The Minimal Gene Complement of *Mycoplasma genitalium*

Claire M. Fraser,* Jeannine D. Gocayne, Owen White, Mark D. Adams, Rebecca A. Clayton, Robert D. Fleischmann, Carol J. Bult, Anthony R. Kerlavage, Granger Sutton, Jenny M. Kelley, Janice L. Fritchman, Janice F. Weidman, Keith V. Small, Mina Sandusky, Joyce Fuhrmann, David Nguyen, Teresa R. Utterback, Deborah M. Saudek, Cheryl A. Phillips, Joseph M. Merrick, Jean-Francois Tomb, Brian A. Dougherty, Kenneth F. Bott, Ping-Chuan Hu, Thomas S. Lucier, Scott N. Peterson, Hamilton O. Smith, Clyde A. Hutchison III, J. Craig Venter

The complete nucleotide sequence (580,070 base pairs) of the *Mycoplasma genitalium* genome, the smallest known genome of any free-living organism, has been determined by whole-genome random sequencing and assembly. A total of only 470 predicted coding regions were identified that include genes required for DNA replication, transcription and translation, DNA repair, cellular transport, and energy metabolism. Comparison of this genome to that of *Haemophilus influenzae* suggests that differences in genome content are reflected as profound differences in physiology and metabolic capacity between these two organisms.

Mycoplasmas are members of the class Mollicutes, a large group of bacteria that lack a cell wall and have a characteristically low G + C content (*1*). These diverse organisms are parasites in a wide range of hosts including humans, animals, insects, plants, and cells grown in tissue culture (*1*). Aside from their role as potential pathogens, *Mycoplasmas* are of interest because of their reduced genome size and content relative to other prokaryotes.

Mycoplasma genitalium is thought to contain the smallest genome for a self-replicating organism (580 kb) and represents an important system for exploring a minimal functional gene set (*2*). *Mycoplasma genitalium* was originally isolated from urethral specimens of patients with non-gonoccocal urethritis (*3*) and has since been shown to exist in parasitic association with ciliated epithelial cells of primate genital and respiratory tracts (*4*).

The strategy and methodology for whole-genome random ("shotgun") sequencing and assembly was similar to that previously described for *Haemophilus influenzae* (*5*, *6*). To facilitate ordering of contigs, each template was sequenced from both ends. A total of 9846 sequencing reactions were performed by five individuals using an average of eight AB 373 DNA sequencers per day for a total of 8 weeks. Assembly of 8472 high-quality *M. genitalium* sequence fragments along with 299 random genomic sequences from Peterson *et al.* (*7*) was performed with the TIGR ASSEMBLER (*8*). The assembly process generated 39 contigs [size range, 606 to 73,351 base pairs (bp)] that contained a total of 3,806,280 bp of primary DNA sequence data. Contigs were ordered by ASM_ALIGN, a program that links contigs on the basis of information derived from forward and reverse sequencing reactions from the same clone.

ASM_ALIGN analysis revealed that all 39 gaps were spanned by an existing template from the small-insert genomic DNA library (that is, there were no physical gaps in the sequence assembly). The order of the contigs was confirmed by comparing the order of the random genomic sequences from Peterson *et al.* (*7*) that were incorporated into the assembly with their known position on the physical map of the *M. genitalium* chromosome (*9*). Because of the high stringency of the TIGR ASSEMBLER, the 39 contigs were searched against each other with GRASTA [a modified FASTA (*10*)] to detect overlaps (<30 bp) that would have been missed during the initial assembly process. Eleven overlaps were detected with this approach, which reduced the total number of gaps from 39 to 28.

Templates spanning each of the sequence gaps were identified, and oligonucleotide primers were designed from the sequences at the end of each contig. All gaps were less than 300 bp; thus, a primer walk from both ends of each template was sufficient for closure. All electropherograms were visually inspected with TIGR EDITOR (*5*) for initial sequence editing. Where a discrepancy could not be resolved or a clear assignment made, the automatic base calls were left unchanged. For each of the 53 ambiguities remaining after editing and the 25 potential frameshifts found after sequence-similarity searching, the appropriate template was resequenced with an alternative sequencing chemistry (dye terminator versus dye primer) to resolve ambiguities.

Ninety-nine percent of the *M. genitalium* genome was sequenced with better than single-sequence coverage, and the mean sequence redundancy was 6.5-fold. Although it is extremely difficult to assess sequence accuracy, we estimate our error rate to be less than 1 base in 10,000 on the basis of frequency of shifts in open reading frames (ORFs), overall quality of raw data, and fold coverage. The *M. genitalium* sequence (version 1.0) has been deposited in the Genome Sequence DataBase (GSDB) with the accession number L43967 (*11*).

Genome Analysis

The *M. genitalium* genome is a circular chromosome of 580,070 bp. The overall G + C content is 32% (A, 34%; C, 16%; G, 16%; and T, 34%). The G + C content across the genome varies between 27 and 37% (using a window of 5000 bp), with the regions of lowest G + C content flanking the presumed origin of replication for this organism (see below). As in *H. influenzae* (*5*), the ribosomal RNA (rRNA) operon (44%) and the transfer RNA (tRNA) genes (52%) in *M. genitalium* contain a higher G + C content than the rest of the genome, which may reflect the necessity of retaining essential G + C base pairing for secondary structure in rRNAs and tRNAs (*12*).

The genome of *M. genitalium* contains 74 Eco RI fragments, as predicted by both cosmid mapping data (*9*) and sequence analysis. The order and sizes of the Eco RI fragments determined by both methods are in agreement, with one apparent discrepancy between coordinates 62,708 and 94,573 in the sequence. However, reevaluation of

C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, and J. C. Venter are at the Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. J. M. Merrick is in the Department of Microbiology, State University of New York at Buffalo, Buffalo, NY 14214, USA. J.-F. Tomb, B. A. Dougherty and H. O. Smith are at the Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, and C. A. Hutchison III are at the University of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, NC 27599, USA.

*To whom correspondence should be addressed.

cosmid hybridization data in light of results from genome sequence analysis confirms that the sequence data are correct, and the extra 4.0-kb Eco RI fragment in this region of the cosmid map reflects a misinterpretation of the overlap between cosmids J-8 and 21 (*13*).

Studies of origins of replication in some prokaryotes have shown that DNA synthesis is initiated in an untranscribed AT-rich region between *dnaA* and *dnaN* (*14*). A search of the M. *genitalium* sequence for "DnaA boxes" around the putative origin of replication with consensus "DnaA boxes" from *Escherichia coli*, *Bacillus subtilis*, and *Pseudomonas aeruginosa* revealed no significant matches. Although we have not been able to localize the origin precisely, the colocalization of *dnaA* and *dnaN* to a 4000-bp region of the chromosome lends support to the hypothesis that it is the functional origin of replication in M. *genitalium* (*14*–*16*). The first base pair of the chromosomal sequence of M. *genitalium* is in an untranscribed region between *dnaA* and *dnaN* and was chosen so that *dnaN* is numbered as the first ORF in the genome. Genes to the right of this region are preferentially transcribed from the plus strand, and those to the left are preferentially transcribed from the minus strand. The apparent polarity in gene transcription is maintained across each half of the genome (Fig. 1), in marked contrast to *H. influenzae*, which displays no apparent polarity of transcription around the origin of replication.

The predicted coding regions of M. *genitalium* were initially defined by searching the entire genome for ORFs greater than 100 amino acids in length. Translations were made with the genetic code for mycoplasma species in which UGA encodes tryptophan. All ORFs were searched with BLAZE (*10*) against a nonredundant bacterial protein database (NRBP) (*5*) developed at TIGR on a MasPar MP-2 massively parallel computer with 4096 microprocessors. Protein matches were aligned with PRAZE, a modified Smith-Waterman (*17*) algorithm. Segments between predicted coding regions of the genome were also searched against all protein sequences from GenPept, Swiss-Prot, and the Protein Information Resource (PIR). The coding potential of 170 unidentified ORFs was analyzed with GeneMark (*18*), which had been trained with 308 M. *genitalium* sequences. Open reading frames that had low coding potential (on the basis of the GeneMark analysis) and were smaller than 100 nucleotides (a total of 53) were removed from the final set of putative coding regions. In a separate analysis, ORFs were searched against the complete set of translated sequences from *H. influenzae* [GSDB accession number L42023 (*5*)]. In total, these processes resulted in the identification of 470 predicted coding regions, of which 374 were putatively identified and 96 had no matches to protein sequences from any other organism. The 374 predicted coding regions with putative identifications were assigned biological roles with the classification system adapted from Riley (*19*).

Twenty-three of the protein matches in Table 1 have been annotated as motifs and represent matches where sequence similarity was confined to short domains in the predicted coding region. Several ORFs in M. *genitalium* displayed lower amino acid similarity to protein sequences in public

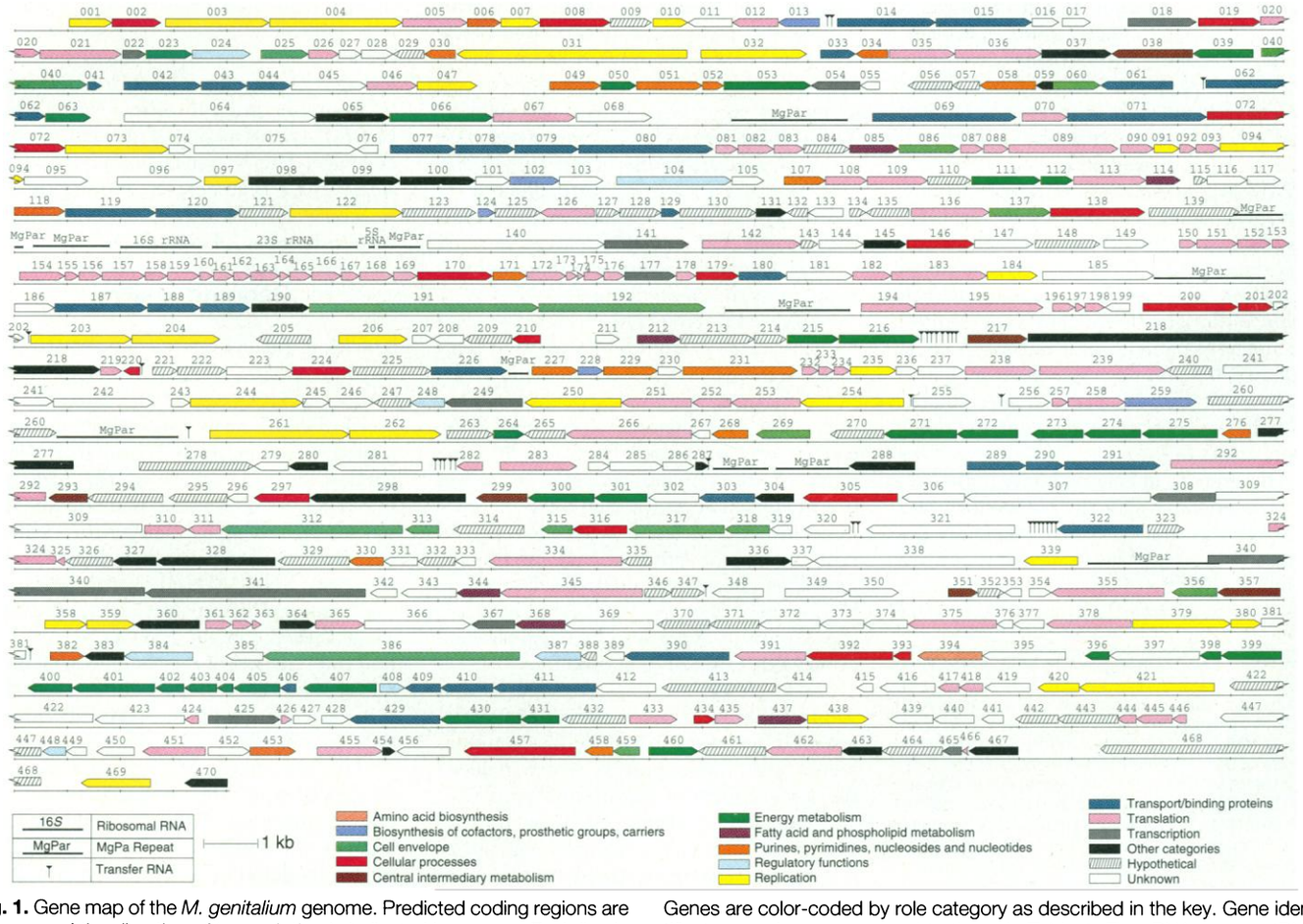

Fig. 1. Gene map of the *M. genitalium* genome. Predicted coding regions are shown, and the direction of transcription is indicated by arrows. Each line in the figure represents 24,000 bp of sequence in the *M. genitalium* genome. Genes are color-coded by role category as described in the key. Gene identificaton numbers correspond to those in Table 1. The rRNA operon, tRNA genes, and adhesin protein (MgPa) operon repeats are labeled.

archives than those observed with the motifs. In these cases, where motif identifications could not be made with confidence, the ORFs were annotated as no database match.

A separate search procedure was used in cases where we were unable to detect genes in the M. *genitalium* genome. Query peptide sequences that were available from eubacteria such as *E. coli*, *B. subtilis*, *Mycoplasma capricolum*, and *H. influenzae* were used in searches against all six reading frame translations of the entire genome sequence, and the alignments were examined by an experienced scientist. The possibility remains that current searching methods, an incomplete set of query sequences, or the subjective analysis of the database matches are not sensitive enough to identify certain M. *genitalium* gene sequences.

One-half of all predicted coding regions in M. *genitalium* for which a putative identification could be assigned display the greatest degree of similarity to a protein

Table 1. Summary of *M. genitalium* genes with putative identifications. Gene numbers correspond to those in Fig. 1. Each identified gene has been classified according to its role category [adapted from Riley (*19*)]. The putative gene identification and the percent amino acid identity are also listed for each entry. Those genes in *M. genitalium* that also match a gene in *H. influenzae* are indicated by an asterisk. An expanded version of this table with additional match information, including species, is available on the World Wide Web at URL http://www.tigr.org. Abbreviations: Bp, binding protein; DHase, dehydrogenase; G3PD, glyceraldehyde-3-phosphate dehydrogenase; MTase, methyltransferase; prt, protein; PRTase, phosphoribosyltransferase; Rdase, reductase; Tase, transferase; Sase, synthase; sub, subunit.

MG#	Identification	%ID
	Amino acid biosynthesis	
	Serine family	
*394	serine hydroxymethyltransferase (glyA)	55
	Biosynthesis of cofactors, prosthetic groups, and carriers	
	Folic acid	
*013	5,10-methylene-tetrahydrofolate DHase (folD)	33
*228	dihydrofolate RDase (dhfr)	33
	Heme and porphyrin	
*259	protoporphyrinogen oxidase (hemK)	31
	Thioredoxin, glutaredoxin, and glutathione	
*124	thioredoxin (trx)	36
*102	thioredoxin RDase (trxB)	39
	Cell envelope	
	Membranes, lipoproteins, and porins	
318	fibronectin-BP (fnbA)	25
040	membrane lipoprotein (tmpC)	31
*086	prolipoprotein diacylglyceryl Tase (lgt)	29
	Surface polysaccharides, lipopolysaccharides and antigens	
137	dTDP-4-dehydrorhamnose RDase (rfbD)	32
*356	lic-1 operon prt (licA) motif	28
*060	LPS biosyn prt (rfbV) motif	36
*269	surface prt antigen precursor (pag) motif	26
025	TrsB	28
	Surface structures	
192	114 kDa prt, MgPa operon (mgp)	100
191	attachment prt, MgPa operon (mgp)	100
315	cytadherence accessory prt (hmw1)	42
312	cytadherence-accessory prt (hmw1)	39
*386	cytadherence-accessory prt (hmw1)	34
313	cytadherence-accessory prt (hmw1)	53
*317	cytadherence-accessory prt (hmw3)	41
*459	surface exclusion prt (prgA) (Plasmid pCF10)	28
	Cellular processes	
	Cell division	
*457	cell division prt (ftsH)	50
*297	cell division prt (ftsY)	36
*224	cell division prt (ftsZ)	31
*434	mukB suppressor prt (smbA)	41
	Cell killing	
*146	hemolysin (tlyC)	26
220	pre-procytotoxin (vacA)	36
	Chaperones	
*019	heat shock prt (dnaJ)	34
002	heat shock prt (dnaJ) motif	40
*200	heat shock prt (dnaJ) motif	34
*392	heat shock prt (groEL)	52
*201	heat shock prt (grpE)	32
*393	heat shock prt 60-like prt (PggroES)	40
*305	heat shock prt 70 (hsp70)	57
	Detoxification	
*008	thiophene and furan oxidizer (tdhF)	32
	Protein and peptide secretion	
*138	GTP-binding membrane prt (lepA)	48
*179	haemolysin secretion ATP-BP (hlyB) motif	35
*072	preprotein translocase (secA)	44
*170	preprotein translocase secY sub (secY)	39
210	prolipoprotein signal peptidase (lsp)	32
*048	signal recognition particle prt (ffh)	43
	Transformation	
316	competence locus E (comE3) motif	30
	Central intermediary metabolism	
	Degradation of polysaccharides	
*217	bifunctional endo-1,4-beta-xylanase xyla precursor (xynA) motif	38
	Other	
*357	acetate kinase (ackA)	43
*038	glycerol kinase (glpK)	47
293	glycerophosphoryl diester phosphodiesterase (glpQ)	30
*299	phosphotransacetylase (pta)	45
	Phosphorus compounds	
*351	inorganic pyrophosphatase (ppa)	39
	Energy metabolism	
	Aerobic	
*039	glycerol-3-phospate DHase (GUT2)	43
*460	L-lactate DHase (ldh)	50
275	NADH oxidase (nox)	39
	ATP-proton motive force interconversion	
*405	adenosinetriphosphatase (atpB)	36
*401	ATP Sase alpha chain (atpA)	63
*403	ATP Sase B chain (atpF)	37
*399	ATP Sase beta chain (atpD)	81
*404	ATP Sase C chain (atpE)	50
*402	ATP Sase delta chain (atpH)	34
*398	ATP Sase epsilon chain (atpC)	37
*400	ATP Sase gamma chain (atpG)	38
	Glycolysis	
*063	1-phosphofructokinase (fruK)	26
*215	6-phosphofructokinase (pfk)	39
*407	enolase (eno)	54
*023	fructose-bisphosphate aldolase (tsr)	46
*301	G3PD (gap)	56
*111	phosphoglucose isomerase B (pgiB)	35
*300	phosphoglycerate kinase (pgk)	51
430	phosphoglycerate mutase (pgm)	45
*216	pyruvate kinase (pyk)	35
*431	triosephosphate isomerase (tim)	40
	Pentose phosphate pathway	
*264	6-phosphogluconate DHase (gnd)	30
*066	transketolase 1 (TK 1) (tktA)	33
	Pyruvate DHase	
*272	dihydrolipoamide acetyltransferase (pdhC)	45
*271	dihydrolipoamide DHase (pdhD)	38
*274	pyruvate DHase E1-alpha sub (pdhA)	43
*273	pyruvate DHase E1-beta sub (pdhB)	55
	Sugars	
*112	D-ribulose-5-phosphate 3 epimerase (cfxEc)	33
*050	deoxyribose-phosphate aldolase (deoC)	83
396	galactosidase acetyltransferase (lacA)	40
*053	phosphomannomutase (cpsG)	39
	Fatty acid and phospholipid metabolism	
*212	1-acyl-*sn*-glycerol-3-phosphate acetyltransferase (plsC)	32
*437	CDP-diglyceride Sase (cdsA)	38
368	fatty acid/phospholipid synthesis prt (plsX)	29
085	hydroxymethylglutaryl-CoA RDase (NADPH)	23
344	lipase-esterase (lip1)	27
*114	phosphatidylglycerophosphate Sase (pgsA)	29
	Purines, pyrimidines, nucleosides, and nucleotides	
	2′-Deoxyribonucleotide metabolism	
*231	ribonucleoside-diphosphate RDase (nrdE)	54
*229	ribonucleotide RDase 2 (nrdF)	50
*227	thymidylate Sase (thyA)	57
	Nucleotide and nucleoside interconversions	
*382	uridine kinase (udk)	34
	Purine ribonucleotide biosynthesis	
*107	5′-guanylate kinase (gmk)	43
*171	adenylate kinase (adk)	32
*058	phosphoribosylpyrophosphate Sase (prs)	44
	Salvage of nucleosides and nucleotides	
*276	adenine PRTase (apt)	34
*052	cytidine deaminase (cdd)	38
*330	cytidylate kinase (cmk)	40
268	deoxyguanosine-deoxyadenosine kinase(I) sub 2	30
*458	hypoxanthine-guanine PRTase (hpt)	38
*049	purine-nucleoside phosphorylase (deoD)	44
*034	thymidine kinase (tdk)	48
051	thymidine phosphorylase (deoA)	53
*006	thymidylate kinase (CDC8)	28
*030	uracil PRTase (upp)	45
	Sugar-nucleotide biosynthesis and conversions	
*118	UDP-glucose 4-epimerase (galE)	34
*453	UDP-glucose pyrophosphorylase (gtaB)	48
	Regulatory functions	
*024	GTP-BP (gtp1)	47
*384	GTP-BP (obg)	40
*387	GTP-BP (era)	27
248	major sigma factor (rpoD)	28
*448	pilin repressor (pilB)	53
*408	pilin repressor (pilB) motif	49
*104	virulence-associated prt homolog (vacB)	29
	Replication	
	Degradation of DNA	
032	ATP-dependent nuclease (addA)	27
	DNA replication, restriction, modification, recombination, and repair	
*469	chromosomal replication initiator prt (dnaA)	31
*004	DNA gyrase sub A (gyrA)	100
*003	DNA gyrase sub B (gyrB)	99
*244	DNA helicase II (mutB1)	36
*254	DNA ligase (lig)	38
*262	DNA polymerase I (polI) motif	30
*031	DNA polymerase III (polC)	38
*261	DNA polymerase III alpha sub (dnaE)	32
*001	DNA polymerase III beta sub (dnaN)	100
*420	DNA polymerase III sub (dnaH)	49
007	DNA polymerase III sub (dnaH) motif	23
*250	DNA primase (dnaE)	27
*010	DNA primase (dnaE) motif	26
*122	DNA topoisomerase I (topA)	39
235	endonuclease IV (nfo)	29
*421	excinuclease ABC sub A (uvrA)	48
*073	excinuclease ABC sub B (uvrB)	48
*206	excinuclease ABC sub C (uvrC)	28
*379	glucose-inhibited division prt (gidA)	40
*380	glucose-inhibited division prt (gidB)	25
*358	Holliday junction DNA helicase (ruvA)	26
*359	Holliday junction DNA helicase (ruvB)	35
184	MTase (ssoIM)	43
*339	recombination prt (recA)	47
*094	replicative DNA helicase (dnaB)	33
438	restriction-modification enzyme EcoD specificity sub (hsdS)	25
*047	S-adenosylmethionine Sase 2 (metX)	44
*091	single-stranded DNA BP (ssb)	22
*204	DNA topoisomerase IV sub A (parC)	100
*203	DNA topoisomerase IV sub B (parE)	100
*097	uracil DNA glycosylase (ung)	33
	Transcription	
	Degradation of RNA	
*367	ribonuclease III (rnc)	30
*465	RNase P C5 sub (rnpA)	40
	RNA synthesis, modification, and DNA transcription	
*308	ATP-dependent RNA helicase (deaD)	23
*425	ATP-dependent RNA helicase (deaD)	32
*018	helicase (mot1) motif	44
*141	N-utilization substance prt A (nusA)	36
*177	RNA polymerase alpha core sub (rpoA)	31
*341	RNA polymerase beta sub (rpoB)	39
*340	RNA polymerase beta' chain (rpoC)	47
022	RNA polymerase delta sub (rpoE)	29
*249	RNA polymerase sigma A factor (sigA)	44
*054	transcription antitermination factor (nusG)	31
	Translation	
	Amino acyl tRNA synthetases and tRNA modification	
*292	Ala-tRNA Sase (alaS)	34
*378	Arg-tRNA Sase (argS)	34
*113	Asn-tRNA Sase (asnS)	41
*036	Asp-tRNA Sase (aspS)	41
*253	Cys-tRNA Sase (cysS)	34
*462	Glu-tRNA Sase (gltX)	43
251	Gly-tRNA Sase	36
*035	His-tRNA Sase (hisS)	31
*345	Ile-tRNA Sase (ileS)	33
*266	Leu-tRNA Sase (leuS)	43
*136	Lys-tRNA Sase (lysS)	46
*365	Met-tRNA formyltransferase (fmt)	24
*021	Met-tRNA Sase (metS)	38
*083	peptidyl-tRNA hydrolase homolog (pth)	38
*195	Phe-tRNA Sase alpha chain (pheT)	26
*194	Phe-tRNA Sase beta chain (pheS)	35
*283	Pro-tRNA Sase (proS)	23
*182	pseudouridylate Sase I (hisT)	27
*005	Ser-tRNA Sase (serS)	43
*375	Thr-tRNA Sase (thrSv)	39
*445	tRNA (guanine-N1)-MTase (trmD)	41
*126	Trp-tRNA Sase (trpS)	41
*455	Tyr tRNA Sase (tyrS)	39
*334	Val-tRNA Sase (valS)	39
	Degradation of proteins, peptides, and glycopeptides	
*391	aminopeptidase	45
*324	aminopeptidase P (pepP)	31
*239	ATP-dependent protease (lon)	44
*355	ATP-dependent protease binding sub (clpB)	48
067	glutamic acid specific protease (SPase)	29
219	IgA1 protease	32
183	oligoendopeptidase F (pepF)	30
020	proline iminopeptidase (pip)	38
310	proline iminopeptidase (pip)	29
*046	sialoglycoprotease (gcp)	36
*238	trigger factor (tig)	25
	Protein modification and translation factors	
*089	elongation factor G (fus)	59
026	elongation factor P (efp)	26
*433	elongation factor Ts (tsf)	39
*451	elongation factor TU (tuf)	100
*106	formylmethionine deformylase (def) motif	37
*173	initiation factor 1 (infA)	49
*172	methionine amino peptidase (map)	36
*258	peptide chain release factor 1 (RF-1)	43
108	prt phosphatase 2C homolog (ptc1) motif	28
109	prt serine-threonine kinase motif	34
*142	prt synthesis initiation factor 2 (infB)	46
*435	ribosome releasing factor (frr)	35
*282	transcription elongation factor (greA)	40
*196	translation initiation factor IF3 (infC)	31
	Ribosomal proteins: synthesis and modification	
*082	ribosomal prt L1	48
*361	ribosomal prt L10	30
*081	ribosomal prt L11	52
*418	ribosomal prt L13	40
*161	ribosomal prt L14	63
*169	ribosomal prt L15	42
*158	ribosomal prt L16	64
*178	ribosomal prt L17	35
*167	ribosomal prt L18	43
*444	ribosomal prt L19	49
*154	ribosomal prt L2	58
*198	ribosomal prt L20	58
*232	ribosomal prt L21	38
233	ribosomal prt L21 homolog	100
*156	ribosomal prt L22	49
153	ribosomal prt L23	39
*162	ribosomal prt L24	45
*234	ribosomal prt L27	64
*426	ribosomal prt L28	36
*159	ribosomal prt L29	42
*151	ribosomal prt L3	37
*257	ribosomal prt L31	48
*363	ribosomal prt L32	58
*325	ribosomal prt L33	67
*466	ribosomal prt L34	60
197	ribosomal prt L35	78
174	ribosomal prt L36	39
*152	ribosomal prt L4	58
*163	ribosomal prt L5	46
*166	ribosomal prt L6	48
*362	ribosomal prt L7/L12 ('A' type)	33
*093	ribosomal prt L9	
*150	ribosomal prt S10	49
*176	ribosomal prt S11	48
*087	ribosomal prt S12	75
*175	ribosomal prt S13	63
*164	ribosomal prt S14	70
*424	ribosomal prt S15	48
*446	ribosomal prt S16	49
*160	ribosomal prt S17	51
*092	ribosomal prt S18	45
*155	ribosomal prt S19	59
*070	ribosomal prt S2	35
*157	ribosomal prt S3	47
*311	ribosomal prt S4	43
*168	ribosomal prt S5	56
*090	ribosomal prt S6	24
*012	ribosomal prt S6 modification prt (rimK) motif	31
*088	ribosomal prt S7	65
*165	ribosomal prt S8	47
*417	ribosomal prt S9	52
*252	rRNA methylase	39
	Transport and binding proteins	
	Amino acids, peptides, and amines	
226	aromatic amino acid transport prt (aroP)	25
*180	membrane transport prt (glnQ)	37
*303	membrane transport prt (glnQ)	32
*079	oligopeptide transport ATP-BP (amiE)	48
*080	oligopeptide transport ATP-BP (amiF)	47
*078	oligopeptide transport permease prt (dciAC)	33
*077	oligopeptide transport permease prt (oppB)	28
*042	spermidine-putrescine transport ATP-BP (potA)	42
*043	spermidine-putrescine transport permease prt (potB)	27
*044	spermidine-putrescine transport permease prt (potC)	29
	Anions	
*410	peripheral membrane prt B (pstB)	51
409	peripheral membrane prt U (phoU)	27
*411	periplasmic phosphate permease homolog (AG88)	31
	Carbohydrates, organic alcohols, and acids	
*187	ATP-BP (msmK)	41
*062	fructose-permease IIBC component (fruA)	43
*033	glycerol uptake facilitator (glpF)	36
061	hexosephosphate transport prt (uhpT)	31
*188	membrane prt (msmF)	22
189	membrane prt (msmG)	27
*119	methylgalactoside permease ATP-BP (mglA)	33
*429	PEP-dependent HPr prt kinase phosphorylTransferase (ptsI)	46
*041	phosphohistidinoprotein-hexose phosphotransferase (ptsH)	49
*069	phosphotransferase enzyme II, ABC component (ptsG)	43
129	PTS glucose-specific permease	25
*120	ribose transport permease prt (rbsC)	27
	Cations	
071	cation-transporting ATPase (pacL)	34
	Other	
*290	ATP-BP P29	32
289	high affinity transport prt P37 (P37)	36
*390	lactococcin transport ATP-BP (lcnDR3)	22
322	Na^+ ATPase sub J (ntpJ)	31
*014	transport ATP-BP (msbA)	28
*015	transport ATP-BP (msbA)	32
291	transport permease prt P69 (P69)	28
*406	transport permease prt P69 (P69) motif	40
	Other categories	
	Adaptations and atypical conditions	
454	osmotically inducible prt (osmC)	28
	phosphate limitation prt (sphX)	31
470	SpoOJ regulator motif	27
*277	spore germination apparatus prt (gerBB) motif	31
383	sporulation prt (outB) motif	36
	Drug and analog sensitivity	
*463	high-level kasgamycin resistance (ksgA)	36
	Other	
*298	115 kD prt (p115)	33
190	29 kDa prt, MgPa operon (mgp)	62
*065	heterocyst maturation prt (devA)	35
*467	heterocyst maturation prt (devA)	40
099	hydrolase (aux2)	32
131	hypothetical prt (GB:M31161_3)	22
*218	macrogolgin	25
*327	magnesium chelatase 30 kD sub (bchO)	27
*304	membrane-associated ATPase (cbiO)	30
364	mobilization prt (mob13) motif	31
*336	nitrogen fixation prt (nifS)	26
287	nodulation prt F (nodF)	35
098	p48 eggshell prt (p48)	23
100	PET112 prt	31
037	pre-B cell enhancing factor (PBEF)	34
288	prt L	31
*328	prt V (fcrV)	28
*145	prt X	29
*280	sensory rhodopsin II transducer (htrII) motif	16
*059	small prt (smpB)	33
360	UV protection prt (mucB)	22

from either a Gram-positive organism (for example, *B. subtilis*) or a *Mycoplasma* species. The significance of this finding is underscored by the fact that NRBP contains 3885 sequences from *E. coli* and only 1975 sequences from *B. subtilis*. In the majority of cases where M. *genitalium* coding regions matched sequences from both *E. coli* and *Bacillus* species, the better match was to a sequence from *Bacillus* (average, 62% similarity) rather than to a sequence from *E. coli* (average, 56% similarity). The evolutionary relationship between *Mycoplasma* and the *Lactobacillus-Clostridium* branch of the Gram-positive phylum has been deduced from small-subunit rRNA sequences (*20*, *21*). Our data from whole-genome analysis support this hypothesis.

Comparative Genomics: *M. genitalium* and *H. influenzae*

A survey of the genes and their organization in M. *genitalium* permits the description of a minimal set of genes required for survival. One would predict that a minimal cell must contain genes for replication and transcription, at least one rRNA operon and a set of ribosomal proteins, tRNAs and tRNA synthetases, transport proteins to derive nutrients from the environment, biochemical pathways to generate adenosine triphosphate (ATP) and reducing power, and mechanisms for maintaining cellular homeostasis. Comparison of the genes identified in M. *genitalium* with those in *H. influenzae* allows for identification of a basic complement of genes conserved in these two species and provides insights into physiological differences between one of the simplest self-replicating prokaryotes and a more complex, Gram-negative bacterium.

The M. *genitalium* genome contains 470 predicted coding sequences as compared with 1727 identified in *H. influenzae* (*5*) (Table 2). The percent of the total genome in M. *genitalium* and *H. influenzae* that encodes genes involved in cellular processes, central intermediary metabolism, energy metabolism, fatty acid and phospholipid metabolism, purine and pyrimdine metabolism, replication, transcription, transport, and other categories is similar, although the total number of genes in these categories is considerably fewer in M. *genitalium*. A smaller percentage of the M. *genitalium* genome encodes genes involved in amino acid biosynthesis, biosynthesis of cofactors, cell envelope, and regulatory functions as compared with *H. influenzae*. A greater percentage of the M. *genitalium* genome encodes proteins involved in translation than in *H. influenzae*, as shown by the similar numbers of ribosomal proteins and tRNA synthetases in both organisms.

The 470 predicted coding regions in M. *genitalium* (average size, 1040 bp) comprise 88% of the genome (on average, one gene every 1235 bp), a value similar to that found in *H. influenzae* where 1727 predicted coding regions (average size, 900 bp) comprise 85% of the genome (one gene every 1042 bp). These data indicate that the reduction in genome size that has occurred in *Mycoplasma* has not resulted in an increase in gene density or a decrease in gene size (*22*). A global search of M. *genitalium* and *H. influenzae* genomes reveals short regions of conservation of gene order, particularly two clusters of ribosomal proteins.

Replication. We have identified genes that encode many essential proteins in the replication process, including M. *genitalium* isologs of the proteins DnaA, DnaB, GyrA,

Table 2. Summary of gene content in *H. influenzae* and *M. genitalium* sorted by functional category. The number of genes in each functional category is listed for *H. influenzae* and *M. genitalium*. The number in parentheses indicates the percent of the putatively identified genes devoted to each functional category. For the category of unassigned genes, the percent of the genome indicated in parentheses represents the percent of the total number of putative coding regions.

Biological role	*H. influenzae*	*M. genitalium*
Amino acid biosynthesis	68 (6.8)	1 (0.3)
Biosynthesis of cofactors	54 (5.4)	5 (1.6)
Cell envelope	84 (8.3)	17 (5.3)
Celluar processes	53 (5.3)	21 (6.6)
Cell division	16	4
Cell killing	5	2
Chaperones	6	7
Detoxification	3	1
Protein secretion	15	6
Transformation	8	1
Central intermediary metabolism	30 (3)	6 (1.9)
Energy metabolism	112 (10.4)	31 (9.7)
Aerobic	4	3
Amino acids and amines	4	0
Anaerobic	24	0
ATP–proton force interconversion	9	8
Electron transport	9	0
Entner-Doudoroff	9	0
Fermentation	8	0
Gluconeogenesis	2	0
Glycolysis	10	10
Pentose phosphate pathway	3	2
Pyruvate dehydrogenase	4	4
Sugars	15	4
TCA cycle	11	0
Fatty acid and phospholipid metabolism	25 (2.5)	6 (1.9)
Purines, pyrimidines, nucleosides, and nucleotides	53 (5.3)	19 (6.0)
2′-Deoxyribonucleotide metabolism	8	3
Nucleotide and nucleoside interconversions	3	1
Purine ribonucleotide biosynthesis	18	3
Pyrimidine ribonucleotide biosynthesis	5	0
Salvage of nucleosides and nucleotides	13	10
Sugar-nucleotide biosynthesis and conversions	6	2
Regulatory functions	64 (6.3)	7 (2.2)
Replication	87 (8.6)	32 (10.0)
Degradation of DNA	8	1
DNA replication, restriction, modification, recombination, and repair	76	31
Transcription	27 (2.7)	12 (3.8)
Degradation of RNA	10	2
RNA synthesis and modification, DNA transcription	17	10
Translation	141 (14)	101 (31.8)
Transport and binding proteins	123 (12.2)	34 (10.7)
Amino acids and peptides	38	10
Anions	8	3
Carbohydrates	30	12
Cations	24	1
Other transporters	22	8
Other categories	93 (9.2)	27 (8.2)
Unassigned role	736 (43)	152 (32)
No database match	389	96
Match hypothetical proteins	347	56

GyrB, a single-stranded DNA-binding protein, and the primase protein DnaE. DnaJ and DnaK, heat shock proteins that may function in the release of the primosome complex, are also found in *M. genitalium*. A gene encoding the DnaC protein, responsible for delivery of DnaB to the primosome, has yet to be identified.

Genes encoding most of the essential subunit proteins for DNA polymerase III in *M. genitalium* were also identified. The *polC* gene encodes the α subunit, which contains the polymerase activity. We have also identified the isolog of *dnaH* in *B. subtilis* (dnaX in *E. coli*) that encodes the γ and τ subunits as alternative products from the same gene. These proteins are necessary for the processivity of DNA polymerase III. An isolog of *dnaN* that encodes the β subunit was previously identified in *M. genitalium* (*15*) and is involved in the process of clamping the polymerase to the DNA template. While we have yet to identify a gene encoding the ε subunit responsible for the 3′-5′ proofreading activity, it is possible that this activity is encoded in the α subunit as previously described (*23*). Finally, we have identified a gene encoding a DNA ligase, necessary for the joining of the Okazaki fragments formed during synthesis of the lagging strand.

While we have identified genes encoding many isologs thought to be essential for DNA replication, some genes encoding proteins with key functions have yet to be identified. Examples of these are Dnaθ and Dnaδ, whose functions are less well understood but are thought to be involved in the assembly and processivity of polymerase III. Also apparently absent is a specific ribonuclease H protein responsible for the hydrolysis of the RNA primer synthesized during lagging-strand synthesis.

DNA repair. It has been suggested that in *E. coli* as many as 100 genes are involved in DNA repair (*24*), and in *H. influenzae* the number of putatively identified DNA-repair enzymes is approximately 30 (*5*). Although *M. genitalium* appears to have the necessary genes to repair many of the more common lesions in DNA, the number of genes devoted to the task is much smaller. Excision repair of regions containing missing bases [apurinic or apyriminic (AP) sites] can likely occur by a pathway involving endonuclease IV (nfo), Pol I, and ligase. The *ung* gene, which encodes uracil-DNA glycosylase, is present. This activity removes uracil residues from DNA that usually arise by spontaneous deamination of cytosine.

All three genes necessary for production of the ultraviolet-resistant ABC exinuclease are present, and along with Pol I, helicase II, and ligase should provide a mechanism for repair of damage such as cross-linking, which requires replacement of both strands. Although *recA* is present, which in *E. coli* is activated as it binds to single-stranded DNA, thereby initiating the SOS response, we find no evidence for a *lexA* gene, which encodes the repressor that regulates the SOS genes. We have not identified photolyase (*phr*) in *M. genitalium*, which repairs ultraviolet-induced pyrimidine dimers, or other genes involved in reversal of DNA damage rather than excision and replacement of the lesion.

Transcription. The critical components for transcription were identified in *M. genitalium*. In addition to the α, β, and β′ subunits of the core RNA polymerase, *M. genitalium* appears to encode a single σ factor, whereas *E. coli* and *B. subtilis* encode at least six and seven, respectively. We have not detected a homolog of the Rho termination factor gene, so it seems likely that a mechanism similar to Rho-independent termination in *E. coli* operates in *M. genitalium*. We have clear evidence for homologs of only two other genes that modulate transcription, *nusA* and *nusG*.

Translation. *Mycoplasma genitalium* has a single rRNA operon that contains three rRNA subunits in the order 16*S* rRNA (1518 bp)–spacer (203 bp)–23*S* rRNA (2905 bp)–spacer (56 bp)–5*S* rRNA (103 bp). The small-subunit rRNA sequence was compared with the Ribosomal Database Project's (*21*) prokaryote database with the program "similarity_rank." Our sequence is identical to the *M. genitalium* (strain G37) sequence deposited there, and the 10 most similar taxa returned by this search are also in the genus *Mycoplasma*.

A total of 33 tRNA genes were identified in *M. genitalium*; these were organized into five clusters plus nine single genes. In all cases, the best match for each tRNA gene in *M. genitalium* was the corresponding gene in *M. pneumoniae* (*25*). Furthermore, the grouping of tRNAs into clusters (trnA, trnB, trnC, trnD, and trnE) was identical in *M. genitalium* and *M. pneumoniae*, as was gene order within the cluster (*25*). The only difference between *M. genitalium* and *M. pneumoniae* with regard to tRNA gene organization was an inversion between trnD and GTG. In contrast to *H. influenzae* and many other eubacteria, no tRNAs were found in the spacer region between the 16S and 23S rRNA genes in the rRNA operon of *M. genitalium*, similar to what has been reported for *M. capricolum* (*26*).

A search of the *M. genitalium* genome for tRNA synthetase genes identified all of the expected genes except glutaminyl tRNA synthetase (*glnS*). In *B. subtilis* and other Gram-positive bacteria, and *Saccharomyces cerevisiae* mitochondria, no glutaminyl tRNA synthetase activity has been detected (*27*). In these organisms, a single glutamyl tRNA synthetase aminoacylates both $tRNA^{Glu}$ and $tRNA^{Gln}$ with glutamate (*28*). The formation of glutaminyl tRNA synthetase is accomplished by amidation of glutamate to glutamine in a reaction that is functionally analogous to that catalyzed by glutamine synthetase (*29*). Because of its evolutionary relationship with Gram-positive organisms (*20*, *21*), it is likely that a similar mechanism is involved in the formation of glutaminyl tRNA synthetase in *M. genitalium*.

Metabolic pathways. The reduction in genome size among *Mycoplasma* species is associated with a marked reduction in the number and components of biosynthetic pathways in these organisms, thereby requiring them to use metabolic products from their hosts. The complex growth requirements of this organism in the laboratory can be explained by the almost complete lack of enzymes involved in amino acid biosynthesis, de novo nucleotide biosynthesis, and fatty acid biosynthesis (Table 1 and Fig. 1). When the number of genes in the categories of central intermediary metabolism, energy metabolism, and fatty acid and phospholipid metabolism are examined, marked differences in gene content between *H. influenzae* and *M. genitalium* are apparent. For example, whereas the *H. influenzae* genome contains 68 genes involved in amino acid biosynthesis, the *M. genitalium* genome contains only 1. In total, the *H. influenzae* genome has 228 genes associated with metabolic pathways, whereas the *M. genitalium* genome has just 44. A recent analysis of 214 kb of sequence from *Mycoplasma capricolum* (*22*), a related organism whose genome size is twice as large as that of *M. genitalium*, reveals that *M. capricolum* contains a number of biosynthetic enzymes not present in *M. genitalium*. This observation suggests that *M. capricolum*'s larger genome confers a greater anabolic capacity.

Mycoplasma genitalium is a facultative anaerobe that ferments glucose and possibly other sugars by way of glycolysis to lactate and acetate. Genes that encode all the enzymes of the glycolytic pathway were identified, including genes for components of the pyruvate dehydrogenase complex, phosphotransacetylase, and acetate kinase. The major route for ATP synthesis may be through substrate-level phosphorylation, because no cytochromes are present. *Mycoplasma genitalium* also lacks all the components of the tricarboxylic acid cycle. None of the genes encoding glycogen or poly-β-hydroxybutryate production were identified, indicating limited capacity for carbon and energy storage. The pentose phosphate pathway also appears limited because only genes encoding 6-phosphogluconate dehydrogenase and transketolase were identified. The limited metabolic capacity of *M. genitalium* contrasts sharply with the complexity of cata-

bolic pathways in *H. influenzae*, reflecting the fourfold greater number of genes involved in energy metabolism found in *H. influenzae*.

Transport. The transporters identified in *M. genitalium* are specific for a range of nutritional substrates. In protein transport, for example, both oligopeptide and amino acid transporters are represented. One interesting peptide transporter is similar to a lactococcin transporter (lcnDR3) and related bacteriocin transporters, suggesting that *M. genitalium* may export a small peptide with antibacterial activity. The *M. genitalium* isolog of the *M. hyorhinis* p37 high-affinity transport system also has a conserved lipid-modification site, providing further evidence that the *Mycoplasma* binding protein–dependent transport systems are organized in a manner analogous to Gram-positive bacteria (*30*).

Genes encoding proteins that function in the transport of glucose by way of the phosphoenolpyruvate:sugar transferase system (PTS) have been identified in *M. genitalium*. These proteins include enzyme I (EI), HPr, and sugar-specific enzyme IIs (EIIs) (*31*). EIIs is a complex of at least three domains: EIIA, EIIB, and EIIC. In some bacteria (for example, *E. coli*) EIIA is a soluble protein, whereas in others (*B. subtilis*) a single membrane protein contains all three domains. These variations in the proteins that make up the EII complex are due to fusion or splitting of domains during evolution and are not considered to be mechanistic differences (*31*). In *M. genitalium*, EIIA, -B, and -C are located in a single protein similar to that found in *B. subtilis*. In *M. capricolum ptsH*, the gene that encodes HPr is located on a monocistronic transcriptional unit, whereas genes encoding EI (*ptsI*) and EIIA (crr) are located on a dicistronic operon (*32*). In most bacterial species studied to date, *ptsI*, *ptsH*, and *crr* are part of a polycistonic operon (*pts* operon). In *M. genitalium*, *ptsH*, *ptsI*, and the gene encoding EIIABC reside at different locations of the genome, and thus each of these genes may constitute monocistronic transcriptional units. We have also identified the EIIBC component for uptake of fructose; however, other components of the fructose PTS were not found. Thus, *M. genitalium* may be limited to the use of glucose as an energy source. In contrast, *H. influenzae* has the ability to use at least six different sugars as a source of carbon and energy.

Regulatory systems. It appears that regulatory systems found in other bacteria are absent in *M. genitalium*. For instance, although two-component systems have been described for a number of Gram-positive organisms, no sensor or response regulator genes are found in the *M. genitalium* genome. Furthermore, the lack of a heat shock σ factor raises the question of how the heat shock response is regulated. Another stress faced by all metabolically active organisms is the generation of reactive oxygen intermediates such as superoxide anions and hydrogen peroxide. Although *H. influenzae* has an *oxyR* homolog, as well as catalase and superoxide dismutase, *M. genitalium* appears to lack these genes as well as an NADH [nicotinamide adenine dinucleotide (reduced)] peroxidase. The importance of these reactive intermediate molecules in host cell damage suggests that some as yet unidentified protective mechanism may exist within the cell.

Antigenic variation. The 140-kD adhesin protein of *M. genitalium* is densely clustered at a differentiated tip and elicits a strong immune response in humans and experimentally infected animals (*33*). The adhesin protein (MgPa) operon in *M. genitalium* contains a 29-kD ORF, the MgPa protein (160 kD), and a 114-kD ORF with intervening regions of six nucleotides and one nucleotide, respectively (*34*). On the basis of hybridization experiments (*35*), multiple copies of regions of the *M. genitalium* MgPa gene and the 114-kD ORF are known to exist throughout the genome.

The availability of the complete genomic sequence from *M. genitalium* has allowed a comprehensive analysis of the MgPa repeats. In addition to the complete operon, nine repetitive elements that are composites of regions of the MgPa operon were found. (Fig. 1) The percent of sequence identity between the repeat elements and the MgPa operon genes ranges from 78 to 90%. The sequences contained in the MgPa operon and the nine repeats scattered throughout the chromosome represent 4.7% of the total genomic sequence. Although this observation might appear to contradict the expectation for a minimal genome, recent evidence for recombination between the repetitive elements and the MgPa operon has been reported (*36*). Such recombination may allow *M. genitalium* to evade the host immune response through mechanisms that induce antigenic variation within the population.

The *M. genitalium* genome contains 90 putatively identified genes that do not appear to be present in *H. influenzae*. Almost 60% of these genes have database matches to known or hypothetical proteins from Gram-positive bacteria or other *Mycoplasma* species, suggesting that these genes may encode proteins with a restricted phylogenetic distribution. Ninety-six potential coding regions in *M. genitalium* have no database match to any sequences in public archives including the entire *H. influenzae* genome; therefore, these likely represent novel genes in *M. genitalium* and related organisms.

The predicted coding sequences of the hypothetical ORFs, the ORFs with motif matches, and the ORFs that have no similarities to known peptide sequences were analyzed with the Kyte-Doolittle algorithm (*37*), with a range of 11 residues, and PSORT, which is available on the World Wide Web at URL http://psort.nibb.ac.jp. PSORT predicts the presence of signal sequences by the methods of McGeoch and von Heijne (*38*), and detects potential transmembrane domains by the method of Klein *et al.* (*39*). Of a total of 175 ORFs examined, 90 potential membrane proteins were found, 11 of which were predicted to have type I signal peptides and 5 to have type II signal peptides. At least 50 potential membrane proteins with role assignments were also identified by this approach, in agreement with previously predicted or confirmed membrane localizations for these proteins. Taken together, these data suggest that the total number of potential membrane proteins in *M. genitalium* may be on the order of 140.

To manage these putative membrane proteins, *M. genitalium* has at its disposal a minimal secretory machinery composed of six functions: two chaperonins (GroEL and DnaK) (*40*, *41*), an adenosine triphosphatase (ATPase) pilot protein (SecA), one integral membrane protein translocase (SecY), a signal recognition particle protein (Ffh), and a lipoprotein-specific signal peptidase (LspA) (*40*). Perhaps the lack of other known translocases (like SecE, SecD, and SecF) that are present in *E. coli* and *H. influenzae* is related to the presence in *M. genitalium* of a one-layer cell envelope. Also, the absence in *M. genitalium* of a SecB homolog, the secretory chaperonin of *E. coli* [it is also absent in *B. subtilis* (*42*)], might reflect a difference between Gram-negative and wall-less Mollicutes in processing nascent proteins destined for the general secretory pathway. Considering the presence of several putative membrane proteins that contain type I signal peptides, the absence of a signal peptidase I (*lepB*) is most surprising. A direct electronic search for the *M. genitalium lepB* gene with the *E. coli lepB* and the *B. subtilis sipS* (*43*) as queries did not reveal any signficant similarities.

A number of possibilities may explain why genes encoding some of the proteins characterized in other eubacterial species appear to be absent in *M. genitalium*. One possibility is that a limited number of proteins in this organism may have become adapted to perform other functions. A second possibility is that certain proteins found in more complex bacteria such as *E. coli* are not required in a simpler prokaryote like *M. genitalium*. Finally, sequences from *M. genitalium* may have such low similarity to known sequences from other species that matches are not detectable above a reason-

able confidence threshold.

The complete sequencing and assembly of other microbial genomes, together with genome surveys using random sequencing, will continue to provide a wealth of information on the evolution of single genes, gene families, and whole genomes. Comparison of these data with the genome sequence of *M. genitalium* should allow a more precise definition of the fundamental gene complement for a self-replicating organism and a more comprehensive understanding of the diversity of life.

REFERENCES AND NOTES

1. S. Razin, *Microbiol. Rev.* **49**, 419 (1985); J. Maniloff, Mycoplasmas: *Molecular Biology and Pathogenesis*, J. Maniloff *et al.*, Eds. (American Society for Microbiology, Washington, DC, 1992), pp. 549–559.
2. S. D. Colman *et al.*, *Mol. Microbiol.* **4**, 683 (1990); C. Su and J. B. Baseman, *J. Bacteriol.* **172**, 4705 (1990).
3. J. G. Tully *et al.*, *Int. J. Syst. Bacteriol.* **33**, 387 (1983).
4. J. B. Baseman *et al.*, *J. Clin. Microbiol.* **26**, 2266 (1988); J. S. Jensen *et al.*, *Genitourin. Med.* **69**, 265 (1993); D. Taylor-Robinson *et al.*, *Lancet* **13**, 1066 (1994).
5. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
6. A total of 50 μg of purified *M. genitalium* strain G-37 DNA (American Type Culture Collection number 33530) were isolated from cells grown in modified Hayflick's medium [L. Hayflick, *Texas Rep. Biol. Med.* **23**, 285 (1965)] containing agamma horse serum, 10% yeast dialysate, and penicillin G (1000 U/ml). A mixture (990 μl) containing 50 μg of DNA, 300 mM sodium acetate, 10 mM tris-HCl, 1 mM EDTA, and 30% glycerol was chilled to 0°C in an Aeromist Downdraft Nebulizer chamber (IBI Medical Products, Chicago, IL) and sheared at 12 psi for 60 s. The DNA was precipitated in ethanol and redissolved in 50 μl of tris-EDTA (TE) buffer to create blunt ends; a 40-μl portion was digested for 10 min at 30°C in 85 μl of BAL31 buffer with 2 U of BAL31 nuclease (New England BioLabs). The DNA was extracted with phenol, precipitated in ethanol, dissolved in 60 μl of TE buffer, and fractionated on a 1.0% low–melting point agarose gel. A fraction (2.0 kb) was excised, extracted with phenol and redissolved in 20 μl of TE buffer. A two-step ligation procedure (5) was used to produce a plasmid library in which 99% of the recombinants contained inserts, of which >99% were single inserts. For this project, a total of 5760 double-stranded DNA plasmid templates were prepared in 96-well plates by a boiling bead method [T. R. Utterback *et al.*, *Genome Sci. Technol.* **1**, 1 (1995)]. Reactions were carried out on the AB Catalyst LabStation with Applied Biosystems PRISM Ready reaction Dye Primer Cycle Sequencing Kits for the M13 forward (M13-21) and the M13 reverse (M13RP1) primers. The success rate and average read length after editing with the M13-21 primer were 88% and 444 bp, respectively, and 84% and 435 bp with the M13RP1 primer.
7. S.N. Peterson *et al.*, *J. Bacteriol.* **175**, 7918 (1993).
8. G. Sutton *et al.*,*Genome Sci. Technol.* **1**, 9 (1995).
9. T. S. Lucier *et al.*, *Gene* **150**, 27 (1994); S. N. Peterson *et al.*, *J. Bacteriol.* **177**, 3199 (1995).
10. D. Brutlag *et al.*, *Comput. Chem.* **1**, 203 (1993). The BLOSUM 60–amino acid substitution matrix was used in all protein-protein comparisons [S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1091 (1992)].
11. The nucleotide sequence and peptide translation of each predicted coding region with identified start and stop codons have also been accessioned by Genetic Sequence Data Bank (GSDB). Additional data, including an enhanced version of Table 1 with information on database accessions that were used to identify the predicted coding regions, additional sequence similarity data, and coordinates of the predicted coding regions in the complete sequence are available on the World Wide Web at URL http://www.tigr.org.
12. M. J. Rogers *et al.*, *Isr. J. Med. Sci.* **20**, 768 (1984).
13. T. S. Lucier, unpublished observation.
14. N. Ogasawara *et al.*, *The Bacterial Chromosome*, K. Drlica and M. Riley, Eds. (American Society for Microbiology, Washington, DC, 1990), pp. 287–295; N. Ogasawara and H. Yoshikawa, *Mol. Microbiol.* **6**, 629 (1992).
15. M. Miyata *et al.*, *Nucleic Acids Res.* **21**, 4816 (1993).
16. C. C. Bailey and K.F. Bott, *J. Bacteriol.* **176**, 5814 (1994).
17. M. S. Waterman, *Methods Enzymol.* **164**, 765 (1988).
18. M. Borodovsky and J. McIninch, *ibid.*, p. 123.
19. M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
20. M. J. Rogers *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1160 (1985); W. G. Weisburg *et al.*, *J. Bacteriol.* **171**, 6455 (1989).
21. B. L. Maidak *et al.*, *Nucleic Acids Res.* **22**, 3485 (1994).
22. P. Bork *et al.*, *Mol. Microbiol.* **16**, 955 (1995).
23. B. Sanjanwala and A. T. Ganesan, *Mol. Gen. Genet.* **226**, 467 (1991); B. Sanjanwala and A. T. Ganesan, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4421 (1989).
24. A. Kornberg and T. A. Baker, *DNA Replication* (Freeman, New York, ed. 2, 1992).
25. P. Simoneau *et al.*, *Nucleic Acids Res.* **21**, 4967 (1993).
26. M. Sawada *et al.*, *Mol. Gen. Genet.* **182**, 502 (1981).
27. 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).
28. M. L. Proulx *et al.*, *J. Biol. Chem.* **258**, 753 (1983); J. L. Lapointe *et al.*, *J. Bacteriol.* **165**, 88 (1986).
29. M. A. Strauch *et al.*, *J. Bacteriol.* **170**, 916 (1988).
30. E. Gilson *et al.*, *EMBO J.* **7**, 3971 (1988).
31. P. W. Postma *et al.*, *Microbiol. Rev.* **57**, 543 (1993).
32. P. P. Zhu *et al.*, *Protein Sci.* **3**, 2115 (1994); P. P. Zhu *et al.*, *J. Biol. Chem.* **268**, 26531 (1993).
33. A. M. Collier *et al.*, *Zentralbl. Bakteriol. Suppl.* **20**, 73 (1992); P.-C. Hu *et al.*, *Infect. Immun.* **55**, 1126 (1987).
34. J. M. Inamine *et al.*, *Gene* **82**, 259 (1989).
35. S. F. Dallo and J. B. Baseman, *Microb. Pathog.* **8**, 371 (1990).
36. S. N. Peterson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press (1995).
37. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**, 105 (1982).
38. D. J. McGeoch, *Virus Res.* **3**, 271 (1985; G. von Heijne, *Nucleic Acids Res.* **14**, 4683 (1986).
39. P. Klein *et al.*, *Biochim. Biophys. Acta.* **815**, 468 (1985).
40. A. P. Pugsley, *Microbiol. Rev.* **57**, 50 (1993).
41. B. Guthrie and W. Wickner, *J. Bacteriol.* **172**, 5555 (1990).
42. D. N. Collier, *ibid.* **176,** 4937 (1994).
43. J. M. van Dijl *et al.*, *EMBO J.* **11**, 2819 (1992).

44. Supported in part by a Department of Energy Cooperative Agreement DE-FC02-95ER61962.A000 (J.C.V.), a core grant to TIGR from Human Genome Sciences, American Cancer Society grant NP-838C (H.O.S.), and NIH grants AI08998 (C.A.H.), AI33161 (K.F.B.), and HL19171 (P.-C.H.). We thank A. Glodek, M. Heaney, J. Scott, R. Shirley, and J. Slagel for software and database support; J. Kelley, T. Dixon, and V. Sapiro for computer system support; and C. A. Harger for assistance with the submission of the *Mycoplasma* accession into GSDB. H.O.S. is an American Cancer Society research professor.

11 August 1995; accepted 15 September 1995