

The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain R_{low}

Leka Papazisi,^{1,2} Timothy S. Gorton,^{1,2} Gerald Kutish,³ Philip F. Markham,⁴ Glenn F. Browning,⁴ Di Kim Nguyen,⁵ Steven Swartzell,⁶ Anup Madan,⁷ Greg Mahairas⁶ and Steven J. Geary^{1,2}

Correspondence
Steven J. Geary
geary@uconnvm.uconn.edu

^{1,2}Center of Excellence for Vaccine Research¹ and Department of Pathobiology and Veterinary Science², The University of Connecticut, Storrs, CT 06269-3089, USA

³Plum Island Animal Disease Center, US Department of Agriculture, Greenport, NY 11944, USA

⁴Department of Veterinary Science, University of Melbourne, Parkville, Victoria 3010, Australia

^{5,6}Department of Medicine, University of Washington⁵ and Regulome Corporation⁶, Seattle, WA 98195, USA

⁷The Institute for Systems Biology, Seattle, WA 98103-8904, USA

The complete genome of *Mycoplasma gallisepticum* strain R_{low} has been sequenced. The genome is composed of 996 422 bp with an overall G+C content of 31 mol%. It contains 742 putative coding DNA sequences (CDSs), representing a 91 % coding density. Function has been assigned to 469 of the CDSs, while 150 encode conserved hypothetical proteins and 123 remain as unique hypothetical proteins. The genome contains two copies of the rRNA genes and 33 tRNA genes. The origin of replication has been localized based on sequence analysis in the region of the *dnaA* gene. The *vlhA* family (previously termed pMGA) contains 43 genes distributed among five loci containing 8, 2, 9, 12 and 12 genes. This family of genes constitutes 10.4 % (103 kb) of the total genome. Two CDSs were identified immediately downstream of *gapA* and *crmA* encoding proteins that share homology to cytoadhesins GapA and CrmA. Based on motif analysis it is predicted that 80 genes encode lipoproteins and 149 proteins contain multiple transmembrane domains. The authors have identified 75 proteins putatively involved in transport of biomolecules, 12 transposases, and a number of potential virulence factors. The completion of this sequence has spawned multiple projects directed at defining the biological basis of *M. gallisepticum*.

Received 16 April 2003
Revised 9 June 2003
Accepted 16 June 2003

INTRODUCTION

Phylogenetic analyses indicate that mycoplasmas (class Mollicutes) have undergone a degenerative evolution from related, low G+C content, Gram-positive eubacteria (Rogers *et al.*, 1985; Woese *et al.*, 1980). The reduction of the mycoplasma genome has resulted in the loss of the cell wall and has limited the biosynthetic capabilities of these organisms. As a consequence of this loss of biosynthetic machinery, mycoplasmas are obligate parasites and rely

on the uptake of many essential molecules from their environment.

Mycoplasmas have long been considered model systems for defining the minimal set of genes required for a living cell (Morowitz, 1984). For this reason, it was not surprising when *Mycoplasma genitalium* (580 kb) was selected as one of the first targets for complete genome sequencing (Fraser *et al.*, 1995). Since this initial report, the genomes of four additional mycoplasmas have been sequenced, *Mycoplasma pneumoniae* (816 kb; Dandekar *et al.*, 2000; Himmelreich *et al.*, 1996), *Ureaplasma urealyticum* (752 kb; Glass *et al.*, 2000), *Mycoplasma pulmonis* (964 kb; Chambaud *et al.*, 2001) and *Mycoplasma penetrans* (1358 kb; Sasaki *et al.*, 2000). Theoretical and experimental approaches have estimated the minimum number of essential mycoplasma genes to be between 265 and 350 (Hutchison *et al.*, 1999; Mushegian & Koonin, 1996).

Abbreviations: CDS, coding DNA sequence; COGs, conserved orthologous groups.

The GenBank accession number for the sequence reported in this paper is AE015450.

The online version of this paper (at <http://mic.sgmjournals.org>) contains a supplementary table showing additional information concerning the features and homologies of the *M. gallisepticum* strain R_{low} genome.

Mycoplasma gallisepticum is an avian pathogen involved in chronic respiratory disease in chickens resulting in considerable economic losses in poultry production. Infection with this bacterium is spread by aerosol exposure and egg transmission. Outbreaks spread rapidly through flocks, establish chronic infections, and are difficult to control with antimicrobial therapy. The chronic nature of mycoplasma infections demonstrates a failure of the host immune system to deal effectively with these organisms. Antigenic variation of surface proteins allows *M. gallisepticum* to evade the host's immune response through the generation of escape variants (Glew *et al.*, 2000; Gorton & Geary, 1997; Levisohn *et al.*, 1995). Intracellular invasion and survival within eukaryotic cells by *M. gallisepticum* may contribute to this organism's resistance to the host's immune response and antimicrobial therapy (Winner *et al.*, 2000).

M. gallisepticum colonizes the respiratory system of chickens, but has also been isolated from the reproductive organs, brain and eyes of several avian species. Cytoadherence to the epithelial surfaces of these host tissues is a requirement for successful colonization. Research into the molecular mechanisms of *M. gallisepticum* cytoadherence has identified a coordinate action between the primary cytoadhesin, GapA, and at least one cytoadherence-related molecule, CrmA (Papazisi *et al.*, 2002).

Whole-genome sequencing establishes a solid foundation from which to begin a myriad of experimental projects directed at examining the pathogenic mechanisms of an organism. Here we present the complete sequence and initial analysis of the *M. gallisepticum* genome.

METHODS

Construction of *M. gallisepticum* genomic library. A clonal isolate, designated R_{low}c2, of the virulent *M. gallisepticum* strain R_{low} was chosen for sequencing. *M. gallisepticum* genomic DNA was prepared by the method of Hempstead (1990). Genomic DNA (4 µg) was sheared using a Hydroshear device (GeneMachines) to a mean size of 3–5 kb. Sheared ends were repaired using T4 DNA kinase and Klenow fragment of DNA polymerase (New England Biolabs). The reactions were stopped and the resulting repaired fragments purified using Qiaquick (Qiagen) columns according to the manufacturer's instructions. Plasmid pBlueScript SKII(+) (Stratagene) was used as the vector and prepared by digestion with *EcoRV* (New England Biolabs), treatment with alkaline phosphatase and electrophoresis in an agarose gel (0.6%). A single discrete band was excised from the gel and purified using a Gene Clean Column (Qbiogene). Ligation reactions were prepared using a 2:1 insert to vector ratio and T4 DNA ligase (New England Biolabs), then transformed into *Escherichia coli* DH10B Max competent cells (Invitrogen). Plates with 1000–3000 recombinant clones were subsequently picked and arrayed into Genetix square-well 384-well plates containing SOB medium (Difco) supplemented with 20% (v/v) glycerol. Following overnight (18 h) incubation these plates were sealed and stored at –80 °C until further use.

Sequencing and assembly. Each quadrant of a stock 384-well plate was freshly inoculated into a deep-well 96-well plate containing 1 ml 2 × YT medium (Difco) supplemented with 100 µg

carbenicillin ml⁻¹. These inoculated plates were shaken at 250 r.p.m. for 14 h at 37 °C followed by centrifugation at 3000 r.p.m. for 10 min. Plasmid preparations were processed automatically using an Eppendorf PerfectPrep-96 VAC Robotic Workstation. Sequencing reactions were carried out using 1 µl DNA (350 ng), 4 pmol forward and reverse primers (–21M13 and M13Reverse) and 3 µl ABI Prism Big Dye Terminators (Applied Biosystems/Perkin Elmer) in a total volume of 5 µl. Reactions were carried out in the ABI GeneAmp PCR System 9700 (Applied Biosystems/Perkin Elmer). The DNA from sequencing reactions was precipitated with 2-propanol, resuspended in 10 µl sterile distilled H₂O, and loaded onto an ABI 3700 capillary electrophoresis sequencer. Sequence data were assembled and analysed using the Phred/Phrap/Consed assembly package (Ewing & Green, 1998; Ewing *et al.*, 1998; Gordon *et al.*, 1998) and CAP3 (Huang & Madan, 1999). CAP3 uses forward and reverse sequence reads from the same plasmid sequences to place constraints on assembled sequence and is useful to break false joins and assure correct sequence assembly (Huang & Madan, 1999). Additional reactions were performed to close gaps in the assembly. Custom primers were designed near gaps and sequence was obtained from plasmid clones that spanned the gaps. Where no clone was available to span the gap, PCR products were generated using custom primers on genomic DNA and the subsequent PCR products were sequenced and added to the assembly. Sequencing coverage, of both strands, was 11 × (or 22 reads per kb), done to an error rate less than 1 per 10 kb.

The sequence data have been submitted to the NCBI database under accession number AE015450.

Identification of CDSs and annotation. Coding DNA sequences (CDSs), of at least 195 nt in length, were initially identified using ORF Finder software (with genetic code 4) provided by the National Center for Biotechnology information (NCBI; <http://www.ncbi.nlm.nih.gov/>). GLIMMER (Delcher *et al.*, 1999) was trained to recognize a potential CDS whose translated product had at least 33 amino acids and did not overlap more than 30 nt with neighbouring CDSs. BLAST (Zhang *et al.*, 1998) and FASTA (Pearson, 1999) searches were performed on the NCBI non-redundant protein databases, conserved orthologous groups (COGs) database (Tatusov *et al.*, 2001), and proteins from the mycoplasma genomes sequenced to date. A CDS was assigned a designation based on homology when its translated product presented a BLAST score of at least 100 in the COG database and 150 in the non-redundant database. Scores of 300 or more were considered definitive. MSP-Crunch (Sonnhammer & Durbin, 1994) was used to filter low-complexity sequences in order to eliminate spurious homologies. FASTrNA (<http://bioweb.pasteur.fr/seqanal/interfaces/fastrna.html>) (el-Mabrouk & Lisacek, 1996) was used for finding tRNA coding sequences. The PROSITE (Falquet *et al.*, 2002), Pfam (Bateman *et al.*, 2000), BLOCKS (Henikoff *et al.*, 1999) and TIGERfam (<http://www.tigr.org/>) databases were searched to identify motifs and other shared features of proteins. Specific motif searches were performed against the *M. gallisepticum* genome using FUZZPRO from the Jemboss software package (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Jemboss/index.html>). Protein topology and location were predicted using TMPred (Hofmann & Stoffel, 1993), MEMSAT (Jones *et al.*, 1994), PSORT (<http://psort.nibb.ac.jp/>) and SAPS (Brendel *et al.*, 1992).

RESULTS AND DISCUSSION

General features of the genome

The general features of the *M. gallisepticum* strain R_{low} genome are shown in Fig. 1 and Table 1. The online version of this paper (at <http://mic.sgmjournals.org>) contains a

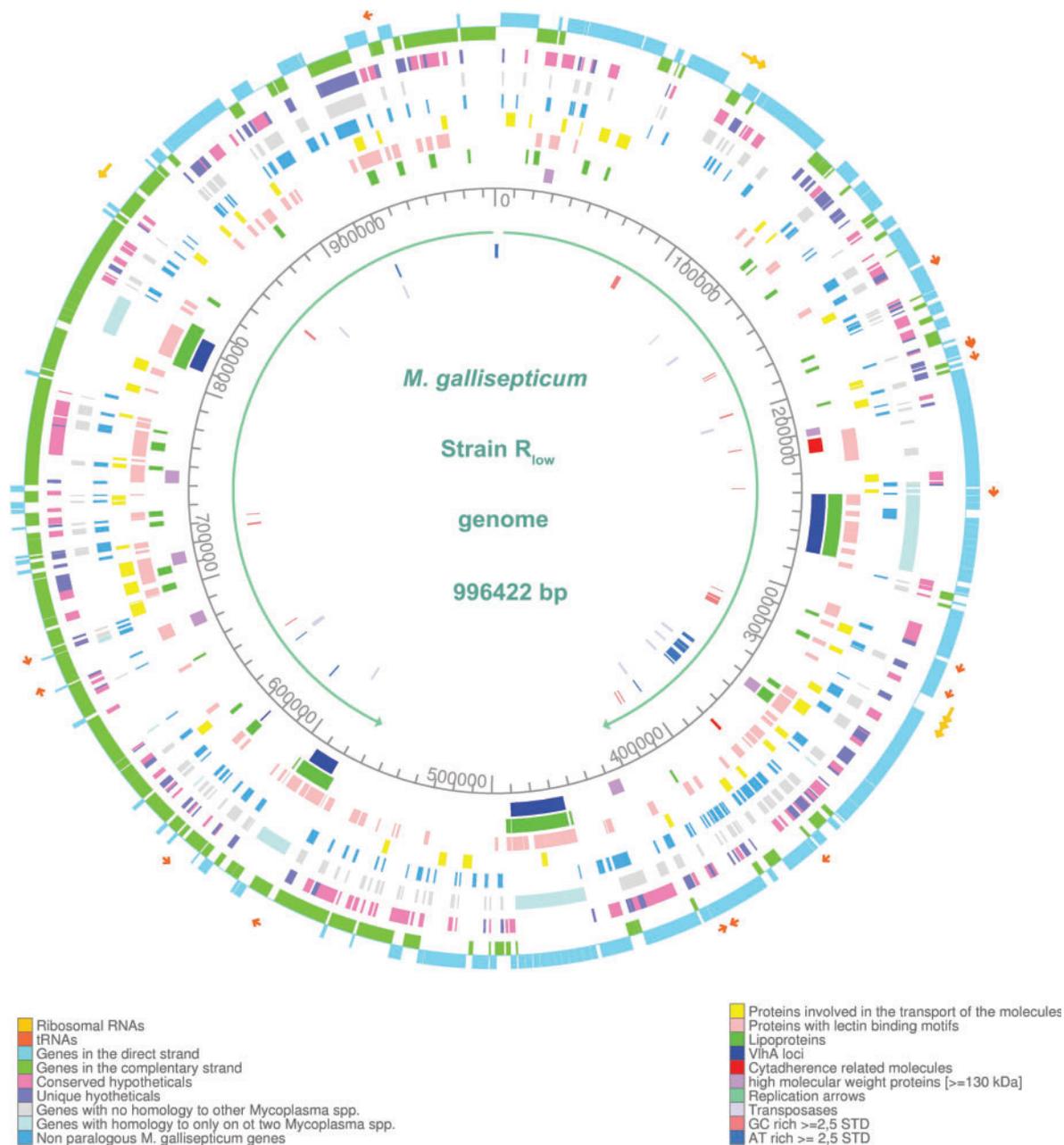


Fig. 1. Circular representation of the *M. gallisepticum* strain R_{low} genome. This figure was generated using GenVision (DNASar).

supplementary table with additional information concerning the features and homologies of the genome. The genome is composed of 996 422 bp with an overall G + C content of 31 mol%. It contains 742 putative coding DNA sequences (CDSs), representing a 91 % coding density. Function has been assigned to 469 of the CDSs, while 150 are conserved hypothetical proteins and 123 remain as hypothetical proteins. The mean CDS length was 1206 nt (108–5928 nt) and the mean CDS G + C content was 32 mol% (17–45 mol%). The mean G + C content of the

third nucleotide position in codons was 24 %. The predicted codon usage within *M. gallisepticum* varies with statistical significance from the other mycoplasma genomes sequenced to date, with the exception of *M. genitalium* (data not shown). A set of 33 tRNA genes was identified, corresponding to all amino acids. A single polypeptide release factor (*prfA*) was identified, consistent with the use of only UAA and UAG as stop codons. The mean occurrence of the amino acid tryptophan was 3.6 per CDS and the usage of the codons TGA and TGG was 2.8 and 0.8 per CDS,

Table 1. General features of *Mycoplasma* species

	<i>M. gallisepticum</i>		<i>M. genitalium</i> *		<i>M. pneumoniae</i> †		<i>M. pulmonis</i> ‡		<i>U. urealyticum</i> §		<i>M. penetrans</i>	
Genome size (bp)	996 422		580 074		816 394		963 879		751 719		1 358 633	
G + C content (mol%)	31.5		32		40		26.6		25.5		25.7	
Sets of 5S 16S and 23S rRNA	2		1		1		1		2		1	
tRNA	33		36		37		29		30		30	
Total no. of predicted CDSs	742		484		689		782		614		1038	
Functional category in COGs												
[J] Translation, ribosomal structure and biogenesis	103	14 %	99	20 %	100	15 %	102	13 %	102	17 %	108	10 %
[K] Transcription	14	2 %	14	3 %	14	2 %	18	2 %	18	3 %	23	2 %
[L] DNA replication, recombination and repair	53	7 %	41	8 %	57	8 %	72	9 %	56	9 %	78	8 %
[D] Cell division and chromosome partitioning	7	1 %	5	1 %	5	1 %	4	1 %	4	1 %	11	1 %
[O] Posttranslational modification, protein turnover, chaperones	33	4 %	20	4 %	20	3 %	18	2 %	19	3 %	26	3 %
[M] Cell envelope biogenesis, outer membrane	8	1 %	10	2 %	10	1 %	5	1 %	4	1 %	13	1 %
[N] Cell motility and secretion	9	1 %	11	2 %	12	2 %	13	2 %	15	2 %	12	1 %
[P] Inorganic ion transport and metabolism	14	2 %	17	4 %	17	2 %	15	2 %	28	5 %	22	2 %
[T] Signal transduction mechanisms	4	1 %	4	1 %	4	1 %	3	0 %	4	1 %	4	0 %
[C] Energy production and conversion	28	4 %	20	4 %	20	3 %	29	4 %	17	3 %	31	3 %
[G] Carbohydrate transport and metabolism	25	3 %	24	5 %	34	5 %	58	7 %	14	2 %	50	5 %
[E] Amino acid transport and metabolism	21	3 %	15	3 %	24	3 %	23	3 %	20	3 %	26	3 %
[F] Nucleotide transport and metabolism	25	3 %	21	4 %	21	3 %	24	3 %	21	3 %	39	4 %
[H] Coenzyme metabolism	11	1 %	13	3 %	14	2 %	12	2 %	9	1 %	12	1 %
[I] Lipid metabolism	11	1 %	8	2 %	9	1 %	8	1 %	8	1 %	15	1 %
[Q] Secondary metabolites biosynthesis, transport and catabolism	13	2 %	4	1 %	5	1 %	10	1 %	7	1 %	20	2 %
[R] General function prediction only	51	7 %	44	9 %	48	7 %	49	6 %	41	7 %	95	9 %
[S] Function unknown	26	4 %	14	3 %	15	2 %	26	3 %	22	4 %	25	2 %
(-) not in COGs	286	39 %	100	21 %	260	38 %	293	37 %	205	33 %	428	41 %

*Fraser *et al.* (1995).†Himmelreich *et al.* (1996); Dandekar *et al.* (2000).‡Chambaud *et al.*, (2001).§Glass *et al.* (2000).||Sasaki *et al.* (2002).

respectively. As previously reported (Chen & Finch, 1989; Gorton *et al.*, 1995; Scamrov & Beabealashvilli, 1991), the *M. gallisepticum* genome contains two copies of the rRNA genes. One set is organized as an operon, with adjacent 16S, 23S and 5S genes; and a second copy of the 16S rRNA gene lies 221 kb upstream of the 23S and 5S rRNA genes.

Origin of replication

The origin of replication (*oriC*) for most bacteria is located in the region of the *dnaA* gene. Comparative analysis of the *oriC* regions of sequenced mycoplasma genomes predicts putative DnaA boxes in the area surrounding the *dnaA* gene (Cordova *et al.*, 2002). Functional mollicute *oriC* regions have been identified in *Spiroplasma citri* and *M. pulmonis* (Cordova *et al.*, 2002; Ye *et al.*, 1994).

In silico analysis localized the predicted *oriC* region of *M. gallisepticum* based on several criteria. First, despite an apparent overall lack of conserved gene order in the mollicute *oriC* region (Cordova *et al.*, 2002), the gene order in the *oriC* region does appear to be conserved within the phylogenetic cluster containing *M. pneumoniae*, *M. genitalium* and *M. gallisepticum* (Cordova *et al.*, 2002; Hilbert *et al.*, 1996; Zou & Dybvig, 2002). These three mycoplasmas have the *gyrA*, *gyrB*, *dnaJ*, *dnaN* and *soj* genes upstream of the *dnaA* gene, and ABC transporter genes, *rpl34* and *rpnA*, downstream of *dnaA*. As in *M. pneumoniae*, there is a unique hypothetical coding region immediately upstream of the *dnaA* gene in *M. gallisepticum*. Second, we identified putative DnaA boxes in the vicinity of the *dnaA* gene (Fig. 2). The region around *dnaA* was searched for 9 nt long sequences resembling a consensus sequence (5'-TTWTMHAMA-3') based on DnaA box sequences from *M. pulmonis*, *M. capricolum* and *M. pneumoniae* (Chambaud *et al.*, 2001; Cordova *et al.*, 2002; Hilbert *et al.*, 1996; Zou & Dybvig, 2002). Finally, the region between *dnaN* and *soj* contains higher than average AT base pair frequency (80%) and contains AT-rich repeats. These criteria are considered to be characteristic of the origin of replication in prokaryotes (Baker & Wickner, 1992).

***vlhA* gene family**

Perhaps the best-described example of a multi-gene family in mycoplasmas is that encoding the VlhA or pMGA lipoproteins (Baseggio *et al.*, 1996; Liu *et al.*, 1998; Markham *et al.*, 1993). It has been well established that *M. gallisepticum* generally expresses a single member of the family at any one time (Glew *et al.*, 1995) and that the specific gene expressed can be influenced by growth in the presence of cognate antibody (Markham *et al.*, 1998). The probable role of this family in generating antigenic variation has been demonstrated in infected chickens, with both phase variation demonstrated during the acute stages of disease, and antigenic switching during the chronic stages (Glew *et al.*, 2000). These findings have led to the suggestion that the principal function of this family is to generate antigenic diversity and hence facilitate immune evasion during chronic infections. These genes formed the largest paralogous gene family in the *M. gallisepticum* genome.

To establish consistency, and in accordance with standard nomenclature, we have annotated this family of genes as *vlhA* (Noormohammadi *et al.*, 1998). This family contains 43 genes, constituting a total of 103 kb or 10.4% of the genome. The 43 *vlhA* genes are distributed among five loci containing 8, 2, 9, 12 and 12 genes, respectively, and have been numbered according to their locus and position (e.g. *vlhA1.01*) (Fig. 3). Of the 43 genes, 38 possess the signature *vlhA* gene features, which include a GAA repeat motif 5' of a GTG start codon and conserved regions flanking the start codon (Markham *et al.*, 1994). The sequence identity among these 38 genes ranges between 41 and 99%. Five of the genes (1.05, 2.01, 2.02, 5.01 and 5.02) possess sequence homology to *vlhA* but lack the *vlhA* signature motifs. The genes within each of the five loci are in the same transcriptional orientation, with the exception of *vlhA1.05*. In addition to the lack of *vlhA* signature motifs, a putative transposase (MGA_0073) has been identified adjacent to *vlhA1.05*, suggesting that rearrangements by a transposon may account for the differences observed in this region. Five of the genes (*vlhA1.01*, 1.08, 3.01, 4.03 and 5.01) have been

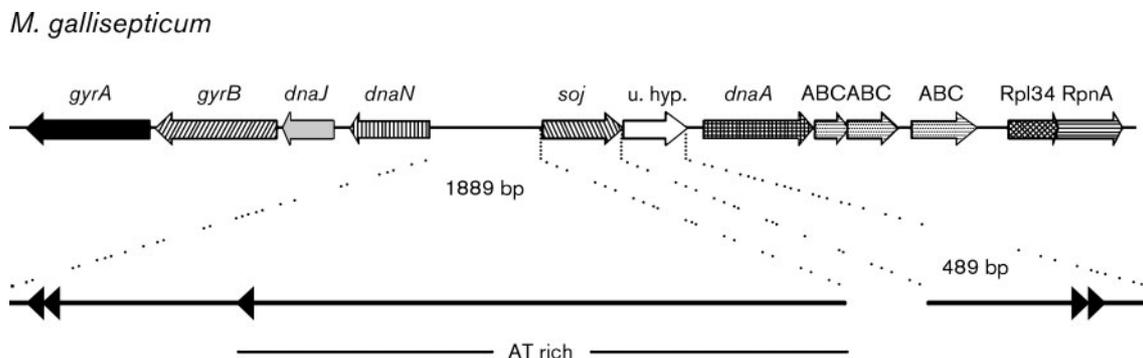


Fig. 2. Organization of the putative origin of replication of *M. gallisepticum*. The black triangles represent putative DnaA boxes with 9 out of 9 matches to the consensus sequence 5'-TTWTMHAMA-3'.

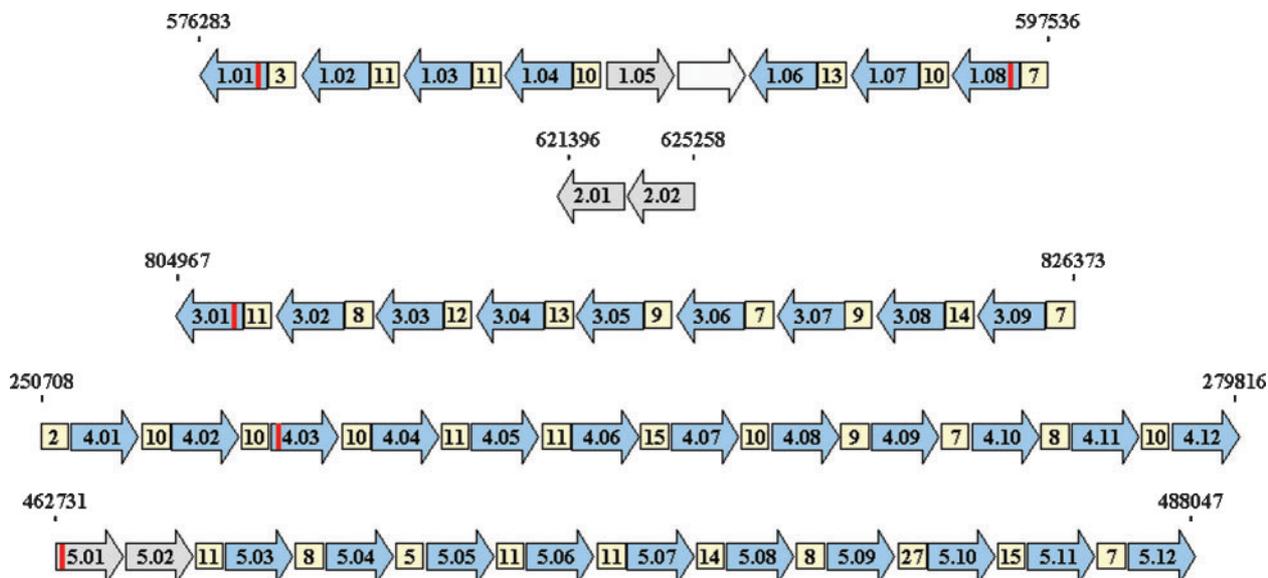


Fig. 3. *M. gallisepticum* strain R_{low} *vlhA* gene loci. The 43 *vlhA* genes (blue and grey arrows) are distributed among five loci containing 8, 2, 9, 12 and 12 genes, respectively, and have been numbered according to their locus and position (e.g.: *vlhA*1.01). Of the 43 genes, 38 (blue arrows with yellow boxes) possess the signature *vlhA* gene features, which include a GAA repeat motif 5' of a GTG start codon and conserved regions flanking the start codon. The numbers within the yellow boxes denote the number of GAA repeats (e.g. *vlhA*3.03 has 12 GAA repeats). Grey arrows represent *vlhA* genes that lack the *vlhA* signature motifs. The empty arrow represents a putative transposase (MGA_0073). Red vertical bars within the arrows represent genes that have been interrupted by mutations that introduce shifts in the reading frames. The six-digit numbers correspond to the map locations of each locus.

interrupted by mutations that introduce shifts in the reading frames.

Transcription of *vlhA* genes has been shown to correlate with the occurrence of 12 repeats within the GAA motif (Glew *et al.*, 1998; Liu *et al.*, 2000). The size of the GAA motifs ranges between 2 and 27 repeats, with a median of 9 repeats. Of the 38 *vlhA* genes possessing GAA motifs, only *vlhA*3.03 possesses the 12 repeats reportedly required for transcription. Interestingly, of all the *vlhA* genes sequenced from different *M. gallisepticum* strains, four of the five that are preceded by 12 GAA repeats are predicted to encode proteins that share very high levels of amino acid sequence identity (95–98%). That different strains express near-identical genes from the large repertoire of *vlhA* genes when cultured *in vitro* suggests that the product of these very closely related genes may have a specific function that is required during growth *in vitro*.

Adhesin-related proteins

A paralogous family of five adhesin-like proteins has been identified in the *M. gallisepticum* genome. Papazisi *et al.* (2002) have recently complemented a cytodherence-deficient, $\text{GapA}^- \text{CrmA}^-$ isolate of *M. gallisepticum* strain R with a functional *gapAcrmA* operon (MGA_0934 and MGA_0939), and have demonstrated that expression of both genes is required for cytodherence and pathogenesis of

M. gallisepticum. The *gapA* gene was initially characterized as the *M. gallisepticum* orthologue of the *M. pneumoniae* cytodhesin P1 (Goh *et al.*, 1998). Attachment inhibition assays using anti-GapA Fab fragments established the role of GapA in *M. gallisepticum* cytodherence (Goh *et al.*, 1998). A comparative analysis of virulent and avirulent isolates of *M. gallisepticum* strain R identified a frameshift mutation, resulting in a premature termination codon, in the *gapA* gene of an avirulent, GapA-deficient isolate (Papazisi *et al.*, 2000). With the initial characterization of *gapA*, a potential CDS was identified and partially sequenced immediately downstream of *gapA* (Goh *et al.*, 1998). During the characterization of a cytodherence-deficient isolate of *M. gallisepticum* strain R, Papazisi *et al.* (2000) matched this partial CDS with amino acid data from a 116 kDa protein that was expressed in a cytodherence-positive isolate of strain R, but not expressed in the cytodherence-deficient isolate. This CDS (designated *crmA*) was sequenced and found to encode a protein that exhibits 41% amino acid identity with the *M. pneumoniae* and *M. genitalium* ORF6 proteins (Papazisi *et al.*, 2000). These proteins have been shown to play an accessory role in *M. pneumoniae* cytodherence (Krause & Balish, 2001; Krause *et al.*, 1982; Layh-Schmitt & Harkenthal, 1999; Seto *et al.*, 2001).

Immediately downstream of the *gapAcrmA* operon there are two CDSs (MGA_0943 and MGA_0945; designated *crmB* and C) that, although not showing significant nucleotide

homology, are predicted to encode proteins sharing homology to GapA and CrmA. Analysis shows that the majority of the similarity lies within the C-terminal region of these four proteins. A fifth paralogue (MGA_1117) shares similarity with the N-terminus of GapA and CrmA.

Membrane-associated proteins

Motif analysis of the *M. gallisepticum* genome has predicted a large repertoire of membrane-associated proteins. Of these proteins, 149 contain multiple transmembrane domains and 18 possess ten or more transmembrane domains. These 18 proteins include molecules involved in amino acid transport (PotE), phosphate transport (Pts) and protein translocation (SecY), as well as five conserved hypothetical and two unique hypothetical proteins. Based on the prokaryotic membrane lipoprotein lipid attachment site motif (PROSITE PS00013), 80 CDSs are predicted to encode lipoproteins (10.8% of all CDSs). The second largest paralogous gene family in *M. gallisepticum* consists of 24 ATP-binding proteins belonging to the ABC transporter superfamily. This family accounts for nearly one-third of the 75 proteins predicted to be involved in transport of biomolecules.

Components of the membrane protein secretion and translocation pathways were found in *M. gallisepticum*. The Sec pathway of *M. gallisepticum* consists of SecA (MGA_670), SecE (MGA_0474), SecY (MGA_0740), YidC (MGA_0631), trigger factor (Tig, MGA_1297) and DnaK (MGA_0279). The signal recognition particle pathway contains FtsY (MGA_0919) and Ffh (MGA_1143). Although the Sec pathway is incomplete when compared to the secretion pathway in *E. coli* (Cao & Saier, 2003; Driessen *et al.*, 2001; van Wely *et al.*, 2001), these proteins are consistent with the minimal set of secretion and translocation proteins proposed by Mushegian & Koonin (1996).

Signal peptidase I (LepB; SPase I) was identified in *M. gallisepticum* (MGA_1091) and was found to possess two motifs similar (one mismatch) to PROSITE SPase I motifs (PS00501; PS00761). The only other known SPase I orthologue in mycoplasmas is in *M. pulmonis* (MYPU_6300). Phylogenetic analysis using FASTA shows that these two mycoplasma signal peptidases cluster separately from other prokaryotic LepB proteins (data not shown).

Protoprotein signal peptidase (LspA; SPase II) was also identified in *M. gallisepticum* (MGA_0997). Orthologues of this protein have been identified in all mycoplasmas sequenced to date. Although these proteins have been annotated as 'signal peptidase type 2' based on the homology that they share with their prokaryotic homologues, all of the mycoplasma LspAs appear to form a divergent phylogenetic cluster using FASTA analysis (data not shown). Additionally, the SPase II motif for all mycoplasmas is

similar, but not identical, to the PROSITE SPase II motif (data not shown).

A potential cleavage motif (PXXR₀₋₅SS, with 1 mismatch; for review see Krause, 1998) was found in GapA, 26 of the VlhA proteins, and three high-molecular-mass proteins (MGA_0306, MGA_0928 and MGA_0205), which have homology to *M. pneumoniae* proteins HMW1, HMW3 and P200, respectively.

Collectively, these findings indicate that post-translational modification of membrane proteins in *M. gallisepticum*, and perhaps other mycoplasma species, occurs in a divergent manner from that of other prokaryotes.

Transposases

Twelve putative transposases were identified within the *M. gallisepticum* genome. Ten of these have homology to transposases found in both Gram-positive and Gram-negative bacteria, while two (MGA_1109 and MGA_1329) have homology with the transposases found in *M. mycoides* subsp. *mycoides* Small Colony type and *M. hyopneumoniae*. In eight cases the presence of a transposase was found to be associated with the rearrangement or disruption of a neighbouring gene. An example of this is seen in the rearrangement found in the *vlhA* locus 1 (Fig. 3). Six of the putative transposases have a motif similar to PROSITE PS01007 'transposase/mutator' (one mismatch). However, it remains to be proven if any of the 12 *M. gallisepticum* predicted transposases are part of any active mobile element.

PvpA

PvpA is a putative haemagglutinin that undergoes phase-variable expression and exhibits size variation among strains of *M. gallisepticum* (Boguslavsky *et al.*, 2000; Liu *et al.*, 2001; Yogev *et al.*, 1994). Analysis of the *pvpA* gene from the genomic sequence revealed a duplication of 37 nt compared to the *M. gallisepticum* strain R sequence submitted as GenBank accession number AF224059 (Boguslavsky *et al.*, 2000). The region encompassing the *pvpA* gene has been amplified by PCR and sequenced to confirm this duplication. The 37 nt duplication results in a frameshift and predicted premature termination of PvpA expression.

Virulence factors

Putative virulence factors in *M. gallisepticum* were identified based on protein motif analysis. Nine of the 20 proteases were found to possess between one and five transmembrane domains; three of them appear to be zinc metalloproteases. In addition, three nucleases and a lipase-like (GDLSL motif) protein were found to possess between one and four transmembrane domains. We found 133 membrane-associated proteins containing putative lectin-binding motifs (Elgavish & Shaanan, 1997; Loris *et al.*, 1998); 51 of them are lipoproteins and 34 belonged to the VlhA family. This suggests that there are a number of membrane-associated proteins that may bind sugar moieties for the

purpose of nutrient uptake or cytodherence. Two proteins, MGA_0090 and MGA_0091, showed similarity to phospholipid-binding proteins. These findings indicate that *M. gallisepticum*, in addition to GapA, CrmA and the VlhAs, has a wide array of proteins predicted to be involved in binding biomolecules. A motif similar to the aerolysin motif (PROSITE PS00274, one mismatch) was found in four membrane proteins: GapA (MGA_0934), two hypotheticals (MGA_0226 and MGA_1162) and VlhA5.04 (MGA_1243). Although this remains speculative, it may help to explain the reported invasiveness of GapA⁺ variants (Much *et al.*, 2002; Winner *et al.*, 2000). Proteins MGA_0313 and MGA_0931 were found to contain a motif similar to the staphylococcal/streptococcal pyrogenic exotoxin signature motif (PROSITE PS00277). A number of heat-shock proteins were identified, including GroEL (MGA_0152), GroES (MGA_0153), Clp (MGA_0178), DnaK (MGA_0279), GrpE (MGA_1232), and seven DnaJ-class proteins (MGA_0617, MGA_0877, MGA_0885, MGA_1131, MGA_1135, MGA_1228, MGA_1324). The number of DnaJ-class proteins is higher than in other mycoplasmas (e.g. *M. pneumoniae* and *M. genitalium* each have three DnaJ-class proteins).

Regulatory proteins

The reduction of the mycoplasma genome has resulted in biosynthetic limitations affecting alterations in regulatory pathways. No component of any 'classical' bacterial two-component regulatory system was found in *M. gallisepticum*. However, a search for motifs identified several potential regulatory proteins. It is known that regulatory proteins that bind DNA possess helix–turn–helix (HTH) motifs. We found 21 proteins (excluding sigma 70, recombinases, helicases, and other enzymes involved in nucleic acid base modification) that contained HTH motifs similar to the PROSITE AraC (PS00041), LysR (PS00044), GntR (PS00043) and LuxR (PS00622) motifs. Another protein, MGA_1295, was found to share homology with the Fur-like class of regulatory proteins involved in iron/zinc uptake (Escolar *et al.*, 1998). Two multi-transmembrane proteins, MGA_1037 and MGA_0337, were found to share homology with chemoreceptors. A motif similar to the PROSITE sigma 54 interaction domain (PS00675) was found in several proteins, including PhnL and HatB (MGA_0655 and MGA_1018, respectively), ATPases that are components of two different ABC transporter systems), UvrC (MGA_1269; recombinase), OppF (MGA_0218 and MGA_0230; oligopeptide/dipeptide uptake transporter), Gmk (MGA_0462; guanylate kinase) and a HprK (MGA_0599; predicted serine kinase). This finding indicates that, in the absence of identified alternative sigma factors in mycoplasmas, proteins exhibiting similar function may exist.

Conclusions

M. gallisepticum strain R contains a 996 kb genome with 742 CDSs. Among the 742 CDSs, function has been assigned to 469 genes, while approximately one-third of the genes

remain undefined in terms of function. Nearly 17 % of the genes appear to be unique to *M. gallisepticum*.

We identified all of the members of the pMGA family, and have renamed this family *vlhA* in accordance with standard nomenclature. The *vlhA* family, totalling 43 genes, constitutes the largest paralogous set of genes in *M. gallisepticum*. Our analysis has identified three putative adhesin-related molecules that share homology with the GapA and CrmA cytodherence molecules of *M. gallisepticum*. We have found that 10 % of the CDSs are putative lipoproteins, and nearly 20 % of the CDSs contain multiple transmembrane domains. With 24 CDSs, ABC transporter molecules make up the second-largest paralogous family in *M. gallisepticum*. A large percentage of the *M. gallisepticum* genome is devoted to membrane-associated molecules.

The completion of this sequence has spawned multiple projects directed at defining the biological basis of *M. gallisepticum*. We have developed microarray chips covering the entire genome of *M. gallisepticum* and are currently performing experiments to analyse gene expression patterns. Experiments are currently under way to examine the proteome profile of *M. gallisepticum*. This sequence is greatly aiding our transposon mutagenesis studies of mycoplasma pathogenicity.

The most salient feature arising from this sequencing project is the large number (36 %) of CDSs annotated as unique (126) or conserved (151) hypothetical, a total of 277 CDSs whose functions remain speculative at this time. This emphasizes the vast amount we do not know about the inner workings of this pathogen. Further analysis of these CDSs and the functions of the proteins that they encode will add significantly to our ever-expanding body of knowledge regarding virulence mechanisms in *M. gallisepticum*.

ACKNOWLEDGEMENTS

This work was supported by USDA grant 58-1940-0-007 (S. J. G.) and the Egg Industry and Chicken Meat Programmes of the Rural Industries Research and Development Corporation, Australia (G. F. B.). This work was also supported by the Center of Excellence for Vaccine Research (CEVR #83).

REFERENCES

- Baker, T. A. & Wickner, S. H. (1992). Genetics and enzymology of DNA replication in *Escherichia coli*. *Annu Rev Genet* **26**, 447–477.
- Baseggio, N., Glew, M. D., Markham, P. F., Whithear, K. G. & Browning, G. F. (1996). Size and genomic location of the pMGA multigene family of *Mycoplasma gallisepticum*. *Microbiology* **142**, 1429–1435.
- Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res* **28**, 263–266.
- Boguslavsky, S., Menaker, D., Lysnyansky, I., Liu, T., Levisohn, S., Rosengarten, R., Garcia, M. & Yogev, D. (2000). Molecular characterization of the *Mycoplasma gallisepticum* *pvpA* gene which

- encodes a putative variable cytoadhesin protein. *Infect Immun* **68**, 3956–3964.
- Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E. & Karlin, S. (1992).** Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci U S A* **89**, 2002–2006.
- Cao, T. B. & Saier, M. H., Jr (2003).** The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. *Biochim Biophys Acta* **1609**, 115–125.
- Chambaud, I., Heilig, R., Ferris, S. & 9 other authors (2001).** The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res* **29**, 2145–2153.
- Chen, X. & Finch, L. R. (1989).** Novel arrangement of rRNA genes in *Mycoplasma gallisepticum*: separation of the 16S gene of one set from the 23S and 5S genes. *J Bacteriol* **171**, 2876–2878.
- Cordova, C. M., Lartigue, C., Sirand-Pugnet, P., Renaudin, J., Cunha, R. A. & Blanchard, A. (2002).** Identification of the origin of replication of the *Mycoplasma pulmonis* chromosome and its use in *oriC* replicative plasmids. *J Bacteriol* **184**, 5426–5435.
- Dandekar, T., Huynen, M., Regula, J. T. & 10 other authors (2000).** Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res* **28**, 3278–3288.
- Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999).** Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636–4641.
- Driessen, A. J., Manting, E. H. & van der Does, C. (2001).** The structural basis of protein targeting and translocation in bacteria. *Nat Struct Biol* **8**, 492–498.
- Elgavish, S. & Shaanan, B. (1997).** Lectin-carbohydrate interactions: different folds, common recognition principles. *Trends Biochem Sci* **22**, 462–467.
- el-Mabrouk, N. & Lisacek, F. (1996).** Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the yeast genome. *J Mol Biol* **264**, 46–55.
- Escolar, L., Perez-Martin, J. & de Lorenzo, V. (1998).** Binding of the fur (ferric uptake regulator) repressor of *Escherichia coli* to arrays of the GATAAT sequence. *J Mol Biol* **283**, 537–547.
- Ewing, B. & Green, P. (1998).** Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186–194.
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998).** Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175–185.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002).** The PROSITE database, its status in 2002. *Nucleic Acids Res* **30**, 235–238.
- Fraser, C. M., Gocayne, J. D., White, O. & 26 other authors (1995).** The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
- Glass, J. I., Lefkowitz, E. J., Glass, J. S., Heiner, C. R., Chen, E. Y. & Cassell, G. H. (2000).** The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407**, 757–762.
- Glew, M. D., Markham, P. F., Browning, G. F. & Walker, I. D. (1995).** Expression studies on four members of the pMGA multigene family in *Mycoplasma gallisepticum* S6. *Microbiology* **141**, 3005–3014.
- Glew, M. D., Baseggio, N., Markham, P. F., Browning, G. F. & Walker, I. D. (1998).** Expression of the pMGA genes of *Mycoplasma gallisepticum* is controlled by variation in the GAA trinucleotide repeat lengths within the 5' noncoding regions. *Infect Immun* **66**, 5833–5841.
- Glew, M. D., Browning, G. F., Markham, P. F. & Walker, I. D. (2000).** pMGA phenotypic variation in *Mycoplasma gallisepticum* occurs *in vivo* and is mediated by trinucleotide repeat length variation. *Infect Immun* **68**, 6027–6033.
- Goh, M. S., Gorton, T. S., Forsyth, M. H., Troy, K. E. & Geary, S. J. (1998).** Molecular and biochemical analysis of a 105 kDa *Mycoplasma gallisepticum* cytoadhesin (GapA). *Microbiology* **144**, 2971–2978.
- Gordon, D., Abajian, C. & Green, P. (1998).** Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195–202.
- Gorton, T. S. & Geary, S. J. (1997).** Antibody-mediated selection of a *Mycoplasma gallisepticum* phenotype expressing variable proteins. *FEMS Microbiol Lett* **155**, 31–38.
- Gorton, T. S., Goh, M. S. & Geary, S. J. (1995).** Physical mapping of the *Mycoplasma gallisepticum* S6 genome with localization of selected genes. *J Bacteriol* **177**, 259–263.
- Hempstead, P. G. (1990).** An improved method for the rapid isolation of chromosomal DNA from *Mycoplasma* spp. *Can J Microbiol* **36**, 59–61.
- Henikoff, S., Henikoff, J. G. & Pietrokovski, S. (1999).** Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15**, 471–479.
- Hilbert, H., Himmelreich, R., Plagens, H. & Herrmann, R. (1996).** Sequence analysis of 56 kb from the genome of the bacterium *Mycoplasma pneumoniae* comprising the *dnaA* region, the *atp* operon and a cluster of ribosomal protein genes. *Nucleic Acids Res* **24**, 628–639.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B. C. & Herrmann, R. (1996).** Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**, 4420–4449.
- Hofmann, K. & Stoffel, W. (1993).** TMbase – a database of membrane spanning protein segments. *Biol Chem Hoppe-Seyler* **347**, 166–173.
- Huang, X. & Madan, A. (1999).** CAP3: a DNA sequence assembly program. *Genome Res* **9**, 868–877.
- Hutchison, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O. & Venter, J. C. (1999).** Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994).** A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049.
- Krause, D. C. (1998).** *Mycoplasma pneumoniae* cytoadherence: organization and assembly of the attachment organelle. *Trends Microbiol* **6**, 15–18.
- Krause, D. C. & Balish, M. F. (2001).** Structure, function, and assembly of the terminal organelle of *Mycoplasma pneumoniae*. *FEMS Microbiol Lett* **198**, 1–7.
- Krause, D. C., Leith, D. K., Wilson, R. M. & Baseman, J. B. (1982).** Identification of *Mycoplasma pneumoniae* proteins associated with hemadsorption and virulence. *Infect Immun* **35**, 809–817.
- Layh-Schmitt, G. & Harkenthal, M. (1999).** The 40- and 90-kDa membrane proteins (ORF6 gene product) of *Mycoplasma pneumoniae* are responsible for the tip structure formation and P1 (adhesin) association with the Triton shell. *FEMS Microbiol Lett* **174**, 143–149.
- Levisohn, S., Rosengarten, R. & Yogeve, D. (1995).** *In vivo* variation of *Mycoplasma gallisepticum* antigen expression in experimentally infected chickens. *Vet Microbiol* **45**, 219–231.
- Liu, L., Payne, D. M., van Santen, V. L., Dybvig, K. & Panangala, V. S. (1998).** A protein (M9) associated with monoclonal antibody-mediated agglutination of *Mycoplasma gallisepticum* is a member of the pMGA family. *Infect Immun* **66**, 5570–5575.

- Liu, L., Dybvig, K., Panangala, V. S., van Santen, V. L. & French, C. T. (2000). GAA trinucleotide repeat region regulates M9/pMGA gene expression in *Mycoplasma gallisepticum*. *Infect Immun* **68**, 871–876.
- Liu, T., Garcia, M., Levisohn, S., Yogev, D. & Kleven, S. H. (2001). Molecular variability of the adhesin-encoding gene *pvpA* among *Mycoplasma gallisepticum* strains and its application in diagnosis. *J Clin Microbiol* **39**, 1882–1888.
- Loris, R., Hamelryck, T., Bouckaert, J. & Wyns, L. (1998). Legume lectin structure. *Biochim Biophys Acta* **1383**, 9–36.
- Markham, P. F., Glew, M. D., Whithear, K. G. & Walker, I. D. (1993). Molecular cloning of a member of the gene family that encodes pMGA, a hemagglutinin of *Mycoplasma gallisepticum*. *Infect Immun* **61**, 903–909.
- Markham, P. F., Glew, M. D., Sykes, J. E., Bowden, T. R., Pollocks, T. D., Browning, G. F., Whithear, K. G. & Walker, I. D. (1994). The organisation of the multigene family which encodes the major cell surface protein, pMGA, of *Mycoplasma gallisepticum*. *FEBS Lett* **352**, 347–352.
- Markham, P. F., Glew, M. D., Browning, G. F., Whithear, K. G. & Walker, I. D. (1998). Expression of two members of the pMGA gene family of *Mycoplasma gallisepticum* oscillates and is influenced by pMGA-specific antibodies. *Infect Immun* **66**, 2845–2853.
- Morowitz, H. J. (1984). The completeness of molecular biology. *Isr J Med Sci* **20**, 750–753.
- Much, P., Winner, F., Stipkovits, L., Rosengarten, R. & Citti, C. (2002). *Mycoplasma gallisepticum*: influence of cell invasiveness on the outcome of experimental infection in chickens. *FEMS Immunol Med Microbiol* **34**, 181–186.
- Mushegian, A. R. & Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**, 10268–10273.
- Noormohammadi, A. H., Markham, P. F., Duffy, M. F., Whithear, K. G. & Browning, G. F. (1998). Multigene families encoding the major hemagglutinins in phylogenetically distinct mycoplasmas. *Infect Immun* **66**, 3470–3475.
- Papazisi, L., Troy, K. E., Gorton, T. S., Liao, X. & Geary, S. J. (2000). Analysis of cytoadherence-deficient, GapA-negative *Mycoplasma gallisepticum* strain R. *Infect Immun* **68**, 6643–6649.
- Papazisi, L., Frasca, S., Jr, Gladd, M., Liao, X., Yogev, D. & Geary, S. J. (2002). GapA and CrmA coexpression is essential for *Mycoplasma gallisepticum* cytoadherence and virulence. *Infect Immun* **70**, 6839–6845.
- Pearson, W. R. (1999). Flexible similarity searching with the FASTA3 program package. In *Bioinformatics Methods and Protocols*, pp. 185–219. Totowa, NJ: Humana Press.
- Rogers, M. J., Simmons, J., Walker, R. T. & 8 other authors (1985). Construction of the mycoplasma evolutionary tree from 5S rRNA sequence data. *Proc Natl Acad Sci U S A* **82**, 1160–1164.
- Sasaki, Y., Ishikawa, J., Yamashita, A. & 8 other authors (2002). The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. *Nucleic Acids Res* **30**, 5293–5300.
- Scamrov, A. & Beabealashvili, R. (1991). *Mycoplasma gallisepticum* strain S6 genome contains three regions hybridizing with 16 S rRNA and two regions hybridizing with 23S and 5S rRNA. *FEBS Lett* **291**, 71–74.
- Seto, S., Layh-Schmitt, G., Kenri, T. & Miyata, M. (2001). Visualization of the attachment organelle and cytoadherence proteins of *Mycoplasma pneumoniae* by immunofluorescence microscopy. *J Bacteriol* **183**, 1621–1630.
- Sonnhammer, E. L. & Durbin, R. (1994). A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* **10**, 301–307.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V. & 7 other authors (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22–28.
- van Wely, K. H., Swaving, J., Freudl, R. & Driessen, A. J. (2001). Translocation of proteins across the cell envelope of Gram-positive bacteria. *FEMS Microbiol Rev* **25**, 437–454.
- Winner, F., Rosengarten, R. & Citti, C. (2000). In vitro cell invasion of *Mycoplasma gallisepticum*. *Infect Immun* **68**, 4238–4244.
- Woese, C. R., Maniloff, J. & Zablen, L. B. (1980). Phylogenetic analysis of the mycoplasmas. *Proc Natl Acad Sci U S A* **77**, 494–498.
- Ye, F., Renaudin, J., Bove, J. M. & Laigret, F. (1994). Cloning and sequencing of the replication origin (*oriC*) of the *Spiroplasma citri* chromosome and construction of autonomously replicating artificial plasmids. *Curr Microbiol* **29**, 23–29.
- Yogev, D., Menaker, D., Strutzberg, K., Levisohn, S., Kirchhoff, H., Hinz, K. H. & Rosengarten, R. (1994). A surface epitope undergoing high-frequency phase variation is shared by *Mycoplasma gallisepticum* and *Mycoplasma bovis*. *Infect Immun* **62**, 4962–4968.
- Zhang, Z., Schaffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res* **26**, 3986–3990.
- Zou, N. & Dybvig, K. (2002). DNA replication, repair and host response. In *Molecular Biology and Pathogenicity of Mycoplasmas*, pp. 303–321. Edited by S. Razin & R. Herrmann. New York: Kluwer/Plenum.