Journal of Bacteriology, Dec. 1993, p. 7918–7930
0021-9193/93/127918-13$02.00/0

Vol. 175, No. 24

A Survey of the *Mycoplasma genitalium* Genome by Using Random Sequencing

SCOTT N. PETERSON,[1]* PING-CHUAN HU,[2,3] KENNETH F. BOTT,[1,3] AND CLYDE A. HUTCHISON III[1,3]

Curriculum in Genetics,[1] Department of Pediatrics and Infectious Disease,[2] and Department of Microbiology and Immunology,[3] University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Received 1 July 1993/Accepted 7 October 1993

A total of 508 random clones from five *Mycoplasma genitalium* genomic libraries were partially sequenced and analyzed. This resulted in the identification of 291 unique contigs. Sequence information from these clones (100,993 nucleotides), representing approximately 17% of this pathogen's genome, was analyzed by comparison to the DNA and protein sequence data bases. The frequency with which clones could be identified, by virtue of possessing homology to another data base entry, was 46%. Sequence analysis indicated the following. (i) The *M. genitalium* genome contains many genes involved in various metabolic processes. (ii) Repetitive DNA may comprise as much as 4% of this genome. (iii) The MgPa adhesin gene may be the result of horizontal transfer from an unknown origin. (iv) Not all dinucleotide pairs are present in this genome at the expected frequency. (v) This genome potentially encodes approximately 390 proteins and makes very efficient use of its limited amount of DNA. In addition, this study allowed us to estimate the number of genes involved with various cellular functions.

Mycoplasma genitalium is a bacterial pathogen with a 570- to 600-kb genome (3, 27). This constitutes the smallest genome of any known free-living organism (15, 29). All mycoplasmas lack a cell wall and have small genomes and a characteristically low G+C content (21). Mycoplasmas have a specialized codon usage whereby UGA encodes tryptophan rather than serving as a stop codon (11, 28, 32). Much of the focus with regard to this organism and the closely related *M. pneumoniae* has centered around the characterization of the MgPa and P1 adhesin operons (for a review, see reference 22). Expression of this operon allows adherence to the human host cells (8, 9). It has become clear that other proteins or accessory factors are also required for adherence (14). It is of interest that all of the known repetitive DNA identified in *M. genitalium* and the majority of repetitive DNA in *M. pneumoniae* is in the form of truncated, dispersed copies of various regions of the MgPa and P1 operons, respectively (2, 4, 24). The function or relevance of this repetitive DNA is not understood.

M. genitalium has a single circular chromosome (3) and is proposed to have evolved through a reduction of genetic material from an ancestor common to gram-positive bacteria (23, 30). Although it has been stated, it is not clear whether the current *M. genitalium* genome represents a "minimal genome" or if it is undergoing changes toward reducing its genome even further. The mechanism by which segments of DNA were deleted and what selective pressures exist to fix these events throughout the evolution of this genome are not understood. By obtaining and comparing large amounts of sequence information from several species of *Mycoplasma*, it may be possible to address this point based on examination of breakpoints in regions that differ between *Mycoplasma* species.

Molecular characterization of the *M. genitalium* genome has been hampered by the inability to express *M. genitalium* genes containing UGA trp codons in *Escherichia coli* or other hosts. This is coupled with the difficulty in applying classical genetic tools to the study of this and other mycoplasmas. No auxotrophic mutants have been defined, and the lack of a system for genetic exchange has precluded "reverse" genetic approaches. It is for this reason that sequence determination on a large scale, if not complete, offers a good alternative for characterizing the contents of this genome, as well as shedding light on other novel features of this unique organism. Determining the complete sequence of the *M. genitalium* genome, although arguably worthwhile, is a time-consuming and laborious project. Previously we used a random sequencing approach as a means of defining putative homologs which could then be used as markers on the physical map (20). By surveying this genome in a random manner and analyzing sequences representative of many portions of the chromosome, general features of the genome can be elucidated. As the amount of sequence data analyzed approaches the total amount of sequence information present, the conclusions become more clear and representative. It is for this reason that we chose to apply a random sequencing strategy of this genome on a reasonably large scale.

MATERIALS AND METHODS

***M. genitalium* DNA isolation.** Exponential *M. genitalium* cultures, strain G-37 (approximately 10^9 cells) grown in Hayflick's medium were harvested. The cells were washed in 1× (PBS) and resuspended in 2 ml of 1× PBS. An equal volume of 0.5 M EDTA, pH 9.0–1% sodium dodecyl sulfate–100 μg of proteinase K (Boehringer Mannheim) per ml was added to the cells, and the mixture was incubated at 50°C for 3 h. Two phenol-chloroform extractions, followed by two chloroform extractions, were then performed. DNA was then desalted and concentrated using a Centricon 30 filter (Amicon). Finally, DNA was ethanol precipitated and resuspended at a concentration of 0.5 to 1.0 μg/μl. Chromosomal DNA to be separated by pulsed-field gel electrophoresis was embedded in InCert agarose (FMC Bioproducts) (3). Agarose blocks equilibrated

* Corresponding author.

in restriction enzyme buffer were incubated overnight with 40 U of restriction enzyme at the appropriate temperature.

***M. genitalium* libraries.** Five separate genomic libraries were prepared; four were constructed by digesting genomic DNA to completion with the following enzyme(s): (i) *Eco*RV and *Sma*I, clones 1 to 68 (Table 1); (ii) *Hin*cII and *Sma*I, clones 69 to 109; (iii) *Xba*I, clones 110 to 154; (iv) partially with *Sau*3AI, clones 155 to 266; and (v) *Hin*dIII, clones 267 to 282. DNA from these digests were size fractionated on 1% SeaKem low-melting-point agarose gels (FMC Bioproducts) to select for fragments larger than 300 bp, except in the case of the *Sau*3AI library, which was size selected for fragments between 2 and 4 kb. Ligation reactions were performed by using the vector pUC118, digested with an appropriate restriction enzyme, followed by dephosphorylation with alkaline phosphatase (Boehringer Mannheim). Pulsed-field gel electrophoresis was performed as described previously (20), except gels were 1% SeaKem low-melting-point agarose (FMC Bioproducts). Bands representing X5/X6 from an *Xho*I digestion and S4, S5, S6, and S7/S8 from a *Sma*I digestion were excised (20). The DNA in agarose blocks was treated with 20 U of β-agarase according to the method of the manufacturer (New England Biolabs). DNA was recovered by ethanol precipitation and then digested with *Hin*dIII to produce clones 283 to 291. Fragments generated from this second digestion were then cloned into pUC118.

Sequencing and sequence analysis. Single-stranded templates were prepared in microtiter dishes (10) by using the helper phage M13CO7 (6). Sequencing was performed using the dideoxynucleotide method (25), with the M13 universal primer and DNA polymerase I large fragment (Gibco BRL). Sequences were run on 60-cm 6% polyacrylamide buffer gradient gels (5× to 0.5× TBE). Sequence data were analyzed by using the Genetics Computer Group (GCG) computer program package running on the UNCVX1 system (7). In order to minimize gel reading errors, autoradiographs were read twice by using the GCG program SEQED. The two readings were compared by using GAP. Discrepancies between the two readings were then reexamined to arrive at a final sequence. Sequence files were then converted to Staden format using TOSTADEN. Individual sequences were compared with each other by using the Staden programs for shotgun sequencing projects (26). Redundant sequence information or the presence of overlapping sequence was used to further improve the reliability of sequence information. Unique contigs were identified, and DNA sequence was used to search for sequence homologies in the GenEMBL data base (releases 71.0 to 73.0), by using the program FASTA (19). DNA sequences were translated by using the program MAP and a translation table for mycoplasmas. Long open reading frames (ORFs) were identified, and the deduced amino acid sequence from ORFs were used for comparison to the same versions of the data base using the program FASTA. In cases where significant matches were found, the sequence of the best match was extracted from the data base by using the program FETCH. DNA and amino acid sequence alignments were improved by using the program GAP. The program PILEUP was used in certain cases to compare multiple sequences of homologous genes from different organisms. The GCG program COMPOSITION was used to determine and analyze the G+C and dinucleotide frequency of all sequence data. A codon usage table was made using the program CODONFREQUENCY.

Nucleotide sequence accession numbers. DNA sequences reported here have been submitted to GenBank. Accession numbers assigned are listed in Table 1.

RESULTS

Sequencing and sequence analysis. *M. genitalium* genomic DNA was digested with various restriction enzymes in order to make five different genomic libraries in the vector pUC118. The rationale was to decrease the bias inherent in cloning small DNA inserts produced from any single restriction enzyme. Single-stranded DNA was prepared from white colonies grown in 96-well microtiter dishes (10). Sequencing reactions were performed on a total of 508 clones. Thirty-six of these reactions resulted in no readable sequence. Typically, a single sequencing reaction was performed and nucleotide sequence was read in one orientation from every clone. From the 472 readable sequences, 12 were found to be that of the cloning vector, containing no insert. The Staden programs (26) for shotgun sequencing were used to compare all sequences to one another. This defined 291 unique contigs; 121 clones were the result of cloning the same genomic fragment two or more times; 48 clones contained a sequence which partially overlapped another clone and so were combined to make a single contig. Redundant and overlapping data provided a means of assessing the quality of the sequence data, which we found to be greater than 99% accurate. Redundancy also served as an indicator for determining when continued sequencing of any particular library would be inefficient. All unique sequences were compared with the DNA sequence data base (GenEMBL releases 71.0 to 73.0) by using the program FASTA (19). Sequences were then translated using a translation table modified to account for the fact that in mycoplasmas UGA encodes tryptophan rather than serving as a stop codon (11, 32). Whenever long ORFs were identified, the deduced amino acid sequence was used for comparison to translations of data base entries in all six reading frames by using FASTA. In certain cases, short ORFs at either the beginning or the end of a contig, which plausibly encode the N or C terminus of a protein, were also used for searches. In some instances these resulted in the identification of putative homologs. The term homolog is used here to indicate the strong probability that the sequences in question are derived from a common ancestor.

The results of these searches are summarized in Table 1. In all cases where significant matches were found in FASTA searches, alignments were repeated using the program GAP. The percentages of identity and similarity obtained by these alignments are those reported in Table 1. We found that the data base searches provide an extremely useful method for identifying potential homologs in *M. genitalium*. In 46% of the clones, a significant data base match was found. In some cases, one contig contained sequence information for two ORFs and in 14 cases provided matches to two genes of separate function. The largest number of matches were found with *Bacillus* species (34 matches), and *E. coli* (33 matches). We believe that the large number of matches with genes of gram-negative bacteria represents an artifact of overrepresentation of the *E. coli* genome in the GenEMBL data base. In most cases where homologs were present in both gram-negative and gram-positive organisms, the best score was obtained for the gram-positive bacteria. The other striking but perhaps expected feature of the data is the large percentage (96%) of random clones containing long ORFs. Only 11 clones were encountered which neither were homologous to RNA species nor contained ORFs of significant length.

In some cases further analysis was necessary to either eliminate or gain greater support for matches of questionable significance. This was done in two ways. Frequently, data base alignments from FASTA were obtained where strong levels of identity or similarity existed but only in a portion of the

TABLE 1. Summary of data base searches

Clone[a]	Accession no.[b]	Length (nucleotides)	ORFs[c]	Homology/accession no.[d]	% Identity/match length		% Similarity
					Nucleotides[e]	Amino acids[f]	
1. esa1	U01692	291	1-291	ECOTGASNS/M33145	53	49/96	68
2. esa2	U01695	285	1-285				
3. esa3	U01696	294	1-294	BACORIC/X02369	55	47/97	67
4. esa4*	U01697	338	0				
5. esa5*	U01698	345	1-345	STRUVS402A/M80215	56	59/114	77
6. esa6+	U01699	480	1-309				
7. esa7+	U01700	410	1-410				
8. esa8	U01701	334	1-334				
9. esa9	U01702	313	1-255				
10. esa10	U01693	350	160-350	STRATPASEA/M90060*	44	37/61	68
11. esa11	U01694	290	1-290	MYCMGP/M31431	100	100/96	100
12. esb1+	U01703	552	1-527	PRPUNC2/X58461	43	22/175	48
13. esb2+*	U01707	640	1-640	ECOPHOS/K01992	53	51/213	69
14. esb3+*	U01708	750	1-750	LBALLDHD/D90340	48	41/249	65
15. esb4	U01709	297	35-297				
16. esb5+*	U01710	645	1-645				
17. esb6+	U01711	618	1-312	BACPHEST/X53057	59	49/104	67
			336-618	BACPHEST	41	28/94	62
18. esb7*	U01712	387	1-387	BACPOLC/M22996	59	57/129	74
19. esb8*	U01713	366	1-366				
20. esb10*	U01704	279	1-279	SMARECA/M22935	53	58/93	72
21. esb11+	U01705	662	1-662				
22. esb12	U01706	303	1-303				
23. esc1	U01714	293	1-293				
24. esc5+*	U01718	439	1-285	STATN4003/X13290	61	58/95	71
			329-439				
25. esc6+	U01719	405	70-405	ECOAPAH/X04711	42	30/111	52
26. esc7+	U01720	362	1-362	MUSESKK/M86377	48	33/120	59
27. esc8+	U01721	299	1-296				
28. esc10+	U01715	576	1-83				
			107-576				
29. esc11	U01716	325	1-81	MYCHMW3A/M82965	67	58/23	71
			100-325				
			100-325				
30. esc12	U01717	223	1-223				
31. esd1+	U01722	688	1-688	TTHFUS/X16278	52	57/229	76
32. esd2	U01726	260	1-129	BACSPCR/M31102	50	65/43	49
			132-260	BACSPCR	58	46/39	54
33. esd3+	U01727	377	1-377	MYCATPA/M29168	68	59/125	75
34. esd4	U01728	299	45-299				
35. esd5+	U01729	454	1-420				
36. esd6	U01730	297	1-297				
37. esd7	U01731	307	96-307				
38. esd8+*	U01732	623	1-623	BACSECA/D90218	46	39/207	66
39. esd10	U01723	304	1-44	BACHSP/M84964	63	75/13	92
			90-304	YSCMOT1/M83224	46	42/71	63
40. esd11+*	U01724	712	1-712	BORGRPEPLS/M96847	47	35/237	55
			1-330				
41. esd12+	U01725	638	500-638	BACLDHA/M19395	50	37/44	63
42. ese3*	U01735	369	1-369	STYRPOBG/X04860	52	57/123	75
43. ese8	U01736	292	1-292				
44. ese11+	U01733	600	1-351	BACALPHA/M26414	57	60/117	73
			354-600	BACALPHA	51	40/81	61
45. ese12	U01734	305	27-305				
46. esf2	U01739	344	21-344	STYPROVW/X52693*	52	48/102	62
47. esf4	U01740	319	1-319	CORXLYSA/X54740*	32	39/105	63
48. esf8	U01741	313	1-313				
49. esf11	U01737	338	1-338	STYRPOBZ/M38311	43	42/112	67
50. esg3	U01746	229	1-229				
51. esg6	U01748	303	1-273				
52. esg7	U01749	284	1-284				
53. esg8	U01751	288	1-243				
54. esg10	U01742	303	1-303				
55. esh3	U01756	186	1-186				
56. esh5	U01757	225	1-225				
57. esh8+	U01758	306	1-306				

Continued on following page

TABLE 1—*Continued*

Clone[a]	Accession no.[b]	Length (nucleotides)	ORFs[c]	Homology/accession no.[d]	% Identity/match length		% Similarity
					Nucleotides[e]	Amino acids[f]	
58. esh9	U01759	311	196-311				
59. esh10*	U01753	366	1-366	MYCMGP/M31431	87	74/112	82
60. esh12	U01754	265	1-222	BACMBR/M77837	51	34/66	57
61. esf1a	U01738	284	1-284				
62. esg1a+	U01744	620	1-117	ECORPSI/X02130	50	41/39	62
			127-520	ECORPSI	48	47/131	62
			561-620				
63. esg2a+	U01745	524	1-478	PSELEPALEP/X56466	53	48/159	73
64. esg3a	U01747	135	20-135				
65. esg7a	U01750	295	1-177				
			165-295				
66. esg9a+	U01752	406	1-406	CYTATPB/M22535	69	74/135	86
67. esg12a	U01743	365	1-150				
			120-365	BACCSBA/M80473	56	56/77	68
68. esh1a	U01755	217	1-170				
69. hsa1+	U01760	501	1-450	SMESPIRG/M31161	41	38/144	59
70. hsa2	U01762	171	1-171				
			1-171				
71. hsa3	U01763	300	1-300				
72. hsa4	U01764	340	1-340				
73. hsa5	U01765	129	1-129	BACIF2G/X04399	51	38/43	60
74. hsa6	U02115	201	1-201				
75. hsa7+	U01766	467	1-104				
			108-467	MYCMGP/M31431	84	79/119	85
76. hsa8+	U01767	1,134	1-1134				
77. hsa9+	U01768	705	1-374				
			425-625				
78. hsa11	U01761	330	1-180	TTHDNALGS/M74792	48	48/60	65
			180-330	TTHDNALGS	42	34/50	56
79. hsb1+	U01769	541	1-323				
80. hsb2	U01772	229	1-229	ECOTIG/M34066	39	29/76	53
81. hsb3	U01773	302	1-206	YSCFUR1A/M36485	45	35/68	58
			162-302				
82. hsb4	U01774	289	1-236				
83. hsb5′	U01775	420	1-420				
84. hsb6	U01776	224	1-224	BACOPPOPER/X56347	37	34/74	59
85. hsb8	U01777	264	1-264				
86. hsb9+	U01778	652	1-652	ECONUSA/X00513	39	24/217	49
87. hsb10	U01770	308	2-282	ECOSPOT/M24503	43	29/94	54
88. hsb12+	U01771	572	1-292				
			340-572				
89. hsc3	U01781	292	1-218				
			252-292				
90. hsc4+	U01782	431	115-431				
91. hsc6	U01783	269	1-78				
			81-269				
92. hsc7	U01784	301	1-301	ECOACE/V01498	47	32/99	52
93. hsc8+	U01785	423	38-423				
94. hsc11	U01779	165	1-65	MYCMGP/M31431	100	100/55	100
95. hsc12	U01780	210	1-210	BACLEUS/M88581	58	57/69	80
96. hsd1	U01786	280	1-114	ECOAPT/M14040	42	39/38	53
			170-280				
97. hsd3*	U01789	381	1-324	MYCMGP/M31341	79	54/108	67
98. hsd5	U01790	312	1-291				
99. hsd9	U01791	326	1-326				
			1-326				
100. hsd11*	U01787	403	5-403				
101. hsd12	U01788	327	1-327				
102. hse1	U01795	277	1-277				
103. hse2	U01796	291	1-75				
			113-291				
104. hse3	U01797	361	1-361	ECORPOBC/V00339	54	58/120	77
105. hse4	U01798	329	1-329	ECOPK1/M24636	52	53/109	63
106. hse6*	U01799	296	1-296				
107. hse7	U01800	342	1-342	ECORF1X/M11519	50	49/113	72
108. hse8	U01801	321	1-321				

Continued on following page

TABLE 1—*Continued*

Clone[a]	Accession no.[b]	Length (nucleotides)	ORFs[c]	Homology/accession no.[d]	% Identity/match length		% Similarity
					Nucleotides[e]	Amino acids[f]	
109. hse9+	U01802	324	1-324	RIRPEPA/M68966	52	35/108	58
110. x1*	U01803	336	1-336	CHTDNAC/Y00505	47	54/112	36
111. x3	U01808	322	1-322				
112. x4	U01809	276	1-276				
113. x5+*	U01810	917		MYCTGWB/M32341	100/182		
			352-533				
			662-917	MYCMGP/M31431	83	78/84	85
114. x6	U01811	345	1-345				
115. x7	U01812	285	1-285	BACORIGS/X62539	61	59/94	78
116. x8	U01813	192	1-192				
117. x9+	U01814	1,006	1-530	ECOASPS/X53863	46	33/176	61
			660-1006				
118. x10	U01804	305	1-305				
119. x11	U01805	220	11-220				
120. x16	U01806	182	1-182				
121. x17	U01807	229	1-229	BACPOLC/M22996	48	41/76	64
			1-229				
122. x19	U02266	180	1-180				
123. x21	U02267	214	1-214				
124. x23*	U02268	472	1-236	BACHSPA/M84965	57	48/78	65
			247-472				
125. x24	U02269	315	56-315				
126. x29	U02218	350	1-350				
127. x30	U02219	320	1-280				
128. x34	U02220	360	1-360	ECOAMSG/M62747*	43	29/119	61
129. xfa4	U02244	263	0				
130. xfb3+	U02245	515	1-145				
			126-515				
131. xfb5	U02246	270	1-270				
132. xfc5	U02247	247	1-247	BACTYRSBR1/M77668	50	43/81	62
133. xfc7	U02248	227	1-227	YSCGAP1P/X52633	43	37/75	56
134. xa6	U02225	246	0				
135. xa7+	U02226	326	1-326	BACPGK/X54519	54	34/108	66
136. xa8	U02227	323	0	ACLTRNA11/X61068	73/323		
137. xa9+	U02228	304	76-304				
138. xa10	U02224	341	1-341	MYCHMW3A/M82965	57	54/113	69
139. xb8	U02230	323	0				
140. xb12	U02229	333	1-201				
			165-333	TTHTRSYN/M64273	54	49/42	70
141. xc2	U02232	250	0				
142. xc3	U02233	265	1-265				
143. xc4	U02234	305	1-305	BACPGK/X54519	54	49/101	66
144. xc5	U02235	326	3-326				
145. xc12	U02231	322	1-322				
146. xd3+	U02238	349	1-349	ECOFMT/X63666	47	31/116	61
147. xd5	U02239	320	62-320				
148. xd6	U02240	348	17-348				
149. xd10	U02236	276	43-276				
150. xd12	U02237	310	1-129				
			126-310				
151. xe5	U02241	314	1-314				
			1-314				
152. xf1	U02242	394	1-394	ECOTOPA/X04475	47	30/131	52
153. xf10	U02243	337	1-337				
154. xh1	U02249	305	1-111				
			143-292				
155. sc4	U02144	345	1-115				
			221-345				
156. sc5+	U02146	418	1-418	BACDNAE/M10040	42	21/139	50
157. sc12+	U02140	367	1-367	MYCMGP/M31431	71	63/122	72
158. sd3	U02156	308	1-308				
159. sd4	U02158	301	1-301				
160. sd5	U02160	313	1-313				
161. sd6	U02162	326	1-326				
162. sd7+	U02163	387	1-387				
163. sd8	U02165	309	1-309				

Continued on following page

TABLE 1—*Continued*

Clone[a]	Accession no.[b]	Length (nucleotides)	ORFs[c]	Homology/accession no.[d]	% Identity/match length		% Similarity
					Nucleotides[e]	Amino acids[f]	
164. sd9	U02167	336	1-336	ECOLEUS/X06331	49	42/112	60
165. sd11	U02152	294	1-294	TTHDNALIG/M36417	38	40/98	63
166. sd12	U02153	325	1-325	MYCRPCLUS/X06414	56	50/108	70
167. se1	U02168	309	1-309				
168. se2	U02173	353	1-353				
169. se4	U02176	377	1-74	ECOHIST1/X02743	33	23/24	45
			70-377	ECOHIST1	39	29/101	47
170. se7	U02179	305	1-305	YSCMOT1/M83224	50	37/101	60
171. se8	U02181	267	1-267				
172. se9	U02183	371	1-371	BACGLTXA/M55073	49	43/123	61
173. se11	U02169	361	1-361				
174. se12	U02171	346	1-305	MYCP372969/M37339	48	33/92	54
175. sf1	U02185	373	27-373				
176. sf2	U02192	355	1-355	STRPAGA/D90354	43	32/110	52
177. sf5	U02194	344	1-344				
178. sf6	U02196	334	1-334	YSCILSI/M30942	49	32/110	53
179. sf7+	U02198	309	1-309				
180. sf8	U02200	364	1-265				
			275-364				
181. sf9*	U02201	475	1-475	YSCUNG1A/J04470	48	35/158	54
182. sf10	U02186	302	0				
183. sf12	U02189	303	1-303				
184. sg1	U02202	330	1-330	BACVALS/M16318	50	34/109	56
185. sg2	U02208	347	1-347	BACPOLC/M22996	52	48/115	70
186. sg3	U02209	367	1-367	MYCMGP/M31431	100	100/122	100
187. sg4	U02210	322	1-322				
188. sg6	U02213	364	1-247	BACGAPDHA/M24493	52	49/80	65
			268-364				
189. sg7	U02215	366	1-245				
			235-366				
190. sg8*	U02217	409	11-409	MYCMGP/M31431	85	84/127	91
191. sg9+	U02251	403	1-403				
192. sg10	U02203	356	1-356				
193. sg11	U02205	346	1-263				
			216-346				
194. sg12	U02206	345	1-213	TMONUSG/Z11839	41	24/71	53
			240-345	STYRPLJL/X53072	56	38/34	65
195. sh2	U02258	311	1-311	ABCCELA/M76548	41	34/103	50
196. sh5	U02260	342	1-342				
197. sh7+	U02262	328	1-328				
198. sh8	U02264	347	1-347				
199. sh9	U02265	339	1-339				
200. sh11+	U02253	649	1-381				
			385-649				
201. sh12	U02255	342	1-342	MYCENTUF/X16463	100	100/114	100
202. sa1	U02122	379	9-379	BACGLTXA/M55072	45	26/124	48
203. sa3	U02126	174	1-174				
204. sa4	U02127	234	49-234				
205. sa5	U02128	299	1-299				
			1-299				
206. sa7	U02129	315	1-315	BACOPPOPER/X56347	56	47/105	67
207. sa8	U02130	342	1-342	BACTRNASB/M36594	53	43/114	67
208. sa9	U02131	356	1-356	RHBGLYA/X54638	50	57/118	70
209. sa10	U02123	284	1-284	ECOMETX/M98266	47	40/94	64
210. sa11*	U02124	475	1-224	MYCMGP/M31431	88	88/71	90
211. sa12	U02125	212	1-212				
			1-212				
212. sb8	U02135	260	0				
213. sb9*	U02136	410	1-180	TTHFUS/X16278	50	57/60	67
			184-410	ECORPSFRI/X04022	47	29/57	53
214. sb10+	U02132	571	0				
215. sb11+	U02133	301	1-301	ECOLEP/K00426	52	53/99	65
216. sb12	U02134	251	1-251	ECOTOPA/X04475	35	25/83U45	
217. sc1	U02137	269	1-192				
218. sc2*	U02142	404	1-404	MYCMGP/M31431	82	73/134	84
219. sc3	U02143	295	1-69				

Continued on following page

TABLE 1—*Continued*

Clone[a]	Accession no.[b]	Length (nucleotides)	ORFs[c]	Homology/accession no.[d]	% Identity/match length		% Similarity
					Nucleotides[e]	Amino acids[f]	
			72-295				
220. sc4a	U02145	352	1-352	MYCDNAA/D90426	43	21/117	47
221. sc6a+	U02147	301	75-301				
222. sc7a+	U02148	370	1-370	ECOLONA/M38347	50	47/123	67
223. sc8a+	U02149	681	1-195	ECOMGLABCO/M59444	53	47/65	61
			243-681	ECOGALET/X06226	44	30/146	56
224. sc9a	U02150	350	1-350				
225. sc10a	U02138	323	1-323	XANFRUKAA/M69242	48	46/107	63
226. sc11a	U02139	312	1-312				
227. sc12a*	U02141	750	117-437	BACSPOIVFO/X59528	51	36/100	64
			415/729	BACSPOIVFO*	41	27/99	53
228. sd2a	U02155	308	1-308				
229. sd3a*	U02157	576	1-218	MYCMGP/M31431	89	92/72	92
				MYCRRNOP/M21374	76/360		
230. sd4a+	U02159	549	1-549	MYCMGP/M31431	99	99/183	99
231. sd5a	U02161	335	1-335	MYCMGP/M31431	100	100/111	100
232. sd7a	U02164	370	1-370				
233. sd8a	U02166	378	1-378				
234. sd10a	U02151	309	1-309				
235. sd12a	U02154	354	1-129	STRRECP/M31296	53	33/41	55
			134-354				
236. se2a+	U02174	333	1-333				
			1-333				
237. se3a	U02175	335	1-335				
238. se4a	U02177	271	1-201	MYCP115A/M34956	54	48/67	61
			209-271				
239. se5a	U02178	333	1-177	TTHYT1GAP/X16595	47	38/59	58
			158-333	BACPGK/X54519	52	49/49	62
240. se7a	U02180	340	1-340	TTHFUS/X16278	56	64/113	83
241. se8a	U02182	341	1-341				
242. se9a+	U02184	338	1-338	STARECF/M86227	64	63/112	75
243. se11a	U02170	369	1-369				
244. se12a	U02172	318	18-303	ECOUVRA/M13495	58	71/82	87
254. sf1a	U02191	183	1-103				
			99-183				
246. sf2a	U02193	272	1-272	VIBHPT/X53382	44	26/90	54
247. sf5a	U02195	290	1-290	ECOPBPBRR/X52063	41	25/96	53
248. sf6a	U02197	322	1-322	CLORUB/M60116	52	36/107	57
249. sf7a	U02199	316	1-316	MYCGYRBA/X53555	78	96/104	98
250. sf10a	U02187	321	1-321	MYCGYRBA/X53555	79	85/106	94
251. sf11a	U02188	287	1-287				
252. sf12a	U02190	294	1-252				
253. sg4a+	U02211	387	1-139	MYCGYRBA/X53555	80	96/46	96
			157-387	MYCGYRBA/X53555	77	77/76	84
254. sg5a	U02212	394	1-309	TTHS127FU/X52165	50	55/101	72
			326-394	TTHS127FU	42	48/22	52
255. sg6a	U02214	359	1-359	ECOAMSG/M62747*	37	38/119	61
256. sg7a	U02216	321	1-273	BACORIGS/X62539	45	35/91	61
257. sg8a	U02250	337	0				
258. sg9a	U02252	297	1-187	CLOGROESL/X62914	73	67/59	85
			197-297	CHTGROE/M58027	43	28/33	52
259. sg10a	U02204	327	1-327	CLOHSP70G/X62915	73	74/108	82
260. sg12a	U02207	279	1-276				
			1-279				
261. sh1a	U02257	296	1-296				
262. sh3a	U02259	299	1-299	ECODNAAOP/J01602	47	37/99	59
263. sh6a	U02261	382	1-382				
			1-382				
264. sh7a	U02263	341	1-341				
265. sh11a	U02254	324	1-324				
266. sh12a	U02256	272	1-272				
267. ha6	U02100	380	31-380	ECOHIST1/X02743	39	24/116	50
268. ha7	U02101	113	1-113				
269. ha8	U02102	345	1-345				
270. ha10+	U02099	201	1-201				
			1-201				

Continued on following page

TABLE 1—*Continued*

Clone[a]	Accession no.[b]	Length (nucleotides)	ORFs[c]	Homology/accession no.[d]	% Identity/match length		% Similarity
					Nucleotides[e]	Amino acids[f]	
271. hb4	U02103	309	1-309				
272. hb5+	U02104	314		MYCTGTYQK/M18050	75/163		
			212-314				
273. hb7	U02105	277	157-277	MYCMGP/M31431	92/117	92/37	92
274. hc8	U02107	196	0				
275. hc10	U02106	284	1-76	MYCMGP/M31431	91/53	93/14	93
				LBATRNA2/X15246	82/70		
276. he1	U02108	212	1-71				
			65-212				
277. hg1	U02109	277	1-270	PFATPIX/L01654	60	54/90	66
278. hg4	U02110	218	1-59	MYCMGP/M31431	88/40	92/12	92
			116-218				
279. hg7	U02111	215	1-54				
			33-215				
280. hg9	U02112	229	1-229				
281. hh4	U02113	278	1-278	TMONUSG/Z11839	58	51/90	74
282. hh9	U02114	298	1-298				
283. s7s8a10	U02120	166	1-166				
284. x5x6e3	U02222	193	1-193				
285. x5x6e6	U02223	117	1-117				
			1-117				
286. s4h10	U02118	317	1-317	STAHVR/X52594	48	31/105	52
287. s7s8b3	U02121	231	1-231				
			1-231				
288. s6d5	U02119	391	1-391	ECOUVRB2/X03678	48	37/130	57
289. x5x6d11	U02221	393	1-393				
290. s4a6	U02116	167	1-167				
291. s4a8	U02117	174	1-174				

[a] *(next to the clone name) indicates clones which were sequenced twice for clarify or longer readings, or primed a second time with a specific oligonucleotide. indicates that two or more clones overlapped to form that contig.

[b] Each of the 291 sequences was submitted to the National Center for Biotechnology Information by using AUTHORIN.

[c] The length of an ORF was calculated from the number of nucleotides between stop codons. In cases where two long ORFs were found in any single clone, they are both listed.

[d] GenBank homologous file. * next to the accession name of the putative homolog indicates that the data base sequence was referred to as ORF X.

[e] Only the percentage identity (at the nucleotide level) is given in cases where the reported match corresponds exactly to an amino acid sequence match. In those cases where a match length is stated, it is in nucleotides (75/163 means a 75% match over a region of 163 nucleotides).

[f] Percentage identity and match length in amino acids were calculated by using the program GAP. The similarity scores for each amino acid match, calculated by the same program, are listed in the next column.

alignment. In such cases the sequence for the strongest match was compared with the *M. genitalium* sequence by using the GCG program GAP. Often this treatment extended the significant similarities between the two proteins through the entire sequence, thus enhancing the confidence of the match. In cases where this was not true, the homology was considered dubious and not entered into Table 1. As a general rule, alignments were improved by placing gaps on the order of 1 to 10 amino acids in the *M. genitalium* protein rather than the converse.

The second method employed to gain confidence in matches required that three or more homologous sequences from different organisms be aligned to the target *M. genitalium* sequence. The GCG program PILEUP was used to align all of the amino acid sequences of interest. By examining the data in this manner, the degree of amino acid conservation could be assessed. This was especially useful for protein homologs where a relatively small number of scattered amino acids were conserved in different species. Invariant amino acids in the multiple alignment output were checked visually against the *M. genitalium* sequence. In cases where conservation at these key positions was maintained, the clone was considered a significant match and is included in Table 1. These homologs can be further classified according to the major cellular function they may perform (Table 2).

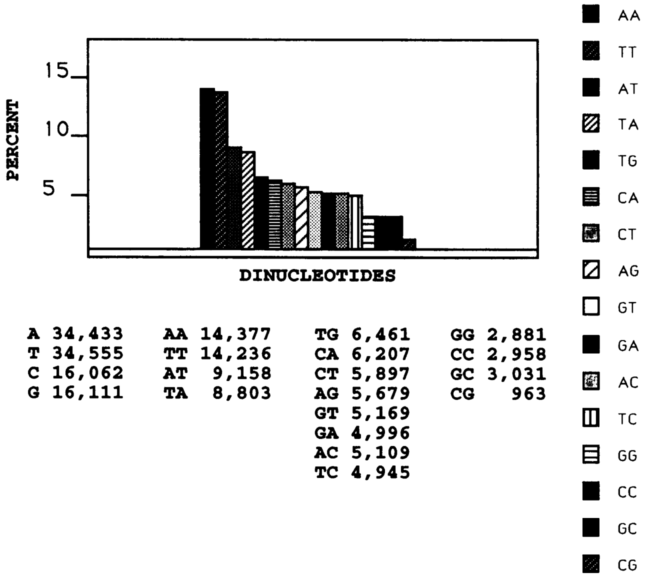
To establish that the sequencing data approximate a random sampling of the genome, we counted the number of sequences in existing contigs that contain overlapping sequences. In this experiment, 339 nonidentical clones contributed to the definition of 291 unique contigs. In other words, 48 (or 16%) of the sequences overlapped existing contigs. This is in close agreement with the estimate, based on sequence length, that we have sequenced approximately 17% of the genome, given a genome size estimate of 580 kbp (3, 27). Taking this to indicate that no particular bias is present in the representation of sequence data, it is instructive to extend our results to the remainder of the genome in order to gain insight into the coding capacity of this organism.

In the 148 data base matches, 97 different proteins, 8 tRNAs, 1 rRNA, and 12 clones representing repetitive DNA were identified. By taking the predicted lengths of the nucleotide sequences required to code for each of the 97 protein matches identified and adding them together, we can estimate the percentage of the genome's coding capacity that our sequence represents. The number obtained is 145,858 nucleotides or 25% of the genome. Since this only represents the number of nucleotides present from data base matches (46%), ignoring for the moment RNAs and MgPa repetitive DNA, then the 54% of the random sequences for which we did not find significant homology to data base entries may represent

TABLE 2. Distribution of *M. genitalium* clones by function

- Adherence
  - 11,[a] esa11; 94, hsc11; 186, sg3; 230, sd4a; 231, sd5a adherence MgPa
  - 29, esc11; 138, xa10 accessory adherence proteins HMW3A
- Membrane transport
  - 13, esb2 phosphate transport
  - 38, esd8 secretion protein
  - 60, esh12 membrane binding protein
  - 84, hsb6, *oppC* oligopeptide transport
  - 133, xfc7 general amino acid permease
  - 174, se12 *M. hyorhinis* p69 membrane protein
  - 176, sf2 surface protein antigen
  - 206, sa7 *oppD* oligopeptide transport
  - 223, sc8a galactose binding protein
  - 238, se4a *M. hyorhinis* 115-kDa protein
- Recombination/repair
  - 5, esa5; 67, esg12a; 288, s6d5 *uvrB* excision repair
  - 20, esb10 *recA* homologous recombination
  - 181, sf9 uracil-*N*-glycosylase
  - 235, sd12a *recP*
  - 244, se12a *uvrA* excision repair
- Metabolic pathways
  - Glycolytic enzymes
    - 14, esb3; 41, esd12 lactate dehydrogenase
    - 69, hsa1; 105, hse4 pyruvate kinase
    - 135, xa7; 143, xc4; 239, se5a phosphoglycerate kinase
    - 188, sg6; 239, se5a glyceraldehyde-3-phosphate dehydrogenase
    - 277, hg1 triosephosphate isomerase
  - Other
    - 24, esc5 thymidylate synthase
    - 81, hsb3 uracil phosphoribosyltransferase
    - 87, hsb10 *spoT* (p) ppGpp 3′pyrophosphohydrolase
    - 92, hsc7 lipoamide dehydrogenase
    - 96, hsd1 adenine phosphoribosyltransferase
    - 195, sh2 UDP pyrophosphorylase
    - 208, sa9 glycine hydroxymethyl transferase
    - 223, sc8a UDP-galactose-4-epimerase
    - 225, sc10a PTS enzyme-II fructose permease
    - 246, sf2a hypoxanthine phosphoribosyl transferase
    - 248, sf6a thioredoxin reductase
- Translation
  - tRNA synthetases
    - 1, esa1 asparaginyl-tRNA synthetase
    - 17, esb6 phenylalanine-tRNA synthetase α subunit
    - 17, esb6 phenylalanine-tRNA synthetase β subunit
    - 95, hsc12; 164, sd9 leucyl-tRNA synthetase
    - 117, x9 aspartyl-tRNA synthetase
    - 132, xfc5 tyrosyl-tRNA synthetase
    - 140, xb12 methionyl-tRNA synthetase
    - 146, xd3 methionyl-*N*-formyl-tRNA synthetase
    - 172, se9; 202, sa1 glutamyl-tRNA synthetase
    - 178, sf6 isoleucyl-tRNA synthetase
    - 184, sg1 valyl-tRNA synthetase
    - 207, sa8 threonyl-tRNA synthetase
    - 209, sa10 *s*-adenosylmethionine synthetase
  - Ribosomal proteins
    - 32, esd2 ribosomal proteins S5
    - 32, esd2 ribosomal protein L15
    - 44, ese11 ribosomal proteins S13
    - 44, ese11 ribosomal protein S11
    - 62, esg1a ribosomal proteins L13
    - 62, esg1a ribosomal protein S9
    - 166, sd12 ribosomal protein L3
    - 194, sg12 ribosomal protein L7
    - 213, sb9 ribosomal protein S6
    - 227, sc12a ribosomal protein L21
    - 254, sg5a ribosomal protein S7
    - 281, hh4 ribosomal protein L1
  - Other
    - 25, esc6 16S rRNA methyltransferase
    - 31, esd1; 213, sb9; 240, se7a elongation factor G
    - 73, hsa5 translation initiation factor
    - 107, hse7 peptide chain release factor
    - 113, x5 tryptophan tRNA
    - 136, xa8 leucine, lysine, threonine, valine tRNA
    - 201, sh12 elongation factor Tu
    - 229, sd3a 16s rRNA promoter
    - 272, hb5 glutamine, tyrosine tRNAs
    - 275, hc10 arginine tRNA
- DNA synthesis/cell division
  - 3, esa3; 253, sg4a gyrase A
  - 18, esb7; 121, x17; 185, sg2 DNA polymerase III
  - 39, esd10; 170, se7 helicase
  - 78, hsa11; 165, sd11 DNA ligase
  - 80, hsb2 trigger factor
  - 110, x1 *dnaB* primosome protein
  - 115, x7 *gidA*, replication initiation
  - 152, xf1; 216, sb12 topoisomerase
  - 156, sc5 *dnaE* primase
  - 220, sc4a; 256, sg7a; 262, sh3a *dnaA* (initiation factor)
  - 242, se9a; 249, sf7a; 250, sf10a; 253, sg4a gyrase B
  - 247, sf5a cell division regulation?
- ATP production and utilization
  - 12, esb1 *uncG* F1 subunit ATP synthetase pathway
  - 33, esd3 ATP synthetase
  - 66, esg9a ATP synthetase β subunit
- Heat shock
  - 39, esd10; 40, esd11 *dnaJ*
  - 124, x23; 258, sg9a *groEL*
  - 222, sc7a heat shock protease
  - 258, sg9a *groES*
  - 259, sg10a *dnaK*
- Transcription
  - 42, ese3; 49, esf11, RNA polymerase β subunit
  - 86, hsb9 N utilization factor
  - 104, hse3 RNA polymerase β′ subunit
  - 194, sg12 nusG
- Protein modification
  - 26, esc7 protein kinase
  - 63, esg2a; 215, sb11 leader peptidase
  - 109, hse9 aminopeptidase
- Repetitive DNA
  - 59, esh10; 75, hsa7; 97, hsd3; 113, x5; 157, sc12; 190, sg8; 210, sa11; 218, sc2; 229, sd3a; 273, hb7; 275, hc10; 278, hg4
- Unknown
  - 10, esa10
  - 46, esf2
  - 47, esf4
  - 128, x34
  - 169, se4
  - 227, sc12a
  - 286, s4h10
  - 255, sg6a

[a] Numbers correspond to those in Table 1.



A	34,433	AA	14,377	TG	6,461	GG	2,881
T	34,555	TT	14,236	CA	6,207	CC	2,958
C	16,062	AT	9,158	CT	5,897	GC	3,031
G	16,111	TA	8,803	AG	5,679	CG	963
				GT	5,169		
				GA	4,996		
				AC	5,109		
				TC	4,945		

FIG. 1. Dinucleotide analysis of *M. genitalium* random clones. Totals were counted from 100,993 nucleotides by using the program COMPOSITION.

171,225 nucleotides. Thus our coding region sequence may represent a sampling of genes occupying 317,082 nucleotides or approximately 55% of the genome. If to this we add 800 nucleotides for 8 tRNAs, 5,000 nucleotides for one rRNA operon and 23,200 nucleotides of repetitive DNA (see below), we estimate that genes occupying 340,282 nucleotides or 59% of the genome's coding capacity are potentially represented in these sequence data.

Having estimated that the 97 protein coding genes identified by data base searches represent approximately 25% of the genome, we can estimate that the total number of proteins potentially encoded by the *M. genitalium* genome is 388. Two-dimensional polyacrylamide gel electrophoresis experiments performed with *Mycoplasma capricolum* identified approximately 350 polypeptides (13). It is possible that this number represents an underestimate, given that the genome of this species is as much as twice the size of *M. genitalium* (16).

Sequences such as tRNAs, rRNAs, ribosomal proteins, and in this organism, MgPa and repetitive DNA having homology to the MgPa operon, are well represented in the data base and possess strong sequence conservation across species. For this reason such sequences are highly identifiable in data base searches whenever they are used as a query sequence. By virtue of this fact, we are able to predict that the *M. genitalium* genome possesses about 32 tRNAs, which is in good agreement with 29 tRNAs present in the *M. capricolum* genome, where the complete set of tRNAs has been identified (1). We estimate that there are about 52 ribosomal proteins, which is identical to the number of different proteins found in the *E. coli* ribosome (31). The number of rRNA genes is known to be three, as there is only one rRNA operon in this genome (33). Likewise, there is only one copy of the MgPa operon (12). We have estimated the fraction of repetitive DNA in this genome to be approximately 4%. We arrived at this estimate by dividing the frequency of repetitive clones in this data set (12) by the 291 unique clones analyzed.

Dinucleotide analysis. The G+C content of the sequence data was determined to be 32%, which is identical to that determined previously by chromatographic analysis of hydrolyzed nucleotides from the entire genome (29). While the majority of dinucleotides are found in their expected frequencies for a genome of low G+C content, there are two striking discrepancies (Fig. 1). The dinucleotides AA and TT are present at greater than expected frequencies. The relevance of this finding is not clear. Of greater interest was the observation that the dinucleotide CpG is present three times less frequently than GpC. This inequality led us to speculate that cytosine methylation may exist in *M. genitalium*. Methylated cytosines, when deaminated, yield thymine or a T-G base pair. After DNA replication the dinucleotide CpG becomes TpG; on the other strand a CpA is formed. These two dinucleotides, TpG and CpA, are the most abundant in their class.

CpG methylation is a phenomenon normally associated with eukaryotes; however, it has been reported in at least one mycoplasma, *Mycoplasma hyorhinis*, and some spiroplasmas (18). That study showed that *Spiroplasma* sp. strain MQ-1 had over 95% of its cytosines methylated in the context CpG. By nearest-neighbor analysis it was shown that the dinucleotide CpG was underrepresented (0.45% found versus 2.25% predicted). Strain MQ-1 was also shown to possess a methylase activity. In our analysis, restriction enzyme digestions of *M. genitalium* genomic DNA, using *Msp*I and *Hpa*II, did not support the fact that CpG methylation currently exists in this genome as evidenced by the identical pattern produced by both restriction enzymes (data not shown). Whether the disparity in CpG dinucleotides in the *M. genitalium* genome is the result of a now extinct CpG methylase activity or related instead to the codon usage of this organism will require further analysis.

Codon usage in *M. genitalium*. A codon usage table was constructed from all of the sequences which were found to have data base homologs, with the exception of matches to MgPa and MgPa repetitive DNAs (Table 3). This codon usage table will assist in identifying the most likely ORF in these and future sequences, which are unidentifiable in data base searches, so that alternate approaches may be employed for determining their function. The data, derived from 12,680 amino acids, are positioned next to the codon usage information of the MgPa and P1 adhesin genes (5). Examining the data in this manner shows clear differences in the codon bias between putative *M. genitalium* genes when compared with adhesin genes from *M. genitalium* and *M. pneumoniae*. It can be seen that the MgPa and P1 genes do not discriminate as strongly against G or C in third positions of codons as does the remainder of the genome. *M. genitalium* protein coding sequences are more strongly biased against use of these nucleotides in the third position. The codon usage data derived from non-MgPa random sequences is consistent with codon usage data from *M. capricolum* (16). Another feature to note is the low frequency of the dinucleotides CpG in *M. genitalium* non-MgPa proteins and MgPa codons. This is not true, however, for P1 codon usage. The significance of this observation is not clear, but it may serve as an evolutionary landmark for the identification of these two species.

A study conducted by Muto and Osawa (17) demonstrated that codon usage in eubacteria is dictated most strongly by the G+C content of the genome. This was shown by plotting the G+C content of the three codon positions against the G+C content of the genome of several bacteria with G+C contents ranging from 25% to over 70%. Organisms with high G+C contents in their genomes preferentially use G and C containing codons. This was particularly the case in third positions. The frequency of G+C in first, second, and third positions in *M. genitalium* non-MgPa protein codons agrees well with the data from that study (data not shown). When codon informa-

TABLE 3. Codon usage table of *M. genitalium* random clones compared with the MgPa and P1 genes

Codon	No. of codons[a] (% of total codons)	% of total codons		Codon	No. of codons[a] (% of total codons)	% of total codons	
		MgPa	P1			MgPa	P1
TTT-Phe	561 (4.42)	4.23	2.52	TAT-Tyr	299 (2.36)	1.87	0.80
TTC-Phe	77 (0.60)	1.11	1.35	TAC-Tyr	99 (0.78)	0.97	1.66
TTA-Leu	560 (4.42)	3.53	2.10	TAA-End	24 (0.19)	0.07	0.00
TTG-Leu	194 (1.53)	1.39	2.21	TAG-End	7 (0.06)	0.00	0.06
CTT-Leu	231 (1.82)	1.18	0.80	CAT-His	158 (1.25)	0.42	0.18
CTC-Leu	51 (0.40)	1.04	2.76	CAC-His	77 (0.61)	0.69	1.10
CTA-Leu	157 (1.24)	1.52	0.12	CAA-Gln	450 (3.55)	3.33	3.44
CTG-Leu	48 (0.38)	0.35	1.10	CAG-Gln	90 (0.71)	1.32	2.33
ATT-Ile	691 (5.45)	1.87	1.35	AAT-Asn	463 (3.65)	4.02	2.27
ATC-Ile	237 (1.87)	1.94	1.17	AAC-Asn	344 (2.71)	4.99	4.48
ATA-Ile	168 (1.33)	0.55	0.25	AAA-Lys	873 (6.89)	4.30	2.03
ATG-Met	230 (1.81)	1.11	0.80	AAG-Lys	322 (2.54)	3.05	3.13
GTT-Val	472 (3.72)	2.15	1.29	GAT-Asp	567 (4.47)	4.16	2.89
GTC-Val	49 (0.39)	0.55	1.29	GAC-Asp	97 (0.77)	0.97	2.95
GTA-Val	186 (1.47)	1.87	0.74	GAA-Glu	603 (4.76)	2.08	1.54
GTG-Val	110 (0.87)	1.32	2.64	GAG-Glu	162 (1.28)	1.59	1.41
TCT-Ser	155 (1.22)	1.11	0.55	TGT-Cys	105 (0.83)	0.00	0.00
TCC-Ser	56 (0.44)	0.97	2.40	TGC-Cys	34 (0.27)	0.00	0.00
TCA-Ser	192 (1.51)	1.52	0.74	TGA-Trp	61 (0.48)	1.11	1.29
TCG-Ser	24 (0.19)	0.14	1.10	TGG-Trp	37 (0.29)	0.83	0.98
CCT-Pro	206 (1.63)	2.43	1.04	CGT-Arg	108 (0.85)	0.21	0.61
CCC-Pro	58 (0.46)	1.80	2.83	CGC-Arg	48 (0.38)	0.21	1.97
CCA-Pro	143 (1.13)	1.94	1.60	CGA-Arg	15 (0.12)	0.14	0.37
CCG-Pro	12 (0.10)	0.35	1.41	CGG-Arg	13 (0.10)	0.07	0.43
ACT-Thr	305 (2.41)	3.33	1.04	AGT-Ser	255 (2.01)	4.57	3.01
ACC-Thr	127 (1.00)	3.05	4.91	AGC-Ser	73 (0.58)	0.62	1.17
ACA-Thr	193 (1.52)	1.52	0.74	AGA-Arg	223 (1.76)	1.11	0.12
ACG-Thr	18 (0.14)	0.49	2.40	AGG-Arg	73 (0.58)	0.69	0.43
GCT-Ala	370 (2.92)	2.22	2.21	GGT-Gly	353 (2.78)	2.77	2.76
GCC-Ala	51 (0.40)	0.49	2.40	GGC-Gly	90 (0.71)	0.97	2.27
GCA-Ala	318 (2.51)	2.29	0.74	GGA-Gly	174 (1.37)	1.39	0.98
GCG-Ala	35 (0.28)	0.14	2.52	GGG-Gly	98 (0.77)	2.01	2.58

[a] Number of codons found in non-MgPa data base matches, excluding repetitive DNA.

tion from the MgPa and P1 genes were plotted relative to the G+C content of their respective genomes, 32% for *M. genitalium* and 42% for *M. pneumoniae*, we observed that the percentage of G+C in the three codon positions do not fit, or approximate data expected (data not shown).

The observation that the MgPa and P1 genes have G+C contents and codon usage which are very different from *M. genitalium* and other mycoplasmas suggests that these sequences were obtained through a horizontal transfer mechanism. This point is substantiated further and more strongly by the sharp discrepancy between the G+C frequency found in the three codon positions of the MgPa and P1 genes, when plotted against G+C content representative of *M. genitalium* and *M. pneumoniae* genomic DNA. These deviations seen in the MgPa and P1 genes may be what is predicted when a sequence from a genome with a given G+C content is transferred to another genome, with a vastly different A/T mutational pressure.

DISCUSSION

The *M. genitalium* chromosome is the smallest of any free-living organism described to date. This makes it an excellent model for characterizing the minimal requirements for life. Inherent in the success of random genomic sequencing is the assumption that the sequence data bases contain several examples of many different types of genes from a wide range of organisms. Having shown previously that random sequencing is a useful means of identifying putative genes which can serve as markers on the physical map of this genome (20), we have extended this analysis to a much larger scale in order to perform a survey of the contents and coding capacity of this genome.

We expected that the organization of this genome would be quite conservatively arranged, containing a high density of essential genes required for host-independent existence. This does appear to be the case because of the high percentage of ORFs found in randomly selected clones. Additionally it was observed that the arrangement of ORFs in sequences containing more than one ORF was such that there was rarely more than a few nucleotides between the stop codon of one ORF and the methionine of the next. This also suggests that this organism makes heavy use of operon systems, potentially reducing the number of regulatory factors required for controlling transcription of genes. In fact, no potential transcriptional regulatory proteins were found in this study. It is not possible to state whether this absence is meaningful.

Another major class of sequences which were not encountered at expected frequencies in the random sequences were proteins involved with amino acid metabolism. Only one homolog was found, this being the gene for glycine hydroxymethyl transferase. It is interesting that this particular gene function is located in a position which connects major pathways. We speculate that *M. genitalium* maintains some selected genes which confer greater flexibility in utilizing host substrates by simple metabolic conversions. The apparent small number of amino acid metabolism proteins seems to be a real phenomenon since these sequences are in the data base from a large array of eubacteria and might be expected to be identified if they were encountered in this survey. It appears likely that de novo amino acid synthesis is not possible for many if any amino acids in *M. genitalium* cells. The precise details of this issue are difficult to address because of the inability to grow *M. genitalium* in defined medium.

M. genitalium is thought to have a "minimal" genome. In analyzing the deduced amino acid sequences of proteins from this organism, it was expected that sequences would be identified with homologies to proteins that carry out required cellular functions, such as DNA replication, protein synthesis, and transcription. It was surprising to find a reasonably large number of genes involved with intermediary metabolism, since it might be assumed that in most cases the products that are made by these genes could be obtained from the host cell.

One example of such an occurrence is the presence of several genes encoding glycolytic enzymes. It is well known that mycoplasmas are facultative anaerobes. The presence of cytochromes have never been reported in members of the class *Mollicutes*. This being the case, two other means of ATP production for the cell are glycolysis and de novo synthesis by ATP synthetases. We have found evidence for both. It may be pertinent to ask why a minimal genome would maintain an inefficient system for ATP production, especially in light of the fact that proteins in an ATP synthetase pathway were identified in the data base searches. While it is possible that *M. genitalium* could survive without the ability to perform glycolysis, it is reasonable to assume that there is a good reason for maintenance of this gene system.

Another group of metabolic genes for which potential homologs potentially exist in this organism are those involved in hexose conversion and alternate mono- and disaccharide use. By inference it might be assumed that both fructose and galactose can be utilized by *M. genitalium*. This may represent an example of the need to retain some metabolic gene functions to increase the adaptability of the cell to potential raw materials available from the host.

It is with regard to this new information that one must potentially reevaluate what a minimal genome is. A cell with a truly minimal genome would be perfectly parasitic, in that it might preserve functions for DNA replication and cell division, transcription, translation, and DNA maintenance, but would acquire all building blocks from the extracellular milieu. This clearly is not the reality of the *M. genitalium* genome. It is not yet clear what selective pressures caused the genomes of *Mycoplasma* spp. to reduce in size so dramatically. It is also not clear whether further reductions could be tolerated or if they would be strongly selected against. If the latter were the case, we might redefine our idea of a minimal genome to that of the genes currently contained in the *M. genitalium* genome. The answers to these questions can only be addressed when the ability to create targeted deletions or disruption mutations in this organism becomes feasible.

ACKNOWLEDGMENTS

We thank Dennis Pollack, Michael Linhoff, Elizabeth Smith, and Camella Bailey for useful suggestions on the manuscript and C. Bailey for contributing unpublished data on genomic DNA methylation and Christine Doyle for assistance on data base submissions.

Support was provided from a Genetics Curriculum Training Grant, T32GM07092. Additional support was obtained from AI33161 to K.F.B. and grants AI08998 and GM21313 to C.A.H.

REFERENCES

1. **Andachi, Y., F. Yamao, A. Muto, and S. Osawa.** 1991. Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*. J. Mol. Biol. **209:**37–54.
2. **Colman, S. D., P.-C. Hu, and K. F. Bott.** 1990. Prevalence of novel repeat sequences in and around the P1 operon in the genome of *Mycoplasma pneumoniae*. Gene **87:**91–96.
3. **Colman, S. D., P.-C. Hu, W. Litaker, and K. F. Bott.** 1990. A physical map of the *Mycoplasma genitalium* genome. Mol. Microbiol. **4:**683–687.
4. **Dallo, S. F., and J. B. Baseman.** 1991. Adhesion gene of *Mycoplasma genitalium* exists as multiple copies. Microb. Pathog. **10:** 475–480.
5. **Dallo, S. F., A. Chavoya, C.-J. Su, and J. B. Baseman.** 1989. DNA and protein sequence homologies between the adhesins of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. Infect. Immun. **57:**1059–1065.
6. **Davies, C. J., and C. A. Hutchison, III.** 1991. A directed DNA sequencing strategy based upon Tn3 transposon mutagenesis: application to the ADE1 locus on *Saccharomyces cerevisiae* chromosome I. Nucleic Acids Res. **19:**5731–5738.
7. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:**387–395.
8. **Hu, P.-C., R. M. Cole, Y.-S. Huang, J. A. Graham, D. E. Gardner, A. M. Collier, and W. A. Clyde, Jr.** 1982. *Mycoplasma pneumoniae* infection: role of a surface protein in the attachment organelle. Science **216:**313–315.
9. **Hu, P.-C., U. Schaper, A. M. Collier, W. A. Clyde, M. Horikawa, Y.-S. Huang, and M. F. Barile.** 1987. A *Mycoplasma genitalium* protein resembling the *Mycoplasma pneumoniae* attachment protein. Infect. Immun. **55:**1126–1131.
10. **Hutchison, C. A., III, R. Swanstrom, and D. D. Loeb.** 1991. Mutagenesis of protein coding domains. Methods Enzymol. **202:** 356–390.
11. **Inamine, J. M., K.-C. Ho, S. Loechel, and P.-C. Hu.** 1990. Evidence that UGA is read as tryptophan rather than as a stop codon by *Mycoplasma pneumoniae*, *Mycoplasma genitalium*, and *Mycoplasma gallisepticum*. J. Bacteriol. **172:**504–506.
12. **Inamine, J. M., S. Loechel, A. M. Collier, F. M. Barile, and P.-C. Hu.** 1989. Nucleotide sequence of the MgPa (*mgp*) operon of *Mycoplasma genitalium* and comparison to the P1 (*mpp*) operon of *Mycoplasma pneumoniae*. Gene **82:**259–267.
13. **Kawauchi, Y., A. Muto, and S. Osawa.** 1982. The protein composition of *Mycoplasma capricolum*. Mol. Gen. Genet. **188:**7–11.
14. **Krause, D. C., and K. K. Lee.** 1991. Juxtaposition of the genes encoding *Mycoplasma pneumoniae* cytadherence-accessory proteins HMW 1 and HMW 3. Gene **107:**83–89.
15. **Krawiec, S., and M. Riley.** 1990. Organization of the bacterial chromosome. Microbiol. Rev. **54:**502–539.
16. **Muto, A.** 1987. The genome structure of *Mycoplasma capricolum*. Isr. J. Med. Sci. **23:**334–341.
17. **Muto, A., and S. Osawa.** 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA **84:**166–169.
18. **Nur, I., M. Szyf, A. Razin, G. Glasser, S. Rottem, and S. Razin.** 1985. Eukaryotic and prokaryotic traits of DNA methylation in spiroplasmas (mycoplasmas). J. Bacteriol. **164:**19–24.
19. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
20. **Peterson, S. N., N. Schramm, P.-C. Hu, K. F. Bott, and C. A. Hutchison, III.** 1991. A random sequencing approach for placing markers on the physical map of *Mycoplasma genitalium*. Nucleic Acids Res. **19:**6027–6031.
21. **Razin, S.** 1985. Molecular biology and genetics of mycoplasmas (*Mollicutes*). Microbiol. Rev. **49:**419–455.
22. **Razin, S., and E. Jacobs.** 1992. Mycoplasma adhesion. J. Gen. Microbiol. **138:**407–422.
23. **Rogers, M. J., J. Simmons, R. T. Walker, W. G. Weisburg, C. R. Woese, R. S. Tanner, I. M. Robinson, D. A. Stahl, G. Olsen, R. H. Leach, and J. Maniloff.** 1985. Construction of the mycoplasma evolutionary tree from 5S rRNA sequence data. Proc. Natl. Acad. Sci. USA **82:**1160–1164.
24. **Ruland, K., R. Wenzel, and R. Herrmann.** 1990. Analysis of three

different repeated DNA elements present in the P1 operon of *Mycoplasma pneumoniae*: size, number and distribution on the genome. Nucleic Acids Res. **18:**6311–6317.

25. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
26. **Staden, R.** 1982. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10:**4731–4751.
27. **Su, C. J., and J. B. Baseman.** 1990. Genome size of *Mycoplasma genitalium*. J. Bacteriol. **172:**4705–4707.
28. **Tanaka, R., Y. Andachi, and A. Muto.** 1991. Evolution of tRNAs and tRNA genes in *Acholeplasma laidlawii*. Nucleic Acids Res. **19:**6787–6792.
29. **Tully, J. G., D. Taylor-Robinson, D. L. Rose, R. M. Cole, and J. M. Bove.** 1983. *Mycoplasma genitalium*, a new species from the human urogenital tract. Int. J. Syst. Bacteriol. **33:**387–396.
30. **Weisburg, W. G., J. G. Tully, D. L. Rose, J. P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T. G. Lawrence, J. Van Etten, J. Maniloff, and C. R. Woese.** 1989. A phylogenetic analysis of the mycoplasmas: basis for their classification. J. Bacteriol. **171:**6455–6467.
31. **Wittman, H. G.** 1982. Components of bacterial ribosomes. Annu. Rev. Biochem. **51:**155–183.
32. **Yamao, F., A. Muto, Y. Kawauchi, M. Iwami, S. Iwagami, Y. Azumi, and S. Osawa.** 1985. UGA is read as tryptophan in *Mycoplasma capricolum*. Proc. Natl. Acad. Sci. USA **82:**2306–2309.
33. **Yogev, D., and S. Razin.** 1986. Common deoxyribonucleic acid sequences in *Mycoplasma genitalium* and *Mycoplasma pneumoniae* genomes. Int. J. Syst. Bacteriol. **36:**426–430.