JOURNAL OF BACTERIOLOGY, Nov. 1997, p. 7135–7155
0021-9193/97/$04.00+0
Copyright © 1997, American Society for Microbiology

Vol. 179, No. 22

Complete Genome Sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional Analysis and Comparative Genomics

DOUGLAS R. SMITH,[1*] LYNN A. DOUCETTE-STAMM,[1] CRAIG DELOUGHERY,[1] HONGMEI LEE,[1] JOANN DUBOIS,[1] TYLER ALDREDGE,[1] ROMINA BASHIRZADEH,[1] DERRON BLAKELY,[1] ROBIN COOK,[1] KATIE GILBERT,[1] DAWN HARRISON,[1] LIEU HOANG,[1] PAMELA KEAGLE,[1] WENDY LUMM,[1] BRYAN POTHIER,[1] DAYONG QIU,[1] ROB SPADAFORA,[1] RITA VICAIRE,[1] YING WANG,[1] JAMEY WIERZBOWSKI,[1] RENE GIBSON,[1] NILOFER JIWANI,[1] ANTHONY CARUSO,[1] DAVID BUSH,[1] HERSHEL SAFER,[1] DONIVAN PATWELL,[1] SHASHI PRABHAKAR,[1] STEVE MCDOUGALL,[1] GEORGE SHIMER,[1] ANIL GOYAL,[1] SHMUEL PIETROKOVSKI,[2] GEORGE M. CHURCH,[3] CHARLES J. DANIELS,[4] JEN-I MAO,[1] PHIL RICE,[1] JÖRK NÖLLING,[1] AND JOHN N. REEVE[4]

Genome Therapeutics Corporation, Collaborative Research Division, Waltham, Massachusetts 02154,[1] Howard Hughes Medical Institute, Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115,[3] Fred Hutchinson Cancer Research Center, Seattle, Washington 98109,[2] and Department of Microbiology, The Ohio State University, Columbus, Ohio 43210[4]

Received 2 July 1997/Accepted 3 September 1997

The complete 1,751,377-bp sequence of the genome of the thermophilic archaeon *Methanobacterium thermoautotrophicum* ΔH has been determined by a whole-genome shotgun sequencing approach. A total of 1,855 open reading frames (ORFs) have been identified that appear to encode polypeptides, 844 (46%) of which have been assigned putative functions based on their similarities to database sequences with assigned functions. A total of 514 (28%) of the ORF-encoded polypeptides are related to sequences with unknown functions, and 496 (27%) have little or no homology to sequences in public databases. Comparisons with *Eucarya*-, *Bacteria*-, and *Archaea*-specific databases reveal that 1,013 of the putative gene products (54%) are most similar to polypeptide sequences described previously for other organisms in the domain *Archaea*. Comparisons with the *Methanococcus jannaschii* genome data underline the extensive divergence that has occurred between these two methanogens; only 352 (19%) of *M. thermoautotrophicum* ORFs encode sequences that are >50% identical to *M. jannaschii* polypeptides, and there is little conservation in the relative locations of orthologous genes. When the *M. thermoautotrophicum* ORFs are compared to sequences from only the eucaryal and bacterial domains, 786 (42%) are more similar to bacterial sequences and 241 (13%) are more similar to eucaryal sequences. The bacterial domain-like gene products include the majority of those predicted to be involved in cofactor and small molecule biosyntheses, intermediary metabolism, transport, nitrogen fixation, regulatory functions, and interactions with the environment. Most proteins predicted to be involved in DNA metabolism, transcription, and translation are more similar to eucaryal sequences. Gene structure and organization have features that are typical of the *Bacteria*, including genes that encode polypeptides closely related to eucaryal proteins. There are 24 polypeptides that could form two-component sensor kinase-response regulator systems and homologs of the bacterial Hsp70-response proteins DnaK and DnaJ, which are notably absent in *M. jannaschii*. DNA replication initiation and chromosome packaging in *M. thermoautotrophicum* are predicted to have eucaryal features, based on the presence of two Cdc6 homologs and three histones; however, the presence of an *ftsZ* gene indicates a bacterial type of cell division initiation. The DNA polymerases include an X-family repair type and an unusual archaeal B type formed by two separate polypeptides. The DNA-dependent RNA polymerase (RNAP) subunits A′, A″, B′, B″ and H are encoded in a typical archaeal RNAP operon, although a second A′ subunit-encoding gene is present at a remote location. There are two rRNA operons, and 39 tRNA genes are dispersed around the genome, although most of these occur in clusters. Three of the tRNA genes have introns, including the tRNAPro (GGG) gene, which contains a second intron at an unprecedented location. There is no selenocysteinyl-tRNA gene nor evidence for classically organized IS elements, prophages, or plasmids. The genome contains one intein and two extended repeats (3.6 and 8.6 kb) that are members of a family with 18 representatives in the *M. jannaschii* genome.

Methanobacterium thermoautotrophicum ΔH, isolated in 1971 from sewage sludge in Urbana, Ill. (72), is a lithoautotrophic, thermophilic archaeon that grows at temperatures ranging from 40 to 70°C and optimally at 65°C. *M. thermoautotrophicum* conserves energy by using H_2 to reduce CO_2 to CH_4 and synthesizes all of its cellular components from these same gaseous substrates plus N_2 or NH_4^+ and inorganic salts, but despite this impressive biosynthetic capacity, *M. thermoautotrophicum* ΔH and related strains have very small genomes (~1.7 ± 0.2 Mb [57, 58]). *M. thermoautotrophicum* ΔH, Marburg, and Winter are the foci of many methanogenesis, archaeal physiology, and molecular biology investigations, and *M. thermoautotrophicum* ΔH was chosen as a representative of this group for genome sequencing. These thermophilic methanogens have mesophilic and hyperthermophilic relatives, *Methanobacterium formicicum* and *Methanothermus fervidus*, respectively, so that comparisons can be made of homologous

[*] Corresponding author. Mailing address: Genome Therapeutics Corporation, Collaborative Research Division, 100 Beaver St., Waltham, MA 02154. Phone: (617) 398-2378. Fax: 1-617-893-9535. E-mail: doug.smith@genomecorp.com.

genes and gene products in these closely related species, which grow at temperatures ranging from 30 to 90°C. In addition, the complete genome sequence is available from the distantly related methanogen *Methanococcus jannaschii* (9) so that comparisons could also be made of all genes and their genome organizations in two organisms in the domain *Archaea*. Here we report the sequence of the *M. thermoautotrophicum* ΔH genome, identify and annotate genes and gene functions, and provide an initial comparison with the *M. jannaschii* genome.

MATERIALS AND METHODS

Construction and isolation of small-insert libraries in multiplex sequencing vectors. DNA, isolated from *M. thermoautotrophicum* ΔH as previously described (66), was nebulized to a median size of 2 kb (5). These fragments were concentrated, and molecules in the 2- to 2.5-kb size range were purified by electrophoresis through 1% agarose gels followed by the GeneClean procedure (Bio 101, Inc., La Jolla, Calif.). Single-stranded ends were filled by using T4 DNA polymerase, and the DNA molecules were then ligated with a 100- to 1,000-fold molar excess of *Bst*XI-linker adapters with the sequences 5′GTCTTCACCACG GGG and 5′GTGGTGAAGAC. When *Bst*XI digested, these adapters are complementary to *Bst*XI-cleaved pMPX vectors (11) but are not self-complementary. Linker-adapted DNA molecules were separated from unincorporated linkers by electrophoresis through 1% agarose gels and ligated, in separate reaction mixtures, to 20 different pMPX vectors to generate 20 small-insert libraries. The pMPX vectors contain an out-of-frame *lacZ* gene which becomes in-frame if an adapter-dimer is cloned, and such clones, recognized as blue colony formers on X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside)-containing plates, were removed from the analysis (10).

The 20 pMPX libraries were transformed into *Escherichia coli* DH5α, and dilutions of the transformed cell suspensions were plated and incubated overnight at 37°C on Luria-Bertani plates that contained 200 μg of either ampicillin or methicillin/ml, IPTG (isopropyl-β-D-thiogalactopyranoside), and X-Gal. One clone from each of the 20 libraries was inoculated into the same 40 ml of L broth. Following incubation overnight at 37°C, plasmid DNA preparations (~100 μg) were isolated from these mixed cultures by using midi-prep kits and Tip-100 columns (Qiagen, Inc., Chatsworth, Calif.) and were stored in the wells of microtiter plates. Sufficient pMPX clones were collected for 5- to 10-fold genome coverage assuming an average sequence read-length of ~275 bp.

Small-insert sequencing. DNA sequences were obtained by using the multiplex sequencing procedure (10) with either chemical degradation (31 membranes) or Sequitherm (Epicenter Technologies, Madison, Wis.) dideoxy termination sequencing (37 membranes). The products of 24 sequencing reactions were separated by electrophoresis through 40-cm gels and transferred by electrophoresis directly onto nylon membranes (48). Following UV cross-linking, the membranes were hybridized with a ^{32}P-labeled oligonucleotide with a sequence complementary to a tag sequence of one of the pMPX vectors (10), washed, and used to expose autoradiograms. The probe was then removed by incubation at 65°C, and the hybridization cycle was repeated with a probe complementary to a different tag sequence. Membranes were first hybridized with a probe complementary to an internal control sequence added to each plasmid pool. Membranes were probed, stripped and reprobed up to 41 times.

Image processing, proofreading, and data storage. Digitized images of the autoradiograms, generated with a laser-scanning densitometer (Molecular Dynamics, Sunnyvale, Calif.), were processed on VaxStation 4000 computers by using REPLICA (11) and Xgel programs (Genome Therapeutics Corporation [GTC]) to obtain lane straightening, contrast adjustment, and resolution enhancement. Base cells made by REPLICA were displayed for visual confirmation before being stored in a project database. Multiple, independent sequence reads, covering the same region of the genome, provided the redundancy that facilitated and legitimized visual editing. Each sequence was assigned an identification number based on the microtiter plate, probe, gel, and gel lane, and all original data are retained in a permanent archive.

Construction of a large-insert cosmid library. A library of *M. thermoautotrophicum* DNA was constructed in the SuperCos1 cosmid vector (Stratagene, La Jolla, Calif.). Following *Xba*I digestion and dephosphorylation, SuperCos1 DNA was ligated overnight at 4°C with *M. thermoautotrophicum* DNA that had been partially digested with *Bam*HI to obtain fragments with lengths ranging from 35 to 45 kb. Ligation mixtures were packaged into lambda particles by using the Packagene system (Promega, Madison, Wis.), infected into *E. coli* XL1-blue, and plated on Luria-Bertani plates that contained 100 μg of ampicillin/ml (Stratagene). Ampicillin-resistant clones were inoculated into 10 ml of L broth supplemented with 100 μg of ampicillin/ml and incubated overnight at 37°C. Cosmid preparations were isolated from these cultures (50), and sequences from the ends of the cloned DNAs were obtained by using dideoxy chain-terminating technology (51) with primers complementary to the flanking T3 and T7 promoter sequences.

Sequence assembly and metacontig construction. At a statistical coverage of ~6.5-fold, the first assembly by using Phrap (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) with default parameters and without quality scores yielded 570 contigs. Random sequencing was continued until the statistical coverage was eightfold. To merge contigs, sequences at the ends of contigs were PCR amplified from the appropriate pMPX pool and sequenced directly by using primers chosen manually in GelAssemble (GA) (a GTC-modified version of the Genetics Computer Group Wisconsin package program [17]) or chosen automatically by Autoprimer (GTC), and short read-lengths at the ends of contigs were extended to ~500 nucleotides by resequencing.

As more sequence was accumulated, the Phrap assembly was repeated, yielding 321, 204, 160, and finally 90 contigs based on the statistical equivalent of ~eightfold genome coverage plus 685 walk and extension sequences. IncAsm (GTC), which employs a directed global alignment algorithm based on the position of a primer's parent fragment, was then used to insert sequences into the Phrap assembly. IncAsm searches a window of user-specified size to insert fragments into the alignment and adds insertions or deletions to the fragment or multi-alignment as necessary. CheckMates (GTC) identified pairs of contigs that contained the opposite ends of a single multiplex clone, and the linking regions were PCR amplified and sequenced from both ends by using dye terminator technology and ABI 377 machines. EndMatch, a program that uses FASTA alignments to compare contig ends and identify overlaps (GTC), identified contig pairs that could be merged in GA, which included some merges rejected initially by Phrap. CheckMates also prevented the misassembly of repetitive sequences by identifying the ends of each originating clone. Identical sequences that originated from clones with different ends were separated, and each was PCR amplified, by using unique flanking sequences, and resequenced to confirm their separate identities. At this point, 23 metacontigs (assemblies of the smaller contigs) remained without order or bridging information.

Metacontig assembly. Forty-six primers, with sequences complementary to sequences present at the ends of the 23 metacontigs, were combined into 47 mixtures. One mixture contained all 46 primers, and 46 mixtures each lacked one primer. PCRs were performed to amplify *M. thermoautotrophicum* genomic DNA, and the products obtained were separated by electrophoresis through 1% agarose gels. Comparing the products obtained with the complete mixture of primers with the products obtained with the mixtures lacking one primer identified products generated by that primer. By identifying two primers that generated the same product, and by knowing which metacontigs contained those primer sequences, metacontigs were ordered with respect to each other. The order was verified by using the primer pairs to PCR amplify the intervening region which was then sequenced. Primer pairs that yielded information were removed, and the combinatorial PCR procedure was repeated until 16 metacontigs remained.

All possible pairwise combinations of the 32 remaining primers were then used in PCRs to amplify *M. thermoautotrophicum* genomic DNA, and the amplified products were sequenced directly using ABI technology. This strategy, in some cases using primers complementary to different sequences at the ends of the metacontigs, closed all of the remaining physical gaps and resulted in a single circular contig.

Confirmation of the assembly and sequence summary. Sequences were obtained from the ends of cosmid inserts (see Fig. 1) to confirm the assembly. The program COVERAGE (GTC) was used to identify regions that had been sequenced in only one direction or by only one chemistry. These regions were resequenced, both in the complementary direction and by using ABI dye terminator chemistry as needed to resolve sequence anomalies. Primer pairs were also used to PCR amplify problematic regions, and sequencing the resulting products resolved almost all remaining uncertainties.

Overall, 36,935 sequence reads, 15,350 and 21,585 with chemical and dideoxy sequencing, respectively, were generated by MPX technology, resulting in a total of ~13.3 Mb with an average read-length of 361 nucleotides. An additional ~1.5 Mb of sequence was generated during the finishing process by 2,884 reads of ABI dye-terminated sequences. The final total of ~14.8 Mb of sequence corresponded to an ~8.5-fold statistical coverage of the *M. thermoautotrophicum* genome, with 97.5% of the genome confirmed by sequencing in both directions and an additional 2.2% confirmed by sequencing in the same direction but with an alternate chemistry (>99.7% of the total).

Sequence analysis and annotation. Contig sequences representing the entire genome were analyzed using GenomeBrowser tools (54) to identify all ORFs of >180 bp in length, compute dicodon usages, and automate BLASTP2 searches (1, 71). Gapped alignments were generated against all nonredundant protein (nrprotein) sequences in the SwissProt, PIR, and GenPept databases. Graphical views of the output were constructed which provided immediate access to HTML summaries of the BLAST output. The contig sequences were then joined in a text editor, and overlapping regions were removed. To facilitate ongoing GenomeBrowser analyses, the genome was evaluated as 10 nonoverlapping, artificially created contigs separated within noncoding regions.

Custom Perl scripts were used to filter the data generated by GenomeBrowser by using BLAST and dicodon usage scores to define potential gene sequences. The results were tabulated in an Excel spreadsheet with the direction of translation, start and stop codons, contig names, codon usage statistics, BLASTP2 similarity scores, *P* values, and database hit descriptions listed for each gene. Annotators reviewed the data and made corrections in GenomeBrowser, assigning product names, deleting spurious entries, and adding information not detected by the automated analyses.

ORF-encoded sequences were aligned with the sequences in the eight functionally annotated genomes in the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg). Functional categories, gene names, and enzyme commission numbers so assigned were imported into the Excel table and reevaluated with reference to the BLAST output before final assignments were made. All intergenic regions of >200 bp were researched against the nrprotein and GenBank databases to identify additional genes and conserved sequences. Start codons (ATG, GTG, and TTG) were putatively identified by their proximity to ribosome binding sequences (RBSs) (8, 53) and by compatibility with BLAST alignment data that minimized or eliminated overlaps. The BLIMPS multiple alignment program (19) was used to search the *M. thermoautotrophicum* protein sequences for inteins, class II DNA-mediated transposases, and homing endonucleases (44).

Overlapping ORFs, adjacent genes with hits to the same database sequence, and genes that were substantially shorter in length than their database homologs were routinely evaluated for frameshifts. The Bic_FrameSearch program (Compugen Bioccelerator, Petach-Tikva, Israel) (17) was used to generate gapped alignments of the *M. thermoautotrophicum* sequence with the corresponding database sequence to identify regions likely to contain errors. These were reinspected in GA, and most frameshifts were identified and resolved by manual editing. When necessary PCR amplification and product sequencing were also undertaken to evaluate potential frameshifts.

BLASTP2 and the parameters listed by Bult et al. (9) were used to compare gene families in *M. thermoautotrophicum* and *M. jannaschii*. Pairs of sequences with at least 30% identity over 50 amino acids were identified, and the resulting clusters were aligned by using Bic_Pileup (Compugen Bioccelerator) (17). These multi-alignments were examined to remove poorly aligned sequences and to separate well-aligned families that were tenuously joined by sequences with marginal homologies to one or both of the families.

The sequences of all *M. thermoautotrophicum* gene products were also aligned separately with only *M. jannaschii* sequences and with only the bacterial, eucaryal, and archaeal sequences (minus the *M. thermoautotrophicum* sequences) in the GenPept databases. These comparisons used Bic_SW, a fast implementation of the Smith-Waterman (SW) algorithm, and the data from the best alignment of each query sequence were tabulated. The fraction of query amino acids present in each alignment was calculated (query amino acids in alignment/total query amino acids), and the values so obtained were multiplied by this fraction to provide a normalized estimate of the identity (% ID) of each *M. thermoautotrophicum* sequence to each target sequence reported. These normalized values (SW %IDs) were used to rank sequences in the databases according to their overall identity to each *M. thermoautotrophicum* sequence. Raw SW %IDs, calculated from only the aligned regions of sequences, were not used for ranking comparisons.

Repetitive sequences were identified by Cross_Match, a fast SW algorithm (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) that compared all of the *M. thermoautotrophicum* contigs to each other. The program COMPOSITION (14) was used to count nucleotides and dinucleotides and to calculate %G+C contents, and the program tRNAscan was used to identify tRNA genes. A Perl script was used to generate a table with enzyme commission numbers which summarized the *M. thermoautotrophicum* genes present in pathways defined in the Ecocyc database (http://www.ai.sri.com/ecocyc/ecocyc.html). PerlTK programs (Genome_map and Gene_map [GTC]) were written to draw circular and linear genome maps (see Fig. 1 to 3), and graphical representations with annotated summaries (gene name, direction, position and putative function), similarities (SW %IDs), %G+C contents, and cosmid end sequences (based on FASTA alignments) were continuously generated and automatically updated.

Nucleotide sequence accession number. The sequence of the *M. thermoautotrophicum* ΔH genome has been deposited with GenBank under accession no. AE000666.

RESULTS

Nucleotide composition and codon usage. The genome of *M. thermoautotrophicum* ΔH was found to be a single, circular DNA molecule 1,751,377 bp in length (Fig. 1). Nucleotide 1 was assigned arbitrarily in a noncoding region upstream of a large cluster of genes, which included 31 ribosomal protein (r-protein)-encoding genes, all arranged in the same direction. Overall, the *M. thermoautotrophicum* genome is 49.5% G+C but several regions have higher G+C contents, including the rRNA and tRNA genes and several polypeptide-encoding regions dispersed around the genome (Fig. 1 and 2). More regions have lower G+C contents, some of which contain clusters of genes that have codon usages atypical for *M. thermoautotrophicum*, indicating regions that may have been acquired by lateral transfer (Fig. 1 and 2). One such region, at approximately nucleotide 49,000, is formed by two directly repeated copies of an ~8-kb sequence that has an ~40% G+C content. Together, these duplicated sequences contain >30 genes, including the adjacent genes MTH0067-MTH0068 and MTH0082-MTH0083, which encode polypeptides with sequences related to polypeptides in *M. jannaschii* that have motifs in common with transcription initiation factor TFIIIC and with a cell division protein (9).

The dinucleotide 5′CG and the CG-containing tetranucleotides 5′CGCG and 5′GCGC are substantially underrepresented in the genome of *M. thermoautotrophicum* ΔH, although as previously noted (34), 5′CTAG is even less common than these CG-containing tetranucleotides. The infrequent occurrence of 5′CTAG in microbial genomes has been previously reported (4, 25) and is proposed to result from the repair of G-T mismatches generated either by the spontaneous deamination of 5′ methyl-cytosine residues or by inaccurate recombination and/or replication. A mismatch repair mechanism could also be the basis for the 5′CTAG deficiency in *M. thermoautotrophicum*, although genes encoding mismatch-repair enzymes related to the Vsr systems thought to be responsible for the G-T mismatch repairs were not detected in the genome.

Genes and domain relationships. A total of 1,855 polypeptide-encoding genes and 47 stable RNA genes have been putatively identified in *M. thermoautotrophicum* (Fig. 3 and 4). Most ORFs (63%) have ATG translation initiating codons, although 22% are predicted to start with GTG and 15% are predicted to start with TTG. Of these putative polypeptide-encoding genes, 1,350 (73%) encode sequences with significant similarities to sequences in public databases (BLASTP2 scores against nrprotein databases of at least 100), 357 (19%) have limited similarity (BLASTP2 scores of 60 to 99), and 148 (8%) have no obvious database homologs (BLASTP2 scores of <60). In terms of function, 844 (46%) of the ORF-encoded sequences have been assigned putative functions based on their similarities to database sequences with assigned functions, 514 (28%) are classified as conserved, having similarities to database sequences with no assigned function (BLASTP2 scores of >100), and 496 (27%) are classified as unknown, having limited or no similarity to database sequences (BLASTP2 of <100). Sixteen ORFs that appear to result from frameshifts are not included in the list of putative genes.

Comparisons with databases that contain only archaeal, bacterial, and eucaryal sequences revealed that 1,013 (55%) of the *M. thermoautotrophicum* polypeptide sequences are most similar to previously documented archaeal sequences, 210 (11%) of which only have archaeal homologs. These include many of the enzymes directly involved in methanogenesis (see below); however, functions could not be assigned for 140 of these 210 archaeal-specific proteins. A total of 1,149 (62%) of the *M. thermoautotrophicum* ORF-encoded sequences have homologs in *M. jannaschii* with SW %IDs that are >30, although only 352 (19%) have SW %IDs of >50, and only 14 (<1%) have SW %IDs of >70. Most orthologous genes in the two methanogens have therefore undergone extensive divergence. When evaluated in terms of their similarities to bacterial versus eucaryal polypeptide sequences, 786 (42%) of the *M. thermoautotrophicum* ORF-encoded sequences are more similar to bacterial sequences and 241 (13%) are more similar to eucaryal sequences. Considering only the strongest matches within these groups, 490 (26%) of the *M. thermoautotrophicum* ORFs encode sequences with SW %IDs that are ≥ twofold higher with bacterial than with eucaryal sequences, whereas only 24 (1%) have SW %IDs that are ≥ twofold higher with eucaryal than with bacterial sequences. Most of the *M. thermoautotrophicum* proteins predicted to participate in cofactor and small molecule biosyntheses, intermediary metabolism, transport, nitrogen fixation, regulatory functions, and interactions with the environment have sequences that are more similar to bacterial

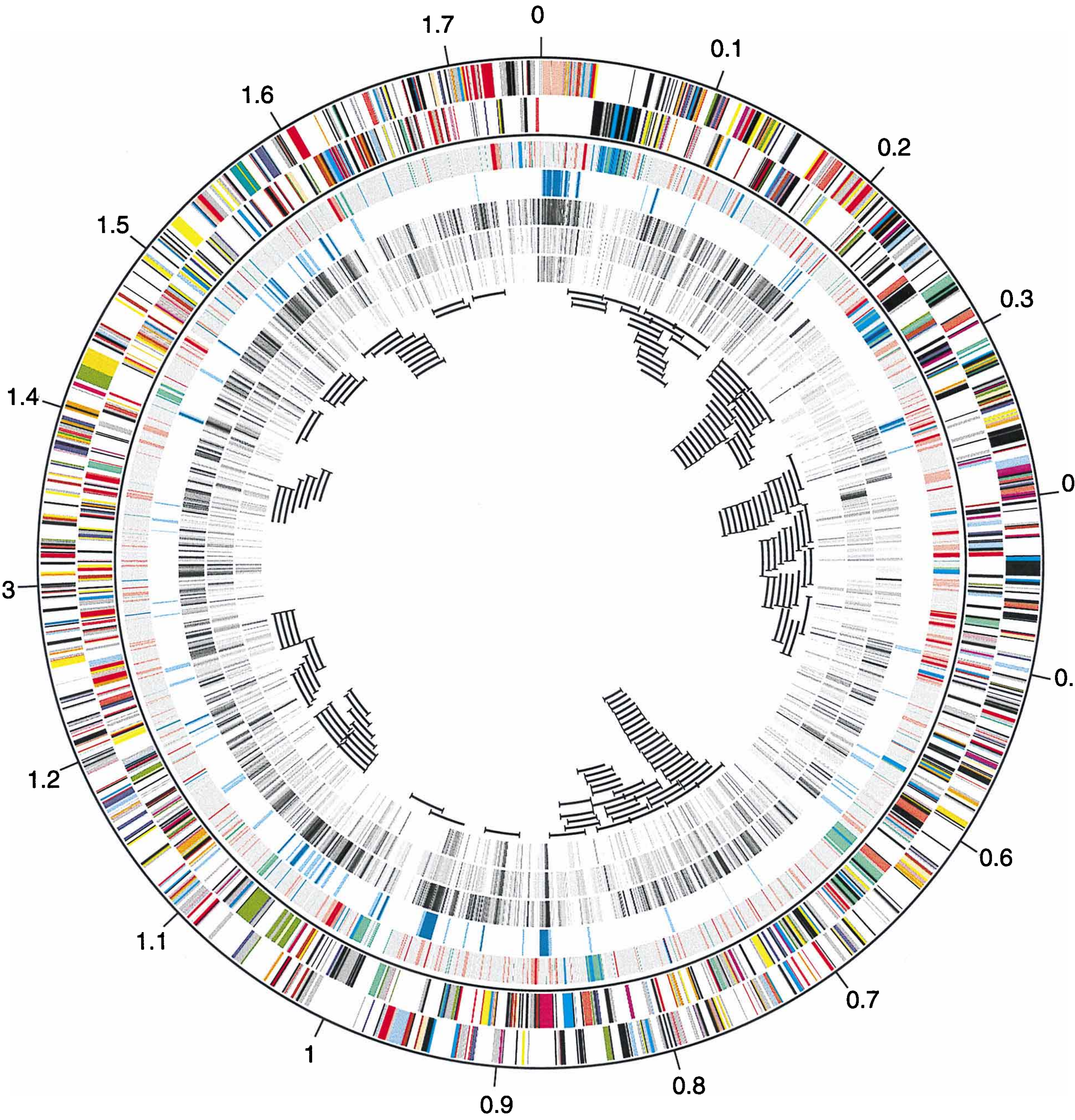

FIG. 1. Circular map of the *M. thermoautotrophicum* ΔH genome and summary of comparative analyses. The outer two rings flanked by dark lines show the positions of genes, color coded by function, on the forward and complementary strands, respectively. Moving inwards, the third ring displays the %G+C content of each putative gene (blue-violet, <32%; blue, 32 to 36%; turquoise, 36 to 41%; light green, 41 to 45%; gray, 45 to 54%; pink, 54 to 57%; red, >57%). The fourth ring identifies genes with conserved order in *M. jannaschii* (light blue, one neighbor conserved; dark blue, two neighbors conserved). The fifth ring displays SW %IDs for the best alignment of each gene product with polypeptides encoded in the *M. jannaschii* genome. The SW %IDs are mapped to a linear gray scale ranging from white to black for ID values of 20 to 86%, respectively. The sixth ring displays SW %IDs for the best alignment of each gene product with all bacterial polypeptides present in the GenPept database. The seventh ring displays SW %IDs for the best alignment of each gene product with all eucaryal polypeptides present in GenPept. The line segments arrayed around the center of the figure indicate the positions of cosmid clones; the tic marks at one or both ends of the segments indicate cosmid ends that were sequenced. The color code for functional categories is as follows: carbohydrate metabolism, sienna; methane metabolism, olive drab; carbon fixation, blue-green; oxidative phosphorylation and other energy metabolism, navajo white; sulfur metabolism, light yellow; nitrogen metabolism, gold; lipid metabolism, medium blue; nucleotide metabolism, orange; amino acid metabolism, yellow; vitamin and cofactor-related activities, light red; transcription and nucleoproteins, light blue; ribosomal proteins, pink; rRNA and tRNA metabolism and translation factors, red; DNA replication, cell division, and repair, light blue; DNA, RNA, and protein degradation, cyan; cell envelope, light green; transport, purple; general regulatory functions, magenta; other identifiable functions, lilac; conserved proteins, black; hypothetical proteins, gray.

sequences, whereas many of the *M. thermoautotrophicum* proteins predicted to be involved in DNA metabolism, transcription, and translation have sequences more similar to eucaryal than bacterial sequences. The similarities of each *M. thermoautotrophicum* sequence to *M. jannaschii*, eucaryal, and bacterial sequences are depicted in Fig. 1 and 2 by gray scales in which darkness corresponds to sequence similarity. The SW %ID values generated by the archaeal database comparisons

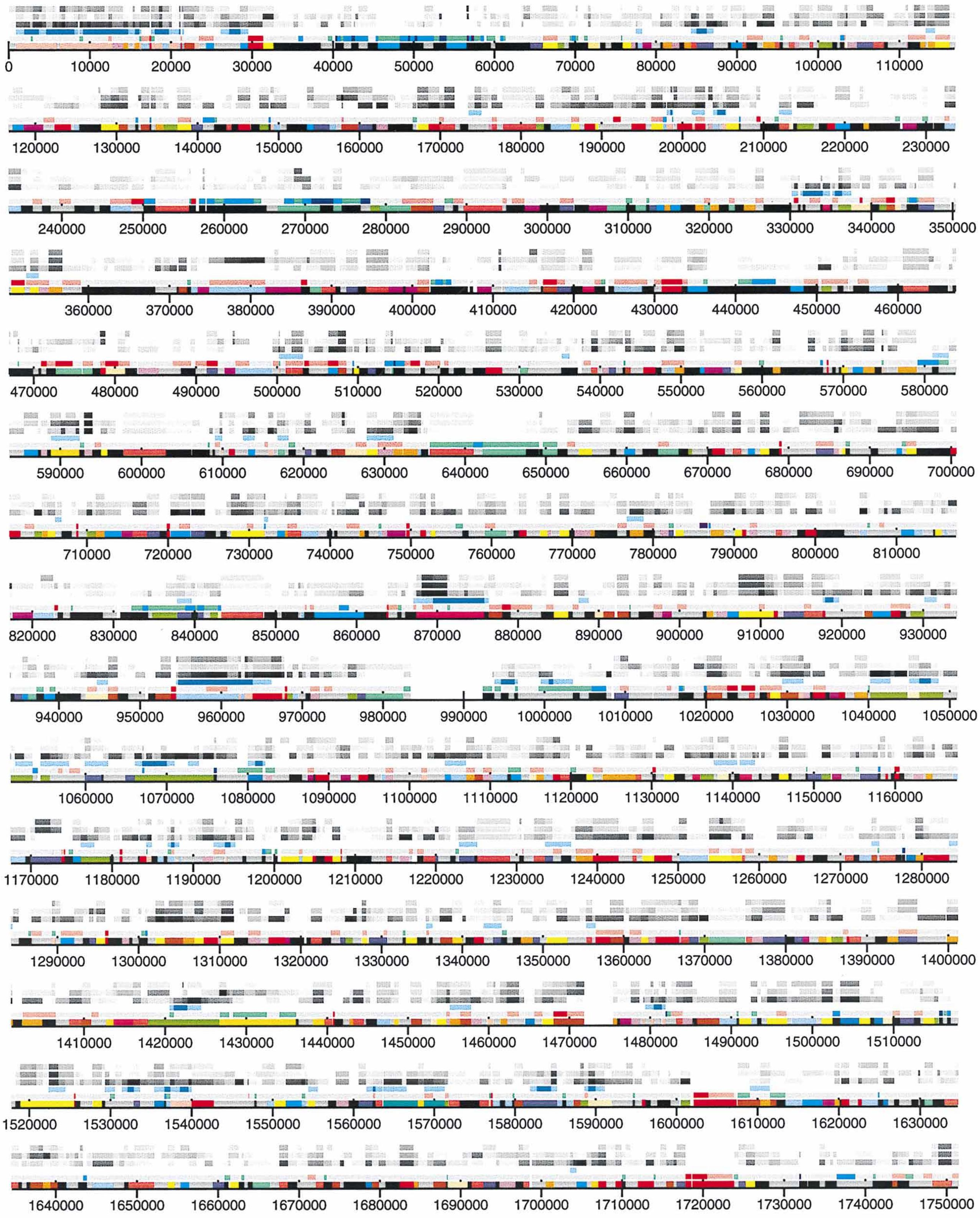

FIG. 2. Linear map of the *M. thermoautotrophicum* ΔH genome and summary of comparative analyses. This map is essentially an expanded, linear version of Fig. 1 that allows the results of comparative analyses associated with particular genes to be visualized more clearly. Individual genes are identified using the band order and colors corresponding to the rings and functional groups in Fig. 1 (see legend to Fig. 1 for a description), with the two coding strands and cosmid locations omitted.

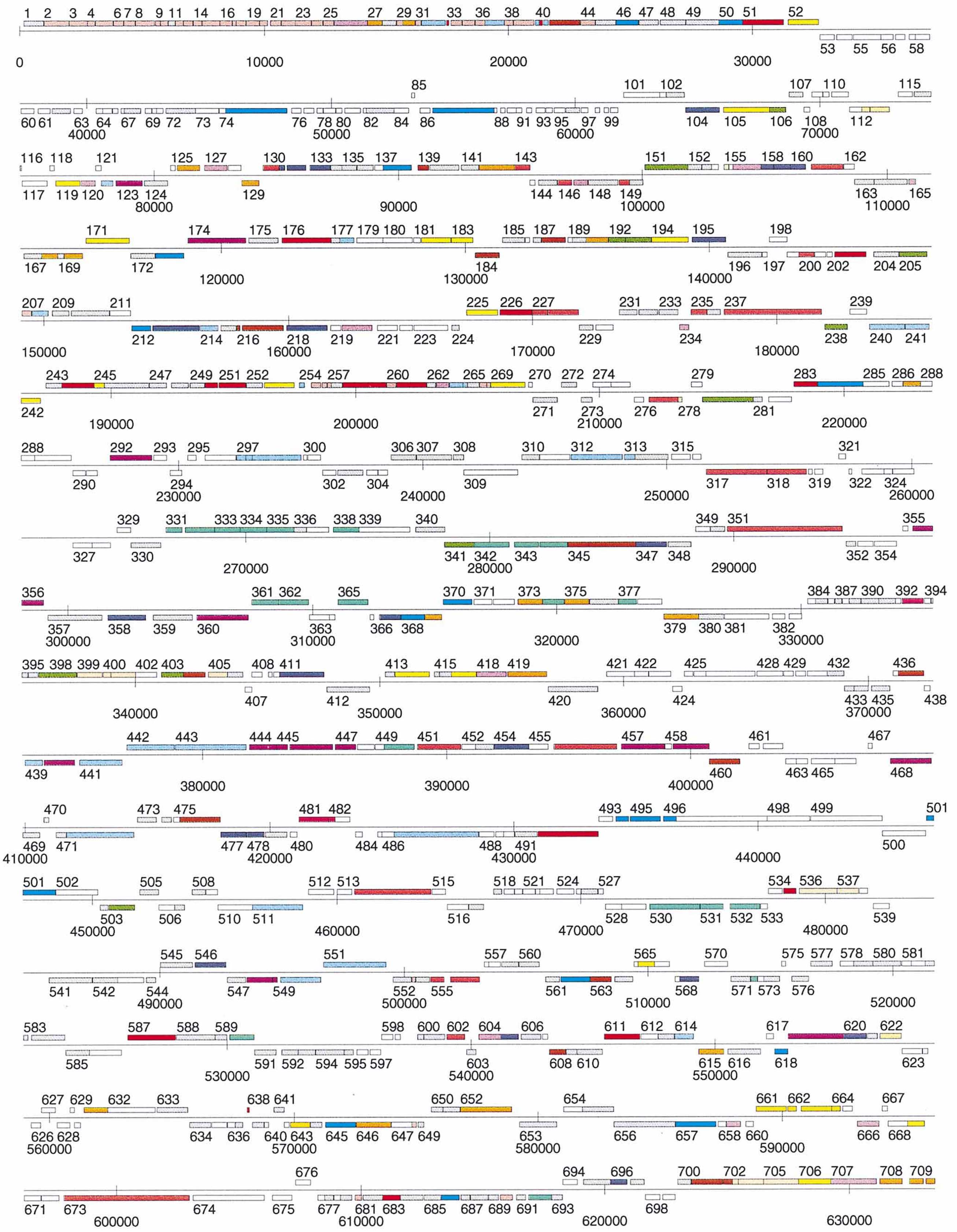
1 2 3 4 6 7 8 9 11 14 16 19 21 23 25 27 29 31 33 36 38 40 44 46 47 48 49 50 51 52
53 55 56 58
0 10000 20000 30000
85 101 102 107 110 115
60 61 63 64 67 69 72 73 74 76 78 80 82 84 86 88 91 93 95 97 99 104 105 106 108 112
40000 50000 60000 70000
116 118 121 125 127 130 133 135 137 139 141 143 151 152 155 158 160 162
117 119 120 123 124 129 144 146 148 149 163 165
80000 90000 100000 110000
171 174 175 176 177 179 180 181 183 185 187 189 192 194 195 198
167 169 172 184 196 197 200 202 204 205
120000 130000 140000
207 209 211 225 226 227 231 233 235 237 239
212 214 216 218 219 221 223 224 229 234 238 240 241
150000 160000 170000 180000
243 245 247 249 251 252 254 257 260 262 265 269 270 272 274 279 283 285 286 288
242 271 273 276 278 281
190000 200000 210000 220000
288 292 293 295 297 300 306 307 308 310 312 313 315 321
290 294 302 304 309 317 318 319 322 324
230000 240000 250000 260000
329 331 333 334 335 336 338 339 340 349 351 355
327 330 341 342 343 345 347 348 352 354
270000 280000 290000
356 361 362 365 370 371 373 375 377 384 387 390 392 394
357 358 359 360 363 366 368 379 380 381 382
300000 310000 320000 330000
395 398 399 400 402 403 405 408 411 413 415 418 419 421 422 425 428 429 432 436
407 412 420 424 433 435 438
340000 350000 360000 370000
442 443 444 445 447 449 451 452 454 455 457 458 461 467
439 441 460 463 465 468
380000 390000 400000
470 473 475 481 482 493 495 496 498 499 501
469 471 477 478 480 484 486 488 491 500
410000 420000 430000 440000
501 502 505 508 512 513 515 518 521 524 527 534 536 537
503 506 510 511 516 528 530 531 532 533 539
450000 460000 470000 480000
545 546 551 557 560 565 570 575 577 578 580 581
541 542 544 547 549 552 555 561 563 568 571 573 576
490000 500000 510000 520000
583 587 588 589 598 600 602 604 606 611 612 614 617 620 622
585 591 592 594 595 597 603 608 610 615 616 618 623
530000 540000 550000
627 629 632 633 638 641 650 652 654 661 662 664 667
626 628 634 636 640 643 645 646 647 649 653 656 657 658 660 666 668
560000 570000 580000 590000
676 694 696 700 702 705 706 707 708 709
671 673 674 675 677 681 683 685 687 689 691 693 698
600000 610000 620000 630000

FIG. 3.

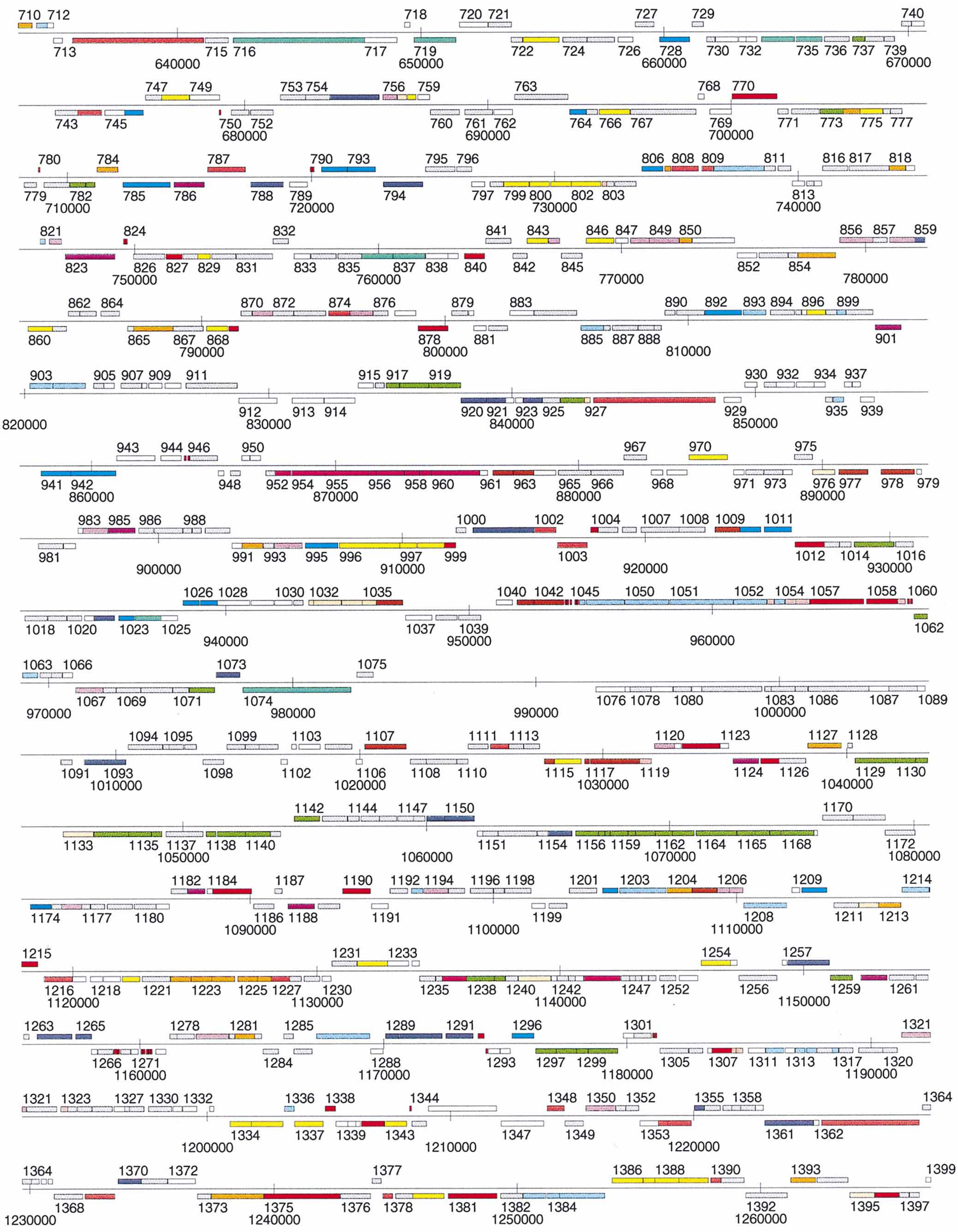

FIG. 3—*Continued.*

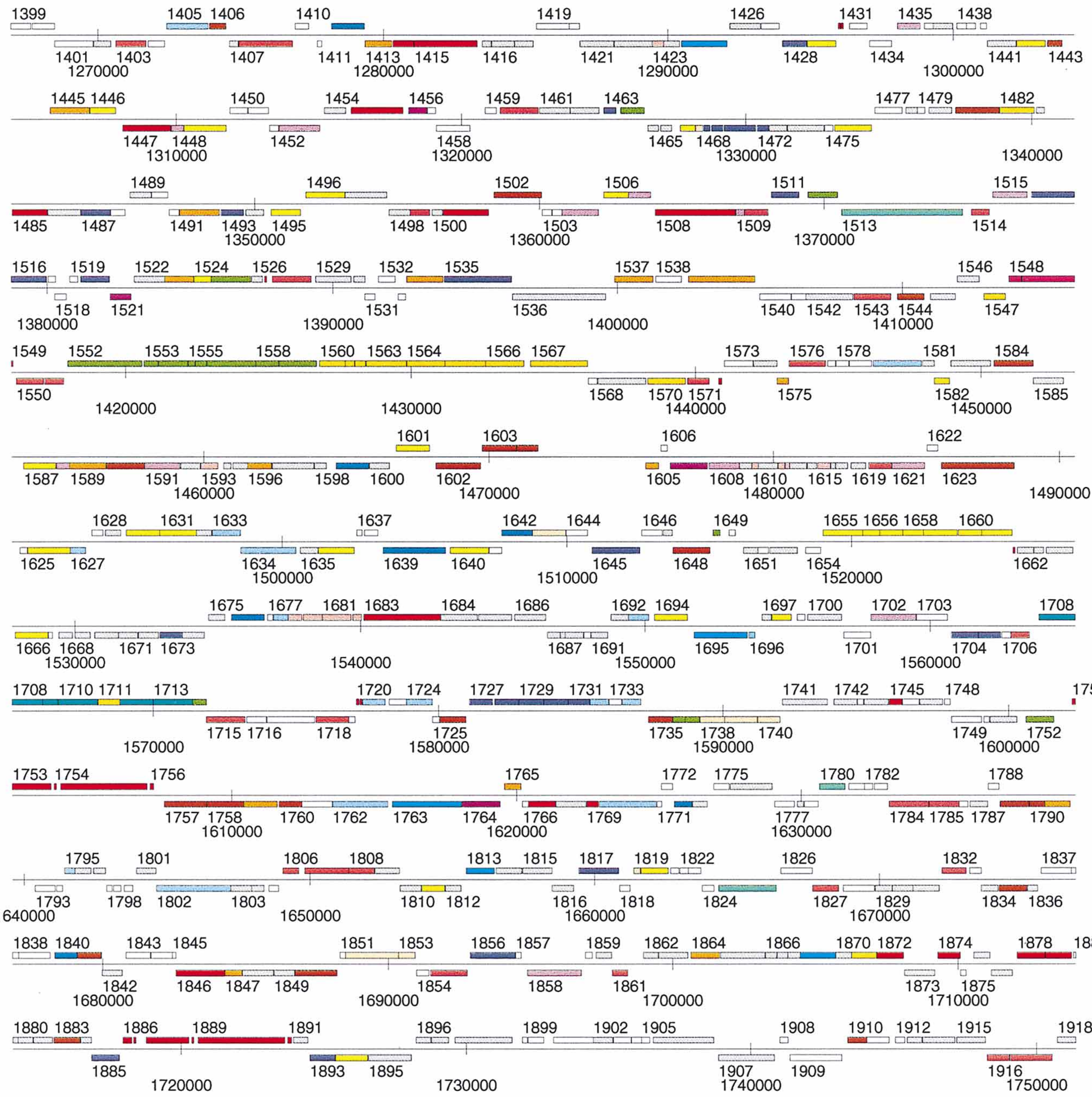

FIG. 3. Gene map of the *M. thermoautotrophicum* ΔH genome. A total of 1,918 putatively identified genes, including 16 that appeared to be caused by frameshifts, are shown with the genes transcribed from the forward strand above the central line in each row and those transcribed from the complementary strand below the line. Genome positions are given by numbers below the periodically spaced tic marks in each row. The genes are color coded according to function as described in the legend to Fig. 1, except that conserved genes are gray and genes with unknown functions are indicated in white. Gene numbers are placed above or below the left end of genes to which they correspond on the forward and complementary strands, respectively. Some gene numbers have been omitted to avoid overlaps in tightly packed regions.

and the SW%IDs graphically represented in Fig. 1 and 2 are available at the GTC web site (http//www.cric.com). As SW%IDs of $<$30 often result from spurious alignments with many gaps, comparative analyses are only reported of aligned sequences with a SW%IDs of $>$30.

Genome organization. Genes are distributed evenly around the *M. thermoautotrophicum* genome, with $\sim$51% being transcribed from one strand and $\sim$49% being transcribed from the complementary strand. Approximately 92% of the genome is predicted to encode gene products, and intergenic regions average $\sim$75 bp. There are two rRNA operons and two regions that contain a large number of repeated sequences (see below).

Functionally related genes are often clustered, and most polypeptide-encoding genes are preceded by sequences consistent with RBSs. Despite these bacterial operon-like features, some of the genes in these clusters have only eucaryal homologs, suggesting that either there has been a selection for clustering or that these genes were clustered in a common ancestor of the domain *Eucarya* and *M. thermoautotrophicum*. Uncoupling of translation and transcription, and the fusion of adjacent genes during the evolution of the eucaryal lineage, may have removed the need for cotranscription and RBSs as few functionally related genes are adjacent in the yeast genome.

A very large transcriptional unit may be formed by 51 genes, including 31 r-protein genes that constitute the region from 0 to 30 kb, and two operons that contain 14 methane genes that total ~9 kb beginning at 1.07 Mbp are cotranscribed under high growth-rate conditions (Fig. 3) (45). Fifteen additional clusters contain at least four functionally related genes which, therefore, are also likely to be single transcriptional units (designated operons). When compared with the *M. jannaschii* genome, related genes occur within conserved operons, but only 14% of orthologous genes have the same neighbor in the two genomes (Fig. 1 and 2). The 8-kb region of the *M. thermoautotrophicum* genome that is only ~40% G+C (see above) is not present in *M. jannaschii*, and an ~29-kb region that contains 36 unidentified genes (MJ0327 to MJ0362) in *M. jannaschii* is not present in *M. thermoautotrophicum*. The cluster of *M. thermoautotrophicum* r-protein genes beginning at position 1 is essentially a sequential fusion of the S10, spc, alpha, and L13 ribosomal operons in *E. coli*, and most of these r-protein genes occur in the same order in two clusters in *M. jannaschii*, one corresponding to the central part and one to the two ends of the *M. thermoautotrophicum* cluster. Five of these *M. thermoautotrophicum* r-protein genes are dispersed as single genes and as a three-gene cluster at separate locations in the *M. jannaschii* genome.

Gene families. A total of 409 (22%) of the *M. thermoautotrophicum* genes group into 111 families with two or more members, by using the alignment parameters established by Bult et al. (9). This is less than the 136 gene families detected in *M. jannaschii*, and only 59 families are conserved in both methanogens. The largest gene family in *M. jannaschii* has 16 members of unknown function that together account for almost 1% of the genome's coding capacity. Surprisingly, there are no members of this family in *M. thermoautotrophicum*, and the largest *M. thermoautotrophicum* family, which encodes 24 two-component sensor kinase-response regulator proteins, has no representatives in *M. jannaschii*. Other large and conserved families in *M. thermoautotrophicum* encode 15 ferredoxin-related proteins, 9 members of the ABC transporter family, 11 IMP dehydrogenase-related proteins, and 6 proteins related to magnesium chelatases. The complete list of gene families is available on the GTC web site.

Methane genes. The enzymes that catalyze the seven steps in the H_2-dependent pathway of CO_2 reduction to CH_4 were characterized primarily through studies of *M. thermoautotrophicum* (Fig. 5) (60, 69), and most of their encoding methane genes were sequenced prior to the completion of the genome sequence (46). *M. thermoautotrophicum* was known to have two step 1-catalyzing enzymes, a tungsten and a molybdenum formylmethanofuran dehydrogenase (W-FMD and Mo-FMD, respectively), two step 4-catalyzing methylene tetrahydromethanopterin dehydrogenases (HMD and MTD), and two step 7-catalyzing methyl coenzyme M reductase isoenzymes (MRI and MRII). The genome sequence predicts the presence of a second step 2-catalyzing formylmethanofuran: tetrahydromethanopterin formyltransferase (FTR) and two additional step 4-catalyzing enzymes. The *ftrII*-encoded amino acid sequence is 38% identical to the *ftr*-encoded protein (14). Similarly, *hmdII* and *hmdIII* encode amino acid sequences which are 24 and 32% identical, respectively, to the sequence of the *hmd*-encoded HMD (36). Based on the conservation of methane genes, *M. jannaschii* apparently employs the same H_2-dependent pathway of CH_4 synthesis from CO_2 and also has three *hmd* genes, but it contains only one *ftr* and only genes for a W-FMD. The only conservation in methane gene organization in both genomes, above the level of related genes within similarly organized operons, is the adjacent positioning of the *mcrBDCGA* and *mtrEDCBAFGH* operons. These operons encode MRI and methyltetrahydromethanopterin:coenzyme M methyltransferase (MTR), which catalyze steps 7 and 6 in methanogenesis, respectively. Read-through transcription of the *mtr* operon from the *mcr* promoter has been documented in *M. thermoautotrophicum* (45), and as this adjacent organization is widespread in methanogens, this suggests functional significance (37). Both methanogens have *mrt* operons that encode MRII, the isoenzyme of MRI, that catalyzes step 7 in *M. thermoautotrophicum* when excess H_2 is available (45). The *mrt* operon in *M. thermoautotrophicum* is organized *mrtBDGA*, whereas *mrtD* is separated by ~37 kb from an *mrtBGA* operon in *M. jannaschii*. The *mcrBGA*/*mrtBGA* genes encode the three polypeptide subunits of MRI/MRII; however, the functions of the *mcrD*, *mrtD*, and *mcrC* gene products remain unknown. The sequences of MJ0094 and MTH1161 suggest that they may be very divergent *mrtC* genes.

M. thermoautotrophicum and *M. jannaschii* have genes related to the *fdhAB* genes that encode formate dehydrogenases (FDH) in formate-catabolizing methanogens but neither of them grows on formate (23, 56). *M. thermoautotrophicum* appears to have lost an *fdhCAB* operon (38), and the *flpECBDA* operon encodes only FDH-like gene products (36). The sequence of the *M. jannaschii fdhBA* operon is, however, consistent with a functional FDH.

Based on homologies with *Methanococcus voltae* (18, 55) *M. jannaschii* synthesizes a [Ni,Fe,Se]-hydrogenase with in-frame UGA codons directing the incorporation of selenocysteinyl (Se-cys) residues (67). An in-frame UGA codon in *hdrA* in *M. jannaschii* predicts that Se-cys is also incorporated into the large subunit of the heterodisulfide reductase (HDR) of this methanogen. The *M. thermoautotrophicum* genome does not encode the translation machinery needed for Se-cys incorporation, and the [Ni,Fe]-hydrogenase genes (*frhDBGA* and *mvhDGAB*) and *hdrA* of *M. thermoautotrophicum* have cysteine codons at the sites of the Se-cys UGA codons in *M. jannaschii*. In both methanogens HDR is encoded by unlinked *hdrA* and *hdrCB* operons. *M. thermoautotrophicum* has one *hdrCB* operon plus an *hdrB*-related gene, MTH0139, while *M. jannaschii* has two *hdr*CB operons.

Cofactor F_{390} levels have been proposed to regulate the expression of alternative methane genes in *M. thermoautotrophicum* (36, 62). However, the presence of *ftsAII* and *ftsAIII*, two additional homologs of the *ftsA* gene known to encode cofactor F_{390} synthetase in *M. thermoautotrophicum*, makes this issue problematic, and the absence of *ftsA* homologs in *M. jannaschii* argues against a generic role for cofactor F_{390} synthesis in methane gene regulation.

Carbon metabolism, nitrogen fixation, and anabolic pathways. Genes encoding several of the enzymes required to catalyze glycolysis, gluconeogenesis, and the pentose phosphate pathway have not been identified in the *M. thermoautotrophicum* genome. Therefore, either these pathways do not exist in *M. thermoautotrophicum* and functionally equivalent but different pathways must be used or the sequences of the *M. thermoautotrophicum* phosphofructokinase, pyruvate kinase, phosphoglucoisomerase, fructose bisphosphatase, fructose 1,6-diphosphoaldolase, phosphoglyceromutase, ribulose phosphate epimerase, transketolase, transaldolase, and 6-phosphodehydrogenase are so different from database sequences that they are unrecognizable. These conclusions were also reached for several "missing" enzymes needed to catalyze steps in cen-

Hydrogen metabolism and methanogenesis

1528	CoF390 Sase
1855	CoF390 Sase II
161	CoF390 Sase III
1464	CoF420-dependent N5,N10-methylene H4MPT DHase
1752	CoF420-dependent N5,N10-methylene H4MPT RDase
1300	CoF420-reducing hydrogenase, α sub
1297	CoF420-reducing hydrogenase, β sub
193	CoF420-reducing hydrogenase, β sub homolog
280	CoF420-reducing hydrogenase, β sub homolog
341	CoF420-reducing hydrogenase, β sub homolog
1299	CoF420-reducing hydrogenase, δ sub
737	CoF420-reducing hydrogenase, δ sub homolog
1298	CoF420-reducing hydrogenase, γ sub
1212	cytochrome-c3 hydrogenase, γ sub
1552	formate DHase, α sub homolog
1140	formate DHase, α sub rel prot FlpC
1139	formate DHase, β sub rel prot FlpB
1714	formate hydrogenlyase, iron-sulfur sub 2
1736	formate hydrogenlyase, iron-sulfur sub 2
1737	formate hydrogenlyase, iron-sulfur sub I
398	formate hydrogenlyase, sub 5
1238	formate hydrogenlyase, sub 5
397	formate hydrogenlyase, sub 7
1239	formate hydrogenlyase, sub 7
1259	formylmethanofuran:H4MPT formyl-Tase
403	formylmethanofuran:H4MPT formyl-Tase II
1142	H(2)-dependent N5,N10-methylene-H4MPT DHase
1512	H(2)-dependent N5,N10-methylene-H4MPT DHase II
504	H(2)-dependent N5,N10-methylene-H4MPT DHase III
139	heterodisulfide RDase sub B rel prot
1381	heterodisulfide RDase, sub A
1879	heterodisulfide RDase, sub B
1878	heterodisulfide RDase, sub C
783	hydrogenase expression/formation prot HypA
782	hydrogenase expression/formation prot HypB
1649	hydrogenase expression/formation prot HypC
1072	hydrogenase expression/formation prot HypD
205	hydrogenase expression/formation prot HypE
1525	hydrogenase expression/formation prot HypE rel prot
1164	methyl CoM RDase I, α sub
1168	methyl CoM RDase I, β sub
1166	methyl CoM RDase I, C prot
1167	methyl CoM RDase I, D prot
1165	methyl CoM RDase I, γ sub
1129	methyl CoM RDase II, α sub
1132	methyl CoM RDase II, β sub
1131	methyl CoM RDase II, D prot
1130	methyl CoM RDase II, γ sub
1015	methyl CoM RDase system, component A2
454	methyl CoM RDase system, component A2 homolog
151	methyl CoM RDase system, component A2 homolog
1134	MV-reducing hydrogenase, α sub
1136	MV-reducing hydrogenase, δ sub
1138	MV-reducing hydrogenase, δ sub homolog FlpD
1135	MV-reducing hydrogenase, γ sub
919	molybdenum formylmethanofuran DHase, sub B
918	molybdenum formylmethanofuran DHase, sub C
917	molybdenum formylmethanofuran DHase, sub E
773	N5,N10-methenyl-H4MPT cyclohydrolase
1159	N5-methyl-H4MPT:CoM MTase, sub A
1062	N5-methyl-H4MPT:CoM MTase, sub A homolog
1160	N5-methyl-H4MPT:CoM MTase, sub B
1161	N5-methyl-H4MPT:CoM MTase, sub C
1162	N5-methyl-H4MPT:CoM MTase, sub D
1163	N5-methyl-H4MPT:CoM MTase, sub E
1158	N5-methyl-H4MPT:CoM MTase, sub F
1157	N5-methyl-H4MPT:CoM MTase, sub G
1156	N5-methyl-H4MPT:CoM MTase, sub H
1548	NADP-reducing hydrogenase, sub A
1549	NADP-reducing hydrogenase, sub C
1557	tungsten formylmethanofuran DHase, sub A
1559	tungsten formylmethanofuran DHase, sub B
1558	tungsten formylmethanofuran DHase, sub C
106	tungsten formylmethanofuran DHase, sub C homolog
192	tungsten formylmethanofuran DHase, sub C homolog
238	tungsten formylmethanofuran DHase, sub C homolog
1556	tungsten formylmethanofuran DHase, sub D
1554	tungsten formylmethanofuran DHase, sub F
926	tungsten formylmethanofuran DHase, sub F homolog
1555	tungsten formylmethanofuran DHase, sub G
1553	tungsten formylmethanofuran DHase, sub H

Electron transport and redox metabolism

536	2-oxoacid:ferredoxin oxidoRDase, α sub
537	2-oxoacid:ferredoxin oxidoRDase, β sub
1033	2-oxoglutarate oxidoRDase, α sub
1034	2-oxoglutarate oxidoRDase, β sub
1035	2-oxoglutarate oxidoRDase, γ sub
1032	2-oxoglutarate oxidoRDase, δ sub
705	2-oxoisovalerate oxidoRDase, α sub
704	2-oxoisovalerate oxidoRDase, β sub
703	2-oxoisovalerate oxidoRDase, γ sub
278	ferredoxin
854	ferredoxin
927	ferredoxin
1106	ferredoxin
1468	ferredoxin
1719	ferredoxin
1819	ferredoxin
1240	ferredoxin-like prot
1350	flavoprotein A
220	flavoprotein A II
157	flavoprotein A III
1852	indolepyruvate oxidoRDase, α sub
1853	indolepyruvate oxidoRDase, β sub
120	NADPH-oxidoRDase
399	polyferredoxin
400	polyferredoxin*
401	polyferredoxin*
405	polyferredoxin
1241	polyferredoxin
1133	polyferredoxin (MvhB)
1586	pyruvate formate-lyase activating enzyme
976	pyruvate formate-lyase activating enzyme rel prot
1395	pyruvate formate-lyase activating enzyme rel prot
1643	pyruvate formate-lyase activating enzyme rel prot
1739	pyruvate oxidoRDase, α sub
1738	pyruvate oxidoRDase, β sub
1740	pyruvate oxidoRDase, γ sub
156	rubredoxin
155	rubredoxin
1352	rubredoxin rel prot
757	rubredoxin oxidoRDase
756	rubrerythrin
822	rubrerythrin
807	thioredoxin
708	thioredoxin RDase

ATPases

1511	arsenical pump-driving ATPase
955	ATP Sase, sub A
954	ATP Sase, sub B
957	ATP Sase, sub C
953	ATP Sase, sub D
958	ATP Sase, sub E
956	ATP Sase, sub F
960	ATP Sase, sub I
959	ATP Sase, sub K
411	cadmium efflux ATPase
1493	cation transporting P-type ATPase rel prot
1001	cation-transporting P-ATPase PacL
1516	cation-transporting P-ATPase PacL
481	H+-transporting ATPase
482	H+-transporting ATPase
755	heavy-metal transporting CPx-type ATPase
1535	heavy-metal transporting CPx-type ATPase
1176	nucleotide-bind prot (putative ATPase)

Glycolysis/Gluconeogenesis

1883	2-phosphoglycerate kinase
1835	2-phosphoglycerate kinase homolog
1042	3-phosphoglycerate kinase
1648	dihydrolipoamide DHase
43	enolase
1009	glyceraldehyde 3-phosphate DHase
188	lactate DHase
978	NADP-dependent glyceraldehyde-3-phosphate DHase
1041	triosephosphate isomerase

Citrate cycle

962	citrate Sase I
1726	citrate Sase I
1115	fumarate hydratase, class I
1735	fumarate hydratase, class I
963	fumarate hydratase, class I rel prot
1910	fumarate hydratase, class I rel prot
1850	fumarate RDase
184	isocitrate DHase
1205	malate DHase
1502	succinate DHase, flavoprot sub
563	succinyl-CoA Sase, α sub
1036	succinyl-CoA Sase, β sub

Pentose phosphate cycle

404	ribokinase
1544	ribokinase
1841	ribokinase
608	ribose 5-phosphate isomerase

Pyruvate and acetyl-CoA metabolism

701	acetyl-CoA Sase rel prot*
702	acetyl-CoA Sase rel prot*
216	acetyl-CoA Sase*
217	acetyl-CoA Sase*
1603	acetyl-CoA Sase*
1604	acetyl-CoA Sase*
346	formate acetyl-Tase 2
1406	fuculose-1-phosphate aldolase
1481	isopropylmalate Sase
1107	oxaloacetate decarboxylase, α sub
460	phosphoenolpyruvate Sase homolog
1117	phosphoenolpyruvate Sase*
1118	phosphoenolpyruvate Sase*
476	pyruvate DHase/acetolactate Sase
345	pyruvate formate-lyase 2 activating enzyme

Butanoate metabolism

1444	acetolactate Sase, large sub
1602	acetolactate Sase, large sub homolog
1443	acetolactate Sase, small sub

Carbon fixation

1708	carbon monoxide DHase, α sub
1710	carbon monoxide DHase, α sub
1709	carbon monoxide DHase, β sub
1582	carbonic anhydrase

Nitrogen metabolism

1567	catalytic sub of nitrate RDase
1570	glutamine Sase
1120	NifH/MinD rel prot
1389	NifS prot
1547	nitrate assimilation prot, NarQ
662	nitrogen regulatory prot P-II
664	nitrogen regulatory prot P-II
1561	nitrogenase GlnBa sub
1562	nitrogenase GlnBb sub
1871	nitrogenase Mo-Fe cofactor biosyn prot NifB
1565	nitrogenase Mo-Fe cofactor biosyn prot NifE
1482	nitrogenase Mo-Fe cofactor biosyn prot NifE homolog
1564	nitrogenase Mo-Fe prot, NifK sub
1566	nitrogenase Mo-Fe prot, NifN sub
1563	nitrogenase NifD sub
1522	nitrogenase NifD sub rel prot
643	nitrogenase NifH sub
1560	nitrogenase NifH sub
1711	nitrogenase RDase rel prot

Sulfur metabolism

113	arylsulfatase regulatory prot*
114	arylsulfatase regulatory prot*
622	thiosulfate sulfur-Tase

Fructose and mannose metabolism

1790	dTDP-4-dehydrorhamnose 3,5-epimerase
1792	dTDP-4-dehydrorhamnose RDase
1789	dTDP-glucose 4,6-DTase
1584	phosphomannomutase
1590	phosphomannomutase
1758	phosphomannomutase

Di-saccharide metabolism

1757	α,α-trehalose-phosphate Sase
1760	trehalose-6-phosphate phosphatase rel prot

Polysaccharide and starch metabolism

437	endo-1,4-β-glucanase
977	endo-1,4-β-glucanase rel prot
1623	oligosaccharyl Tase STT3 sub rel prot

Alanine, aspartate and glutamate metabolism

183	acetylglutamate kinase
269	argininosuccinate lyase
1254	argininosuccinate Sase
414	asparagine Sase
1601	aspartate amino-Tase
1894	aspartate amino-Tase homolog
52	aspartate amino-Tase rel prot
1694	aspartate amino-Tase rel prot
799	aspartate-semialdehyde DHase
802	aspartokinase II α sub
997	carbamoyl-phosphate Sase, large sub*
996	carbamoyl-phosphate Sase, large sub*
998	carbamoyl-phosphate Sase, small sub
860	glucosamine-fructose-6-phosphate amino-Tase
1116	glutamate decarboxylase
182	glutamate NAc-Tase
105	glutamate Sase (NADPH), α sub
194	glutamate Sase (NADPH), α sub
1666	glutamate Sase (NADPH), α sub rel prot

FIG. 4. Functional classification of *M. thermoautotrophicum* gene products. Gene product names and functional categories are based on the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg). Gene numbers correspond to those shown in Fig. 3. An expanded version of this table with additional information is available on the GTC web site (http://www.cric.com). Asterisks indicate genes which may contain frameshifts. Abbreviations: bind, binding; biosyn, biosynthesis; Co, coenzyme; dinuc, dinucleotide; DHase, dehydrogenase; DTase, dehydratase; fam, family; GlcNAc, N-acetylglucosamine; H4MPT, tetrahydromethanopterin; LPS, lipopolysaccharide; m5C, 5-methylcytosine; Mo-Fe, molybdenum-iron; MTase, methyltransferase; MV, methylviologen; MurNAc, N-acetylmuramyl; NAc, *N*-acetyl; PQQ, pyrrolo-quinoline-quinone; PR, phosphoribosyl; PRPP, phosphoribosylpyrophosphate; PRTase, phosphoribosyltransferase; prot, protein; RDase, reductase; rel, related; Sase, synthetase or synthase; sub, subunit; Tase, transferase; triP, triphosphate.

171 glutamine-fructose-6-phosphate transaminase
225 histidinol DHase
1467 imidazoleglycerol-phosphate DTase
1524 imidazoleglycerol-phosphate Sase
706 L-asparaginase I
1337 NAc-ornithine amino-Tase

Glycine, serine and threonine metabolism

1232 homoserine DHase
417 homoserine DHase homolog
970 phosphoglycerate DHase
1626 phosphoserine phosphatase
1380 serine hydroxy-MTase
253 threonine Sase

Methionine metabolism

775 cobalamin-independent methionine Sase
1820 homoserine O-acetyl-Tase
1636 S-adenosylhomocysteine hydrolase

Valine, leucine and isoleucine metabolism

723 2-isopropylmalate Sase
1630 2-isopropylmalate Sase
1387 3-isopropylmalate DTase, LeuC sub
1631 3-isopropylmalate DTase, LeuC sub
1386 3-isopropylmalate DTase, LeuD sub
829 3-isopropylmalate DTase, LeuD sub
1388 3-isopropylmalate DHase
1430 branched-chain amino acid amino-Tase
1449 dihydroxy-acid DTase
1442 ketol-acid reductoisomerase

Lysine metabolism

800 dihydrodipicolinate RDase
801 dihydrodipicolinate Sase

Arginine and proline metabolism

868 agmatine ureohydrolase
1698 δ 1-pyrroline-5-carboxylate Sase
1446 ornithine carbamoyl-Tase
1495 ornithine cyclodeaminase
897 pyrroline-5-carboxylate RDase

Histidine metabolism

1506 ATP PRTase
119 ATP PRTase rel prot
1587 histidinol-phosphate amino-Tase
1343 imidazoleglycerol-phosphate Sase (cyclase)
245 PR-AMP cyclohydrolase
843 PR-formimino-5-aminoimidazole carboxamide ribotide isomerase
669 PR-formimino-5-aminoimidazole carboxamide ribotide isomerase rel prot

Phenylalanine, tyrosine and tryptophan metabolism

566 3-dehydroquinate DTase
1658 5'-PR anthranilate isomerase
766 5-enolpyruvylshikimate 3-phosphate Sase
1661 anthranilate PRTase
1655 anthranilate Sase component I
1656 anthranilate Sase component II
1220 chorismate mutase
1640 chorismate mutase
748 chorismate Sase
1657 indole-3-glycerol phosphate Sase
242 shikimate 5-DHase
1659 tryptophan Sase, β sub
1476 tryptophan Sase, β sub homolog
1660 tryptophan Sase, sub α

Purine metabolism

1492 5'-nucleotidase
866 adenine deaminase
27 adenylate kinase
1663 adenylate kinase homolog
1537 adenylosuccinate lyase
615 adenylosuccinate Sase
646 amido-PRTase
1539 anaerobic ribonucleoside-triP RDase
287 anaerobic ribonucleoside-triP RDase activating prot
1445 glycinamide ribonucleotide Sase
709 GMP Sase, sub A
710 GMP Sase, sub B
142 inosine-5'-monophosphate DHase
1222 inosine-5'-monophosphate DHase rel prot I
1223 inosine-5'-monophosphate DHase rel prot II
1224 inosine-5'-monophosphate DHase rel prot III
1225 inosine-5'-monophosphate DHase rel prot IV
992 inosine-5'-monophosphate DHase rel prot IX
1226 inosine-5'-monophosphate DHase rel prot V
1282 inosine-5'-monophosphate DHase rel prot VI
126 inosine-5'-monophosphate DHase rel prot VII
855 inosine-5'-monophosphate DHase rel prot VIII
1575 inosine-5'-monophosphate DHase rel prot X
1393 PR-aminoimidazole carboxylase
1286 PR-aminoimidazole carboxylase rel prot
170 PR-aminoimidazolesuccinocarboxamide Sase
1204 PR-formylglycinamidine cyclo-ligase
168 PR-formylglycinamidine Sase I
1374 PR-formylglycinamidine Sase II
1864 PR-formylglycinamidine Sase II rel prot
652 ribonucleotide RDase, large sub

Pyrimidine metabolism

1413 aspartate carbamoyl-Tase
850 aspartate carbamoyl-Tase regulatory sub
419 CTP Sase
1847 deoxycytidine-triP deaminase
1605 deoxycytidine-triP deaminase rel prot
1127 dihydroorotase
1213 dihydroorotate oxidase
1605 deoxyuridne 5'-triP nucleotidohydrolase rel prot
129 orotidine 5' monophosphate decarboxylase
840 pseudouridylate Sase I
434 UMP/CMP kinase rel prot
879 UMP kinase
1114 uracil phosphoribosyl-Tase
1860 uridine 5'-monophosphate Sase

Nucleotide sugar metabolism

373 dTDP-glucose 4,6-DTase rel prot
1523 glucose-1-phosphate adenylyl-Tase rel prot
1791 glucose-1-phosphate thymidylyl-Tase
1589 glucose-1-phosphate thymidylyl-Tase homolog
1759 mannose-1-phosphate guanyl-Tase
634 UTP-glucose-1-phosphate uridylyl-Tase
631 UDP-glucose 4-epimerase
380 UDP-glucose 4-epimerase homolog
375 UDP-glucose 4-epimerase rel prot

Salvage and interconversion pathways

30 cytidylate kinase
818 deoxyribose-phosphate aldolase
1596 methylthioadenosine phosphorylase
784 ribose-phosphate pyrophosphokinase
1765 thymidylate kinase
774 thymidylate Sase

Cofactor metabolism

1917 biotin carboxylase
1916 biotin [acetyl-CoA carboxylase] ligase/biotin operon repressor bifunctional prot
1785 Co PQQ synthesis prot
1227 Co PQQ synthesis prot III
1713 corrinoid/iron-sulfur prot, large sub
1712 corrinoid/iron-sulfur prot, small sub
228 glutamate-1-semialdehyde amino-Tase
1499 GTP cyclohydrolase II
393 NADH DHase (ubiquinone), sub 1 rel prot
1237 NADH DHase (ubiquinone), sub 1 rel prot
1246 NADH DHase I, sub N
1354 NADH oxidase
1510 NH(3)-dependent NAD+ Sase
1216 pantothenate metabolism flavoprot
1807 phytoene DHase
1808 phytoene Sase
227 precorrin isomerase
1832 quinolinate PRTase
1827 quinolinate Sase
1390 riboflavin Sase β sub
235 riboflavin-specific deaminase
758 S-D-lactoylglutathione methylglyoxal lyase
1543 thiamine biosyn prot
1576 thiamine biosyn prot
1620 thiamine biosynthetic prot
1396 thiamine monophosphate kinase

Porphyrin metabolism

277 bacteriochlorophyll Sase 43 kDa sub
1718 bacteriochlorophyll Sase 43 kDa sub
1098 bacteriochlorophyll Sase rel prot
1112 cobalamin (5'-phosphate) Sase
808 cobalamin biosyn prot D
1409 cobalamin biosyn prot B
1408 cobalamin biosyn prot G
1002 cobalamin biosyn prot J
130 cobalamin biosyn prot M
1707 cobalamin biosyn prot M
200 cobalamin biosyn prot M rel prot
514 cobalamin biosyn prot N
673 cobalamin biosyn prot N
1363 cobalamin biosyn prot N
787 cobyric acid Sase
1460 cobyrinic acid a,c-diamide Sase
1497 cobyrinic acid a,c-diamide Sase rel prot
237 magnesium chelatase sub
351 magnesium chelatase sub
456 magnesium chelatase sub
714 magnesium chelatase sub
928 magnesium chelatase sub
317 magnesium chelatase sub *
318 magnesium chelatase sub *
555 magnesium chelatase sub *
451 magnesium chelatase sub Chl I
556 magnesium chelatase sub Chl I*
1784 Mg-protoporphyrin IX monomethylester oxidative cyclase
1378 phycocyanin α phycocyanobilin lyase CpcE
1806 phycocyanin α phycocyanobilin lyase CpcE
1715 phycocyanin α phycocyanobilin lyase CpcE rel prot
874 porphobilinogen deaminase
744 porphobilinogen Sase
1348 precorrin-2 MTase
602 precorrin-3 methylase
1403 precorrin-3 methylase
1514 precorrin-6Y methylase
146 precorrin-8W decarboxylase
167 S-adenosyl-L-methionine uroporphyrinogen MTase
166 uroporphyrinogen III Sase

Molybdopterin metabolism

1550 molybdenum cofactor biosyn MoaA
62 molybdenum cofactor biosyn MoaA rel prot
1861 molybdenum cofactor biosyn MoaB
1369 molybdenum cofactor biosyn MoeA
809 molybdenum cofactor biosyn prot MoaC
149 molybdenum cofactor biosyn prot MoaE
1003 molybdenum cofactor biosyn prot MoeA
1571 molybdopterin biosyn prot MoeB homolog
143 molybdopterin-guanine dinucl biosyn MobA rel prot
1551 molybdopterin-guanine dinucl biosyn prot B rel

Fatty acid metabolism

272 acetyl/acyl Tase rel prot
50 bifunctional short chain isoprenyl diphosphate Sase
657 long-chain-fatty-acid-CoA ligase
46 mevalonate kinase

Sterol metabolism

562 3-hydroxy-3-methylglutaryl CoA RDase
792 3-hydroxy-3-methylglutaryl-CoA Sase
1869 activator of (R)-2-hydroxyglutaryl-CoA
793 lipid-transfer prot

Diaminopimelate metabolism

1335 diaminopimelate decarboxylase
1334 diaminopimelate epimerase

Glycerolipid metabolism

1027 CDP-diacylglycerol-serine O-phosphatidylTase
368 glycerol-3-phosphate DHase (NAD)
610 glycerol-1-phosphate DHase
1026 phosphatidylserine decarboxylase

Cell envelope and membrane

604 adhesion prot
362 capsular polysaccharide biosyn prot
1825 cell surface glycoprot
716 cell surface glycoprot (s-layer prot)
719 cell surface glycoprot (s-layer prot)
1513 cell surface glycoprot (s-layer prot) rel prot
374 dolichyl-phosphate mannose Sase rel prot
377 dolichyl-phosphate mannose Sase rel prot
335 galactosyl-Tase RfpB rel prot
333 GDP-D-mannose DTase
138 GlcNAc-phosphatidylinositol rel biosynthetic prot
590 GlcNAc-1-phosphate Tase
173 LPS biosyn RfbU rel prot
370 LPS biosyn RfbU rel prot
332 LPS biosyn RfbU rel prot
338 LPS biosyn RfbU rel prot
450 LPS biosyn RfbU rel prot
331 mannosyl Tase
334 perosamine Sase
735 phospho-NAcmuramoyl-pentapeptide-Tase
572 polysaccharide biosyn prot
1074 putative membrane prot
1092 putative membrane prot
343 rhamnosyl Tase
1024 rod shape-determining prot
1702 secretory prot kinase
692 stomatin-like prot

FIG. 4—*Continued.*

1780	stomatin-like prot
342	succinoglycan biosyn transport prot
361	teichoic acid biosyn prot RodC rel prot
365	teichoic acid biosyn prot RodC rel prot
344	UDP-galactopyranose mutase
836	UDP-NAc-D-mannosaminuronic acid DHase
837	UDP-GlcNAc 2-epimerase
369	UDP-GlcNAc pyrophosphorylase rel prot
530	UDP-MurNAc tripeptide Sase rel prot
531	UDP-MurNAc tripeptide Sase rel prot
532	UDP-MurNAc tripeptide Sase rel prot
734	UDP-MurNAc tripeptide Sase rel prot

Cell division

1639	cell division control prot Cdc48
1840	cell division inhibitor
1173	cell division inhibitor rel prot
1174	cell division inhibitor rel prot
1642	cell division prot
1676	cell division prot FtsZ
1773	cell division prot J
32	centromere/microtubule-bind prot

Chaperones

218	chaperonin
794	chaperonin
1291	DnaJ prot
1290	DnaK prot (Hsp70)
1289	heat shock prot GrpE
569	heat shock prot X
1817	heat shock prot X rel prot
859	heat shock prot, class I
686	proteasome, α sub
1202	proteasome, β sub

Protein and peptide secretion

26	preprot translocase SecY
849	prot-export membrane prot, SecD
848	prot-export membrane prot, SecF
1448	signal peptidase
165	signal recognition particle 19 kDa prot
1608	signal recognition particle prot (docking prot)
1321	signal recognition particle prot SRP54

Protein modification and degradation

728	ATP-dependent 26S protease regulatory sub 4
1011	ATP-dependent 26S protease regulatory sub 8
284	ATP-dependent Clp protease regulatory sub
785	ATP-dependent protease LA
892	ATP-dependent protease LA rel prot
645	collagenase
1763	collagenase
827	L-isoaspartyl prot carboxyl MTase
995	lysyl endopeptidase
1296	methionine aminopeptidase
999	N-terminal acetyl-Tase complex, sub ARD1
1425	O-sialoglycoprot endopeptidase
535	peptide methionine sulfoxide RDase
1125	peptidyl-prolyl cis-trans isomerase
1338	peptidyl-prolyl cis-trans isomerase B
806	protease IV
1745	prot disulphide isomerase
283	prot kinase
1414	prot-L-isoaspartate MTase homolog
1918	prot Mtase rel prot
1813	serine protease HtrA
1485	serine/threonine prot kinase
75	surface protease rel prot
87	surface protease rel prot

Detoxification

875	3-chlorobenzoate-3,4-dioxygenase DHase rel prot
159	alkyl hydroperoxide RDase
1355	arsenate RDase
1428	bacitracin resistance prot*
1429	bacitracin resistance prot*
1893	cation efflux system prot (zinc/cadmium)
1509	divalent cation tolerance prot
195	efflux pump antibiotic resistance prot
659	epoxidase
1505	N-ethylammeline chlorohydrolase homolog
994	N-ethylammeline chlorohydrolase rel prot
147	phenylacrylic acid decarboxylase
160	superoxide dismutase (Fe/Mn)
1435	survival prot SurE

Regulatory functions

936	iron repressor
214	iron repressor
707	PET112-like prot
1280	PET112-like prot
1732	phosphate transport system regulator
1734	phosphate transport system regulator
1724	phosphate transport system regulator rel prot
1188	pleiotropic regulatory prot DegT
1634	transcriptional control factor (enhancer-bind prot)
614	transcriptional regulator
1193	transcriptional regulator
313	transcriptional regulator
711	transcriptional regulator
899	transcriptional regulator
1795	transcriptional regulator
1287	transcriptional regulator HypF homolog
178	transcriptional regulator Icc rel prot
1722	transcriptional regulator Icc rel prot
1063	transcriptional regulator rel prot

Two-component signal transductions proteins

123	sensory transduction histidine kinase
174	sensory transduction histidine kinase
292	sensory transduction histidine kinase
356	sensory transduction histidine kinase
360	sensory transduction histidine kinase
444	sensory transduction histidine kinase
459	sensory transduction histidine kinase
468	sensory transduction histidine kinase
619	sensory transduction histidine kinase
823	sensory transduction histidine kinase
902	sensory transduction histidine kinase
985	sensory transduction histidine kinase
1124	sensory transduction histidine kinase
1260	sensory transduction histidine kinase
786	sensory transduction histidine kinase rel prot
440	sensory transduction regulatory prot
445	sensory transduction regulatory prot
446	sensory transduction regulatory prot
447	sensory transduction regulatory prot
457	sensory transduction regulatory prot
548	sensory transduction regulatory prot
549	sensory transduction regulatory prot
901	sensory transduction regulatory prot
1607	sensory transduction regulatory prot
1764	sensory transduction regulatory prot

Transport of organic compounds

605	ABC transporter
1645	ABC transporter
1370	ABC transporter (ATP-bind prot)
1093	ABC transporter (ATP-bind; daunorubicin resistance)
1487	ABC transporter (ATP-bind; daunorubicin resistance)
696	ABC transporter (glutamine transport ATP-bind prot)
1463	ABC transporter rel prot
1149	ABC transporter sub Ycf16
1150	ABC transporter sub Ycf24
1022	biopolymer transport prot
546	cationic amino acid transporter rel prot
540	intracellular prot transport prot
104	multidrug transporter homolog
347	O-antigen transporter
367	O-antigen transporter
379	O-antigen transporter rel prot
1471	O-antigen transporter rel prot*
1472	O-antigen transporter rel prot*
1673	sn-glycerol-3-phosphate transport ATP-bind prot
1856	sodium/proline symporter (proline permease)

Transport of inorganic compounds

661	ammonium transporter
663	ammonium transporter
1073	cation antiporter
1172	cation transporter rel prot
1704	cobalt transport ATP-bind prot O
133	cobalt transport ATP-bind prot O
1705	cobalt transport membrane prot
131	cobalt transport prot N
132	cobalt transport prot Q
358	glutathione-regulated K+/H+ antiporter
158	ferritin like prot RsgA
213	ferrous iron transport prot B
1361	ferrous iron transport prot B
620	Mg2+ transporter
924	molybdate-bind periplasmic prot
1469	molybdenum transport ATP-bind prot homolog
1470	molybdenum transport prot ModA rel prot
1155	Na+/Ca+ exchanging prot rel
788	Na+/dicarboxylate or sulfate cotransporter
1731	phosphate transport system ATP-bind
1729	phosphate transporter permease PstC
1730	phosphate transporter permease PstC homolog
1727	phosphate-bind prot PstS
1728	phosphate-bind prot PstS homolog
1258	potassium channel rel prot
1520	potassium channel rel prot
505	potassium channel rel prot
1885	sodium-dependent phosphate transporter
920	sulfate permease
477	sulfate transport system ATP-bind
921	sulfate transport system permease prot
478	sulfate transport system permease prot
1265	TRK system potassium uptake prot TrkA
1264	TRK system potassium uptake prot TrkH

DNA metabolism, modification and replication

312	ATP-dependent helicase
1802	ATP-dependent helicase
1347	ATP-dependent helicase rel prot
1412	Cdc6 rel prot
1599	Cdc6 rel prot
1456	chromosome partitioning prot Soj
904	DNA deoxyribodipyrimidine photolyase
472	DNA helicase II
511	DNA helicase II
551	DNA helicase II rel prot
487	DNA helicase rel prot
810	DNA helicase rel prot
1580	DNA ligase
1762	DNA mismatch recognition prot MutS
1405	DNA polymerase δ small sub
1633	DNA repair prot Rad2
1693	DNA repair prot Rad51 homolog
1383	DNA repair prot RadA
541	DNA repair Rad32 rel prot
1770	DNA replication initiator (Cdc21/Cdc54)
1624	DNA topoisomerase I
1208	DNA-dependent DNA polymerase fam B (PolB1)
208	DNA-dependent DNA polymerase fam B (PolB2)
550	DNA-dependent DNA polymerase fam X
764	endonuclease III
496	endonuclease III homolog
746	endonuclease III rel prot
1010	endonuclease IV
443	excinuclease ABC sub A
442	excinuclease ABC sub B
441	excinuclease ABC sub C
212	exodeoxyribonuclease
821	histone HMtA1
1696	histone HMtA2
254	histone HMtB
893	integrase-recombinase prot
501	m5C-specific restriction enzyme McrB rel prot
495	modification MTase, cytosine-specific
1210	Mrr restriction system rel prot
1315	mutator MutT prot
1336	mutator MutT prot homolog
122	mutator MutT rel prot
618	O6-methylguanine-DNA MTase
1342	8-oxoguanine DNA glycosylase
903	photoreactivation-associated prot
1312	proliferating-cell nuclear antigen
439	recombinase
1384	replication factor A rel prot*
1385	replication factor A rel prot*
240	replication factor C, large sub
241	replication factor C, small sub
164	single-stranded DNA exonuclease RecJ rel prot
494	thermonuclease precursor
940	type I restriction enzyme
942	type I restriction modification enzyme, sub M
941	type I restriction modification system, sub S

Transcription and RNA processing

203	ATP-dependent RNA helicase, eIF-4A fam
1415	ATP-dependent RNA helicase, eIF-4A fam
492	ATP-dependent RNA helicase, eIF-4A fam
656	ATP-dependent RNA helicase rel prot
1203	cleavage and polyadenylation specificity factor
1052	DNA-dependent RNA polymerase, sub A"
1051	DNA-dependent RNA polymerase, sub A'1a
297	DNA-dependent RNA polymerase, sub A'1b *
298	DNA-dependent RNA polymerase, sub A'1b *
299	DNA-dependent RNA polymerase, sub A'1b *
1050	DNA-dependent RNA polymerase, sub B'
1049	DNA-dependent RNA polymerase, sub B"
37	DNA-dependent RNA polymerase, sub D
264	DNA-dependent RNA polymerase, sub E'
265	DNA-dependent RNA polymerase, sub E"
1048	DNA-dependent RNA polymerase, sub H
42	DNA-dependent RNA polymerase, sub K

FIG. 4—*Continued.*

1317	DNA-dependent RNA polymerase, sub L
40	DNA-dependent RNA polymerase, sub N
1215	fibrillarin-like pre-rRNA processing prot
1190	N2,N2-dimethylguanosine tRNA MTase
1214	pre-mRNA splicing prot PRP31
1023	ribonuclease HII
683	ribonuclease PH
1695	RNase L inhibitor
1627	TATA-bind transcription initiation factor
1314	transcription elongation factor TFIIS
885	transcription initiation factor TFIIB
1054	transcription termination factor NusA
1678	transcription termination factor NusG
584	tRNA nucleotidyl-Tase
250	tRNA intron endonuclease
176	tRNA-guanine transglycosylase
Aminoacyl tRNA Synthetases	
1683	alanyl-tRNA Sase
1447	arginyl-tRNA Sase
226	aspartyl-tRNA Sase
51	glutamyl-tRNA Sase
1846	glycyl-tRNA Sase
244	histidyl-tRNA Sase
1375	isoleucyl-tRNA Sase
1508	leucyl-tRNA Sase
587	methionyl-tRNA Sase
770	phenylalanyl-tRNA Sase
742	phenylalanyl-tRNA Sase
1501	phenylalanyl-tRNA Sase α sub
611	prolyl-tRNA Sase
1122	seryl-tRNA Sase
1455	threonyl-tRNA Sase
251	tryptophanyl-tRNA Sase
1767	tyrosyl-tRNA Sase
767	valyl-tRNA Sase
Ribosomal proteins	
1119	ribosomal prot L10
1680	ribosomal prot L10a (*E.coli* L1)
16	ribosomal prot L11 (*E.coli* L5)
1679	ribosomal prot L12 (*E.coli* L11)
31	ribosomal prot L14
690	ribosomal prot L15
7	ribosomal prot L17 (*E.coli* L22)
38	ribosomal prot L18 (*E.coli* L17)
1610	ribosomal prot L18a
21	ribosomal prot L19
1323	ribosomal prot L21
13	ribosomal prot L23 (*E.coli* L14)
4	ribosomal prot L23a (*E.coli* L23)
257	ribosomal prot L24
14	ribosomal prot L26 (*E.coli* L24)
25	ribosomal prot L27a (*E.coli* L15)
2	ribosomal prot L3 (*E.coli* L3)
1053	ribosomal prot L30
1612	ribosomal prot L31
20	ribosomal prot L32
29	ribosomal prot L34 (*E.coli* L36)
9	ribosomal prot L35 (*E.coli* L29)
1310	ribosomal prot L36a
648	ribosomal prot L37
681	ribosomal prot L37a
1613	ribosomal prot L39
3	ribosomal prot L4
553	ribosomal prot L40
22	ribosomal prot L5 (*E.coli* L18)
24	ribosomal prot L7 (*E.coli* L30)
255	ribosomal prot L7a
5	ribosomal prot L8 (*E.coli* L2)
19	ribosomal prot L9 (*E.coli* L6)
1681	ribosomal prot Lp0 (*E.coli* L10)
1682	ribosomal prot Lp1
12	ribosomal prot S11 (*E.coli* S17)
1423	ribosomal prot S13 (*E.coli* S15)
36	ribosomal prot S14 (*E.coli* S11)
6	ribosomal prot S15 (*E.coli* S19)
18	ribosomal prot S15a (*E.coli* S8)
39	ribosomal prot S16 (*E.coli* S9)
803	ribosomal prot S17
34	ribosomal prot S18 (*E.coli* S13)
1616	ribosomal prot S19
23	ribosomal prot S2 (*E.coli* S5)
1059	ribosomal prot S20 (*E.coli* S10)
1055	ribosomal prot S23 (*E.coli* S12)
267	ribosomal prot S24
1309	ribosomal prot S27
268	ribosomal prot S27a
256	ribosomal prot S28
17	ribosomal prot S29 (*E.coli* S14)
8	ribosomal prot S3 (*E.coli* S3)
1593	ribosomal prot S3a
15	ribosomal prot S4
1056	ribosomal prot S5 (*E.coli* S7)
260	ribosomal prot S6
1199	ribosomal prot S7
207	ribosomal prot S8
35	ribosomal prot S9 (*E.coli* S4)
44	ribosomal prot Sa (*E.coli* S2)
10	ribosomal prot SUI1
Translation factors	
871	extragenic suppressor prot SuhB homolog
191	glutamine PRPP amido-Tase
1012	glutamyl-tRNA RDase
827	L-isoaspartyl prot carboxyl MTase
999	N-terminal acetylTase complex, sub ARD1
878	peptide chain release factor eRF, sub 1
535	peptide methionine sulfoxide RDase
1125	peptidyl-prolyl cis-trans isomerase
1338	peptidyl-prolyl cis-trans isomerase B
1745	prot disulphide isomerase
283	prot kinase
1414	prot-L-isoaspartate MTase homolog
259	translation initiation factor IF2 homolog
1485	serine/threonine prot kinase rel prot
1058	translation elongation factor EF-1 α
1185	translation elongation factor EF-1 α rel prot
1699	translation elongation factor EF-1 β
1057	translation elongation factor EF-2
1308	translation initiation factor eIF-2, α sub
1769	translation initiation factor eIF-2, β sub
261	translation initiation factor eIF-2, γ sub
1872	translation initiation factor eIF-2B, α sub
1004	translation initiation factor eIF-1A
869	translation initiation factor eIF-5A
RNA gene products	
1753	16S rRNA (1)
1888	16S rRNA (2)
1755	23S rRNA (1)
1890	23S rRNA (2)
1756	5S rRNA (1)
1891	5S rRNA (2)
1886	7S RNA
1292	RNaseP RNA
1754	tRNA-Ala (1) (ugc)
1889	tRNA-Ala (2) (ugc)
780	tRNA-Ala (ggc)
1304	tRNA-Arg (ccu)
1527	tRNA-Arg (gcg)
1344	tRNA-Arg (ucg)
1303	tRNA-Arg (ucu)
1276	tRNA-Asn (guu)
1046	tRNA-Asp (guc)
1269	tRNA-Cys (gca)
945	tRNA-Gln (cug)
946	tRNA-Gln (uug)
1274	tRNA-Glu (uuc)
790	tRNA-Gly (gcc)
791	tRNA-Gly (ucc)
1272	tRNA-His (gug)
1662	tRNA-Ile (gau)
638	tRNA-Leu (gag)
1720	tRNA-Leu (uaa)
1273	tRNA-Leu (uag)
1047	tRNA-Lys (uuu)
1275	tRNA-Met (1) (cau)
1572	tRNA-Met (2) (cau)
1293	tRNA-Met (i) (cau)
825	tRNA-Phe (gaa)
41	tRNA-Pro (ggg)
1044	tRNA-Pro (ugg)
33	tRNA-Ser (1) (gga)
1061	tRNA-Ser (2) (gga)
1887	tRNA-Ser (gcu)
1060	tRNA-Ser (uga)
750	tRNA-Thr (cgu)
1721	tRNA-Thr (ggu)
1043	tRNA-Thr (ugu)
1268	tRNA-Trp (cca)
1045	tRNA-Tyr (gua)
1432	tRNA-Val (cac)
824	tRNA-Val (gac)
1431	tRNA-Val (uac)
Unclassified functions	
1194	acetylpolyamine aminohydolase
1067	acetyl-Tase
1496	amidase
1474	D-arabino 3-hexulose 6-phosphate formaldehyde rel prot
1534	aryldialkylphosphatase rel prot
844	autotrophic growth prot
127	deoxyhypusine Sase
666	ethylene-inducible prot
1588	ferripyochelin-bind prot
234	γ-carboxymuconolactone decarboxylase
1515	GTP-bind prot
1621	GTP-bind prot, GTP1/OBG fam
858	GTP-bind prot, GTP1/OBG fam
765	GTP-binding prot Rab rel prot
1507	2-hydroxyhepta-2,4-diene-1,7-dioate isomerase
263	inorganic pyrophosphatase
724	Mtase rel prot
1329	Mtase rel prot
846	NAc-γ-glutamyl-phosphate RDase
1811	N-carbamoyl-D-amino acid amidohydrolase
1858	phage infection prot homolog
1183	pheromone shutdown prot TraB
1591	phosphonopyruvate decarboxylase
418	phosphonopyruvate decarboxylase rel prot
1206	phosphonopyruvate decarboxylase rel prot*
1207	phosphonopyruvate decarboxylase rel prot*
1453	[6Fe-6S] prismane-containing prot
911	probable surface prot
984	1,3-propanediol DHase
816	sporulation prot IVFB rel prot
1521	sugar fermentation stimulation prot
103	water channel prot
856	zinc metalloprotase

FIG. 4—*Continued.*

tral carbon metabolism in *M. jannaschii*; however, some of the missing genes in *M. jannaschii* have been identified in *M. thermoautotrophicum* and vice versa. Genes encoding all of the tricarboxylic acid cycle enzymes, except α-ketoglutarate dehydrogenase, have been identified in the *M. thermoautotrophicum* genome including two almost identical citrate synthetase genes, indicating a recent duplication event. Carbon monoxide dehydrogenase-encoding genes are present; however, unlike *M. jannaschii*, there is no evidence for a second pathway of CO_2 assimilation using ribulose bisphosphate carboxylase.

As in *M. thermoautotrophicum* Marburg (20), nitrogen fixation genes that encode a molybdenum-iron nitrogenase are clustered immediately downstream and transcribed in the same direction as the W-FMD-encoding *fwdHFGDACB* operon in strain ΔH. A second *nifH* is located at a remote site.

Based on database comparisons, *M. thermoautotrophicum* enzymes involved in amino acid, purine, pyrimidine, and vitamin biosynthetic pathways generally have sequences most similar to their bacterial homologs. Some enzymes required for these pathways do, however, appear to be missing, but since *M. thermoautotrophicum* synthesizes all of the products of these pathways from CO_2, H_2, and salts, it seems likely that the missing enzymes are present but have sequences sufficiently different from database sequences that they have not been recognized. Some of the unidentified ORFs conserved in both *M. thermoautotrophicum* and *M. jannaschii* presumably encode

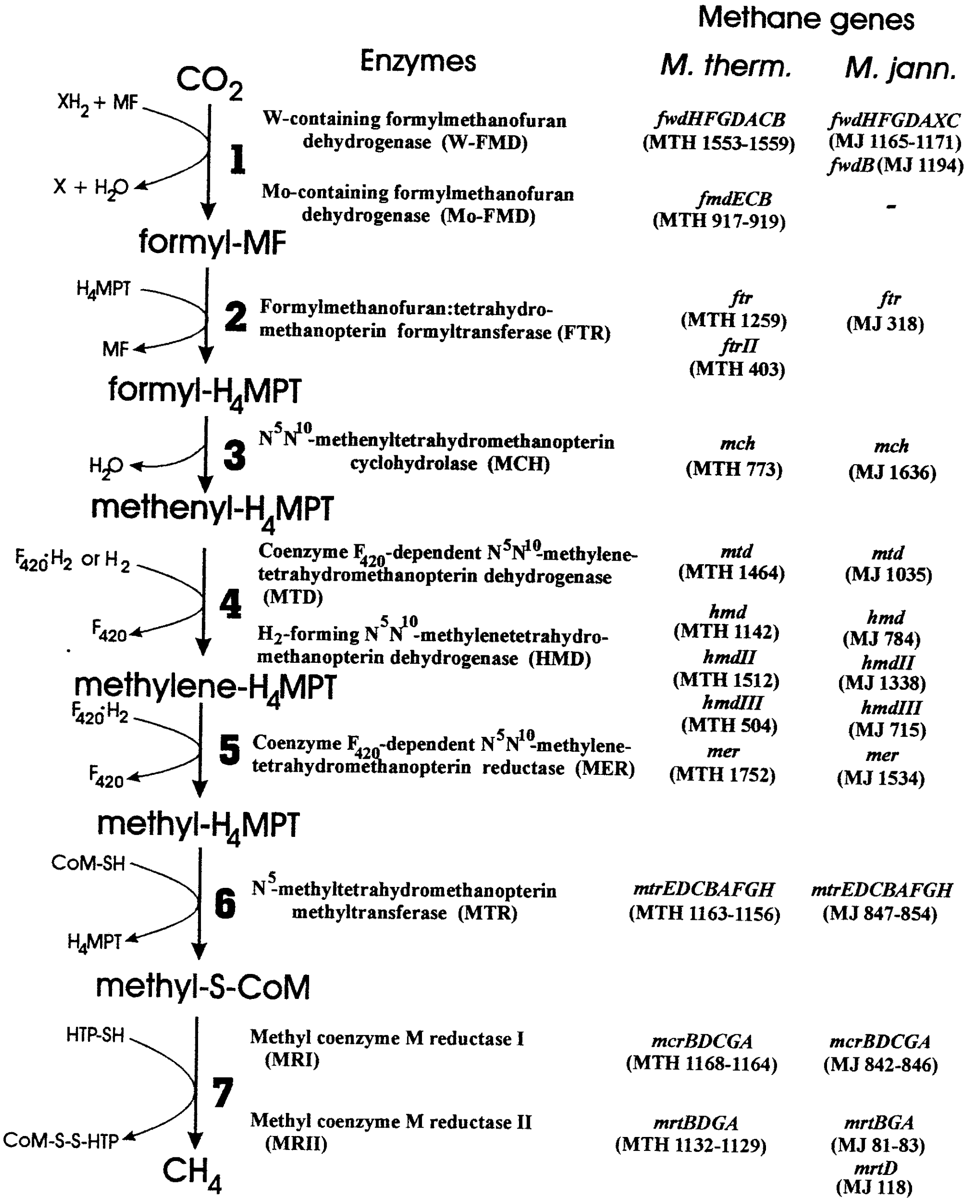
Methane genes
Enzymes
M. therm.
M. jann.
CO2
XH2 + MF
X + H2O
1
W-containing formylmethanofuran dehydrogenase (W-FMD)
fwdHFGDACB (MTH 1553-1559)
fwdHFGDAXC (MJ 1165-1171) fwdB(MJ 1194)
Mo-containing formylmethanofuran dehydrogenase (Mo-FMD)
fmdECB (MTH 917-919)
-
formyl-MF
H4MPT
MF
2
Formylmethanofuran:tetrahydromethanopterin formyltransferase (FTR)
ftr (MTH 1259) ftrII (MTH 403)
ftr (MJ 318)
formyl-H4MPT
H2O
3
N5N10-methenyltetrahydromethanopterin cyclohydrolase (MCH)
mch (MTH 773)
mch (MJ 1636)
methenyl-H4MPT
F420·H2 or H2
F420
4
Coenzyme F420-dependent N5N10-methylenetetrahydromethanopterin dehydrogenase (MTD)
mtd (MTH 1464)
mtd (MJ 1035)
H2-forming N5N10-methylenetetrahydromethanopterin dehydrogenase (HMD)
hmd (MTH 1142) hmdII (MTH 1512) hmdIII (MTH 504)
hmd (MJ 784) hmdII (MJ 1338) hmdIII (MJ 715)
methylene-H4MPT
F420·H2
F420
5
Coenzyme F420-dependent N5N10-methylenetetrahydromethanopterin reductase (MER)
mer (MTH 1752)
mer (MJ 1534)
methyl-H4MPT
CoM-SH
H4MPT
6
N5-methyltetrahydromethanopterin methyltransferase (MTR)
mtrEDCBAFGH (MTH 1163-1156)
mtrEDCBAFGH (MJ 847-854)
methyl-S-CoM
HTP-SH
CoM-S-S-HTP
7
Methyl coenzyme M reductase I (MRI)
mcrBDCGA (MTH 1168-1164)
mcrBDCGA (MJ 842-846)
Methyl coenzyme M reductase II (MRII)
mrtBDGA (MTH 1132-1129)
mrtBGA (MJ 81-83) mrtD (MJ 118)
CH4

FIG. 5. Biochemical pathway of H_2-dependent reduction of CO_2 to CH_4. The C_1 moiety is transferred from CO_2 via methanofuran (MF), tetrahydromethanopterin (H_4MPT), and coenzyme M (CoM-SH) into CH_4. The immediate source(s) of reductant (XH_2) used in step 1 is unknown (46, 60). The enzymes that catalyze each step, their encoding transcriptional units in *M. thermoautotrophicum* (*M. therm.*) and *M. jannaschii* (*M. jann.*), and their corresponding gene identification numbers are listed. The genes designated *ftrII*, *hmdII*, and *hmdIII* are homologs of *ftr* and *hmd*, respectively, but their gene products and functions in vivo remain to be identified.

enzymes that catalyze the synthesis of the unique cofactors employed in methanogenesis, an area of methanogen molecular biology that awaits investigation.

Cell envelope biosynthesis, protein secretion, solute uptake, and electron transport. The rod-shape of the *M. thermoautotrophicum* cell is maintained by a rigid layer of pseudomurein, a structure analogous but not chemically identical to the murein layer in the domain *Bacteria* (24). The presence of genes encoding sequences conserved in enzymes involved in murein and teichoic biosyntheses, bacterial shape determination (*mreB*), and cell division (notably *ftsZ* [63]) nevertheless suggests that cell envelope biosynthesis and the reconfiguration of the *M. thermoautotrophicum* cell during cell division do have features in common with their bacterial counterparts. Four genes encode proteins predicted to form the outer surface (S layer) of the *M. thermoautotrophicum* cell, and these include homologs of S layer proteins that are glycosylated in the hyperthermophilic methanogens *M. fervidus* and *Methanothermus sociabilis* (7).

The mechanisms of preprotein processing, membrane insertion, and protein secretion are widely conserved in biology, and ~12% of *M. thermoautotrophicum* ORFs encode polypeptides with N-terminal amino acid sequences consistent with signal peptides and ~20% have motifs indicative of membrane-spanning regions (see GTC web site for specific details). The majority of these proteins belong to the group for which functions could not be assigned, consistent with most biochemical studies of *M. thermoautotrophicum* having focused to date primarily on cytoplasmic enzymes. It appears that *M. thermoautotrophicum* may secrete a substantial number of proteins and may also have many membrane-associated proteins that await investigation. The *M. thermoautotrophicum* genome encodes homologs of the bacterial *secY* (preprotein translocase), *secD*, and *secF* (membrane-located protein export proteins) genes, a signal peptidase-encoding gene, and genes encoding homologs of eucaryal signal recognition particle proteins and of their associated RNA component (known as the 7S RNA). The same complement of protein processing and secretion genes is present in the *M. jannaschii* genome; however, *M. jannaschii* is motile and synthesizes flagellins that appear to be processed by a separate system (22). *M. thermoautotrophicum* is nonmotile and does not have *fla*, *mot*, or *che* gene homologs.

M. thermoautotrophicum is predicted to have a large number of transport systems for inorganic solutes, many of which have components related to the ABC family of ATP-dependent transporters. However, consistent with the autotrophic lifestyle, *M. thermoautotrophicum* does not appear to have many transport systems for organic molecules. There are also many genes that encode proteins predicted to have [4Fe-4S] centers, including nine ferredoxins and five polyferredoxins, some of which are probably membrane-located electron transport proteins. Similarly, a large family of genes is predicted to encode two-component sensor kinase-response regulator systems, and at least some of the sensor proteins appear to be membrane located (see below).

Two-component sensor kinase-response regulator systems. Although genes encoding two-component sensor kinase-response regulator systems have been documented in bacterial, archaeal, and eucaryal species, none were identified in the *M. jannaschii* genome. In contrast, the *M. thermoautotrophicum* genome appears to encode 14 sensor kinases, 9 response regulators, and 1 protein that is a fusion of a sensor kinase and a response regulator (MTH0901). Based on the presence of C-terminal blocks of conserved amino acids, designated H, N, G1, F, and G2, the sensor kinase encoded by MTH0444 is most similar to established bacterial sensor kinases, whereas the remaining *M. thermoautotrophicum* sensor kinases lack block F and contain a conserved region of 24 residues that has only limited sequence similarity to block H (Fig. 6). Except in the MTH1260 gene product, this region does, however, contain a histidyl residue appropriately located for autophosphorylation. An H block with a similar, atypical sequence has also been identified as a sensor kinase encoded in the *Synechocystis* sp. strain PCC6803 genome (24a) (Fig. 6). This *Synechocystis* protein also shares a number of other residues with the *M. thermoautotrophicum* sensors, including 12 amino acids located between blocks H and N, designated block E, consistent with the existence of a conserved subfamily of sensor kinases (Fig. 6). Although sequence conservation is very limited in the different two-component proteins in *M. thermoautotrophicum*, the MTH0292 and MTH0356 gene products are similar over their entire lengths, consistent with similar structures and the sensing of similar signals. Eight of the sensor kinases are predicted to contain N-terminal membrane-spanning helices within the region expected to function as the signal receptor, consistent with these being membrane-located proteins (Fig. 7).

The sensor kinase and response regulator genes MTH0901 and MTH0902 are adjacent and presumably form a single transcriptional unit, and one sensor kinase and four response regulator-encoding genes are clustered at position 378,000 (Fig. 1). MTH0549 is included in the list of response regulator genes although it does not encode the lysine-containing C-terminal region that is conserved in all documented response regulators (Fig. 6).

Translation machinery. There are two rRNA operons, designated *rrnA* and *rrnB*, separated by only ~110 kb in the *M. thermoautotrophicum* genome. Both have a 16S-23S-5S rRNA gene organization, with a tRNA^{Ala}(UGC) gene between the 16S and 23S rRNA genes. They encode 16S and 23S rRNAs with sequences that are 99.9 and 99.5% identical, respectively. The 7S RNA gene and a tRNA^{Ser} (GCU) gene are located immediately upstream of *rrnB*, which therefore may be part of a longer transcriptional unit. In both operons, the 16S and 23S rRNA genes are flanked by large inverted repeats capable of forming the bulge-helix-bulge secondary structure motif recognized by archaeal intron tRNA endonucleases (15, 27, 30, 61). This intron endonuclease probably catalyzes rRNA maturation in *M. thermoautotrophicum* as there is no evidence for a RNaseIII-like processing enzyme in the genome.

Thirty-nine tRNA genes have been identified. Ten are isolated, apparently forming single-gene transcriptional units; however, 16 are in eight operons that contain two tRNA genes, and 10 are in two five-tRNA gene operons. As in *M. jannaschii*, an elongator tRNA^{Met} (CAU) gene and the tRNA^{Trp} (CCA) gene contain introns located between positions 37 and 38 of the anticodon loop of the mature tRNAs. The tRNA^{Pro}(GGG) gene also contains an intron at this site plus a second intron uniquely located between positions 32 and 33. The presence of two introns in a single tRNA gene is unprecedented. All four *M. thermoautotrophicum* tRNA introns have flanking sequences capable of forming the bulge-helix-bulge secondary structure needed for archaeal tRNA intron processing.

Genes for members of all 20 tRNA families are present, although there is no Se-cys-tRNA(UCA) gene. Except for tRNA^{Ser} (GGA), elongator tRNA^{Met}(CAU), and the rRNA operon-associated tRNA^{Ala}(UGC) genes, there is only one copy of each tRNA gene. Two tRNAs are synthesized for amino acids encoded by four codons, one for codons ending in pyrimidines, and one for codons ending in purines, except for tRNA^{Val}(CAC) and tRNA^{Thr}(CGU) which translate only the codons with third-position guanines. For amino acids encoded by two codons, there is a single tRNA gene except that genes for both tRNAs^{Gln} are present. The six leucine and six serine codons are decoded by three tRNAs, and there are four arginine tRNA genes for the six arginine codons, one of which is specific for AGG. All three isoleucine codons are apparently translated by tRNA^{Ile}(GAU), although it is also possible that one of the two putative elongator methionine tRNAs decodes AUA isoleucine codons. Such a minor isoleucine-decoding tRNA species has been found in *Bacillus subtilis* that has a C*AU anticodon in which the first residue of the anticodon is replaced by the modified nucleotide, lysidine (31). *M. thermoautotrophicum* has tRNA^{Thr}(CGU) and tRNA^{Arg}(CCU) genes that are not present in *M. jannaschii*, presumably reflecting the higher %G+C content of the *M. thermoautotrophicum* genome and the different codon usage pattern.

Aminoacyl-tRNA synthetase genes have been identified for 16 tRNA families, but as in *M. jannaschii*, genes encoding asparaginyl-, glutaminyl-, cysteinyl- and lysyl-tRNA synthetases are not recognizable. As for organisms known to lack asparaginyl- and glutaminyl-tRNA synthetases, it is likely that *M. thermoautotrophicum* acylates tRNA^{Gln} and tRNA^{Asn} with glutamyl and aspartyl residues, respectively, which are then converted to glutaminyl and asparaginyl residues by amidotransferases. Consistent with this hypothesis, MTH1496, MTH1280, and MTH0415 are homologs of *gatA*, *gatB*, and *gatC*, which encode the three subunits of the glu-tRNA^{Gln} amidotransferase in *B. subtilis* (12).

The *M. thermoautotrophicum* r-protein-encoding genes were identified and named based on alignments with their rat homologs (70). Only 2 of the 61 r-protein-encoding genes, L12 and L10a, encode proteins with sequences more similar to

A

```
MTH1260   DELKNTINGL YRQIDRNLQL ITSIVNLQFP YIKDKDDYEL LRDTQNRL.. ...KSIRKAY EKLIYEG... ...SSDTINF GAYARSIVSG ILSTYSPEPG   242
MTH0360   EEKELLLREI HHRVKNNLQV ISSLLNLQSS YIDDPGITGV LRDSQGRI.. ...MSMSMIH EKLYRSG... ...NLADVDV RGYIEGLARS IMFSYMRPDQ   583
MTH0123   EEKEMLLREI NHRVKNNLMI ISSILNIQSR YVKDRDDLML FREAQSKA.. ...RAMAMLH ERLYTSG... ...KERRVDF GEYLRGLVRD LYHSFIQDSG   242
MTH0823   AEKELLLKEI HHRVKNNLMI ISSLLSLQSR QAKDRETMDL FRESENRT.. ...RSMVLIH ERLYRSE... ...DLKNIDL AEYLGRLASE IFRSYSADS.   550
MTH0902   REKEFLLSEI HHRVKNNLQL ISSLLRLQSR YIEDERSLEI FMECQNRV.. ...KSIALVH EKLYGSG... ...DMMVVNL AEYIEELLSE L.RNMCRGRD   351
MTH0459   REKEVLLREI HHRVKNNLQI VASLLSLQTA YTDNQETLNV LRDSQMRV.. ...RAMAVAH EKIYRSS... ...SLSMINV GDYLRAIAEE MTTLQSTGGL   384
MTH0292   REKEALLREL QHRVRNNLQI ITSLINIQLQ DADG.PVKEA LLATQTRV.. ...RAMTIIQ ESLYSTD... ...GYSSVHI ESCISRMTEH LKSLLGAHGV   449
MTH0356   SERDALLAEV HHRVKNNLQI IMSLLNIQAM NASE.EAREV LRDAQSRV.. ...RAMAILH ETIYDSG... ...NFTGVDM GSFITRLIER LVSAYGVYGI   458
MTH0174   REKEALLREV HHRVKNNFQV ISSLINLQLD DA...EDPAP LRDLQSRI.. ...QSMALVH ELLYESE... ...DLTSIDM GRYIERLTSS IVNSH..HNG   669
MTH0468   SEKELLLREV HHRVKNNLQI ISSLLNLQSL GTEGKEVRDV LMESQGRI.. ...KVMAMIH EHLYRSE... ...SLASINF RDYVERLVED IIISH..GS.   444
MTH0619   QEKELLLREI HHRVKNNLQI ISSLLSIQER QLESEELSDV LRESRERI.. ...RSIALVH EHLYRST... ...NLRTIRI RNYLNNILSK LSQGQTHGK.   635
MTH0985   RENEVLLSEI HHRVKNNLQI ISSLLSLQSH GIDDPSCRSL LSESQDRI.. ...RSMALIH EQLYRSG... ...DFSSIEF SSYASRLLKN LKRSYAPGK.   246
MTH0901   EEKEQLLREL HHRVKNNLQL IISLLSLQIR YIEDPGVEEF FRDYVNQL.. ...RSIAMIH ERAYPSS... ...GTYIIDF QEYVRSLSSH LISAHGRAS.   234
MTH1124   EANRTLLAEL HHRVKNNLQI ISSLISIQSS KM.PREHAEI MRSLQLRI.. ...KSIAVIH EMLLSSP... ...ESSSISF ASYVSGLTGY LRDMY..QSA   264
Syn.      EQKKVLLKEI HHRVKNNLQI MSSLLYLQFS KA.SPAIQQL SEEYQNRI.. ...QSMALIH EQLYRSE... ...DLANIDF SQYLKNLTHN ICQSYGCNTD   744
MTH0444   KELEAFAYSV SHDLRVPLRA IDGFSRILVE DYEDKLDDEG VR.ILGIIRD NTRKMGQLID DILLLSRAGR QEMNLAMLDM ...RELAE.S TYRELASQEE   244
Bac.      KDFVA...NV SHELKTPITS IKGFTETLLD GAME..DKEA LSEFLSIILK ESERLQSLVQ DLLDLSKIEQ QNFTLSIETF EPAKMLGEIE TLLKHKADEK   446

                        H-Block                                               E-Block

MTH1260   RVRLEMYFED VDMGL.DLAV PLGIILSELL SNSFRHAFTE DQDGRIRAVF MDKGDHYMLE VRDNGRGFPE GFDFE..... .......... .EADSLGLQL   325
MTH0560   RVDLRFEIED IKLNV DTIM PLGLIVNELV TNAFKYAFPD G.GGEVRVSL GRDGDGFLLT VADDGVGLPD DFNLD..... .......... .SLKSLGMLL   665
MTH0123   RIGLETDIDD AELDI.NTVV PLALIVNEVF TNAIKHGFPE GRGGIIRVSF KRSDDGYLLE IFDNGVGLPE DFDPM..... .......... .STSTMGMQL   325
MTH0823   RIRLKLEIDE LKVDV.ETAV PLGLIVNELL TNAVKHAFPD .GEGTVTVSL RKRNGTVTLE VSDDGAGFPE DIDWE..... .......... .SSPSLGLQL   632
MTH0902   TV.FRTELDE VRVGI.NTAV SIGLIVNELV TNAINHGIDS HGEVRITLSV ..SDGRGTLV VADNGCGLPQ DFEVS..... .......... .DSPGFGLKL   431
MTH0459   LVDLDVHYDD IMAEM.DRCI PLGLITNEII SNSIKHAFTG .DRGRIVISL KREDDLGILE ISDNGRGLPE DFNID..... .......... .ELESLGMQL   466
MTH0292   GFNIR...AD LRLNL.ETAM PLCLMVNELV TNAIKHAFPE .GKGEVHIEI DEGESGYHMR FADDGIGFSG E......... .......... .GEGT.GLKL   523
MTH0356   HFRVD...AD VRVNL.ETAI PLGLLINEAV TNSIRHAFPS .GEGSITVTM .ESDGLLYLR VEDDGTGMEG I......... .......... .PDGTVGLSL   532
MTH0174   EIEVEVAVGD ITLPL.ETAI PLGLIINELV TNSFKHAFT. .SGGMISVEL EEHGGEFTLT ITDNGVGLPP DFIIE..... .......... .DSDSLGLRL   750
MTH0468   SIRKVIEVDD IKPDI.DTAI PLGLIINELV TNSVKYAFPD .GTGSVTVRI RSHDDDVSLV VADDGVGLPE DIEPE..... .......... .NTDTLGLSL   526
MTH0619   DVRISSSIED LEFNL.ETSL PIGLMVNELV SNSLKHS... .GADNITVEL RSLNGTLELT VKDDGIGLES PEVLE..... .......... .KSGSMGWYL   714
MTH0985   NIELSLDTEN LKLSL.ETSI PLGLMLSELV TNALKHAFKG RDSGNIIVKF KKDGDYCVLE VRDDGVGFNE EKIRN..... .......... ..STSLGFRL   328
MTH0901   DVRVTVSGDT AELNM.DTAV PLALITAELI SNSLKHALS. .GGGEIHIEI RRFNGRHRLV YRDSGPGLPE DVSFP..... .......... .EGGSFGFRM   315
MTH1124   A.EFELDVPD VEFNI.ETAV PLGLIVGELV SNSLRHAFT. .DGGTIRISL EARDDGFILV VADNGGAFPI TSAFR..... .......... .NQPASAWSL   344
Syn.      SIKIKLLVEQ VKVPL.EQSI PLGLIIQELV SNALKHAFPT .TEGEISIKF TSMNSHYSLQ VWDNGVGISR DIDLE..... .......... .NTDSLGMQL   806
MTH0444   GRSIEFSVAD LPPAMADRAL .MGQVMGNLL SNAIKFT.RD RDPAVIEVGY MDGGDEHTYY VKDNGAGFDM KYASKLFGLF QRLHSQ..EE FEGTGVGLSI   340
Bac.      GISLHLNVPK DPQYVSGDPY RLKQVFLNLV NNALTYT.PE GGSVAINVKP REKDIQIE.. VADSGIGIQK EEIPRIFERF YRVDKDRSRN SGGTGLGLAI   543

                                            N-Block                              G1-Block      F-Block            G2-Block

MTH1260   VRNLINQIEA RVDY..KLSP GTCFRVRVLK P*                354
MTH0360   VRNLTEQLNG ELEY...TSN GGAEFRVRFS EIQYKKRF*         700
MTH0123   IRSLSEQMNG DLKI...ESH GGTRVSIEFR DWNH*             356
MTH0823   VRKPH*                                             637
MTH0902   VNFMLRRVNG SV.V..AENR DGAVFTVTFD AGGE*             462
MTH0459   VSNLVMQIGG ELEY..G.NR DGAFFRVTFP LE*               495
MTH0292   VRILVEQLEG DLKILVDEEK GGTEILVPFR ELQYRERT*         561
MTH0356   MRALADQLEG ELEI...ESD QGTVVSLRFR ELEYMKRT*         567
MTH0174   VAGLVDQIDG TLEV...SGE DGTRFRLTFG VVPYRRRV*         785
MTH0468   VSILTEQLDG TLTI...RRD HGTEFRISFP V*                554
MTH0619   IRALTDQLDG ELKI...ETE DGLSVSLRFR ELGYRERY*         749
MTH0985   VEILTEQLDG SLTY...SGE NGGLFRIRFR EPLYKDRLTN*       365
MTH0901   MDNLAGQLGG HIKV..ESSD DGVVFIFEFF EQFYADRIT*        352
MTH1124   LRIW*                                              348
Syn.      IYSLTEQLQG ELHY...EYV GGAQFGLEFS L*                834
MTH0444   VQRIIKRHGG RVWG.EGKVD GGATIYFTLP KVVK*             373
Bac.      VKHLIEAHEG KIDV.TSELG RGTVFTVTLK RAAEKSA*          579
```

B

```
MTH0549              VTSILIVED EALIAADLRT RL........ ....ERMGYE VVGAAGDGRE ALKLIAEKRP DLVLMDDGTG CFSHSIL*                64
MTH0440            M SPTSLLVVED ESIVAMDIKH RA........ ....EGLGYR VVGIAASGED AIKLAREEKP DLVLMDIVLK GEMDGIEAAE VIREE...MD   76
MTH0447              MAKILVVED EAIVAMGITH KL........ ....ESMGHR VVETVSTGKD AIMACKVHEP DLVLMDIVLK GEMDGIEAAR RIRDQ...FN   74
MTH1764        LRWTM SRAKVMVVED ESIVAIDISQ RL........ ....QSLGYE VTATVSSGEK AVEMAEKTRP DIILMDIVLK GEMGGIEAAE EINKR...MK   80
MTH1607   MSHYPCGDFH MGVRILLVED EAITAMDLQR KL........ ....EFWGYD VVGVAYSGET AVELAQKHHP DLILMDIVLK GPLNGVDAAK EIRS....LD   84
MTH0901            M MRGRVVIVED EELVAQDIRY IL........ ....EDAGYE VAAIFHSAED LLESLEKLEP DSIIMDIMLE GELDGIDAAR IIKKK...MD   76
MTH0457              MPSALVVED EAVTSLELLR LL........ ....ESWGYE TVSV.KTGED AIETALRMKP DVILMDIVLP SDVDGVTAAR AIKKE...MD   73
MTH0548          VRG PVPGVLIVED EAIIAADLKQ KL........ ....ENAGFR VLGVHDTGEG AIAAASELEP DVVIMDVYIR GEMDGIKAAE RIQER...YG   145
MTH0446           MS EKLKVLILED VPLDAELVIR ELQ....... ....RDGIEF EHLTVDSEDS FRRALEEFSP DIILADHALP S.FDGVSALR IVREN...YD   77
MTH0445           MM TDADILLVED NPTDAELTIR ALKKNNLANK LHWVKDGAEA LDYIFASGSY SDRDPENL.P KLILLDLRMP K.VDGLEVLQ EIKRNDSTSK   90

MTH0440   IPVVYLTAYS DEKTLSRAKL TGPFGYIIKP FEDRELHSAI EVALY..... .....KHKMD   136
MTH0447   IPIIYLTAYA DEEMLTRAKV TEPYGYIVKP FKSSELNANI EMAIY..... .....RHR..   134
MTH1764   VPIVYITAYS DEETLRRAKV TGPFGYIIKP FEDRELHSVI EVALY..... .....KHELE   140
MTH1607   IPVVFLSAHS EGSTMERARE VEPYGYLIKP FDEKELLFSI ELAVQ..... .....RHRSQ   144
MTH0901   IPVLYLTAYS SDEIVKRARE TEPYAYILKP FHERDIRVNL EMALY..... .....KHEAK   136
MTH0457   IPLIFITAYS SREVFERAAE VEPEAYLLKP FNSRELGYAM ELAIY..... .....KNRIQ   133
MTH0548   IPVIYLTAYS DDATLSRILE SEPYGYLLKP LNTEQLQAEI EVVLE..... .....NLRTT   205
MTH0446   IPFIFVSGKI GEEFAVDMLK AGATDYVLK. .......NNL SKLPLAFRRA LQEAEEERKI   129
MTH0445   IPVVVLTSSK EDRDIVESYK LGVNSYVSKP VEFDEFISAV STLGFYWMII NQPPE*      145
```

FIG. 6. Alignments of the conserved regions in putative sensor kinase (A) and response regulator (B) proteins in *M. thermoautotrophicum* ΔH. The alignments were generated by PILEUP (17), and residue positions are listed to the right. Completely conserved residues are shaded black, and regions with ≥75% sequence similarity are shaded gray. In panel A the *M. thermoautotrophicum* sequences have been grouped and aligned to emphasize their similarity to the putative sensor protein encoded by *Synechocystis* sp. PCC6803 (ethylene sensor response protein, GenPept gene identification no. g162472) and to the PhoR sensor of *B. subtilis* (Swiss-Prot P23545). The sensor kinase motifs H, N, G1, F, and G2, and a previously unrecognized block of conserved amino acid residues designated motif E, are identified below the sequences.

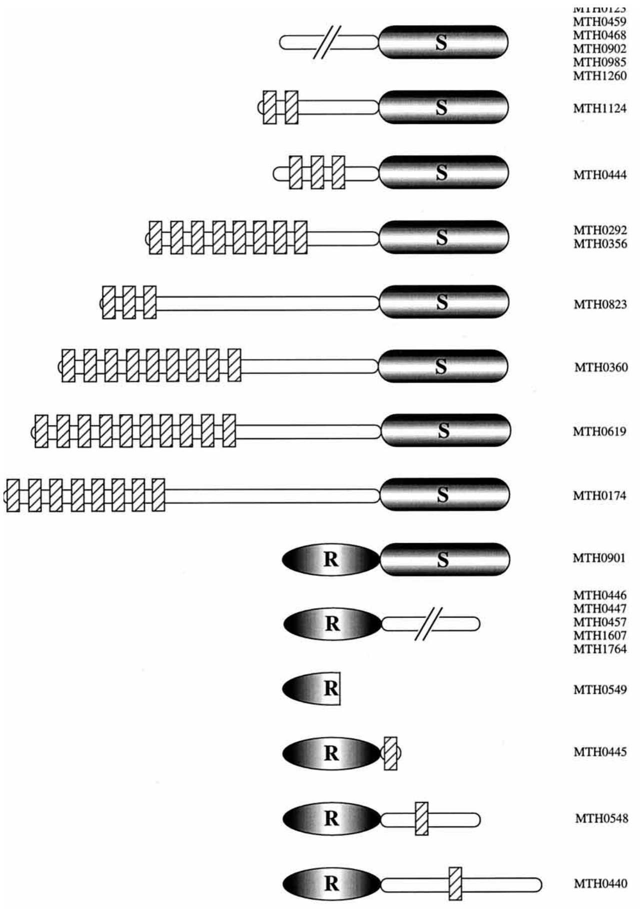

FIG. 7. Structures of putative sensor kinases and response regulator proteins in *M. thermoautotrophicum* ΔH. Conserved domains identified in the sequence alignments in Fig. 6A and B are shown as gray blocks labeled S and R, respectively. Open boxes indicate nonconserved regions with variable lengths (-//-), and hatched boxes identify membrane-spanning helices predicted by TMpred (www.microbiolgy.adelaide.edu.au/learn/tmpred.htm).

their bacterial homologs (L11 and L1, respectively) than to their eucaryal homologs. Seven genes in the *M. thermoautotrophicum* genome encode r-proteins that have eucaryal but not bacterial homologs, and homologs of 23 *E. coli* r-protein-encoding genes have not been identified in the *M. thermoautotrophicum* genome.

RNA-processing enzymes. Genes encoding the RNA component of RNaseP, a tRNA intron endonuclease, a tRNA nucleotidyltransferase, and proteins associated with the modification of nucleotides in tRNAs and rRNAs have been identified. The two physically adjacent genes MTH1214 and MTH1215 respectively encode homologs of the eucaryal nuclear proteins PRP31 and fibrillarin. Fibrillarin associates with small nucleolar RNAs in complexes that participate in endonuclease processing of rRNA primary transcripts and in the addition of 2′O-methyl groups to rRNAs (26). PRP31 is required for mRNA processing and *prp31* is an essential gene in yeast (65). MTH0032 is predicted to encode a homolog of a centromere-microtubule binding protein whose precise function in *Eucarya* remains to be determined, although members of this family include the nucleolar protein NAP57 and bacterial proteins involved in pseudouridylation. The conservation of the same RNA processing enzymes in *M. thermoautotrophicum* and *M. jannaschii*, and the fact that archaeal and eucaryal tRNA intron endonucleases employ a conserved biochemistry, indicates that these RNA processing systems probably predate the divergence of the *Archaea* and *Eucarya.*

DNA-dependent RNAP and transcription factors. Genes encoding the large A′, A″, B′, and B″ and small D, E′, E″, H, I, K, L, and N subunits of the *M. thermoautotrophicum* RNA polymerase (RNAP) have been identified, but homologs of the *Sulfolobus acidocaldarius* G and F subunit-encoding genes are not present. The sequences of these large RNAP subunits and of subunit D are more similar to their eucaryal than to their bacterial counterparts, and there are only eucaryal homologs of the E′, E″, H, K, L, and N subunits (29). As in *M. jannaschii*, the *M. thermoautotrophicum* homolog of the *S. acidocaldarius* subunit E-encoding gene is split into *rpoE1* and *rpoE2* genes that encode E′ and E″ subunits, respectively. However, unlike *M. jannaschii*, the *M. thermoautotrophicum* genome contains a second subunit A′ gene, designated *rpoA1b*, located ~500 kb from the *rpoA1a* gene in the *rpoHB2B1A1aA2* operon. The *rpoA1a* and *rpoA1b* genes have sequences that are ~2.6-kb long and 82% identical, but except for 10 bp immediately preceding the TTG start codons that contain RBSs, the genes are not flanked by conserved sequences. The *rpoA1a* gene encodes a single 98-kDa polypeptide whereas the *rpoA1b* sequence contains frameshifts suggesting a pseudogene, frameshifting, or possibly the synthesis of three separate polypeptides with sizes of 10, 15, and 75 kDa. The frameshifts have been confirmed by PCR amplification from genomic DNA and resequencing, and cotranscription of *rpoA1b* with the unidentified upstream gene (MTH0296) has also been documented (13).

Transcription initiation in *Archaea* follows the eucaryal paradigm but with a reduced preinitiation complex (47). Consistent with this, the *M. thermoautotrophicum* genome encodes a TATA-binding protein and transcription factors TFIIB and TFIIS but no homologs of the eucaryal general transcription factors TFIIA, TFIIF, and TFIIH that form part of most preinitiation complexes assembled in *Eucarya.*

DNA-dependent DNA polymerases. *M. thermoautotrophicum* apparently contains two DNA polymerases, a member of the X family (synonymous to the polymerase β family) of DNA repair enzymes, and an archaeal group I B-type DNA polymerase. *M. jannaschii*, in contrast, contains only a B family DNA polymerase encoded by a gene with two inteins. Family X polymerases are usually ~350 residues long with common motifs that form the active site for nucleotidyl transfer (52). These motifs are present in the MTH0550 gene product, but this polypeptide also has an ~200-amino-acid C-terminal extension with a sequence similar to sequences contained in several bacterial proteins of unknown function, including a *B. subtilis* protein that also has an N-terminally located PolX domain (68).

The *M. thermoautotrophicum* B-type DNA polymerase is typical in having exonuclease and polymerase domains; however, unlike other archaeal B-type polymerases that are single polypeptide enzymes (16), the *M. thermoautotrophicum* ΔH polymerase apparently contains two polypeptides encoded by two genes, *polB1* and *polB2*, that are separated by ~650 kb. Although DNA polymerases with physically separate exonuclease and polymerase domains, encoded by separate genes, have been described previously (21), the break-site in the *M. thermoautotrophicum* enzyme is uniquely located within the polymerase domain. The two PolB1 and PolB2 polypeptides are predicted to contain 586 (68.0 kDa)- and 223 (25.5 kDa)-amino-acid residues, respectively, which if added together would give a length very similar to that of the single polypeptide archaeal B-type polymerases. The DNA polymerase purified from *M. thermoautotrophicum* Marburg was reported to be a single polypeptide with a molecular mass of ~72 kDa, although DNA polymerase activity was also associated with an ~38-kDa polypeptide that was considered to be a degradation product of the ~72-kDa polypeptide (28).

Mobile genetic elements. There is no evidence for typical insertion sequence (IS) elements, prophages, or homing endonucleases (3), although the *M. thermoautotrophicum* genome does appear to encode one intein within the alpha chain of ribonucleoside-diphosphate reductase (MTH652). This intein, designated Mth RIR1, has readily recognizable protein-splicing motifs but lacks an endonuclease domain, and with only 134 amino acid residues, it is the shortest intein so far identified (40). Although the *M. jannaschii* genome does not appear to encode a ribonucleoside diphosphate reductase, genes homologous to MTH652 are present in *Thermoplasma acidophila* (59) and *Pyrococcus furiosus* (49). There is no intein in the *T. acidophila* homolog whereas the *P. furiosus* ribonucleoside diphosphate reductase alpha subunit gene encodes two inteins, one integrated at the same position as the Mth RIR1 intein (Fig. 8). The sequence of the Pfu RIR1 intein is only 31% identical, over 103 residues, to that of the Mth RIR1 intein, and it does have an endonuclease domain. Inteins with only limited sequence similarity, but integrated at identical sites, have also been identified in the DnaB proteins of a cyanobacterium and a red algal chloroplast (42).

Repetitive sequences. A list of the repetitive sequences present in the *M. thermoautotrophicum* genome, including gene duplications, is available on the GTC web site. Two remarkable repeats, R1 and R2, which are separated by ~480 kb, orientated in opposite directions, and 3.6 and 8.6 kb in length, respectively, belong to a family designated the LS_n repeat family. R1 and R2 contain a 372-bp long repeat (LR) sequence, which is 88% identical in R1 and R2, followed by 47 and 124 copies, respectively, of the same 30-bp short repeat (SR) sequence. These SR sequences are separated by unique sequences 34 to 38 bp in length, and larger repeating units consisting of blocks of several SR sequences plus their intervening sequences are detectable within R1 and R2.

There are also 18 LS_n repeats in the *M. jannaschii* genome, with LR sequences unrelated to the LR sequences in *M. thermoautotrophicum* but with SR sequences that are 76% (23 of 30 nucleotides) identical to the *M. thermoautotrophicum* SR sequence. Although the number of SR elements per LS_n repeat is smaller in *M. jannaschii*, ranging from 1 to 25, the total number of SR sequences is very similar in both genomes.

Plasmid-related sequences. Although *M. thermoautotrophicum* ΔH does not contain extrachromosomal DNA elements, plasmids have been isolated and sequenced from closely related thermophilic *Methanobacterium* species, including plasmid pME2001 from *M. thermoautotrophicum* Marburg (6) and the related plasmids pFV1 and pFZ1 from *Methanobacterium thermoformicicum* THF and Z-245, respectively (33). There are no pME2001-related sequences in the *M. thermoautotrophicum* ΔH genome but pFV1 and the strain ΔH genome both contain one copy of a sequence that is present in several copies in the genomes of other thermophilic methanobacterial isolates (35). In addition, five pFV1 genes (*orf1*, *orf4*, *orf5*, *orf9*, and *orf10*) have homologs in the *M. thermoautotrophicum* ΔH genome (MTH1412/MTH1599, MTH0350, MTH1074, MTH0471, and MTH0764/MTH0496, respectively). Three of these genes (*orf1*, *orf4*, and *orf5*) also have homologs in pFZ1, and the *orf10*-related genes MTH0764 and MTH0496 encode endonuclease III homologs. MTH1074 encodes 1,474 amino acid residues including 10 repeats of a block of ~90 residues, and this gene therefore appears to be an expanded version of *orf5*, which encodes 499 amino acid residues with four of the ~90-bp repeats. Similar repeats are present in a 60-kDa outer membrane protein of *Chlamydia psittaci* (64). These methanogen proteins may also be membrane located, possibly with a similar function, as they have N-terminal amino acid sequences that resemble bacterial signal sequences. The plasmid-encoded *orf1* gene products are likely to be involved in plasmid replication (33) as they are members of the Cdc18-Cdc6 family of proteins that directs the initiation of DNA replication in *Eucarya* (32). The *M. thermoautotrophicum* genome encodes two members of this family and a homolog of the eucaryal DNA replication initiation protein Cdc54. Cdc6-encoding genes are not present in the *M. jannaschii* genome, although genes encoding proteins related to other eucaryal DNA replication and DNA repair enzymes are conserved in both genomes and both genomes encode DNA restriction and modification systems.

```
            248                       265
Mth RIR1   -dpgilfedrinrynptpql..grieatnpCVSGDTIVMTSGGPRTVAELE--15aa--
            |||::| | ||| |       | | |||||| ||| ::|  |   : |:
Pfu RIR1   -dpgviffdvinrrnvlkkakggpiratnpCVVGDTRILTPEGYLKAEEIF--56aa--
            886                       914

            302
Mth RIR1    PSGFFRTCERDVYDLRTREGHCLRLTHDHRVLVMDGGLEWRAAGELERGDRLVM--9aa--
            |:  ::   : |  ::|:||: :  | ||:::  :|   |: :|:|: ||::::
Pfu RIR1    PAYVWKVGRKKVARVKTKEGYEITATLDHKLMTPEG...WKEVGKLKEGDKILL--219aa-
            992

            365                                400
Mth RIR1    LATFRGLRGAGRQDVYDATVYGASAFTANGFIVHNcgeqpllthescnlgsvnlslmv-
            ::|   :   | : ||| ||     : :|||: |||||:||  :|||||:|:||: :|
Pfu RIR1    IVTVESVEVLGEEIVYDFTVPNYHMYISNGFMSHNcgeeplyeyescnlasinlakfv-
            1262                               1297
```

FIG. 8. Alignment of RIR1 intein sequences and their integration points in ribonucleoside diphosphate reductase in *M. thermoautotrophicum* (Mth) and *P. furiosus* (Pfu) (gil1688292). Intein sequences are shown in uppercase letters with the ribonucleoside diphosphate reductase flanking sequences in lowercase letters. The numbers above and below the sequences indicate residue positions in the full-length ORFs (host protein and intein). The numbers of residues in the unaligned intein regions are indicated between the aligned regions. Lines mark alignment of identical residues and colons mark conservative substitutions. Gaps introduced to optimize the alignment are indicated by dots.

DISCUSSION

This is the seventh publication reporting the complete sequence of a procaryotic genome, and trends are now becoming apparent. In each case, ~90% of the genome is predicted to encode gene products, the average ORF length is ~1 kb, and a complement of tRNA genes is present which is adequate to decode all sense codons. Many genes appear to be organized into multigene transcriptional units, inaccurately but conveniently designated operons, and RBSs precede most ORFs. The relative locations of genes and operons within these genomes show little conservation, consistent with most gene expression being coordinated in *trans* by soluble intracellular signals. The origins of DNA replication have not been identified in the two methanogen genomes; however, there is no detectable bias in gene orientation and the lack of conservation of gene location suggests that genome position is not a generically important parameter for gene expression. There is also little evidence for the direction of transcription being consistently coordinated with or against the direction of DNA replication.

M. thermoautotrophicum seems to have an unusually low number of mobile DNA elements. There are no recognizable prophages, plasmids, or IS elements and only one, very short, intein. By contrast, *M. jannaschii* has two plasmids, 19 inteins, and 11 members of an IS family (9, 43). The difference in the abundance of inteins might be correlated with the absence of homing endonucleases in *M. thermoautotrophicum*. These enzymes have been proposed to drive the mobility of prokaryotic introns and inteins (2), and homing endonucleases are encoded in *M. jannaschii* as independent genes (41a) and within almost all of its inteins (40, 43), but they do not occur in *M. thermoautotrophicum*.

M. thermoautotrophicum synthesizes all of its cellular components and conserves energy from just CO_2, H_2, and salts but, nevertheless, has a genome that is only ~40% the size of the *E. coli* genome and only three times the size of the *Mycoplasma genitalium* genome. Considerable discussion has been focused on the concept of identifying the minimum number of genes needed for a minimal cell but identifying the minimum number of genes needed to constitute a fully independent autotrophic cell is an equal challenge and potentially has more practical value. When compared with the similar sized genome of *M. jannaschii*, it appears that both methanogens still harbor more genes than they need for their lithoautotrophic lifestyles. Both contain duplicated genes which presumably provide nonessential metabolic flexibility, and 20% of *M. thermoautotrophicum* genes do not have homologs in *M. jannaschii* whereas ~15% of *M. jannaschii* genes do not have homologs in *M. thermoautotrophicum*. These two methanogens do have very different cell envelope structures (24), so some of the species-specific genes probably are essential for the methanogen in which they exist but this is unlikely to be predominantly the case. There are, for example, 24 two-component system genes in *M. thermoautotrophicum*, none of which are present in *M. jannaschii*, and both genomes encode several different DNA repair and DNA restriction-modification systems and a large number of small solute transport systems.

In the context of this initial report, discussing every gene, all the novelties, and all the questions raised by the genome is impossible and inappropriate. A few of the interesting differences between *M. thermoautotrophicum* and *M. jannaschii* do, however, warrant noting. *M. thermoautotrophicum* has a *grpE dnaJ dnaK* heat shock operon in addition to genes that encode an archaeal proteasome-chaperonin structure, and it has additional DNA repair enzymes, DNA helicases, nitrogenase subunits, an Fe-Mn superoxide dismutase, a ribonucleotide reductase, three coenzyme F_{390} synthetases, and proteases that are absent in *M. jannaschii*. Unique features predicted for *M. thermoautotrophicum* are the presence of two Cdc6 homologs, an archaeal B-type DNA polymerase with a novel subunit structure, the possibility of two RNAP A′ subunits, hinting at a previously unsuspected mechanism of gene selection, and two introns in the same $tRNA^{Pro}$(CCC) gene, which establishes a precedent and a new location for tRNA introns.

Phylogenetics is dominated by the small subunit rRNA (ssu rRNA) tree which groups organisms into three domains, *Bacteria*, *Archaea*, and *Eucarya* (39). Inherent in this concept is the idea that these groups must have other group-specific features, and the −10 and −35 structure of the promoter and promoter recognition by sigma factors in *Bacteria*, ether-linked lipids and methanogenesis in *Archaea*, and the nuclear membrane and the complex pathways of mRNA processing in *Eucarya* are frequently cited as examples. Phylogenetic trees based on the sequences of conserved enzymes, however, are often not consistent with the ssu rRNA tree, and defining a gene product as bacterial, archaeal, or eucaryal because its sequence is most similar to the sequence of a gene product previously established from a bacterial, archaeal, or eucaryal species based on the ssu rRNA tree promotes the idea that this tree is valid for that gene product. Based on the genome sequences available, it appears that it might now be more appropriate to consider phylogenetic arguments and analyses separately for metabolic pathways and for components of the genetic information storage, retrieval, and expression systems. Are there biochemical pathway phylogenies that correlate precisely with the ssu rRNA tree or is this tree only congruent with the phylogenies of genes that encode products involved in genetic information processing? Most proteins in the two methanogens, and almost all of the metabolic pathway enzymes, have sequences that are more similar to sequences in other *Archaea* and/or in *Bacteria* than in *Eucarya*. However, the presence of genes that encode homologs of proteins that exist only in *Eucarya*, namely TATA-binding and transcription factor IIB proteins, histones, DNA replication factors, transcript-processing systems, and ribosomal proteins, reinforces the conclusion that these functions must have evolved in a lineage separate from the bacterial lineage that gave rise only to the *Archaea* and *Eucarya*. Lateral transfer and assimilation of all of these different levels of genetic information processing seems very unlikely, and their correlation with the ssu rRNA tree argues that this tree is valid as an indicator of the underlying phylogeny of whole organisms. Data from genome-sequencing projects should now make it possible to superimpose on this tree the phylogenies of all the other subcellular components and biochemical pathways. For example, it should be possible to track the phylogenetic history of nitrogen fixation, which is conserved in *Archaea* and *Bacteria* but which does not appear to exist in *Eucarya*. Was nitrogen-fixing ability lost in the eucaryal lineage after divergence from the archaeal lineage or did nitrogen fixation evolve in one lineage, say in the bacterial lineage, and was then transferred to only the archaeal lineage? This latter scenario would be analogous to the chloroplast endosymbiont theory often evoked to explain why photosynthesis occurs in *Bacteria* and *Eucarya* but not in *Archaea*. Sequencing more genomes will address and resolve these fundamentally important and very interesting issues.

ACKNOWLEDGMENTS

This work was supported by research grant DE-FG02-95ER-61967.

We thank T. Conway (OSU) for the analysis of metabolic pathway genes and D. Graham (U. Illinois) for providing an independent evaluation of the *M. thermoautotrophicum* genome sequence.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. F. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Belfort, M., Reaban, M. E., Coetzee, T. and J. Z. Dalgaard.** 1995. Prokaryotic introns and inteins: a panoply of form and function. J. Bacteriol. **177:**3897–3903.
3. **Belfort, M. and R. Roberts.** 1997. Homing endonucleases—keeping the house in order. Nucleic Acids Res. **25:**3379–3388.
4. **Bhagwat, A. S., and M. McClelland.** 1992. DNA mismatch correction by very short patch repair may have altered the abundance of oligonucleotides in the *Escherichia coli* genome. Nucleic Acids Res. **20:**1663–1668.
5. **Bodenteich, A., S. Chissoe, Y. F. Wang, and B. A. Roe.** 1994. Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. *In* M. Adams, C. Fields, and J. C. Venter (ed.), Automated DNA sequencing and analysis techniques. Academic Press, San Diego, Calif.
6. **Bokranz, M., A. Klein, and L. Meile.** 1990. Complete nucleotide sequence of plasmid pME2001 from *Methanobacterium thermoautotrophicum* (Marburg). Nucleic Acids Res. **18:**363.
7. **Brockl, G., M. Behr, S. Fabry, R. Hensel, H. Kaudewitz, E. Biendl, and H. König.** 1991. Analysis and nucleotide sequence of the genes encoding the surface-layer glycoproteins of the hyperthermophilic methanogens *Methanothermus fervidus* and *Methanothermus sociabilis*. Eur. J. Biochem. **199:**147–152.
8. **Brown, J. W., C. J. Daniels, and J. N. Reeve.** 1989. Gene structure, organization and expression in archaebacteria. Crit. Rev. Microbiol. **16:**287–338.
9. **Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, E. A. Presley, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, M. A. Hurst, K. M. Roberts, B. P. Kaine, M. Borodovsky, H.-P.

Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273:**1058–1073.

10. **Church, G. M., and S. Kieffer-Higgins.** 1988. Multiplex DNA sequencing. Science **240:**185–188.

11. **Church, G. M., G. Gryan, N. Lakey, S. Kieffer-Higgins, L. Mintz, M. Temple, M. Rubenfield, L. Jaehn, H. Ghazizadeh, K. Robison and P. Richterich.** 1994. Automated multiplex sequencing, p. 11–16. *In* M. Adams, C. Fields, and J. C. Venter (ed.), Automated DNA sequencing and analysis techniques. Academic Press, San Diego, Calif.

12. **Curnow, A. W., K. Kwang-won, R. Yuan, S.-I. Kim, O. Martins, W. Winkler, T. M. Henkin, and D. Söll.** Glu-tRNA^{Gln} amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. Proc. Natl. Acad. Sci. USA **94**, in press.

13. **Darcy, T. J., R. M. Morgan, J. Nölling, and J. N. Reeve.** 1997. Unpublished results.

14. **DiMarco, A. A., K. A. Sment, J. Konisky, and R. S. Wolfe.** 1990. The formylmethanofuran: tetrahydromethanopterin formyltransferase from *Methanobacterium thermoautotrophicum* ΔH. J. Biol. Chem. **265:**472–476.

15. **Durovic, P., and P. P. Dennis.** 1994. Separate pathways for excision and processing of 16S and 23S rRNA from the primary rRNA operon transcript from the hyperthermophilic archaebacterium *Sulfolobus acidocaldarius*: similarities to eukaryotic rRNA processing. Mol. Microbiol. **13:**229–242.

16. **Edgell, D., H.-P. Klenk, and W. F. Doolittle.** 1997. Gene duplication in evolution of archaeal family B DNA polymerases. J. Bacteriol. **179:**2632–2640.

17. **Genetics Computer Group.** 1995. Wisconsin package version 8.1. Genetics Computer Group, Madison, Wis.

18. **Halboth, S., and A. Klein.** 1992. *Methanococcus voltae* harbors four gene clusters potentially encoding two [NiFe] and two [NiFeSe] hydrogenases, each of the cofactor F$_{420}$-reducing or F$_{420}$-non-reducing types. Mol. Gen. Genet. **233:**217–224.

19. **Henikoff, S., Henikoff, J. G., Alford, W. J. and S. Pietrokovski.** 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene **163:**17–26.

20. **Hochheimer, A., R. A. Schmitz, R. K. Thauer, and R. Hedderich.** 1995. The tungsten formylemthanofuran dehydrogenase from *Methanobacterium thermoautotrophicum* contains sequence motifs characteristic for enzymes containing molybdopterin dinucleotide. Eur. J. Biochem. **234:**910–920.

21. **Ito, J., and D. K. Braithwaite.** 1991. Compilation and alignment of DNA polymerase sequences. Nucleic Acids Res. **19:**4045–4057.

22. **Jarrell, K. J., D. P. Bayley, and A. S. Kostyukova.** 1996. The archaeal flagellum: a unique motility structure. J. Bacteriol. **178:**5057–5064.

23. **Jones, W. J., J. A. Leigh, F. Mayer, C. R. Woese, and R. S. Wolfe.** 1983. *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. Arch. Microbiol. **136:**254–261.

24. **Kandler, O., and K. König.** 1993. Cell envelopes of archaea: structure and chemistry, p. 223–259. *In* M. Kates, D. J. Kushner, and A. T. Matheson (ed.), The Biochemistry of *Archaea* (*Archaebacteria*). Elsevier Science Publishers B.V., Amsterdam, The Netherlands.

24a. **Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabata.** 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3:**109–139.

25. **Karlin, S., J. Mrázek, and A. M. Campbell.** 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179:**3899–3913.

26. **Kiss-Laszlo, Z., Y. Henry, J. P. Bachellerie, M. Caizergues-Ferrer, and T. Kiss.** 1996. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell **85:**1077–1088.

27. **Kleman-Leyer, K., D. A. Armbruster, and C. J. Daniels.** 1997. Properties of the *H. volcanii* tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. Cell **89:**839–847.

28. **Klimczak, L. J., F. Grummt, and K. J. Burger.** 1986. Purification and characterization of DNA polymerase from the archaebacterium *Methanobacterium thermoautotrophicum*. Biochemistry **25:**4850–4855.

29. **Langer, D., J. Hain, P. Thuriaux, and W. Zillig.** 1997. Transcription in *Archaea*: similarity to that in *Eucarya*. Proc. Natl. Acad. Sci. USA **92:**5768–5772.

30. **Lykke-Andersen, J., and R. A. Garrett.** 1994. Structural characteristics of the stable RNA introns of archaeal hyperthermophiles and their splicing junctions. J. Mol. Biol. **243:**846–855.

31. **Matsugi, J., K. Murao, and H. Ishikura.** 1996. Characterization of a B. subtilis minor isoleucine tRNA deduced from tDNA having a methionine anticodon CAT. J. Biochem. **119:**811–816.

32. **Muzi-Falconi, M., and T. J. Kelly.** 1995. Orp1, a member of the Cdc18/Cdc6 family of S-phase regulators, is homologous to a component of the origin recognition complex. Proc. Natl. Acad. Sci. USA **92:**12475–12479.

33. **Nölling, J., F. J. M. van Eeden, R. I. L. Eggen, and W. M. de Vos.** 1992. Modular organization of related archaeal plasmids encoding different restriction-modification systems in *Methanobacterium thermoformicicum*. Nucleic Acids Res. **20:**5047–5052.

34. **Nölling, J.** 1993. Mobile genetic elements in *Methanobacterium thermoautotrophicum*. Ph.D. thesis. Wageningen Agicultural University, The Netherlands.

35. **Nölling J., F. J. M. van Eeden, and W. M. de Vos.** 1993. Distribution and characterization of plasmid-related sequences in the chromosomal DNA of different thermophilic *Methanobacterium* strains. Mol. Gen. Genet. **240:**81–91.

36. **Nölling, J., T. D. Pihl, A. Vriesema, and J. N. Reeve.** 1995. Organization and growth phase-dependent transcription of methane genes in two regions of the *Methanobacterium thermoautotrophicum* genome. J. Bacteriol. **177:**2460–2468.

37. **Nölling, J., A. Elfner, J. R. Palmer, V. J. Steigerwald, T. D. Pihl, J. A. Lake, and J. N. Reeve.** 1996. Phylogeny of *Methanopyrus kandleri* based on methyl coenzyme M reductase operons. Int. J. System. Bacteriol. **46:**1170–1173.

38. **Nölling, J., and J. N. Reeve.** 1997. Growth and substrate-dependent transcription of the formate dehydrogenase (*fdhCAB*) operon in *Methanobacterium thermoformicicum* Z-245. J. Bacteriol. **179:**899–908.

39. **Olsen, G. J., C. R. Woese, and R. Overbeek.** 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176:**1–6.

40. **Perler, F. B., Olsen, G. J. and E. Adam.** 1997. Compilation and analysis of intein sequences. Nucleic Acids Res. **25:**1087–1093.

41. **Pietrokovski, S.** 1994. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. Protein Sci. **3:**2340–2350.

41a. **Pietrokovski, S.** Unpublished data.

42. **Pietrokovski, S.** 1996. A new intein in Cyanobacteria and its significance for the spread of inteins. Trends Genet. **12:**287–288.

43. **Pietrokovski, S.** Modular organization of inteins and C-terminal autocatalytic domains. Protein Sci., in press.

44. **Pietrokovski, S., and S. Henikoff.** 1997. A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. Mol. Gen. Genet. **254:**689–695.

45. **Pihl, T. D., S. Sharma, and J. N. Reeve.** 1994. Growth phase-dependent transcription of the genes that encode the two methylcoenzyme M reductase isoenzymes and N^5-methyltetrahydromethanopterin:coenzyme M methyltransferase in *Methanobacterium thermoautotrophicum* ΔH. J. Bacteriol. **176:**6384–6391.

46. **Reeve, J. N., J. Nölling, R. M. Morgan, and D. R. Smith.** 1997. Methanogenesis: genes, genomes, and who's on first. J. Bacteriol. **179:**5975–5986.

47. **Reeve, J. N., K. Sandman, and C. J. Daniels.** 1997. Archaeal histones, nucleosomes and transcription initiation. Cell **89:**999–1002.

48. **Richterich, P. and G. M. Church.** 1993. DNA sequencing with direct transfer electrophoresis and non-radioactive detection. Methods Enzymol. **218:**187–222.

49. **Riera, J., F. T. Robb, R. Weiss, and M. Fontecave.** 1997. Ribonucleotide reductase in the archaeon Pyrococcus furiosus: a critical enzyme in the evolution of DNA genomes? Proc. Natl. Acad. Sci. USA **94:**475–478.

50. **Sambrook, J. E., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

51. **Sanger, F., S. Nicklen, and A. R. Coulsen.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.

52. **Sawaya, M. R., H. Pelletier, A. Kumar, S. H. Wilson, and J. Kraut.** 1994. Crystal structure of rat DNA polymerase β: evidence for a common polymerase mechanism. Science **264:**1930–1935.

53. **Shine, J., and L. Dalgarno.** 1975. Correlation between the 3′-terminal-polypyrimidine sequence of 16S RNA and translational specificity of the ribosome. Eur. J. Biochem. **57:**221–230.

54. **Smith, D. R., P. Richterich, M. Rubenfield, P. W. Rice, C. Butler, H.-M. Lee, S. Kirst, K. Gundersen, K. Abendschan, Q. Xu, M. Chung, C. Deloughery, T. Aldredge, J. Maher, R. Lundstrom, C. Tulig, K. Falls, J. Imrich, D. Torrey, M. Engelstein, G. Breton, D. Madan, R. Nietupski, B. Seitz, S. Connelly, S. McDougall, H. Safer, R. Gibson, L. Doucette-Stamm, K. Eiglmeier, S. Bergh, S. T. Cole, K. Robison, L. Richterich, J. Johnson, G. M. Church, and J. Mao.** 1997. Multiplex sequencing of 1.5 Mb of the *Mycobacterium leprae* genome. Genome Res. **7:**802–819.

55. **Sorgenfrei, O., S. Müller, M. Pfeiffer, I. Sniezko, and A. Klein.** 1997. The [NiFe] hydrogenases of Methanococcus voltae: genes, enzymes, and regulation. Arch. Microbiol. **167:**189–195.

56. **Stams, A. J.** 1994. Metabolic interactions between anaerobic bacteria in methanogenic environments. Antonie Leeuwenhoek **66:**271–294.

57. **Stettler, R., and T. Leisinger.** 1992. Physical map of the *Methanobacterium thermoautotrophicum* Marburg chromosome. J. Bacteriol. **174:**7227–7234.

58. **Stettler, R., G. Erauso, and T. Leisinger.** 1995. Physical and genetic map of the *Methanobacterium wolfei* genome and its comparison with the updated map of *Methanobacterium thermoautotrophicum* Marburg. Arch. Microbiol. **163:**205–210.

59. **Tauer, A., and S. A. Benner.** 1997. The B12-dependent ribonucleotide reductase from the archaebacterium Thermoplasma acidophila: an evolution-

ary solution to the ribonucleotide reductase conundrum. Proc. Natl. Acad. Sci. USA **94:**53–58.

60. **Thauer, R. K., R. Hedderich, and R. Fischer.** 1993. Reactions and enzymes involved in methanogenesis from CO_2 and H_2, p. 209–252. *In* J. M. Ferry (ed.), Methanogenesis, ecology, physiology, biochemistry and genetics. Chapman and Hall, New York, N.Y.
61. **Thompson, L. D., and C. J. Daniels.** 1990. Recognition of exon-intron boundaries by the *Halobacterium volcanii* tRNA intron endonuclease. J. Biol. Chem. **265:**18104–18111.
62. **Vermeij, P., E. Vinke, J. T. Keltjens, and C. van der Drift.** 1995. Purification and properties of the coenzyme F_{390} hydrolase from *Methanobacterium thermoautotrophicum* (strain Marburg). Eur. J. Biochem. **234:**592–597.
63. **Wang, X., and J. Lutkenhaus.** FtsZ ring: the eubacterial division apparatus conserved in archaebacteria. Mol. Microbiol. **21:**313–319.
64. **Watson, M. W., P. R. Lamden, and I. N. Clarke.** 1990. The nucleotide sequence of the 60 kDa cysteine rich outer membrane protein of *Chlamydia psittaci* strain EAE/A22/M. Nucleic Acids Res. **18:**5300.
65. **Weidenhammer, E. M., M. Singh, M. Ruiz-Noriega, and J. L. Woolford, Jr.** 1996. The PRP31 gene encodes a novel protein required for pre-mRNA splicing in *Saccharomyces cerevisiae*. Nucleic Acids Res. **24:**1164–1170.
66. **Weil, C. F., D. S. Cram, B. A. Sherf, and J. N. Reeve.** 1988. Structure and comparative analysis of the genes encoding component C of the methyl coenzyme M reductase in the extremely thermophilic archaebacterium *Methanothermus fervidus*. J. Bacteriol. **170:**4718–4726.
67. **Wilting, R., S. Schorling, B. C. Persson, and A. Böck.** 1997. Selenoprotein synthesis in Archaea: identification of an mRNA element of *Methanococcus jannaschii* probably directing selenocysteine insertion. J. Mol. Biol. **266:**637–641.
68. **Wipat, A., N. Carter, S. C. Brignell, B. J. Guy, K. Piper, J. Sanders, P. T. Emmerson, and C. R. Harwood.** 1996. The dnaB-pheA (256 degrees-240 degrees) region of the *Bacillus subtilis* chromosome containing genes responsible for stress responses, the utilization of plant cell walls and primary metabolism. Microbiology **142:**3067–3078.
69. **Wolfe, R. S.** 1991. My kind of biology. Annu. Rev. Microbiol. **45:**1–35.
70. **Wool, I. G., Y. L. Chan, and A. Gluck.** 1995. Structure and evolution of mammalian ribosomal proteins. Biochem. Cell Biol. **73:**933–947.
71. **Washington University School of Medicine.** 1997. Washington University Blast2, version 2.0a10. Washington University School of Medicine, St. Louis, Mo.
72. **Zeikus, J. G., and R. S. Wolfe.** 1972. *Methanobacterium thermoautotrophicus* sp. n., an anaerobic, autotrophic, extreme thermophile. J. Bacteriol. **109:**707–713.