Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*

Carol J. Bult, Owen White, Gary J. Olsen, Lixin Zhou, Robert D. Fleischmann, Granger G. Sutton, Judith A. Blake, Lisa M. FitzGerald, Rebecca A. Clayton, Jeannine D. Gocayne, Anthony R. Kerlavage, Brian A. Dougherty, Jean-Francois Tomb, Mark D. Adams, Claudia I. Reich, Ross Overbeek, Ewen F. Kirkness, Keith G. Weinstock, Joseph M. Merrick, Anna Glodek, John L. Scott, Neil S. M. Geoghagen, Janice F. Weidman, Joyce L. Fuhrmann, Dave Nguyen, Teresa R. Utterback, Jenny M. Kelley, Jeremy D. Peterson, Paul W. Sadow, Michael C. Hanna, Matthew D. Cotton, Kevin M. Roberts, Margaret A. Hurst, Brian P. Kaine, Mark Borodovsky, Hans-Peter Klenk, Claire M. Fraser, Hamilton O. Smith, Carl R. Woese, J. Craig Venter*

The complete 1.66–megabase pair genome sequence of an autotrophic archaeon, *Methanococcus jannaschii*, and its 58– and 16–kilobase pair extrachromosomal elements have been determined by whole-genome random sequencing. A total of 1738 predicted protein-coding genes were identified; however, only a minority of these (38 percent) could be assigned a putative cellular role with high confidence. Although the majority of genes related to energy production, cell division, and metabolism in *M. jannaschii* are most similar to those found in Bacteria, most of the genes involved in transcription, translation, and replication in *M. jannaschii* are more similar to those found in Eukaryotes.

The discovery of the Archaea in 1977 (*1*) created a quandary for biologists because it was then widely believed that the deepest, most significant evolutionary distinctions were those between Prokaryotes and Eukaryotes. Yet the Archaea, although cytologically prokaryotic, are not specifically related to the Bacteria; at the molecular level, the Archaea are in many respects more like Eukaryotes and may be specifically related to them (*2*). The nature of the Archaea and their relationships to Eukaryotes and Bacteria have posed an intriguing and incompletely resolved puzzle, one that until now has been addressed on the basis of evidence from individual genes (*2*). We now report the first complete genome sequence for a representative of the Archaea, *Methanococcus jannaschii*. The *M. jannaschii* genome sequence provides the first opportunity to compare complete genetic complements and biochemical pathways among the three domains of life from which all extant life forms evolved. *Methanococcus jannaschii* also represents the first complete genome of an autotrophic organism. Its genome sequence, therefore, should provide valuable information on the genetic basis for encoding the metabolic capacity to synthesize de novo all of the building blocks essential for cellular life from inorganic constituents.

The era of true comparative genomics has been ushered in by complete genome sequencing and analysis. We recently described the first two complete bacterial genome sequences, those of *Haemophilus influenzae* and *Mycoplasma genitalium* (*3*). In addition, the complete genome of a Eukaryote, *Saccharomyces cerevisiae*, was recently reported to have been completed (*4*). Large-scale DNA sequencing also has produced an extensive collection of sequence data from *Homo sapiens* (*5*) and *Caenorhabditis elegans* (*5*). The lack of archaeal sequence data has hampered construction of a comprehensive comparative evolutionary framework for assessing the molecular basis of the origin and diversification of cellular life.

Methanococcus jannaschii was originally isolated by J. A. Leigh from a sediment sample collected from the sea floor surface at the base of a 2600-m-deep "white smoker" chimney located at 21°N on the East Pacific Rise (*6*). *Methanococcus jannaschii* grows at pressures of up to more than 200 atm and over a temperature range of 48° to 94°C, with an optimum temperature near 85°C (*6*). It is a strict anaerobe, and, as the name implies, it produces methane.

A whole-genome random sequencing method (*3*) was used to obtain the complete genome sequence for *M. jannaschii*. A small-insert plasmid library [average insert size, 2.5 kilobase pairs (kbp)] and a large-insert λ library (average insert size, 16 kbp) were used as substrates for sequencing. The λ library was used to form a genome scaffold and to verify the orientation and integrity of the contigs formed from the assembly of sequences from the plasmid library. All clones were sequenced from both ends to aid in ordering of contigs during the sequence assembly process. The average length of sequencing reads was 481 bp. A total of 36,718 sequences were assembled by means of the TIGR Assembler (*3*, *7*). Sequence and physical gaps were closed by a combination of strategies (*3*). The colinearity of the in vivo genome to the genome sequence was confirmed by comparison of restriction fragments from six rare-cutter restriction enzymes (Aat II, Bam HI, Bgl II, Kpn I, Sma I, and Sst II) to those predicted from the sequence data. Additional confidence in the colinearity was provided by the genome scaffold produced by sequence pairs from 339 large-insert λ clones, which covered 88% of the main chromosome. Open reading frames (ORFs) and predicted protein-coding regions were identified as described (*3*) with modification (*8*).

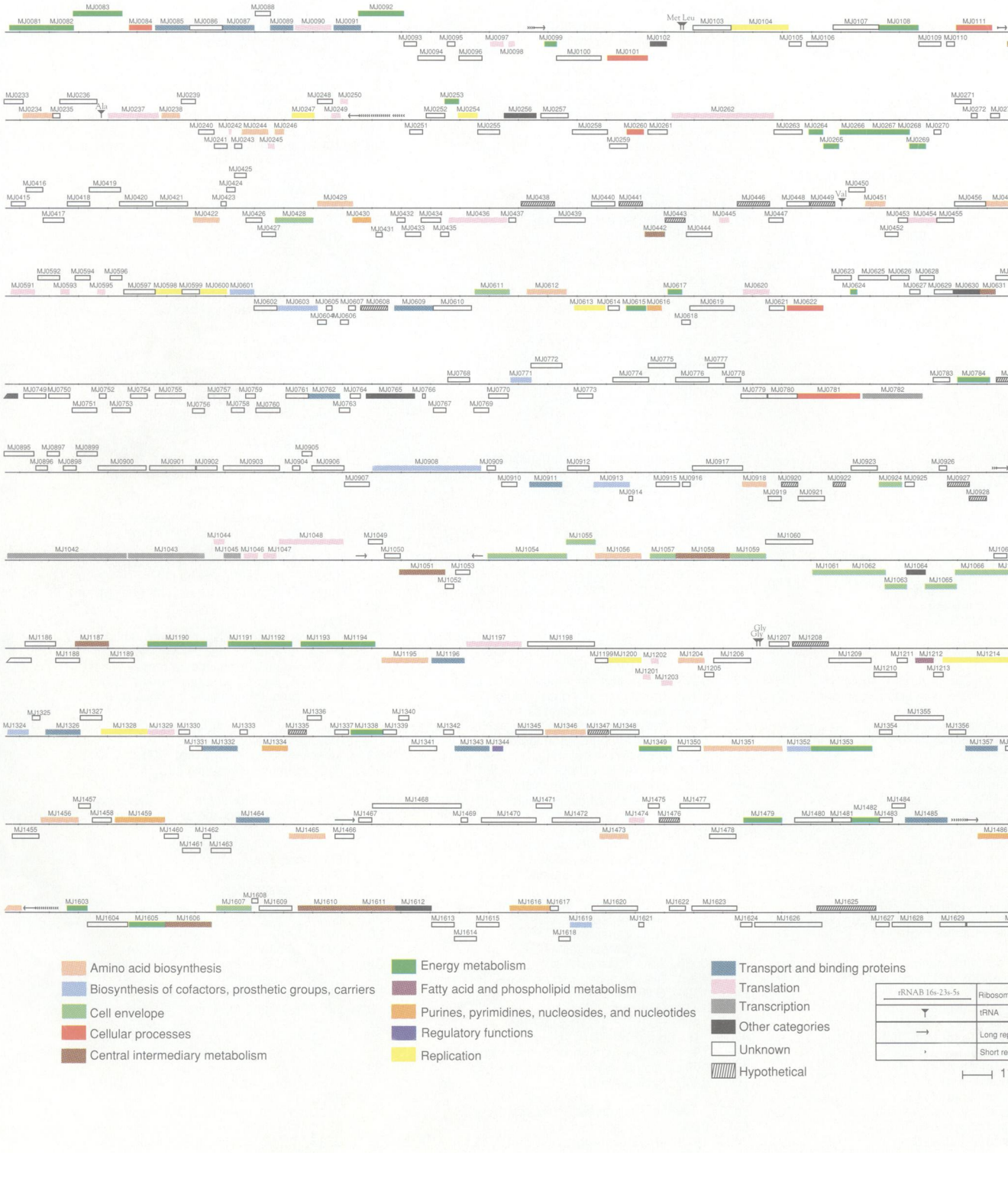
The *M. jannaschii* genome consists of three physically distinct elements: (i) a large circular chromosome of 1,664,976 base pairs (bp) (Fig. 1), which contains 1682 predicted protein-coding regions and has a G+C content of 31.4%; (ii) a large circular extrachromosomal element (ECE) (*9*) of 58,407 bp, which contains 44 predicted protein-coding regions and has a G+C content of 28.2% (Fig. 2); and (iii) a small circular ECE (*9*) of 16,550 bp, which

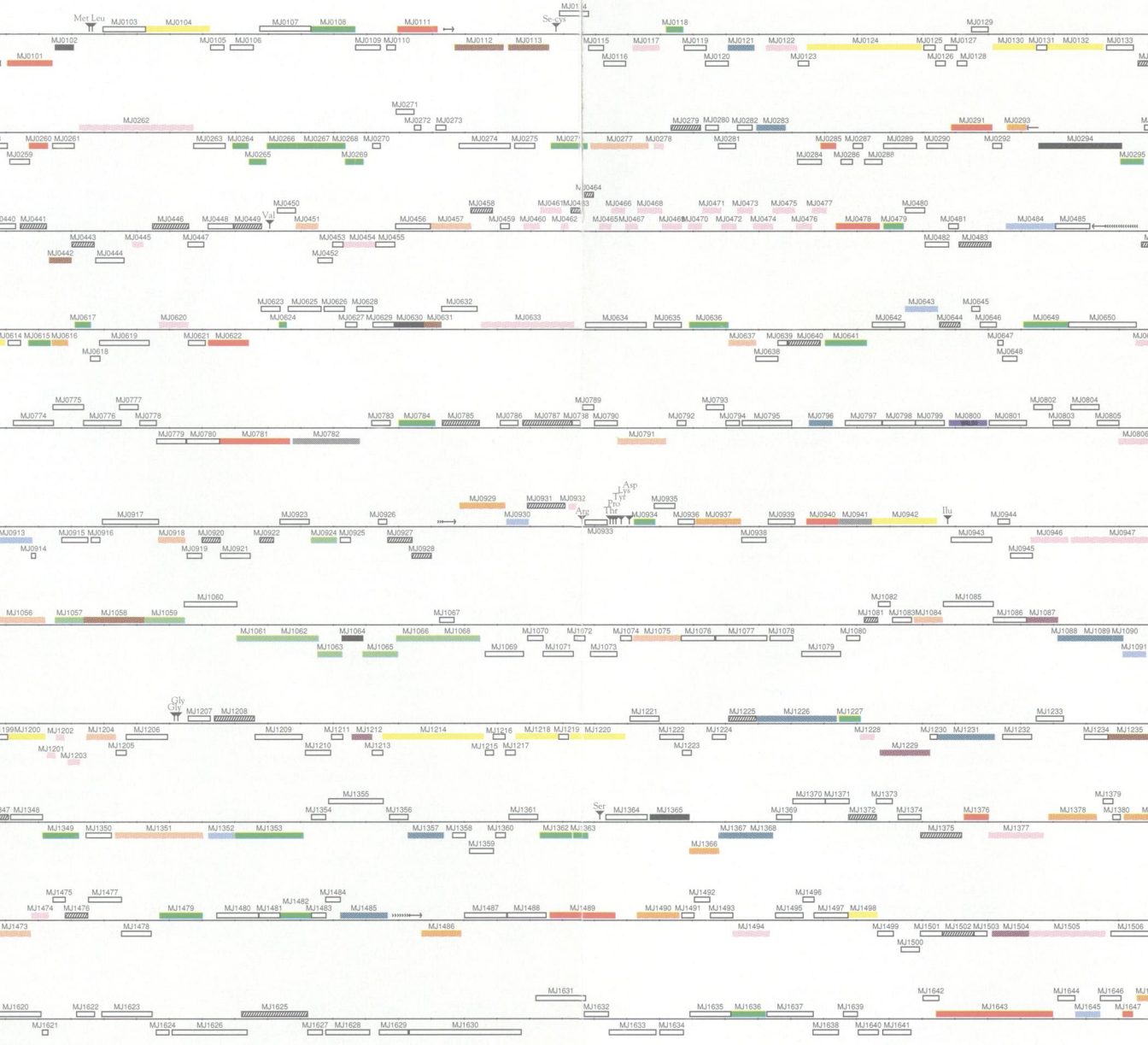
G. J. Olsen, C. I. Reich, B. P. Kaine, and C. R. Woese are in the Microbiology Department, University of Illinois, Champaign-Urbana, IL 61801, USA. R. Overbeek is with the Division of Mathematics and Computer Science, Argonne National Laboratory, Argonne, IL 60439, USA. J. M. Merrick is in the Department of Microbiology, State University of New York, Buffalo, NY 14214, USA. M. Borodovsky is with the School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA. H. O. Smith is in the Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. All other authors are with The Institute for Genomic Research (TIGR), Rockville, MD 20850, USA.

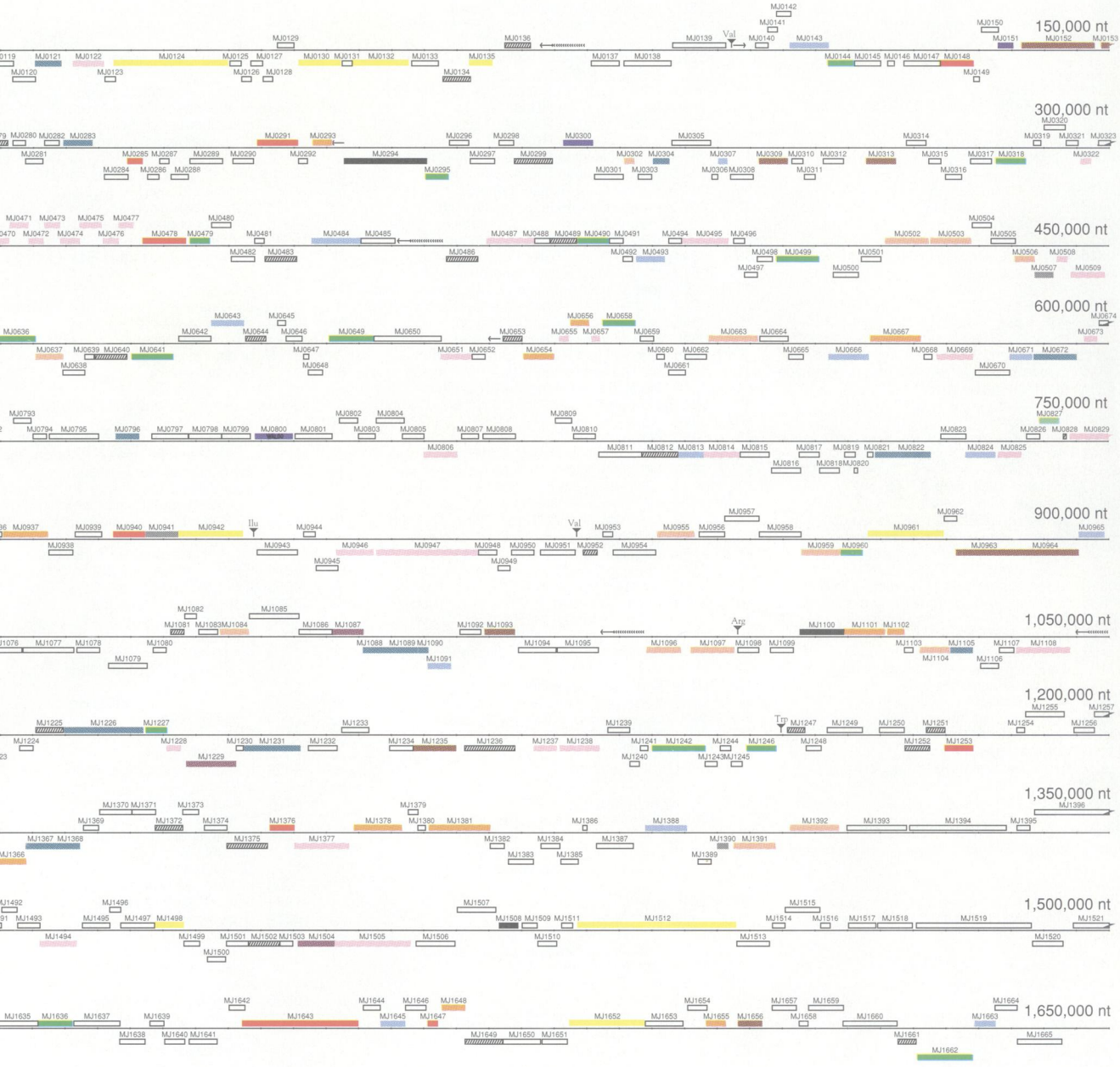
*To whom correspondence should be addressed at The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

MJ#	Gene description	%Id
	Amino acid biosynthesis	
	Aromatic amino acid family	
1454	3-dehydroquinate Dtase	34
0502	5-enolpyruvylshikimate 3-phosphate Sase	37
1075	anthranilate Sase, sub I	49
0234	anthranilate Sase, sub II'	45
0238	anthranilate Sase, sub II"	52
0246	chorismate mutase, sub A	38
0612	chorismate mutase, sub B	34
1175	chorismate Sase	49
0918	indole-3-glycerol phosphate Sase	43
0451	N-phosphoribosyl anthranilate isomerase	42
0637	prephenate DTase	39
1084	shikimate 5-DHase	35
1038	tryptophan Sase, alpha sub	50
1037	tryptophan Sase, beta sub	63
	Aspartate family	
1116	Asn Sase	34
1056	Asn Sase	35
1391	Asp ATase	31
0684	Asp ATase	38
0001	Asp ATase	42
0205	Asp-semialdehyde DHase	52
0571	aspartokinase I	39
1473	cobalamin-independent Met Sase	48
1097	diaminopimelate Dcase	42
1119	diaminopimelate epimerase	37
0422	dihydrodipicolinate RDase	45
0244	dihydrodipicolinate Sase	48
1003	homoaconitase	36
1602	homoserine DHase	41
1104	homoserine kinase	31
0020	L-asparaginase I	35
0457	succinyl-diaminopimelate desuccinylase	28
1465	Thr Sase	52
	Glutamate family	
0069	acetylglutamate kinase	45
0791	argininosuccinate lyase 41	
0429	argininosuccinate Sase	72
0186	Glu N-acetylTase	44
1351	Glu Sase (NADPH), alpha sub	38
1346	Gln Sase	71
1096	N-Ac-gamma-glutamyl-phosphate RDase	41
0721	N-acetylornithine ATase	47
0881	ornithine carbamoylTase	43
	Histidine family	
1204	ATP PRTase	34
1456	histidinol DHase	47
0955	histidinol-phosphate ATase	38
0698	imidazoleglycerol-phosphate DHase	51
0506	imidazoleglycerol-phosphate Sase	47
0411	imidazoleglycerol-phosphate Sase	62
1430	phosphoribosyl-AMP cyclohydrolase	64
0302	phosphoribosyl-ATP pyrophosphohydrolase	48
1532	PRAC ribotide isomerase	57
	Pyruvate family	
1392	2-isopropylmalate Sase	44
0503	2-isopropylmalate Sase	45
1271	3-isopropylmalate DTase	42
1277	3-isopropylmalate DTase	50
0663	acetolactate Sase, large sub	35
0277	acetolactate Sase, large sub	51
0161	acetolactate Sase, small sub	50
1008	branched-chain amino acid Atase	43
1276	dihydroxy-acid DTase	45
1195	isopropylmalate Sase	43
1543	ketol-acid reductoisomerase	54
	Serine family	
1597	Gly hydroxy MTase	69
1018	phosphoglycerate DHase	43
1594	phosphoserine phosphatase	43
0959	Ser ATase	55
	Biosynthesis of cofactors, prosthetic groups, and carriers	
0603	Glu-1-semialdehyde ATase	52
0569	porphobilinogen deaminase	42
0493	quinolinate PRTase	40
0407	quinolinate Sase	41
1388	S-adenosylhomocysteine hydrolase	61
	Biotin	
1297	6-carboxyhexanoate-CoA ligase	43
1298	8-amino-7-oxononanoate Sase	45
1300	DAPA ATase	40
1619	bifunctional prot	62
1296	biotin Sase	39
1299	dethiobiotin Sase	37
	Heme and porphyrin	
1438	cobalamin (5'-phosphate) Sase	28
0552	cobalamin biosyn prot J	27
1314	cobalamin biosyn prot D	35
0022	cobalamin biosyn prot D	35
1569	cobalamin biosyn prot M	30
1091	cobalamin biosyn prot M	52
0908	cobalamin biosyn prot N	38
0484	cobyric acid Sase	74
1421	cobyrinic acid a,c-diamide Sase	36
0143	glutamyl-tRNA RDase	48
0643	porphobilinogen Sase	63
0930	precorrin isomerase	39
0771	precorrin-2 MTase	31
0813	precorrin-3 methylase	45
1578	precorrin-3 methylase	55
1522	precorrin-6Y methylase	31
0391	precorrin-8W DCase	28
0965	uroporphyrin-III C-MTase	55
0994	uroporphyrinogen III Sase	29
	Menaquinone and ubiquinone	
1645	CoPQQ synthesis prot III	34
	Molybdopterin	
0824	molybdenum cofactor biosyn moaA prot	32
0167	molybdenum cofactor biosyn prot moaB	38
1135	molybdenum cofactor biosyn prot moaC	48
0666	molybdenum cofactor biosyn prot moeA	35
0886	molybdenum cofactor biosyn prot moeA	35
1663	mop-guanine dinucl biosyn prot A	30
1324	mop-guanine dinucl biosyn prot B	33
	Pantothenate	
0913	pantothenate metabolism flavoprotein	33
	Pyridine nucleotides	
1352	NH(3)-dep NAD+ Sase	38
	Riboflavin	
0055	GTP cyclohydrolase II	39
0671	riboflavin-specific deaminase	42
	Thiamine	
1026	thiamine biosyn prot	46
0601	thiamine biosynthetic enzyme	36
	Thioredoxin, glutaredoxin, and glutathione	
1536	thioredoxin RDase	39
0530	thioredoxin-2	33
0307	glutaredoxin-like prot	53
	Cell envelope	
	Membranes, lipoproteins, and porins	
0544	dolichyl-phosphate mannose Sase	35
1057	glycosyl Tase	32
0827	membrane prot	43
0611	membrane prot	34
	Pseudomurein sacculus	
1160	amidase 41	
0204	amidoPRTase	52
	Surface polysaccharides, lipopolysaccharides, and antigens	
0924	capsular polysaccharide biosyn prot B	55
1061	capsular polysaccharide biosyn prot D	52
1055	capsular polysaccharide biosyn prot I	51
1059	capsular polysaccharide biosyn prot M	32
1607	LPS biosyn rel rfbu-prot	34
1113	GLcNAc-1-phosphate Tase	28
0399	phosphomannomutase	37
1068	put O-antigen transporter	24
1066	spore coat polysaccharide biosyn prot C	55
1065	spore coat polysaccharide biosyn prot E	38
1063	spore coat polysaccharide biosyn prot F	39
1062	spore coat polysaccharide biosyn prot G	33
0211	UDP-glucose 4-epimerase	38
1054	UDP-glucose DHase	43
0428	UDP-N-Ac-D-mannosaminuronic acid DHase	47
	Surface structures	
0891	flagellin B1	56
0892	flagellin B2	61
0893	flagellin B3	60
	Cellular processes	
	Cell division	
1489	cell division control prot 54	35
0363	cell division control prot 21	30
1156	cell division control prot 48	52
0579	cell division inhibitor minD-rel prot	32
0169	cell division inhibitor minD	35
0547	cell division inhibitor minD	37
0084	cell division inhibitor minD	29
0174	cell division prot	29
0370	cell division prot ftsZ 49	
1376	cell division prot ftsJ 41	
0622	cell division prot ftsZ 51	
0148	centromere/microtubule-BP	44
1647	DNA BP	58
1643	P115 prot	31
	Chaperones	
0999	chaperonin	73
0285	heat shock prot 31	
0278	rotamase, peptidyl-prolyl cis-trans isomerase	39
0825	rotamase, peptidyl-prolyl cis-trans isomerase	32
	Chromosome-associated proteins	
ECL17	archaeal histone	59
ECL29	archaeal histone	59
0932	archaeal histone	68
0168	archaeal histone	68
1258	archaeal histone	72
	Detoxification	
0736	alkyl hydroperoxide RDase	48
1541	N-ethylammeline chlorohydrolase	30
	Protein and peptide secretion	
0478	preprot translocase SecY	71
0111	protein-export membrane prot SecD	29
1253	protein-export membrane prot SecF	32
0260	signal peptidase	35
0101	signal recognition particle prot	41
0291	signal recognition particle prot	49
	Transformation	
0781	klbA prot	35
0940	transformation sensitive prot	31
	Central intermediary metabolism	
	Amino sugars	
1420	Gln–fructose-6-phosphate transaminase	42
	Carbon fixation	
0153	carbon monoxide DHase, alpha sub	48
0152	carbon monoxide DHase, alpha sub	43
0156	carbon monoxide DHase, alpha sub	48
0728	carbon monoxide DHase, beta sub	36
0112	corrinoid/iron-sulfur prot, large sub	34
0113	corrinoid/iron-sulfur prot, small sub	38
1235	ribulose bisphosphate carboxylase, large sub	41
	Degradation of polysaccharides	
1611	alpha-amylase	28
0555	endoglucanase	0
1610	glucoamylase	27
	Nitrogen metabolism	
1187	ADP-ribosylglycohydrolase	30
0214	hydrogenase accessory prot	32
0713	hydrogenase accessory prot	35
0676	hydrogenase expression/formation prot E	45
0442	hydrogenase expression/formation prot B	43
0200	hydrogenase expression/formation prot C	40
0993	hydrogenase expression/formation prot D	43
0631	hydrogenase maturation protease	34
1093	nifB prot	44
0879	nitrogenase RDase	78
0685	nitrogenase RDase rel prot	32
1051	nodulation factor production prot	33
1058	nodulation factor production prot	35
	Phosphorus compounds	
0963	N-methylhydantoinase	33
0964	N-methylhydantoinase	36
	Polyamine biosynthesis	
0535	acetylpolyamine aminohydolase	34
0313	spermidine Sase	39

MJ#	Gene description	%Id
Polysaccharides (cytoplasmic)		
1606	glycogen Sase	32
Other		
1656	2-hydroxyhepta-2,4-diene-1,7-dioate isomerase	41
0406	ribokinase	24
0309	ureohydrolase	41
Energy metabolism		
0479	adenylate kinase	100
Aerobic		
0649	NADH oxidase	28
0520	NADH-ubiquinone oxidoRDase, sub 1	29
Anaerobic		
0092	fumarate RDase	41
ATP–proton motive force interconversion		
0217	ATP Sase, A sub	61
0216	ATP Sase, B sub	68
0219	ATP Sase, C sub	29
0615	ATP Sase, D sub	39
0220	ATP Sase, E sub	33
0218	ATP Sase, F sub	22
0222	ATP Sase, I sub	28
0221	ATP Sase, K sub	46
Electron transport		
1446	cytochrome-c3 hydrogenase, gamma sub	41
0741	desulfoferredoxin	44
0722	ferredoxin	43
0099	ferredoxin	40
0061	ferredoxin	43
0199	ferredoxin	75
0578	ferredoxin	50
0533	ferredoxin 2[4Fe-4S] homolog	37
0624	ferredoxin 2[4Fe=4S]	48
0276	ferredoxin oxidoRDase, alpha sub	45
0267	ferredoxin oxidoRDase, alpha sub	33
0537	ferredoxin oxidoRDase, beta sub	41
0266	ferredoxin oxidoRDase, beta sub	33
0268	ferredoxin oxidoRDase, delta sub	59
0536	ferredoxin oxidoRDase, gamma sub	32
0269	ferredoxin oxidoRDase, gamma sub	64
0732	flavoprotein	41
1192	MVR hydrogenase, alpha sub	78
1191	MVR hydrogenase, gamma sub	72
1362	NADH DHase, sub 1	26
0934	polyferredoxin	41
1303	polyferredoxin	40
0514	polyferredoxin	36
1193	polyferredoxin	62
1227	pyruvate formate-lyase activating enzyme	31
0735	rubredoxin	60
0740	rubredoxin	62
Fermentation		
0007	2-hydroxyglutaryl-CoA DTase, beta sub	23
Gluconeogenesis		
1479	Ala ATase 2	30
0542	phosphoenolpyruvate Sase	61
Glycolysis		
1482	2-phosphoglycerate kinase	48
0641	3-phosphoglycerate kinase	58
0232	enolase	59
1605	glucose-6-phosphate isomerase	33
1146	G3PDHase	60
0490	lactate DHase	40
1411	NADP-dep G3PDHase	40
0108	pyruvate kinase	39
1528	triosephosphate isomerase	30
Methanogenesis		
0253	F420-reducing hydrogenase, delta sub	50
1035	F420-dep N5,N10-methylene-H4MPT DHase	68
0030	F420-reducing hydrogenase, alpha sub	66
0727	F420-reducing hydrogenase, alpha sub	28
0029	F420-reducing hydrogenase, alpha sub	51
0725	F420-reducing hydrogenase, beta sub	43
1349	F420-reducing hydrogenase, beta sub	36
0032	F420-reducing hydrogenase, beta sub	72
0870	F420-reducing hydrogenase, beta sub	43
0726	F420-reducing hydrogenase, gamma sub	44
0031	F420-reducing hydrogenase, gamma sub	77
0295	formate DHase (fdhD)	36
0006	formate DHase, alpha sub	42
1353	formate DHase, alpha sub	56
0005	formate DHase, beta sub GB:J02581_2 0.0	49
0155	formate DHase, iron-sulfur sub	38
0264	formate hydrogenlyase, sub 2	40
0265	formate hydrogenlyase, sub 2	41
0515	formate hydrogenlyase, sub 5	32
1027	formate hydrogenlyase, sub 5	35
1363	formate hydrogenlyase, sub 7	38
0516	formate hydrogenlyase, sub 7	49
0318	formylmethanofuran:H4MPT formylTase	71
1338	H2-dep methylene-H4MPT DHase-rel prot	30
0715	H2-form N5,N10-methylene-H4MPT DHase-rel prot	30
0784	H2-form N5,N10-methylene-H4MPT DHase	75
1190	heterodisulfide RDase, A sub	60
0743	heterodisulfide RDase, B sub	61
0863	heterodisulfide RDase, B sub	64
0864	heterodisulfide RDase, C sub	53
0744	heterodisulfide RDase, C sub	56
0118	methyl CoM RDase II operon, prot D	54
0083	methyl CoM RDase II, alpha sub	88
0081	methyl CoM RDase II, beta sub	80
0082	methyl CoM RDase II, gamma sub	83
0844	methyl CoM RDase operon, prot C	83
0843	methyl CoM RDase operon, prot D	60
1662	methyl CoM RDase system, component A2	38
1242	methyl CoM RDase system, component A2	61
0846	methyl CoM RDase, alpha sub	86
0842	methyl CoM RDase, beta sub	76
0845	methyl CoM RDase, gamma sub	79
1636	N5,N10-methenyl-H4MPT cyclohydrolase	69
1534	N5,N10-methylene-H4MPT RDase	67
0849	N5-methyl-H4MPT:CoM MTase, C sub	40
0848	N5-methyl-H4MPT:CoM MTase, D sub	64
0850	N5-methyl-H4MPT:CoM MTase, B sub	37
0847	N5-methyl-H4MPT:CoM MTase, E sub	62
0852	N5-methyl-H4MPT:CoM MTase, F sub	38
0854	N5-methyl-H4MPT:CoM MTase, H sub	63
0851	N5-methyl-H4MPT:CoM MTase, A sub	56
0853	N5-methyl-H4MPT:CoM MTase, G sub	50
1169	tungsten formyl-MFR DHase, A sub	70
1194	tungsten formyl-MFR DHase, B sub	70
1171	tungsten formyl-MFR DHase, C sub	52
0658	tungsten formyl-MFR DHase, C sub rel prot	36
1168	tungsten formyl-MFR DHase, D sub	58
1165	tungsten formyl-MFR DHase, E sub	45
1166	tungsten formyl-MFR DHase, F sub	48
1167	tungsten formyl-MFR DHase, G sub	60
Pentose phosphate pathway		
0680	pentose-5-phosphate-3-epimerase	46
1603	ribose 5-phosphate isomerase	42
0960	transaldolase	60
0681	transketolase, A sub	42
0679	transketolase, B sub	38
Pyruvate dehydrogenase		
0636	dihydrolipoamide DHase	29
Sugars		
1418	fuculose-1-phosphate aldolase	30
TCA cycle		
0499	aconitase	30
1294	fumarate hydratase, class I, A sub	35
0617	fumarate hydratase, class I, B sub	44
1596	isocitrate DHase	43
0720	isocitrate DHase (NADP)	48
1425	malate DHase	61
0033	succinate DHase, flavoprotein sub	41
1246	succinyl-CoA Sase, alpha sub	59
0210	succinyl-CoA Sase, beta sub	49
Fatty acid and phospholipid metabolism		
0705	3-hydroxy-3-methylglutaryl CoA RDase	48
1546	acyl carrier prot Sase	65
0860	bifunctional short chain isoprenyl diP Sase	49
1229	biotin carboxylase	59
1212	CDP-diacylglycerol–Ser O-phosphatidylTase	44
1504	lipopolysaccharide biosyn prot D	43
1087	melvalonate kinase	34
1549	nonspecific lipid-transfer prot	47
Purines, pyrimidines, nucleosides, and nucleotides		
2'-Deoxyribonucleotide metabolism		
0832	anaerobic ribonucleoside-triP RDase	28
0430	deoxycytidine triP deaminase	39
1102	put deoxycytidine triP deaminase	32
0511	deoxyuridylate hydroxymethylase	40
0937	glycinamide ribonucleotide Sase	38
Purine ribonucleotide biosynthesis		
0929	adenylosuccinate lyase	43
0561	adenylosuccinate Sase	43
1131	GMP Sase	53
1575	GMP Sase	41
1616	inosine-5'-monophosphate DHase	62
1265	nucleoside diP kinase	55
0616	PRAD carboxylase	57
1592	PRAD succinocarboxamide Sase	48
0203	phosphoribosylformylglycinamidine cyclo-ligase	40
1648	phosphoribosylformylglycinamidine Sase I	52
1264	phosphoribosylformylglycinamidine Sase II	43
1486	phosphoribosylglycinamide formylTase 2	64
1366	ribose-phosphate pyrophosphokinase	35
Pyrimidine ribonucleotide biosynthesis		
1581	Asp carbamoyl Tase, catalytic sub	50
1406	Asp carbamoyl Tase, regulatory sub	37
1378	carbamoyl-phosphate Sase, large sub	60
1381	carbamoyl-phosphate Sase, large sub	55
1019	carbamoyl-phosphate Sase, small sub	49
1174	CTP Sase	58
0656	cytidylate kinase	34
1490	dihydroorotase	35
0654	dihydroorotase DHase	42
0293	thymidylate kinase	32
1109	uridine 5'-monophosphate Sase	39
1259	uridylate kinase	31
Salvage of nucleosides and nucleotides		
1459	adenine deaminase	36
1655	adenine PRTase	34
0060	methylthioadenosine phosphorylase	42
0667	thymidine phosphorylase	31
Sugar-nucleotide biosynthesis and conversions		
1101	glucose-1-phosphate thymidylyl Tase	34
1334	UDP-glucose pyrophosphorylase	47
Regulatory functions		
0800	(R)-2-hydroxyglutaryl-CoA DTase activator	32
0004	(R)-2-hydroxyglutaryl-CoA DTase activator	38
0059	nitrogen regulatory prot P-II	57
1344	nitrogen regulatory prot P-II	57
0300	put transcriptional regulator	32
0723	put transcriptional regulator	51
0151	put transcriptional regulator	52
Replication		
Degradation of DNA		
1434	endonuclease III	27
0613	endonuclease III	42
1439	thermonuclease precursor	37
DNA replication, restriction, modification, recombination, and repair		
1029	dimethyladenosine Tase	40
0104	put DNA helicase	36
0171	DNA ligase	36
0869	DNA repair prot 45	
1444	DNA repair prot RAD2	38
0254	DNA repair prot RAD51	34
0961	DNA replication initiator prot	33
1652	DNA topoisomerase I	34
0885	DNA-dep DNA polymerase, fam B	47
1529	methylated DNA protcysteine MTase	37
0598	modification methylase	36
1328	modification methylase	36
1200	modification methylase	41
1498	modification methylase	31
0563	modification methylase	35
0985	modification methylase	54
1149	mutator mutT prot	41
0942	put ATP-dep helicase	35
0247	proliferating-cell nuclear antigen	32
0026	proliferating-cell nucleolar antigen, 120 kD	47
1422	replication factor C	46
0884	replication factor C, large sub	38
1220	restriction modification enzyme, M1 sub	33
0132	type I restriction enzyme	37
0130	restriction modification system S sub	29
1512	reverse gyrase	42
0135	ribonuclease HII	40
ECL42	type I restriction enyzme ECOR124/3 I M prot	40
1214	type I restriction enzyme	28
0124	type I restriction enzyme	32
ECL40	type I restriction enzyme	37
1531	type I restriction enzyme Cfrl, specificity sub	39


Met Leu
Ala
Val
Gly
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid and phospholipid metabolism
Purines, pyrimidines, nucleosides, and nucleotides
Regulatory functions
Replication
Transport and binding proteins
Translation
Transcription
Other categories
Unknown
Hypothetical
rRNA 16s-23s-5s
Ribosom
tRNA
Long re
Short re


Met Leu
Ser Cys
Val
Gly Gly
Ser
Asp Lys Tyr Pro Thr Arg
Ile
Transport and binding proteins
Translation
Transcription
Other categories
Unknown
Hypothetical
bolism
s, and nucleotides
rRNA 16s-23s-5s
Ribosomal operon
tRNA
Long repeat
Short repeat
1 kb


150,000 nt
300,000 nt
450,000 nt
600,000 nt
750,000 nt
900,000 nt
1,050,000 nt
1,200,000 nt
1,350,000 nt
1,500,000 nt
1,650,000 nt

MJ#	Gene description	%Id
1218	type I restriction-modification enzyme, S sub	30
0984	type II restriction enzyme	48
0600	type II restriction enzyme DPNII	41
	Transcription	
	DNA-dependent RNA polymerase	
1042	DNA-dep RNA polymerase, A' sub	75
1043	DNA-dep RNA polymerase, A'' sub	65
1041	DNA-dep RNA polymerase, B' sub	74
1040	DNA-dep RNA polymerase, B'' sub	71
0192	DNA-dep RNA polymerase, D sub	41
0397	DNA-dep RNA polymerase, E' sub	42
0396	DNA-dep RNA polymerase, E'' sub	36
1039	DNA-dep RNA polymerase, H sub	50
1390	DNA-dep RNA polymerase, I sub	54
0197	DNA-dep RNA polymerase, K sub	44
0387	DNA-dep RNA polymerase, L sub	36
0196	DNA-dep RNA polymerase, N sub	54
	RNA processing	
0697	fibrillarin-like pre-rRNA processing prot	76
	Transcription factors	
0941	put transcription initiation factor IIIC	21
1045	put transcription term-antiterm factor nusA	48
0372	put transcription term-antiterm factor nusG	25
0507	TATA-binding transcription initiation factor	48
0782	transcription initiation factor IIB	64
1148	transcription-associated prot 'TFIIS'	59
	Translation	
0160	PET112 prot	34
	Amino acyl tRNA synthetases	
0564	alanyl-tRNA Sase	28
0237	arginyl-tRNA Sase	32
1555	aspartyl-tRNA Sase	58
1377	glutamyl-tRNA Sase	52
0228	glycyl-tRNA Sase	46
1000	histidyl-tRNA Sase	36
0947	isoleucyl-tRNA Sase	53
0633	leucyl-tRNA Sase	36
1263	methionyl-tRNA Sase	37
0487	phenylalanyl-tRNA Sase, alpha sub	41
1108	phenylalanyl-tRNA Sase, beta sub	32
1238	prolyl-tRNA Sase	40
1197	threonyl-tRNA Sase	30
1415	tryptophanyl-tRNA Sase	31
0389	tyrosyl-tRNA Sase	39
1007	valyl-tRNA Sase	37
1077	seryl-tRNA Sase	18
	Degradation of proteins, peptides, and glycopeptides	
1176	ATP-dep 26S protease regulatory sub 4	47
1494	ATP-dep 26S protease regulatory sub 8	54
1417	ATP-dep protease La	30
0090	collagenase	33
1130	O-sialoglycoprotein endopeptidase	51
0651	protease IV	35
0591	proteasome, alpha sub	58
1237	proteasome, beta sub	49
0806	XAA-PRO dipeptidase, M24B fam	34
0996	Zn protease	34
	Protein modification	
0814	deoxyhypusine Sase	50
1274	diphthine Sase	42
0172	L-isoaspartyl prot carboxyl MTase	46
1329	Met aminopeptidase	36
1530	N-terminal acetylTase complex, ARD1 sub	40
1591	selenium donor prot	35
	Ribosomal proteins: synthesis and modification	
0509	acidic ribosomal prot P0 (L10E)	63
0242	ribosomal prot HG12	64
1203	ribosomal prot HS6-type	47
0510	ribosomal prot L1	65
0373	ribosomal prot L11	47
0508	ribosomal prot L12	73
0194	ribosomal prot L13	46
0466	ribosomal prot L14	75
0657	ribosomal prot L14B	37
0477	ribosomal prot L15	65
0983	ribosomal prot L15B	55
0474	ribosomal prot L18	74
0473	ribosomal prot L19	69
0179	ribosomal prot L2	74
0040	ribosomal prot L21	55
0460	ribosomal prot L22	40
0178	ribosomal prot L23	70
0467	ribosomal prot L24	73
1201	ribosomal prot L24E	54
0462	ribosomal prot L29	52
0193	ribosomal prot L29E	49
0176	ribosomal prot L3	47
1044	ribosomal prot L30	65
0049	ribosomal prot L31	41
0472	ribosomal prot L32	57
0655	ribosomal prot L34	37
0098	ribosomal prot L37	52
0593	ribosomal prot L37a	45
0177	ribosomal prot L4	50
0707	ribosomal prot L40	58
0249	ribosomal prot L44	39
0689	ribosomal prot L46	52
0469	ribosomal prot L5	72
0471	ribosomal prot L6	67
0476	ribosomal prot L7	72
0595	ribosomal prot LX	39
0322	ribosomal prot S10	68
0191	ribosomal prot S11	62
1046	ribosomal prot S12	83
0036	ribosomal prot S13	50
1474	ribosomal prot S15A	22
0465	ribosomal prot S17	72
0245	ribosomal prot S17B	52
0189	ribosomal prot S18	43
0180	ribosomal prot S19	57
0692	ribosomal prot S19S	46
0394	ribosomal prot S24	43
0250	ribosomal prot S27	41
0393	ribosomal prot S27A	58
0461	ribosomal prot S3	47
1202	ribosomal prot S33	63
0980	ribosomal prot S3a	29
0190	ribosomal prot S4	52
0468	ribosomal prot S4E	71
0475	ribosomal prot S5	74
1260	ribosomal prot S6	37
0620	ribosomal prot S6 modification prot	35
1001	ribosomal prot S6 modification prot II	25
1047	ribosomal prot S7	63
0470	ribosomal prot S8	74
0673	ribosomal prot S8E	50
0195	ribosomal prot S9	51
	Translation factors	
0829	peptide chain release factor, eRF, sub 1	33
1574	ATP-dep RNA helicase, eIF-4A fam	34
1505	ATP-dep RNA helicase, eIF-4A fam	32
0669	ATP-dep RNA helicase, eIF-4A fam	44
0495	put translation factor, EF-TU/1 alpha fam	37
0262	put translation initiation factor, FUN12/IF-2 fam	40
0324	translation elongation factor, EF-1 alpha	80
1048	translation elongation factor, EF-2	75
0445	translation initiation factor, eIF-1A	49
0117	translation initiation factor, eIF-2, alpha sub	34
0097	translation initiation factor, eIF-2, beta sub	33
1261	translation initiation factor, eIF-2, gamma sub	53
0454	translation initiation factor, eIF-2B, alpha sub	38
0122	translation initiation factor, eIF-2B, delta sub	30
1228	translation initiation factor, eIF-5A	50
	tRNA modification	
0946	N2,N2-dimethylguanosine tRNA MTase	33
1675	pseudouridylate Sase I	34
0436	queuine tRNA ribosylTase	30
	Transport and binding proteins	
1572	ABC transporter ATP-BP	36
0719	ABC transporter ATP-BP	50
1023	ABC transporter ATP-BP	50
0035	ABC transporter sub	38
1508	ABC transporter, probable ATP-binding sub	44
1326	GTP-BP	52
1332	GTP-BP	40
1408	GTP-BP, GTP1/OBG-fam	31
1464	put GTP-BP	35
1033	magnesium and cobalt transport prot	43
0091	sodium-calcium exchanger prot	32
0283	nucleotide-BP	45
	Amino acids, peptides, and amines	
0609	amino acid transporter	22
1343	ammonium transport prot AMT1	36
0058	ammonium transporter	35
1269	branched-chain amino acid transport prot livH	31
1266	branched-chain amino acid transport prot livJ	28
1270	branched-chain amino acid transport prot livM	31
1196	cationic amino acid transporter MCAT-2	25
0304	ferripyochelin BP	53
0796	Gln transport ATP-BP Q	48
1267	branched-chain amino acid transport ATP-BP	35
1268	branched-chain amino acid transport ATP-BP	40
	Anions	
0412	nitrate transport ATP-BP	45
0413	nitrate transport permease prot 35	
1012	phosphate transport system ATP-BP	60
1013	phosphate transport system permease prot A	37
1014	phosphate transport system permease prot C	39
1009	phosphate transport system regulatory prot	29
1015	phosphate-BP	42
	Carbohydrates, organic alcohols, and acids	
0576	malic acid transport prot	24
0762	malic acid transport prot	25
0121	SN-glycerol-3-phosphate transport ATP-BP	33
1319	sodium-dep noradrenaline transporter	40
	Cations	
1088	cobalt transport ATP-BP O	46
1090	cobalt transport prot N	46
1089	cobalt transport prot Q	29
0089	ferric enterobactin transport ATP-BP	34
0873	ferric enterobactin transport ATP-BP	36
0566	ferrous iron transport prot B	36
0877	hemin permease	34
0087	hemin permease	38
0085	iron transport system BP	33
0876	iron(III) dicitrate transport system permease prot	32
1441	magnesium chelatase sub 36	
0911	magnesium chelatase sub 56	
1275	sodium-hydrogen antiporter	30
0672	sodium transporter	40
1231	oxaloacetate DCase, alpha sub	53
1357	put potassium channel prot	30
1367	sulfate permease	38
1368	sulfate/thiosulfate transport prot	30
1485	TRK system potassium uptake prot	29
1105	TRK system potassium uptake prot A	35
	Other	
1142	ATPase, arsenical pump-driving	35
0822	ATPase, vanadate-senstive	49
0718	chromate resistance prot A	28
1226	ATPase, hydrogen transporting	45
1560	quinolone resistance norA prot	29
	Other categories	
	Drug and analog sensitivity	
1538	toxin sensitivity prot KTI12	29
0102	phenylacrylic acid DCase	46
	Phage-related functions and prophages	
0630	sodium-dep phosphate transporter	33
	Transposon-related functions	
0367	integrase	31
0017	transposase	30
1466	transposase	30
	Other	
1064	acetylTase	47
1612	phosphonopyruvate DCase	32
0677	ethylene-inducible prot homolog	67
0534	flavoprotein	35
0748	flavoprotein	68
0256	phosphonopyruvate DCase	30
1682	heat shock prot X	31
0866	HIT prot, member of the HIT-fam	40
0294	large helicase rel prot, LHR	32
0010	phosphonopyruvate DCase	28
0734	rubrerythrin	49
0559	survival prot surE	35
1100	urease operon prot	34
0543	Wilm's tumor suppressor homolog	44
0765	[6Fe-6S] prismane-containing prot	61
1365	pheromone shutdown prot	31
ECL24	SOJ prot	36

contains 12 predicted protein-coding regions and has a G+C content of 28.8% (Fig. 2). The sequences of the M. *jannaschii* chromosome and of the large and small ECEs have been deposited in the Genome Sequence DataBase with the accession numbers L77117, L77118, and L77119, respectively. The annotated genome sequence data and clone information for M. *jannaschii* are available on the World Wide Web (http://www.tigr.org/tdb/mdb/mjdb/mjdb.html).

Of the 1743 predicted protein-coding regions reported previously for *H. influenzae*, 78% had a match in the public sequence database (*3*). Of these, 58% were matches to genes with reasonably well defined function, whereas 20% were matches to genes whose function was undefined. Similar observations were made for the M. *genitalium* genome (*3*). Of the predicted protein-coding regions from M. *genitalium*, 83% have a counterpart in the *H. influenzae* genome. In contrast, only 38% of the predicted protein-coding regions from M. *jannaschii* match a gene in the database that could be assigned a putative cellular role with high confidence; 6% of the predicted protein-coding regions had matches to hypothetical proteins (Fig. 3 and Table 1). Approximately 100 genes in M. *jannaschii* had marginal similarity to genes or segments of genes from the public sequence databases and could not be assigned a putative cellular role with high confidence. Only 11% of the predicted protein-coding regions from *H. influenzae* and 17% of the predicted protein-coding regions from M. *genitalium* matched a predicted protein-coding region from M. *jannaschii*.

Energy production in M. *jannaschii* occurs by the reduction of CO_2 with H_2 to produce methane. Genes for all of the known enzymes and enzyme complexes associated with methanogenesis (*10*) were identified in M. *jannaschii*, the sequence and order of which are typical of methanogens. *Methanococcus jannaschii* appears to use both H_2 and formate as substrates for methanogenesis, but lacks the genes to use methanol or acetate. The ability to fix nitrogen has been demonstrated in a number of methanogens (*11*), and all the genes necessary for this pathway have been identified in M. *jannaschii* (Table 1). In addition to its anabolic pathways, several scavenging molecules have been identified in M. *jannaschii* that probably play a role in importing small organic compounds, such as amino acids, from the environment (Table 1).

Three different pathways control the fixation of CO_2 into organic carbon: the noncyclic, reductive acetyl–coenzyme A–carbon monoxide dehydrogenase pathway (Ljungdahl-Wood pathway), the reductive trichloroacetic acid cycle, and the Calvin cycle. Methanogens fix carbon by the Ljungdahl-Wood pathway (*12*), which is facilitated by the carbon monoxide dehydrogenase enzyme complex (*13*). The complete Ljungdahl-Wood pathway, encoded in the M. *jannaschii* genome, depends on the methyl carbon in methanogenesis; however, methanogenesis can occur independently of carbon fixation.

Although genes encoding two enzymes required for gluconeogenesis (glucopyruvate oxidoreductase and phosphoenolpyruvate synthase) were found in the M. *jannaschii* genome, genes encoding other key intermediates of gluconeogenesis (fructose bisphosphatase and fructose 1,6-bisphosphate aldolase) were not identified. Glucose catabolism by glycolysis also requires the aldolase, as well as phosphofructokinase, an enzyme that also was not found in M. *jannaschii* and has not been detected in any of the Archaea. In addition, genes specific for the Entner-Doudoroff pathway, an alternative pathway used by some microbes for the catabolism of glucose, were not identified in the genomic sequence. The presence of a number of nearly complete metabolic pathways suggests that some key genes are not recognizable at the sequence level, although we cannot exclude the possibility that M. *jannaschii* may use alternative metabolic pathways.

In general, the M. *jannaschii* genes that encode proteins involved in the transport of small inorganic ions into the cell are homologs of bacterial genes. The genome includes many representatives of the ABC transporter family, as well as genes for exporting heavy metals (for example, the chromate-resistance protein) and other toxic compounds (for example, the *norA* drug efflux pump locus).

More than 20 predicted protein-coding regions have sequence similarity to polysaccharide biosynthetic enzymes. These genes have only bacterial homologs or are most closely related to their bacterial counterparts. The identified polysaccharide biosynthetic genes in M. *jannaschii* include those for the interconversion of sugars, activation of sugars to nucleotide sugars, and glycosyltransferases for the polymerization of nucle-

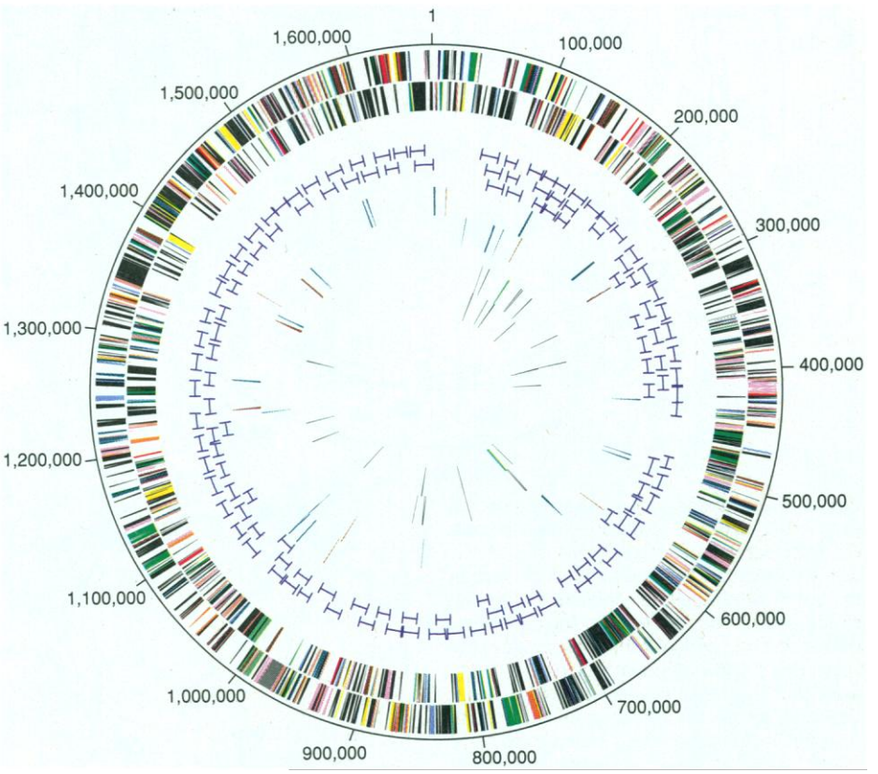


Fig. 1. A circular representation of the *M. jannaschii* chromosome illustrating the location of each predicted protein-coding region as well as selected features of the genome. Outer concentric circle: predicted protein-coding regions on the plus strand color-coded according to role as indicated in Fig. 3. Second concentric circle: predicted protein-coding regions on the minus strand color-coded according to role as indicated in Fig. 3. Third concentric circle: coverage by λ clones (three levels of blue range bars). Fourth and fifth concentric circles representing the plus and minus strands, respectively: members of the IS*AMJ1* family (red) and repetitive elements (cyan). Sixth and seventh concentric circles representing the plus and minus strands, respectively: transfer RNAs (black) and ribosomal RNAs (green).

otide sugars into oligo- and polysaccharides that are subsequently incorporated into surface structures (*14*). In an arrangement similar to that of bacterial polysaccharide biosynthesis genes, many of the genes for M. *jannaschii* polysaccharide production are clustered together (Table 1 and Fig. 3). The G+C content in this region is <95% of that in the rest of the M. *jannaschii* genome. A similar observation was made in *Salmonella typhimurium* (*15*), in which the gene cluster for lipopolysaccharide O antigen has a significantly lower G+C ratio than that in the rest of the genome. In that case, the difference in G+C content was interpreted as meaning that the region originated by lateral transfer from another organism.

Of the three main multicomponent information-processing systems (transcription, translation, and replication), translation appears to be the most universal in its overall makeup in that the basic translation machinery is similar in all three domains of life. *Methanococcus jannaschii* has two ribosomal RNA operons, designated A and B, and a separate 5S RNA gene that is associated with several transfer RNAs (tRNAs). Operon A has the organization 16S-23S-5S, whereas operon B lacks the 5S component. An alanine tRNA is situated in the spacer region between the 16S and 23S subunits in both operons. The majority of proteins associated with the ribosomal subunits (especially the small subunit) are present in both Bacteria and Eukaryotes. However, the relatively protein-rich eukaryotic ribosome contains additional ribosomal proteins not found in the bacterial ribosome. A smaller number of Bacteria-specific ribosomal proteins exist as well. The M. *jannaschii* genome contains all ribosomal proteins that are common to Eukaryotes and Bacteria. It shows no homologs of the bacterial-specific ribosomal proteins, but does possess homologs of a number of the eukaryotic-specific ones. Homologs of all archaeal-specific ribosomal proteins that have been reported to date (*16*) are found in M. *jannaschii*.

As shown for other Archaea (*2*), the *Methanococcus* translation elongation factors EF-1α (EF-Tu in Bacteria) and EF-2 (EF-G in Bacteria) are most similar to their eukaryotic counterparts. In addition, the M. *jannaschii* genome contains 11 translation-initiation factor genes. Three of these genes encode the subunits homologous to those of the eukaryotic IF-2 and are reported here in the Archaea for the first time. A fourth initiation factor gene that encodes a second IF-2 is also found in M. *jannaschii*. This additional IF-2 gene is most similar to the yeast protein FUN12 (*17*) which, in turn, appears to be a homolog of the bacterial IF-2. It is not known which of the two IF-2–like initiation factors identified in M. *jannaschii* plays a role in directing the initiator tRNA to the start site of the mRNA. The fifth identified initiation factor gene in M. *jannaschii* encodes IF-1A, which has no bacterial homolog. The sixth gene encodes the hypusine-containing initiation factor eIF-5a. Two subunits of the translation initiation factor eIF-2B were identified in M. *jannaschii*. Finally, three putative adenosine triphosphate–dependent helicases were identified that belong to the eIF-4a family of translation initiation factors.

Thirty-seven tRNA genes were identified in the M. *jannaschii* genome. Almost all amino acids encoded by two codons have a single tRNA, except for glutamic acid, which has two. Both an initiator and an internal methionyl tRNA are present. The two pyrimidine-ending isoleucine codons are covered by a single tRNA, whereas the third (AUA) seems covered by a related tRNA having a CAU anticodon. A single tRNA appears to cover the three isoleucine codons. Those amino ac-

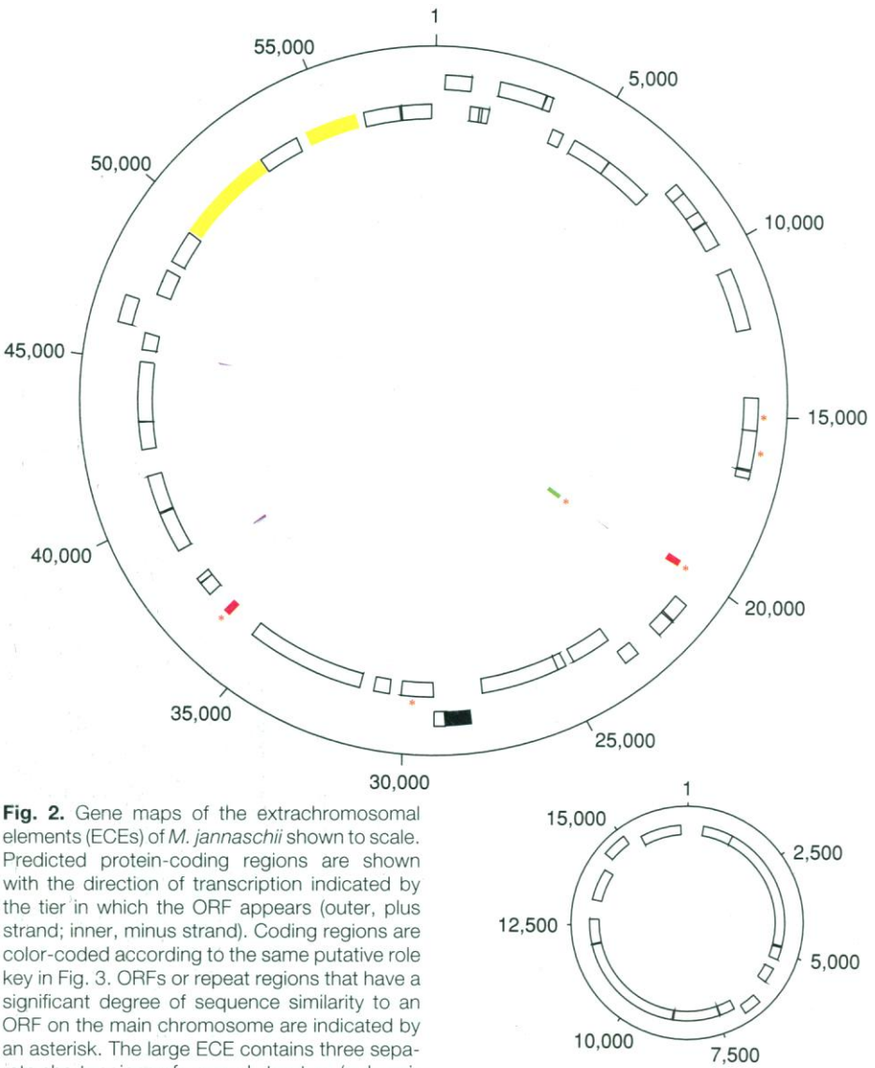


Fig. 2. Gene maps of the extrachromosomal elements (ECEs) of *M. jannaschii* shown to scale. Predicted protein-coding regions are shown with the direction of transcription indicated by the tier in which the ORF appears (outer, plus strand; inner, minus strand). Coding regions are color-coded according to the same putative role key in Fig. 3. ORFs or repeat regions that have a significant degree of sequence similarity to an ORF on the main chromosome are indicated by an asterisk. The large ECE contains three separate short regions of unusual structure (polypyrimidine or polypurine stretches of high G+C content) that show mirror symmetry. These regions are indicated in the third concentric circle by blue rectangles. In all three instances, the core of the mirror structure has the sequence CCCTCTCGGG-CTCTCCC (or its complement). Approximate mirror symmetry extends beyond this core, for stretches of (total length) 15, 19, and 21 pyrimidines (or purines) on either side of the center of symmetry. Mirror symmetry of this sort is characteristic of DNA capable of forming triple-stranded structures (*33*). The green rectangle in the innermost concentric circle of the large ECE indicates the location of the group C member of the IS*AMJ1* family of insertion elements.

ids encoded by four codons each have two tRNAs, one to cover the Y-, the other the R-ending, codons. Valine has a third tRNA, which is specific for the GUG codon; and alanine has three tRNAs (two of which are in the spacer regions separating the 16S and 23S subunits in the two ribosomal RNA operons). Leucine, serine, and arginine, all of which have six codons, each possess three corresponding tRNAs. The genes for the internal methionine and tryptophan tRNAs contain introns in their anticodon loops.

A tRNA also exists for selenocysteine (UGA codon). At least four genes in M. *jannaschii* contain internal stop codons that are potential selenocysteine codons: the α chain of formate dehydrogenase, coenzyme F420-reducing hydrogenase, B-chain tungsten formyl–methanofuran dehydrogenase, and a heterodisulfide reductase. Three genes with a putative role in selenocysteine metabolism were identified by their similarity to the *sel* genes from other organisms (Table 1).

Recognizable homologs for four of the aminoacyl-tRNA synthetases (glutamine, asparagine, lysine, and cysteine) were not identified in the M. *jannaschii* genome. The absence of a glutaminyl-tRNA synthetase is not surprising given that a number of organisms, including at least one archaeon, have none (*18*). In these instances, glutaminyl tRNA charging involves a post-charging conversion mechanism whereby the tRNA is charged by the glutamyl-tRNA synthetase with glutamic acid, which then is enzymatically converted to glutamine. A post-charging conversion is also involved in selenocysteine charging by the seryl-tRNA synthetase. A similar mechanism has been proposed for asparagine charging, but has not been demonstrated (*18*). The inability to find homologs of the lysine and cysteine aminoacyl-tRNA synthetases is surprising because bacterial and eukaryotic versions in each instance show clear homology.

Aminoacyl-tRNA synthetases of M. *jannaschii* and other Archaea resemble eukaryotic synthetases more closely than they resemble bacterial forms. The tryptophanyl synthetase is one of the more notable examples, because the M. *jannaschii* and eukaryotic versions do not appear to be specifically related to the bacterial version (*19*). Two versions of the glycyl synthetase are present in Bacteria, one that is very unlike the version found in Archaea and Eukaryotes and one that is an obvious homolog of it (*20*).

Eleven genes encoding subunits of the DNA-dependent RNA polymerase were identified in the M. *jannaschii* genome. The sequence similarity between the subunits and their homologs in *Sulfolobus acidocaldarius* supports the evolutionary unity of the archaeal polymerase complex (*21*). All of the subunits found in M. *jannaschii* show greater similarity to their eukaryotic counterparts than to the bacterial homologs. The genes encoding the five largest subunits (A′, A″, B′, B″, D) have homologs in all organisms. Six genes encode subunits shared only byArchaea and Eukaryotes (E, H, K, L, and N). The M. *jannaschii* homolog of the S. *acidocaldarius* subunit E is split into two genes designated E′ and E″. *Sulfolobus acidocaldarius* also contains two additional small subunits of RNA polymerase, designated G and F, that have no counterparts in either Bacteria or Eukaryotes. No homolog of these F subunits was identified in M. *jannaschii*.

The archaeal transcription initiation system is essentially the same as that found in Eukaryotes and is radically different from the bacterial version (*22*). The central molecules in the former systems are the TATA-binding protein (TBP) and transcription factor B (TFIIB and TFIIIB in Eukaryotes, or simply TFB). In the eukaryotic systems, TBP and TFB are parts of larger complexes, and additional factors (such as TFIIA and TFIIF) are used in the transcription process. However, the M. *jannaschii* genome does not contain obvious homologs of TFIIA and TFIIF.

Several components of the replication machinery were identified in M. *jannaschii*. The M. *jannaschii* genome appears to encode a single DNA-dependent polymerase that is a member of the B family of polymerases (*23*). The polymerase shares sequence similarity and three motifs with other family B polymerases, including eukaryotic α, γ, and ε polymerases, bacterial polymerase II, and several archaeal polymerases. However, it is not homologous to bacterial polymerase I and has no homologs in *H. influenzae* or M. *genitalium*.

Primer recognition by the polymerase takes place through a structure-specific DNA binding complex, the replication factor complex (rfc) (*23*). In humans and yeast, the rfc is composed of five proteins: a large subunit and four small subunits that have an associated adenosine triphosphatase (ATPase) activity stimulated by proliferating cell nuclear antigen (PCNA). Two genes in M. *jannaschii* are putative members of a eukaryotic-like replication factor complex. One of the genes in M. *jannaschii* is a putative homolog of the large subunit of the rfc, whereas the second is a putative homolog of one of the small subunits. Among Eukaryotes, the rfc proteins share sequence similarity in eight signature domains (*23*). Domain I is conserved only in the large subunit among Eukaryotes and is similar in sequence to DNA ligases. This domain is missing in the large-subunit homolog in M. *jannaschii*. The remaining domains in the two M. *jannaschii* genes are well conserved relative to the eukaryotic homologs. Two features of the sequence similarity in these domains are of particular interest. First, domain II (an ATPase domain) of the small-subunit homolog is split between two highly conserved amino acids (lysine and threonine) by an intervening sequence of unknown function. Second, the sequence of domain VI has regions that are useful for distinguishing between bacterial and eukaryotic rfc proteins (*23*); the rfc sequence for M. *jannaschii* shares the characteristic eukaryotic signature in this domain.

We attempted to identify an origin of replication by searching the M. *jannaschii* genome sequence with a variety of bacterial and eukaryotic replication-origin consensus sequences. Searches with oriC, ColE1, and autonomously replicating sequences from yeast (*23*) did not identify an origin of replication. With respect to the related cellular processes of replication initiation and cell division, the M. *jannaschii* genome contains two genes that are putative homologs of Cdc54, a yeast protein that belongs to a family of putative DNA replication initiation proteins (*24*). A third potential regulator of cell division in M. *jannaschii* is 55% similar at the amino acid level to *pelota*, a *Drosophila* protein involved in the regulation of the early phases of meitoic and mitotic cell division (*25*).

In contrast to the putative rfc complex and the initiation of DNA replication, the cell division proteins from M. *jannaschii* most resemble their bacterial counterparts (*26*). Two genes similar to that encoding FtsZ, a ubiquitous bacterial protein, are found in M. *jannaschii*. FtsZ is a polymer-forming, guanosine triphosphate (GTP)–hydrolyzing protein with tubulin-like elements; it is localized to the site of septation and forms a constricting ring between the dividing cells. One gene similar to FtsJ, a bacterial cell division protein of undetermined function, also is found in M. *jannaschii*. Three additional genes (MinC, MinD, and MinE) function in concert in Bacteria to determine the site of septation during cell division. In M. *jannaschii*, three MinD-like genes were identified, but none for MinC or MinE. Neither spindle-associated proteins characteristic of eukaryotic cell division nor bacterial mechanochemical enzymes necessary for partitioning the condensed chromosomes were detected in the M. *jannaschii* genome. Taken together, these observations raise the possibility that cell division in M. *jannaschii* might occur by

a mechanism specific for the Archaea.

The structural and functional conservation of the signal peptide of secreted proteins in Archaea, Bacteria, and Eukaryotes suggests that the basic mechanisms of membrane targeting and translocation may be similar among all three domains of life. The secretory machinery of M. *jannaschii* appears to be a rudimentary apparatus relative to that of bacterial and eukaryotic systems and consists of (i) a signal peptidase (SP) that cleaves the signal peptide of translocating proteins, (ii) a preprotein translocase that is the major constituent of the membrane-localized translocation channel, (iii) a ribonucleoprotein complex (signal recognition particle, SRP) that binds to the signal peptide and guides nascent proteins to the cell membrane, and (iv) a docking protein that acts as a receptor for the SRP. The 7S RNA component of the SRP from M. *jannaschii* shows a highly conserved structural domain shared by other Archaea, Bacteria, and Eukaryotes (27). However, the predicted secondary structure of the 7S RNA SRP component in Archaea is more like that found in Eukaryotes than in Bacteria (27). The SP and docking proteins from M. *jannaschii* are most similar to their eukaryotic counterparts; the translocase is most similar to the SecY translocation-associated protein in *Escherichia coli*.

A second distinct signal peptide is found in the flagellin genes of M. *jannaschii*. Alignment of flagellin genes from M. *voltae* (28) and M. *jannaschii* reveals a highly conserved NH_2-terminus (31 of the first 50 residues are identical in all of the mature flagellins). The peptide sequence of the M. *jannaschii* flagellin indicates that the protein is cleaved after the canonical Gly-12 position, and it is proposed to be similar to type-IV pilins of Bacteria (28).

Five histone genes are present in the M. *jannaschii* genome—three on the main chromosome and two on the large ECE. These genes are homologs of eukaryotic histones (H2a, H2b, H3, and H4) and of the eukaryotic transcription-related CAAT-binding factor CBF-A (29). The similarity between archaeal and eukaryotic histones suggests that the two groups of organisms resemble one another in the roles histones play both in genome supercoiling dynamics and in gene expression. The five M. *jannaschii* histone genes show greatest similarity among themselves even though a histone sequence is available from the closely related species, *Methanococcus voltae*. This intraspecfic similarity suggests that the gene duplications that produced the five histone genes occurred on the M. *jannaschii* lineage per se.

Self-splicing portions of a peptide sequence that generally encode a DNA endonuclease activity are called inteins, in analogy to introns (30). The sequences remaining after an intein is excised are called exteins, in analogy to exons. Exteins are spliced together after the excision of one or more inteins to form functional proteins. The biological significance and role of inteins are not clearly understood (30). Fourteen genes in the M. *jannaschii* genome contain 18 putative inteins, a significant increase in the approximately 10 intein-containing genes that have been described (30) (Table 2). The only previously described inteins in the Archaea are in the DNA polymerase genes of the Thermococcales (30). The M. *jannaschii* DNA polymerase gene has two inteins in the same locations as those in *Pyrococcus* sp. strain KOD1. In this case, the exteins exhibit 46% amino acid identity, whereas intein 2 of the two organisms has only 33% identity. This divergence suggests that intein 2 has

Table 2. Genes of *M. jannaschii* that contain inteins.

Gene no.	Putative identification	No. of inteins
MJ0043	Hypothetical protein (*Bacillus subtilis*)	1
MJ0262	Putative translation initiation factor, FUN12/bIF-2 family	1
MJ0542	Phosphoenolpyruvate synthase	1
MJ0682	Hypothetical protein (*Escherichia coli*)	1
MJ0782	Transcription initiation factor IIB	1
MJ0832	Anaerobic ribonucleoside-triphosphate reductase	2
MJ0885	DNA-dependent DNA polymerase, family B	2
MJ1042	DNA-dependent RNA polymerase, subunit A′	1
MJ1043	DNA-dependent RNA polymerase, subunit A″	1
MJ1054	UDP-glucose dehydrogenase	1
MJ1124	Hypothetical protein (*Saccharomyces cerevisiae*)	1
MJ1420	Glutamine–fructose-6-phosphate transaminase	1
MJ1442	Replication factor C, 37-kD subunit	3
MJ1512	Reverse gyrase	1

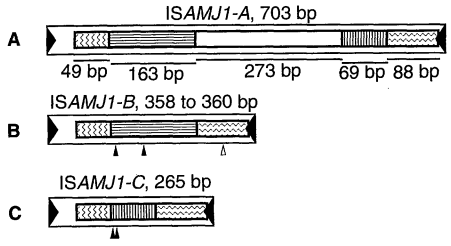


Fig. 4. Structure of a putative family of insertion sequence (IS) elements in the *M. jannaschii* genome. The family of elements has been named IS*AMJ1* and contains 11 members distributed among three groups (**A**, **B**, and **C**). The outer rectangle indicates the entire IS element; the interior rectangles indicate the predicted coding regions, oriented with the NH_2-termini to the left. DNA immediately adjacent to the NH_2-termini is 75 to 100% identical over 50 bp; DNA sequence similarity at the COOH-termini ends immediately after the stop codon. Black triangles indicate terminal inverted repeats. Fill patterns indicate which regions are missing from the elements in groups B and C. (A) Two copies of this family are 642 bp long and are 97% similar to each other at the nucleotide level. They appear to encode a protein 214 amino acids in length (ORFs MJ0017 and MJ1466) that are 27% identical to the IS*240* transposase of *B. thuringiensis* (GenBank accession number: M23741). (B) Eight copies of the family range in length from 358 to 360 bp and are missing a 342-bp internal region relative to the two members of group A. Some members of group B have putative frameshifts (indicated by solid arrows) and in-frame UGA codons (indicated by open arrows). (C) The single copy in group C is 265 bp in length and occurs on the large ECE. The 436-bp internal region missing from this element is different than that of the members of group B.

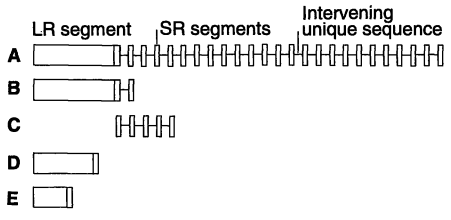


Fig. 5. Structure of a multicopy repetitive element in the *M. jannaschii* genome. Of the 18 copies identified on the main chromosome, 7 are oriented in one direction (plus strand) and 11 are oriented in the opposite strand. Each element consists of a long, 391- to 425-bp repeat segment (designated LR) followed by up to 25 short, 27- to 28-bp repeat segments (designated SR). Each SR segment is separated by 31 to 51 bp of sequence that is unique within and between each complete repeat element. (**A**) The longest repeat element has an LR segment followed by 25 SR segments and spans more than 2 kbp, and (**B**) the shortest complete element has an LR segment followed by two SR segments. (**C**) One element is present in the genome with five SR segments and no LR component. (**D** and **E**) The LR segments of two elements in the genome are truncated at the end adjacent to the SR segments; both are followed by a single SR segment.

not been recently (laterally) transferred between the Thermococcales and *M. jannaschii*. In contrast, the intein 1 sequences are 56% identical, more than that of the gene containing them, and comparable to the divergence of inteins within the Thermococcales. This high degree of sequence similarity might be the result of an intein transfer more recent than the splitting of these species. The large number of inteins found in *M. jannaschii* led us to question whether these inteins have been increasing in number by moving within the genome. If this were so, we would expect to find some pairs of inteins that are particularly similar. Comparisons of these and other available intein sequences showed that the closest relationships are those noted above linking the DNA polymerase inteins to correspondingly positioned elements in the Thermococcales. Within *M. jannaschii*, the highest identity observed was 33% for a 380-bp portion of two inteins. This finding suggests that the diversification of the inteins predates the divergence of the *M. jannaschii* and *Pyrococcus* DNA polymerases.

Three families of repeated genetic elements were identified in the *M. jannaschii* genome. Within two of the families, at least two members were identified as ORFs with a limited degree of sequence similarity to bacterial transposases. Members of the first family, designated ISAMJ1, are repeated 10 times on the main chromosome and once on the large ECE (Fig. 4). There is no sequence similarity between the IS elements in *M. jannaschii* and the ISM1 mobile element described previously for *Methanobrevibacter smithii* (*31*). Two members of this family were identified as ORFs and are 27% identical (at the amino acid sequence level) to a transposase from *Bacillus thuringiensis* (IS*240*; GenBank accession number M23741). Relative to these two members, the remaining members of the ISAMJ1 family are missing an internal region of several hundred nucleotides (Fig. 4). With one exception, all members of this family end with 16-bp terminal inverted repeats typical of insertion sequences. One member is missing the terminal repeat at its 5′ end. The second family consists of two ORFs that are identical across 928 bp. The ORFs are 23% identical at the amino acid sequence level to the COOH-terminus of a transposase from *Lactococcus lactis* (IS982; GenBank accession number L34754). Neither of the members of the second family contains terminal inverted repeats.

Eighteen copies of the third family of repeated genetic structures (Fig. 5) are distributed fairly evenly around the *M. jannaschii* genome (Fig. 3). Unlike the genetic elements described above, none of the components of this repeat unit appears to have coding potential. The repeat structure is composed of a long segment followed by 1 to 25 tandem repetitions of a short segment. The short segments are separated by sequence that is unique within and among the complete repeat structure. Three similar types of short segments were identified; however, the type of short repeat is consistent within each repeat structure, except for variation of the last short segment in six repeat structures. Similar tandem repeats of short segments have been observed in Bacteria and other Archaea (*32*) and have been

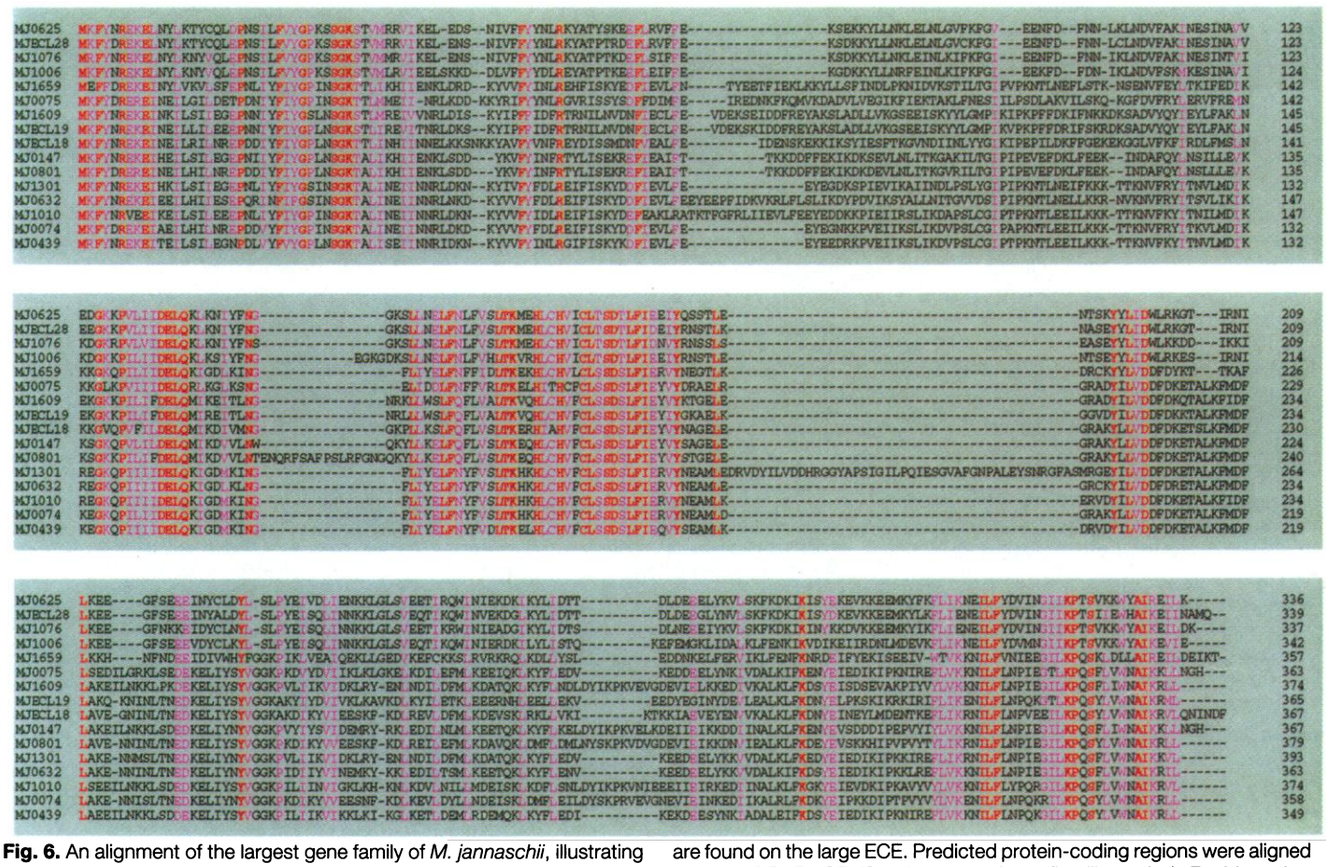


Fig. 6. An alignment of the largest gene family of *M. jannaschii*, illustrating 16 paralogous genes that have no database matches or recognizable motifs relative to previously published sequences. These proteins contain many charged residues; no regions of hydrophobicity were detected. Three members of the gene family, those designated by MJECL numbers, are found on the large ECE. Predicted protein-coding regions were aligned with the GENEWORKS software package (Intelligenetics). Residues that are invariant among the 16 sequences are shaded red; residues that are invariant in >80% of the sequences (or are substituted conservatively) are shaded pink.

hypothesized to participate in chromosome partitioning during cell division.

The 16-kbp ECE from M. *jannaschii* contains 12 ORFs, none of which had a significant full-length match to any published sequence (Fig. 2). The 58-kbp ECE contains 44 predicted protein-coding regions, 5 of which had matches to genes in the database. Two of the genes are putative archael histones, one is a sporulation-related protein (SOJ protein), and two are type I restriction modification enzymes. There are several instances in which predicted protein-coding regions or repeated genetic elements on the large ECE have similar counterparts on the main chromosome of M. *jannaschii* (Fig. 2). The degree of nucleotide sequence similarity between genes present on both the ECE and the main chromosome ranges from 70 to 90%, suggesting that there has been relatively recent exchange of at least some genetic material between the large ECE and the main chromosome.

All the predicted protein-coding regions from M. *jannaschii* were searched against each other in order to identify families of paralogous genes (genes related by gene duplication, not speciation). The initial criterion for grouping paralogs was >30% amino acid sequence identity over 50 consecutive amino acid residues. Groups of predicted protein-coding regions were then aligned and inspected individually to ensure that the sequence similarity extended over most of their lengths. This curatorial process resulted in the identification of more than 100 gene families, half of which have no database matches. The largest identified gene family (16 members) (Fig. 6) contains almost 1% of the total predicted protein-coding regions in M. *jannaschii*. The gene family alignments for M. *jannaschii* are available on the World Wide Web (http://www.tigr.org/tdb/mdb/mjdb/).

Despite the availability for comparison of two complete bacterial genomes and several hundred megabase pairs of eukaryotic sequence data, the majority of genes in M. *jannaschii* cannot be identified on the basis of sequence similarity. Previous evidence for the shared common ancestry of the Archaea and Eukaryotes was based on a small set of gene sequences (2). The complete genome of M. *jannaschii* allows us to move beyond a "gene by gene" approach to one that encompasses the larger picture of metabolic capacity and cellular systems. The anabolic genes of M. *jannaschii* (especially those related to energy production and nitrogen fixation) reveal an ancient metabolic world shared largely by Bacteria and Archaea. That many basic autotrophic pathways appear to have a common evolutionary origin suggests that the most recent universal common ancestor to all three domains of extant life had the capacity for autotrophy. The Archaea and Bacteria also share structural and organizational features that the most recent universal prokaryotic ancestors also likely possessed, such as circular genomes and genes organized as operons. In contrast, the cellular information-processing and secretion systems in M. *jannaschii* demonstrate the common ancestry of Eukaryotes and Archaea. Although components of these systems are present in all three domains, their apparent refinement over time—especially transcription and translation—indicate that the Archaea and Eukaryotes share a common evolutionary trajectory independent of the lineage of Bacteria.

REFERENCES AND NOTES

1. G. E. Fox *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4537 (1977); C. R. Woese and G. E. Fox, *ibid.*, p. 5088; C. R. Woese *et al.*, *ibid.* **87**, 4576 (1990).
2. N. Iwabe *et al.*, *ibid.* **86**, 9355 (1989); J. P. Gogarten *et al.*, *ibid.*, p. 6661; W. Zillig *et al.*, *Endocytobiosis Cell Res.* **6**, 1 (1989); J. R. Brown and W. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 2441 (1995).
3. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995).
4. N. Williams, *ibid.* **272**, 481 (1996).
5. M. D. Adams *et al.*, *Nature* **377**, 3 (1995); R. Wilson *et al.*, *ibid.* **368**, 32 (1994).
6. W. Jones *et al.*, *Arch. Microbiol.* **136**, 254 (1983).
7. G. Sutton *et al.*, *Genome Sci. Tech.* **1**, 9 (1995).
8. The statistical prediction of *M. jannaschii* genes was performed with GeneMark [M. Borodovsky and J. McIninch, *Comput. Chem.* **17**, 123 (1993)]. Regular GeneMark uses nonhomogeneous Markov models derived from a training set of coding sequences and ordinary Markov models derived from a training set of noncoding sequences. Only a single 16S ribosomal RNA sequence of *M. jannaschii* was available in the public sequence databases before the whole genome sequence described here. Thus, the initial training set to determine parameters of a coding sequence Markov model was chosen as a set of ORFs >1000 nucleotides (nt). As an initial model for noncoding sequences, a zero-order Markov model with genome-specific nucleotide frequencies was used. The initial models were used at the first prediction step. The results of the first prediction were then used to compile a set of putative genes used at the second training step. Alternate rounds of training and predicting were continued until the set of predicted genes stabilized and the parameters of the final fourth-order model of coding sequences were derived. The regions predicted as noncoding were then used as a training set for a final model for noncoding regions. Cross-validation simulations demonstrated that the GeneMark program trained as described above was able to correctly identify coding regions of at least 96 nt in 94% of the cases and noncoding regions of the same length in 96% of the cases. These values assume that the self-training method produced correct sequence annotation for compiled control sets. Comparison with the results obtained by searches against a nonredundant protein database (*3*) demonstrated that almost all genes identified by sequence similarity were predicted by the GeneMark program as well. This observation provides additional confidence in genes predicted by GeneMark whose protein translations did not show significant similarity to known protein sequences. The predicted protein-coding regions were searched against the Blocks database [S. Henikoff and J. G. Henikoff, *Genomics* **19**, 97 (1994)] by means of BLIMPS [J. C. Wallace and S. Henikoff, *Comput. Appl. Biosci.* **8**, 249 (1992)] to verify putative identifications and to identify potential functional motifs in predicted protein-coding regions that had no database match. Genes were assigned to known metabolic pathways. When a gene appeared to be missing from a pathway, the unassigned ORFs and the complete *M. jannaschii* genome sequence were searched with specific query sequences or motifs from the Blocks database. Hydrophobicity plots were performed on all predicted protein-coding regions by means of the Kyte-Doolittle algorithm [J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**, 105 (1982)] to identify potentially functionally relevant signatures in these sequences. The results of the Blocks and Kyte-Doolittle analyses are available on the World Wide Web (http://www.tigr. org/tdb/mdb/mjdb/mjdb.html).
9. H. Zhao *et al.*, *Arch. Microbiol.* **150**, 178 (1988).
10. A. A. DiMarco *et al.*, *Annu. Rev. Biochem.* **59**, 355 (1990).
11. N. Belay *et al.*, *Nature* **312**, 286 (1984).
12. H. G. Wood *et al.*, *Trends Biochem. Sci.* **11**, 14 (1986).
13. M. Blaat, *Antonie Leewenhoek* **66**, 187 (1994).
14. E. Hartmann and H. König, *Arch. Microbiol.* **151**, 274 (1989).
15. X. M. Jiang *et al.*, *Mol. Microbiol.* **5**, 695 (1991).
16. K. Lechner *et al.*, *J. Mol. Evol.* **29**, 20 (1989); A. K. E. Köpke and B. Wittmann-Liebold. *Can. J. Microbiol.* **35**, 11 (1989).
17. P. Keeling *et al.*, *Syst. Appl. Microbiol.*, in press.
18. M. Wilcox, *Eur. J. Biochem.* **11**, 405 (1969); N. C. Martin *et al.*, *J. Mol. Biol.* **101**, 285 (1976); N. C. Martin *et al.*, *Biochemistry* **16**, 4672 (1977); A. Schon *et al.*, *Biochimie* **70**, 391 (1988); D. Soll and U. Raj Bhandary, Eds. *tRNA: Structure, Biosynthesis, and Function* (American Society for Microbiology, Washington, DC, 1995).
19. R. de Pouplana *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 166 (1996).
20. E. A. Wagner *et al.*, *J. Bacteriol.* **177**, 5179 (1995); D. T. Logan *et al.*, *EMBO J.* **14**, 4156 (1995).
21. C. R. Woese and R. S. Wolfe, Eds. *The Bacteria* (Academic Press, New York, 1985), vol. 8; D. Langer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5768 (1995); M. Lanzendoerfer *et al.*, *Syst. Appl. Microbiol.* **16**, 656 (1994).
22. H.-P. Klenk and W. F. Doolittle, *Curr. Biol.* **4**, 920 (1994).
23. A. Bernard *et al.*, *EMBO J.* **6**, 4219 (1987); G. Cullman *et al.*, *Mol. Cell. Biol.* **15**, 4661 (1995); T. Uemori *et al.*, *J. Bacteriol.* **177**, 2164 (1995); M. Delarue *et al.*, *Protein Eng.* **3**, 461 (1990); K. A. Gavin, M. Hidaka, B. Stillman, *Science* **270**, 1667 (1995).
24. L. A. Whitbred and S. Dalton, *Gene* **155**, 113 (1995).
25. C. G. Eberhart and S. A. Wasserman, *Development* **121**, 3477 (1995).
26. L. Rothfield and C.-R Zhao, *Cell* **84**, 183 (1996); J. Lutkenhaus, *Curr. Opin. Genet. Dev.* **3**, 783 (1993).
27. B. P. Kaine and V. L. Merkel, *J. Bacteriol.* **171**, 4261 (1989); M. A. Poritz *et al.*, *Cell* **55**, 4 (1988).
28. D. M. Faguy *et al.*, *Can. J. Microbiol.* **40**, 67 (1994): M. L. Kalmokoff *et al.*, *Arch. Microbiol.* **157**, 481 (1992).
29. K. Sandman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 5788 (1990).
30. P. M. Kane *et al.*, *Science* **250**, 651 (1990); R. Hirata *et al.*, *J. Biol. Chem.* **265**, 6726 (1990); A. A. Cooper and T. Stevens, *Trends Biochem. Sci.* **20**, 351 (1995); M.-Q Xu *et al.*, *Cell* **75**, 1371 (1993); F. Perler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5577 (1992); Cooper *et al.*, *EMBO J.* **12**, 2575 (1993); F. Michel *et al.*, *Biochimie* **64**, 867 (1982); S. Pietrokovski, *Protein Sci.* **3**, 2340 (1994). Most inteins in the *M. jannaschii* genome were identified by (i) similarity of the bounding exteins to other proteins, (ii) similarity of the inteins to those previously described, (iii) presence of the dodecapeptide endonuclease motifs, and (iv) canonical intein-extein junction sequences. In two instances (MJ0832 and MJ0043), the similarity to other database sequences did not unambiguously define the NH_2-terminal extein-intein junction, so it was necessary to rely on consensus sequences to select the putative site. The inteins in MJ1042 and MJ0542 have previously uncharacterized COOH-terminal splice junctions, GNC and FNC, respectively.
31. P. T. Hamilton *et al.*, *Mol. Gen. Genet.* **200**, 47 (1985).
32. F. J. M. Mojica *et al.*, *Mol. Microbiol.* **17**, 85 (1995).

33. G. Felsenfeld *et al.*, *J. Am. Chem. Soc.* **79**, 2023 (1957); A. G. Letai *et al.*, *Biochemistry* **27**, 9108 (1988).
34. M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
35. Supported in part by Department of Energy Cooperative Agreements DE-FC02-95ER61962 (J.C.V.) and DEFC02-95ER61963 (C.R.W. and G.J.O), NASA grant NAGW 2554 (C.R.W.), and a core grant to TIGR from Human Genome Sciences. G.J.O. is the recipient of the National Science Foundation Presidential Young Investigator Award (DIR 89-57026). M.B. is supported by National Institutes of Health grant GM00783. We thank M. Heaney, C. Gnehm, R. Shirley, J. Slagel, and W. Hayes for software and database support; T. Dixon and V. Sapiro for computer system support; K. Hong and B. Stader for laboratory assistance; and B. Mukhopadhyay for helpful discussions. The *M. jannaschii* source accession number is DSM 2661, and the cells were a gift from P. Haney (Department of Microbiology, University of Illinois).

RESEARCH ARTICLES

Universal Quantum Simulators

Seth Lloyd

Feynman's 1982 conjecture, that quantum computers can be programmed to simulate any local quantum system, is shown to be correct.

Over the past half century, the logical devices by which computers store and process information have shrunk by a factor of 2 every 2 years. A quantum computer is the end point of this process of miniaturization—when devices become sufficiently small, their behavior is governed by quantum mechanics. Information in conventional digital computers is stored on capacitors. An uncharged capacitor registers a 0 and a charged capacitor registers a 1. Information in a quantum computer is stored on individual spins, photons, or atoms. An atom can itself be thought of as a tiny capacitor. An atom in its ground state is analogous to an uncharged capacitor and can be taken to register a 0, whereas an atom in an excited state is analogous to a charged capacitor and can be taken to register a 1.

So far, quantum computers sound very much like classical computers; the only use of quantum mechanics has been to make a correspondence between the discrete quantum states of spins, photons, or atoms and the discrete logical states of a digital computer. Quantum systems, however, exhibit behavior that has no classical analog. In particular, unlike classical systems, quantum systems can exist in superpositions of different discrete states. An ordinary capacitor can be either charged or uncharged, but not both: A classical bit is either 0 or 1. In contrast, an atom in a quantum superposition of its ground and excited state is a quantum bit that in some sense registers both 0 and 1 at the same time. As a result, quantum computers can do things that classical computers cannot.

Classical computers solve problems by using nonlinear devices such as transistors to perform elementary logical operations on the bits stored on capacitors. Quantum computers can also solve problems in a similar fashion; nonlinear interactions between quantum variables can be exploited to perform elementary quantum logical operations. However, in addition to ordinary classical logical operations such as AND, NOT, and COPY, quantum logic includes operations that put quantum bits in superpositions of 0 and 1. Because quantum computers can perform ordinary digital logic as well as exotic quantum logic, they are in principle at least as powerful as classical computers. Just what problems quantum computers can solve more efficiently than classical computers is an open question.

Since their introduction in 1980 (*1*) quantum computers have been investigated extensively (*2–29*). A comprehensive review can be found in (*15*). The best known problem that quantum computers can in principle solve more efficiently than classical computers is factoring (*14*). In this article I present another type of problem that in principle quantum computers could solve more efficiently than a classical computer—that of simulating other quantum systems. In 1982, Feynman conjectured that quantum computers might be able to simulate other quantum systems more efficiently than classical computers (*2*). Quantum simulation is thus the first classically difficult problem posed for quantum computers. Here I show that a quantum computer can in fact simulate quantum systems efficiently as long as they evolve according to local interactions.

Feynman noted that simulating quantum systems on classical computers is hard. Over the past 50 years, a considerable amount of effort has been devoted to such simulation. Much information about a quantum system's dynamics can be extracted from semiclassical approximations (when classical solutions are known), and ground state properties and correlation functions can be extracted with Monte Carlo methods (*30–32*). Such methods use amounts of computer time and memory space that grow as polynomial functions of the size of the quantum system of interest (where size is measured by the number of variables—particles or lattice sites, for example—required to characterize the system). Problems that can be solved by methods that use polynomial amounts of computational resources are commonly called tractable; problems that can only be solved by methods that use exponential amounts of resources are commonly called intractable. Feynman pointed out that the problem of simulating the full time evolution of arbitrary quantum systems on a classical computer is intractable: The states of a quantum system are wave functions that lie in a vector space whose dimension grows exponentially with the size of the system. As a result, it is an exponentially difficult problem merely to record the state of a quantum system, let alone integrate its equations of motion. For example, to record the state of 40 spin-½ particles in a classical computer's memory requires $2^{40} \approx 10^{12}$ numbers, whereas to calculate their time evolution requires the exponentiation of a $2^{40} \times 2^{40}$ matrix with $\approx 10^{24}$ entries. Feynman asked whether it might be possible to bypass this exponential explosion by having one quantum system simulate another directly, so that the states of the simulator obey the same equations of motion as the states of the simulated system. Feynman gave simple examples of one quantum system simulating another and conjectured that there existed a class of universal quantum simulators capable of simulating any quantum system that evolved according to local interactions.

The answer to Feynman's question is, yes. I will show that a variety of quantum systems, including quantum computers, can be "programmed" to simulate the behavior of arbitrary quantum systems whose dynamics are determined by local interactions. The programming is accomplished by inducing interactions between the variables of the simulator that imitate the interactions between the variables of the system to be simulated. In effect, the dynamics of the properly programmed simulator and the dynamics of the system to be simulated are one and the same to within any desired accuracy. So, to simulate the time evolution of 40 spin-½ particles over time t requires a simulator with 40 quantum bits evolving

The author is at the D'Arbeloff Laboratory for Information Systems and Technology, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: slloyd@mit.edu