**Mycoplasma pneumoniae** and **Mycoplasma genitalium**: a comparison of two closely related bacterial species

Richard Herrmann* and Berta Reiner†

The rapid progress in sequencing large quantities of DNA will provide an increasing number of complete genome sequences of closely related bacterial species as well as of pairs of isolates from the same species with different features, such as a pathogenic and an apathogenic representative. This opens the way to apply subtractive comparative analysis as a tool to select from the large pool of all bacterial genes a relatively small set of genes that can be correlated with the expression of a certain phenotype. These selected genes can then be the target for further functional analyses.

**Abbreviations**

C cytosine

G guanine

IP isoelectric point

ORF open reading frame

**Introduction**

The genus mycoplasma is one of several genera belonging to the bacterial class Mollicutes. The common features of all mycoplasma species are the complete lack of a bacterial cell wall, the absence of a periplasmic space, a limited metabolic potential and a parasitic lifestyle. The cells are surrounded only by a cytoplasmic membrane which contains, uniquely among bacteria, cholesterol [1,2••]. The term mycoplasma is used in this paper to name only the members of the genus mycoplasma and not, as frequently done, all the members of the class Mollicutes. Mycoplasmas have lost substantial parts of their genomes during evolution and are presently the bacteria with the smallest known genomes, with sizes between 580 and 1250 kbp [3]. The cells have a diameter of 0.5 µm or even less and are too small to be clearly seen in a light microscope. Species of the genus mycoplasma can be found in many different hosts including humans, cattle, rodents, birds, reptiles and fish, but individual species have a strict host, organ and tissue specificity [2••].

Mycoplasmas have received new attention through the recent progress in sequencing their entire bacterial genomes. After *Haemophilus influenzae* [4], *Mycoplasma genitalium* was the second bacterium for which the complete genome sequence was published [5]. Later, when the genome sequence of *Mycoplasma pneumoniae* was also available [6], it became possible to compare two closely related bacterial species with different features at the genome level (genomics) [7••]. Both species show a high level of similarity, but also sufficient differences to make this comparison worthwhile. This review is intended to show that a thorough comparative genetic analysis will allow correlation of genetic properties with current knowledge on the physiology, molecular biology and microbiology of each of the two mycoplasma species but will also provide useful information for planning and executing experiments designed to reveal the function of newly proposed mycoplasmal genes that are functionally unassigned. *M. genitalium* and *M. pneumoniae* are already included in a number of comparative studies that also deal with the complete sequences of other bacterial or archaeal genomes (for examples, see [8•,9•,10••,11••]).

**Habitat and pathogenicity**

Humans are the natural host for *M. genitalium* and *M. pneumoniae*. For a long time, these bacteria were considered as surface parasites, but recent reports have indicated that they may also enter the host cell and persist intracellularly [12,13]. This observation could explain the tendency of *M. pneumoniae* and *M. genitalium* to survive in their hosts after antibiotic treatment and clinical recovery [14]. Both mycoplasmas possess a specialized terminal organelle (tip structure, Figure 1), the correct formation of this tip structure is the prerequisite for a successful infection [15,16]. As the name implies, *M. pneumoniae* is preferentially found in the respiratory tract and *M. genitalium* in the urogenital tract, although not exclusively, *M. genitalium* has also been isolated from the respiratory tract and *M. pneumoniae* from the urogenital tract [14]. It is beyond the scope of this review to evaluate the consequences of *M.
genitalium and M. pneumoniae infections for individual cells, tissue or the complete organism, but it should be remembered that the ability of M. pneumoniae to cause respiratory tract disease has been confirmed by many studies, whereas while M. genitalium seems to be associated with nongonococcal urethritis it has not been strictly proven to cause this infection. For more details see the recent clinical paper by Taylor-Robinson [14], or reviews on mycoplasma virulence factors and the immune responses [17,18]. The key to understanding the species-specific tissue prevalence and pathogenic mechanisms should be found in the genetic differences between M. genitalium and M. pneumoniae that have been revealed by the complete sequences of their genomes.

**Sequence analysis and coding capacity**

The genome of M. genitalium has a size of 580,070 bp [5] and is 236,324 bp smaller than the genome of M. pneumoniae (816,394 bp) [6]. The first publications reported 470 open reading frames (ORFs) for M. genitalium and 677 ORFs for M. pneumoniae. Since then a number of corrections have been made, including new functional assignments and the proposal of new ORFs. None of these modifications has any significant influence on this comparative study, because it has been shown that all the ORFs that were proposed for M. genitalium are contained in the larger genome of M. pneumoniae [7••]. Therefore, in this review, we do not consider all the proposed corrections and new functional classification, and we refer always to the publications by Fraser et al. [5] and Himmelreich et al. [6] unless otherwise stated.

To compare the number of ORFs in both bacteria directly one has to make an adjustment in annotation. The list of 677 ORFs form M. pneumoniae includes 26 ORFs which are coded for by the repetitive sequences RepMP2/3 (10 copies), RepMP4 (eight copies) and RepMP5 (eight copies) [22]. These repetitive sequences, called the MgPar repeat [23,24], are also contained in M. genitalium at a lower copy number (14 copies all together) but they were not incorporated in the ORF index. Therefore, the actual numbers of ORFs to be compared are 470 and 663. The repetitive sequences are part of a genetic exchange mechanism that generates, probably by gene conversion, variable copies of two proteins which are involved in cytadherence [25,26].

These two proteins are coded for by two genes (gene P1, ORF6) of the P1 operon in M. pneumoniae [27] and its homologs (MG191, 192) in M. genitalium which are organized in the MgPa operon [28]. Only the copies that reside in the P1 operon and the MgPa operon are expressed.

The average identity between M. pneumoniae and M. genitalium orthologs is 66.1% at the nucleotide level and 67.4% at the amino acid level, and the 16S and 23S ribosomal RNAs are 98% identical. Furthermore, a comparison of theoretical two-dimensional protein maps, based on the calculated isoelectric point (IP) and molecular weight for each proposed protein, showed a very similar distribution. The relatively high number of proteins with IPs between 10 and 13 in M. genitalium and M. pneumoniae compared to Bacillus subtilis or Haemophilus influenzae is notable (Figure 2) and should be considered when analyzing protein extracts from these bacteria by methods based on charge differences, such as in two-dimensional gel electrophoresis.

This high degree of similarity raises the question of whether these two bacteria should belong to different species. According to the still accepted rules for species definition in Bergey's Manual, a difference in G + C content higher than 2 mol% and DNA homology values lower than 40% justify separation into two species [29], but these rules were established before the prospect of the use of complete genomic sequences for taxonomic purposes. Therefore, these two bacteria could serve as a good paradigm to check, and maybe also redefine, the rules for species definition. The DNA guanine (G) and cytosine (C) content differs by 8 mol% between the genomes of M. pneumoniae and M. genitalium, and this is somewhat surprising given the high agreement in gene order and coding capacity. Detailed studies revealed that the difference is mainly due to variation at the third codon position, which is with 19 mol% difference in G + C the most variable. The third codon position in about 32,000 codons (of a total of 170,400) in M. genitalium could be altered from A/T to G/C to increase the genomic G + C content by 5.6 mol% G + C without changing a single amino acid. A G + C plot of the complete genomes of M. genitalium and M. pneumoniae showed a similar slightly uneven distribution, which was mainly attributed to the repetitive DNA sequences which in both organisms have a G + C content above average [7••]. Kerr et al. [30] and McInerney [31•] discovered independently that a more discriminatory difference is seen when only the G + C content of the third position of the codons (GC3) is plotted with respect to the position of the proposed gene on the genome. The GC3 distribution is random in M. pneumoniae but in M. genitalium it has a maximum between nucleotide positions 190,000 and 205,000 and minimum around nucleotide position 450,000. Thus it seems that codon usage and genome position are correlated in M. genitalium but not in M. pneumoniae. The significance of this finding is unclear.

**Functional assignment**

Mycoplasmas developed by degenerate evolution from Gram-positive bacteria of the Lactobacillus group [32]. During this ongoing process they have lost many functions, probably only after they had adapted to a parasitic life style in an environment that is physiologically fairly constant. It has been shown by biochemical experiments [33••] and was confirmed and extended by the genome sequence analyses that neither M. genitalium nor M. pneumoniae code for proteins involved in the biosynthesis of amino acids, fatty acids, cholesterol, or purine and pyrimidine bases. This makes them very dependent on import systems. Unexpectedly, the number of transport systems in the M. pneumoniae and M. genitalium is relative low compared to other bacteria, therefore one has to conclude that the
mycoplasma transport systems must have low substrate specificities [6,34••]. Several metabolic pathways were present in mycoplasmas in a truncated form or completely absent. The glycolysis pathway is completely present, but only two components of the pentose-phosphate pathway and one of the TCA cycle could be identified [35••]. None of the proteins involved in cell wall synthesis have been conserved. Instead both species have an identical set of proteins which are believed to be part of a cytoskeleton-like structure (Table 1; [15]). These proteins do not show significant homologies to proteins from other organisms by standard similarity searches in databases. This is consistent with the observation that the orthologs of these proteins in \textit{M. genitalium} and \textit{M. pneumoniae} share relatively low levels of similarity. With a single exception they are less than 50% identical at the amino acid level, compared to at least 70% to 97% identity for conserved housekeeping proteins (Table 1). The degree of identity can be used to predict whether a protein with unknown function is a housekeeping protein. For instance, the original annotations for MG353 and its homolog in \textit{M. pneumoniae}, G12_orf109, did not predict a function despite exhibiting almost 92% identity. Fold assignments for these genes suggested that they code for the histone-like protein HU [36••].

One of the obvious differences from other bacteria is the low number of proteins detected for regulation of gene expression in \textit{M. genitalium} and \textit{M. pneumoniae}. So far, only one sigma factor has been identified, and no members of the almost ubiquitous two component signal transduction system. How these mycoplasmas regulate the synthesis of many gene products, which are present in different concentrations
The direction of transcription of genes is very uniform in \textit{M. genitalium} and \textit{M. pneumonia}. Starting from the proposed origin of replication [5,6,7**] 85% of all proposed ORFs are transcribed bidirectionally in the same way as DNA replication proceeds (Figure 3). In both species only 15% of genes are transcribed against this general direction (Figure 3). Preliminary transcription studies on \textit{M. pneumoniae} did not show a detectable correlation between the location of a gene with respect to the proposed origin of replication and the concentration of transcripts in logarithmically growing cells (HWH Göhlmann, personal communication).

The analysis of the genome sequences of both \textit{M. genitalium} and \textit{M. pneumoniae} failed to identify genes encoding for activities that had to be present in these bacteria; the most prominent examples being RNaseH, which removes RNA primers during replication of chromosomal DNA, nucleoside diphosphate kinase, which is essential for the synthesis of nucleoside triphosphates, or glycosyl transferases. The standard procedure to functionally assign a proposed ORF is to search for orthologs in other organism. If a genome does not encode the corresponding ortholog one has to screen whether non-orthologous genes could be candidates for replacement of the missing orthologs. Based on this approach Koonin \textit{et al.} [20] proposed a number of ORFs in \textit{M. genitalium} which could code for the above mentioned missing RNaseH and kinase activities. Given the close relationship between \textit{M. genitalium} and \textit{M. pneumoniae} the corresponding \textit{M. pneumoniae} ORFs should also be considered. These proposals have yet to be verified experimentally.

\textbf{\textit{M. pneumoniae} specific functions}

Because all the ORFs proposed for \textit{M. genitalium} are present in \textit{M. pneumoniae}, it is fair to assume that they code for products with identical functions. It seems very likely that the differences in physiology, tissue specificity and pathogenicity have to be encoded by the \textit{M. pneumoniae} specific ORFs which were absent in \textit{M. genitalium}. Therefore, a careful functional analysis of the 209 ORFs that are encoded by the 236 kbp additional DNA in \textit{M. pneumoniae} should explain the features specific to \textit{M. pneumoniae} (Tables 2 and 3). Surprisingly, only 110 ORFs of the 209 additional ORFs are new and \textit{M. pneumoniae} specific, whereas the residual 99 ORFs are either gene duplications, which exist mainly as single copies in \textit{M. genitalium}, or represent ORFs derived from the repetitive DNA sequences RepMP2/3, RepMP4 and RepMP5. Among the single copy genes in \textit{M. genitalium} that are amplified in \textit{M. pneumoniae} are the lipoproteins of the murein lipoprotein type found in \textit{Escherichia coli}. \textit{M. genitalium} codes only for 21 of these lipoproteins but \textit{M. pneumoniae} for 46. The 25 additional lipoproteins in \textit{M. pneumoniae} are only amplifications, which are frequently organized in sequential order. It might give \textit{M. pneumoniae} a greater reservoir for antigenic variation in these frequently surface exposed proteins, but the phenomenon of antigenic variation through differential expression of lipoproteins, which is documented for several mycoplasma species [37], has not yet been observed in \textit{M. pneumoniae}. We do not assume that the additional lipoproteins and the other gene duplications contribute to the species specific features of \textit{M. pneumoniae}.

From the 110 \textit{M. pneumoniae} specific ORFs only 26 can be assigned functionally by sequence similarity to other proteins (Table 3). This is a rather low value considering the relatively high number of available complete bacterial genome sequences including that of \textit{Bacillus subtilis}, which is supposed to be the closest phylogenetically and which should have homologs of all genes of \textit{M. pneumoniae} except for those which were acquired by \textit{M. pneumoniae} from other sources (e.g. by horizontal gene transfer). A possible explanation for this low value could be that many \textit{M. pneumoniae} specific genes have diverged far. The 16 ORFs that show the highest similarity to eukaryotic proteins are shared by both \textit{M. pneumoniae} and \textit{M. genitalium}. There is one exception which is only found in \textit{M. pneumoniae}.  

\begin{table}[h]
\centering
\caption{Sequence similarities of selected proteins from \textit{M. pneumoniae} and \textit{M. genitalium}.}
\begin{tabular}{lcc}
\hline
Proposed function/protein name & \textit{M. pneumoniae} & \textit{M. genitalium} \\
\hline
Cytadherence accessory proteins & & \\
HMW1 & H08\_orf1018 & MG312 \\
HMW2 & F10\_orf1818 & MG218 \\
HMW3 & H08\_orf672 & MG317 \\
P65 & F10\_orf405 & MG217 \\
P200 & D02\_orf10360 & MG368 \\
P1 & E07\_orf1627 & MG191 \\
ORF6 (P1 operon) & E07\_orf1218 & MG192 \\
P90 & H08\_orf274 & MG318 \\
Proteins with the highest similarities & & \\
Elongation factor TK & K05\_orf394 & MG451 \\
Ribosomal protein S12 & G07\_orf139 & MG087 \\
Ribosomal protein L14 & GT9\_orf122 & MG161 \\
Ribosomal protein L33 & P01\_orf53 & MG325 \\
Ribosomal protein S87 & VXpSPT7\_orf87 & MG155 \\
Ribosomal protein L16 & VXpSPT7\_orf1390 & MG158 \\
Heat shock protein DnaK & A05\_orf595 & MG305 \\
RNA polymerase β′RpoC & F04\_orf1290 & MG340 \\
\hline
\end{tabular}
\end{table}
(CO9_orf600, carnitine palmitoyltransferase II precursor [crt2]). From the 26 functionally assigned *M. pneumoniae* specific ORFs, 12 ORFs code for components of a *hsd*-type restriction-modification system. Two are components of the phosphoenolpyruvate/carbohydrate phosphotransferase system (PTS), five are enzymes comprising the arginine dihydrolase pathway (arginine deiminase, ornithine carbamoyltransferase, carbamate kinase), with the potential to synthesize ATP from arginine, and one is an NADP-dependent alcohol dehydrogenase.

In correlation with the absence of the restriction modification system and the arginine dihydrolase pathway in *M. genitalium*, the restriction endonuclease and two proposed arginine deiminase genes are truncated and probably not active in *M. pneumoniae* due to frame shift mutations. It is unclear whether the full length ORF (H03_orf438), coding for an arginine deiminase synthesizes, is an active enzyme *in vivo*, as such an activity has not been shown for *M. pneumoniae* [33**], and hence *M. pneumoniae* has been classified as an arginine nonfermenting species.

Some of these *M. pneumoniae* specific functions (e.g. the transport system) may explain why *M. pneumoniae* is easier to grow in the laboratory than *M. genitalium*, but none of them provides convincing indications for the observed differences in pathogenicity and tissue specificity of the two mycoplasmas. We assume that the key to understanding these phenomena might be found in the 37 *M. pneumoniae* specific ORFs that do not show any significant homologies to other genes/proteins in databases. To proof this and to do functional analyses, one has now to apply the methods of biochemistry, biophysics and genetics.

**Genome organization**

The complete bacterial genome sequences so far published show convincingly that there is no long range colinearity between genomes from distantly related bacteria [38] except in the closely related pair *M. pneumoniae* and *M. genitalium* [7**]. The gene order in both mycoplasmas is

---

### Table 2

Differences between *M. pneumoniae* and *M. genitalium*.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Excess in <em>M. pneumoniae</em> compared to *M. genitalium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size</td>
<td>236 kbp</td>
</tr>
<tr>
<td>Number of ORFs</td>
<td>209</td>
</tr>
<tr>
<td>Number of unique <em>M. pneumoniae</em> specific ORFs</td>
<td>110</td>
</tr>
<tr>
<td>Gene duplications</td>
<td>99</td>
</tr>
</tbody>
</table>

### Table 3

Classification of *M. pneumoniae* specific ORFs.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number of ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function assigned</td>
<td>25</td>
</tr>
<tr>
<td>Paralogs or conserved motifs</td>
<td>27</td>
</tr>
<tr>
<td>Repetitive DNA derived</td>
<td>21</td>
</tr>
<tr>
<td>Without homology to other genes/proteins</td>
<td>37</td>
</tr>
</tbody>
</table>

---

Comparative presentation of the organization of the DNA segments with conserved gene order and general direction of transcription in *M. genitalium* (MG) and *M. pneumoniae* (MP). Both genomes have been oriented the same way as *M. genitalium* was originally presented [5], starting with the dnaN gene as MG001. The black arrows indicate the direction of transcription, starting with dnaA and dnaN in opposite directions from the proposed origin of replication (black dot). The MG ORF numbers at the beginning and end of each bar give the first and last *M. genitalium* ORF of each of the six segments with conserved gene order. For simplification, the ORFs that were transcribed against the general direction are not indicated, but the region of the genome in which the direction of transcription changes frequently is shown by the interrupted lines following the arrowhead. The bottom lines indicate the nucleotide positions in the published genome sequences. R indicates rRNA operon; ▼ indicates repetitive DNA sequences.
very well conserved; not over the entire genome, but within six segments which are in different order on each genome. Within each segment the order of homologous genes with respect to each other is conserved, with the additional \textit{M. pneumoniae} specific ORFs interspaced between them (Figure 3). The direction of transcription does not change in the genomic segments despite their different location on the genomes of both bacteria. This could indicate that an advantage exists in maintaining the strongly biased direction of transcription. Such a strong bias in the direction of transcription has not been found in other bacteria. The translocation of the genomic segments can be explained by homologous recombination between individual copies of the repetitive DNA sequences which flank these segments. Not all the copies of the repetitive DNA sequences seem to be involved in translocation of segments (see Figure 3). It is obvious that in the larger part of both genomes the order of genes is colinear although four repetitive sequences are located in this region. We assume that a selection pressure exists against interruption of this part of the genome by translocation of segments from outside, as they would interfere with the predominantly bidirectional transcription.

Siefert \textit{et al.} [39] identified 16 gene clusters, categorized in five functional groups, which are conserved in many bacteria including both mycoplasmas. Among these clusters are several ribosomal protein operons, including the FσF₁ ATPase operon, the ribosomal RNA operon, an ABC transporter with a proposed specificity for spermidine and putrescine. In addition, these authors identified a set of clusters which were found in either the \textit{Haemophilus/Mycoplasma} comparison or the \textit{Haemophilus/Synechocystis} comparison, but not in both. Of the 14 proposed clusters only one, the dnaA/dnaN region, is relevant in context of this review. The dnaA/dnaN region is part of the proposed origin of replication (Figure 4). The genes dnaA and dnaN are located next to each other in many bacterial genomes and transcribed in the same direction [40].

This is not true for \textit{M. genitalium} [5,41] and \textit{M. pneumoniae} [42]. We find here, so far, a unique divergent organisation of the dnaA region but with an identical gene order in both mycoplasma genomes (Figure 4). The identical organisation is not obvious from the original publications, because the annotation for several ORFs differed and some ORFs were proposed for \textit{M. pneumoniae} but not for \textit{M. genitalium}. The comparative analysis at the nucleotide level shows convincingly that the dnaA/dnaN clusters contain the same ORFs organized in the same order, except that the homolog of ORFMP155 is truncated in \textit{M. genitalium}.

This organization, however, is not conserved among mycoplasmas, as the cluster and their order of genes in \textit{Mycoplasma capricolum} differ from those in \textit{M. pneumoniae} and \textit{M. genitalium} and appears more similar to the cluster in \textit{B. subtilis} [40,43**].

**Mycoplasmas as experimental systems**

The main advantage of mycoplasmas as research models is their small genome size and accordingly the rather limited number of gene products. \textit{M. genitalium} in particular, but also \textit{M. pneumoniae}, could serve as model organisms [44] to define experimentally by further genome reduction the essential functions for a self replicating cell under a particular set of environmental conditions [45]. In addition, the two species are excellent model organisms to test and develop strategies for global expression analyses, because transcription profiles or protein maps based on two-dimensional gel electrophoresis are much easier to establish for 500–700 transcripts or proteins than for organisms which code for several thousand gene products.

A clear disadvantage for many experimental approaches is the lack of an established gene transfer system and of a defined culture medium. Transposon mutagenesis and integration of foreign DNA into the genome via transposons can be achieved albeit with limited efficiency.
For growth both *M. genitalium* and *M. pneumoniae* need a rich medium which contains 5–20% serum making pulse-labelling experiments and growth under defined conditions impossible. In addition, the medium is expensive and the yields of mycoplasma cells are low compared to *E. coli* yields. Last but not least, the codon usage in mycoplasmas differs from that in *E. coli*. The codon UGA is read as tryptophan in mycoplasmas [49] but as a stop codon in most other bacteria. Expression of mycoplasma proteins in *E. coli* therefore requires a UGA-suppressor gene [50] and is further improved by using an *E. coli* mutant with an inactive release factor 2 [51]. Even then, the suppression of several UGAs in one gene is limited and the synthesis of full length proteins is severely hampered. Considering the advantages and disadvantages it seems obvious that mycoplasmas are not well suited to study those genes or functions which can be analyzed in bacteria like *E. coli* or *B. subtilis* with highly developed genetic tools and which also give high yields of cells.

**Conclusions**

The comparative analysis of *M. genitalium* and *M. pneumoniae* on the basis of the information provided by their complete genome sequences showed that the two bacterial species were much more similar than anticipated from a difference of 30% in genome size, 8 mol% in DNA G + C content, and their different tissue specificity and pathogenicity. Not only were all the proposed ORFs of the smaller *M. genitalium* genome found in *M. pneumoniae*, but genome organization and the rather uniform direction of transcription were highly conserved. The additional 236 kbp DNA of the *M. pneumoniae* genome coded for only 105 new ORFs, of which 37 were particularly interesting as they showed no significant similarities to proteins from other organisms and were considered to be *M. pneumoniae* specific. Combining the results from comparative analyses with studies on global expression analyses at the transcriptional and translational level under varying growth conditions should reveal which of the proposed ORFs of *M. pneumoniae* and *M. genitalium* are involved in species-specific cellular responses.

Finally, the broad agreement concerning the defined ORFs and their annotation in the independent sequencing projects on *M. genitalium* and *M. pneumoniae* also confirms the results for each of both analyses.

**Acknowledgements**

We thank P-C Hu for the micrograph, B Browning and S Razin for critically reading the manuscript, D Hofmann for preparing some of the figures and I Schmid for secretarial work.

**References and recommended reading**

Papers of particular interest, published within the annual period of review, have been highlighted as: **of special interest** and **of outstanding interest**.


Recommended for specialists and general readers. A very clearly written comprehensive summary.


