA. Custom-made primers designed by PRIDE (28) were used for these purposes.

Sequence Analysis and Annotation. Annotation of the completed sequence was performed with the aid of various software packages. Individual ORFs greater than 150 bp with homology to BLAST hits (expect value greater than $e - 10$) were examined for the presence of ribosomal binding site(s). Potential start sites were decided by considering ATG, GTG, and TTG start sites. FrameD, a hidden Markov chain method trained for *Rhizobium* (ref. 29; see also http//www.toulouse.inra.fr/FrameD.html), was also used in conjunction with BLAST results to determine potential ORFs and start sites. Final editing of the annotation included a search for homology with genes within the three replicons, and annotations were confirmed with PRODOM information (30).

Annotation at the University of Bielefeld used the GEN-DB annotation database environment developed at Bielefeld. ORFs were predicted by using GLIMMER 2.0 (31). Each ORF was subject to comparison with sequences in public databases by using BLAST (32). Additionally, PFSCAN (www.isrec.isb-sib.ch) and SIGNALP (www.cbs.dtu.dk/services/SignalP) were used to search protein sequence profile databases and to identify signal peptides. Predicted ORFs were reviewed individually by annotators for misprediction start-codon assignment based on contextual information and assignment of categories based on the functional gene classification of *Escherichia coli* (33). Rhizobium-specific intergenic mosaic elements, motif A, B, C palindromic elements (34, 35) and insertion sequence elements were included in the annotation of the complete sequence.

The complete pSymB sequence can be viewed on the consortium web site at http://sequence.toulouse.inra.fr/meliloti.html.

Results and Discussion

General Features of the Nucleotide Sequence and Replicon Organization. The total length of the *S. meliloti* pSymB megaplasmid is 1,683,333 bp, with an overall G + C content of 62.4%. Variation of G + C content throughout the replicon is evident, ranging from as low as ≈56% in the *rkp* gene region (surface polysaccharide-associated export protein) to as high as ≈66% (within windows of 10,000- and 5,000-bp steps). Ninety percent of the nucleotide sequence is predicted as protein coding, constituting 1,570 ORFs with a mean length of 959 bp. The distribution of ORFs on the forward and reverse strands is symmetrical, with 52% on the forward strand and 48% on the reverse strand (see Fig. 1). A limited number of insertion sequence elements (12), rhizobium-specific intergenic mosaic elements (26), and motif A, B, C elements (5) were also identified throughout the megaplasmid.

Origin of Replication and *rep* Genes. The origin of replication of pSymB is predicted to lie within the r*epA1B1C1* gene cluster, and subclones of this region are sufficient to allow its autonomous replication in *Agrobacterium* (23). There is a second *repAB* gene set (but lacking the *repC* gene) (*repA3B3*) lying 131 kb from the *repA1B1C1* cluster. RepC is believed to be directly involved in the control of the initiation of plasmid replication, whereas RepA and RepB appear to play a role in plasmid segregation. Whether the *repA3B3* genes play redundant roles in plasmid segregation in *S. meliloti* remains to be established. We note that RepA1 has higher homology to RepA from other organisms than to the pSymB RepA3 protein or the RepA2 protein encoded by pSymA.

Cell Division, Chaperonins, and Arg-tRNA. The MinCDE proteins play important roles in the placement of the cell division site in *E. coli* (36). The only proteins in *S. meliloti* homologous to the *Neisseria, E. coli*, and *Xylella fastidiosa* MinCDE are located downstream of a two-component sensor histidine kinase and response regulator (but coded on the opposite strand) and a glycine-rich hypothetical protein. Whether the *S. meliloti minCDE* genes are essential has yet to be established; however, we note that these genes appear to lie in a region for which deletion derivatives have not previously been identified (21).

The FtsK protein plays a role in septa formation and recombinational resolution of dimeric chromosomes in *E. coli*. Two FtsK homologs in *S. meliloti* share 74% amino acid identity; one is located on the chromosome, and the other (FtsK2) (homolog of SpoIIIE) is located on pSymB close to the *repA3B3* genes. We believe the chromosomal *ftsK1* copy is functional for cell division, because deletion derivatives that remove the *ftsK2* region of pSymB showed no obvious growth defects. It is interesting that FtsK2 is more similar in sequence and length to the *Rickettsia prowazekii* (AJ235273) FtsK homolog than is FtsK1.

There are four chaperonins on pSymB: *groEL5*, *htpG*, which is a member of heat-shock-protein Hsp90 family (37), and two members of the Hsp20 family (SMb21294 and SMb21295).

The sole copy of the tRNA specific for the second most frequently used arginine codon, CCG, is located on pSymB next to a putative transposase gene (SMb20905) followed by a 3-kb region containing seven small, possibly degenerating ORFs of unknown function. In view of this sequence context, it is possible that the location of this tRNA gene resulted from a transposition event.

EPS and Lipopolysaccharide (LPS) Biosynthetic Genes. Gram-negative bacteria exhibit complex sets of surface polysaccharides, including LPS (38), capsular polysaccharides (CPS) (39), EPS (40, 41), and periplasmic glucans (42). Genes involved in the biosynthesis and export of cell surface carbohydrates are often clustered, and the *exo*/*exs* and *exp* gene clusters directing the synthesis of the EPS succinoglycan (EPS I) (43, 44) and galactoglucan (EPS II) (13, 14) were previously mapped on pSymB of *S. meliloti*. The production of surface polysaccharides is essential for successful nodule invasion by rhizobia.

Analysis of the DNA sequence of pSymB revealed many genes whose products are typically involved in the synthesis of cell surface carbohydrates. Most of these genes are organized in 11 clusters (Table 1), but surprisingly, the existence of nine of these clusters ranging in size from 5 to 42 kb was previously unknown. In addition to these gene clusters, there are several isolated genes whose products also appear to be involved in the synthesis of cell surface carbohydrates. The 11 gene clusters contain 188 predicted genes and have a total size of 223 kb. Hence over 12% of the genes on pSymB are involved in the synthesis of cell surface carbohydrates.

The pSymB gene clusters 3, 4, 8 (*exo*/*exs*), and 9 (Table 1) comprise genes encoding proteins of the Wzy-dependent polymerization mechanism, although the specific target polysaccharide cannot be specified. In clusters 3 and 4, some genes encoding key functions of the Wzy-dependent export mechanism are missing. Cluster 2 contains ATP-binding cassette (ABC)-2 export proteins, and several of these find their closest homologs in

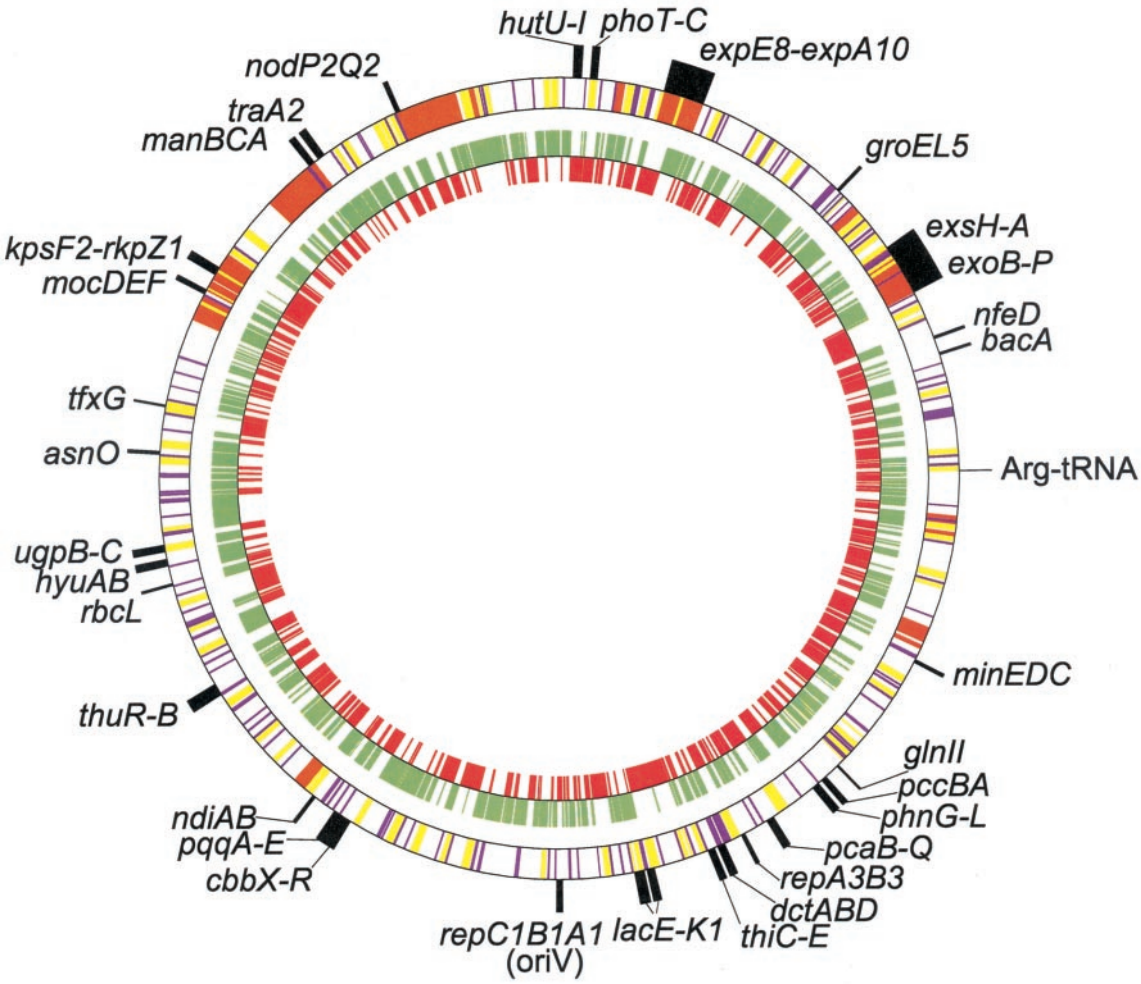

Fig. 1. Genome map of the pSymB megaplasmid of *Sinorhizobium meliloti* strain 1021. The inner circle displays ORFs on the leading (red) and lagging (green) strands of pSymB. The outer circle shows predicted gene regions encoding transcriptional regulators (pink), ABC transport systems (yellow), and genes involved in polysaccharide biosynthesis (orange). The positions of specific genes or sets of genes (e.g., *exp*, *exs*, and *exo*) are also shown on the outer edge of the map.

the export machinery of CPS, implying that this cluster may be involved in the export of CPS (39). Cluster 2 is homologous to *rpk* loci in *S. meliloti* strain 41 (45) and comprises genes *rkpRSTZ* and probable homologs of *rkpLM*. Other *rpk* genes (*rkpAGHIS* and *rkpK-lpsL*) are located on the chromosome of strain 1021. The *rkp* genes code for K antigens, CPS that have been implicated in root nodule invasion, and can functionally replace EPS I or EPS II during the invasion process (46).

Besides the ABC-transporter and Wzy-dependent export mechanisms, a third mechanism involving the ABC-transporter protein MsbA has been proposed to mediate the export of nascent lipopolysaccharide (LPS) molecules consisting of the lipid A and core components (47). A MsbA-like protein encoded by the *S. meliloti ndvA* gene is involved in the export of β-glucan (48). Other MsbA-homologs have been identified in pSymB gene clusters 1, 5, and 8 (*exo*/*exs*). Although cluster 1 MsbP protein may be involved

MICROBIOLOGY

Table 1. Overview on cell surface carbohydrate synthesis gene clusters of pSymB

Cluster	Size, kb	Gene number	ORFs	Surface carbohydrate
1	14	11	SMb20803–SMb20816	Possibly LPS core/ lipid A
2	26	27	SMb20821–SMb21013	Possibly CPS
3	40	33	SMb21050–SMb21082	Unidentified
4	42	33	SMb21223–SMb21256	Unidentified
5	6	4	SMb21188–SMb21191	Unidentified
6	28	21	expE8-expA10	EPS II
7	5	5	SMb21581–SMb21585	Unidentified
8	25	23	exsH-exoP	EPS I
9	13	10	SMb21499–SMb2150 + SMb21512–SMb21513	Unidentified
10	14	14	SMb21416–SMb21428 + SMb20803–SMb20834	Unidentified
11	9	8	SMb20238–SMb20245	Unidentified

in the synthesis and export of nascent LPS molecules, the functions of the other MsbA-like proteins are unknown.

Genes Involved in Metabolic Pathways

Biosynthesis. With the exception of thiamin biosynthesis, few pSymB genes are predicted to play roles in amino acid or vitamin biosynthesis (SMb20481, SMb20652, and SMb21494). Of the genes predicted to be involved in amino acid biosynthesis (e.g., *aroE2*, shikimate 5-dehydrogenase; *glnII*, glutamine synthetase II), all have chromosomal homologs. Two pSymB gene regions are involved in thiamin biosynthesis, *thiCOSGE* and *thiD*, and mutations at either of these loci result in thiamin auxotrophy (10). Despite the presence of these genes, it has been determined that availability of thiamin is growth-limiting in the rhizosphere and is a key factor promoting root colonization (49).

One of the *thi* clusters (*thiCOGE*) lies next to the *dct* genes encoding a C_4-dicarboxylate transport system. Because *thi* and *dct* genes are found in the symbiotic island of *Mesorhizobium loti* and *Bradyrhizobium japonicum*, it will be interesting to compare the organization of these regions across these three organisms (50, 51) to determine the extent of gene shuffling and/or horizontal gene transfer that has occurred since the divergence of these species.

Asparagine is synthesized from aspartic acid and ammonia (or glutamine) by the enzyme asparagine synthetase. Two *S. meliloti* genes, *asnB* and *asnO*, encode proteins similar in size and sequence (27–33% identity) to asparagine synthetases of *E. coli*, and both are located on pSymB. Although the activity of these proteins remains to be established, in the absence of an alternate route to synthesize asparagine, it would appear that at least one of these genes is essential for growth of *S. meliloti* in minimal medium. Interestingly, *asnB* is located only 5 kb upstream from the *minCDE* genes, which are also potentially essential to cell viability.

Catabolism. A number of putative enzymes involved in amino acid degradative pathways were identified. These include components of the enzymes 2-oxoisovalerate dehydrogenase (SMb20019), propionyl-CoA carboxylase (*pccAB*), and methylmalonyl-CoA mutase (*bhbA*), which are involved in valine, leucine, and isoleucine degradation. The histidine utilization genes (*hutUGHIL*), whose products are responsible for the conversion of L-histidine to glutamate and formamide, are present as a single operon. Previously identified genes involved in the utilization of poly3-hydroxybutyrate, the α-galactosides, melibose, raffinose (*apaA, agpL, agpT*), and lactose (*lacEFGZK*), including the two previously hypothesized *lacZ* genes, were identified (21, 52–55).

A 20-kb region of pSymB (SMb21277 to SMb21293) encodes 17 genes involved in purine/pyrimidine nucleotide salvage/catabolic pathways. This cluster includes a putative uracil/xanthine permease, a purine/pyrimidine phosphoribosyltransferase, adenine deaminase, xanthine dehydrogenase, uricase, and allantoicase.

Other catabolic gene clusters include pathways for the utilization of aromatic compounds associated with degradation of plant material and the mineralization of lignin. These include a 15-kb gene cluster encoding proteins involved in a pathway for conversion of hydroxyphenylpyruvate and 4-hydroxybenzoate via protocatechuate and β-ketoadipate to tricarboxylic acid cycle intermediates (*pcaB-pcaF*). A gene cluster (*paaGZEDBAX*) similar in sequence and gene order to what is believed to be a multicomponent oxygenase involved in phenylacetic acid catabolism in *E. coli* (55) was also identified. There are four genes encoding proteins with homology to inositol monophosphatases (SMb20150, SMb20159, SMb20362, and SMb21225). At least one of these is believed to be involved in inositol utilization, whereas others may be involved in utilization of rhizopines or possible degradation of the plant-derived phosphate storage compound, phytic acid (myoinositol hexaphosphate).

All members of the *Rhizobiaceae*, including strain Rm1021, have been reported to be able to use phosphonates as a source of phosphate (56); several gene clusters involved in phosphonate degradation are located on pSymB (*phnA, phnGHIJKL, phnM*, and *phoCDET*). In a different strain of *S. meliloti*, the *pta* and *ackA* genes encoding phosphotransacetylase and acetate kinase are part of a single operon whose expression is induced on phosphate starvation (57). Interestingly, in strain Rm1021, we find that the putative *pta* and *ackA* genes are located over 50 kb apart on pSymB. The significance of the latter observation is unclear, and it remains to be determined whether there is considerable gene shuffling within *S. meliloti* strains.

Among many genes with high similarity to alcohol dehydrogenases, one gene cluster (SMb20169 to SMb20174) includes the methanol dehydrogenase structural gene and is clearly involved in methanol utilization. Methanol dehydrogenase requires the redox coenzyme pyrroloquinoline quinone (PQQ), and a complete PQQ biosynthesis gene cluster (*pqqA-E*) is also located on pSymB.

A CO_2-fixation gene cluster, *cbbR-cbbFPTALSX*, encoding the enzymes of the Calvin-Benson-Bassham (CBB) cycle, is located directly upstream of the *pqqABCDE* genes. Ribulose-1, 5-bisphosphate carboxylase (small and large subunits) and phosphoribulokinase are enzymes unique to the CBB cycle and are encoded by *cbbS*, *cbbL*, and *cbbP*. The *cbbR* gene encodes a LysR family transcriptional regulator, and this, together with the other *cbb* genes, is found in both photo- and chemoautrotrophic bacteria. To date, *Bradyrhizobium japonicum* is the only rhizobium that has been shown to grow chemolithoautotrophically by using CO_2 as a carbon source and H_2 as the electron donor (58). It remains to be established experimentally whether *S. meliloti* can grow as an autotroph and under what conditions the *cbb* genes are expressed. However, the close vicinity of the methanol utilization gene cluster to the *cbb* genes is interesting, as growth on methanol is a carbon-limiting condition that could lead to expression of the CO_2-fixation genes (59).

ABC and Other Transport Proteins. A prominent feature of the *S. meliloti* genome is the number of genes encoding for ABC transport systems. These systems contain an ATP-binding protein, one or two integral membrane proteins, and, in the case of uptake systems, a periplasmic solute-binding protein with a N-terminal export-signal sequence. These genes are generally arranged as an operon. Of the 430 ABC transport system genes predicted in the whole genome, 235, or over half, are located on the pSymB megaplasmid. This constitutes 17% of the total coding capacity of pSymB. Almost half of the 64 ABC-transporter systems (some with missing components) are predicted to transport sugars (28), including previously identified lactose and trehalose/maltose transporters (*thuREFGK*). Other predicted solutes include iron (4), amino acids (6), peptides and oligopeptides (6), spermidine/putrescine (2), sulfate (1), aliphatic sulfonate (1), phosphate (1), choline (1), glycerol-3-phosphate (1), rhizopine (1), and taurine (1). Although a number of genes show similarity to C_4-dicarboxylate transporters, they are unlikely to transport C_4-dicarboxylates because mutation of *dctA* (also located on pSymB) has been shown to abolish C_4-dicarboxylate transport in *S. meliloti* (18).

In addition to the ABC family of transporters, we have identified transporter proteins of the Major Facilitator Superfamily, including DctA, and possible nitrate (SMb20436), sulfate (SMb20070), and xanthine/uracil (SMb21281) permeases. Genes encoding transmembrane efflux proteins were identified, several of which appear to be involved in exporting toxic compounds from the bacteria (SMb20071, SMb20338, SMb20345, and SMb21575).

The requirement for specific transport systems in the production of normal N_2-fixing nodules has already been demonstrated for the *dct*, *bacA*, and *pho* loci, highlighting the important role of pSymB in the infection process and in endosymbiosis. The wealth of transport systems uncovered by sequence analysis suggests that pSymB may play a broader role in adaptation of the bacterium to the local environment, both as a free-living saprophyte and as an endosymbiont with host plants.

Transcriptional Regulators. We have identified 134 ORFs on pSymB as transcriptional regulators; this includes the response regulators from two component systems for which there are an additional 15 sensor histidine kinase-like proteins. Nineteen regulators belong to the LysR family that activate transcription in response to coinducers; 21 belong to the GntR family, 16 to the LacI/GalR family, and the remainder to the TetR, AraC, ArsR, AsnC, DeoR, MerR, and SorC families. On the basis of sequence homology and transcription direction/location, we have assigned gene symbols to a number of the LysR regulators (GstR, CbbR, PcaQ, and GbpR), because these activators are generally transcribed divergently from the genes they regulate (60). GntR proteins bind to promoter regions and negatively regulate transcription (61). We were unable to assign genotypes to any of the pSymB GntR-like genes, despite the fact that gene clusters regulated by several GntR-like genes were apparent (e.g., SMb20106, SMb20129, and SMb20441).

In addition to the above regulators, we have identified four *rpoE*-like genes encoding alternative σ factors related to the extracytoplasmic function subfamily of eubacterial RNA polymerase σ factors (SMb21484, SMb20592, SMb20531, and SMb20030) (62).

Nodulation and Nitrogen Metabolism. It was known that genes involved in nodulation and in N_2-fixation are mainly localized on the pSymA replicon. Only a few genes involved in these processes were found on pSymB. A previously identified and functionally interchangeable copy of *nodP1* and *nodQ1* is localized on pSymB (*nodP2*, *nodQ2*) (63). The SMb20472 protein is 65% identical to the NolO nodulation protein; however, this is likely a protein with carbamoyl transferase activity rather than a protein with a direct role in nodulation. Another gene annotated as *nfeD* is involved in nodulation competitiveness (64).

Several genes involved in nitrate/nitrite reduction were identified; these include a potential nitrate/nitrite response regulator (SMb20077-SMb20078), a periplasmic nitrate reductase (SMb20997) similar to NnuR, and a nitrate transporter (SMb21114). A nitrate/nitrite reductase and a siroheme synthase for nitrate assimilation are encoded by a four-gene cluster (SMb20987-SMb20984) (65). One of three glutamine synthetase structural genes is located on pSymB. Transcription of *glnII* requires the RpoN σ factor and the NtrC transcriptional activator, both of which are encoded by chromosomal genes (66).

Protective Response and Antibiotic Resistance. A number of genes whose products could be involved in detoxifying reactions were located on pSymB. These include two nonheme haloperoxidases possibly involved in dehalogenation reactions (SMb20054 and SMb20860) (67) and two glutathione *S*-transferase genes (SMb20005, SMb21149), of which there are a total of 16 in the genome. There are several genes that may be involved in antibiotic resistance. Three genes, SMb20345, SMb20346, and SMb20698, encode multidrug efflux permeases. The genes *acrE* (SMb21497) and *acrF* (SMb21498) encode putative acriflavin resistance proteins, which also have homology to other transmembrane multidrug efflux proteins (68). AmpC is a putative β-lactamase, and AacC4 (SMb21552) is a putative aminoglycoside 6′-*N*-acetyltransferase rendering resistance to amikacin. The hypothetical gene SMb21154 codes for a protein that belongs to the bleomycin resistance protein family.

There are several genes that may play a role in responding to osmotic stress protection, including two trehalose synthases (SMb20099 and SMb20574) (there is a third on pSymA) and a previously reported trehalose/maltose transport gene cluster. These may play a role in generation of the osmolyte trehalose in *S. meliloti*. A metallo-regulatory gene of the *merR* family, *hmrR2*, is located next to the previously identified *atcU2* gene, whose product appears to be a copper export ATPase rendering resistance to copper. A similar gene pair is also located on the pSymA replicon.

Eight genes are assigned functions in DNA modification or degradation (vs. 57 on the chromosome and 10 on pSymA), 4 DNA ligases (SMb20008, SMb206868, SMb20912, and SMb21044), a DNA topoisomerase I (SMb21445), a methylated-DNA-protein-cysteine methyltransferase (SMb20708), a 3-methyladenine DNA glycosylase (SMb20709), and an exodeoxyribonuclease III (*xthA4*).

Dehydrogenases, Oxidoreductases, and Sugar Kinases. A large number of genes encoding proteins potentially involved in oxidative metabolism were located on pSymB. The numbers of pSymB genes predicted to encode dehydrogenases (67), oxidoreductases (41), and dehydratases (19) are similar to the proportions of these genes predicted for the chromosome and pSymA. However, 10 predicted sugar kinase genes were located on pSymB; this contrasts the single predicted sugar kinase on pSymA and the 11 chromosomal genes. Several of the pSymB sugar kinase genes (SMb20852, SMb21217, and SMb21373) appear to be part of sugar catabolic gene clusters that include ABC transport genes.

Conclusions

The 1,683,333 bp size of the pSymB replicon is similar to that predicted from previous genetic and restriction analyses (4, 20). Interestingly, the size of this replicon is comparable to the entire genomes of *Hemophilus influenzae* (1.8 Mb) and *Methanococcus jannaschii* (1.66 Mb) (69, 70). Our annotation revealed that the gene density of pSymB is similar to the *S. meliloti* chromosome and to the density of other bacteria genomes (1 ORF/1.1 kb). Moreover, with few exceptions (see below), we did not find evidence of single genes or gene clusters carrying nonsense or other mutations suggestive of regions on their way to being eliminated from the genome. This is interesting, as our previous deletion studies revealed that much of the pSymB replicon is dispensable for growth in minimal medium in the laboratory. We identified two gene regions, coding for the arg-tRNA gene and the *minCDE* genes that are likely to be essential for growth of the bacteria. The arg-tRNA gene lies within a region of pSymB that could not be deleted in previous experiments (21).

Two major observations to emerge from an analysis of the annotated pSymB sequence are the large number of solute transport systems and genes involved in polysaccharide synthesis. In addition to these, we have observed many genes that have potential catabolic activities such as alcohol dehydrogenases. Thus, it appears that pSymB endows the bacteria with the ability to take up and presumably oxidize many different compounds from the soil environment. The presence of pathways for the oxidation of methanol and plant degradation products such as protocatechuate is consistent with this hypothesis. Although it is noticeable that there are few if any pSymB genes that play a direct role in nodulation and symbiotic N_2-fixation, we note that pSymB does play a role in adaptation to the endosymbiotic lifestyle, as emphasized by the fact that mutations in the *exo*, *dct*, *pho*, and *bacA* genes abolish symbiotic nitrogen fixation. Additionally, pSymB codes for several detoxification and antibiotic resistance functions. Hence, we envisage pSymB as playing an important role in the survival of the bacterium under the presumably diverse nutritional living conditions encountered in the soil and rhizosphere. The increased accessibility to carbon sources and increased surface variability may point to the

particular importance of pSymB in enhancement of competitive abilities of *S. meliloti* in the natural habitat.

The sequence of the pSymB replicon revealed that *S. meliloti* carries many more putative polysaccharide synthesis genes than previously envisioned. This surprising wealth of genes encoding cell surface polysaccharides may reflect the very different conditions (such as desiccation and starvation) and environments *S. meliloti* has had to adapt to, e.g., soil, rhizosphere, and legume nodule. Additionally, these polysaccharides may be important for root nodule invasion, as shown for the previously characterized EPS synthesis gene clusters. The actual polysaccharides synthesized by most of the newly identified gene clusters are still to be discovered. It will be of interest to determine the conditions under which these gene clusters are transcribed and how they cooperate functionally. It has been shown that some key components of the surface carbohydrate export machinery can be involved in the export of more than one polysaccharide structure (71, 72). The absence of key genes in some of the new clusters encoding Wzy-dependent export machineries may indicate that this phenomenon also occurs in *S. meliloti*.

Almost all of the *S. meliloti* genes required for cell growth and viability are located on the 3.6-Mb chromosome (6). However, the presence on pSymB of the single essential genomic copy of the transfer RNA gene, ^{Arg}tRNA $_{CCG}$, together with other possibly essential genes such as *minCDE* and *asn*, clearly suggests that pSymB is indispensable to the cell and hence can justifiably be viewed as a second chromosome. We interpret the biased distribution of rhizobium-specific intergenic mosaic elements and A, B, C palindromic elements (34, 35) on the three replicons together with the G + C content of the replicons as evidence that pSymB was acquired by an ancestral *S. meliloti* before pSymA (5). This is reminiscent of the observations concerning the two chromosomes of *Vibrio* species, where the 2.9-Mb Chromosome 1 carries most genes required for cell growth and viability, and the 1-Mb Chromosome 2 also carries a few essential genes (73, 74).

Although the pSymB nucleotide sequence has revealed a wealth of new genes, for the most part the precise biological functions of these genes remain to be determined. The identification of these functions will lead to a clearer understanding of the interaction of *S. meliloti* with plants and, more generally, how this bacteria lives and survives in the soil environment.

We thank F. Barloy-Hubler, D. Capela, and F. Galibert for providing the pSymB BAC clones, H. Voss and B. Remmel from Lion Bioscience AG Heidelberg (Germany), and W. Arnold, W. Engelhardt, and F. Hecht from IIT Biotech GmbH Bielefeld. Work at McMaster University was supported by Natural Sciences and Engineering Research Council (Canada) Research, Strategic, and Genomics grants to T.M.F. and B.G. Work at Bielefeld was supported by grants from the Bundesministerium für Forschung und Technologie (0311752) to A.P.

1. Jumas-Bilak, E., Michaux-Charachon, S., Bourg, G., Ramuz, M. & Allardet-Servent, A. (1998) *J. Bacteriol.* **180,** 2749–2755.
2. Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L. & Ramuz, M. (1993) *J. Bacteriol.* **175,** 7869–7874.
3. Meade, H. M. & Singer, E. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 2076–2078.
4. Honeycutt, R. J., McClelland, M. & Sobral, B. W. (1993) *J. Bacteriol.* **175,** 6945–6952.
5. 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.
6. 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)
7. Barnett, M. J., Fisher, R. F., Jones, T., Komp, C., Abola, A. P., Barloy-Hubler, F., Bowser, L., Capela, D., Galibert, F., Gouzy, J., *et al*. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 9883–9888. (First Published July 31, 2001; 10.1073/pnas.161294798)
8. Banfalvi, Z., Sakanyan, V., Koncz, C., Kiss, A., Dusha, I. & Kondorosi, A. (1981) *Mol. Gen. Genet.* **184,** 318–325.
9. Rosenberg, C., Boistard, P., Dénarié, J. & Casse-Delbart, F. (1981) *Mol. Gen. Genet.* **184,** 326–333.
10. Finan, T. M., Kunkel, B., De Vos, G. F. & Signer, E. R. (1986) *J. Bacteriol*. **167,** 66–72.
11. Hynes, M. F., Simon, R., Müller, P., Niehaus, K., Labes, M. & Pühler, A. (1986) *Mol. Gen. Genet.* **202,** 356–362.
12. Müller, P., Keller, M., Weng, W. M., Quandt, J., Arnold, W. & Pühler, A. (1993) *Mol. Plant–Microbe Interact.* **6,** 55–65.
13. Glazebrook, J. & Walker, G. C. (1989) *Cell* **56,** 661–672.
14. Becker, A., Rüberg, S., Küster, H., Roxlau, A. A., Keller, M., Ivashina, T., Cheng, H. P., Walker, G. C. & Pühler, A. (1997) *J. Bacteriol.* **179,** 1375–1384.
15. Bardin, S., Dan, S., Østeras, M. & Finan, T. M. (1996) *J. Bacteriol.* **178,** 4540–4547.
16. Finan, T. M., Oresnik, I. & Bottacin, A. (1988) *J. Bacteriol.* **170,** 3396–3403.
17. Glazebrook, J., Ichige, A. & Walker, G. C. (1993) *Genes Dev.* **7,** 1485–1497.
18. Watson, R. J., Chan, Y. K., Wheatcroft, R., Yang, A. F. & Han, S. H. (1988) *J. Bacteriol.* **170,** 927–934.
19. Rosenberg, C., Casse-Delbart, F., Dusha, I., David, M. & Boucher, C. (1982) *J. Bacteriol.* **150,** 402–406.
20. Charles, T. C. & Finan, T. M. (1990) *J. Bacteriol.* **172,** 2469–2476.
21. Charles, T. C. & Finan, T. M. (1991) *Genetics* **127,** 5–20.
22. Vincent, J. M. (1941) *Proc. Linn. Soc. N.S.W.* **66,** 145–154.
23. Chain, P. S., Hernandez-Lucas, I., Golding, B. & Finan, T. M. (2000) *J. Bacteriol.* **182,** 5486–5494.
24. Barloy-Hubler F., Capela, D., Batut, J. & Galibert, F. (2000) *Curr. Microbiol.* **41,** 109–113.
25. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8,** 175–185.
26. Ewing, B. & Green, P. (1998) *Genome Res.* **8,** 186–194.
27. Staden, R. (1996) *Mol. Biotechnol.* **5,** 233–241.
28. Haas, S., Vingron, M., Poustka, A. & Wiemann, S. (1998) *Nucleic Acids Res.* **26,** 3006–3012.
29. Schiex, T., Thébault, P. & Kahn, D. (2000) *Proceedings of the JOBIM Conference, Montpellier, France*, pp. 321–328.
30. Sonnhammer, E. L. & Kahn, D. (1994) *Protein Sci.* **3,** 482–492.
31. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27,** 4636–4641.
32. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
33. Riley, M. & Labedan, B. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, eds. Neidhardt, F., Curtiss, R., III, Lin, E. C. C., Ingraham, J., Low, K. B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. & Umbarger, H. E. (Am. Soc. Microbiol. Press, Washington, DC), 2nd Ed., pp. 2118–2202.
34. Østeras, M., Stanley, J. & Finan, T. M. (1995) *J. Bacteriol.* **177,** 5485–5494.
35. Østeras, M., Boncompagni, E., Vincent, N., Poggi, M. C. & Le Rudulier, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11394–11399.
36. de Boer, P. A., Crossley, R. E. & Rothfield, L. I. (1989) *Cell* **56,** 641–649.
37. Bardwell, J. C. & Craig, E. A. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 5177–5181.
38. Schnaitman, C. A. & Klena, J. D. (1993) *Microbiol. Rev.* **57,** 655–682.
39. Whitfield, C. & Roberts, I. S. (1999) *Mol. Microbiol.* **31,** 1307–1319.
40. Coplin, D. L. & Cook, D. (1990) *Mol. Plant–Microbe Interact.* **3,** 271–279.
41. Stevenson, G., Andrianopoulos, K., Hobbs, M. & Reeves, P. R. (1996) *J. Bacteriol.* **178,** 4885–4893.
42. Kennedy, E. P. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, eds. Neidhardt, F., Curtiss, R., III, Lin, E. C. C., Ingraham, J., Low, K. B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. & Umbarger, H. E. (Am. Soc. Microbiol. Press, Washington, DC), 2nd Ed., pp. 1064–1070.
43. Becker, A., Kleickmann, A., Keller, M., Arnold, W. & Pühler, A. (1993) *Mol. Gen. Genet.* **241,** 367–379.
44. Glucksmann, M. A., Reuber, T. L. & Walker, G. C. (1993) *J. Bacteriol.* **175,** 7045–7055.
45. Kereszt, A., Kiss, E., Reuhs, B. L., Carlson, R. W., Kondorosi. S. & Putnoky, P. (1998) *J. Bacteriol.* **180,** 5426–5431.
46. Reuhs, B. L., Williams, M. N., Kim, J. S., Carlson, R. W. & Cote, F. (1995) *J. Bacteriol.* **177,** 4289–4296.
47. Zhou, Z., White, K. A., Polissi, A., Georgopoulos, C. & Raetz, C. R. (1998) *J. Biol. Chem.* **273,** 12466–12475.
48. Stanfield, S. W., Ielpi, L., O'Brochta, D., Helinski, D. R. & Ditta, G. S. (1988) *J. Bacteriol.* **170,** 3523–3530.
49. Streit, W. R., Joseph, C. M. & Phillips, D. A. (1996) *Mol. Plant–Microbe Interact.* **9,** 330–338.
50. Göttfert, M., Rothlisberger, S., Kundig, C., Beck, C., Marty, R. & Hennecke, H. (2001) *J. Bacteriol.* **183,** 1405–1412.
51. Sullivan, J. T. & Ronson, C. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5145–5149.
52. Aneja, P. & Charles, T. C. (1999) *J. Bacteriol.* **181,** 849–857.
53. Gage, D. J. & Long, S. R. (1998) *J. Bacteriol.* **180,** 5739–5748.
54. Galbraith, M. P., Feng, S. F., Borneman, J., Triplett, E. W., de Bruijn, F. J. & Rossbach, S. (1998) *Microbiology* **144,** 2915–2924.
55. Ferrández, A., Miñambres, B., García, B., Olivera, E. R., Luengo, J. M., García, J. L. & Díaz, E. (1998) *J. Biol. Chem.* **273,** 25974–25986.
56. Liu, C.-M., McLean, P., Sookdeo, C. & Cannon, F. (1991) *Appl. Environ. Microbiol.* **57,** 1799–1804.
57. Summers, M. L., Denton, M. C. & McDermott, T. R. (1999) *J. Bacteriol.* **181**, 2217–2224.
58. Lepo, J. E., Hanus, F. J. & Evans, H. J. (1980) *J. Bacteriol.* **141,** 664–670.
59. Shively, J. M., van Keulen, G. & Meijer, W. G. (1998) *Annu. Rev. Microbiol.* **52,** 191–230.
60. Schell, M. A. (1993) *Annu. Rev. Microbiol.* **47,** 597–626.
61. Haydon, D. J. & Guest, J. R. (1991) *FEMS Microbiol Lett.* **63,** 291–295.
62. Missiakas, D. & Raina, S. (1998) *Mol. Microbiol.* **6,** 1059–1066.
63. Schwedock, J. S. & Long, S. R. (1992) *Genetics* **132,** 899–909.
64. Garcia-Rodriguez, F. M. & Toro, N. (2000) *Mol. Plant–Microbe Interact.* **13,** 583–591.
65. Lin, J. T., Goldman, B. S. & Stewart, V. (1993) *J. Bacteriol.* **175,** 2370–2378.
66. Shatters, R. G., Somerville, J. E. & Kahn, M. L. (1989) *J. Bacteriol.* **171,** 5087–5094.
67. van Pée, K.-H. (1996) *Annu. Rev. Microbiol.* **50,** 375–399.
68. Okusu, H., Ma, D. & Nikaido, H. (1996) *J. Bacteriol.* **178,** 306–308.
69. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J., Dougherty, B. A., Merrick, J. M., *et al*. (1995) *Science* **269,** 496–512.
70. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al*. (1996) *Science* **273,** 1058–1073.
71. Whitfield, C., Amor, P. A. & Köplin, R. (1997) *Mol. Microbiol.* **23,** 629–638.
72. Feldman, M. F., Marolda, C. L., Monteiro, M. A., Perry, M. B., Parodi, A. J. & Valvano, M. A. (1999) *J. Biol. Chem.* **274,** 35129–35138.
73. 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.
74. Yamaichi, Y., Iida, T., Park, K. S., Yamamoto, K. & Honda, T. (1999) *Mol. Microbiol.* **31,** 1513–1521.