Molecular Microbiology (1995) **16**(5), 955–967

Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology

Peer Bork,[1,2*] Christos Ouzounis,[1†] Georg Casari,[1] Reinhard Schneider,[1] Chris Sander,[1] Maureen Dolan,[3] Walter Gilbert[4] and Pat M. Gillevet[5]

[1]*EMBL, Meyerhofstrasse 1, D-69012 Heidelberg, Germany.*

[2]*Max-Delbrück-Centre for Molecular Medicine, Berlin-Buch D-13122, Germany.*

[3]*Massachusetts General Hospital, Boston, Massachusetts 02114, USA.*

[4]*Harvard University, Boston, Massachusetts 02138, USA.*

[5]*George Mason University, Fairfax, Virginia 22030-4444, USA.*

Summary

We report on the analysis of 214 kb of the parasitic eubacterium *Mycoplasma capricolum* sequenced by genomic walking techniques. The 287 putative proteins detected to date represent about half of the estimated total number of 500 predicted for this organism. A large fraction of these (75%) can be assigned a likely function as a result of similarity searches. Several important features of the functional organization of this small genome are already apparent. Among these are (i) the expected relatively large number of enzymes involved in metabolic transport and activation, for efficient use of host cell nutrients; (ii) the presence of anabolic enzymes; (iii) the unexpected diversity of enzymes involved in DNA replication and repair; and (iv) a sizeable number of orthologues (82 so far) in *Escherichia coli*. This survey is beginning to provide a detailed view of how *M. capricolum* manages to maintain essential cellular processes with a genome much smaller than that of its bacterial relatives.

Introduction

Mycoplasmas are a diverse group of Gram-positive bacteria with a characteristically low GC content (Herrmann, 1992) that lack a cell wall. Their genomes vary between about 600 kb and 1800 kb (Wenzel *et al.*, 1992) and are the smallest genomes of living cells (Razin, 1992). As deduced from sequence comparison of 16S rRNA, *Bacillus subtilis* is a well-characterized close relative (Woese, 1987; Weisburg *et al.*, 1989) but has a genome size of about 4.2 Mb, about four or five times larger than that of the mycoplasmas. The main driving force for the condensation of its genome appears to be the parasitic life of mycoplasmas; the use of metabolic products of the host appears to allow a drastic reduction in the number or components of biosynthetic pathways. Although considerable amounts of biochemical data provide an insight into this highly specialized metabolism (for a comprehensive review see Maniloff *et al.*, 1992), direct analysis of the genome is a powerful approach to complement existing knowledge. Large-scale genomic sequencing of *Mycoplasma* species has already successfully begun for *Mycoplasma genitalium* (Peterson *et al.*, 1993; Nowak, 1995), *Mycoplasma pneumoniae* (R. Herrmann, personal communication) and *Mycoplasma capricolum* (Gillevet, 1993). Here we report the initial analysis of 214 kb from *M. capricolum*. Although this is a preliminary data set consisting of multiple short contigs, the analysis has already yielded many insights into the physiology of this small parasitic bacterium.

Received 30 August, 1994; revised 23 December, 1994; accepted 14 February, 1995. †Present address: AI Center, SRI International, Menlo Park, California 94025-3493, USA. *For correspondence. Tel. (6221) 387534; Fax (6221) 387517.

Results and Discussion

Data and data quality

One-pass fluorescent sequencing was used to produce 1505 random clones as starting points for genomic walking techniques (Gillevet, 1993). At the current stage, more than 870 477 bases have been sequenced from *M. capricolum*, (California kid) strain ATCC 27343 and finally assembled into 372 non-overlapping contigs covering 214 528 bp. The length of the contigs varies from 63 to 2049 bases with an average length of 658. The data consist of 13 091 bases (6%) from one-pass fluorescent sequencing that cover 1198 of the 1505 random clones and 201 437 bases (94%) produced by genomic walking. Of the latter, 45 632 bases (23%) are single-stranded regions at the end of the growing contig, 155 805 bases (77%) have multiple coverage on at least one strand while 112 621 bases (56%) are covered on both strands. There is a total of 968 ambiguous calls (nucleotide assignments) in the data set (~0.5%).

We have compared 8868 bases of our data with *M. capricolum* sequences stored in public databases and

note less than 0.7% difference between the two data sets (these include ambiguous calls, insertions, deletions and mismatches). It should be noted that the strand coverage of this latter data set is essentially identical to that of the entire data set. Many of the discrepancies noted between the published data and our data set are in regions where we have multiple coverage on both strands and, therefore, they may be attributed to strain differences. As a result, the true error rate of the walking data may be much lower than 0.7%. Furthermore, we have identified only 97 frameshifts and aberrant stop codons in the 103 000 bases contained in the open reading frames (ORFs) that we have analysed, which indicates that the error rate of this type is probably less than 10^{-3} (details of the sequencing methodology will be published elsewhere; P. M. Gillevet *et al.*, in preparation).

Identification of proteins by similarity searches

Similarity searches are insensitive to error rates at least up to 1% (States and Botstein, 1992) and can even be used for detection of sequencing errors (Posfai and Roberts, 1992). Therefore, a preliminary data set consisting of multiple short contigs with appreciable sequencing errors can still yield many insights into the functionality of putative proteins, as has the one-pass fluorescent data published for *M. genitalium* (Peterson *et al.*, 1993). After correction of frameshifts (see the *Experimental procedures*) and masking compositionally biased regions (low-entropy regions, Claverie and States, 1993), relatively stringent cut-offs were applied in similarity searches (BLAST Altschul *et al.*, 1990): $p \leqslant 10^{-9}$; FASTA (Pearson and Lipman, 1988): opt score $\geqslant 150$) in order to identify matches with proteins that are clearly related to the putative gene products of *M. capricolum*. We have identified 176 distinct regions within the 372 contigs that fulfilled these criteria; their translated primary structures have obvious homologues. Among the 176 regions that match database proteins, several pairs were identified that correspond to non-overlapping segments of a single database protein. Although it is somewhat difficult to judge whether multiple hits (e.g. in the N-terminal and C-terminal portions of a database protein) characterize only a single gene product or different mycoplasma proteins of a multigene family, 22 of these pairs were considered to be from the same protein and were merged and considered in the statistics as a single entry (resulting in 154 unique protein hits, Table 1). In addition to this first fully automatic procedure, various sequence analysis tools (Bork *et al.*, 1992; Koonin *et al.*, 1994) were used to verify more subtle relationships, which revealed another 61 regions of similarity. The 215 distinct proteins thus identified (Tables 1 and 2) represent the minimum number of proteins encoded by the ORFs in the 372 contigs. Not a single gene was found that overlaps significantly with, or is contained in, another gene. There are 72 ORFs longer than 300 bases that do not match any other sequence in the databases and do not overlap with known genes. Most of these might code for additional proteins (Table 1) with as yet unidentified function.

Table 1. Statistics and calculation of gene density in *M. capricolum*.

Contigs	Characteristics	Size (kb)
202	Contain at least one protein and maybe RNA genes	
+43	Contain DNA repeats	
+10	Contain at least one RNA gene	
+117	Have no special characteristics — no sequence similarity	
=372	Total number of contigs	214
Proteins		
154	Have easily identifiable homologues	
+61	Have distant homologues	
=215	Total number of identified proteins	≈103
+72	Other, non-overlapping ORFS >100 aa	+≈34
=287	Estimated number of ORFs	≈137
Capacity		
64%	Coding density (137 kb/214 kb)	
Function		
75%	Function prediction rate (215/287)	

At the DNA level, numerous matches were found (Table 2) with tRNA, rRNA and, surprisingly, snRNA-like sequences (Ushida and Muto, 1993). Also, the six conceptual translations of 43 contigs do not match any protein in the database but do match another *M. capricolum* contig (Table 2). Most of these cases appear to be DNA repeats, a striking feature of mycoplasma genomes. A relatively high number of repetitive non-coding regions has also been reported for *M. pneumoniae* (Wenzel and Herrmann, 1988; Ruland *et al.*, 1990) and *M. genitalium* (Peterson *et al.*, 1993). In spite of the large number of contigs characterized by sequence similarity with DNA repeats, RNA genes or proteins, 117 contigs still remain dissimilar to any sequence in public databases and, therefore, have no special characteristics, i.e. they do not have any ORFs longer than 300 bp or other obvious structural features. These contigs comprise ~77 kb of sequence data.

Coding density

The regions with similarity to database proteins or fragments cover 103 kb, which is a lower limit for estimating the coding density. For a more realistic estimate, the additional 72 ORFs (larger than 100 amino acids) with no database similarity were taken into account (an additional 34 kb). Some of these putative ORFs, though, may not code for proteins. This possible overestimate is compensated by the fact that shorter ORFs may have escaped identification in the remaining 77 kb with no obvious features. Therefore, the coding density can be roughly

estimated as (103 kb coding + 34 kb putative coding)/214 kb, which is about 64% (Table 1). Of course, longer contigs are desirable for such an estimate, but the calculation is supported by the low coding density of *M. genitalium*, estimated at 55% (based on a much smaller data set; Peterson *et al.*, 1993). It appears that the protein coding density of mycoplasmas is much lower than the 82% of *Escherichia coli* (Daniels *et al.*, 1992; Blattner *et al.*, 1993) and, surprisingly, even lower than the approximately 72% coding density in yeast (Fickett and Guigo, 1993; Dujon *et al.*, 1994). Therefore, reduction in genome size does not necessarily lead to an increase in coding density. Instead, *M. capricolum* has apparently retained or accumulated non-coding DNA, the function of which remains to be elucidated.

Genome size and its coverage

Almost 30% of published rare restriction sites are present in the random sample of 214 kb (Table 3). This would seem to support earlier estimates of the genome size based on two-dimensional denaturating gradient gel electrophoresis, i.e. 724 kb (Poddor and Maniloff, 1989). Calculations based on pulsed-field gel electrophoresis (PFGE) (Whitely *et al.*, 1991) predict a genome size of 1070 kb, a possible overestimate caused by the aberrant movement of AT-rich restriction fragments in pulsed-field gels (Poddor and Maniloff, 1989; Maniloff, 1989). This explanation has been rebutted, however (Robertson *et al.*, 1990; Neimark and Lange, 1990), and a final conclusion can only be drawn when the entire genome has been sequenced.

If the estimate of 500 proteins in *M. capricolum* (Muto *et al.*, 1987) is correct (which would correlate with the smaller genome size estimate), then the approximately 287 proteins identified in this work represent about half of the total number of proteins in this organism. Yet, this fraction (287 proteins in 214 kb) cannot be extrapolated to the whole genome, as most ORFs analysed here contain only partial sequences, i.e. future contigs will often contain sequences of genes already identified.

Prediction of function by similarity

The 215 proteins identified as a result of similarity searches reflect a high success rate in predicting function by sequence analysis in randomly sequenced genomes. Considering the additional 72 putative protein-coding ORFs larger than 100 amino acids, the percentage of proteins with homologues in public databases (215/287) is about 75% of those sequenced (Table 1). This level is much higher than in other large-scale sequencing projects, where the percentage varies between 33% (Wilson *et al.*, 1994) and 61% (Koonin *et al.*, 1994), with an average of about 50% throughout all phyla (Blattner *et al.*, 1993; Adams *et al.*, 1993; Honoré *et al.*, 1993; Glaser *et al.*, 1993; Bork *et al.*, 1994). This high percentage is probably the result of the reduction of a parasitic genome to the genes coding for essential enzymes, which are usually well characterized and therefore over-represented in current databases, as well as the extensive characterization of other prokaryotes such as *E. coli*.

Finding homologous proteins in databases does not necessarily imply a precise functional prediction (Bork *et al.*, 1994). Yet, for most of the proteins shown in Table 2, the best matching protein or protein family permits transfer of at least some functional information. This, in turn, enables us to draw conclusions about metabolic pathways and to obtain some insight into how *M. capricolum* manages to maintain cellular processes with such a small genome.

From ORFs to physiology

As long as the complete genome sequence is unavailable, only the presence of a certain protein but not its absence is conclusive. The enzymes already identified are useful markers of biochemical pathways of *M. capricolum* and closely related species.

Intermediary metabolism. As predicted by biochemical methods, glycolysis, as one of the major catabolic pathways, is certainly present in *M. capricolum*. Partial ORFs similar to nearly all glycolytic enzymes were found as well as components of the pyruvate dehydrogenase complex (Table 2). However, only one protein, transketolase, of the pentose phosphate cycle was identified. Together with the fact that no enzymes of the pentose phosphate cycle have been found yet in *M. genitalium* (Peterson *et al.*, 1993), the intactness of this important pathway remains to be verified. The absence so far of enzymes of the citric acid cycle tentatively confirms that mycoplasmas take other routes for energy storage. A major enzyme, ornithine carbamoyltransferase, of the arginine hydrolase pathway was identified and may be utilized for ATP production, although other roles of carbamoylphosphates are possible (Pollak, 1992). Energy metabolism is often associated with the production of reduced NADH (e.g. in glycolysis). NADH might be further processed by flavoproteins and an NADH oxidase (Pollak, 1992). This is an alternative route to the respiratory chain involving cytochromes transferring released protons to oxygen (Pollak, 1992). Indeed, we found at least one protein similar to NADH oxidase (encoded by an ORF on the contig with EMBL accession number Z33133, Table 2). This alternative pathway might be coupled to ATP synthesis. As in *M. genitalium* (Peterson *et al.*, 1993), several components of a membrane-associated ATP synthase were also identified.

Table 2. Sequence similarities of *M. capricolum* contigs with database entries.

Contig	Bases	Best hit	P-value	Frame	Function
Metabolite transport and activation					
Z33018	1201	LKTB_ACTAC	1.5×10^{-22}	+2	ABC transporter
Z33019_2	1444	PT2S_STRMU	6.1×10^{-11}	+1	Phosphotransferase EII (sucrose)
Z33047	855	PT2X_ECOLI	1.7×10^{-9}	+1*	Phosphotransferase EII (maltose)
Z33074_1	1623	PT2G_ECOLI	9.3×10^{-31}	+3	Phosphotransferase EII (glucose)
Z33097	808	ATMB_SALTY	9.6×10^{-30}	+1*	Mg-transporting ATPase, P-type
Z33098_2	848	GYLA_STRCO	4.1×10^{-20}	+3	Glycerol facilitator protein
Z33100	803	PT2G_BACSU	1.9×10^{-15}	+3	Phosphotransferase EII (glucose)
Z33102	1006	ATMB_SALTY	6.9×10^{-33}	−1*	Mg-transporting ATPase
Z33105	1174	ARAG_ECOLI	3.3×10^{-22}	−3*	Arabinose-transport protein
Z33112	271	PT2G_BACSU	6.5×10^{-7}	−3	Phosphotransferase EII (glucose)
Z33141	284	PT2G_BACSU	3.0×10^{-10}	+1	Phosphotransferase EII (glucose)
Z33144	620	GLNQ_BACST	2.0×10^{-19}	+3	ABC transporter (glutamine)
Z33157	495	HMT1_SCHPO	7.9×10^{-27}	−1*	ABC transporter (MDR subfamily)
Z33174	203	P29_MYCHR	1.4×10^{-13}	+1*	ABC transporter
Z33178	579	MDR_PLAFF	4.8×10^{-48}	−3	ABC transporter (MDR subfamily)
Z33179	586	BRAF_PSEAE	1.8×10^{-9}	−1	ABC transporter (amino acids)
Z33187	625	PT3M_ENTFA	1.7×10^{-17}	+3*	Phosphotransferase EIII (mannitol)
Z33210	408	PT2N_ECOLI	1.1×10^{-6}	−2	Phosphotransferase EII (NAG)
Z33222	589	AMIE_STRPN	6.1×10^{-8}	−3	ABC transporter (oligopeptide)
Z33224	353	PT2G_BACSU	7.4×10^{-4}	+2	Similar to phosphotransferase EII
Z33225	188	GLPF_BACSU	7.0×10^{-2}	+3	Glycerol facilitator protein
Z33229	422	OPPD_BACSU	1.1×10^{-21}	−3	ABC transporter (oligopeptide)
Z33249	265	CYSA_SYNP7	2.1×10^{-7}	+2	ABC transporter (sulphate)
Z33251	406	ATC2_YEAST	2.1×10^{-15}	+3	Ca^{2+}-transporting ATPase
Z33253	208	PT2M_ECOLI	4.9×10^{-7}	−1	Phosphotranferase EII (mannitol)
Z33266	580	ATC1_YEAST	1.2×10^{-39}	−1*	Ca^{2+}-transporting ATPases
Z33293	242	PT2G_BACSU	9.3×10^{-13}	+2	Phosphotransferase EII (glucose)
Z33307	183	AROP_ECOLI	1.6×10^{-4}	−3	Aromatic transport protein (permease)
Z33314	283	GLPF_BACSU	2.3×10^{-19}	+1*	Glycerol facilitator protein
Z33339	539	CYSA_SYNP7	1.4×10^{-23}	−2	ABC transporter (sulphate)
Z33340	353	PT2S_VIBAL	5.2×10^{-12}	−1	Phosphotransferase EII (sucrose)
Z33348_2	1708	YHBI_ECOLI	1.7×10^{-16}	+2	Similar to phosphotransferase EII
Z33348_3	1708	PT2F_ECOLI	1.6×10^{-10}	+2	Phosphotransferase EII (fructose)
Z33353	829	PT2G_BACSU	2.5×10^{-34}	+3*	Phosphotransferase EII (glucose)
Z33361	1191	YBBA_ECOLI	7.9×10^{-23}	−2	ABC transporter (MDR subfamily)
Z33363	463	PT3M_ECOLI	3.5×10^{-6}	+2	Similar to phosphotransferase EIII
Amino acid metabolism					
Z33027	1144	AQPETBDCP_3	9.5×10^{-5}	+1	Similar to serine protease RE
Z33032_1	1250	DAPE_ECOLI	1.2×10^{-9}	−2*	SDAP desuccinylase
Z33281	179	METX_ECOLI	2.0×10^{-25}	−2	S-adenosyl-Met synthetase 2
Z33292	295	LON_ECOLI	1.5×10^{-32}	−1	ATP-binding endopeptidase La
Z33299	451	METK_HUMAN	3.9×10^{-27}	−3	S-adenosyl-Met synthetase (alpha)
Z33311	157	OTCC_NEICI	3.0×10^{-15}	+1	Ornithine carbamoyltransferase
Z33349	1106	OTC_ASPNG	3.2×10^{-38}	+3	Ornithine carbamoyltransferase
Z33357	922	LON_ECOLI	3.6×10^{-74}	+3	ATP-binding endopeptidase LA
Nucleotide metabolism					
Z33022	1759	PYRG_ECOLI	3.1×10^{-123}	+2*	CTP synthase
Z33033	566	YC42_CAEEL	6.6×10^{-9}	−3	Similar to dCTP deaminases
Z33044	999	C1TM_YEAST	7.8×10^{-40}	+2*	C1-THF synthase
Z33066	900	DEOK_HUMAN	3.8×10^{-4}	+1*	Pyrimidine kinase
Z33079	971	KTHY_HUMAN	5.2×10^{-5}	+3	Pyrimidine kinase
Z33170	533	RNC_ECOLI	1.5×10^{-16}	+3	Ribonuclease III
Z33218	124	KPRS_ECOLI	2.7×10^{-6}	+1	Phosphoribosyl-PP-kinase
Z33263	127	UPP_ECOLI	2.3×10^{-9}	+2	Uracil-phosphoribosyltransferase
Z33279	297	KITH_AMEPV	2.1×10^{-6}	+3*	Thymidine kinase
Z33336	363	PURA_HUMAN	3.8×10^{-33}	−3	Adenylosuccinate synthetase
Kad_Mycca					Adenylate kinase
Lipid metabolism					
Z33013_1	1904	PGSA_ECOLI	6.2×10^{-10}	+3	Phosphatidylglycerol-P synthase
Z33059_2	1484	TPES_PSEPU	1.2×10^{-4}	−2	Similar to lipase/esterase family
Z33081	1192	GLPE_ECOLI	8.9×10^{-4}	+2	Similar to GlpE and PspE proteins
Z33287	299	ODB2_HUMAN	6.7×10^{-5}	−1	Lipoamide acyltransferase (E2)
Z33350	736	ODO2_BACSU	4.9×10^{-16}	−3	Dihydrolipoamide succinyltransferase (E2)

Table 2. Continued

Contig	Bases	Best hit	P-value	Frame	Function
Carbohydrate metabolism/energy storage					
Z33030_1	1748	OUTB_BACSU	9.8×10^{-18}	−1	ATP pyrophosphatase[a]
Z33051	710	ATPB_PECFR	2.5×10^{-44}	+1*	ATP synthase (beta)
Z33053_1	627	THIO_BACSU	1.6×10^{-19}	+3	Thioredoxin
Z33055	1061	NAGB_ECOLI	3.0×10^{-30}	−2	Glucosamine-6-isomerase
Z33072_1	2057	K1PF_ECOLI	6.0×10^{-11}	−2	1-Phosphofructokinase
Z33074_2	1623	YICI_ECOLI	8.0×10^{-14}	+1	Similar to isomaltase-sucrases
Z33075	225	ALF2_RHOSH	6.4×10^{-6}	−1*	Fructose-bisphosphate aldolase
Z33087_2	859	PMGY_MAIZE	3.8×10^{-17}	+3	Phosphoglyceromutase
Z33089	710	NADO_THEBR	1.3×10^{-4}	−1	Similar to NADH oxidase
Z33096	875	DHAC_RAT	4.7×10^{-26}	+3	Aldehyde dehydrogenase
Z33098_1	848	GLPK_ECOLI	6.8×10^{-20}	+1*	Glycerol kinase
Z33104	1036	O16G_BACCE	3.0×10^{-79}	+3*	Cytoplasmic oligo-1,6-glucosidase
Z33109	1300	TSR_BACSU	1.5×10^{-24}	−1	RNA synthesis protein/fructose-Bi-P-aldolase
Z33110	545	ODPB_BACSU	2.0×10^{-57}	+3	Pyruvate DH E1 (beta)
Z33126	534	MGLC_ECOLI	4.3×10^{-7}	+2	Membrane forming protein
Z33133	153	S26965P	8.8×10^{-13}	−1	Similar to *Streptococcus faecalis* NADH oxidase
Z33226	486	YLP2_PSEPU	1.1×10^{-17}	−3	Similar to dihydrolipoamide DH
Z33228	372	YINL_LISMO	1.7×10^{-13}	+2*	Similar to 7-alpha-hydroxysteroid DH
Z33230	121	TKT_RHOSH	1.2×10^{-11}	+1	Transketolase
Z33232	267	ENO_ARATH	5.4×10^{-22}	−2*	Enolase
Z33234	178	YIEK_ECOLI	2.8×10^{-7}	−3	Similar to glucosamine isomerase
Z33235_1	681	ATPD_SYNY3	1.1×10^{-6}	+1*	ATP synthase (delta)
Z33235_2	681	ATPA_BACME	2.0×10^{-22}	+1	ATP synthase (alpha)
Z33250	427	PT1_STACA	4.8×10^{-21}	−1*	PEP P-transferase
Z33255	187	ENO_ECOLI	1.4×10^{-20}	+1	Enolase
Z33265	152	PT1_STACA	1.9×10^{-5}	+3	PEP P-transferase
Z33277	420	XYLK_STAXY	2.6×10^{-4}	−1	Sugar kinase (hexokinase family)
Z33286	510	IPYR_THEP3	1.9×10^{-28}	+2	Inorganic pyrophosphatase
Z33288	291	PGKY_WHEAT	1.2×10^{-35}	+1	Cytoplasmic phosphoglycerate kinase
Z33313	166	ODP2_BACSU	1.9×10^{-15}	−2	Dihydrolipoamide acetyltransferase (E2)
Z33330	319	ATPA_ECOLI	2.4×10^{-12}	−3	ATP synthase (alpha)
Z33331	176	K6PF_BACST	1.6×10^{-8}	+3	6-Phosphofructokinase
Z33348_1	1708	K1PF_RHOCA	1.5×10^{-9}	+3	1-Phosphofructokinase
Z33368	2167	MGLA_ECOLI	6.4×10^{-15}	−1	Galactoside-binding protein
Z33370	1329	ATPB_THEP3	9.6×10^{-143}	−1*	ATP synthase (beta)
Other metabolic enzymes					
Z33006	897	YIDA_ECOLI	1.3×10^{-16}	+3	Hydrolase (HAD family[b])
Z33015	1156	BAIC_EUBSP	3.8×10^{-14}	+2	Similar to trimethylamine DH
Z33083	567	AMID_RHOER	4.6×10^{-9}	−3	Amidase
Z33087_1	859	YIDA_ECOLI	5.2×10^{-10}	+1	hydrolase (HAD family[2])
Z33177	214	UGPQ_ECOLI	7.8×10^{-3}	−2	Similar to phosphodiesterase
Z33272	283	GSHR_HUMAN	2.4×10^{-4}	+2	FAD/NAD-binding reductase
Z33273	257	PCR_AVESA	8.4×10^{-3}	−1	NAD-binding oxidoreductase
Z33290	236	YHDG_ECOLI	1.2×10^{-21}	−3	Similar to DH subfamily (BAIC_EUBSP)
Z33346	778	FOLC_ECOLI	8.2×10^{-5}	−1*	Similar to folylpoly-Glu synthetase
Z33362_2	1259	YHDG_ECOLI	5.5×10^{-7}	+2	Similar to DH subfamily (BAIC_EUBSP)
DNA replication, repair and recombination					
Z33035	2049	GYRA_BACSU	1.7×10^{-54}	−2*	DNA gyrase (alpha)
Z33054	580	A30868P	2.9×10^{-9}	+2	Transposase
Z33057	752	DNAA_MYCCA	1.6×10^{-119}	+3	DnaA protein
Z33071	1102	GYRA_STAAU	1.8×10^{-13}	−3	DNA gyrase (alpha)
Z33091	594	DP3A_ECOLI	3.6×10^{-30}	+1	DNA polymerase III (alpha)
Z33108	752	GYRB_BACSU	2.3×10^{-35}	−2*	DNA gyrase (beta)
Z33173	307	TOP1_SYNP7	1.5×10^{-2}	−1*	Topoisomerase I
Z33193	234	DPO1_STRPN	1.0×10^{-12}	−1	DNA polymerase I
Z33201	495	DP3A_BACSU	5.1×10^{-5}	−1#	DNA polymerase III (alpha)
Z33239	367	RNH2_ECOLI	7.9×10^{-14}	−3	Ribonuclease HII
Z33252	286	DP3A_BACSU	1.0×10^{-8}	−1	DNA polymerase III (alpha)
Z33256	54	GYRB_ECOLI	1.9×10^{-1}	+3	DNA gyrase (beta)
Z33302	129	DNLJ_ECOLI	3.5×10^{-8}	−3	DNA ligase
Z33305	417	GYRA_BACSU	1.2×10^{-39}	−1	DNA gyrase (alpha)
Z33335	385	DP3A_BACSU	2.5×10^{-30}	+3*	DNA polymerase III (alpha)
Dp3b_Mycca					DNA polymerase III (beta)
Z33060_1	1308	APN1_YEAST	7.9×10^{-3}	−1	Similar to AP-endonucleases

Table 2. Continued

Contig	Bases	Best hit	P-value	Frame	Function
Z33120	621	UVRA_MICLU	9.1×10^{-65}	+1	Excinuclease ABC (A)
Z33128	483	RECM_BACSU	1.3×10^{-11}	+1	DNA repair protein RecM
Z33191	510	UVRC_BACSU	9.9×10^{-7}	+2	Excinuclease ABC (C)
Z33233	585	MTS1_STRSA	1.9×10^{-9}	−1*	Modification methylase
Z33254	340	APN1_YEAST	2.4×10^{-4}	+1	Similar to AP-endonucleases
Z33278	208	S18707P	6.2×10^{-5}	−1	UvrD protein (also in *M. genitalium*)
Z33332	263	JQ0894P	1.4×10^{-13}	+2*	P115 protein, RecF/RecN family
Z33345	1065	UVRB_ECOLI	3.9×10^{-127}	+3	Excinuclease ABC (B)
Z33355	1401	YAT3_RHORU	1.1×10^{-28}	+2	RecF/RecN family
Z33373	299	UVRA_MICLU	6.5×10^{-31}	+3	Excinuclease ABC (A)
Cell division					
Z33086	245	FTSH_ECOLI	3.1×10^{-24}	+2	Cell division protein FTSH
Z33209	186	FTSH_ECOLI	1.0×10^{-5}	+2	Cell division protein FTSH
Z33322	149	FTSZ_BACSU	6.5×10^{-18}	−2	Cell division protein FTSZ
Transcription factors					
Z33019_1	1444	LACR_STAAU	2.7×10^{-24}	+2	Sugar repressor (HTH family)
Z33041	1410	RPOB_PSEPU	1.2×10^{-20}	+2*	DNA-directed RNA polymerase (beta)
Z33050_2	1490	RPOA_BACSU	7.7×10^{-57}	−1	DNA-directed RNA Polymerase (alpha)
Z33052	880	PILB_NEIGO	1.3×10^{-43}	+3	Transcription repressor
Z33063	448	PHNF_ECOLI	7.4×10^{-2}	−2	HTH motif (HTH family)
Z33072_2	2057	LACR_LACLA	2.0×10^{-15}	−2	lac repressor
Z33085	1500	RPOB_ECOLI	3.3×10^{-81}	−3*	DNA-directed RNA polymerase (beta)
Z33123	325	HMGT_ONCMY	9.6×10^{-4}	+3	Similar to HMG proteins
Z33203	290	GREA_ECOLI	3.6×10^{-13}	+1*	Transcription elongation factor
Z33275	209	XYLR_BACSU	1.6×10^{-1}	+2	Similar to xylose repressor (HTH family)
Z33297	160	RPOC_MYCLE	4.1×10^{-19}	−1	DNA-directed RNA polymerase (beta)
Z33338	278	RPOD_NOSCO	8.8×10^{-9}	−3	DNA-directed RNA polymerase (delta)
Translation, protein biosynthesis and ribosomal proteins					
Z33016_1	1476	EFTS_ECOLI	8.5×10^{-27}	−3	Elongation factor EF-TS
Z33034_2	840	IF1_BACSU	5.0×10^{-35}	−2	Initiation factor IF-1
Z33068	852	EFG_SPIPL	1.7×10^{-95}	+1*	Elongation factor G
Z33076_2	1299	SPOU_ECOLI	9.0×10^{-10}	−2	rRNA methylase
Z33271	377	FMT_ECOLI	1.1×10^{-13}	+2*	Met-tRNA formyltransferase
Z33284	487	RF1_ECOLI	8.8×10^{-18}	−2*	Peptide chain release factor 1
Z33303	131	EFTU_ECOLI	1.4×10^{-12}	+3*	Elongation factor TU
Z33352_1	1629	IF3_BACST	3.7×10^{-32}	+3	Initiation factor IF-3
Z33371	1151	IF2_BACST	2.2×10^{-87}	−2*	Initiation factor IF-2
Z33008	878	SYL_ECOLI	4.3×10^{-32}	+3	Leu-tRNA synthetase
Z33021	865	SYL_ECOLI	7.1×10^{-14}	+3	Leu-tRNA synthetase
Z33026	678	SYN_ECOLI	2.5×10^{-57}	+1	Asn-tRNA synthetase
Z33034_3	840	AMPM_BACSU	5.8×10^{-43}	−3	M-aminopeptidase
Z33042	1582	SYA_ECOLI	3.7×10^{-31}	−3	Ala-tRNA synthetase
Z33048	1474	SYD_ECOLI	7.4×10^{-63}	+1	Asp-tRNA synthetase
Z33053_2	627	SYK1_ECOLI	2.8×10^{-17}	−3	Lys-tRNA synthetase
Z33056	1031	SYT1_BACSU	7.3×10^{-41}	−2	Thr-tRNA synthetase
Z33122	790	SYV_BACST	4.3×10^{-38}	−3	Val-tRNA synthetase
Z33127	298	SYI_METTH	1.6×10^{-10}	+1*	Ile-tRNA synthetase
Z33137_1	700	SYH_STREQ	3.6×10^{-14}	+1	His-tRNA synthetase
Z33137_2	700	SYD_ECOLI	1.8×10^{-13}	+3	Asp-tRNA synthetase
Z33153	735	SYFA_BACSU	3.8×10^{-53}	+1*	Phe-tRNA synthetase
Z33154	63	SYE_BACST	4.7×10^{-1}	+1	Glu-tRNA synthetase
Z33244	306	SYL_ECOLI	5.6×10^{-11}	+3	Leu-tRNA synthetase
Z33261	192	SYE_BACST	7.2×10^{-10}	−3*	Glu-tRNA synthetase
Z33276	401	SYY_BACCA	3.5×10^{-28}	+3	Tyr-tRNA synthetase
Z33289	622	SYI_ECOLI	8.5×10^{-33}	+3*	Ile-tRNA synthetase
Z33341	331	SYEP_HUMAN	8.0×10^{-35}	+1	Pro-tRNA synthetase
Z33362_1	1259	SYS_ECOLI	6.2×10^{-39}	+1*	Ser-tRNA synthetase
Z33372	353	SYN_ECOLI	1.7×10^{-33}	−3	Asn-tRNA synthetase
Z33011_1	1803	RL22_MYCCA	3.2×10^{-57}	+1*	50S ribosomal protein
Z33011_2	1803	RS3_MYCCA	7.0×10^{-171}	+3	30S ribosomal protein
Z33011_3	1803	RL16_MYCCA	1.8×10^{-106}	+2	50S ribosomal protein
Z33011_4	1803	RI29_MYCCA	2.0×10^{-54}	+1*	50S ribosomal protein
Z33016_2	1476	RS2_ECOLI	1.3×10^{-75}	−1	30S ribosomal protein
Z33034_1	840	RL36_BACSU	2.1×10^{-8}	−2	50S ribosomal protein
Z33050_1	1490	RL17_BACST	3.8×10^{-24}	−2	50S ribosomal protein

Table 2. Continued

Contig	Bases	Best hit	P-value	Frame	Function
Z33050_3	1490	RS11_BACST	1.7×10^{-7}	−1	30S ribosomal protein
Z33076_1	1299	RL33_BACST	1.3×10^{-9}	−2	50S ribosomal protein
Z33236	396	RS18_BACST	3.3×10^{-23}	−2*	30S ribosomal protein
Z33246	180	RL10_STRAT	1.4×10^{-2}	+3	Similar to 50S ribosomal protein
Z33291	324	RL6_MYCCA	5.8×10^{-55}	−2*	50S ribosomal protein
Z33312	244	RS11_BACSU	1.2×10^{-27}	−1*	30S ribosomal protein
Z33327_1	422	RL23_MYCCA	8.6×10^{-32}	−2*	50S ribosomal protein
Z33327_2	422	RL4_MYCCA	2.3×10^{-38}	−3	50S ribosomal protein
Z33333	273	RL34_PSEAE	5.2×10^{-15}	−3	59S ribosomal protein
Z33343_1	339	RS9_BACST	7.1×10^{-25}	−3	30S ribosomal protein
Z33343_2	339	RL13_HAESO	7.2×10^{-4}	−1	50S ribosomal protein
Z33352_2	1629	RL35_BACST	1.2×10^{-7}	+1*	50S ribosomal protein
Z33352_3	1629	RL20_BACST	5.1×10^{-7}	+3	50S ribosomal protein
Z33366_1	1299	RL11_THEMA	2.3×10^{-42}	+2	50S ribosomal protein
Z33366_2	1299	RL1_BACST	2.6×10^{-95}	+1	50S ribosomal protein
Rl2_Mycca					50S ribosomal protein
Rl3_Mycca					50S ribosomal protein
Rl5_Mycca					50S ribosomal protein
Rs8_Mycca					30S ribosomal protein
Rl15_Mycca					50S ribosomal protein
Rl24_Mycca					50S ribosomal protein
Rs10_Mycca					30S ribosomal protein
Rs17_Mycca					30S ribosomal protein
Rs19_Mycca					30S ribosomal protein
Protein folding					
Z33106	753	DNAK_BACME	5.6×10^{-102}	+1	DnaK protein
Z33267	364	YGRP_BACSU	2.7×10^{-11}	−1	CLPA/CLPB family
Z33269	228	CLPX_ECOLI	1.1×10^{-3}	+3	CLPA/CLPB family
Z33364_1	754	DNAJ_BACSU	4.2×10^{-19}	−1	DnaJ protein
Z33364_2	754	DNAK_BORBU	2.9×10^{-20}	−3*	DnaK protein
Z33365	569	TIG_ECOLI	4.9×10^{-19}	+1	Trigger factor
Z33376	488	CLPB_ECOLI	4.8×10^{-50}	−3	CLPA/CLPB family
Internal transport and translocation					
Z33092	549	SECA_BACSU	2.6×10^{-59}	+3	Preprotein translocase SECA subunit
Z33142	595	PSTC_ECOLI	2.7×10^{-11}	+1	Phosphate transport protein
Z33270	198	SECA_BACSU	7.8×10^{-7}	−2	Preprotein translocase SECA subunit
Z33315	221	AMIC_STRPN	1.2×10^{-4}	−2	Oligopeptide transporter
Z33328	396	SECY_MYCCA	1.1×10^{-72}	−1	Preprotein translocase SECY subunit
Z33334	254	PSTA_ECOLI	6.8×10^{-9}	+3	Phosphate transporter PstA
Z33358	350	SRP5_MYCMY	8.3×10^{-50}	−3*	Signal recognition particle
Z33374_1	2095	AMIE_STRPN	2.8×10^{-96}	−2*	Oligopeptide transporter
Z33374_2	2095	AMID_STRPN	5.5×10^{-30}	−1*	Oligopeptide transporter
Z33375	241	AMIF_STRPN	1.1×10^{-36}	+1	Oligopeptide transporter
Unclassified					
Z33012_1	872	S14881P	6.5×10^{-16}	+2	Hypothetical yeast protein 1
Z33012_2	872	S14882P	3.2×10^{-8}	+3	Hypothetical yeast protein 2
Z33013_2	1904	GIDB_BACSU	1.0×10^{-32}	+1	Methyltransferases
Z33013_3	1904	OBG_BACSU	1.8×10^{-12}	+1	Similar to GTP-binding GTP1/OBG family
Z33023	808	HP9ORFS_7	3.6×10^{-2}	−2	Hypothetical yeast protein 7
Z33024	788	BSGENR_1GP	4.5×10^{-11}	−2	Unknown function
Z33030_2	1748	OBG_BACSU	2.4×10^{-3}	−3	Similar to GTP-binding GTP1/OBG family
Z33032_2	1250	HIS4_METVA	2.2×10^{-1}	−1	Similar to isomerases/dehydrogenases
Z33040	1019	YMG1_MYCGE	6.3×10^{-8}	+1	Unknown function
Z33045_2	640	A44803P	6.9×10^{-20}	−3#	Human pG1 protein
Z33058	719	Y311_BACSU	8.8×10^{-17}	+2	GTP binding CD48/PAS1/SEC18 family
Z33060_2	1308	YAAC_PSEFL	1.2×10^{-5}	−3	Similar protein X in RPST-ILES region
Z33103	1306	A43577P	4.5×10^{-12}	−1	Regulatory *Clostridium* protein PfoR
Z33138	815	SMPB_ECOLI	2.1×10^{-9}	+1	Small protein B
Z33140	583	YJIF_ECOLI	3.6×10^{-6}	−2	Secreted protein downstream SMP
Z33167	935	OBG_BACSU	5.7×10^{-12}	−1*	Similar to GTP-binding GTP1/OBG family
Z33192	443	YP15_STAAU	4.5×10^{-3}	+3	Unknown function (see YDEU_BACSU)
Z33195	617	LEPA_ECOLI	9.7×10^{-21}	+2	GTP-binding proteins
Z33259	443	DBH_BACSU	6.7×10^{-12}	+1	DNA-binding protein HU
Z33329	548	YSR1_MYCMY	7.0×10^{-115}	+2	Protein in SRPM54 5′ region

Table 2. Continued

Contig	Bases	Best hit	P-value	Frame	Function
Z33342	344	P37_MYCHR	3.5×10^{-3}	−3	Surface protein (transport-associated)
Z33354	1538	A43863P	3.3×10^{-25}	−1	Similar to a haemolysin

RNAs

Z33007 (tRNA), Z33009 (tRNA), Z33010_1 (tRNA), Z33025 (5S rRNA), Z33045_1 (16S rRNA), Z33060_3 (tRNA), Z33070 (snRNA-like), Z33094 (16S/23S rRNA), Z33172 (23S rRNA), Z33190 (snRNA-like), Z33223 (rRNA), Z33231 (23S rRNA)

Repeats

Z33005, Z33010_2, Z33014, Z33036, Z33037, Z33038, Z33043, Z33046, Z33059_1, Z33065, Z33078, Z33082, Z33101, Z33116, Z33117, Z33119, Z33124, Z33130, Z33131, Z33152, Z33164, Z33175, Z33185, Z33197, Z33198, Z33200, Z33204, Z33206, Z33211, Z33212, Z33215, Z33217, Z33227, Z33238, Z33243, Z33245, Z33248, Z33257, Z33258, Z33324, Z33351, Z33356, Z33359, Z33360

Contig: EMBL accession number; an underscore followed by a number denotes a particular segment of a contig that has several distinct proteins or other features. The contigs are sorted according to functional categories.
Bases: length of contig in number of base pairs.
Best hit: SWISSPROT codes or PIR accession numbers for the database protein with the best sequence similarity.
P-value: significance of the similarity from the BLASTP program. All similarities with *p*-values above 10^{-9} were verified by other techniques such as multiple alignments.
Frame: position of first codon and orientation of ORF relative to the beginning of the contig DNA sequence.
Similar to: denotes a statistically significant homology without precise assignment of function.
Repeat: used for all contigs with no homologues in protein databases that resemble at least one other *M. capricolum* contig or match known DNA repeats in nucleic acid sequence databases. *M. capricolum* proteins already in sequence databases have the SWISSPROT identifiers (Bairoch and Boeckmann, 1993) in column Contig. The longest continuous DNA segment known so far includes the origin of replication (Miyata *et al.*, 1993).
a. This ATP pyrophosphatase domain is usually associated with other domains within one peptide chain (e.g. in GMP synthetase). Therefore, it might be only a subunit of a larger complex (data not shown).
b. HAD is a superfamily of enzymes that contains functionally diverse hydrolases.
*Putative frameshifts.
#Possibly erroneous stop codons.

Anabolism. While the presence of catabolic enzymes is expected, the detection of putative anabolic enzymes is somewhat surprising for such a highly specialized parasite. Indeed, only a few were found. (i) One of them appears to be involved in amino acid biosynthesis: succinyl-diaminopimilate desuccinylase catalyses a step in the synthesis of diaminopimeline and lysine. (ii) Another anabolic enzyme is a phosphatidyl transferase involved in phospholipid synthesis. (iii) A third one is tetrahydrofolate (THF) synthase. THF and its derivatives are essential for numerous C1 unit (carbon) transfers that occur in a living cell. The last steps in THF biosynthesis involve the reduction of folate, a vitamin for many organisms including the majority of mycoplasmas (Finch and Mitchell, 1992). All three anabolic enzymes act at the level of metabolic intermediates, so the complete pathways may not be necessarily present.

Table 3. Identification of rare restriction sites in the 214 kb of *M. capricolum*.

Enzyme	Recognition site	Total[a]	Observed
*Fsp*I	TGCGCA	5	2
*Bgl*I	GCCNNNNGCC	6	0
*Apa*I	GGGCCC	2	1
*Bss*HII	GCGCGC	1	0
*Sal*I	GTCGAC	2	2
*Sma*I	CCCGGG	2	1
*Xho*I	CTCGAG	2	1
Total		20	7 (35%)

a. The restriction sites mapped onto the *M. capricolum* genome were taken from Whitely *et al.* (1991) and Miyata *et al.* (1991). *Bam*HI was omitted as reciprocal 2D-PFGE was used to map the sites relative to the *Bgl*I site (see Whitely *et al.*, 1991 and refs therein). Extrapolating from the number of observed cutting sites (even by including *Bam*HI for which 1 out of 9 sites were found), a genome size of about 780 kb would result.

DNA repair. Several DNA repair mechanisms exist, and the lack of at least some of them in *M. capricolum* has been claimed (for review see Labarere, 1992). We have found excinucleases and homologues of at least two proteins, RecM and RecN, that are thought to be involved in SOS DNA repair mechanisms (Alonso *et al.*, 1990; van Hoy and Hoch, 1990). RecN of *B. subtilis* and *E. coli* and their *M. capricolum* homologue show high similarity to a yeast chromosome segregation protein (Fig. 1). Indeed, internal coiled coil domains (data not shown) suggest a mechanochemical role of this family. Interestingly, the entire N-terminal ATP-binding domain of the family is similar to that of RecF, yet another *E. coli* protein involved in DNA repair (Gorbalenya and Koonin, 1990).

DNA replication. Although it was thought that *Mycoplasma* species have only one DNA polymerase (Razin, 1992; Labarere, 1992), ORFs with similarity to *E. coli* DNA polymerases I and III reveal the presence of at least two distinct enzymes (Table 2). Therefore, replication in *M. capricolum* may resemble the replication mechanisms of other eubacteria more closely than previously thought.

A

```
family:                                    h hhhG  G GKoohh h
Recf_Ecoli     1 SLTRLLIRDFRNI.ETADLALS.PGFNFLVGANGSGKTSVLEAIYTLGHGRAFRSLQIGRVIRHEQEAFVLHG
Recf_Promi     2 ILSRLLIRHFRNI.EQADLPLA.DGFNFLVGPNGSGKTSILEAIYTLGHGRAFRSAQANRVIQHDENAFILHG
Recf_Actpl     2 PLSRLIINNFRNL.QSLDLELS.PNFNFIVGHNGSGKTSLLEAIFYLGHGRSFKSHISNRIIHYQAEDFVLHA
Recf_Bacsu     2 YIQNLELTSYRNY.DHAELQFE.NKVNVIIGENAQGKTNLMEAIYVLSMAKSHRTSNDKELIRWDKDYAKIEG
Recn_Ecoli     1 MLAQLTISNFAIV.RELEIDFH.SGMTVITGETGAGKSIAIDALGLCLGGRAE.............ADMVRTG
Recn_Bacsu     1 MLAELSIKNFAII.EELTVSFE.RGLTVLTGETGAGKSIIIDAISLLVGGRGS.............SEFVRYG
P115/Mychy     3 KLIKIEIEGFKSFADPISINFD.GSVVGIVGPNGSGKSNINDAIRWVLGEQSAKQLRGLNM.....DDVIFAG
Z33332         ?     IRASGFKXFADLTVMDFN.YDMTGVVGPNGSGKSNITDAIRWTLGXQSTKTLRGSKM.....ADIVXSG
Smc1_Yeast     3 RLVGLELSNFKSYRGVTKVGFGESNFTSIIGPNGSGKSNMMDAISFVLGVRS.NHLRSNIL.....KDLIYRG
Nam1/Yeast     ? YIKKVILRNFMCH.EHFELELG.SRLNFIVGNNGSGKSAILTAITIGLGAKASETNRGSSL.....KDLIREG
consensus:       h  h h tF  h t h htht  th  hhGttGtGKo hhtAh hh tttt              thh  G
```

B

```
family:                                                hSGG    h h             hhhhDE
Recn_Ecoli   394 IDVKFDEHHLGADG.ADRIEFRVTTNPGQPMQPI..SKVASGGELSRIALAIQVITARKMETPALIFDEVDVG
Recn_Bacsu   410 PLVNGQPVQLTEQG.IDLVKFLISTNTGEPLKSL..AKVASGGELSRVMLAIKSIFSSQQDVTSIIFDEVDTG
P115/Mychy   839 KMFGGGKAEIHFTDKNDILNSGVEISAQPPGKTIKNLRLFSGGEKAIIAISLLFAILKARPIPLCILDEVEAA
Z33355         ? KMFGGGYAELIYTDPDNILETGIDIKIFPPGKKITNLNLLSGGEKSLVALSVLFAILKARPLPLVILDEAEAP
Yat3_Rhoru     ? KLFGGGRAHLTLIESDDPLEAGLEIMASPPGKRLQSLGLLSGGEQALTATALLFAVFLTNPAPICVLDEVDAP
Smc1_Yeast  1090 VELAGGNASLTIEDEDEPFNAGIKYHATPPLKRFKDMEYLSGGEKTVAALALLFAINSYQPSPFFVLDEVDAA
consensus:         htGt   hh  t  t ht  hth  ttP Kth  hthhSGGE  hhAh h   h    t P hhhDEh-

family:          hD                    hhh H
Recn_Ecoli       ISGPTAAVVGKLLRQLGE..STQVMCVTHLPQVAGCGHQHYFVSKETDGAMTETHMQSLNKKAR   P05824
Recn_Bacsu       VSGRVAQAIAEKIHKVSI..GSQVLCITHLPQVAAMADTHLYIAKELKDGRTTTRVKPLSKQEK   P17894
P115/Mychy       LDESNVIRYVEFLKLLKE..NTQFLIITHRSGTMSRVDQLLGVTMQKRG.VTSIFSVELSKAKE   M34956
Z33355           LDPVNVERFARYVRHFSD..NTQFIIVTHREGTMTQCDSLFGVTMQTKG.ITKIINVKLVEAKN   Z33355
Yat3_Rhoru       LDDANVDRFCAMLRHLTDTTGTRFLVVTHHRMTMARMDRLFGVTMAERG.VSSLVSVDLCQAED   P15016
Smc1_Yeast       LDITNVQRIAAYIRRHRN.PDLQFIVISLKNTMFEKSDALVGVYRQQQENSSKIITLDLSNYAE   P32908
consensus:       ht   h  hh hht  t   t QhhhhoH t hh   D hhhh  t tt  othh  tL t tt
```

Fig. 1. Multiple alignment of two ORFs in contigs Z33332 and Z33355, with RecN and related proteins. Based on the strong similarity to the P115 proteins of other mycoplasmas, Z33332 and Z33355 might be fragments of the same protein. The role of P115 remains, however, to be elucidated, although the similarity to different repair proteins (RecN, RecF) is striking. SMC1 is involved in chromosome segregation. As in SMC1, RecN and Z33355, the N- and C-terminus is separated by coiled coil regions (cc) of different length, as predicted using the algorithm of Lupas *et al.* (1991). All the proteins shown in the alignments are members of a vast group of functionally diverse ATP-binding proteins (Gorbalenya and Koonin, 1990).
A. Alignment of the ORF in contig Z33332 with the N-terminal (ATP-binding) domain of the closest relatives.
B. Alignment of Z33355 with the closest relatives. The first line (family) indicates the consensus of the superfamily whereas the bottom line (consensus) summarizes conserved features of the alignment in the RecN subfamily: capitals, amino acids conserved in at least all but one sequence; h, hydrophobic position; o, S or T, negatively charged. An X within the alignment denotes an unidentified residue. The first column shows the SWISS-PROT codes (Bairoch and Boeckmann, 1993) of the respective proteins if available (underscore in protein name); the second row indicated the position of the aligned sequences in the respective proteins.

Metabolism of a parasite. The analysis further shows the presence of proteins involved in protein biosynthesis, translation and transcription regulation, protein folding and intracellular protein transport (Table 2). As expected for a parasite, a relatively large number of proteins is involved in the transport of intermediate metabolites and energy sources through the membrane, and their modification (first category, Table 2). As an even more surprising fact for a 'perfect' parasite, several carbohydrate- and protein-degrading enzymes have apparently been retained by *M. capricolum* to produce usable intermediates in metabolism (Table 2).

Quantitative sequence analysis: comparison with M. genitalium *and* E. coli

Quantitative sequence analysis can also reveal how closely related two organisms are, how protein function is conserved and which proteins appear to be most constrained in evolution. The high percentage of *M. capricolum* proteins with homologues of known function in other organisms supports the notion that mainly essential and conserved genes have been retained.

A set of contigs from *M. genitalium* that cover about 101 kb (Peterson *et al.*, 1993) provide a good basis for species comparisons. Although *M. capricolum* and *M. genitalium* both belong to the Mollicutes, mycoplasmas are a polyphyletic group, separated by other species including those with intact cell walls (Weisburg *et al.*, 1989). Indeed, the two data sets are complementary to each other with surprisingly little overlap: comparing our data with the contigs from *M. genitalium*, only 25 protein matches were observed, their sequence identity ranging between 40 and 80%. Because of the few and very short matches, however, quantification and generalization are difficult. Although it is still possible that all *M. genitalium* proteins are present in *M. capricolum*, the considerable variability among mycoplasmas is striking.

Numerous homologues were detected in *E. coli*. In order to quantify these similarities, all matches with *E. coli*

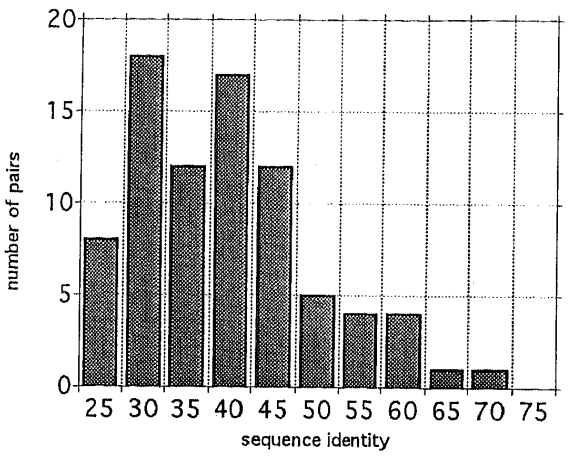

Fig. 2. Histogram of sequence identity of 82 *M. capricolum* proteins with their putative orthologues in *E. coli*. All hits above 50% amino acid identity are shown in Table 4.

proteins were studied in a first attempt to compare considerable fractions of two prokaryotic genomes. As many as 107 out of the 215 *M. capricolum* proteins functionally identified in this study have a homologue in *E. coli*. As it can only be estimated whether these are orthologues (the equivalent gene in another species) or paralogues (implying descent after gene duplication; Fitch, 1970), we have selected 95 *E. coli* sequences that can be readily aligned to their mycoplasma homologues above a certain similarity threshold (see the *Experimental procedures*). Lower scoring proteins, such as ABC transporters and members of the helix-turn-helix DNA repressor family, were excluded from the set of 95 homologous pairs. The similarity between the remaining 82 putative orthologues from *E. coli* and *M. capricolum* ranges from 26% to 72% amino acid identity, with an average of 41% (Fig. 2). Most of the 'genetic' proteins (e.g. in translation, transcription, replication, repair and cell division) and those involved in protein biosynthesis are the most similar, the highest being the cell division protein FtsH with 72.3% amino acid identity over 83 residues (Table 4). Although metabolic enzymes in general have considerably lower similarity levels, enolase also appears highly conserved (Table 4). This conclusion is tentative, as the two short enolase fragments sequenced so far (ORFs in contigs Z33232 and Z33255) may be parts of the most conserved regions. A longer strongly conserved match (59% identity over 95 residues) was found for protease LA (Table 4). Although it is known that protease LA plays an important role in controlled degradation of short-lived or abnormal proteins (Suzuki *et al.*, 1994), the precise reason for the constraints at the sequence level remains to be discovered.

Quantitative sequence analysis may also reveal unusual modes of protein evolution, i.e. contradictions between the generally accepted phylogenetic position of an organism and the evolutionary tree of a particular protein family. One of those cases is a putative *M. capricolum* prolyl-tRNA synthetase with high similarity to mammalian multifunctional aminoacyl-tRNA synthetases but only modest similarity to *E. coli* and other prolyl-tRNA synthetases (Fig. 3). One explanation for this anomaly would be that prokaryotic orthologues have not been sequenced yet. However, more than 60% of the *E. coli* genome is already stored in public databases and several other prokaryotic genome projects have also produced large amounts of

Table 4. Proteins or protein fragments most conserved between *E. coli* and *M. capricolum* (pairwise amino acid identities above 50%).

Contig	*E. coli*	DIST	OPT	%IDE	LEN	Description
Z33086	ftsh_ecoli	47.5	284	72.3	83	Cell division protein
Z33255	eno_ecoli	36.8	175	65.6	61	Enolase
Z33376	clpb_ecoli	38.4	400	63.2	125	clpb (stress protection)
Z33013	ychf_ecoli	36.9	228	63.0	73	OBG-related protein
Z33068	efg_ecoli	36.7	608	61.5	200	Elongation factor EF-G
Z33345	uvrb_ecoli	35.2	980	60.0	305	Excinuclease subunit B
Z33034	if1_ecoli	32.5	240	59.4	69	Initiation factor IF-1
Z33292	lon_ecoli	34.1	263	58.9	95	Cytoplasmic protease LA
Z33370	atpb_ecoli	32.3	917	57.1	345	ATP synthase beta chain
Z33195	lepa_ecoli	29.8	197	56.3	71	GTP-binding protein LepA
Z33330	atpa_ecoli	24.1	161	52.4	63	ATP synthase alpha chain
Z33011	rl22_ecoli	27.0	195	51.8	85	50S ribosomal protein L22
Z33026	syn_ecoli	26.9	459	51.7	172	Asn-tRNA synthetase
Z33011	rl16_ecoli	25.9	367	50.7	134	50S ribosomal protein L16
Z33120	uvra_ecoli	25.7	510	50.5	200	Excinuclease subunit A

Contig: EMBL/GenBank accession numbers.
E. coli: SWISSPROT codes for *E. coli* proteins.
DIST: distance (in percentage points) to the length-dependent threshold of structural homology (25% for length 80 or more residues, higher for shorter alignments (Sander and Schneider, 1991).
OPT: FASTA 'optimized' scores (Pearson and Lipman, 1988).
%IDE: percentage of identical residues.
LEN: extent of the pairwise alignment number of amino acid residues.

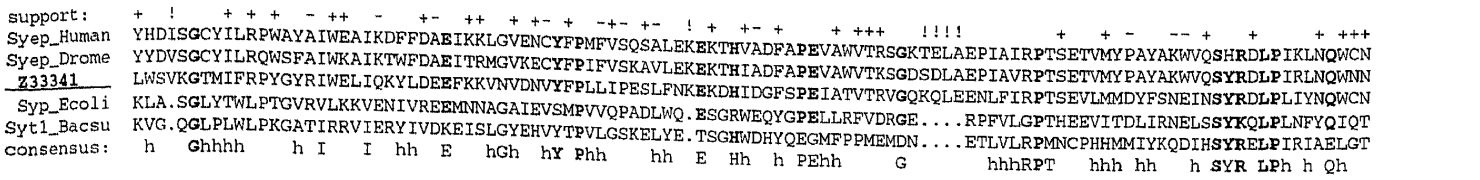

Fig. 3. Alignment of tRNA synthetases similar to an ORF in contig Z33341. The ORF in contig Z33341 is closer to the animal proteins than to other prokaryotic synthetases, which suggests the possibility that horizontal transfer has been involved in the acquisition of the gene for prolyl-tRNA synthetase. The first line shows positions that support grouping with the animal (+) or with the prokaryotic (−) sequences; characteristic insertions are denoted by an exclamation mark (!).

sequence data (for review, see Bork *et al.*, 1994). Therefore, the chance of finding relatively well-characterized proteins which have not yet been described in other prokaryotes is relatively low. A tempting hypothesis for this surprising similarity would be the acquisition of the animal gene by horizontal gene transfer (the lateral transfer of genetic material across species). Although the acquisition of animal genes by bacteria is extremely rare (Heinemann, 1991; Doolittle and Bork, 1993), the transfer of genetic material between bacteria (e.g. via plasmids) appears to be a frequent event and it is known that parasites are able to acquire genes from their hosts (Heinemann, 1991). Interestingly, a similar phenomenon has been independently proposed for the other (N-terminal) part of the multifunctional aminoacyl-tRNA synthetase. This part of the protein codes for a Glu-tRNA synthetase which is more similar to Gln-tRNA than Glu-tRNA synthetase in *E. coli* (Lamour *et al.*, 1994). It has been suggested that Gln-tRNA synthetase in various eubacteria has evolved from eukaryotic Glu-tRNA synthetase after having been acquired by horizontal gene transfer (Lamour *et al.*, 1994). Other candidates for a horizontal transmission of genetic material in *M. capricolum* are U6 snRNA-like sequences (Ushida and Muto, 1993), which are involved in pre-mRNA splicing in eukaryotes, while no such phenomena have been reported for eubacteria.

Conclusion

The detection of at least 215 proteins in 214 kb DNA shows the power of the current sequencing and analysis approaches. The function of many proteins can be predicted by similarity, and biochemical pathways can be reconstructed. Quantitative sequence analysis also indicates an unorthodox similarity between some eukaryotic proteins and *M. capricolum* proteins. As mycoplasmas have been shown to have faster mutation rates than many other prokaryotes (Woese *et al.*, 1984), the conservation profile of numerous proteins provides a wealth of data about functional and structural constraints at the protein sequence level, with proteins acting on DNA or involved in protein biosynthesis being the most constrained.

Comparison of complete genomes of prokaryotes such as *E. coli*, as well as eukaryotes such as *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, will soon enable us to follow in detail the evolution of genomes on the basis of sequence similarity. The study presented here is a first attempt to include *M. capricolum* in such systematic comparisons.

Experimental procedures

Sequencing of the contigs

The sequences were obtained by the genomic walking technique (Gillevet, 1990), i.e. the direct sequencing of the organism by probing Southern blots of sequencing gels (Ohara *et al.*, 1989). These blots contain the DNA of the entire organism that has been cut with restriction enzymes, treated with chemical sequencing reactions, run out on a sequencing gel and transferred to a charged nylon membrane. When such a membrane is probed with a labelled oligonucleotide, the resulting autoradiograph displays sequencing patterns in those lanes in which the oligonucleotide has hybridized near a restriction cut. About 1000 random clones were sequenced using fluorescent technology to supply starting points for the walking process (Gillevet, 1993). We also picked starting points from 64 segments taken from the EMBL nucleotide sequence database. The resulting 372 contigs have been deposited in the EMBL database (accession numbers Z33005 to Z33376).

First-round sequence analysis

The computational and database analysis comprises the following steps.

1. The assembled consensus sequence from the contigs was subjected to BLASTX searches (Gish and States, 1993) using the mycoplasma genetic code in which UGA codes for tryptophan. BLASTX performs a rapid search of all six conceptual translations of the contig consensus against protein sequence databases.
2. In order to check for frameshifts and artificial stop codons within the contigs, BLASTX output was also automatically parsed (program FRAMESHIFT by G. Casari, unpublished) and critical regions extracted.
3. All possible ORFs longer than 10 amino acids were translated. The requirement for recording ORFs within the contigs was the presence of start and stop codons; in terminal fragments, only start (C-terminus) or stop codons (N-terminus) were required.
4. With the resulting 1845 putative ORFs, BLAST similarity

searches (Altschul *et al.*, 1990) against sequence databases were carried out.
5. For ORFs longer than 30 amino acids without clear similarity, other search methods (Bork *et al.*, 1992; Koonin *et al.*, 1994) were applied and possible 'twilight zone' hits studied in detail.
6. Finally, DNA sequence databases were screened for non-protein coding elements such as RNA genes and internal repeats.
7. All results were merged (Table 2) and stored in a relational database (Scharf *et al.*, 1994) for further evaluation.

The last database search was done in January 1994. We realize that the increasing number of characterized proteins from other organisms is a valuable source of information (Bork *et al.*, 1994) and the searches will be repeated in updated databases with a more complete *M. capricolum* data set that we are currently obtaining.

Quantitative comparison

For verifying the significance of subtle similarities or detecting the possibility of horizontal gene transfers, multiple alignments and phylogenetic trees (as implemented in CLUSTALW, a new version of CLUSTALV; Higgins *et al.*, 1992) were used.

The comparison with *E. coli* and *M. genitalium* proteins involved the following steps: (i) extracting all *E. coli* and *M. genitalium* proteins from sequence databases; (ii) all-against-all comparison with the mycoplasma contigs using TFASTA (Pearson and Lipman, 1988); (iii) extraction of putative orthologues; and (iv) ranking the hits according to several scoring schemes such as FASTA 'opt' scores (Pearson and Lipman, 1988), amino acid identity and distance to a length-dependent threshold for structural similarity (Sander and Schneider, 1991).

Acknowledgements

We would like to thank all members of the Harvard Genome Lab., A. Ally, F. Barton, S. E. Benner, R. Clark-Whitehead, N. Dougles, E. Hsu, L. Marquez, M. S. Purzcycki, B. Richter, J. Sartell, S. W. Smith, C. Wang and J. Williams, for their efforts over the course of the *M. capricolum* sequencing project to make this new technology a reality and to produce this data set. We are grateful to B. Altenberg for help during the analysis and R. Wade for critical reading of the manuscript.

References

Adams, M.D., Kerlavage, A.R., Fields, C., and Venter, J.C. (1993) 3400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genetics* **4:** 256–267.

Alonso, C., Shirahige, K., and Oganawana, N. (1990) Molecular cloning, genetic characterization, and DNA sequence analysis of the recM region of *B. subtilis*. *Nucl Acids Res* **18:** 6771–6777.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Bairoch, A., and Boeckmann, B. (1993) The SWISS-PROT protein sequence databank. *Nucl Acids Res* **21:** 3093–3096.

Blattner, F.R., Burland, V., Plunkett, III, G., Sofia, H.J., and Daniels, D.L. (1993) Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.9 minutes. *Nucl Acids Res* **21:** 5408–5417.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. (1992) Comprehensive sequence analysis of the 182 ORFs of yeast chromosome III. *Prot Sci* **1:** 1677–1690.

Bork, P., Ouzounis, C., and Sander, C. (1994) From genome sequences to protein function. *Curr Opin Struct Biol* **4:** 393–403.

Claverie, J.-M., and States, D. (1993) Information enhancement methods for large scale sequencing analysis. *Comp Chem.* **17:** 191–201.

Daniels, D.L., Plunkett, III, G., Burland, V., and Blattner, F. (1992) Analysis of the *E. coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257:** 771–778.

Doolittle, R.F., and Bork, P. (1993) Evolutionarily mobile protein modules. *Sci Am* **269(4):** 50–56.

Dujon, B., Alexandraki, D., Andre, B., Ansorge, B., Baladron, V., Ballesta, J.P.G. *et al.* (1994) Complete DNA sequence of yeast chromosome XI. *Nature* **369:** 371–378.

Fickett, J.W., and Guigo, R. (1993) Estimation of protein coding density in a corpus of DNA sequence data. *Nucl Acids Res* **21:** 2837–2844.

Finch, L.R., and Mitchell, A. (1992) Sources of nucleotides. In *Mycoplasmas — Molecular Biology and Pathogenesis.* Maniloff, J., McElhaney, R.N., Finch, L.R., and Baseman, J.B. (eds). Washington, DC: American Society for Microbiology, pp. 211–230.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool* **19:** 99–113.

Gillevet, P.M. (1990) Chemiluminescent multiplex DNA sequencing. *Nature* **348:** 657–658.

Gillevet, P.M. (1993) Integration of the Wet Lab and Dataflow in multiplex genomic walking. In *Proceedings of the Second International Conference of Bioinformatics, Supercomputing and Complex Genome Analysis.* Lim, H.A., Fickett, J.W., Cantor, C.R., and Robbins, R.J. (eds). River Edge, NJ: World Scientific Publishing, pp. 197–205.

Gish, W., and States, D. (1993) Identification of protein coding regions by database similarity search. *Nature Genetics* **3:** 266–272.

Glaser, P., Kunst, F., Arnaud, M., Coudart, M.-P., Gonzales, W., Hullo, M.-F. *et al.* (1993) *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol Microbiol* **10:** 371–384.

Gorbalenya, A.E., and Koonin, E.V. (1990) Superfamily of UvrA-related NTP-binding proteins. *J Mol Biol* **213:** 583–591.

Heinemann, J.A. (1991) Genetics of gene transfer between species. *Trends Genet* **7:** 181–185.

Herrmann, R. (1992) Genome structure and organization. In *Mycoplasmas — Molecular Biology and Pathogenesis.* Maniloff, J., McElhaney, R.N., Finch, L.R., and Baseman, J.B. (eds). Washington, DC: American Society for Microbiology, pp. 157–168.

Higgins, D.G., Bleasby, A.J., and Fuchs, R. (1992) Clustal V

— an improved software for multiple sequence alignment. *CABIOS* **8:** 189–191.

Honoré, N., Beregh, S., Chanteau, S., Doucet-Populaire, F., Eiglmeier, K., Garnier, T. *et al.* (1993) Nucleotide sequence of the first cosmid from the *Mycobacterium leprae* genome project: structure and function of the Rif–Str regions. *Mol Microbiol* **7:** 207–214.

van Hoy, B.E., and Hoch, J.A. (1990) Characterization of the spoIVB and recN loci in *B. subtilis*. *J Bacteriol* **172:** 1306–1311.

Koonin, E.V., Bork, P., and Sander, C. (1994) Yeast chromosome III: new gene functions. *EMBO J* **13:** 493–503.

Labarere, J. (1992) DNA replication and repair. In *Mycoplasmas — Molecular Biology and Pathogenesis*. Maniloff, J., McElhaney, R.N., Finch, L.R., and Baseman, J.B. (eds). Washington, DC: American Society of Microbiology, pp. 309–323.

Lamour, V., Quevillon, S., Diriong, S., N'Guyen, V.C., Lipinski, M., and Mirande, M. (1994) Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. *Proc Natl Acad Sci USA* **91:** 8670–8674.

Lupas, A., Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science* **252:** 1162–1164.

Maniloff, J. (1989) Anomalous values of *Mycoplasma* genome sizes determined by pulse-field electrophoresis. *Nucl Acids Res* **17:** 1268.

Maniloff, J., McElhaney, R.N., Finch, L.R., and Baseman, J.B. (eds) (1992) *Mycoplasmas — Molecular Biology and Pathogenesis*. Washington, DC: American Society for Microbiology.

Miyata, M., Wang, L., and Fukumura, T. (1991) Physical mapping of the *Mycoplasma capricolum* genome. *FEMS Microbiol Lett* **79:** 329–334.

Miyata, M., Sano, K.-I., Okada, R., and Fukumura, T. (1993) Mapping of replication initiation site in *Mycoplasma capricolum* genome by two-dimensional gel electrophoretic analysis. *Nucl Acids Res* **21:** 4816–4823.

Muto, A., Yamao, F., and Osawa, S. (1987) The genome of *Mycoplasma capricolum*. *Progr Nucl Acids Res* **34:** 28–58.

Neimark, H.C., and Lange, C.S. (1990) Pulse-field electrophoresis indicates full-length mycoplasma chromosomes range widely in size. *Nucl Acids Res* **18:** 5443–5448.

Nowak, R. (1995) Venter wins sequencing race — twice. *Science* **268:** 1273.

Ohara, O., Dorit, R.L., and Gilbert, W. (1989) Direct genomic sequencing of bacterial DNA: the pyruvate kinase I gene of *Escherichia coli*. *Proc Natl Acad Sci USA* **86:** 6883–6887.

Pearson, W.R., and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85:** 3338–3342.

Peterson, S.N., Hu, P.-C., Bott, K.F., and Hutchinson, C.A. III (1993) A survey of the *Mycoplasma genitalium* genome by using random sequencing. *J Bacteriol* **175:** 7918–7930.

Poddar, S.K., and Maniloff, J. (1989) Determination of microbial genome sizes by two-dimensional denaturating gradient gel electrophoresis. *Nucl Acids Res* **8:** 2889–2895.

Pollak, J.D. (1992) Carbohydrate metabolism and energy conservation. In *Mycoplasmas — Molecular Biology and Pathogenesis*. Maniloff, J., McElhaney, R.N., Finch, L.R., and Baseman, J.B. (eds). Washington, DC: American Society for Microbiology, pp. 181–200.

Posfai, J., and Roberts, R. (1992) Finding errors in DNA sequences. *Proc Natl Acad Sci USA* **89:** 4698–4702.

Razin, S. (1992) Peculiar properties of mycoplasmas: the smallest self-replicating prokaryotes. *FEMS Microbiol Lett* **100:** 423–432.

Robertson, J.A., Pyle, L.E., Stemke, G.W., and Finch, L.R. (1990) Human ureaplasmas show diverse genome sizes by pulse-field electrophoresis. *Nucl Acids Res* **18:** 1451–1455.

Ruland, K., Wenzel, R., and Herrmann, R. (1990) Analysis of different repeated DNA elements present in the P1 operon of *M. pneumoniae*: size, number and distribution in the genome. *Nucl Acids Res* **18:** 6311–6317.

Sander, C., and Schneider, R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9:** 56–68.

Scharf, M., Schneider, R., Casari, G., Bork, P., Valencia, A., Ouzounis, C., and Sander, C. (1994) GeneQuiz: a workbench for sequence analysis. *Proc 2nd Int Conf Intell Syst Mol Biol*, Palo Alto, CA: AAAI Press, pp. 348–353.

States, D., and Botstein, D. (1992) Molecular sequence accuracy and the analysis of protein coding regions. *Proc Natl Acad Sci USA* **88:** 5518–5523.

Suzuki, C.K., Suda, K., Wang, N., and Schatz, G. (1994) Requirement for the yeast gene LON in intramitochondrial proteolysis and maintenance of respiration. *Science* **264:** 273–276.

Ushida, C., and Muto, A. (1993) A small RNA of *M. capricolum* that resembles eukaryotic U6 small nuclear RNA. *Nucl Acids Res* **21:** 2649–2653.

Weisburg, W.G., Tully, J.G., Rose, D.L., Petzel, J.P., Oyaizu, H., Yang, D. *et al.* (1989) A phylogenetic analysis of the mycoplasmas: basis for their classification. *J Bacteriol* **171:** 6455–6467.

Wenzel, R., and Herrmann, R. (1988) Repetitive DNA elements in *Mycoplasma pneumoniae*. *Nucl Acids Res* **16:** 8337–8350.

Wenzel, R., Pirkl, E., and Herrmann, R. (1992) Construction of an *Eco*RI restriction map of *Mycoplasma pneumoniae* and localization of selected genes. *Nucl Acids Res* **22:** 7289–7296.

Whitely, J.C., Muto, A., and Finch, L.R. (1991) A physical map for *M. capricolum* Cal.kid with loci for all known tRNA species. *Nucl Acids Res* **19:** 399–400.

Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., Bonfield, J. *et al.* (1994) 2.2 Mb of continuous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368:** 32–38.

Woese, C.R. (1987) Bacterial evolution. *Microbiol Rev* **51:** 221–271.

Woese, C.R., Stackebrandt, E., and Ludwig, W. (1984) What are mycoplasmas: the relationship of tempo and mode in bacterial evolution. *J Mol Evol.* **21:** 305–316.