*Genetics and Molecular Biology*, 30, 1, 169-173 (2007)

www.sbg.org.br

Research Article

# In search of essentiality: Mollicute-specific genes shared by twelve genomes

Rangel Celso Souza[1], Darcy Fontoura de Almeida[2], Arnaldo Zaha[3], David Anderson de Lima Morais[1] and Ana Tereza Ribeiro de Vasconcelos[1]

[1]*Laboratório Nacional de Computação Científica, Petropólis, RJ, Brazil.*
[2]*Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brazil.*
[3]*Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.*

## Abstract

Mollicutes are cell wall-less bacteria with a genome characterized by its small size. Chromosomal rearrangements help these organisms evade host immune surveillance and hence cause disease. Our goal was to determine genes shared by Mollicutes genomes using the bidirectional best hit methodology. The twelve studied Mollicutes share 210 genes, most of which (> 60%) fall into the following COG categories: translation, ribosomal structure and biogenesis; DNA replication, recombination and repair; nucleotide transport and metabolism and energy production and conversion. Thirty Mollicute-specific genes were identified, 22 of them previously described as essential genes in *Mycoplasma genitalium*.

*Keywords:* mollicutes, comparative genomics, synteny, bidirectional best hit.

Received: April 4, 2006; Accepted: October 5, 2006.

## Introduction

Mollicutes show a large potential for recombination due to their large number of repeats (Rocha and Blanchard, 2002). The existence of arrangements and gene clustering indicate that they may confer some evolutionary advantages to individuals or populations. Through these mechanisms, they produce the machinery necessary for cell growth, host-defense invasion, and often survival in the host for indefinite periods (Lo, 1992).

Since the first mycoplasma genome was sequenced (Fraser *et al*., 1995) efforts have been made to find the minimal number of genes required for a self-replicating cell. According to Mushegian and Koonin (1996), a minimal gene set required for a species could be deduced from conserved genes in the analyzed genomes. This minimal gene set has been defined by Koonin (2000 and 2003) as the "smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favorable conditions imaginable, that is, in the presence of a full complement of essential nutrients and in the absence of environmental stress" (quoted by Gil *et al.*, 2004). Another approach is to define the essential functions in a living cell and to list the genes necessary to maintain such functions (Gil *et al*., 2004).

In addition, comparative genomics may be used to detect the set of genes common to all genomes in a phylogenetically coherent group (Charlebois and Doolittle, 2004). Considering that increasing the number of genomes used in a comparison can reduce the number of genes regarded as essential (Gil *et al*., 2004), and assuming that the genes shared by multiple genomes are likely to be essential, we performed a comparative analyses on the complete genomes of twelve Mollicutes. This study was done using the Bidirectional Best Hit (BBH) approach to determine the number of genes shared by all studied genomes and to verify the synteny among these genes. The results were compared with those obtained by Mushegian and Koonin (1996) and by Gil *et al.* (2004). We were able to identify the genes that are Mollicute-specific and to find their respective representation within the COG categories used as reference. Most of the specific gene families or COG functional categories are covered by the accompanying articles in the present issue of *Genetics and Molecular Biology*. In some instances, the focus has been concentrated on the three *Mycoplasma* species whose complete genome sequences have been determined by the Brazilian National Genome Program (Southern Network for Genome Analysis

Send correspondence to Ana Tereza Ribeiro de Vasconcelos. Rua Getúlio Vargas 333, Quitandinha, 25651-075 Petropólis, RJ, Brazil. E-mail: atrv@lncc.br.

and Brazilian National Genome Project Consortium) as reported by Vasconcelos *et al.* (2005).

## Material and Methods

### Sequence and database

The twelve genomes compared were: *Mycoplasma genitalium,* Mge, (Fraser *et al*., 1995); *Mycoplasma pneumoniae,* Mpn, (Himmelreich *et al*., 1996); *Ureaplasma urealyticum,* Uur, (Glass *et al*., 2000); *Mycoplasma pulmonis,* Mpu, (Chambaud *et al*., 2001); *Mycoplasma penetrans,* Mpe, (Sasaki *et al*., 2002); *Mycoplasma gallisepticum,* Mga, (Papazisi *et al*., 2003); *Mycoplasma mycoides* subsp. *mycoides* SC, Mmy, (Westberg *et al*., 2004); *Mycoplasma mobile,* Mmo, (Jaffe *et al*., 2004); *Mycoplasma hyopneumoniae* 232, Mhy-232, (Minion *et al*., 2004); *Mycoplasma synoviae,* Msy, *Mycoplasma hyopneumoniae,* Mhy-J and *Mycoplasma hyopneumoniae* 7448, Mhy-P (Vasconcelos *et al*., 2005). The data on the complete genomes were downloaded from Entrez genome (http://www.ncbi.nlm.nih.gov) except for *M. synoviae* (AE017245), *M. hyopneumoniae* strain J (AE017243) and *M. hyopneumoniae* strain 7448 (AE017244) that were analyzed by the Brazilian National Genome Sequencing Consortium and the Southern Genome Investigation Program - PIGS.

### Comparative analyses

In order to determine the distribution of the genes of the twelve genomes across COG classes, an analysis of each CDS was performed (Tatusov *et al*., 2001). The comparison of CDSs among the studied genomes was done through SABIA (System for Automated Bacterial (genome) Integrated Annotation) software (Almeida *et al*., 2004).

In order to determine genes shared by all genomes the BBH approach was used (Overbeek *et al*., 1999). Given two genes $X_a$ and $X_b$ from two genomes $G_a$ and $G_b$, $X_a$ and $X_b$ are called BBH if and only if recognizable similarity exists between them and there is no gene $Z_b$ in $G_b$ genome that is more similar than $X_b$ is to $X_a$, and if there is no gene $Z_a$ in $G_a$ that is more similar than $X_a$ is to $X_b$.

### Results and Discussion

The BBH methodology revealed 210 shared genes, classified in 17 COG categories. We did not find genes that would have fallen into the T category (signal transduction mechanisms),and four genes could not be classified (Table 1). Most of the 210 genes shared by Mollicutes have a known function or at least a predicted function. Among all BBHs only 7 clusters are formed by conserved hypothetical or only hypothetical proteins in all genomes (Table 1).

For the comparison between the results presented here and those from previous studies, the *M. genitalium* genome was used as reference. When compared with data from Gil *et al.* (2004), 86 BBH-specific genes were found; and the comparison with Mushegian and Koonin's (1996) data set revealed 42 BBH-specific genes. When the comparison was made using the pooled data from both sources, the number of BBH-specific genes was 30. The most represented COG categories were R (General function prediction only) with seven genes; L (DNA replication, recombination and repair) with five genes; F (Nucleotide transport and metabolism) with three genes; K (Transcription) with two and P (Inorganic ion transport and metabolism) also with two genes. These results indicate that these specific genes groups may be a source of useful information regarding properties that could be characteristic of Mollicutes (supplementary material and Table 1). Twenty-two of the Mollicute-specific genes have been previously identified as essential ones in *M. genitalium* (Glass *et al*., 2006).

Shared genes amounted to 168, 124 and 120 when compared with Mushegian and Koonin (1996), with Gil *et al.* (2004) and with both pooled together, respectively. It should be pointed out that Mushegian and Koonin's data were obtained from two genomes only, and that the results from Gil and collaborators included seven genomes, five of them from endosymbionts.

The BBH approach was able to group 118 genes (56% of the total) belonging to the information storage and processing division (categories J, K and L), metabolism

**Table 1** - Genes shared by the twelve mycoplasmas genomes studied

| COG category | Number of genes |
|---|---|
| - - Not in COG | 4 |
| C - Energy production and conversion | 11 |
| D - Cell division and chromosome partitioning | 2 |
| E - Amino acid transport and metabolism | 3 |
| F - Nucleotide transport and metabolism | 15 |
| G - Carbohydrate transport and metabolism | 9 |
| H - Coenzyme metabolism | 2 |
| I - Lipid metabolism | 3 |
| J - Translation, ribosomal structure and biogenesis | 82 |
| K - Transcription | 10 |
| L - DNA replication, recombination and repair | 26 |
| M - Cell envelope biogenesis, outer membrane | 3 |
| N - Cell motility and secretion | 6 |
| O - Posttranslational modification, protein turnover, chaperones | 10 |
| P - Inorganic ion transport and metabolism | 3 |
| R - General function prediction only | 18 |
| S - Function unknown | 3 |

Total number of genes found: 210. The *M. genitalium* genome was used as reference.

(21%) and cellular processes (11%). The poorly characterized COG categories (R and S) contain 21 genes (10%).

The COG category J (translation, ribosomal structure and biogenesis), as expected (Santos *et al*. and Borges *et al*., in the present issue), contains the highest number of genes (82) (Table 1). All COG categories involved in metabolism (E, F, G, H, I and P) were clustered (Arraes *et al*., Balaião *et al.*, and Staats *et al*., in the present issue). The F0F1-Type ATP synthase system is ubiquitous in Mollicutes. Among the nine subunits (from *atpA* to *atpI*) only *atpC* and *atpI* did not form BBH clusters. Two proteins belonging to the energy production and conversion COG category were found to contribute to pyruvate decarboxylation (Nicolás *et al.*, in the present issue). As far as carbohydrate transport and metabolism are regarded, glycolysis is the most conserved pathway and six genes that take part in this pathway are shared by all Mollicutes. It seems to be the main source of ATP in Mollicutes. They also have three genes of the pentose phosphate pathway; however this pathway is not complete in some species.

The COG category F (nucleotide transport and metabolism) is highly conserved. Fifteen BBH clusters were formed, and on average, the studied Mollicutes have about 24 genes in this pathway. Purine biosynthesis and purine salvage, pyrimidine biosynthesis and pyrimidine salvage, and thymidylate biosynthesis, have each at least one protein shared by all Mollicutes (Bizarro *et al*. in the present issue). However, *Haemophilus influenzae* shows more enzymes in these pathways than Mollicutes (Razin *et al*., 1998). The COG L category (DNA replication, recombination and repair) is also well conserved with 26 clusters (Fonseca *et al*., and Brocchi *et al.*, in the present issue). Among the genes responsible for lipid metabolism (COG category I) only three genes that participate in phospholipids biosynthesis formed clusters, the cell division and chromosome partitioning class (COG category D) contains two genes (Alarcon *et al.*, in the present issue).

### Analysis of synteny

Using the MEGA2 program, Kumar *et al.* (2001) obtained a tree through the concatenation of eight ribosomal proteins. The analyzed mollicutes are divided into three groups (Figure 1): the Hominis (Mpu, Mmo, Msy, Mhy-P, Mhy-J and Mhy-232); the Pneumoniae (Mga, Mge, Mpn, Mpe and Uur) and the Spiroplasma group (Mmy). These data are in agreement with those obtained by Weisburg *et al.* (1989) and Yotoko and Bonatto (in the present issue).

Via the tree analysis, the closest pairs of genomes were determined and maps of synteny between them were constructed (Figure 2). Although mollicutes show a high capacity of rearrangement, some patterns emerged. The most evident aspect in the mollicutes genomes is the presence of large clusters (genes that kept the same order) formed mainly by ribosomal proteins (red lines, Figure 2). According to Vasconcelos *et al.* (2005) the total size of this inversion is 243,104 kb.

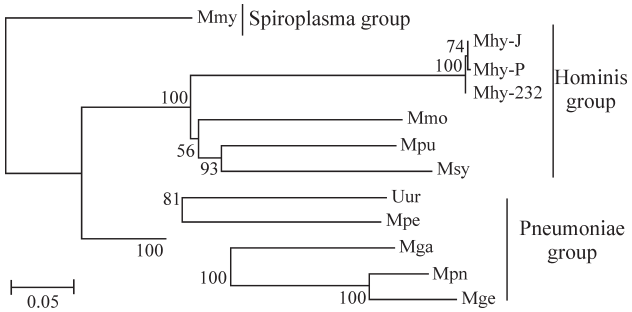


**Figure 1** - Neighbour joining phylogenetic tree based on eight ribosomal genes. The bootstrap values are shown. Mmy *M. mycoides;* Mpu *M. pulmonis;* Msy *M. synoviae;* Mmo *M. mobile;* Mhy-J *M. hyopneumoniae;* Mhy-P *M. hyopneumoniae* 7448; Mhy-232 *M. hyopneumoniae* 232; Mga *M. gallisepticum;* Mge *M. genitalium;* Mpn *M. pneumoniae;* Mpe *M. penetrans;* Uur *U. urealyticum.*

In order to quantify the number of rearrangements between pairs of genomes, the groups of shared genes that kept the same position relative to each other in both genomes were scored (Figure 3). They were classified in three categories: group I (sharing from one to three genes); group II (sharing 4 to 6 genes) and group III (sharing more than six genes). All species analyzed share at least two group III genes, one formed by the ATP synthase family and the other formed by ribosomal proteins. Group I was the prevalent group among mycoplasmas, with numbers varying from 33 to 46. Such findings confirm data from Yokoto and Bonatto (in the present issue) indicating that, even though there is a number of conserved genes among the studied genomes, they are not in general kept in the same relative positions.

## Conclusions

The comparison of twelve Mollicute genomes provides evidence that mycoplasmas have prioritized the conservation of some genes. Genes belonging to the information storage and processing COG division seem to be the most conserved, while genes belonging to the metabolism COG division are less conserved. The BBH methodology identified 210 genes that are shared among the twelve studied genomes. The comparison with the pooled data from other studies showed that the number of BBH-specific genes was 30, out of which 10 (33%) corresponded to conserved hypothetical or putative genes. It will be important to study these Mollicute-specific genes to obtain useful information regarding properties that should be characteristic of Mollicute*s*. Twenty-two of these genes were previously identified as essential genes in

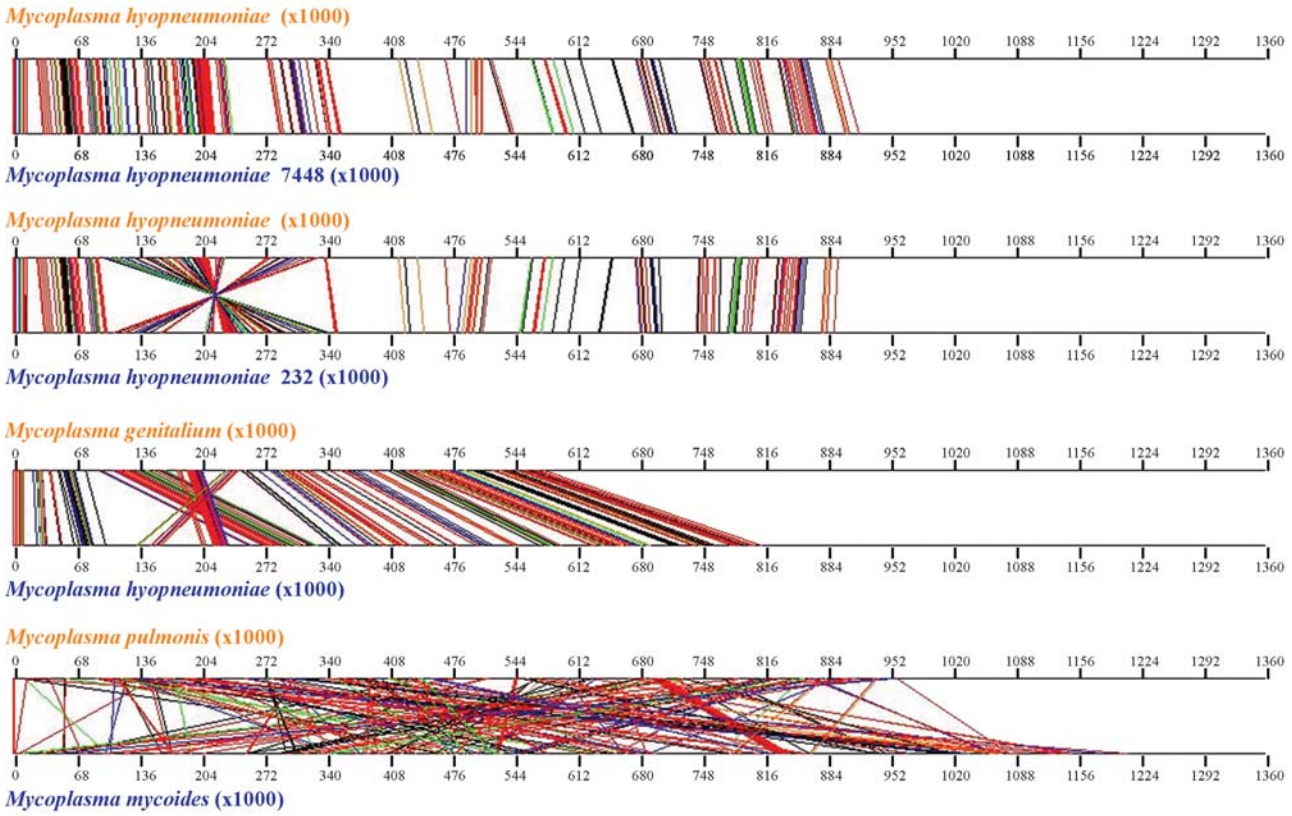


**Figure 2** - Synteny between pairs of genomes. Only the closest pairs of genomes are shown. The genes shown in these maps are BBHs and are single copy genes per genome. Genes not classified by COG were also removed. Each color represents one COG division: Red - Information storage and processing; Blue - Cellular processes; Green - Poorly characterized and Black - Metabolism.

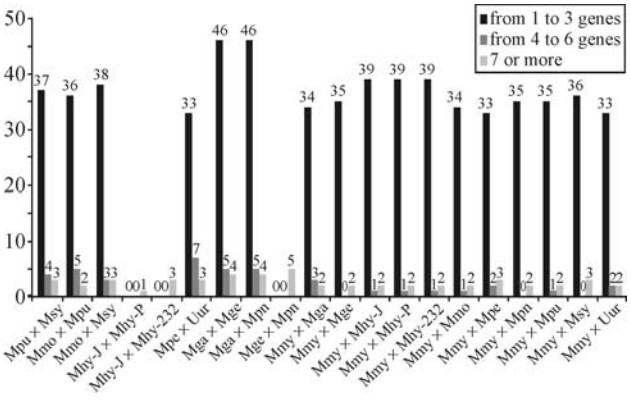


**Figure 3** - Number of rearrangements (changes in the gene order) between a pair of genomes. Only pairs of nearest genomes according to phylogeny are shown. The number over each bar indicates the amount of clusters in each category.

*Mycoplasma genitalium*. The synteny analyses have shown a pattern of clustering in mycoplasmas.

## List of Abbreviations

BBH: bidirectional best hit
COG: Clusters of Orthologous Genes
CDS: Coding sequence
Mge: *Mycoplasma genitalium*
Mpn: *Mycoplasma pneumoniae*
Uur: *Ureaplasma urealyticum*
Mpu: *Mycoplasma pulmonis*
Mpe: *Mycoplasma penetrans*
Mga: *Mycoplasma gallisepticum*
Mmy: *Mycoplasma mycoides* subsp. *mycoides* SC
Mmo: *Mycoplasma mobile*
Mhy-232: *Mycoplasma hyopneumoniae* 232
Msy: *Mycoplasma synoviae*
Mhy-J: *Mycoplasma hyopneumoniae*
Mhy-P: *Mycoplasma hyopneumoniae* 7448

## Acknowledgments

We thank Frank Alarcon and Marcos Oliveira de Carvalho for comments on the manuscript. The present and former staffs from the Ministério da Ciência e Tecnologia (MCT)/Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) are gratefully acknowledged for their strategic vision and enthusiastic support. This work was undertaken by the Brazilian National Genome Program (Southern Network for Genome Analysis and Brazilian National Genome Project Consortium) with funding provided by MCT/CNPq and SCT/FAPERGS (RS).

## References


Almeida LG, Paixão R, Souza RC, Costa GC, Barrientos FJ, Santos MT, Almeida DF and Vasconcelos AT (2004) A System for Automated Bacterial (Genome) Integrated Annotation (SABIA). Bioinformatics 20:2832-2833.

Chambaud I, Heilig R, Ferris S, Barbe V, Samson D, Galisson F, Moszer I, Dybvig K, Wroblewski H, Viari A, *et al*. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res 29:2145-2153.

Charlebois RL and Doolittle WF (2004) Computing prokaryotic gene ubiquity: Rescuing the core from extinction. Genome Res 14:2469-2477.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397-403.

Gil R, Silva FJ, Pereto J and Moya A (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68:518-537.

Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY and Cassell GH (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum.* Nature 407:757-762.

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA 3rd, Smith HO and Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103:425-430.

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC and Herrman R (1996) Complete sequence analysis of the genome of the bacteria *Mycoplasma pneumoniae.* Nucleic Acids Res 24:4420-4449.

Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, *et al.* (2004) The complete genome and proteome of *Mycoplasma mobile.* Genome Res 14:1447-1461.

Koonin EV (2000) How many genes can make a cell: The minimal-gene-set concept. Annu Rev Genomics Hum Genet 1:99-116.

Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature Reviews Microbiol 1:127-136.

Kumar S, Tamura K, Jakobsen IB and Nei M (2001) MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics 17:1244-1245.

Lo SC (1992) Mycoplasmas and AIDS. In: Maniloff J, McElhaney RN, Finch LR and Baseman JB (eds) Mycoplasmas: Molecular Biology and Pathogenesis. American Society for Microbiology Press, Washington, DC, pp 525-545.

Minion FC, Lefkowitz EJ, Madsen ML, Cleary BJ, Swartzell SM and Mahairas, GG (2004) The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. J Bacteriol 186:7123-7133.

Mushegian AR and Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93**:**10268-10273.

Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96:2896-2901.

Papazisi L, Gorton TS, Kutish G, Markham PF, Browning GF, Nguyen DK, Swartzell S, Madan A, Mahairas G and Geary SJ (2003) The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain Rlow. Microbiology 149:2307-2316.

Razin S, Yogev D and Naot Y (1998) Molecular biology and pathogenicity of *Mycoplasmas.* Microbiol Mol Biol Rev 62:1094-1156.

Rocha EPC and Blanchard A (2002) Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res 30:2031-2042.

Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, Furuya K, Yoshino C, Horino A, Shiba T, Sasaki T, *et al*. (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res 30:5293-5300.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV, *et al*. (2001) The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22-28.

Vasconcelos AT, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, Almeida DF, Almeida LG, Almeida R, Alves-Junior L, *et al.* (2005) Swine and poultry pathogens: The complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*. J. Bacteriol 187:5568-5577.

Weisburg WG, Tully JG, Rose DL, Petzel JP, Oyaizu H, Yang D, Mandelco L, Sechrest J, Lawrence TG, Van Etten J, *et al*. (1989) A phylogenetic analysis of the mycoplasmas: Basis for their classification. J Bacteriol 171:6455-6467.

Westberg J, Persson A, Holmberg A, Goesmann A, Lundeberg J, Johansson KE, Pettersson B, Uhlen M, *et al*. (2004) The genome sequence of *Mycoplasma mycoides* subsp *mycoides* SC type strain PG1$^{T}$, the causative agent of contagious bovine pleuropneumonia (CBPP). Genome Res 14:221-227.


## Internet Resources


http://www.ncbi.nlm.nih.gov/COG - Clusters of Orthologous Groups of proteins (verified 01/09/2006).

http://www.ncbi.nlm.nih.gov - GenBank database and BLAST tools (verified 01/09/2006).


*Assistant Editor: Klaus Hartfelder*