Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution

Eduardo P. C. Rocha, Alain A. Blanchard

Nucleic Acids Research, 2002, Vol. 30, No. 9 **2031–2042**

Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution

Eduardo P. C. Rocha[1,2,*] and Alain Blanchard[3]

[1]Atelier de Bioinformatique, 12 Rue Cuvier, 75005 Paris, France, [2]Unité GGB, URA2171, Institut Pasteur, 28 Rue du Dr Roux, 75724 Paris Cedex 15, France and [3]INRA–Université de Bordeaux 2, Institut de Biologie Végétale Moléculaire, 71 Avenue Edouard Bourleaux, BP 81, 33883 Villenave d'Ornon Cedex, France

Received December 20, 2001; Revised and Accepted March 4, 2002

ABSTRACT

Mycoplasmas evolved by a drastic reduction in genome size, but their genomes contain numerous repeated sequences with important roles in their evolution. We have established a bioinformatic strategy to detect the major recombination hot-spots in the genomes of *Mycoplasma pneumoniae*, *Mycoplasma genitalium*, *Ureaplasma urealyticum* and *Mycoplasma pulmonis*. This allowed the identification of large numbers of potentially variable regions, as well as a comparison of the relative recombination potentials of different genomic regions. Different trends are perceptible among mycoplasmas, probably due to different functional and structural constraints. The largest potential for illegitimate recombination in *M.pulmonis* is found at the *vsa* locus and its comparison in two different strains reveals numerous changes since divergence. On the other hand, the main *M.pneumoniae* and *M.genitalium* adhesins rely on large distant repeats and, hence, homologous recombination for variation. However, the relation between the existence of repeats and antigenic variation is not necessarily straightforward, since repeats of P1 adhesin were found to be anti-correlated with epitopes recognized by patient antibodies. These different strategies have important consequences for the structures of genomes, since large distant repeats correlate well with the major chromosomal rearrangements. Probably to avoid such events, mycoplasmas strongly avoid inverse repeats, in comparison to co-oriented repeats.

INTRODUCTION

Mycoplasma (class Mollicutes) are bacteria with a genome characterised by its small size, ranging from 0.58 to 1.35 Mb. This limited genetic information is associated with reduced biosynthetic capabilities and the need for these bacteria to acquire many nutrients from their hosts. As a consequence, mycoplasmas are obligate parasites of a wide range of animals, including humans (1). They usually reside on mucosal surfaces and the ability to adhere to the surface of epithelial cells is a prerequisite for colonisation. The mycoplasmas related to the human pathogen *Mycoplasma pneumoniae* possess a specialised tip structure that mediates attachment to host cells (2). At the surface of these structures there is a cluster of proteins involved in adherence. The major adhesins of *M.pneumoniae* are the proteins P1 and P30 that interact with other membrane proteins to facilitate lateral movement and concentration of the adhesin molecules at the attachment organelle. Both the *P1* and *P30* genes have repeated regions of high similarity dispersed in the genome. Although *M.pneumoniae* is considered to be a very homogeneous species, variation by recombination between these repeats has been associated with antigenic variation (3).

The *Mycoplasma* cell envelope lacks a cell wall, is devoid of lipopolysaccharides and contains large amounts of lipoproteins. The sequencing of four *Mycoplasma* genomes revealed the occurrence of many putative lipoprotein genes (4–7). High frequency variation of lipoproteins seems to be a feature common to all mycoplasmas and results in a changing mosaic of antigenic structures at the bacterial cell surface (8). This is thought to help these microorganisms evade host immune surveillance and hence cause disease (for a review see 9). The mycoplasmas having a reduced machinery for transcriptional regulation of gene expression, the lipoprotein variation is probably based on mechanisms constantly generating diversity in the lipoprotein repertoire during replicative events. This diversity results from both phase and/or size variation of lipoproteins. Although a high frequency of lipoprotein variation has been described in many *Mycoplasma* species, the genetic prerequisite allowing this variation has only been identified in a few cases. A common theme is the association of this variation with some types of genetic repeat. Indeed, a large proportion of the *Mycoplasma* genome can be dedicated to the generation of the surface antigen diversity as exemplified by *Mycoplasma gallisepticum*, for which ~16% of the genome seems to correspond to a reservoir of sequences for variation of a single antigen (10).

Bacteria engaging in frequent sequence variation can be regarded as multicellular populations that continually generate individual cells with variant phenotypes (11). This allows rapid adaptation to environmental changes. The sources of such variation are typically constituted of repeats that engage in recombination events, either dependent on (homologous

*To whom correspondence should be addressed at: Atelier de Bioinformatique, 12 Rue Cuvier, 75005 Paris, France. Tel: +33 1 44 27 65 36; Fax: +33 1 44 27 63 12; Email: erocha@abi.snv.jussieu.fr

recombination) or independent of RecA (illegitimate recombination). Frequently one type of repeat will preferentially engage in one type of recombination, though in certain circumstances (long repeats at short distances) both types of events may compete (12).

Homologous recombination involves exchanges between segments of DNA molecules of (nearly) identical sequence. This induces the rearrangement of genes or parts of genes both within and between chromosomes, limits the divergence of repeated DNA sequences and promotes repair of damaged DNA (13). Illegitimate recombination between short close or tandem repeats proceeds either by slipped mispair, at a replication arrest, or single strand annealing, at a DNA double-strand break (12). Although illegitimate recombination between repeats of 6 bp was observed in bacteriophage T7, it is more common among repeats >8 bp. Also, the frequency of recombination increases exponentially with repeat length from 8 to 20 nt, in both *Escherichia coli* and *Bacillus subtilis* (14,15). In these species the frequency of intramolecular recombination decreases exponentially with the distance between the homologous sequences (16,17). Although repeated sequences are scarcer in bacterial genomes than in eukaryotic genomes, enough exist to produce extensive rearrangement, deletion or multiplication of genetic material (18,19).

The potential of repeats to generate polymorphisms in bacterial genomes is subject to increasing attention (20–23). However, most of these studies have focused on the role of simple sequence repeats (SSR) in one given genome. Here, in a comparative analysis of the repeats within four *Mycoplasma* genomes, we have identified a significant number of SSR, but also small close repeats and large potentially distant repeats. This allows a thorough comparative analysis of these genomes, potentially identifying different strategies for the creation of variability, engaging different repeats in different types of recombination. This is important for an understanding of the pathogenicity and ecology of bacteria. We have analysed the complete genomes of *M.pneumoniae*, *Mycoplasma genitalium*, *Ureaplasma urealyticum* and *Mycoplasma pulmonis*. Although repeats have been experimentally studied in these genomes, no study has been done to exhaustively analyse all kinds of repeats. *Mycoplasma genitalium* and *U.urealyticum* primarily colonise the human urinogenital tract, whereas *M.pneumoniae* is the aetiological agent of community-acquired pneumonia. A very similar respiratory disease is caused by *M.pulmonis* in rats and mice, hence its use as a model. *Mycoplasma pulmonis* also colonises other body sites in rodents, including the urinogenital tract. *Mycoplasma pulmonis* belongs to the phylogenetic branch of *Mycoplasma hominis*, whereas the other species are related to *M.pneumoniae* (24). Mycoplasmas have frequently been given as examples of minimal genomes. Their small genome size renders them dependent on the host for many essential molecules, such as amino acids, lipids and precursors of nucleic acids. Under pressure for genome size minimisation, the existence of a high number of repeats is highly significant, indicating that their conservation/generation has been subject to more intense selection than the many metabolic functions that were lost.

MATERIALS AND METHODS

Genome data mining

The data on the complete genomes of *M.pulmonis* (964 kb, 27% G+C) (7), *M.genitalium* (580 kb, 32% G+C) (4), *M.pneumoniae* (816 kb, 40% G+C) (5) and *U.urealyticum* (752 kb, 26% G+C) (6) were downloaded from Entrez genomes (http://www.ncbi.nlm.nih.gov). The analysis of *M.pulmonis* was facilitated by access to a dedicated web server (http://Genolist.pasteur.fr/MypuList). The *vsa* locus of strain KD735-15 was taken from GenBank (accession no. U23947).

Identification of simple sequence repeats

We searched for motifs X of length 1–5 nt (e.g. 2 in CG) with n consecutive copies (e.g. 3 in CGCGCG) for n high enough that X_n should not occur by chance in the genome. Considering L, the length of the genome, the probability of not finding X_n anywhere in the genome is:

$$P = (1 - f_X^n)^L \quad \mathbf{1}$$

where f_X is the relative frequency of motif X in the genome. Setting a threshold $P < 0.001$, we solved the above equation for all possible motifs X of length 1–5 nt, determining the threshold length for each motif. We then searched for significant SSR elements in all *Mycoplasma* genomes using standard pattern matching methods.

Identification of large repeats

We also searched the genomes for large repeats. To compute the threshold length, we used a statistic of extremes that takes into account the composition in nucleotides and the length of the genome (25). For a large sequence (of length N) we expect the length (L_2) of the largest repeat present at least twice in the genome to have a Gaussian distribution with mean and variance given by:

$$E(L_2^{(N)}) = [\log(N_2)/-\lambda] - \langle\{[\log(1 - \textstyle\sum_{j=1}^{4} p_j^2) + \lambda]/\lambda\} + (0.5772/\lambda)\rangle + 0.5 \quad \mathbf{2}$$

$$\mathrm{var}(L_2^{(N)}) = 1.645(1/\lambda)^2 \quad \mathbf{3}$$

$$\lambda = \log(\textstyle\sum_{j=1}^{4} p_j^2) \quad \mathbf{4}$$

where p_j is the relative frequency of nucleotide j in the sequence. For bacteria this value is in the range 21–26 (for $P < 0.001$) (18), which coincides with the minimal region of strict homology required for homologous recombination in *E.coli* and *B.subtilis* (26). The search for large strictly identical repeats was done using Reputer (27), which outputs all pairs of repeats larger than or equal to the threshold length. Repeats were clustered and classed into direct doublets (DR), inverse doublets (IR), i.e. occurring in different DNA strands, and multiple repeats (MR), occurring in more than two copies, frequently in both DNA strands.

Analysis of large repeats

Repeats were cross-compared in order to construct families of similarity. The alignments were performed using a variant of the classical dynamic programming algorithm for global alignment, where one allots 0 weight for gaps at both ends of the largest sequence pattern (28). This variant is used to 'fit' a sequence into a longer one and, therefore, we shall call it a pattern-fit. This algorithm was also used to analyse the

Table 1. Number of elements identified in the *Mycoplasma* genomes for each class of repeats

	SSR	CR	DR	IR	MR	D/I
Mycoplasma pulmonis	18	29	653	68	3	6.5
Ureaplasma urealyticum	10	8	72	10	0	17.8
Mycoplasma genitalium	14	11	645	19	1	57.2
Mycoplasma pneumoniae	6	36	2476	303	1	9.3

SSR, simple sequence repeats; CR, close repeats; DR, direct repeats; IR, inverse repeats. DR and IR include repeats present in multiple copies. D/I is the ratio of the cumulative lengths of direct and inverse repeats.

sequence conservation between *P1* and *mgpB* shown in Figure 3. In this case we used a sliding window (60 bp with 10 bp steps) on *mgpB* making successive pattern-fits in *P1*. This results in a curve indicating the local similarity between the two sequences.

Identification of close repeats

Close repeats at short distances (typically <1000 bp) may engage in illegitimate recombination. Hence, we searched for repeats in sliding windows of 1000 bp. For each window we determined the length threshold using the above formulae (equations **2**–**4**) (for $P < 0.001$). These values varied slightly from window to window (as a function of the window composition), but typically ranged from 12 to 14 bp. Then we searched for more distinctive close repeats, by identifying repeats with occurrences at <50 bp apart and for those with copies distant by less than three times their length. The exact edges of the close repeats were identified using dotter, a program dedicated to building large dot-plots among sequences (29). When possible, we defined a consensus for the close repeat using multiple alignments in both the DNA and protein (when the repeat is in a gene).

RESULTS AND DISCUSSION

Overview of results

We found a large number of repeats of each type in the four mycoplasmas (Table 1), which suggests the existence of a large potential for recombination in these genomes. It is also clear that different bacteria possess different recombination potentials, as well as specific emphases on certain types of repeats. Repeats engaging in illegitimate recombination are always close, whereas large repeats are dispersed on the chromosome. As a consequence, the distribution of such repeats on the chromosomes is very different (Fig. 1). As we shall discuss, this may have important consequences for stability of the chromosome. The complete list of the repeats found in the genomes can be consulted at http://wwwabi.snv.jussieu.fr/~erocha/mycoreps/.

SSR. The sequencing of the *M.pulmonis* genome revealed several localised sequence polymorphisms, which sometimes hampered our initial attempts at sequence assembly. These highly variable regions correspond to SSR elements or close repeats identified by our analysis (Table 2). Indeed, almost all of the identified SSR of mononucleotide motifs corresponded to polymorphic sites at the sequencing stage (7). As expected, most of these elements are within intergenic regions and may therefore be involved in phase variation of the gene lying 3′ of the repeat. Some elements are quite distant from the start of the genes, e.g. –115 nt for the T_{25} located before the *S1* gene and –191 nt for the SSR before oligo-1,6-glucosidase. The phase variation of these genes, as well as of *tsr* (an aldolase) in *U.urealyticum*, is dubious. SSR elements far from a functional start are common in the four mycoplasmas. However, it should be noted that very little is known about *Mycoplasma* promoters and a recent study indicated that a large distance (>100 nt) between the promoter and the first base of the start codon is not uncommon in *M.pneumoniae* (30). Also, some of these elements are associated with pseudogenes, with strong similarity to proteins such as P1 in *M.pneumoniae* and MgpB in *M.genitalium*.

SSR elements are scarcer in *U.urealyticum* and *M.pneumoniae* than in *M.pulmonis* or *M.genitalium* (Table 1). Among the genes putatively regulated by phase variation in the former genomes, most correspond to lipoproteins, unknown function ORFs (UFO) or restriction and modification systems (RMS). Although abundant, most SSR of *M.genitalium* are composed of trinucleotides inside genes or in intergenic regions flanked by two stops, and only a few genes contain SSR potentially mediating phase variation in this genome. An intriguing example is the *rpoA* gene, which contains a $CAAC_3$. Another gene that would not be expected to present an SSR is the ribosomal protein L7 of *U.urealyticum*. Naturally, frequent slippage of these small SSR must be tested further by experimental work.

Close repeats. The genome of *M.pulmonis* contains a large number of close repeats, ranging from large tandem motifs of 6 nt up to full gene multiplications (e.g. the three tandem genes Mypu_3160–Mypu_3180). With rare exceptions, these elements are in coding sequences (Table 3 and additional web material) and especially in lipoprotein coding genes (see below). *Ureaplasma urealyticum* contains fewer important close repeats. Interestingly, one of these is in the *nusG* gene, which codes for a transcription anti-termination factor (Table 3). The others concern GTP-binding proteins, RMSs and the MBA protein (discussed below). In *M.genitalium* only a lipoprotein presents a clearly identified close repeat (MG309). In contrast, a large number of close repeats were identified in *M.pneumoniae*, of which 10 have well conserved motifs. Most of these repeats are associated with lipoproteins, UFOs and RMSs. The reason for such a discrepancy between two closely related bacteria is intriguing.

Large repeats. The three families of large repeats found in *M.pulmonis* correspond to a set of lipoproteins, a set of RMSs and a set of UFOs. Similarly, most of the direct and inverse two-copy repeats were found to be within lipoproteins (25%), within UFOs (33%) or in intergenic positions (11%, which correlates with the frequency of intergenic sequence, 9%). In *U.urealyticum* most large repeats concern UFOs (28%), transport/ATP-binding proteins (28%), RMSs (11%) and the rRNA operon. There are no repeats occurring more than twice in this genome. In *M.genitalium* the most relevant large repeats concern a multiplet of copies of fragments of the gene coding for an adhesin (*mgpABC* operon). Interestingly, the putative

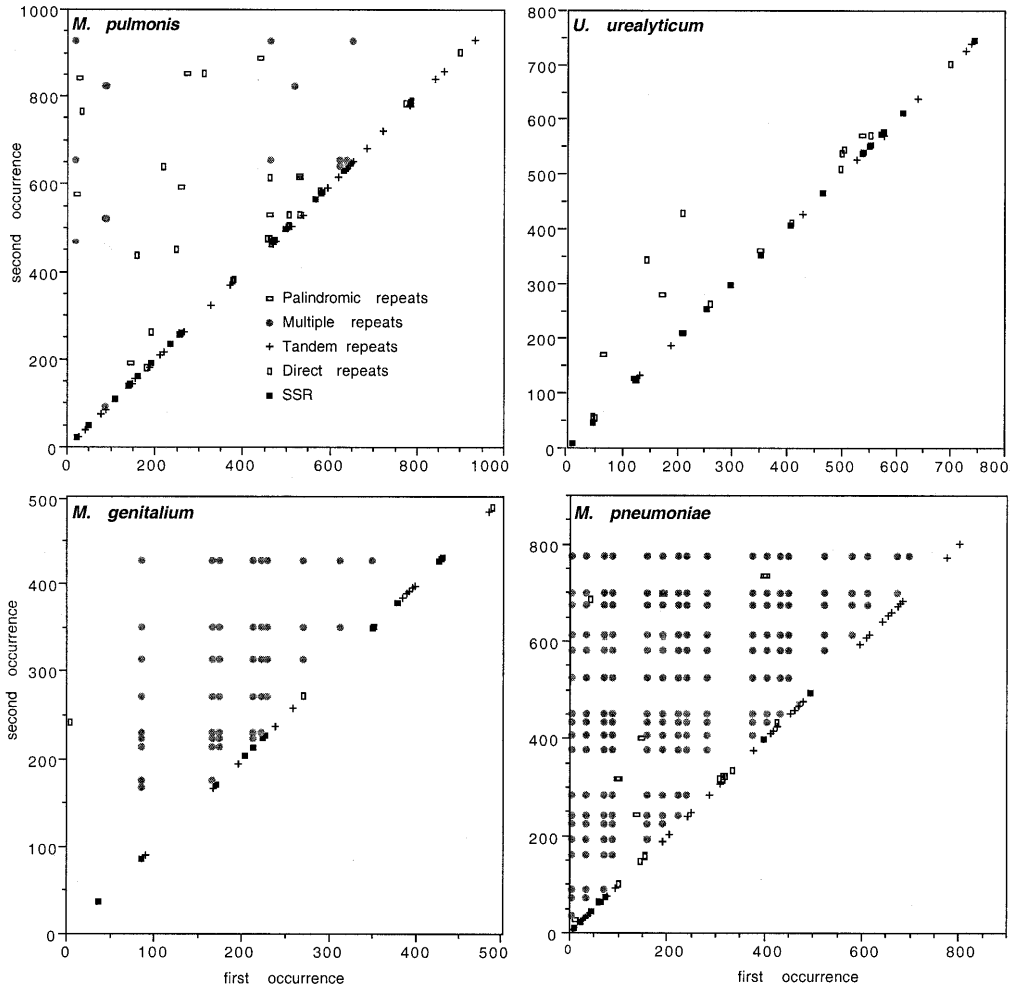

Figure 1. Distribution of repeats in the four *Mycoplasma* genomes. Repeats were classed into five classes, SSRs, close repeats and three classes for large repeats (inverse, direct and multiple). The *x*-axis represents the position of the first occurrence of the repeat in the chromosome and the *y*-axis represents the position of the second occurrence. Naturally, the two copies overlap for SSR and almost do so for close repeats. For multiple repeats all possible pairs among the group are shown.

M.pulmonis products (Mypu_0220/6920), homologous to the MgpB adhesin, also contain large repeats. Finally, there is a very significant number of large repeats in *M.pneumoniae*, mostly concerning UFOs or pseudogenes (77%) and lipoproteins (16%). There is also a large multiplet constituting the P1 adhesin, which occupies a large fraction of the genome (see below).

Variation of immunodominant proteins

In each of the four mycoplasmas some proteins are immunodominant antigens and they seem to be particularly exposed to variation strategies: the Vsa lipoproteins of *M.pulmonis*, the MBA lipoprotein of *U.urealyticum*, the P1 adhesin of *M.pneumoniae* and the Mgp adhesin of *M.genitalium*. Interestingly, the two lipoproteins and the two adhesins are found to have a very different number and type of repeats (Table 4). We have thus tried to understand the basis of these differences.

Sequence variation in the vsa genes. The *vsa* locus encodes highly variable surface lipoprotein antigens, which in strain KD735-15 have been shown to engage in high frequency, site-specific DNA inversions (31), possibly mediated by the nearby recombinase found in the UAB CTIP strain sequence (7). Only the *vsa* gene located in the expression site is transcribed at a given time and the inversions are proposed to follow a model involving DNA strand exchange at conserved *vsa* recombination sites (31). Since this particular locus has been sequenced in two different strains, we performed a comparison between the strains, which have a different origin and host. The UAB CTIP strain is a low passage isolate that is highly virulent in mice and strain KD735-15 is of unknown virulence, but was derived by serial cloning from strain 6510C isolated from infected rats (31). The basic features of repeats in each *vsa* locus have been described elsewhere (7,31) and here we focus on an analysis of the differences between the two strains. The gene order, the gene number and the number of repeated units in individual genes are very different between the two strains (Fig. 2). The genomic sequence in strain UAB CTIP contains seven *vsa* genes, of which only four (*vsaA*, *vsaC*, *vsaE* and *vsaF*) were previously described in strain KD735-15. However, since the KD735-15 locus has not necessarily been completely sequenced, we cannot conclude

Table 2. SSR elements found in the four *Mycoplasma*

Repeat	Location[a]	Gene potentially affected	Gene function	Putative effect of variation[c]
Mycoplasma pulmonis				
T_{31}[b]	I (–16)	Mypu_0190	Lipoprotein	Phase variation
A_{45}[b]	I (–191)	Mypu_1030	Oligo-1,6-glucosidase	Phase variation
T_{25}[b]	I (–115)	S1	Ribosomal protein	Phase variation
T_{23}	I[d]	Mypu_6510–6520	UFO–lipoprotein	Unknown
T_{39}[b]	I (–61)	Mypu_4780	Lipoprotein	Phase variation
A_{20}[b]	I (–165)	*dnaK*	Heat shock	Phase variation
CA_{14}	G (+21)	Mypu_0430	C-methyltransferase	Phase variation
$AG_{9.5}$	G (+208)	Mypu_1850	C-methyltransferase (fragment?)	Unknown
AG_8	G (+1699)	Mypu_3960	RMS type III	Specificity variation
AG_{12}	G (+1675)	Mypu_3970	RMS type III	Specificity variation
AG_7	G (+1701)	Mypu_3980	RMS type III	Specificity variation
AG_{13}[b]	G (+320)	Mypu_4800	RMS type III	Specificity variation
TGT_5	G (+1165)	Mypu_1520	UFO	Unknown
ACA_5	G (+456)	Mypu_4150	Lipoprotein	Sequence variation
$GGA_{3.3}$	G (+1861)	Mypu_5170	UFO	Unknown
TAC_4[b]	G (+950)	Mypu_2130	UFO	Unknown
TTA_4	I[d]	Mypu_6510–6520	UFO–lipoprotein	Unknown
$TTAA_4$[b]	I[d]	Mypu_4660-rpL19	UFO–ribosomal protein	Unknown
Ureaplasma urealyticum				
T_{19}	I[d]	UU497-*rpS4*	UFO–ribosomal protein	Unknown
AT_{10}	I (–30)	UU263	Lipoprotein	Phase variation
AT_{11}	I (–20)	UU474	UFO	Phase variation
AT_9	I[e] (–125/–232)	UU171–172	UFO–MBA N-terminal paralog	Phase variation
AT_9	I[d]	UU403-tnp-2	UFO–integrase/recombinase	Unknown
AG_6	G (+193)	hsdS-2	Type I RMS S subunit	Specificity variation
AG_{10}	G (+328)	UU478	UFO	Sequence variation
GCT_4	G (+123)	*rpL7*	Ribosomal protein	Sequence variation
$ATAAT_3$	I (–46)	UU208	UFO	Phase variation
$ATAAT_3$	G (+872)	*tsr*	Fructose bisphosphate aldolase	Sequence variation
Mycoplasma genitalium				
T_{17}	I[e] (–77/–171)	MG031–032	UFO–UFO	Phase variation
GTA_{18}	I (–6365)			Unknown
AGT_7	G (+2963)	*mgpB*	Adhesin	Sequence variation
AGT_{11}	G (+1223)	*mgpC*	Adhesin	Sequence variation
TGA_{10}	I[d]	MG287–MG288	Acyl carrier–UFO w/frameshift	Unknown
AGT_5	G (+407)	MG307	Lipoprotein	Sequence variation
$TGA_{9.6}$	I (–1272)	*rpoB*	RNA polymerase	Phase variation
CTT_5	I (–2160)	MG069	PTS system	Phase variation
AAG_5	I[d]	MG287–288	Acyl carrier–transport	Unknown
AAG_{16}	I (–1852)	*recA*	Recombination	Phase variation
ACA_{11}	G (+1091)	MG338	Lipoprotein	Sequence variation
$CAAC_3$	G (+38)	*rpoA*	RNA polymerase	Phase variation
$AAACA_4$	G (+1489)	MG185	Lipoprotein	Phase/sequence variation

Table 2. *Continued*

Repeat	Location[a]	Gene potentially affected	Gene function	Putative effect of variation[c]
Mycoplasma pneumoniae				
A_{16}	I (–400)	MPN006	UFO	Phase variation
A_{16}	I (–195)	MPN405	Lipoprotein	Phase variation
AG_{22}	I (–623)	MPN045	UFO	Phase variation
AGT_7	G (+2928)	P1	Adhesin	Sequence variation
ACT_5	G (+66)	MPN052	UFO	Sequence variation
$AAAGC_3$	G (+734)	*uvrC*	DNA repair	Sequence/phase variation

The sequence of the repeat is given from the DNA strand of the gene when it falls within a gene or its putative regulatory region. Elsewhere the sequence is on the direct (published) strand.
[a]G or I indicates that the repeat falls within a gene or between two genes, respectively, and is followed by distance to the presumed translation start. A negative or a positive value indicates that the repeat is located upstream or downstream from this start, respectively.
[b]*Mycoplasma pulmonis* repeats exhibiting confirmed polymorphisms during genome sequencing.
[c]The putative phase variation of the listed genes is solely indicated on the basis of a repeat located upstream of a coding sequence and does not take into account the distance between the repeat and the start codon.
[d]Intergenic repeats 3′ of convergent genes.
[e]Intergenic repeats 5′ of divergent genes.

that there is a deletion in this strain. On the other hand, the latter contains 11 *vsa* genes, four of which are absent in UAB CTIP. Furthermore, the expressed gene in UAB CTIP is the *vsaI* gene, whereas in KD735-15 it is the *vsaA* gene.

All genes presenting repeats in the two strains exhibit either contractions or expansions of their sequences. For example, the *vsaC* gene is present in a single copy in the sequenced strain, where it presents a 36 bp motif repeated 62 times, but in three copies in the KD735-15 strain, presenting the motif in a significantly different number of copies: 8, 40 and 11. Similarly, the *vsaA* gene in UAB CTIP contains a 51 bp motif in 16 tandem copies, whereas in the KD735-15 strain, where it is the expressed gene, there are 33. It would be interesting to determine whether expressed genes are typically larger than the homologous silent genes. Selection for such larger repetitive variants of one such gene was suggested for the Vlp surface lipoproteins of *Mycoplasma hyorhinis*. In this case it has been proposed that a larger protein could more efficiently 'shield' other exposed proteins that are not so free to change (32). Unfortunately, the *vsaI* gene expressed in the UAB CTIP strain is not present in the KD735-15 published sequence, which precludes a comparative analysis of this gene in the two strains.

From a complete genome point of view, and even though the *vsa* locus has been the subject of several studies, it is surprising to observe the diversity of strategies involved in its variation. These include illegitimate recombination, mediated by tandem repeats, site-specific recombination, mediated by a recombinase, and homologous recombination between genes in multiple copies. As we shall see, such variation is much more constrained for the evolution of the immunodominant proteins of *M.pneumoniae* and *M.genitalium*.

Epitope mapping and repeats in the M.pneumoniae and M.genitalium main adhesins. The membrane protein P1 is localised primarily to the attachment organelle of *M.pneumoniae* and is thought to have a major receptor-binding role in *M.pneumoniae* (1). Early observations concerning the existence of several regions homologous to the *P1* gene scattered along the chromosome (33,34) were interpreted as a strategy for antigenic variation through homologous recombination. Hence, we have tried to evaluate the association between the distribution of the repeats and the regions of the protein reacting with IgG serum antibodies from *M.pneumoniae*-infected patients (35). If repeats are selected for antigenic variation, one should expect a larger density of repeats in regions of the protein that overlap with the identified epitopes. We tested this hypothesis by computing the number of repeats found in the genome corresponding to each part of the *P1* gene. The unexpected result is that most of the HAb seem to recognise epitopes located in regions that do not present repeats (Fig. 3). However, these repeated regions have been found to be potentially immunogenic, as evidenced by mAb recognition (35).

Within the current paradigm, two hypotheses can be proposed to explain this apparent contradiction. (i) The repeated regions are so highly variable among *M.pneumoniae* strains that antibodies generated in patients may be specific for another 'form' of the P1 polypeptide. However, such extensive variability is not in accordance with published data (3). (ii) The repeated sequences code for regions displaying conformational rather than linear epitopes; the conformational epitopes would not be detected by epitope mapping using overlapping peptides in the screening assay (35). However, this fails to explain the absence of repeats in the regions that are indeed recognised by the antibodies. Outside the antigenic variation paradigm, the existence of repeats within the cell recognition domains paves the way for diversity of these domains. Two reasons could be given for this: (i) bacteria strive to avoid the action of antibodies that would impede correct recognition of cellular receptors; (ii) it could constitute a way of altering the repertoire of its receptors. Hence, the existence of repeats would result from the needs of shifting ecological niches of *M.pneumoniae*.

The MgpA protein of *M.genitalium* (identified as MgpB in the genome sequence) is the equivalent of the P1 adhesin, with which it shares strong sequence and structural resemblance (36). Surprisingly, when we analysed the density of repeats

Table 3. Most significant DNA close repeats found within coding regions in the four *Mycoplasma* genomes

Position	Length	Motif[a]	Genes	Gene product	Associated amino acid motif
Mycoplasma pulmonis					
75281	179	33 × 5	MYPU_0680	UFO	ILEAQGQREAA
593130	371	63 × 6	MYPU_4870	Lipoprotein	ILEAQGQREAA
640375	826	51 × 16	*vsaA*	Lipoprotein	TPPTTGSGSGSTDTKPQ
641650	1010	33 × 29	*vsaF*	Lipoprotein	PPAKDGDTMAT
644800	351	36 × 61	*vsaC*	Lipoprotein	ANTSQTPSTTTG
647200	1350	33 × 29	*vsaG*	Lipoprotein	NPPAPGGDTMT
650000	1601	57 × 27	*vsaI*	Lipoprotein	PQGNGQMDNQGNDQMGNTN
652750	101	48 × 2	*vsaH*	Lipoprotein	PPADTPAPG
Ureaplasma urealyticum					
125835	48	12 × 3.5	*hsdS3*	RMS	TELE
427328	790	18 × 44	MBA	MBA	GKEQPA
525929	24	9 × 2.5	*obg*	GTP-binding protein	GGN
727188	27	9 × 3.5	*nusG*	Transcription antitermination	IAK
Mycoplasma genitalium					
384461	33	6 × 5/6	MG309	Lipoprotein	NS
Mycoplasma pneumoniae					
4393	100	30 × 3.5	MP002	Lipoprotein	QNQGKKGEGA
27173	163	21 × 7	MP016	UFO	KVDKLEE/A
28168	170	21 × 8	MP017	UFO	DSVEGRL
93812	69	12 × 5	*hsdS*	RMS	ELSA
376502	138	21 × 6	MP318	UFO	[KR][MI]DKMEX
413453	147	21 × 4	MP342	UFO	QGEQI[NK]XL
611990	201	12 × 16	*hsdS*	RMS	ELSA
674527	50	12 × 4	*hsdS1B*	RMS	ELSA
681320	189	12 × 15	*prrB*	RMS	ELSA
775044	24	6 × 4	MP626	P1-like adhesin precursor	QP

The motif characterization is of the form (A × B), where A is the length of the motif and B its multiplicity (e.g. 21 × 2) for a 21 bp motif repeated twice.

along the gene, we observed some important differences (Fig. 3). Although both genes present few repeats at the C-terminus (which corresponds to hydrophobic regions of these proteins), several differences are noticeable. (i) The N-terminus and the region between amino acids 1000 and 1100 is not repeated in MgpB. (ii) The region around amino acid 600 is highly repeated in MgpB, although it is one of the most conserved among the two sequences. This is also the most responsive region in P1 to patient antibodies. Hence, the comparison of the main adhesins of *M.genitalium* and *M.pneumoniae* provides puzzling results, since the region that was predicted to vary in P1 is not repeated, whereas it is in MgpB, for which a similar study using patient antibodies has not been published, at least to our knowledge. If the variation in these proteins is related to tropism variation, the different types of tissues that these bacteria colonise could explain these results.

Variation in lipoproteins. Besides the *vsa* and *MBA* loci, lipoproteins constitute the largest group of genes containing repeats in the four mycoplasmas (Table 4). It is in *M.pulmonis* that we find the largest number of repeated elements falling in lipoproteins (37 genes), which contrasts significantly with the values for *M.pneumoniae* (9), *U.urealyticum* (8) and *M.genitalium* (6). Taking into account the different number of lipoproteins in these genomes, the frequency of repeated elements in *M.pulmonis* lipoproteins is still significantly larger than in the other genomes ($P < 0.001$, χ^2 test). This higher number of potentially variable surface proteins may be related to the ability of *M.pulmonis* to colonise different body sites in at least two rodents (rats and mice), the other mycoplasmas having more restricted ecological niches.

The set of lipoproteins contains all types of repeats analysed in this study, although close and long repeats dominate in *M.pulmonis* and only SSR were identified in *M.genitalium*. Lipoproteins are surface antigens with a potential modulin activity and are preferential targets of the host immune response (8). Therefore, it is not surprising to observe such a recombination potential in this category of genes. Among the

Table 4. Repeats found in the genes encoding immunodominant antigens and lipoproteins

	Immunodominant proteins				Lipoproteins				
	Protein	SSR	CR	DR+IR	No.	SSR	CR	DR+IR	MR
MYPU	Vsa	no	yes	yes	66	4	14	18	1
URUR	MBA	no	yes	no	38	2	1	5	0
MYGE	MgpB	no	no[a]	yes	20	3	2	0	1
MYPN	P1	yes	yes[b]	no	44	1	2	5	1

SSR, simple sequence repeats; CR, close repeats; DR, direct repeats; IR, inverse repeats; MR, multiple repeats.
[a]Two tandem repeats are found in regions coding for fragments of MgpB (nt numbering 86592 and 168175). However, these repeats are degenerate and no CR pattern is distinguishable.
[b]Only one tandem repeat (6 nt × 4) exists within an ORF, probably a pseudogene, putatively encoding a P1 protein homologue.

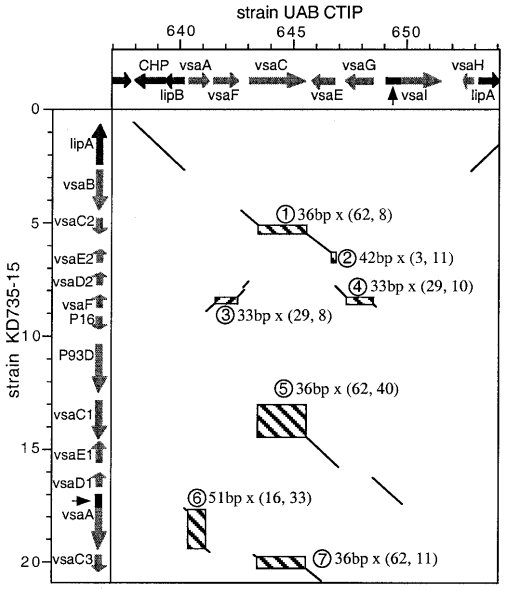

Figure 2. Comparison of the *vsa* locus of the UAB CTIP and KD735-15 strains of *M.pulmonis*. Boxes indicate regions of SSR common to both sequences and diagonal lines regions of high similarity between the two loci. The legends to boxes are of the form A × (B, C), where A indicates the length of the repeated motifs and B and C represent the number of consecutive motifs identified in the UAB CTIP strain and in the KD735-15 strain. The expression site is indicated by an arrow. The gene nomenclature is taken from the original papers describing these loci (7,31). CHP, conserved hypothetical protein.

potential phase variation cases in *M.pulmonis* lipoproteins, two highly related lipoproteins (Mypu_0190 and Mypu_4780) possess a poly(A) tract lying upstream from the start. The phase variation hypothesis is strongly supported by the sequencing, which indicated that the length of these poly(A) tracts was variable (7). A similar polymorphism has been shown to regulate the phase variation of *M.hyorhinis* lipoprotein expression (37,38). An interesting case is provided by the lipoproteins MG307 of *M.genitalium* and MP405 of *M.pneumoniae*, which are very similar in sequence (65% amino acid similarity) but which seem to have adopted different strategies for antigenic variation. Indeed, MG307

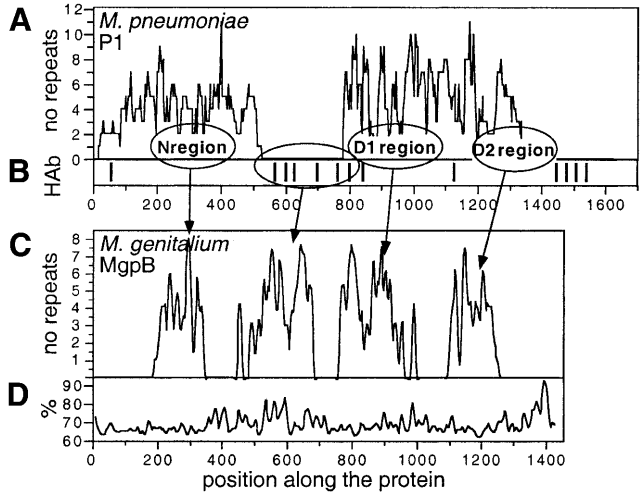

Figure 3. Analysis of the potential for homologous recombination between the P1 and MgpB proteins and their pseudogenes scattered on the chromosome. The *y*-axis represents the number of repeats larger than 25 bp found elsewhere in the genome for sliding windows of 60 bp (10 bp steps) along the genes coding for P1 (**A**) and MgpB (**C**). HAb (**B**) indicates the positions of the binding sites of human IgG serum antibodies of patients suffering from *M.pneumoniae* disease (35). Similarity between *mgpB* and *P1* at the DNA level (**D**) was computed in sliding windows of 30 bp (10 bp steps) of the *mgpB* gene (see Materials and Methods for details).

contains a repeat of AGT codons, which could result in regulation at the translation level, whereas MP405 contains an upstream intergenic poly(A) SSR of 16 nt, which could result in transcriptional repression. Naturally, these strategies have different effects on transcription of the downstream genes in the operon.

Strategies for fast evolution in *Mycoplasma*

Variation, response to stress and repair. Bacteria subject to periodic selection for antigenic variation use phase and sequence variation to adapt. However, many natural populations of pathogenic and commensal bacteria include a small percentage of mutator elements that have a higher probability of finding an adaptive mutation in a maladapted population (39). Mycoplasmas have high mutation rates and lack part of the SOS response and also a considerable number of the DNA repair proteins existing in *E.coli* and *B.subtilis* (1). In particular, the four sequenced genomes reveal absence of the MutSLH system for mismatch repair, which in *E.coli* constitutes a barrier to recombination between divergent sequences and whose inactivation is the basis of most mutator phenotypes in enterobacteria (40). In this context, mycoplasmas could be regarded as constitutive mutators, with predictable higher frequencies of recombination between divergent sequences. Furthermore, absence of the MutSLH system has been shown to induce SSR instability in *E.coli* (41). This probably increases the recombination potential between close repeats in *Mycoplasma*, as seen in the *vsa* locus of *M.pulmonis*.

We have found several large SSR in genes capable of inducing a mutator phenotype when inactivated. For example, the *U.urealyticum rpL7* gene contains a stretch of GCT codons coding for a poly-alanine sequence. This protein seems to be essential for accurate translation. Its alignment with the *B.subtilis* and the *E.coli* homologues reveals that the poly-alanine

region was formed by point mutations, not indels (data not shown). The genes *rpoA* of *M.genitalium* and *uvrC* of *M.pneumoniae* also reveal a potential for variation, as well as several topoisomerases and gyrases in all four genomes. Together with lack of the MutSLH system, inactivation of these genes could contribute to a mutator phenotype for the mycoplasmas.

More puzzling is the presence of an SSR before the gene encoding the chaperone DnaK. This SSR touches a CIRCE (for controlling inverted repeat of chaperone expression) element that is important for the induction of expression of eubacterial heat shock proteins and has also been identified upstream of *dnaK* of *M.pneumoniae* (42). The repressor HrcA (Mypu_1420) is supposed to interact with the CIRCE element to control σ^A factor-dependent expression of *dnaK*. Since the transcriptional start of *dnaK* in *M.pulmonis* is supposed to be upstream of the CIRCE sequences, it would fall in the poly(A) tract. Length variation of this repeat would eventually provide the mycoplasma with an additional level of regulation of *dnaK* expression by phase variation. The advantage that would result from such a mechanism is as yet unknown. However, it is known that DnaK and trigger factor (Mypu_2090 in *M.pulmonis*) cooperate in *de novo* protein folding (43) and *E.coli* can tolerate the loss of either chaperone.

Variation in restriction and modification systems. *Mycoplasma pulmonis* possesses the most complex *hsd* genes described so far (44). The UAB CTIP strain presents three *hsd* loci, one of which seems not to be functional (7). Each of the functional loci encodes specificity (S), methylase (M) and endonuclease (R) subunits highly homologous to known type I RMS. Both loci include site-specific inversion systems containing two *hsdS* genes encoding two different specificity S subunits. We observed an extensive number of repeated elements in other RMSs in *M.pulmonis*, *U.urealyticum* and *M.pneumoniae* (Tables 2 and 3). In fact, the data suggests sequence variation to be of extreme importance for these systems, since they include all types of repeats, i.e. SSR, close repeats and long repeats, besides site-specific recombination sites. We have found that all the repeated elements in *hsd* elements in all genomes are in the S subunit. In type III RMS such repeats are always in the methylase subunit. This subunit includes the sequence coding for sequence-specific recognition and this strongly suggests selection of variability at this level. Even though no functional RMS has been identified in *M.pneumoniae*, it contains repeats in the corresponding pseudogenes *hsdS* and *hsdS1B*, where a 12 bp motif is repeated four times (Table 3). Similar cases of potentially silent, hypervariable RMS have been reported in the genome of *Helicobacter pylori* (45).

Why have such an elaborate apparatus to change the specificity of the recognition sequence in *Mycoplasma*? A correlation has been observed between inversions of the *hsd* and *vsa* genes in *M.pulmonis* (46) and it has been suggested that the change in specificity of the RMS could induce the establishment of inversions in the *vsa* locus. Naturally, this would not explain the existence of repeats in the RMS of *M.pneumoniae* and *U.urealyticum*, but it is in line with the suggestion that RMS might have evolved to become a sort of bacterial genetic engineering tool (47). Recently, a correlation between RMS expression and colonisation of certain tissues was described and it was proposed that RMS might be involved in host cell DNA degradation (48). Although this hypothesis may seem provocative, it is supported by reports showing that mycoplasmal nucleases can induce cell apoptosis (49,50).

SSR were also found within two *M.pulmonis* CDSs encoding putative CpG-specific methyltransferases (Table 2). Among mollicutes, this type of methylase was first reported in a *Spiroplasma* strain (51) and was designated *Sss*I methylase. This enzyme, similarly to its mammalian counterpart, completely methylates CpG-containing sequences (52). The function of CpG-specific methyltransferases in mollicutes is unknown and its elucidation is further complicated by the fact that *Sss*I methylase also exhibits a type I-like topoisomerase activity (53). It is tempting to suggest that these methylases could play a role in the regulation of gene expression in mollicutes, which are known to have a reduced machinery for transcriptional regulation. Analysis of sequenced *Mycoplasma* genomes revealed that both *U.urealyticum* and *M.pulmonis* have genes encoding CpG methylases, although functional identification of the corresponding enzymes has not yet been reported. As evidenced by a high degree of identity with *Sss*I methylase and by the presence of a motif shared by C-5 cytosine-specific DNA methylases, two CpG methylase genes (Mypu_0430 and Mypu_1860) appear functional in the *M.pulmonis* genome. Both are located next to truncated homologous genes (Mypu_0440 and Mypu_1850). Interestingly, Mypu_0430 and Mypu_0440 overlap, with the end of the overlap corresponding exactly to the SSR identified in our analysis (CA_{14}, Table 2). This organisation suggests that mutational frameshifting may occur in this region, resulting in fusion of the two CDSs. The possible consequence of this variation would be a change in the transcriptional regulation of genes regulated by CpG methylation. In eukaryotes, where such regulation is known to happen, CpG is a mutational hotspot and therefore a rare dinucleotide in regions of low gene densities (54). We have therefore analysed the dinucleotide content of 60 completely sequenced bacteria to identify genomes where CpG is strongly avoided. Taking into account the nucleotide composition of genomes, as described in Rocha *et al*. (55), we found that *M.pulmonis* is the bacterial genome presenting the highest avoidance of CpG, *M.genitalium* being the third and *U.urealyticum* the sixth (data not shown). This is consistent with an important impact of CpG methylases on these genomes.

Repeats versus genome stability

Recombination events may conflict with genome stability by causing chromosomal rearrangements. Among the currently sequenced small bacterial genomes (<1 Mb), mycoplasmas exhibit the highest density of large repeats (56). For example, the density of large repeats in *M.genitalium* and *M.pneumoniae* is six and 17 times higher, respectively, than in *B.subtilis*. These repeats could be particularly susceptible to changes in the chromosome structure. Indeed, the *Mycoplasma* genomes sequenced so far present a poor conservation of gene order, with the exception of *M.pneumoniae* and *M.genitalium*, which have recently diverged (6). This may indicate either weak counter-selection of chromosomal rearrangements or strong selection for repeats (which inevitably produce rearrangements by recombination). However, the genomes of mycoplasmas have an extreme distribution of genes between the two strands of the chromosomal DNA, since ~80% of them are in the

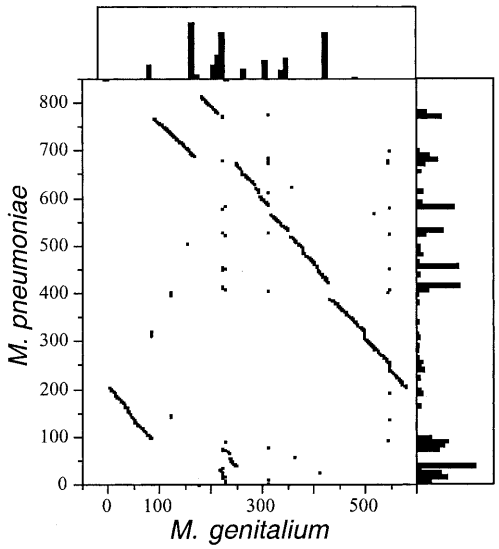

Figure 4. Synteny between the chromosomes of *M.pneumoniae* and *M.genitalium* (inner square) and distribution of large direct repeats capable of engaging in homologous recombination (histograms). Dots represent the position of orthologous genes in the respective genomes.

replicating leading strand, compared to 75% for *B.subtilis*, 59% for *Mycobacterium tuberculosis* and 55% for *E.coli*. Since this suggests a strong counter-selection for repeats capable of engaging in inversions, i.e. long inverse repeats, we computed the ratio between the numbers of direct and inverse repeats observed in mycoplasmas. We found that direct repeats are six to 60 times more frequent than inverse repeats in these genomes (Table 1). These ratios are to be compared to only 1.8 for *B.subtilis* and *M.tuberculosis* and 1.1 for *E.coli* (data not shown). The difference between direct and inverse repeats is particularly large for genomes with a smaller number of repeats, *M.genitalium* and *U.urealyticum*, which might reflect a stronger selection for genome stability.

The probable role of adhesin repeats in DNA translocations has been described for rearrangements between the genomes of *M.pneumoniae* and *M.genitalium* (57). Our statistical analysis further indicates a coincidence between the regions with high densities of large repeats and the edges of regions that have been rearranged or deleted (Fig. 4). Thus, such highly repeated regions constitute recombination hot-spots at the basis of all major chromosomal plasticity events in these genomes (which lack functional insertion sequences). Many of these regions correspond to the partial repeats of the *mgpB* and *P1* genes, but also to lipoproteins and repeated UFO genes. Therefore, genome structure in these species is the result of a trade-off between genome stability and the necessity of having elements engaging in recombination events. The high avoidance of inverse repeats described above could then be the result of selective pressure aimed at avoiding chromosome inversions. Such inversions disrupt chromosomal structure and, given the preference of *Mycoplasma* to code genes on the leading strand, they negatively affect viability of the bacteria.

CONCLUSION

Bacteria can respond to unanticipated challenges because natural selection has selected those that best survived unexpected changes in the past. Although transient mutator phenotypes may suffer second order selection in certain circumstances, it seems that bacterial pathogen adaptation can often be based on the changing of circumscribed 'contingency loci' (22). Recombination in such loci allows for antigenic and tissue tropism variation, but it may have negative consequences, such as replication pausing, in illegitimate recombination hot-spots, and genome rearrangements, by recombination between distant repeats. As a consequence, one may expect to observe large repeats in genomes that do not strongly counter-select against genome rearrangements. Also, if replication pausing is not a major hurdle to bacterial proliferation, SSR and close repeats would be expected to abound. The type of gene that is subject to variation is also probably a determinant of the observed type of repeat. For example, the precise sequence of the *vsa* genes of *M.pulmonis* does not seem to be important for its function and therefore large close repeat motifs do the job efficiently. As a consequence, such proteins evolve freely and should not be regarded as ideal targets for potential vaccines. On the other hand, the major adhesins of *M.pneumoniae* and *M.genitalium* perform an essential function during host colonisation and variability is constrained by functional requirements. In such cases variation by homologous recombination with pseudogenes may be positively selected. Naturally, such pseudogenes are not devoid of selective importance and, therefore, they are not expected to diverge as freely as pseudogenes typically do.

Finally, we have identified a large potential for variation in unknown function ORFs. In genomes suffering pressure for minimality such elements are probably not devoid of functionality. The analysis of gene function taking into account the potential for variation of these ORFs could then be a precious source of information on the many details of *Mycoplasma* virulence and evolution that remain to be elucidated.

ACKNOWLEDGEMENTS

We would like to thank Christine Citti and Guillaume Achaz for carefully reading the manuscript and Alain Viari for earlier discussions. This work was funded by the INRA, the Université Victor Segalen Bordeaux 2 and the Région Aquitaine.

REFERENCES

1. Razin,S., Yogev,D. and Naot,Y. (1998) Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.*, **62**, 1094–1156.
2. Krause,D.C. and Balish,M.F. (2001) Structure, function and assembly of the terminal organelle of *Mycoplasma pneumoniae*. *FEMS Microbiol. Lett.*, **198**, 1–7.
3. Kenri,T., Taniguchi,R., Sasaki,Y., Okazaki,N., Narita,M., Izumikawa,K., Umetsu,M. and Sasaki,T. (1999) Identification of a new variable sequence in the P1 cytadhesin gene of *Mycoplasma pneumoniae*: evidence for the generation of antigenic variation by DNA recombination between repetitive sequences. *Infect. Immun.*, **67**, 4557–4562.
4. 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.
5. Himmelreich,R., Hilbert,H., Plagens,H., Pirki,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. 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.
7. Chambaud,I., Heilig,R., Ferris,S., Barbe,V., Samson,D., Galisson,F., Moszer,I., Dybvig,K., Wroblenski,H., Viari,A., Rocha,E.P.C. and Blanchard,A. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.*, **29**, 2145–2153.
8. Chambaud,I., Wroblewski,H. and Blanchard,A. (1999) Interactions between mycoplasma lipoproteins and the host immune system. *Trends Microbiol.*, **7**, 493–499.
9. Rosengarten,R., Citti,C., Glew,M., Lischewski,A., Droesse,M., Much,P., Winner,F., Brank,M. and Spergser,J. (2000) Host-pathogen interactions in mycoplasma pathogenesis: virulence and survival strategies of minimalist prokaryotes. *Int. J. Med. Microbiol.*, **290**, 15–25.
10. Baseggio,N., Glew,M.D., Markham,P.F., Whithear,K.G. and Browning,G.F. (1996) Size and genomic location of the pMGA multigene family of *Mycoplasma gallisepticum*. *Microbiology*, **142**, 1429–1435.
11. Saunders,J.R. (1995) Population genetics of phase variable antigens. In Baumberg,S., Young,J.P.W., Wellington,E.M.H. and Saunders,J.R. (eds), *Population Genetics of Bacteria*. Cambridge University Press, Cambridge, UK, pp. 247–268.
12. Michel,B. (1999) Illegitimate recombination in bacteria. In Charlebois,R.L. (ed.), *Organization of the Prokaryotic Genome*. ASM Press, Washington, DC, pp. 129–150.
13. Lloyd,R.G. and Low,K.B. (1996) Homologous recombination. In Curtiss,R., Edmund,J.L.I., Lin,C.C., Brooks Low,K., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, Washington, DC, pp. 2236–2255.
14. Pierce,J.C., Kong,D. and Masker,W. (1991) The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7. *Nucleic Acids Res.*, **19**, 3901–3905.
15. Peeters,B.P., de Boer,J.H., Bron,S. and Venema,G. (1988) Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. *Mol. Gen. Genet.*, **212**, 450–458.
16. Lovett,S.T., Gluckman,T.J., Simon,P.J., Sutera,V.A. and Drapkin,P.T. (1994) Recombination between repeats in *E. coli* by a recA-independent, proximity-sensitive mechanism. *Mol. Gen. Genet.*, **245**, 294–300.
17. Chédin,F., Dervyn,E., Ehrlich,S.D. and Noirot,P. (1994) Frequency of deletion formation decreases exponentially with distance between short direct repeats. *Mol. Microbiol.*, **12**, 561–569.
18. Rocha,E.P.C., 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.
19. Hughes,D. (2000) Co-evolution of the tuf genes links gene conversion with the generation of chromosomal inversions. *J. Mol. Biol.*, **297**, 355–364.
20. Hancock,J.M. (1996) Simple sequences in a "minimal" genome. *Nature Genet.*, **14**, 14–15.
21. Field,D. and Wills,C. (1998) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae* and the different distributions of microsatellites in 8 prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc. Natl Acad. Sci. USA*, **95**, 1647–1652.
22. Field,D., Magnasco,M.O., Moxon,E.R., Metzgar,D., Tanaka,M.M., Wills,C. and Thaler,D.S. (1999) Contingency loci, mutator alleles and their interactions. *Ann. N. Y. Acad. Sci.*, **870**, 378–382.
23. 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.
24. Weisburg,W.G., Tully,J.G., Rose,D.L., Petzel,J.P., Oyaizu,H., Yang,D., Mandelco,L., Sechrest,J., Lawrence,T.G., Van Etten,J. *et al.* (1989) A phylogenetic analysis of the mycoplasmas: basis for their classification. *J. Bacteriol.*, **171**, 6455–6467.
25. Karlin,S. and Ost,F. (1985) Maximal segmental match length among random sequences from a finite alphabet. In Cam,L.M.L. and Olshen,R.A. (eds), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*. Wadsworth Inc., Belmont, CA, Vol. I, pp. 225–243.
26. Cohan,F.M. (1994) Genetic exchange and evolutionary divergence in prokaryotes. *Trends Ecol. Evol.*, **9**, 175–180.
27. Kurtz,S. and Schleiermacher,C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.
28. Erickson,B.W. and Sellers,P.H. (1983) Recognition of patterns in genetic sequences. In Sankoff,D. and Kruskal,J.B. (eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, pp. 55–91.
29. 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–GC10.
30. Weiner,J., Herrmann,R. and Browning,G.F. (2000) Transcription in *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **28**, 4488–4496.
31. 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.
32. Citti,C., Kim,M.F. and Wise,K.S. (1997) Elongated versions of Vlp surface lipoproteins protect *Mycoplasma hyorhinis* escape variants from growth-inhibiting host antibodies. *Infect. Immun.*, **65**, 1773–1785.
33. Wenzel,R. and Herrmann,R. (1988) Repetitive DNA sequences in *Mycoplasma pneumoniae*. *Nucleic Acids Res.*, **16**, 8337–8350.
34. Su,C.J., Dallo,S.F., Chavoya,A. and Baseman,J.B. (1993) Possible origin of sequence divergence in the P1 cytadhesin gene of *Mycoplasma pneumoniae*. *Infect. Immun.*, **61**, 816–822.
35. Razin,S. and Jacobs,E. (1992) Mycoplasma adhesion. *J. Gen. Microbiol.*, **138**, 407–422.
36. Opitz,O. and Jacobs,E. (1992) Adherence epitopes of *Mycoplasma genitalium* adhesin. *J. Gen. Microbiol.*, **138**, 1785–1790.
37. Yogev,D., Rosengarten,R., Watson-McKown,R. and Wise,K.S. (1991) Molecular basis of *Mycoplasma* surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5′ regulatory sequences. *EMBO J.*, **10**, 4069–4079.
38. Citti,C. and Wise,K.S. (1995) *Mycoplasma hyorhinis vlp* gene transcription: critical role in phase variation and expression of surface lipoproteins. *Mol. Microbiol.*, **18**, 649–660.
39. Taddei,F., Matic,I., Godelle,B. and Radman,M. (1997) To be a mutator, or how pathogenic and commensal bacteria can evolve rapidly. *Trends Microbiol.*, **5**, 427–429.
40. Denamur,E., Lecointre,G., Darlu,P., Tenaillon,O., Acquaviva,C., Sayada,C., Sunjevaric,I., Rothstein,R., Elion,J., Taddei,F. *et al.* (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*, **103**, 711–721.
41. Eckert,K.A. and Yan,G. (2000) Mutational analyses of dinucleotide and tetranucleotide microsatellites in *Escherichia coli*: influence of sequence on expansion mutagenesis. *Nucleic Acids Res.*, **28**, 2831–2838.
42. Hecker,M., Schumann,W. and Volker,U. (1996) Heat-shock and general stress response in *Bacillus subtilis*. *Mol. Microbiol.*, **19**, 417–428.
43. Teter,S.A., Houry,W.A., Ang,D., Tradler,T., Rockabrand,D., Fischer,G., Blum,P., Georgopoulos,C. and Hartl,F.U. (1999) Polypeptide flux through bacterial Hsp70: DnaK cooperates with trigger factor in chaperoning nascent chains. *Cell*, **97**, 755–765.
44. Dybvig,K., Sitaraman,R. and French,C.T. (1998) A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene arrangements. *Proc. Natl Acad. Sci. USA*, **95**, 13923–13928.
45. Xu,Q., Morgan,R.D., Roberts,R.J. and Blaser,M.J. (2000) Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proc. Natl Acad. Sci. USA*, **97**, 9671–9676.
46. Bhugra,B., Voelker,L.L., Zou,N., Yu,H. and Dybvig,K. (1995) Mechanism of antigenic variation in *Mycoplasma pulmonis*: interwoven, site-specific DNA inversions. *Mol. Microbiol.*, **18**, 703–714.
47. Arber,W. (2000) Genetic variation: molecular mechanisms and impact on microbial evolution. *FEMS Microbiol. Rev.*, **24**, 1–7.
48. Gumulak-Smith,J., Teachman,A., Tu,A.H., Simecka,J.W., Lindsey,J.R. and Dybvig,K. (2001) Variations in the surface proteins and restriction enzyme systems of *Mycoplasma pulmonis* in the respiratory tract of infected rats. *Mol. Microbiol.*, **40**, 1037–1044.
49. 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.
50. Paddenberg,R., Weber,A., Wulf,S. and Mannherz,H. (1998) Mycoplasma nucleases able to induce internucleosomal DNA degradation in cultured cells possess many characteristics of eukaryotic apoptotic nucleases. *Cell Death Differ.*, **5**, 517–528.

51. Renbaum,P., Abrahamove,D., Fainsod,A., Wilson,G.G., Rottem,S. and Razin,A. (1990) Cloning, characterization and expression in *Escherichia coli* of the gene coding for the CpG DNA methylase from *Spiroplasma* sp. strain MQ1 (M.SssI). *Nucleic Acids Res.*, **18**, 1145–1152.
52. Renbaum,P. and Razin,A. (1995) Footprint analysis of M.Sssl and M.Hhal methyltransferases reveals extensive interactions with the substrate DNA backbone. *J. Mol. Biol.*, **248**, 19–26.
53. Matsuo,K., Silke,J., Gramatikoff,K. and Schaffner,W. (1994) The CpG-specific methylase SssI has topoisomerase activity in the presence of Mg^{2+}. *Nucleic Acids Res.*, **22**, 5354–5359.
54. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
55. Rocha,E.P.C., Viari,A. and Danchin,A. (1998) Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic Acids Res.*, **26**, 2971–2980.
56. Rocha,E.P.C., Danchin,A. and Viari,A. (1999) Functional and evolutionary roles of long repeats in prokaryotes. *Res. Microbiol.*, **150**, 725–733.
57. 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.*, **25**, 701–712.