Nucleic Acids Research, 2001, Vol. 29, No. 10 **2145–2153**

The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*

Isabelle Chambaud[1,2], Roland Heilig[5], Stéphane Ferris[2], Valérie Barbe[5], Delphine Samson[5], Frédérique Galisson[3], Ivan Moszer[4], Kevin Dybvig[6], Henri Wróblewski[7], Alain Viari[8], Eduardo P.C. Rocha[4,9] and Alain Blanchard[1,*]

[1]INRA–Université de Bordeaux 2, Institut de Biologie Végétale Moléculaire, 71 avenue Edouard Bourleaux, BP 81, 33883 Villenave D'Ornon Cedex, France, [2]Unité d'Oncologie Virale, [3]Service d'Informatique Scientifique and [4]Unité de Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris Cedex 15, France, [5]Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, BP 191, 91006 Evry Cedex, France, [6]Department of Genomics and Pathology, University of Alabama at Birmingham, 1670 University Boulevard, Volker Hall, Room 418A, Birmingham, AL 35294-0019, USA, [7]Université de Rennes 1, UMR CNRS 6026, Campus de Beaulieu, F-35042 Rennes Cedex, France, [8]INRIA Rhone-Alpes-Projet HELIX, 655 Avenue de l'Europe, 38330 Montbonnot-Saint Martin, France and [9]Atelier de Bioinformatique, 12 Rue Cuvier, 75005 Paris, France

Received January 8, 2001; Revised and Accepted March 19, 2001

DDBJ/EMBL/GenBank accession no. AL445566

ABSTRACT

***Mycoplasma pulmonis* is a wall-less eubacterium belonging to the *Mollicutes* (trivial name, mycoplasmas) and responsible for murine respiratory diseases. The genome of strain UAB CTIP is composed of a single circular 963 879 bp chromosome with a G + C content of 26.6 mol%, i.e. the lowest reported among bacteria, *Ureaplasma urealyticum* apart. This genome contains 782 putative coding sequences (CDSs) covering 91.4% of its length and a function could be assigned to 486 CDSs whilst 92 matched the gene sequences of hypothetical proteins, leaving 204 CDSs without significant database match. The genome contains a single set of rRNA genes and only 29 tRNAs genes. The replication origin *oriC* was localized by sequence analysis and by using the G + C skew method. Sequence polymorphisms within stretches of repeated nucleotides generate phase-variable protein antigens whilst a recombinase gene is likely to catalyse the site-specific DNA inversions in major *M.pulmonis* surface antigens. Furthermore, a hemolysin, secreted nucleases and a glycoprotease are predicted virulence factors. Surprisingly, several of the genes previously reported to be essential for a self-replicating minimal cell are missing in the *M.pulmonis* genome although this one is larger than the other mycoplasma genomes fully sequenced until now.**

INTRODUCTION

Mycoplasmas are prokaryotes that cause slowly progressive, chronic diseases in human and animals. Mycoplasmal infections in humans are associated with a variety of respiratory and urogenital diseases, and those in farm animals are responsible for important economic losses (1–3). Because current antibiotic treatments often fail to eradicate mycoplasmas from the host, better control of these infections is needed. As for other pathogens, it is expected that the availability of fully sequenced genomes will help in understanding the physiology and the pathogenesis mechanisms of these bacteria. This should subsequently allow the development of new and efficient means for the control of these pathogens.

Mycoplasmas (class *Mollicutes*) lack a cell wall but are related to Gram-positive eubacteria from which they evolved by a drastic reduction of genome size, resulting in the loss of many biosynthetic abilities. They are, thus, considered as the best representatives of the concept of a minimal cell (4). Three mycoplasma genomes have been sequenced to date, two from closely related species, *Mycoplasma pneumoniae* (Mpn) (5) and *Mycoplasma genitalium* (Mge) (6), and the third from *Ureaplasma urealyticum* (Uu) (7), which belongs to the same phylogenetic group.

Mycoplasma pulmonis, the subject of our work, is the etiologic agent of murine respiratory mycoplasmosis in rats and mice. This natural pathogen is considered as the most useful model for studying mycoplasmal respiratory infections, including those commonly caused by Mpn in humans. Mpn is a leading cause of respiratory tract infections, especially in children and young adults. Although it has been estimated that only 3–10% of Mpn-infected patients develop pneumonia, up to 30% of cases involving all age groups may be caused by this bacterium (for review see 8). In accordance with observations made during Mpn infections, innate immunity provides anti-*M.pulmonis* defence of the lungs and humoral immunity plays a role against the dissemination of the infection (9). A key factor in the ability of mycoplasmas to establish a chronic infection is their genome flexibility, which allows them to produce a highly

*To whom correspondence should be addressed. Tel: +33 5 57 12 33 20; Fax: +33 5 56 84 31 59; Email: ablancha@bordeaux.inra.fr

variable mosaic of surface antigens (10,11). One of the better-understood systems exhibiting this antigenic variation is the one responsible for the expression of the Vsa lipoproteins in *M.pulmonis* (12). Here we report the complete *M.pulmonis* genome sequence. Its analysis revealed hypervariable domains relevant to pathogenesis and provides clues for deciphering the respective roles of putative virulence factors.

MATERIALS AND METHODS

Mycoplasma strain

The strain UAB CT of *M.pulmonis* was kindly provided by Maureen Davidson (University of Florida, Gainesville, FL). This strain was passaged twice in a modified PPLO broth medium in our laboratory and the obtained stock named UAB CTIP. Growth medium contained per liter: 22.5 g of PPLO broth powder (Difco), 0.2 g of salmon sperm DNA (Sigma), 200 ml of horse serum, 5 ml of Isovitalex solution (Becton Dickinson), 10 ml of 25% (w/v) glucose, 4 ml of 1% (w/v) phenol red and 120 mg of ampicillin. Genomic DNA was isolated by lysing mycoplasma cells with SDS and proteinase K followed by phenol–chloroform extractions. The same DNA preparation was used for the construction of the different libraries.

Construction of genomic libraries

A preliminary library, named A (1–2 kb inserts obtained by partial *Sau*3AI digestion of genomic DNA and cloned into pUC18), was constructed for a pilot study. We have subsequently generated three other libraries of different insert size to favor the scaffold formation between sequence contigs during the assembly stage. Library C was a short insert (2–3 kb) library constructed by partial *Sau*3AI digestion followed by gel sizing and ligation to the *Bam*HI site of pUC18. Library R was made by partial *Tsp*509I digestion, sizing of fragments to 5–6 kb and insertion into the *Eco*RI site of the pBAM3 vector (derived from pBluescript II SK+; R.Heilig, unpublished material). Library B was a 'miniBAC' library, constructed by partial *Tsp*509I digestion of genomic DNA, sizing >15 kb and ligation to the gel-purified 8.7 kb *Eco*RI-linearized pBACe3.6 BAC vector.

Shotgun sequencing and sequence assembly

Cycle sequencing was performed, from both ends of the plasmid and miniBAC clones using flanking vector sequences as primers. Dye primer reactions were analyzed on a Li-Cor 4200L (4000 clones from library R) and dye terminator reactions on an ABI 377 (500, 2000 and 1500 clones, from libraries A, C and B, respectively). The whole genome sequence assembly was performed using the PHRED and PHRAP software packages (13). A total of 6018 end sequences from library R (mean length 930 nt; 5.8 genome equivalents) and 574, 3840 and 1920 reads from libraries A, C and B, respectively (mean length 600 nt; 4.0 genome equivalents), were incorporated into contigs. Moreover, the *vsa*-containing region, spanning 10–15 kb, revealed a high level of heterogeneity between the subclones, which hindered the correct sequence assembly and was attributed to sequence heterogeneity in the *M.pulmonis* culture used for the genomic DNA preparation. To circumvent this problem, we performed restriction digest analysis of all the miniBACs covering this region to identify an over-represented population of clones and selected two of them for subsequent subcloning and complete sequencing. Finally, the five individual assemblies were combined to form a consensus, which was used to generate the full genomic sequence.

Finishing and validation

One cloning gap (39 bp) and several regions with uncertainties were resolved by sequencing duplicate PCR products. To fullfil the sequence quality criteria (a minimum of three reads for each position, sequencing on both strands or using two types of chemistry and a PHRAP quality value of 40 or more), a total of 1144 finishing reactions, requiring 481 primers, were performed. The assembly was validated by restriction digest analysis, both on total genomic DNA by pulsed-field gel electrophoresis using the enzymes *Eag*I, *Kpn*I, *Nru*I and *Sac*II, and on a selection of overlapping miniBACs and plasmid clones using the enzymes *Eco*RI, *Eco*RV, *Hin*dIII and *Bss*HII.

Identification of CDSs and annotation

The annotation was mainly performed using the integrated computer environment specialized for large-scale sequence annotation, Imagene (14). The see text on p.1 were identified using GLIMMER, software for compositional analysis with an interpolated Markov model (15). Putative coding sequences (CDSs) extracted from 300 kb of the *M.pulmonis* genome were used to train the Markov model. Once trained, the model was applied to the genome using 100 bp as the cut-off value for the smallest CDSs. A total of 875 CDSs were detected using this method but some of them were removed because either they significantly overlapped other CDSs and/or they were of small size (100–150 bp) with a deduced polypeptide presenting no detectable similarity with sequences and patterns in databases. CDSs in intergenic spaces were also searched using BLASTX2 (16). The CDSs were numbered starting with *dnaA* as the first gene (MYPU_10010) and *recD* as the last (MYPU_7820).

Search of homology of the products deduced from these CDSs was performed using BLASTP2 (16) against SWISS-PROT and SP-TrEMBL. Additional investigations with CDSs showing no homologs using this first strategy were performed using FASTA (17) and Smith and Waterman algorithms. Motifs were also searched on translated genes using the ProfileScan server (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html) against the databases PROSITE, Pfam-A and Gribskov (18–20). For putative membrane proteins and ABC transporters, transmembrane domains were searched using TMpred (21). In the comparison of mollicute genomes, two genes were regarded as homologous if the proteins they encode were similar both in sequence and in size. For this, we made pair-wise comparisons of all proteins of all proteome pairs, filtering potential homologs using a threshold in BlastP of E-value $<10^{-5}$ and in maximal difference of protein lengths of 20%. Subsequently, we aligned the sequences, using a variant of the classical dynamic programming algorithm for global alignment, where one counts 0-weight for gaps at both ends of the largest sequence, using the BLOSUM62 matrix (22). Finally, we retained pairs of proteins presenting similarity >40%. The set of orthologous genes (supposed to have diverged following a speciation event) was defined by adding a further criterion of double best hit, i.e. two genes are defined as orthologous if they are homologous and if they are the best

match of one another in the respective genomes. For the comparaison of genomes, we also used the information provided in the KEGG catalog of genes for Mpn, Mge and Uu (http://star.scl.genome.ad.jp/kegg/kegg2.html#genes) (23). Terminators were searched using PETRIN (24). The tRNAs were located using the software fastRNA (25) and rRNA by detecting regions of homology with mycoplasmal rRNAs. Analysis of codon usage through factorial correspondence analysis was done using software developed at the Atelier de BioInformatique (26). The search for close repeats was done through the use of the KMR algorithm in windows as explained elsewhere (27). Dot-plots were drawn by using the program dotter from the Sanger Center (28).

Database submission and web-based access to the sequences

These sequence data have been submitted to the EMBL/GenBank/DDBJ databases under accession number AL445566.

The information of individual CDSs, data retrieval and graphical views are available on the web (http://genolist.pasteur.fr/MypuList). This server, which has an organization identical to that described for Subtilist, the *Bacillus subtilis*-dedicated web server (29), also provides the possibility of performing similarity searches, using FASTA or BLAST, and motif searches using pattern matching.

RESULTS AND DISCUSSION

General features of the genome

The general features of the genome are listed in Table 1. The genome size is 963 879 bp and its average G + C content is 26.6%, which is, along with that of Uu, the lowest of all bacterial genomes (7). In *M.pulmonis*, the base composition is rather homogeneous along the entire chromosome (Fig. 1A), with regions corresponding to rRNA and *vsa* genes showing the lowest A + T values. The average A + T content of the coding regions is 73% with 85.3% at the third position of the codon. An analysis of the codon usage among the 782 identified CDSs using factorial correspondence analysis (26) indicated that the code is homogeneously used among all the CDSs, the only exception being the *vsa* genes and a few other lipoprotein genes (data not shown). This result suggests that there was, if any, a low level of horizontal transfer during the recent evolution of this mollicute. Like all species belonging to the genus *Mycoplasma*, *M.pulmonis* reads UGA as a tryptophan codon instead of a stop codon. In fact, the UGA codon is strongly preferred to UGG (91.2 versus 8.8%). There is also a strong preference for the usage of the UAA stop codon (86.2%) as compared with 13.8% for UAG, the other stop codon. This bias confirms the association between the use of these two stop codons and the G + C% of the entire genome which was recently reported for different bacterial species (31). The use of only two stop codons is in accordance with the finding of a single peptide chain release factor-encoding gene (*prfA*), capable of recognizing both UAA and UAG.

The *M.pulmonis* genome encodes a single set of rRNA genes. In contrast to Mpn, Mge and Uu, the three genes are not organized as an operon: the 16S rRNA and 23S rRNA genes are adjacent, without intervening tRNA genes, which seems to be common to all mollicutes, the phytoplasmas excepted. The

Table 1. General features of the *M.pulmonis* genome

Length (bp)	963 879
G + C ratio	26.6%
Putative protein coding sequences (CDSs)	782
Coding region (% of chromosome size)	91.4
Average CDS length (bp)	1115
CDSs with functional assignment[a]	486
CDSs with matches to conserved hypothetical proteins	92
CDSs without significant database match	204
Ribosomal RNA operons	1[b]
tRNAs	29[c]

[a]Functional assignment was performed using the classification scheme from Riley (30).
[b]The 5S rRNA gene is located 350 kb upstream of the 16S rRNA and 23S rRNA gene cluster.
[c]This set of tRNAs corresponds to all amino acids.

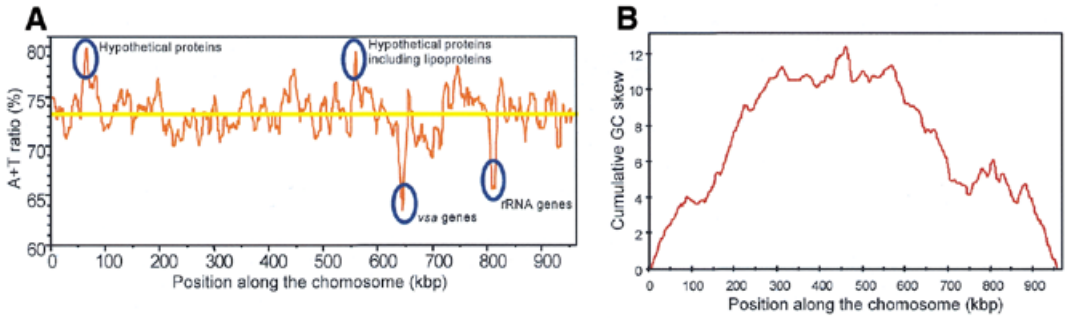

Figure 1. A + T ratio and cumulated GC skew along the *M.pulmonis* chromosome. (**A**) The curve indicates the A + T ratio along the *M.pulmonis* chromosome. This ratio was calculated with a sliding window of 10 kb with a step of 1 kb. The yellow line indicates the average A + T ratio for the genome. The identity of regions showing unusual A + T ratio is indicated. (**B**) The curve indicates the cumulated GC skew along the *M.pulmonis* chromosome starting at position 1 as defined in the text. The transition in GC skew is easily identified around position 1, the putative origin of replication of the chomosome. In contrast, the other transition which would correspond to the termination of replication is more difficult to identify because there are distinct peaks in the region centered around 460 kb.

Table 2. Functional classification of proteins in the completely sequenced mollicute genomes

Functional category	*M.pulmonis* (964 kb)	*M.genitalium* (580 kb)	*M.pneumoniae* (816 kb)	*U.urealyticum* (752 kb)
Amino acid biosynthesis	0	0	0	1
Purine, pyrimidine, nucleoside and nucleotide metabolism	23	19	19	20
Fatty acid and phospholipid metabolism	15	8	9	8
Biosynthesis of co-factors, prosthetic groups and carriers	10	4	8	12
Central intermediary metabolism	9	7	6	13
Energy metabolism	55	33	39	24
Transport and binding proteins	83	33	44	48
DNA metabolism	66	29	46	46
Transcription	12	13	13	17
Protein synthesis	97	90	91	91
Protein fate	22	21	17	12
Regulatory functions	4	5	8	6
Cell envelope	60	29	54	19
Cellular processes	11	6	11	13
Other categories	8	0	3	18
Unknown	11	12	46	22
Hypothetical				
Database match	92	168	188	106
No database match	204	6	86	171
Total number of proteins	**782**	**483**	**688**	**613**

5S rRNA is located 350 kb from the two other rRNA genes, such a separation having previously been described for two other mycoplasmas, *Mycoplasma hyopneumoniae* and *Mycoplasma flocculare* (32). The set of 29 tRNA genes with specificity for all amino acids was identified in the *M.pulmonis* genome and currently represents the smallest set of tRNAs among the sequenced bacterial genomes. Likewise, the four sequenced mollicute genomes share a single gene for a sigma factor, the one encoding σ70.

With the software Petrin, that uses *Escherichia coli* default parameters, we were able to identify 281 ρ-independent terminators in the genome of *M.pulmonis*. Although these terminators were detected in the genome of Mpn (5), they have not been reported in the other sequenced mycoplasma genomes and they have been predicted not to exist in such low G + C genomes (33).

Origin of replication

The origin of replication of the *M.pulmonis* chromosome has not been experimentally identified. However, as described for other bacterial genomes, but not for the Mpn and Mge genomes, the tandem arrangement of *dnaA* and *dnaN* genes was identified. However, the two other genes, *gyrA* and *gyrB*, commonly found in the vicinity of the origin of replication, are located elsewhere on the genome. We have designated as base 1 of the *M.pulmonis* genome the first non-coding base upstream *dnaA*. We identified five putative DnaA boxes (consensus sequence 5′-TTATCCACA-3′), three being upstream *dnaA* and the two other downstream of this gene. This organization was not found in the Mpn, Mge and Uu genomes (5–7), but is reminiscent of that described for other mollicutes including *Spiroplasma citri* (34) and *Mycoplasma capricolum* (35). It is, therefore, possible that this divergence from the typical eubacterial *oriC* is restricted to the phylogenetic cluster of mycoplasma species that includes Mpn, Mge and Uu.

The analysis of the G + C skew, defined as [(G – C)/(G + C)] (36), on the *M.pulmonis* genome provided additional evidence that the origin of replication was properly located (Fig. 1B). There was indeed a marked inversion of G + C skew upstream of *dnaA*. This bias is stronger than in Mpe and Mge but weaker when compared with other bacterial genomes (37). On both sides of the putative origin of replication, there is also a strong bias in the polarity of the genes transcribed on the two DNA strands.

Comparison with other sequenced mollicute genomes

The *M.pulmonis* genome is currently the largest completely sequenced genome from a mollicute. Consequently, the number of CDSs identified is larger than for the other three mollicutes: 782 as compared with 483, 688 and 605 for Mpn (38), Mge and Uu, respectively (Table 2). This increased number of genes is not evenly distributed in all the functional categories. In some categories, including those governing nucleotide metabolism, intermediary metabolism and protein synthesis and fate, the number of genes is equivalent in the four

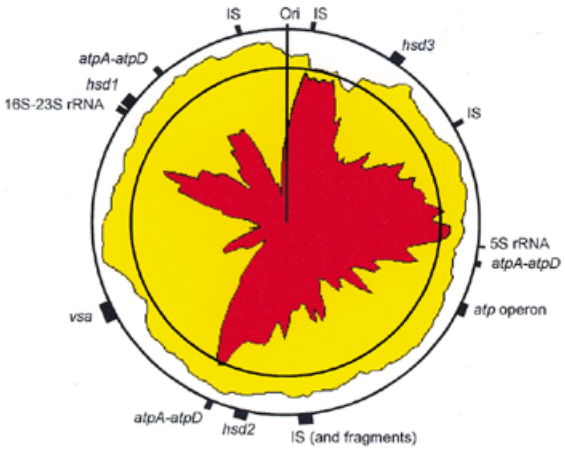

Figure 2. Density of coding sequences along the *M.pulmonis* chromosome. Yellow indicates the density of coding sequences in both strands of the sequence and red the density of the coding sequences in the clockwise strand (published sequence). This density was calculated with a sliding window of 50 kb. The first circle represents 100%, the second (inner circle) is 80%. The positions of rRNA genes and of other genes discussed in the text are indicated by black boxes.

mollicute genomes (Table 2). In these categories, these genomes appear to approach, to a similar level, the concept of a minimal cell. Furthermore, and quite surprisingly, in contrast to Mpn, Mge and Uu which share the same set of genes for transcription, *M.pulmonis* has adopted in this category a still more minimalist organization. Indeed, the gene *rpoE* encoding the δ subunit of RNA polymerase is lacking; this gene has been shown by mutagenesis experiments to be dispensable in *B.subtilis* (39). The larger number of genes in *M.pulmonis* as compared with the other three mollicutes can be mainly attributed to an increased number of membrane proteins, in particular those devoted to transport. It is difficult to predict from the sequence analysis the specificity of ABC (ATPase) transporters (40), 56 *M.pulmonis* CDSs were found to potentially encode subunits for these transporters as compared with 32 and 33 CDSs for Mpn and Uu, respectively. Differences were also found in the 'Energy metabolism' category (Table 2) and are due, in part, to the presence of four CDSs encoding α (*atpA*) and β (*atpD*) ATP synthase subunits. A set of *atpA-atpD* genes is located within the operon encoding the other six subunits of the ATP synthase and the three other *atpA-atpD* tandem copies are scattered on the chromosome (Fig. 2).

Of the 179 *M.pulmonis* CDSs, which do not have an ortholog in the Mpn, Mge and Uu genomes, 123 encode putative proteins showing no significant homology in the databases. Additionally, eight other *M.pulmonis* CDSs encode conserved hypothetical proteins with the closest homologs among the polypeptides deduced from Gram-positive bacterial genomes, as expected from the mycoplasmal phylogenetic position. One noteworthy exception is Mypu_4730, which shares homology with a putative *Thermotoga maritima* polypeptide. Some other *M.pulmonis* polypeptides with no ortholog deduced from the Mpn, Mge and Uu genomes include ribosomal protein S1 (Mypu_1300), glucokinase (GlcK; Mypu_2280), 5′-nucleosidase precursor (UshA; Mypu_0550) and signal peptidase I (SipS; Mypu_6300). It should be noted that re-annotating the Mpn genome suggested that one of the identified proteases in this organisms (MPN386) could act as a signal peptidase I (38). *Mycoplasma pulmonis* also seems to have the ability to metabolize sugars other than those metabolized by Mpn, Mge and Uu. Indeed, a set of genes involved in the catabolism of maltodextrins has been identified; these genes are clustered together in one region of the genome and encode a maltodextrin ABC transport system (Mypu_6390–6410), a β-D-phosphate-glucomutase (PgmB; Mypu_6350), a dextrinase with unknown specificity (Mypu_6330) and an α-amylase 3 (AmyC, Mypu_6320). This finding suggests that *M.pulmonis* can use glycogen as a source of energy and carbon which is in accordance with reports that all glucose-fermenting mycoplasmas ferment glycogen. However, these results were based on experiments using media containing serum or serum fractions, which contain saccharolytic enzymes (41). Therefore, the use of complex sugars by *M.pulmonis* will have to be experimentally verified. Interestingly, the closely located Mypu_6220 CDS potentially encodes a polypeptide that belongs to the *Lac*I family of transcriptional regulators. This polypeptide, with no ortholog in Mpn, Mge and Uu, together with the heat-inducible transcription repressor (Mypu_1420) is the only indication for transcriptional regulation in this mycoplasma.

Among the *M.pulmonis* CDSs for which an ortholog was found in Uu but not in Mpn, are CTP synthase (Mypu_6660), cardiolipin synthetase (Mypu_6650) and a poorly understood enzyme (*Psl*X; Mypu_1620) involved in the phospholipid metabolism. In addition, six IS-elements including three which are incomplete were found in the *M.pulmonis* genome.

Although the gene order in Mpn and Mge is highly conserved (42), this is the case neither in Uu (7) nor in *M.pulmonis* (data not shown). The regions of syntheny between the four genomes consist in most cases of genes encoding products such as ribosomal proteins, ABC transporters or ATP synthase subunits.

Comparison with the set of genes for a minimal cell

The Mge genome is the smallest bacterial genome and as such is considered as the best representative for the concept of a minimal cell (43). With the assumption that genes conserved across large phylogenetic distances are likely to be essential, a minimal set of genes for a living cell was defined by others (4). Among the 256 Mge genes belonging to this minimal set (http://www.ncbi.nlm.nih.gov/Complete_Genomes/Minset/), eight genes do not have an ortholog in the *M.pulmonis* genome [MG278, guanosine 5′-diphosphate 3′-diphosphate pyrophosphohydrolase (*spoT*); MG453, UTP-glucose-1-phosphate uridylyltransferase (*gtaB*); MG053, phosphomannomutase (*cpsG*); MG063, 1-phosphofructokinase (*fruK*); MG393, heat shock protein (*groES*); MG392, heat shock protein (*groEL*); MG382, uridine kinase (*udk*); MG118, UDP-glucose 4-epimerase (*galE*)]. Although the Mpn and Mge genomes have genes for the synthesis of the heat shock proteins-chaperonins GroEL and GroES, the absence of these genes has also been reported for Uu (7). GroEL has also been shown to be dispensable for Mpn *in vitro* (44). The fact that the *M.pulmonis* genome lacks *fruK* but has a probably functional *fruA* gene has to be regarded in the light of some recently reported data showing that in the mollicute *S.citri*, fructose catabolism requires *fruA* but not *fruK* (45). Another striking absence in *M.pulmonis* is that of

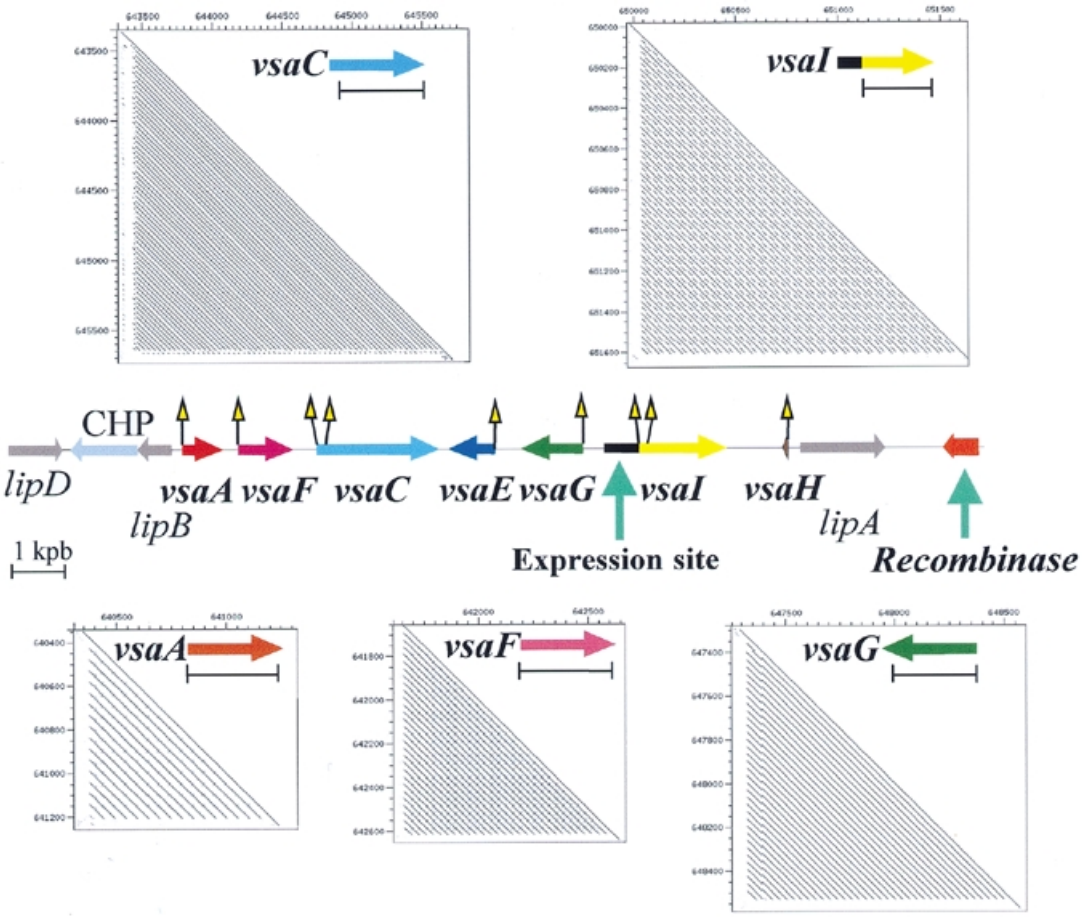

Figure 3. General organization of the *vsa* locus and visualization of repeats within individual *vsa* genes Within this locus encoding seven *vsa* genes, the transcribed gene is *vsaI* as indicated by the location of the expression site which is fused to this gene. The *vrs* boxes which are the sites for DNA site-specific inversions are indicated by yellow triangles. We suggest that the recombinase responsible for the rearrangements within this locus is encoded by the gene located at the 3′-end. In addition to the *vsa* genes, three genes putatively encoding lipoproteins (LipA, LipB and LipD) and a gene encoding a conserved hypothetical protein (CHP) were identified. The repeats within each *vsa* gene were individually analyzed by dot-plot. Within each of the five boxes, the bracket below the colored gene indicates the region of the gene with repeated units.

SpoT, an enzyme playing a key role in the stringent response. Strong evidence supports that there is only one *relA/spoT* product in Gram-positive bacteria (in contrast to two in Gram-negative bacteria) that displays activity for both synthesis and degradation of (p)ppGpp (46). SpoT was shown by transposon mutagenesis to be dispensable for Mpn and Mge (44). There are only a few examples of eubacteria, which lack the stringent response, in particular *Helicobacter pylori* (47) and *Chlamydia trachomatis* (48).

Highly recombinogenic loci in the *M.pulmonis* genome

The *vsa* locus encodes highly variable surface lipoprotein antigens. It has been shown in strain KD735-15 that high-frequency, site-specific DNA inversions serve to regulate the phase-variable expression of individual *vsa* genes (12). Only one *vsa* gene, located in the expression site, is transcribed at a given time, the other *vsa* genes being silent. A model involving DNA strand exchange at conserved *vsa* recombination sites (*vrs* boxes) has been proposed to mediate these inversions (12). The completed genomic sequence allows for the first time analysis of the complete *vsa* locus of the UAB CTIP strain and reveals seven *vsa* genes (Fig. 3), with four of them (*vsaA*, *vsaC*, *vsaE* and *vsaF*) being previously described in the strain KD735-15 containing 11 *vsa* genes (12). As expected from such a highly recombinogenic locus, neither the gene order nor the number of repeated units (see below) in individual genes are conserved between strains. The expressed gene in the sequenced locus is the newly described *vsaI* gene. At the 5′-end of each *vsa* gene a *vrs* box was identified, an additional *vrs* box laying within the genes *vsaC* and *vsaI* (Fig. 3). The dot-plot analysis of this locus highlights repeated elements within each individual *vsa* gene; for example, a 57 bp unit (19 amino acids) is repeated 27 times in the *vsaI* gene. There are also common repeats between different *vsa* genes such as between *vsaF* and *vsaG* (data not shown). Within the *vsa* locus, three lipoprotein-encoding CDSs (*lipA*, *lipB* and *lipD*; Fig. 3) were also identified. Although the deduced sequences share significant similarity with Vsas, they do not exhibit a repeated C-terminus nor *vrs* boxes within their genes. Furthermore, it was predicted that the putative site-specific recombinase that promotes DNA inversion within this locus would be a member of the λ integrase family. Indeed a gene, the product of which potentially

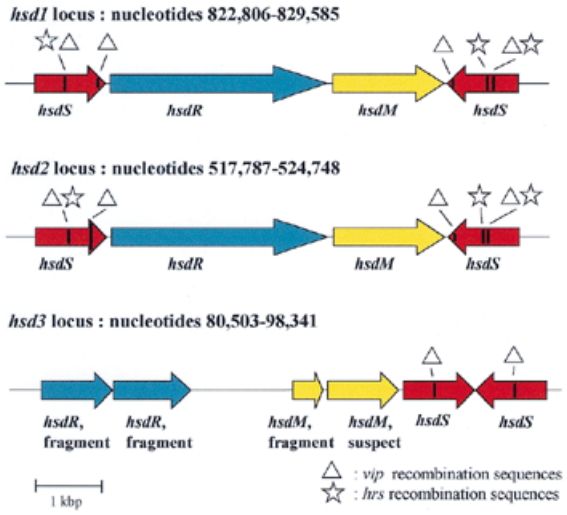

Figure 4. Comparative organization of the three *hsd* loci encoding type I restriction modification systems in *M.pulmonis*. The organization of the three *hsd* is depicted on this figure. Within the *hsdS* genes the two types of recombination sequences are indicated by stars (*hrs* recombination sequences) and by triangles (*vip* recombination sequences). The position of the *hsd* loci on the genome is indicated and can also be found on a circular representation of the genome in Figure 2.

encodes this type of recombinase, was located immediately downstream of the *vsa* locus (Fig. 3). Experiments are underway to confirm that this gene is indeed responsible for the variability of cell surface antigens and also to determine if antigenic variation confers some protection from the host's immune system.

Previous studies have shown that *M.pulmonis* strain KD735-15 contains at least two highly recombinogenic *hsd* loci that code for type I restriction and modification systems (49). The UAB CTIP genome has two *hsd* loci similar to that of KD735-15 and an additional third *hsd* locus that is not completely functional (Fig. 4). At the *hsd3* locus, the two *hsdS* are most likely functional but the *hsdR* and *hsdM* have been disrupted by frameshift mutations and are predicted to be non-functional. The three loci are not co-located on the chromosome (Fig. 2) and, in contrast to the *vsa* locus, no CDS encoding a putative recombinase involved in the site-specific DNA inversions was found in their vicinity or elsewhere on the chromosome. Although detected in the Mpn and Mge genomes, the *hsd* systems do not seem to be active in these species. In addition to these *hsd* genes, we identified a complex locus of a type III RM system and CpG methylase genes that were not identified previously within the genus *Mycoplasma*.

Lipoproteins, polymorphisms and repeats

In addition to the expressed *vsa* product (VsaI), 55 other lipoprotein CDSs (7% of the total number of CDSs), were identified as compared with 46, 21 and 42 in Mpn, Mge and Uu, respectively. Very few of these lipoproteins had orthologs in the databases and a putative function could be attributed to only three of them: MYPU_1390 and MYPU_6930 as membrane nucleases, and MYPU_6320 as an amylase. Three other lipoproteins with unknown function (MYPU_0070, MYPU_3430 and MYPU_5260) have orthologs in other mycoplasmas. However, several of the lipoproteins could be classified into highly homologous families. Two of the enzymes involved in the maturation of bacterial lipoproteins could be deduced from the genome, the prolipoprotein diacylglycerol transferase (*lgt*) and the prolipoprotein signal peptidase (*lsp*). A third enzyme, the apolipoprotein:phospholipid *N*-acyl transferase (*lnt*), was not found, which substantiates the finding that lipoprotein acylation may not be complete in some mycoplasma species (50,51). However, one cannot rule out that other types of acylation occur, similar to that recently described for *Borrelia burgdorferi* (52).

The expression of several mycoplasmal lipoprotein genes has been shown to be phase-variable (for review see 11) and one of the underlying mechanisms is due to variation in the length of homopolymeric nucleotide tracts located either in the promoter or in the coding sequence itself. A stretch of repeated T (T_{31}) and A (A_{39}) nucleotides is located upstream of two *M.pulmonis* genes encoding the putative lipoproteins MYPU_0190 and MYPU_4780, respectively. These homopolymeric tracts are highly suggestive of phase-variation in the expression of these genes because they are hotspots for frameshift mutations generated by replication slippage. Additional evidence for phase-variation was provided by polymorphisms in the sequence of these repeats. The assembly revealed variations in these tracts among sequences from otherwise identical clones. We verified that this was not due to sequencing difficulties. For MYPU_0190, T_{31} was found six times and T_{29} two times (6×T31/2×T29). The situation was more complicated for MYPU_4780, with the following polymorphism: 1×A36/2×A39/2×A40/2×A41/2×A44. The detection of variation in these homopolymeric tracts was favored because the DNA for generating the sequencing libraries (see Materials and Methods) was obtained from a *M.pulmonis* isolate rather than from a single clone. Should this phase-variation experimentally be confirmed, this would be the first indication that two different mechanisms (length variation in a homopolymeric tract and site-specific inversions in the *vsa* genes) driving phase-variation of surface antigens operate in a single mycoplasma cell.

Virulence factors

Like other mollicutes, this organism contains few recognizable genes likely to be involved in virulence. Exceptions are a hemolysin (MYPU_1710), a set of membrane-associated or secreted nucleases (MYPU_6930–6940 and MYPU_1390) and a potentially secreted protease (MYPU_1180). The hemolytic activity of *M.pulmonis* was previously described (53). The MYPU_1710 product has no homolog in Mpn or Mge, but in Uu, another mollicute displaying a hemolytic activity (7). This hemolysin belongs to the *hlyA* family, members of which are recognized virulence factors in *Serpulina hyodysenteriae* and in mycobacteria (54,55). Mollicutes have limited biosynthetic capacity, and membrane-associated or secreted proteases and nucleases are thought to help these microorganisms to acquire metabolic precursors from the host. The study of the nuclease produced by *Mycoplasma penetrans* clearly indicated that mycoplasmal nucleases could act as virulence factors (56,57). A thiol peroxidase (MYPU_7080) is also of interest because the production of hydrogen peroxide has been suggested to be

a virulence factor during *M.pulmonis* infections (58). This enzyme was not deduced from the genomes of other mollicutes.

CONCLUSIONS

The complete sequence of the *M.pulmonis* genome opens a wide range of perspectives for the study of the mollicutes, and, more generally, of bacterial evolution related to pathogenicity. The analysis of this genome will be more easily exploited than those of the preceding mollicutes because of the availability of tools, transposons and *oriC*-based replicative plasmids, to genetically manipulate *M.pulmonis* (59,60) (A.Blanchard, unpublished material). Furthermore, the mouse, one of the natural hosts of *M.pulmonis*, appears to be a particularly useful model system to dissect mycoplasma–host interactions notably owing to the existence of mice strains differing in their susceptibility to mycoplasmal disease. *Mycoplasma pulmonis* belongs to a phylogenetic branch distinct from that of the other sequenced mycoplasmas. We are aware of at least six other ongoing mollicute genome projects (*M.mycoides* subsp. mycoides SC, *M.hyopneumoniae*, *M.gallisepticum*, *M.penetrans*, *S.citri* and *S.kunkelii*). From an evolutionary perspective, the availability of the genome sequence of these mycoplasmal species will make this group of organisms an ideal target for genome comparisons and studies of microbial evolution.

Still, the most striking feature of the genome of *M.pulmonis* is that it reveals an astonishing diversity of strategies related to the evolution of pathogenicity. It contains genes that can participate in intra-chromosomal homologous recombination, like in *M.pneumoniae* and *M.genitalium*, antigenic variation through slipped-strand mispairing of repetitive motifs in control regions, as found in *H.pylori* (61) and *Haemophilus influenzae* (62), highly variable repetitive proteins, like in *Mycobacterium tuberculosis* (63), and variable restriction and modification systems, as proposed for *Neisseria gonorrhoea* (64,65). Hence, in spite of its minimalist character in terms of biochemical potential, *M.pulmonis* is an extremely suited model organism to study pathogenesis and co-evolution of bacteria with the host.

This work paves the way for a wide variety of experimental studies. One obvious experiment is to inactivate the gene(s) potentially controlling the inversions within the *vsa* locus to confirm the contribution of this enzyme in the variability of cell surface antigens and also to determine whether antigenic variation confers protection against the host's immune system. The fact that a single recombinase gene was found in the genome suggests that this enzyme catalyzes DNA inversions in both the *vsa* and *hsd* loci.

ACKNOWLEDGEMENTS

We would like to thank J.Weissenbach and L.Montagnier for continuous support during this program, Paul Scott for carefully reading the manuscript and Sophie Oztas for technical help during the sequencing at Genoscope. This work was funded by the Institut Pasteur, Genoscope, INRA, Région Aquitaine and the Université Victor Segalen Bordeaux 2. It was also supported in part by the US Public Health Service grant AI41113 to K.D. from the National Institutes of Health.

REFERENCES

1. Jacobs,E. (1997) Mycoplasma infections of the human respiratory tract. *Wien. Klin. Wochensch.*, **109**, 574–577.
2. Taylor-Robinson,D. and Furr,P.M. (1998) Update on sexually transmitted mycoplasmas. *Lancet*, **351**, 12–15.
3. Frey,J. and Nicolet,J. (1997) Molecular identification and epidemiology of animal mycoplasmas. *Wien. Klin. Wochensch.*, **109**, 600–603.
4. Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
5. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **24**, 4420–4449.
6. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R. D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
7. Glass,J.I., Lefkowitz,E.J., Glass,J.S., Heiner,C.R., Chen,E.Y. and Cassell,G.H. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature*, **407**, 757–762.
8. Mansel,J.K., Rosenow,E.C.D., Smith,T.F. and Martin,J.W.,Jr (1989) *Mycoplasma pneumoniae* pneumonia. *Chest*, **95**, 639–646.
9. Cartner,S.C., Lindsey,J.R., Gibbs-Erwin,J., Cassell,G.H. and Simecka,J.W. (1998) Roles of innate and adaptive immunity in respiratory mycoplasmosis. *Infect. Immun.*, **66**, 3485–3491.
10. Citti,C. and Rosengarten,R. (1997) Mycoplasma genetic variation and its implication for pathogenesis. *Wien. Klin. Wochensch.*, **109**, 562–568.
11. Chambaud,I., Wroblewski,H. and Blanchard,A. (1999) Interactions between mycoplasma lipoproteins and the host immune system. *Trends Microbiol.*, **7**, 493–499.
12. Shen,X., Gumulak,J., Yu,H., French,C.T., Zou,N. and Dybvig,K.(2000) Gene rearrangements in the *vsa* locus of *Mycoplasma pulmonis*. *J. Bacteriol.*, **182**, 2900–2908.
13. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
14. Medigue,C., Rechenmann,F., Danchin,A. and Viari,A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, **15**, 2–15.
15. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.
18. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
19. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.
20. Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.
21. Hofmann,K. and Stoffel,W. (1993) TMbase—a database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*, **347**, 166.
22. Henikoff,S. and Henikoff,J.G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.
23. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
24. d'Aubenton Carafa,Y., Brody,E. and Thermes,C. (1990) Prediction of ρ-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem–loop structures. *J. Mol. Biol.*, **216**, 835–858.
25. el-Mabrouk,N. and 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.
26. Moszer,I., Rocha,E.P. and Danchin,A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.*, **2**, 524–528.
27. Rocha,E.P., Danchin,A. and Viari,A. (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in

Bacillus subtilis and other competent prokaryotes. *Mol. Biol. Evol.*, **16**, 1219–1230.
28. Sonnhammer,E.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–10.
29. Moszer,I. (1998) The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. *FEBS Lett.*, **430**, 28–36.
30. Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.
31. Rocha,E.P., Danchin,A. and Viari,A. (1999) Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Res.*, **27**, 3567–3576.
32. Stemke,G.W., Huang,Y., Laigret,F. and Bove,J.M. (1994) Cloning the ribosomal RNA operons of *Mycoplasma flocculare* and comparison with those of *Mycoplasma hyopneumoniae*. *Microbiology*, **140**, 857–860.
33. Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
34. Ye,F., Renaudin,J., Bove,J.M. and 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.
35. Fujita,M.Q., Yoshikawa,H. and Ogasawara,N. (1992) Structure of the *dnaA* and DnaA-box region in the *Mycoplasma capricolum* chromosome: conservation and variations in the course of evolution. *Gene*, **110**, 17–23.
36. Lobry,I.R. (1996) Asymmetric distribution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.
37. Rocha,E.P., Danchin,A. and Viari,A. (1999) Universal replication biases in bacteria. *Mol. Microbiol.*, **32**, 11–16.
38. Dandekar,T., Huynen,M., Regula,J.T., Ueberle,B., Zimmermann,C.U., Andrade,M.A., Doerks,T., Sanchez-Pulido,L., Snel,B., Suyama,M. *et al.* (2000) Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res.*, **28**, 3278–3288.
39. Lopez de Saro,F.J., Yoshikawa,N. and Helmann,J.D. (1999) Expression, abundance, and RNA polymerase binding properties of the δ factor of *Bacillus subtilis*. *J. Biol. Chem.*, **274**, 15953–15958.
40. Quentin,Y., Fichant,G. and Denizot,F. (1999) Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. *J. Mol. Biol.*, **287**, 467–784.
41. Miles,R.J. (1992) Catabolism in mollicutes. *J. Gen. Microbiol.*, **138**, 1773–1783.
42. Himmelreich,R., Plagens,H., Hilbert,H., Reiner,B. and Herrmann,R. (1997) Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res.*, **24**, 701–712.
43. Peterson,S.N. and Fraser,C.M. (2001) The complexity of simplicity. *Genome Biol.*, **2**, 2002.1–2002.7.
44. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Venter,J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma genome*. *Science*, **286**, 2165–2169.
45. Gaurivaud,P., Laigret,F., Garnier,M. and Bové,J.M. (2000) Fructose utilization and pathogenicity of *Spiroplasma citri*: characterization of the fructose operon. *Gene*, **252**, 61–69.
46. Wendrich,T.M. and Marahiel,M.A. (1997) Cloning and characterization of a relA/spoT homologue from *Bacillus subtilis*. *Mol. Microbiol.*, **26**, 65–79.
47. Scoarughi,G.L., Cimmino,C. and Donini,P. (1999) *Helicobacter pylori*: a eubacterium lacking the stringent response. *J. Bacteriol.*, **181**, 552–555.
48. Stephens,R.S., Kalman,S., Lammel,C., Fan,J., Marathe,R., Aravind,L., Mitchell,W., Olinger,L., Tatusov,R.L., Zhao,Q. *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science*, **282**, 754–759.
49. Dybvig,K., Sitaraman,R. and French,C.T. (1998) A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 13923–13928.
50. Muhlradt,P.F., Kiess,M., Meyer,H., Sussmuth,R. and Jung,G. (1997) Isolation, structure elucidation, and synthesis of a macrophage stimulatory lipopeptide from *Mycoplasma fermentans* acting at picomolar concentration. *J. Exp. Med.*, **185**, 1951–1958.
51. Piec,G., Mirkovitch,J., Palacio,S., Muhlradt,P.F. and Felix,R. (1999) Effect of MALP-2, a lipopeptide from *Mycoplasma fermentans*, on bone resorption *in vitro*. *Infect. Immun.*, **67**, 6281–6285.
52. Beermann,C., Lochnit,G., Geyer,R., Groscurth,P. and Filgueira,L. (2000) The lipid component of lipoproteins from *Borrelia burgdorferi*: structural analysis, antigenicity, and presentation via human dendritic cells. *Biochem. Biophys. Res. Com.*, **267**, 897–905.
53. Jarvill-Taylor,K.J. and Minion,F.C. (1995) The effect of thiol-active compounds and sterols on the membrane-associated hemolysin of *Mycoplasma pulmonis*. *FEMS Microbiol. Lett.*, **128**, 213–218.
54. Hyatt,D.R., ter Huurne,A.A., van der Zeijst,B.A. and Joens,L.A. (1994) Reduced virulence of *Serpulina hyodysenteriae* hemolysin-negative mutants in pigs and their potential to protect pigs against challenge with a virulent strain. *Infect. Immun.*, **62**, 2244–2248.
55. Wren,B.W., Stabler,R.A., Das,S.S., Butcher,P.D., Mangan,J.A., Clarke,J.D., Casali,N., Parish,T. and Stoker,N.G. (1998) Characterization of a haemolysin from *Mycobacterium tuberculosis* with homology to a virulence factor of *Serpulina hyodysenteriae*. *Microbiology*, **144**, 1205–1211.
56. Bendjennat,M., Blanchard,A., Loutfi,M., Montagnier,L. and Bahraoui,E. (1997) Purification and characterization of *Mycoplasma penetrans* Ca2+/Mg2+-dependent endonuclease. *J. Bacteriol.*, **179**, 2210–2220.
57. Bendjennat,M., Blanchard,A., Loutfi,M., Montagnier,L. and Bahraoui,E. (1999) Role of *Mycoplasma penetrans* endonuclease P40 as a potential pathogenic determinant. *Infect. Immun.*, **67**, 4456–4462.
58. Brennan,P.C. and Feinstein,R.N. (1969) Relationship of hydrogen peroxide production by *Mycoplasma pulmonis* to virulence for catalase-deficient mice. *J. Bacteriol.*, **98**, 1036–1040.
59. Dybvig,K. and Cassell,G.H. (1987) Transposition of Gram-positive transposon Tn*916* in *Acholeplasma laidlawii* and *Mycoplasma pulmonis*. *Science*, **235**, 1392–1394.
60. Dybvig,K., French,C.T. and Voelker,L.L. (2000) Construction and use of derivatives of transposon Tn*4001* that function in *Mycoplasma pulmonis* and *Mycoplasma arthritidis*. *J. Bacteriol.*, **182**, 4343–4347.
61. Wang,G., Humayun,M.Z. and Taylor,D.E. (1999) Mutation as an origin of genetic variability in *Helicobacter pylori*. *Trends Microbiol.*, **7**, 488–493.
62. Hood,D.W., Deadman,M.E., Jennings,M.P., Bisercic,M., Fleischmann,R.D., Venter,J.C. and Moxon,R. (1996) DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA*, **93**, 11121–11125.
63. Cole,S., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S., Eiglmeier,K., Gas,S., Barry,C.R. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
64. Parkhill,J., Achtman,M., James,K.D., Bentley,S.D., Churcher,C., Klee,S.R., Morelli,G., Basham,D., Brown,D., Chillingworth,T. *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, **404**, 502–506.
65. Saunders,N.J., Jeffries,A.C., Peden,J.F., Hood,D.W., Tettelin,H., Rappuolli,R. and Moxon,E.R. (2000) Repeat-associated phase-variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Microbiol.*, **37**, 207–215.