Universitat Oberta de Catalunya uoc.edu

Universitat Autònoma de Barcelona

Deciphering host-pathogen interactions at the biomembrane interface, key to target therapeutical approaches against *Mycoplasma genitalium*

Guillem Macip Sancho

Màster universitari en Bioinformàtica i bioestadística UOC-UB

Genómica comparativa y evolución

Nom Consultors

José Luis Villanueva-Cañas

Alex Perálvarez Marín

06/2019

FITXA DEL TREBALL FINAL

Títol del treball:	Deciphering host-pathogen interactions at the biomembrane interface, key to target therapeutical approaches against *Mycoplasma genitalium*
Nom de l'autor:	*Guillem Macip Sancho*
Nom del consultor/a:	José Luis Villanueva-Cañas
Nom del consultor d'empresa	Alex Perálvarez Marín
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament (mm/aaaa):	*06/2019*
Titulació o programa:	*Master's Degree in Bioinformatics and Biostatistics UOC-UB*
Àrea del Treball Final:	*TFM-Bioinformàtica i Bioestadística Àrea 2*
Idioma del treball:	*Anglès*
Paraules clau	*Therapeutical target, Mycoplasma genitalium, host-pathogen interactions*

Abstract:

Mycoplasma genitalium is a pathologic bacterium that lives in the epithelial cells of the urogenital tracts in humans. It is a sexually transmitted infection with a prevalence that increases drastically. It presents resistance to an increasing number of antibiotics, rendering the most reliable treatment (azithromycin) ineffective. The focus of this project is to find likely therapeutical targets against the infection by *M.genitalium* to *Homo sapiens*. For this regard, we analyse the membrane proteome of several strains and near species of *M.genitalium* and calculate their evolutionary rate.

In this project we choose 54 membrane proteomes (5 from *M.genitalium* and 49 from *M.pneumoniae*) compromised of 78 to 93 and 116 to 137 proteins respectively. They were inputted to an orthology predictor to obtain groups of orthologous proteins. Those that contained *M.genitalium* proteins (97 from 136) were aligned and then analysed looking specifically for gaps and identity between the sequences. Of the 97, it was finally decided that 25 were fitted to calculate the nonsynonymous (amino acid–altering) to synonymous (silent) substitution ratio (dN/dS).

We also did a parallel research inside a host-pathogen interaction database with two different approaches. The first consisted in contrasting the data inside the database regarding *M.pneumoniae* to *M.genitalium* proteins with a BLAST. The second consisted in submitting the canonical membrane proteome of *M.genitalium* (G37).

With these three approaches, we obtained a total 34 possible targets: 4 from the dN/dS calculation, 7 from proteins exclusive to *M.genitalium,* 11 from the first approach to the database and 13 from the second.

Índex

Contents

Llista de figures

1. Introduction

1.1 Context and project justification

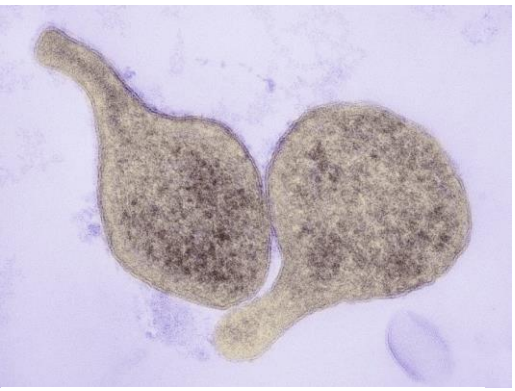


Figure 1. *Mycoplasma genitalium* cells. Photo Credit: Thomas Deerinck, NCMIR / Science Source

Mycoplasma genitalium is a pathologic bacterium that lives in the epithelial cells of the urogenital tracts in humans [Weinstein SA et al., 2012]. *In vitro* they are capable of infecting any human cell line and even other mammals. It is a sexually transmitted infection (STI) with a prevalence that has increased drastically in the recent times [Manhart LE et al., 2007]. It presents resistance to a lot of antibiotics [Jensen JS et al., 2015] and with each passing year the number keeps increasing, rendering the most reliable treatment (*azithromycin*) ineffective.

The infection with *M.genitalium* usually causes urethritis, both in men and women, and cervicitis and pelvic inflammatory diseases (PID) in women. Because of that, it tends to be confused with *Chlamydia trachomatis* infections [Wiesenfeld HC et al., 2017]. In extreme cases, the infection with *M.genitalium* provokes spontaneous abortion, female infertility and it seems that it can play a role in the development of prostate and ovarian cancers. In coinfection with other pathogens, it heightens the virulence of the other pathogen and it is highly associated with HIV-1 [World Health Organization (WHO), 2013].

Until 2003 it was regarded as the species with the smallest genome and it's the organism of choice for research into the *minimal genome* [Fraser CM et al., 1995].

Given the small number of genes it has, specifically 525 for the canon strain G37, and mostly focusing our view in the proteins of the cellular membrane since it's the first point of contact for the infections, we decided to perform an analysis of its membrane proteome, which has 84 proteins annotated. The vast majority are essential for the bacterium survival and many have an unknown function, branded as *Uncharacterized protein* in the Uniprot database.

Another complication the pathogen presents for its treatment is the MG281 protein, commonly known as protein M [Lerner et al., 2014] [Grover RK et al., 2014], which is virtually a universal antibody-binding protein as it is known to be reactive against all antibody types tested.

The focus of this project is to find likely therapeutical targets against the infection by *M.genitalium to Homo sapiens*. For this regard, we analyse the membrane proteome of several strains and near species of *M.genitalium*.

1.2 Objectives of the project

1. Analysis of the similarities between *Mycoplasma genitalium* and close species at the cell membrane level to obtain data.
    1.1. Select strains and related species from *M.genitalium* to obtain their proteomes.
    1.2. Select proteins that have transmembrane segments to obtain the membrane proteome of each species/strains.
    1.3. Obtain the groups of orthologs of the membrane proteins from each species/strain.
2. Discover likely therapeutical targets involved in the infection by *Mycoplasma genitalium*.
    2.1. Obtain the alignment of each orthogroup.
    2.2. Identify gens of membrane proteins with a high evolutionary rate in *M.genitalium*.
    2.3. Obtain possible proteins from *M.genitalium* that interact with the host.
3. Learn to carry out a purely bioinformatics research to guide my professional career to the bioinformatics world.
    3.1. Analyse how other projects with similar approaches are done.
    3.2. Shape my results in a TFM format.

1.3 Approach and method followed

Ideally, we would want to do this analysis with the whole proteome, but due to the times constraints, we focused only in the cell membrane, since this is the likely first point of interaction between the *M.genitalium* and the *Homo sapiens* cells.

The closest sequenced species to *M.genitalium* is *M.pneumoniae* and it is widely studied. The real difference between the two organism is the loss of around 200 genes that *M.genitalium* suffered while adapting from the respiratory tract to the urogenital. Most of the proteins are conserved between these two species, and those with distinct evolutionary rate will be targets for therapeutical approaches since that might signify a host-pathogen competition.

To calculate the evolutionary rate of proteins, it is usually used the Codon-Substitution Models [Yang Z et al., 2002] [Nowick DK, 2012]. The most robust test is the branch-site test [Yang Z et al., 2011], but since that requires a lot of time and power to run, it was finally decided to use the free-ratio model to calculate the nonsynonymous (amino acid–altering) to synonymous (silent) substitution rate ratio (ω = dN/dS). This provides measure of the natural selection at a protein level, with omega = 1, >1, and <1, indicating neutral evolution, purifying selection, and positive selection, respectively.

Since the HPIDB 3.0 (Host-Pathogen Interaction DataBase) is a big repository of the information we seek [Ammari MG et al.,2016], we decided to run a parallel research using the host-pathogen interaction data from *M.pneumoniae* and searching with the canonical membrane proteome of *M.genitalium*.

1.4 Work planification

The following tasks are based on the Methodology Chapter:

	18/3 - 31/3	01/4 - 14/4	15/4 - 28/4	29/4 - 12/5	13/5 - 26/5	27/5 - 05/6
2.1						
2.2						
2.3						
2.4						
2.5						
2.6						
2.7						
2.8						

2.1. Computing environment preparation

2.2. Data acquisition and conversion

2.3. Topology prediction

2.4. Dataset creation

2.5. Orthology prediction

2.6. Multiple sequence alignment

2.7. dN/dS calculation

2.8. Parallel research

1.5 Brief summary of the products obtained

- 54 **membrane proteome datasets**, 5 from *Mycoplasma genitalium* and 49 from *Mycoplasma pneumoniae.*
- **Phylogenetic tree** of all species/strains.
- 97 **orthogroups** of membrane proteins from *Mycoplasma genitalium* and *Mycoplasma pneumoniae.*
- 97 **multiple sequence alignment** with its **phylogenetics trees** of membrane proteins from *Mycoplasma genitalium* and *Mycoplasma pneumoniae.*
- 2 datasets of possible **host-pathogen interactions** between *Mycoplasma genitalium* and *Homo sapiens*.
- 34 **proteins** from *Mycoplasma genitalium* that could interact with *Homo sapiens.*

1.6 Brief description of the chapters of the memory

- Methodology→ Materials and methods explained in detail.
- Results→ Presentation of the results of each task.
- Discussion→ Analysis of possible therapeutic targets against *M.genitalium* and conclusions.

2. Methodology

2.1. Computing environment preparation

This project was mostly done with the help of two programming languages (R, and Python) and a command language (Bash). The following parts of this section are an explanation of the computing environment preparation.

R and Rstudio

R is a programming language and a free software environment. It has gained wide acceptance as a reliable computational environment for statistical computing and visualization. One of the main reasons that computational biologists use R is the Bioconductor project, which is a set of packages for R to analyse genomic data [Eglen SJ et al., 2009]. In this project it is not used Bioconductor, since the packages needed are found on the CRAN repository (Comprehensive R Archive Network).

The installation of this program was done in Windows 10. It is also installed RStudio, an integrated development environment, which facilitates de usage of R and enables the use of Rmarkdown. The following packages are needed to run the scripts.

- readr→ Fast and friendly wat to read data in format “csv”, “tsv” among others.
- rentrez→ “Connect” the R session to the NCBI’s EUtls’API, allowing searches into its databases, like GenBank and PubMed.
- seqinr→ Data analysis and visualization for DNA and protein sequences.
- dplyr→ Consistent tool for working with data frame objects.

Oracle VM VirtualBox and Ubuntu Mate 18.04

Since we need to use several programs that require Linux to operate correctly, one of the fastest solutions is to install a virtual machine inside our computer. For this purpose, we use the Oracle VM VirtualBox a powerful x86 and AMD64/Intel64 virtualization product, which we can run on Windows, Linux, Macintosh… and supports a large number of guest operating systems.

For this project, we install it on Windows 10 and run the latest stable version of Ubuntu Mate as a virtual machine.

Orthofinder2

OrthoFinder2 is a fast, accurate and comprehensive platform for comparative genomics. It finds orthogroups and orthologs, infers rooted gene trees for all orthogroups and identifies the gene duplication events in those gene trees [D.M. E et al., 2018].

We installed the last release of the program (beta 2.3.1), in the Ubuntu virtual machine from its github repository.

Python and Jupyter Notebook

The programs we want to execute on Linux can be run using command line (Bash), but since we need to do it for many files, one solution is to do a quick script in python. Python is the other programming language we use in this project. Like most of the programs we install on Linux, it can be done with bash. We also install the Jupyter Notebook, which servers for python in the same way than Rstudio does to R, enabling Markdown.

The following lines are needed for the installation of both tools with the Ubuntu terminal:

```
sudo apt Update
sudo apt install python3-pip python3-dev
python3 -m pip install --upgrade pip
python3 -m pip install jupyter
```

For its usage, we go to the desired folder and opening the terminal there, we type

```
jupyter notebook
```

Which will open our web browser with the notebook ready to use.

PRANK

PRANK is a probabilistic multiple alignment program for DNA, codon and amino-acid sequences. It's based on a novel algorithm that treats insertions correctly and avoids over-estimation of the number of deletion events. In addition, PRANK borrows ideas from maximum likelihood methods used in phylogenetics and correctly takes into account the evolutionary distances between sequences [Löytynoja A. et al., 2014].

The following lines are needed for the installation of this tool with the Ubuntu terminal:

```
sudo apt-get update
sudo apt-get install prank
```

triamAl

trimAl is a tool for automated alignment trimming, which is especially suited for large-scale analyses. Its speed and the possibility for automatically adjusting the parameters to optimize the phylogenetic signal-to-noise ratios for different families, makes trimAl especially suited for large-scale phylogenomic analyses, involving thousands of large multiple sequence alignments [Capella-Gutiérrez S et al., 2009]. In this project we will use this tool to calculate identity and gaps in the alignments.

The installation of this tool on Ubuntu is done with the same code as ETE (see below) since it's part of the ete3 toolkit.

PAL2NAL

PAL2NAL is a program that converts a multiple sequence alignment of proteins and the corresponding DNA (or mRNA) sequences into a codon alignment. The program automatically assigns the corresponding codon sequence even if the input DNA sequence has mismatches with the input protein sequence, or contains UTRs, polyA tails [Suyama M et al., 2006].

The following lines are needed for the installation of this tool with the Ubuntu terminal:

```
sudo apt-get update
sudo apt-get install pal2nal
```

ETE

The Environment for Tree Exploration (ETE) is a computational framework that simplifies the reconstruction, analysis, and visualization of phylogenetic trees and multiple sequence alignments [Huerta-Cepas J et al., 2007]. In this project we will be using ete-evol for the calculation of dN/dS with the free-ratios model (fb) in CodeML.

The following lines are needed for the installation of this pro with the Ubuntu terminal:

```
wget http://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh -O
Miniconda3-latest-Linux-x86_64.sh
bash Miniconda3-latest-Linux-x86_64.sh -b -p ~/anaconda_ete/
export PATH=~/anaconda_ete/bin:$PATH;
conda install -c etetoolkit ete3 ete_toolchain
```

2.2. Data acquisition and conversion

Since we needed to obtain the proteomes of several strains to evaluate the similarities between *M.genitalium* and *M.pneumoniae,* we accessed several webs services. The data was finally obtained from the web of the National Center for Biotechnology (NCBI) from the database "genome". The keywords were "*Mycoplasma genitalium*" and "*Mycoplasma pneumoniae*". From the respective pages, we accessed the "Genome Assembly and Annotation report" link, which leaded us to a table with all the strains available for the organism. From the column "Protein" we then accessed and downloaded a new table with the information of the proteome of each strain minus the sequence. The column "FTP", through the green rhombus, allowed us to access a directory with the RefSeq information of the strain. From then, we downloaded the "cds_from_genomic.fna". We chose to use the information available only from strains that had the complete sequence, indicated by the column "Level". Both files from each strain were renamed with the name of the species and the strain.

Seeing that the table does not contain the sequence of the proteins, we then wrote a script that, with the help of the "rentrez" package, allowed us to obtain the fasta sequence using the "Protein.product" column, which contains the accession number from RefSeq. The protein sequences of each strain were added as a new column to the corresponding dataset as well to a new file in fasta format.

The cds file was introduced to Rstudio, with the "seqnir" package `read.fasta` function. With a script we then changed the identifier of each sequence to contain only the RefSeq accession number.

2.3. Topology prediction

Once we have the sequences of the proteomes of each strain, we then need to select those with transmembrane segments to obtain the membrane proteome. TOPCONS is a web server for consensus prediction of membrane protein topology [Tsirigos KD et al., 2015]. We submit the proteome fasta file of each strain.

When the prediction is finished, the server allows us to download a compressed file with all the information. From all the files inside it, we only need "finished_seqs.txt", which contains a table with the summary of the predictions of each fasta. Since it does not contain the sequence of the protein, but the identifier of the fasta, we need to merge it with the original dataset to extract the sequences.

2.4. Dataset creation

Calling the files from TOPCONS, cds and the dataset, we wrote a script to merge all the information from each strain, creating a new table with the protein sequence, cds and its number of transmembrane segments. We then selected only the proteins with transmembrane segments to obtain the dataset of the membrane proteome of each strain. From each dataset, we create a protein sequence file and a cds file, both in fasta format.

2.5. Orthology prediction

Now that we have the membrane proteins of each strain, we need to group them by homology. Since the species are close related, those homologs are called orthologs, as they come from the same ancestry. To calculate those orthologs we use Orthofinder.

Putting the protein sequence files of all the strains into a folder, and using the Ubuntu virtual machine, we run Orthofinder. From all the files the programs outputs we need “Orthogrups.tsv” and the files from the folder “Orthogroup_Sequences”. The program also generates phylogenetic trees for each protein and merge them all into a single one.

With the help of Rstudio we create a script to eliminate the orthogroups that do not contain any *M.genitalium* protein.

2.6. Multiple sequence alignment

To compare the proteins of each orthogroup, we need to perform multiple sequence alignment, for that purpose we use PRANK.

Using Jupyter Notebook inside the Ubuntu virtual machine, we create a script to run PRANK on each Orthogroup_Sequence’s files. The output is an alignment of each orthogroups and a phylogenetic tree between the proteins of each file.

Automated alignment analysis.

Since we have a lot of groups, we need a quick way to analyse of all them. Using Jupyter Notebook inside the Ubuntu virtual machine, we create a script to run trimAl so that it calculates the gaps and the identity of the sequences. With that we can select those orthogroups more likely to be meaningful and hence those ready to run in the free-ratios model.

Even if some orthogroups are not optimized for dN/dS calculation, the analysis of the trimAl output can help us find interesting proteins. In our case, proteins exclusive to *M.genitalium* or that few *M.pneumoniae* strains contain, proteins that are larger in *M.genitalium* than *M.pneumoniae* or proteins really changed between the two species.

2.7. dN/dS calculation

To calculate the evolution rate of the proteins, we need to perform a dN/dS calculation using the free ratios model. For this we need to use ete3-evol module. Since the program runs on nucleotides, we first need to convert our alignments of proteins into DNA.

Protein to DNA conversion

Using Jupyter Notebook inside the Ubuntu virtual machine, we create a script to run pan2nal so that the alignments are translated from amino acids to DNA. For that, we need first to obtain the cds for the proteins of each orthogroup.

Using Rstudio we create several scripts. First, we need to merge all the dataset strains into one. Since the datasets have a column called “Identifier” that corresponds exactly with the fasta identifiers, both protein and cds, we need another script that obtains that “Identifier” from each alignment, extracts from the dataset the correct cds sequence and output each alignment file with DNA instead of amino acids.

Then we can run the Jupyter script for pal2nal with the alignment files and the corresponding fasta cds to translate the alignments to DNA.

Free-ratios model

Using Jupyter Notebook inside the Ubuntu virtual machine, we create a script to run ete3 evol with the free-ratios model. The inputs needed for this are the files obtained from pal2nal and the corresponding phylogenetic tree obtained from PRANK in newick format.

2.8. Parallel research

Since our goal is to find therapeutical targets against the infection of *M.genitalium* to *Homo sapiens* we can search for already stablished host-pathogen interactions. HPIDB 3.0 is the biggest database available. Since there is no information annotated about *M.genitalium*. we decided to use the information about M.pneumoniae it contains. Searching the site by keyword, changing the query type to “Taxon Name / Species” and introducing “*Mycoplasma pneumoniae”*, we can download a table that contains which proteins interact with *Homo sapiens*.

Using Rstudio we create a script to eliminate the duplicated proteins and retrieve from the table the accession number from Uniprot. We then access the Uniprot database to retrieve the fasta sequence of each protein. To see whether these proteins could interact with human proteins, we run protein BLAST in the NCBI web service against *M.genitalium* to infer from its identity and query coverage their likelihood of homology.

From HPIDB 3.0 we can also input the *M.genitalium* membrane proteomes into the database using the search by Homologous HPI to see if there are other pathogens that have homology with these proteins. Since we were evaluating membrane proteins, we changed the Percent Identity Filter to 20% since for that type of proteins having more than that is really relevant [Chandonia J-M et al., 2006].

3. Results

3.1. Initial Data

During the data acquisition phase, we obtained the protein tables from 5 strains of *M.genitalium* and 49 strains of *M.pneumoniae*.

Mycoplasma genitalium strains→G37, M2288, M2321, M6282 and M6320.

Mycoplasma pneumoniae strains→309, 519, 549, 685, 1006, 1134, 1801, 19294, 39443, 51494, 54089, 54524, 85084, 85138, C267, CO3, CO103, E16, E57, FH, FH_2, FH_2009, FH-tet-R, FL1, FL8, GA3, K27, KCH-402, KCH-405, M29, M129, M129_2002, M129-B7, M1139, M2192, M2592, MAC, NCTC10119, PI_1428, PO1, RI3, S4-tet-R, S12-tet-R, S34-tet-R, S55-tet-R, S63-tet-R, S68-tet-R, S91-tet-R and S355.

The starting datasets were comprised of 11 columns containing the information of the replicon name (X.Replicon.Name), the replicon accession (Replicon.Accession), the start of the gene (Start), the end of the gene (Stop), the strand (Strand), the gene id (GeneID), the locus (Locus), the locus tag (Locus.tag), the Refseq accession (Protein.product), the length of the protein (Lenght) and its name (Protein.name).

Each dataset contains between 451 to 515 proteins for *M.genitalium* strains and 680 to 743 for *M.pneumoniae*.

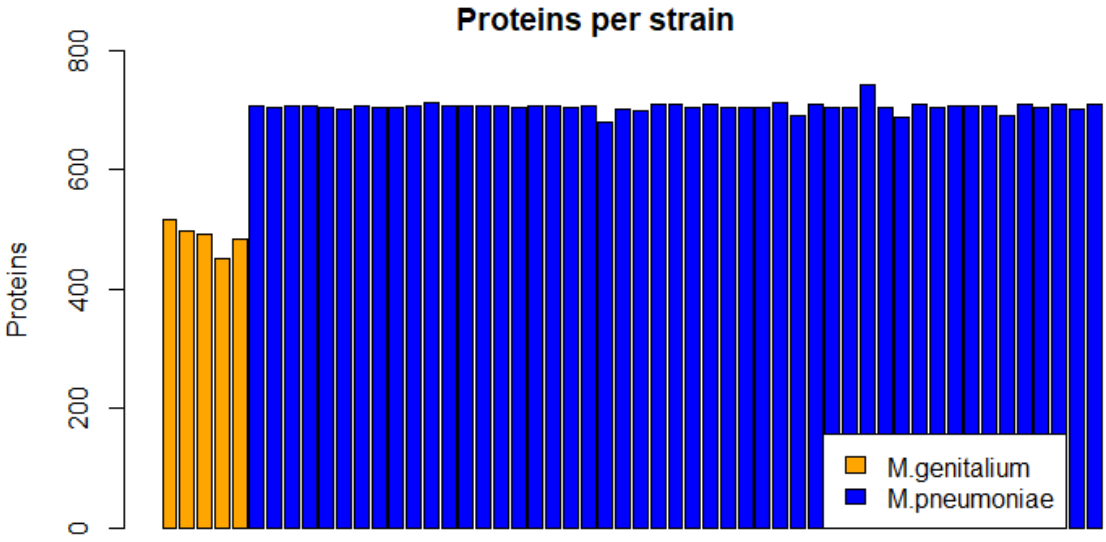


Figure 2. Number of proteins per strain. We can clearly see that *M.genitalium* has around 200 less than *M.pneumoniae*.

With the help of a script, a new column (Fasta) was added with the sequence of each protein, which then was used to obtain files containing the proteomes of the 54 species/strains in fasta format.

The other file we downloaded from each strain is the cds of the proteins in fasta format, which we modified the identifier of so that later we can match them to its respective protein. We saved each file as a 2-column table, one containing the identifier (protein_id) and the other the nude nucleotide sequence (sequence).

3.2. TOPCONS prediction

The file from each prediction is comprised of 8 columns containing the information of the sequence number (V1), the length (V2), its transmembrane segments (V3), the presence of signal peptide (V4), if the sequence was already in the cache of the web server (V5), the time needed to complete the prediction (V6), the identifier of the fasta sequence (V7) and the date of the run (V8).

3.3. "All-information" membrane proteome strain dataset

Merging the proteome table, TOPCONS and cds files from each strain, we created new datasets containing all the relevant information for each membrane protein. Those columns from each file that do not contain relevant information were dropped before the merging. Below there is list of the changed made to accomplish the final dataset:

- Proteome tables→ The columns X.Replicon.Name, GeneID and Locus were discarded.
- TOPCONS files→ Only the columns V3, V4 and V7 are added to the corresponding dataset. They were renamed to "numTM", "SignalPeptide" and "Identifier" respectively. The column Identifier was modified, adding the name of the strain at the start, just after the ">" symbol.
- cds files→ The columns were renamed "Protein.product" and "cds". Protein.product is then merged completely into the column of the same name from the Proteome tables, so that cds and protein match completely.

Since the Protein.product (Refseq accession) of these proteins are sometimes the same between the different strains, to facilitate the differentiation of each protein, two new columns were created. Both in fasta format, they contain the protein sequence (Soca) and the cds sequence (cds_fasta) respectively, with the Identifier column as the fasta identifier.

We then subset the datasets to contain only those proteins with transmembrane segments.

Each dataset contains between 78 to 93 membrane proteins for *Mycoplasma genitalium* strains and 116 to 137 for *Mycoplasma pneumoniae*.

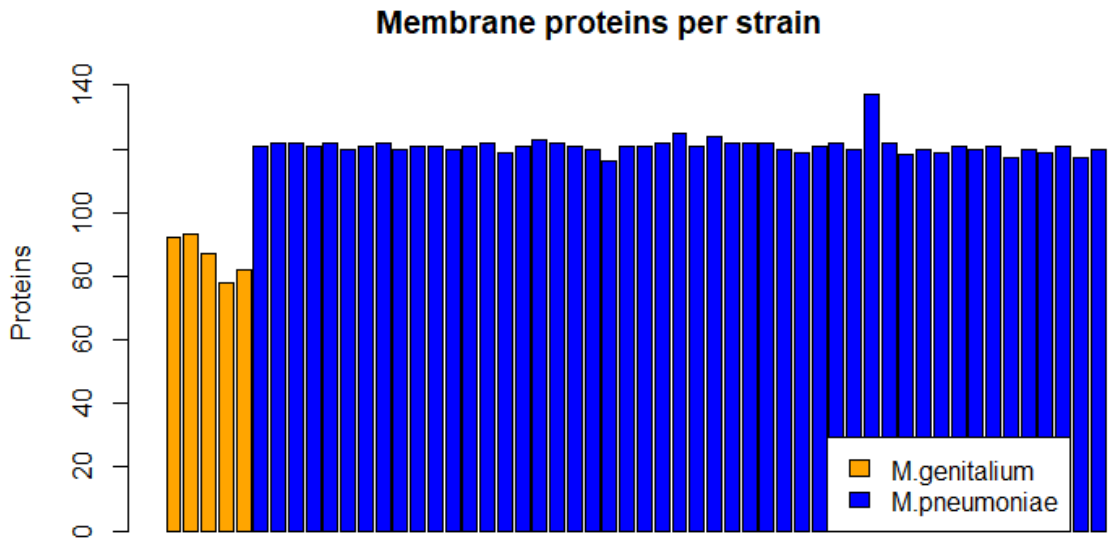


Figure 3. Number of membrane proteins per strain. We can clearly see that *M.genitalium* has around 30 less than *M.pneumoniae*.

From this dataset we extract the “Soca” column to obtain the files of the membrane protein sequences in fasta format.

3.4. Orthogroups

After the orthology prediction with the membrane protein, we obtained a total of 136 groups. These groups contain what the program predicted as orthologs. Since we only need those with *M.genitalium* proteins, we run a script to detect the orthogroups that do not have any, so that we can discard those that contains exclusively *M.pneumoniae* proteins. This tool also calculates a phylogenetic tree between the input strains.

We obtain a total of 97 groups, ranging from 103 to 2 proteins per group. Most of orthogroups contain around 54 to 50 sequences meaning that the proteins inside are one-to-one orthologs. This kinds of orthologs are really relevant because they signify a close relation between species.

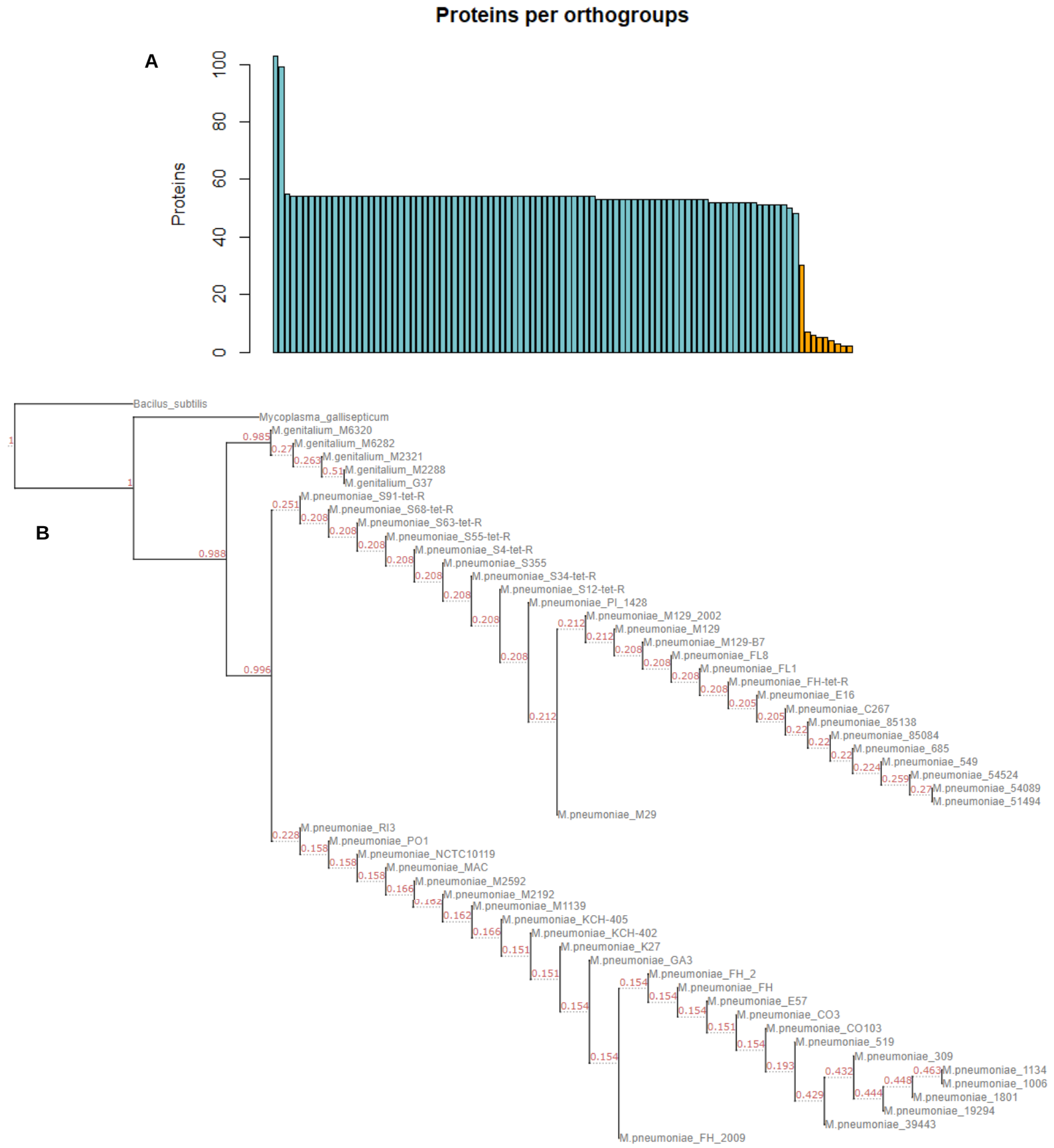


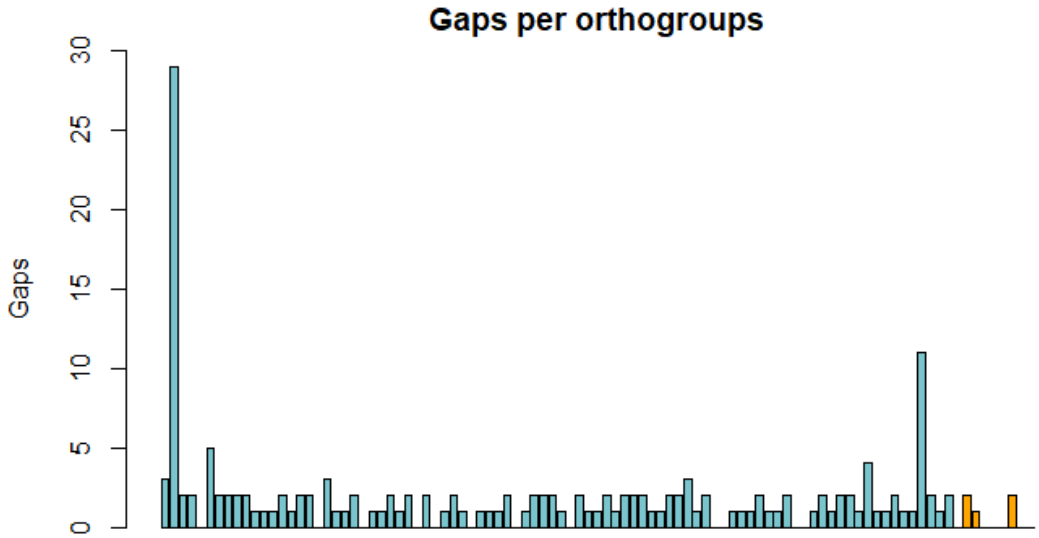
Figure 4. A. Number of proteins per orthogroups. The final columns, filled with orange, contain a higher percentage of *M.genitalium* proteins. **B.** Phylogenetic tree of the strains/species used. It includes *Bacillus subtilis* and *M.gallisepticum* as reference species to improve the distance calculation.

3.5. Alignment analysis

After aligning all 97 orthogroups using PRANK, we obtained the alignment of each in fasta format as well as its phylogenetic tree.

We used trimAl to output the gaps of the alignments and the identity of each sequence against the others present in the alignment. Since we were planning to do a free-ratio model, we selected those with fewer gaps and better identity. Of the original 97 alignments, 25 were the best fitted, and from those not fitted, 7 were interesting enough to do bibliographic research on them.

A



B

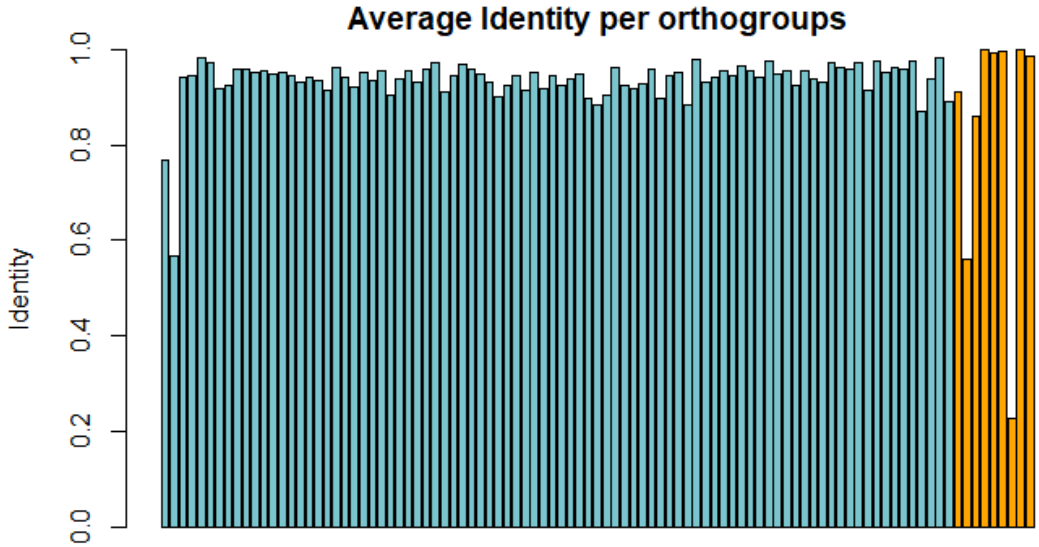
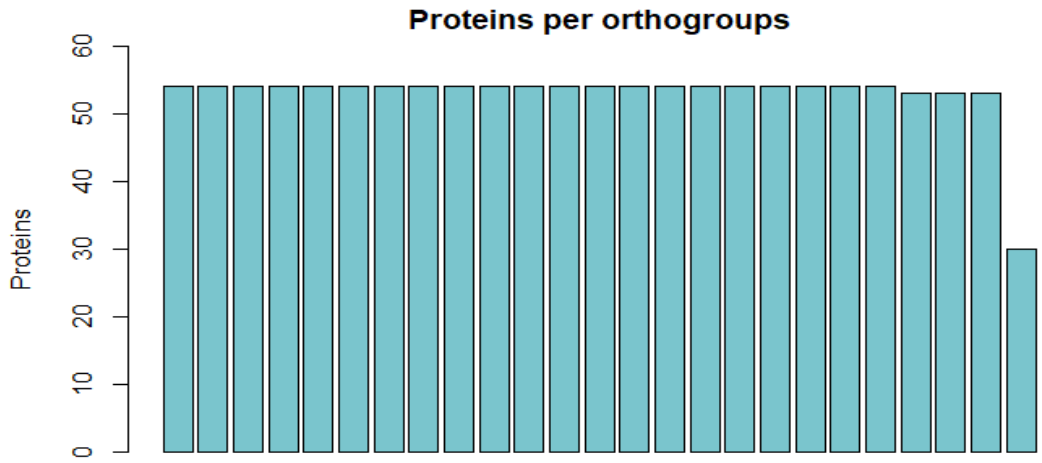


Figure 5. A. Number of gaps in each alignment. **B.** Average identity between all the proteins in the alignments. Since vast majority of proteins in most of the orthogroups was from *M.pneumoniae* the identity percentages are higher than usual. We took as limiting point 95% instead of 90% for choosing which to input in the ete3-evol free-ratios model.

3.6. ete3 evol

Since the program has problem with large Identifier in the fasta, we shorten them so that they only contain the strain and the Refseq Accession. Using the 25 alignments and its corresponding trees obtained from PRANK to run the free-ratios model, we obtain 4 alignments with a dN/dS higher than one and 21 of the alignments that had a ratio lower than one, meaning that most of the proteins are really conserved.

A



B

Orthogroup	dN	dS	ω (dN/dS)
OG0000003	12.14	6562.37	0.0018
OG0000004	0.02	203.45	0.0001
OG0000005	4.28	5827.11	0.0007
OG0000009	0.54	5350.12	0.0001
OG0000010	8.31	0.08	104.5878
OG0000011	18.10	1794.83	0.0101
OG0000012	15.41	5843.50	0.0026
OG0000013	9.67	5775.02	0.0017
OG0000014	0.64	6440.45	0.0001
OG0000019	3.55	0.00	861.5005
OG0000022	0.15	1465.46	0.0001
OG0000027	0.00	15.09	0.0001
OG0000029	6.92	0.02	287.8374
OG0000030	2.38	5682.10	0.0004
OG0000033	1.94	0.00	999.0000
OG0000034	8.55	6571.63	0.0013
OG0000035	14.82	5992.28	0.0025
OG0000039	10.11	6157.48	0.0016
OG0000041	17.02	716.80	0.0237
OG0000043	19.86	765.40	0.0259
OG0000050	5.92	6454.17	0.0009
OG0000062	19.44	431.37	0.0451
OG0000066	21.02	5698.24	0.0037
OG0000067	0.00	0.00	0.8320
OG0000115	21.63	476.24	0.0454

Figure 6. A. Number of proteins of each orthogroup introduced in the free-ratios model. **B.** Output of the free-ratios model for the *Mycoplasma genitalium* branch of the tree. Values are rounded to two decimals for dN, dS and to four for ω**.**

3.7. HPIDB 3.0

After the search with "*Mycoplasma pneumoniae"* as a keyword in the HPIDB 3.0, we obtain a table comprised of 15 columns containing the information about the interactions between *Homo sapiens* and *M.pneumoniae* proteins. The columns contain the following information:

The Uniprot id of the human (protein_xref_1) and the *M.pneumoniae* protein (protein_xref_2), alternative identifiers of the proteins from column 1 and 2 (alternative_identifiers_1, alternative_identifiers_2), alias of the proteins from column 1 and 2 (protein_alias_1, protein_alias_2), the detection method of the interaction (detection_method), the author (author_name), the PUBMED Identifier (pmid), the taxon of the proteins from column 1 and 2 (protein_taxid_1, protein_taxid_2), the type of interaction (interaction_type), the id from the source database (sourcer_database_id), the identifier of the database (database_identifier) and the confidence in the interaction (confidence).

Of the 45 rows in the table, we found that several proteins were duplicated since the database saves each publication where the proteins were found to interact. After running the script to eliminate the duplicated *M.pneumoniae* proteins, we found 12 singular proteins that interact with 8 from humans.

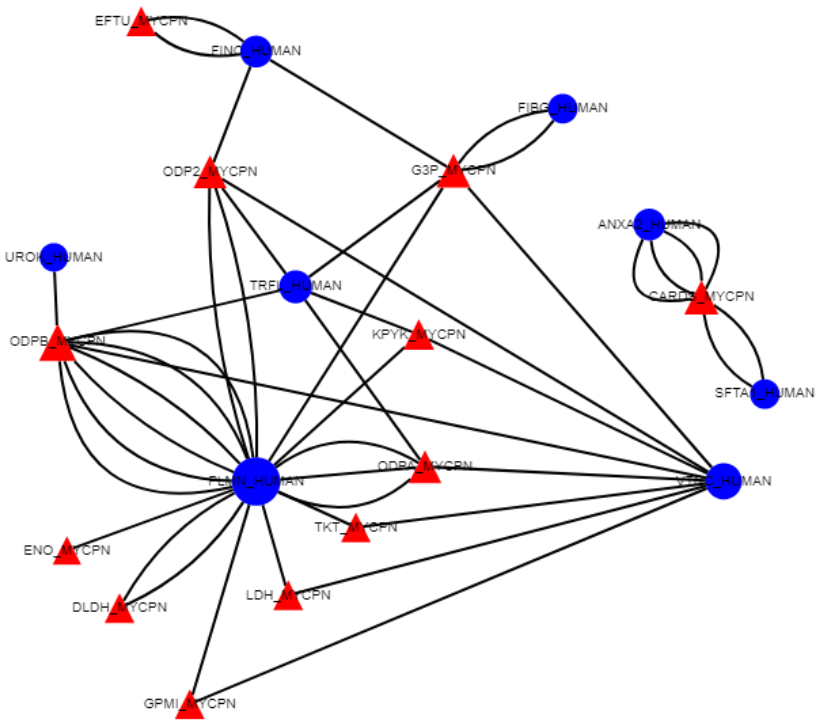


Figure 7. HPIDB Network Visualization based on HPIDB Interaction Records from *Mycoplasma pneumoniae.* Each line corresponds to a publication that found the interaction.

From these 12 proteins, we obtain their fasta thanks to the column protein_xref_2 and the Uniprot database. After performing the protein BLAST in the NCBI web service, we obtain 11 proteins from *M.genitalium* with a query cover higher than 97%. If we reproduced a new network visualization, the disappearance of three isolated proteins (two from humans and one from *M.pneumoniae*) found on the right, would be the only change from the figure above.

M.pneumoniae	**Display ID**	***M.genitalium***	**Query cover**	**Identity**
uniprotkb:P23568	EFTU_MYCPN	uniprotkb:P13927	100	96.70
uniprotkb:P75392	ODP2_MYCPN	uniprotkb:P47514	100	71.43
uniprotkb:P75358	G3P_MYCPN	uniprotkb:P47543	100	82.20
uniprotkb:P75391	ODPB_MYCPN	uniprotkb:P47515	100	88.99
uniprotkb:P78031	KPYK_MYCPN	uniprotkb:P47458	100	78.15
uniprotkb:P75390	ODPA_MYCPN	uniprotkb:P47516	100	88.27
uniprotkb:P75409	CARDS_MYCPN	uniprotkb:P47572	4	39.29
uniprotkb:P75393	DLDH_MYCPN	uniprotkb:P47513	99	72.81
uniprotkb:P78007	LDH_MYCPN	uniprotkb:P47698	100	82.69
uniprotkb:P75189	ENO_MYCPN	uniprotkb:P47647	97	79.73
uniprotkb:P75611	TKT_MYCPN	uniprotkb:P47312	100	71.76
uniprotkb:P75167	GPMI_MYCPN	uniprotkb:P47669	99	71.43

Submitting the whole membrane proteome of *M.genitalium* G37, since it is its canonical strain, on a HPIDB 3.0 internal BLAST, we obtained 13 proteins that could interact with 59 proteins from *Homo sapiens*.

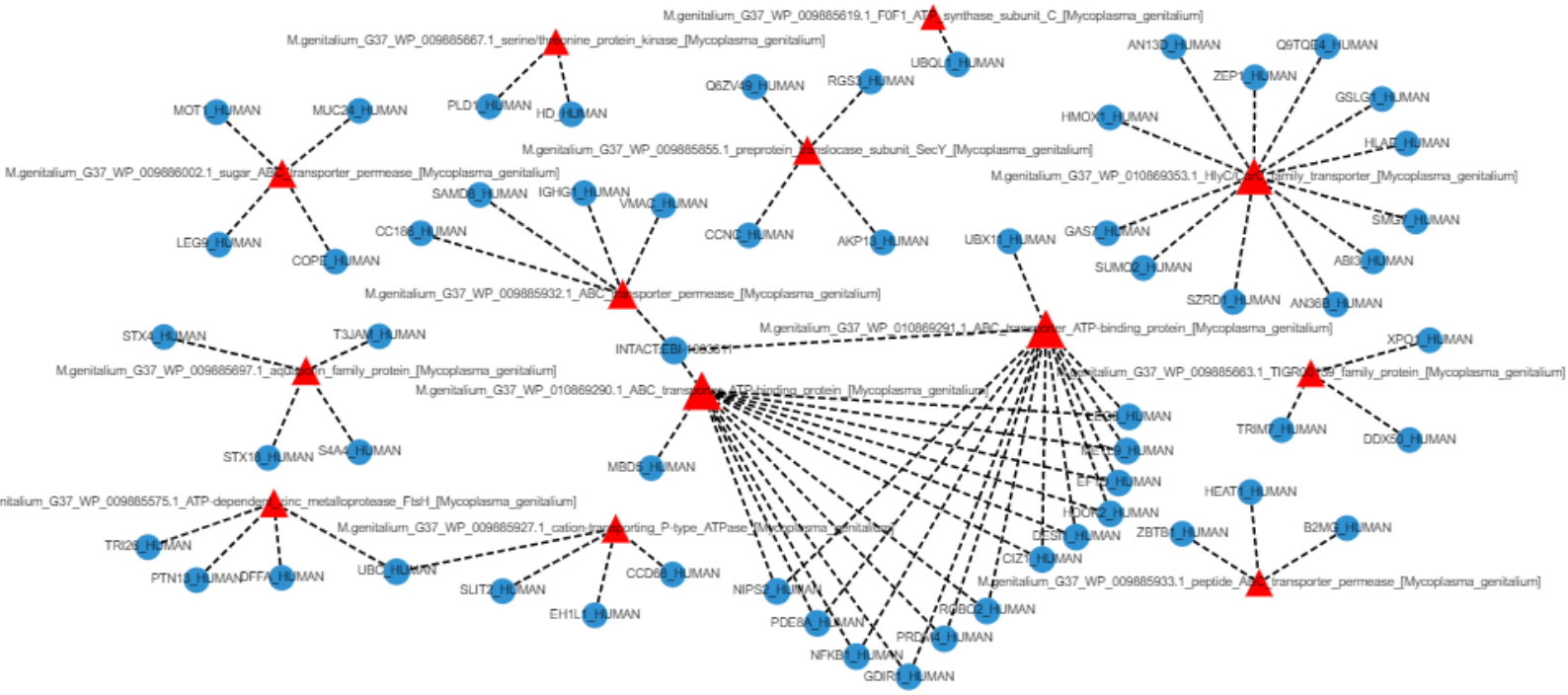


Figure 8. HPIDB Network Visualization based on HPIDB Interaction Records from a BLAST with Mycoplasma genitalium G37 membrane proteome against pathogens with *Homo sapiens* interactions.

M.genitalium	**Query cover**	**Identity**	**Uniprot Pathogen protein**	**Pathogen name**
WP_009885575.1	66	53.715	A0A0J1HL43	Bacillus anthracis
WP_009885619.1	65	37.879	P61829	Saccharomyces cerevisiae
WP_009885663.1	75	39.073	A0A2A8KPE5	Bacillus anthracis
WP_009885667.1	76	28.912	Q81WH6	Bacillus anthracis
WP_009885697.1	90	33.740	Q5NIE2	Francisella tularensis ssp. tularensis
WP_009885855.1	87	37.915	A0A1Q4MAW6	Bacillus anthracis
WP_009885927.1	81	33.377	A0A384KPX3	Yersinia pestis
WP_009885932.1	74	27.526	A0A1Q4M2S2	Bacillus anthracis
WP_009885933.1	99	28.191	A0A0J1I2N1	Bacillus anthracis
WP_009886002.1	83	26.316	Q7CKV6	Yersinia pestis
WP_010869290.1	97	30.807	A0A0J1HNX0	Bacillus anthracis
WP_010869291.1	83	27.038	Q0WER9	Yersinia pestis
WP_010869353.1	78	24.157	A0A0F7RM01	Bacillus anthracis

3.8 Summary of potential therapeutical targets

The following tables are the presentation of all the proteins that we consider worth mentioning. The first column is the name of the file it contains them or its Uniprot/NCBI Accession name and the second is its conversion to the Ordered Locus name of the canonical strain of *M.genitalium.*

Alignment analysis results	**Ordered Locus**
OG0000058	MG144
OG0000085	MG415
OG0000127	MG137
OG0000128	MG441
OG0000129	Missing→ WP_009885728.1
OG0000130	MG255 + MG494 (protein fusion)?
OG0000134	MG279

Ete3-evol results	**Ordered Locus**
OG0000010	MG403
OG0000019	MG044
OG0000029	MG077
OG0000033	MG335.1

HPIDB 3.0 results (*M.pneumoniae* data)	**Ordered Locus**
P13927	MG451
P47514	MG272
P47543	MG301
P47515	MG273
P47458	MG216
P47516	MG274
P47513	MG271
P47698	MG460
P47647	MG407
P47312	MG066
P47669	MG430

HPIDB 3.0 results (proteome BLAST)	Ordered Locus
WP_009885575.1	MG457
WP_009885619.1	MG404
WP_009885663.1	MG105
WP_009885667.1	MG109
WP_009885697.1	MG033
WP_009885855.1	MG170
WP_009885927.1	MG071
WP_009885932.1	MG077
WP_009885933.1	MG078
WP_009886002.1	MG188
WP_010869290.1	MG014
WP_010869291.1	MG015
WP_010869353.1	MG146

4. Discussion

4.1 Final analysis

Alignment analysis results

During the analysis of the alignments, several of the files that did not make it into the cut for the next step, drew our attention. Upon further research we decided that they might good candidates:

- MG144→This protein displays a series of large gaps if we compare it to *M.pneumoniae*. Since previously Orthofinder grouped them together, marking them as orthologs, it most likely implies that the one from *M.genitalium* suffered heavy modifications under some kind of evolutionary pressure since between strains of the same specie it really conserved.
- MG415→ The N-term of the *M.genitalium* strains are almost 300 aa larger than their *M.pneumoniae* counterpart. After an NCBI protein BLAST, it might be that this protein is in truth a fusion of two other from *M.pneumoniae*.
- MG137, MG441 and WP_009885728.1→These proteins share that they are unique of *M.genitalium*. In specific, of the 49 strains of *M.pneumoniae*, only a MG137 ortholog appears in *M.pneumoniae* FL8.

- MG255 + MG494→ This protein seems to be an anomaly. Inside the orthogroup predicted with Orthofinder, of the 5 strains, only the canonical strain *M.genitalium* G37 was missing. After an NCBI protein BLAST, it was found that in this strain, the protein is separated into two loci. It is unclear if this is a fusion of proteins or if the canonical strain, for some reason, truncated it into two, but since the strains of *M.pneumoniae* were present in the group, we think the later.
- MG279→This one is the less likely to be a key factor for the infection of *M.genitalium* to *Homo sapiens* since it only appears in three of the five strains, but since it is in the canonical strain, it might be interesting to in research it.

Ete3-evol results

From the 25 alignments, after the free ratios test, only MG403, MG044, MG077 and MG335.1, had a ratio higher than 1 for the branch containing the *M.genitalium* strains. Does this signify that these proteins are under selective pressure? Without using the branch site test for positive selection and with more species, we cannot be sure of that, but with this information we know that these proteins evolve at a higher rate than the other tested. Usually branches with dS < 0.01 are discarded since they give inflated omega values [Villanueva-Cañas et al., 2013] but since none of the results entered in that criteria, it was decided not to apply this criterion.

HPIDB 3.0 results from M.pneumoniae data

Since the information of the host-pathogen interactions between *Homo sapien*s and *M.pneumoniae* are about cytosolic or secreted proteins, they might not be useful as therapeutic targets, but since the query coverage and the identity of the sequence with some of the *M.genitalium* proteins are high enough, we decided to present them here:

MG066, MG216, MG271, MG272, MG273, MG274, MG301, MG407, MG451, MG430 and MG460. They might not be useful in creating drugs to prevent the infection but can maybe aid in the research of blockers of *M.genitalium* activity.

HPIDB 3.0 results from membrane proteome BLAST

The evolution of the membrane proteins is usually significantly higher than other subsets of proteins. When comparing two of them, any homology or structure conserved between the two, and in specific when that homology is found inside the membrane, means that the structure or sequence is doing an important function. Having a higher than 20% of homology for membrane proteins is really significant. All of the following proteins were found with at least a 20% of homology with the corresponding protein inside the database: MG014, MG015, MG033, MG071, MG077, MG078, MG105, MG109, MG146, MG170, MG188, MG404 and MG457.

MG077 is especially significant because it also appeared in the free-ratio model results.

4.2 Conclusions

- We found 7 proteins that are exclusive to *M.genitalium*.
- From the 25 proteins inputted in the dN/dS calculation, 4 reported an evolutionary rate higher than 1. Because of the dS<0.01 in all of the cases, the dN/dS estimates may be inaccurate, therefore we cannot truly trust these results. Even taking this into account, MG077 appears again as a possible target on the HPIDB second approach.
- The HPIDB 3.0 research from *M.pneumoniae* data reported 11 cytosolic/secreted proteins from *M.genitalium* that could interact with 6 proteins from *Homo sapiens*.
- From the HPIDB 3.0 internal BLAST we found 13 membrane proteins from *M.genitalium* that could interact with 59 proteins from *Homo sapiens*.

4.3 Future Work

- To further validate the results of the approach presented during the calculation of the free-ratio model with the ete3-evol program, it might be worth to lessen the number of *M.pneumoniae* strains and add some other species like *M.gallispecticum* and *Bacillus subtilis*. From that, repeat the methodology until the usage of the ete3-evol and perform the branch-site test instead of the free-ratio.
- 34 proteins are still a large number. If we eliminate the 11 that are not membrane proteins, we end up having 23 that we still need to curate. One of this curation methods can be protein modelling.

4.4 Personal thoughts

The usage of bioinformatics to analyse large quantities of data and hence reduce the load of work that the conventional research needs to perform is one of the main reasons that attracted me to it. In this project we demonstrate that by going from more than 500 proteins to 34, that can be even less depending on the reasons explained before or the preferences of the researcher.

The complications of this project focused less in what to accomplish with the data and more on how to accomplish it. Most of the programs and tools were new to me and some of them proved difficult to setup correctly. Another problem added to this was the corruption of the computing environment, which prompted me to install from scratch a

new virtual machine. Since the power of the computer used during the project was on the mid-low end, some of the programs, in particular PRANK and ete3-evol were instead run on more potent computers.

Apart from the 34 proteins to use as therapeutical targets, we also think that some of the datasets created during the project can be helpful to other *M.genitalium* related research since they contain a really varied array of information.

5. Glossary

- **Codon-Substitution Models** use a series of matrices to translate nucleotides from an alignment to amino acids and then check between the different sequences whether that change was synonymous or not. The nonsynonymous (amino acid-altering) to synonymous (silent) substitution rate ratio (omega = d(N)/d(S)) provides a measure of natural selection at the protein level, with omega = 1, >1, and <1, indicating neutral evolution, purifying selection, and positive selection, respectively
- **HPIDB 3.0** is a resource that helps annotate, predict and display host-pathogen interactions (HPI). HPI that underpin infectious diseases are critical for developing novel intervention strategies. Currently our database contains 69,441 curated entries
- **R** is a programming language and a free software environment.
- **Python** is an interpreted, high-level, general-purpose programming language.
- **Orthogroup** is a group made of orthologous proteins.
- **PRANK** is a probabilistic multiple alignment program for DNA, codon and amino-acid sequences.
- **trimAl** is a tool for automated alignment trimming, which is especially suited for large-scale analyses.
- **PAL2NAL** is a program that converts a multiple sequence alignment of proteins and the corresponding DNA (or mRNA) sequences into a codon alignment
- **The Environment for Tree Exploration (ETE)** is a computational framework that simplifies the reconstruction, analysis, and visualization of phylogenetic trees and multiple sequence alignments.

6. Bibliography

- Eglen SJ. A Quick Guide to Teaching R Programming to Computational Biology Students. Lewitter F, editor. PLoS Comput Biol [Internet]. 2009 Aug 28 [cited 2019 May 21];5(8):e1000482. Available from: https://dx.plos.org/10.1371/journal.pcbi.1000482
- Grover RK, Zhu X, Nieusma T, Jones T, Boreo I, MacLeod AS, et al. A structurally distinct human mycoplasma protein that generically blocks antigen-antibody union. Science [Internet]. 2014 Feb 7 [cited 2019 Mar 12];343(6171):656–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24503852
- Yang Z, Nielsen R. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol [Internet]. 2002 Jun 1 [cited 2019 May 24];19(6):908–17. Available from: http://academic.oup.com/mbe/article/19/6/908/1094851
- Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol [Internet]. 2016 Jun 1 [cited 2019 May 22];33(6):1635–8. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw046
- Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host–pathogen interactions. Database [Internet]. 2016 Jul 3 [cited 2019 May 22];2016:baw103. Available from: https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw103
- Jensen JS, Bradshaw C. Management of Mycoplasma genitalium infections – can we hit a moving target? BMC Infect Dis [Internet]. 2015 Dec 19 [cited 2019 Mar 18];15(1):343. Available from: http://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-015-1041-6
- Manhart LE, Holmes KK, Hughes JP, Houston LS, Totten PA. Mycoplasma genitalium among young adults in the United States: an emerging sexually transmitted infection. Am J Public Health [Internet]. 2007 Jun 10 [cited 2019 Mar 18];97(6):1118–25. Available from: http://ajph.aphapublications.org/doi/10.2105/AJPH.2005.074062
- Wiesenfeld HC, Manhart LE. Mycoplasma genitalium in Women: Current Knowledge and Research Priorities for This Recently Emerged Pathogen. J Infect Dis [Internet]. 2017 Jul 15 [cited 2019 Mar 18];216(suppl_2):S389–95. Available from: https://academic.oup.com/jid/article/216/suppl_2/S389/4040973
- D.M. E, S. K. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. bioRxiv [Internet]. 2018 Nov 8 [cited 2019 May 22];466201. Available from: https://www.biorxiv.org/content/10.1101/466201v1
- Dessimoz C. Orthology : definitions , inference , and impact on species phylogeny inference. [cited 2019 Mar 17];1–17. Available from: https://arxiv.org/ftp/arxiv/papers/1903/1903.04530.pdf
- Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. [cited 2019 May 22]; Available from: http://www.bork.embl.de/pal2nal

- Löytynoja A. Phylogeny-aware alignment with PRANK. In Humana Press, Totowa, NJ; 2014 [cited 2019 May 22]. p. 155–70. Available from: http://link.springer.com/10.1007/978-1-62703-646-7_10
- Weinstein SA, Stiles BG. Recent perspectives in the diagnosis and evidence-based treatment of Mycoplasma genitalium. Expert Rev Anti Infect Ther [Internet]. 2012 Jan 10 [cited 2019 Mar 18];10(4):487–99. Available from: http://www.tandfonline.com/doi/full/10.1586/eri.12.20
- Nowick DK. Sequence analysis and genomics - 7. Tests for selection. Bioinforma Leipzig [Internet]. 2012; Available from: https://www.bioinf.uni-leipzig.de/Leere/WS1213/Sequenzanalyse_Genomik/Vorlesung_PositiveSelection.pdf
- Yang Z, Dos Reis M. Statistical properties of the branch-site test of positive selection. Mol Biol Evol [Internet]. 2011 Mar 1 [cited 2019 May 15];28(3):1217–28. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msq303
- Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of Mycoplasma genitalium. Science [Internet]. 1995 Oct 20 [cited 2019 Mar 18];270(5235):397–403. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7569993
- Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res [Internet]. 2015 Jul 1 [cited 2019 May 22];43(W1):W401-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25969446
- Lerner, Wilson G and Z. The Ultimate Decoy: Scripps Research Institute Scientists Find Protein that Helps Bacteria Misdirect Immune System | Scripps Research [Internet]. www.scripps.edu. 2014 [cited 2019 Mar 18]. Available from: https://www.scripps.edu/news-and-events/press-room/2014/20140206lerner.html
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma Appl NOTE [Internet]. 2009 [cited 2019 May 22];25(15):1972–3. Available from: http://phylemon2.bioinfo.cipf.es/
- World Health Organization (WHO). Laboratory diagnosis of sexually transmitted infections, including human immunodeficiency virus. Switzerland: World Health Organization 2013
- Chandonia J-M, Kim S-H. Structural proteomics of minimal organisms: conservation of protein fold usage and evolutionary implications. BMC Struct Biol [Internet]. 2006 Mar 28 [cited 2019 Jun 4];6:7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16566839Eglen SJ. A Quick Guide to Teaching R Programming to Computational Biology Students. Lewitter F, editor. PLoS Comput Biol [Internet]. 2009 Aug 28 [cited 2019 May 21];5(8):e1000482. Available from: https://dx.plos.org/10.1371/journal.pcbi.1000482

- Villanueva-Cañas JL, Laurie S, Albà MM. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol Evol [Internet]. 2013 Feb 1 [cited 2019 Jun 5];5(2):457–67. Available from: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evt017

7. Annexes

7.1 Possible therapeutical target sequences in Fasta format

MG144

```
>sp|P47390|Y144_MYCGE Uncharacterized protein MG144 OS=Mycoplasma
genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG144
PE=4 SV=1
MDPQNKSPKPQVKSTRLVVKKQPAGVVFPKLSIPVNDFEKTVTLTRAQKKEAKLLKKAQR
KANKLNNKQDSTFFNSASGETNNTILPPGVKNQADNKTNRFSKFISFFTSSKNKQPDEIT
ERLVDDPTVKNRFSAFNKKLIWVLKDKKLRARAWKIVGYTNLVIVAFFAGLLAVMNKFIT
LSSVEYPAIALQLPINNALWGISIFVISIVTLPFWTMFILFLMGVKDVRTSRSIHYFIWI
VLIINVVLLLVSCLLMIAAYAHLDGYNIWRNLESLNPNN
```

MG415

```
>sp|P47655|Y415_MYCGE Uncharacterized protein MG415 OS=Mycoplasma
genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG415
PE=3 SV=2
MSWKRYLKWVSFAIIPLLFANTSIKSKLIDTNLYLVKDDFQSQNQLTIATNQLAKIIVNQ
IEFDSNSLIANPTTVLNKELIGSKITPKLKFSDQFSNAIEMVSKLNQEFDQLANKDKTFF
QFALDLLEKQEESKFDFEPKDERIDAIFFLSNLININPKQEKTLNFIRILPNLIKSIFKD
TTITINIKIGGKNKVITFIENGSNVFLLSDVENFLNADQTGINFYEIEFLTFDFIVVNKT
GWTLKNQPVDSFFKSVKNLPSIQKTKNGFQYSLKFRSEYNEHHILKDHFLIPIVTNQKNF
SVNDIEKNGLNSYQREQITYAIKNSFTSQKENNLNISSATIKYIKDPEKLIKKSLIKPSV
KNGIFYVSAQIINSNDLTKWGSKNDSEIIKDKMYFLEQNKNFPAIRTYLFQMRTKKLVLN
VNDIWFKSSGDKLRVIVNNVEIDEFNPKENNTSFFESYEVHINDYFSLANKELLIKKLNL
ALSEMNLLIDKKKSSLDLFPKEIKLTTLKINSSLHFYLNVDAIKNQLNIEVNISKNRLTS
LVYDIAIKNENELQIRTTNNYLNKYIWFDLDKKNNQKLKNELKLFLSLKKFQFKKEPNFS
LKKNSYSFQIDKIIQSNSEDKKTDIIVYLIIGFSVLVLFITVFIYFHKWNKKQKMIKNKT
RDNF
```

MG137

```
>sp|Q49398|GLF_MYCGE UDP-galactopyranose mutase OS=Mycoplasma
genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=glf
PE=3 SV=1
MNVILSVMLFSSPSCVNINSFDILIVGAGISGIVLANILANHNKRVLIVEKRDHIGGNCY
DKVDSKTQLLFHQYGPHIFHTNNQTVINFISPFFELNNYHHRVGLKLKNNLDLTLPFDFQ
QIYKLMGKDGRKLVSFFKENFSLNTHLSLAELQLIDNPLAQKLYQFLISNVYKPYSVKMW
GLPFAMINENVINRVKIVLSEQSSYFPDAIIQGLPKSGYTNSFLKMLANPLIDVQLNCKD
NLLVYQDEKLFFNNNLIEKPVVYCGLIDKLFNFCFGHLQYRSLAFSWKRFNQKKYQTYPV
VNMPLAKSITRSVEYKQLTNQGSFKPQTIVSFETPGSYAINDPRFNEPYYPINNTLNDTL
FKKYWKKASKLKNLHLLGRLATYQYIDMDKAILLSIKKAQQLLS
```

MG441

```
>sp|P47679|Y441_MYCGE Uncharacterized protein MG441 OS=Mycoplasma
genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG441
PE=4 SV=1
MSVSFLRSKFSLKASVFAFFVLFLFCLKIILVLFRNFGKRFKHFLFNQTSLYLLVRLFQK
TEIVWNLIANIHFFIKTQIQNLGIRLSRESISNETFQAVKLFHVNNLGLQEQEVINSKLS
DYFCFFKYRNLLFVNW
```

WP_009885728.1

>WP_009885728.1 hypothetical protein [Mycoplasma genitalium]
MVKRKRKPKLNSRNILTIQIVLTIFSMIFFLTLLSLILFLSLQSNLATALVENRNKAVELVDNIVFF

MG255 , MG494 and the "fused"

>sp|P47497|Y255_MYCGE Uncharacterized protein MG255 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG255 PE=4 SV=1
MESENQIAILDYIFNQVNQPNQPKIVWFSGEGEDEKINFLIRLNDFFKPKFVENTNDSSF
LLSFRNHVETKNSTPLTQANFANIANKLLAVLFGSLQWKQLNKPTGNWFLVILFLALLWL
RQCWLKLQLTKISKFVNQKGILSFIKQQWPILTTLVTVGTTLGTPVFSLTIAQQDGIKQN
AGNDVFIFLIIFSVFSISLGLVSSLIFLVSSLFSIRQKKTLDALDKVLSKFIDKYFFLDE
KEIKKQLKYQFKNNGVCFFYGFDFDQAEFLEQSMNLMLLLKQTNCFILVGCKESEMTLIK
NKIEPNINLKQNSFYLDLSNEISQVEQISKFNLLFRQLRLSSELFYLEDFFDYLTTAKQI
VNFLF

>sp|Q9ZB77|Y255A_MYCGE Uncharacterized protein MG255.1 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG_494 PE=4 SV=2
MLWSFNNHFVIDNSFKQEYDKPNITAFFNRSLQFFQENSLVLKPELFSLQKYTKDVYGLN
VINQLNLNKHPMLIPLTWDKKQKFISFIESCVQKYSQVKKDNQVFSLTVGKRVFFLLLIN
KQFKQIKLETALKYLGFKTSLGAMDSTTES

>WP_050980911.1 hypothetical protein [Mycoplasma genitalium]
MESENQIAILDYIFNQVNQPNQPKIVWFSGEGEDEKINFLIRLNDFFKPKFVENTNDSSFLLSFRNHVET
KNSTPLTQANFANIANKLLAVLFGSLQWKQLNKPTGNWFLVILFLALLWLRQCWLKLQLTKISKFVNQKG
ILSFIKQQWPILTTLVTVGTTLGTPVFSLTIAQQDGIKQNAGNDVFIFLIIFSVFSISLGLVSSLIFLVS
SLFSIHQKKTLDALDKVLSKFIDKYFFLDEKEIKKQLKYQFKNNGVCFFYGFDFDQAEFLEQSMNLMLLL
KQTNCFILVGCKESEMTLIKNKIEPNINLKQNSFYLDLSNEISQVEQISKFNLLFRQLRLSSELFYLEDF
FDYLTTAKQIVNFLFKTKPNLDQFQENKKKLFDFLALWALVIGTDFEFNNVLWSFNNHFVIDNSFKQEYD
KPNITAFFNRSLQFFQENSLVLKPELFSLQKYTKDVYGLNVINQLNLNKHPMLTPLTWDKKQKFISFIES
CVQKYSQVKKDNQVFSLTVGKRVFFLLLINKQFKQIKLETALKYLGFKTSLGAMDSTTES

MG279

>sp|P47521|Y279_MYCGE Uncharacterized protein MG279 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG279 PE=4 SV=1
MIRLLKKLAVFLIILVGILLLGGIATAGYFAFTYREPINNYYKEGYNKISEYNTEIKKIS
QNIFQNNLVKTLSEVEKSLNEGRKLTQNNSFASGLDSSLNALEGSLKKINNFDSNAAFTQ
IKHTLNNITSFVDQMLEKFPNPNQNDDFKRYLTEVSQILFYTGISIIGAFFVSGFLLILF
TKKVYGVRVSRFNPQRLLKKHLVLLLRDEEVYDAVFGN

MG403

>sp|P47643|ATPF_MYCGE ATP synthase subunit b OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=atpF PE=3 SV=1
MVKAKKLVFKWSLLVFSFFTLSLFLVSCTENVREIKSSSVINELFPNFWVFITHLLAFFI
LLTLMIFLFWKPTQRFLNNRKNLLEAQIKQANELEKQARNLLEESNQRHEKALIVSKEIV
DQANYEALQLKSEIEKTANRQANLMIFQARQEIEKERRSLKEQSIKESVELAMLAAQELI
LKKIDQKSDREFIDKFIRDLEANETEDD

MG044

>sp|P47290|POTC_MYCGE Spermidine/putrescine transport system permease protein PotC homolog OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=potC PE=3 SV=1
MKKHFKNLIKNSYFFLLITLIYLPLLIVVLVSLNGSSSRGNIVLDFGNVLNPNPDSKSAY
LRLGETDFATPLINSIIIGVITVLVSVPIAVISAFALLRTRNALKKTIFGITNFSLATPD
IITAISLVLLFANTWLSFNQQLGFFTIITSHISFSVPYALILIYPKIQKLNPNLILASQD
LGYSPLKTFFHITLPYLMPSIFSAVLVVFATSFDDYVITSLVQGSVKTIATELYSFRKGI
KAWAIAFGSILILISVLGVCLITLQKYLREKRKEIIKIRQWKNS

MG077

>sp|P47323|OPPB_MYCGE Oligopeptide transport system permease protein OppB OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=oppB PE=3 SV=1
MFKYILKRLGLAVVAMFIVMSIVFFLVNATGNVPLSATSARDIAAVQAQLQEFGFNDPII
VRYFRYWAKLFSFQADALGIYYANPNQTIGEIVFARVPNTLYVVLISFLIGSLLGIFLGM
VSGLNRGKFLDAAINVLVVLFVSIPSFVVGLGLLKLAGFLNLPPRFINFDDAFFSFDRFL
LASIIPILSLVFYSSAAFTYRIRNEVVEVMNQDYIKTAKSKGLGMFAVARYHIFRNSIIP
SIPLFVFGISGAFSGGFIIESLFGVQGVSRILIDSVQVNETNMVMFNILFIQGIPLLASV
FIEFIYVLVDPRIRIANSSNVSLLTKLKFLSSRHQWLMKWNKINSDNAQNIVFNSPLHHQ
LLELNAIDYKTKTVQLTTEQKTALNISATANFILLGNKCLKLKTIHG

MG335.1

>sp|Q49310|Y335A_MYCGE UPF0154 protein MG335.1 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG335.1 PE=3 SV=2
MNDLALALGLGIPLSLLVGMILGYFISIKIFKKQMRDNPPITENQIKAMYAKMGRKLSET
QVKEIMRSIKNQK

MG451

>sp|P13927|EFTU_MYCGE Elongation factor Tu OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=tuf PE=3 SV=1
MAREKFDRSKPHVNVGTIGHIDHGKTTLTAAICTVLAKEGKSAATRYDEIDKAPEEKARG
ITINSAHVEYSSDKRHYAHVDCPGHADYIKNMITGAAQMDGAILVVSATDSVMPQTREHI
LLARQVGVPKMVVFLNKCDIASDEEVQELVAEEVRDLLTSYGFDGKNTPIIYGSALKALE
GDPKWEAKIHDLIKAVDEWIPTPTREVDKPFLLAIEDTMTITGRGTVVTGRVERGELKVG
QEVEIVGLKPIRKAVVTGIEMFKKELDSAMAGDNAGVLLRGVERKEVERGQVLAKPGSIK
PHKKFKAEIYALKKEEGGRHTGFLNGYRPQFYFRTTDVTGSIALAENTEMVLPGDNASIT
VELIAPIACEKGSKFSIREGGRTVGAGTVTEVLE

MG272

>sp|P47514|ODP2_MYCGE Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=pdhC PE=3 SV=1
MANEFKFTDVGEGLHEGKVTEILKQVGDQIKIDEALFVVETDKVTTELPSPFAGTISAIN
VKVGDVVSIGQVMAVIGEKTSTPLVEPKPQPTEEVAKVKEAGASVVGEIKVSDNLFPIFG
VKPHATPAVKDTKVASSTNITVETTQKPESKTEQKTIAISTMRKAIAEAMTKSHAIIPTT
VLTFYVNATKLKQYRESVNGYALSKYSMKISYFAFFVKAIVNALKKFPVFNASYDPDQNE
IVLNDDINVGIAVDTEEGLIVPNIKQAQTKSVVEIAQAIVDLANKARTKKIKLTDLNKGT
ISVTNFGSLGAAVGTPIIKYPEMCIVATGNLEERIVKVENGIAVHTILPLTIAADHRWVD
GADVGRFGKEIAKQIEELIDLTVA

MG301

>sp|P47543|G3P_MYCGE Glyceraldehyde-3-phosphate dehydrogenase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=gapA PE=3 SV=1
MAAKNRTIKVAINGFGRIGRLVFRSLLSKANVEVVAINDLTQPEVLAHLLKYDSAHGELK
RKITVKQNILQIDRKKVYVFSEKDPQNLPWDEHDIDVVIESTGRFVSEEGASLHLKAGAK
RVIISAPAKEKTIRTVVYNVNHKTISSDDKIISAASCTTNCLAPLVHVLEKNFGIVYGTM
LTVHAYTADQRLQDAPHNDLRRARAAAVNIVPTTTGAAKAIGLVVPEANGKLNGMSLRVP
VLTGSIVELSVVLEKSPSVEQVNQAMKRFASASFKYCEDPIVSSDVVSSEYGSIFDSKLT
NIVEVDGMKLYKVYAWYDNESSYVHQLVRVVSYCAKL

MG273

>sp|P47515|ODPB_MYCGE Pyruvate dehydrogenase E1 component subunit beta OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=pdhB PE=3 SV=1
MSKIQVNNIEALNNAMDLALERDQNVVLYGQDAGFEGGVFRATKGLQQKYGSERVWDCPI
AENSMAGIGVGAAIGGLKPIVEIQFSGFSFPAMFQIFVHAARIRNRSRGVYTAPLVVRMP
MGGGIKALEHHSETLEAIYAQIAGLKTVMPSNPYDTKGLFLAAIESPDPVIFFEPKKLYR
AFRQEIPSDYYTVPIGEANLISEGSELTIVSYGPTMFDLINLVYSGELKDKGIELIDLRT
ISPWDKQTVFNSVKKTGRLLVVTEAVKSFTTSAEIITSVTEELFTYLKKAPQRVTGFDIV
VPLARGEKYQFEINARVIDAVNQLLK

MG216

>sp|P47458|KPYK_MYCGE Pyruvate kinase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=pyk PE=3 SV=1
MIDHLKRTKIIATCGPALTKSLVSLKMLDDNEYAAIKKVAYANIEAIIKSGVSVIRLNFS
HGTHEEQQVRIKIVRDVAKAMNIPVSIMLDTNGPEIRIVETKKEGLKITKDSEVIINTMS
KMIASDNQFAVSDASGKYNMVNDVNIGQKILVDDGKLTLVVTRVDKQHNQVICVAKNDHT
VFTKKRLNLPNAQYSIPFLSEKDLKDIDFGLSQGIDYIAASFVNTVADIKQLRDYLKLKN
ASGVKIIAKIESNHALNNIDKIIKASDGIMVARGDLGLEIPYYQVPYWQRYMIKACRFFN
KRSITATQMLDSLEKNIQPTRAEVTDVYFAVDRGNDATMLSGETASGLYPLNAVAVMQKI
DKQSETFFDYQYNVNYYLKNSTANKSRFWHNVVLPLTKKTVPKRKLVNSAFKYDFIVYPT
NNINRIYALSNARLAAAVIILTNNKRVYTGHGVDYGIFCYLIDKNPNQLTKAELIELAWK
AINHYQAYGDLEKLKQCLAVYNETIINL

MG274

>sp|P47516|ODPA_MYCGE Pyruvate dehydrogenase E1 component subunit alpha OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=pdhA PE=3 SV=1
MAILIKNKVPTTLYQVYDNEGKLIDPNHKITLTDEQLKHAYYLMNLSRMMDKKMLVWQRA
GKMLNFAPNLGEEALQVGMGLGLNENDWVCPTFRSGALMLYRGVKPEQLLLYWNGNEKGS
QIDAKYKTLPINITIGAQYSHAAGLGYMLHYKKQPNVAVTMIGDGGTAEGEFYEAMNIAS
IHKWNTVFCINNNQFAISTRTKLESAVSDLSVKAIACGIPRVRVDGNDLIASYEAMQDAA
NYARGGNGPVLIEFFSYRQGPHTTSDDPSIYRTKQEEEEGMKSDPVKRLRNFLFDRSILN
QAQEEEMFSKIEQEIQAAYEKMVLDTPVSVDEVFDYNYQELTPELVEQKQIAKKYFKD

MG271

>sp|P47513|DLDH_MYCGE Dihydrolipoyl dehydrogenase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=pdhD PE=3 SV=1
MDYDLIILGAGPAGYIAAEYAGKHKLKTLVIEKQYFGGVCLNVGCIPTKTLLKRAKIIDY
LVHAKDYGITINGQAKLDWKQLLKQKQEVVDKLVAGVKTIIKGAKVESIEGEATVIDKNK
VQVNNTTYTTNNIIVATGSRPRYLTLPGFEKAQQAGFIIDSTQALALEGVPKKFVVVGGG
VIGVEFAFLFASLGSEVTIIQGVDRILEVCDSDVSELISKTLKNKGVQIITNAHVVRAEN
NQLFYTVNGVEQSVIGDKILVSIGRIANTECLDQLDLKRDHNNKIVLNEKLQTSTTNIYL
IGDVNTQMMLAHYAYQQGRYAVDQILNQNQVKPAEKNKCPACIYTNPEVAFVGYSEMELQ
KEKIDYVKSSLPFIYSGKAIADHETNGFVKMMFNPKTGAILGGCIIASTASDIIAELALV
MENNLTVFDIANSISPHPTMNEMVTDVCKKAIFDYFS

MG460

>sp|P47698|LDH_MYCGE L-lactate dehydrogenase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=ldh PE=3 SV=1
MKGPKIAIVGSGAVGTSFLYAAMTRALGSEYMIIDINEKAKVGNVFDLQDASSSCPNFGK
VVAGEYSQLKDYDFIFISAGRPQKQGGETRLQLLEGNVEIMKSIAKEIKKSGFNGVTLIA
SNPVDIMSYTYLKVTGFEPNKVIGSGTLLDSARLRYAIATKYQMSSKDVQAYVIGEHGDS
SVSIISSAKIAGLSLKHFSKASDIEKEFGEIDQFIRRRAYEIIERKGATFYGIGEASADV
AEQILKDTKEVRVVAPLLTGQYGAKDMMFGTPCVLSRKGIEKILEIELSNTEKVALENSI
KVLKDNIKLAKL

MG407

>sp|P47647|ENO_MYCGE Enolase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=eno PE=3 SV=1
MGSSNLNINSKITDIFAYQVFDSRGVPTVACVVKLASGHVGEAMVPSGASTGEKEAIELR
DNDPKNYFGKGVNEAVDNVNKVIAPKLIGLNAFDQLTVDQAMIKLDNTPNKAKLGANAIL
SVSLAVSKAAAKAQNSSLFQYISNKLIGLNTTNFVLPVPMLNVINGGAHADNYIDFQEFM
IMPLGAKKMHEALKMASETFHALQNLLKKRGLNTNKGDEGGFAPNLKLAEDALDLDIMVEAI
KLAGYKPWDDIAIAIDVAASEFYDEDKKLYVFKKGIKANILNAKDWSLTSKEMIAYLEKL
TKKYPIISIEDGLSENDWEGMNQLTKTIGSHIQIVGDDTYCTNAELAKKGVAQNTTNSIL
IKLNQIGSISETIQTIEVAKKANWSQVISHRSGETEDTTIADLAVAAQTGQIKTGSMSRS
ERIAKYNRLLYIEIELGDKGKYLGWNTFTNIKPKNFNI

MG066

>sp|P47312|TKT_MYCGE Transketolase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=tkt PE=3 SV=1
MKYLYATQHLTLNAIKHAKGGHVGMAIGASPILFSLFTKHFHFDPDQPKWINRDRFVLSA
GHGSMALYSIFHFAGLISKQEILQHKHGQINTSSHPEYAPNNFIDASTGPLGQGFGMAVG
MVLAQKLLANEFKELSDKLFDHYTYVVVGDGDLQEGVSYEVSQIAGLYKLNKLIVLHDSN
RVQMDSEVKKVANENLKVRFENVGWNYIHTDDQLENIDQAIIKAKQSDKPTFIEVRTTIA
KNTHLEDQYGGHWFIPNEVDFQLFEKRTNTNFNFFNYPDSIYHWFKQTVIERQKQIKEDY
NNLLISLKDKPLFKKFTNWIDSDFQALYLNQLDEKKVAKKDSATRNYLKDFLNQINNPNS
NLYCLNADVSRSCFIKIGDDNLHENPCSRNIQIGIREFAMATIMNGMALHGGIKVMGGTF
LAFADYSKPAIRLGALMNLPVFYVYTHDSYQVGGDGPTHQPYDQLPMLRAIENVCVFRPC
DEKETCAGFNYGLLSQDQTTVLVLTRQPLKSIDNTDSLKTLKGGYILLDRKQPDLIIAAS
GSEVQLAIEFEKVLTKQNVKVRILSVPNITLLLKQDEKYLKSLFDANSSLITIEASSSYE
WFCFKKYVKNHAHLGAFSFGESDDGDKVYQQKGFNLERLMKIFTSLRN

MG430

>sp|P47669|GPMI_MYCGE 2,3-bisphosphoglycerate-independent phosphoglycerate mutase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=gpmI PE=3 SV=2
MHKKVLLAILDGYGISNAIYGNAVQNANTPMLDELINSYPCVLLDASGEAVGLPMGQIGN
SEVGHLNIGAGRVVYTGLSLINQHIKDRSFFANKAFLKTIEHVEKNHSKIHLIGLFSNGG
VHSHNEHLLALIELFSKHAKVVLHLFGDGRDVAPCSLKQDLEKLMIFLKNYPNVVIGTIG
GRYYGMDRDQRWDREMIAYKALLGVSKNKFNDPIGYIETQYQNQITDEFIYPAINANLNS
DQFALNNNDGVIFFNFRPDRARQMSHLIFNSNYYNYQPELKRKENLFFVTMMNYEGIVPS
EFAFPPQTIKNSLGEVIANNNLKQLRIAETEKYAHVTFFFDGGFEVNLSNETKTLIPSLK
VATYDLAPEMSCKAITDALLEKLNNFDFTVLNFANPDMVGHTGNYQACIKALEALDVQIK
RIVDFCKANQITMFLTADHGNAEVMIDNNNNPVTKHTINPVPFVCTDKNVNFNQTGILAN
IAPTILEYLNLSKPKEMTAKSLLKNNN

MG457

>sp|P47695|FTSH_MYCGE ATP-dependent zinc metalloprotease FtsH OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=ftsH PE=3 SV=1
MKKRNKGLVEQTTTEKNNFSRKTAWKVFWWVIILAVVIGVLAYIFSPRAATAVVESWKLN
GGSNSTLTAKVSGFSNELTFKQINGSTYVTDTILQVSITFDGLNSPLTVTAHKTVNSNGN
VIFNIANLSINQSNGQITVNSNGTMMNGGSSNNTKSIAGFETLGTFIAPDTRARDVLNGL
FGLLPIIIFVVFFLLFWRSARGISAGGREEDNIFSIGKTQAKLAKSTVKFTNIAGLQEEK
HELLEIVDYLKNPLKYAQMGARSPRGVILYGPPGTGKTLLAKAVAGEAGVPFFQSTGSGF
EDMLVGVGAKRVRDLFNKAKKAAPCIIFIDEIDSVGSKRGRVELSSYSVVEQTLNQLLAE
MDGFTSRTGVVVMAATNRLDVLDDALLRPGRFDRHIQINLPDIKEREGILKVHAENKNLS
SKISLLDVAKRTPGFSGAQLENVINEATLLAVRDNRTTININDIDEAIDRVIAGPAKKSR
VISDEDRKLVAYHEAGHALVGLHVHSNDEVQKITIIPRGQAGGYTLSTPKSGDLNLKRKS
DLLAMIATAMGGRAAEEEIYGNLEITTGASSDFYKATNIARAMVTQLGMSKLGQVQYVPS
QGTLPSNVKLYSEQTAKDIDNEINFIIEEQYKKAKTIIKSNRKELELLVEALLIAETILK
SDIDFIHKNTKLPPEILLQKQEQQAKQKLNKSEVKPESETNS

MG404

>sp|P47644|ATPL_MYCGE ATP synthase subunit c OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=atpE PE=3 SV=1
MEHVNEILATVGVILQQTQTTQDVNASAKLGAYIGAGVTMIAGSTVGIGQGYIFGKAVEA
IARNPEVEKQVFKLIFIGSAVSESTAIYGLLISFILIFVAGA

MG105

>sp|P47351|DACB_MYCGE Diadenylate cyclase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=dacB PE=3 SV=2
MVVNILLFITLIFLLLLFVFLIAFAFLNKRVRNYVVRTWTSVFSKSKQNLDKKNFFDNLT
STLLRLSVDKIGAIIAIEKRDSLDPYINIGYRVSSDFSPELLVTIFYNKSSPLHDGAVIV
RDYKIISVSSYFPMTRQLIDVSYGSRHRSALGLSEKSDAVVFIVSETTGKISVALKGVIK
TLSSNSDRLQDEIIHYLSSK

MG109

>sp|P47355|PKNS_MYCGE Putative serine/threonine-protein kinase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG109 PE=3 SV=2
MEAILKIGDIVENKYQIEKLLNRGGMDSYLFLAKNLNLKNYGPVQKKQYGHLVLKVVQKN
PKINENNWKKFLDEMVTTTRVHHSNLVKSFDVVNPFLKIVRGNKTIALNQIVMIAMEYVD
GPSLRQLLNRKGYFSVSEVVYYFTKIVKAIDYLHSFKHQIIHRDLKPENILFTSDLTDIK
LLDFGIASTVVKVAEKTEVLTDENSLFGTVSYMIPDVLESTVNKAGKKVRKPPNAQYDIY
SLGIILFEMLVGRVPFNKSINPNKERETIQKARNFDLPLMQATRSDIPNSLENIAFRCTA
VKRENNKWLYSSTKELLEDLANWENEQAMIKPANERVLEGQVEIREMMLEKPLAWYFKTW
ALSIFTIVFIGLIIAAIVLLLIFNARF

MG033

>sp|P47279|GLPF_MYCGE Probable glycerol uptake facilitator protein OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=glpF PE=3 SV=2
MYFQNSTQLGWWFLAELIGTFILIIFGNGAVAQVNLKKMATSETKAKFLTVALTWGIGVL
FGVLTANAIFKGSGHLNPAISLFYAINGSIKSPTALIWPGFVIGILAQFLGAMIAQTTLN
FLFWKQLSSTDPQTVLAMHCTSPSVFNITRNFLTEFIATLILIGGVVAASHFLHNNPNSV
PPGFMGLWLVAGIIIAFGGATGSAINPARDLGTRIVFQLTPIKNKDANWKYSWIPVIAPL
SAGLVLSIIIGFSPAPVL

MG170

>sp|P47416|SECY_MYCGE Protein translocase subunit SecY OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=secY PE=3 SV=1
MQTVSSPKQKLNFGQRLLTLLQNRDFMVSLVLTVVLLILFRVLAIIPLPGIRINESVLDR
NSNDFFSLFNLLGGGGLNQLSLFAVGISPYISAQIIMQLLSTDLIPPLSKLVNSGEVGRR
KIEMITRIITLPFALVQAFAVIQIATNAGTGSSPISLANSGSEFIAFYIIAMTAGTYMAV
FLGDTISKKGVGNGITLLILSGILSQLPQGFIAAYNVLSGIVITLTPQLTAAISFFIYFL
AFLVLLFATTFITQATRKIPIQQSGQGLVSEVKTLPYLPIKVNAAGVIPVIFASSIMSIP
VTIAQFQPQTESRWFVEDYLSLSTPVGIFLYAVLVILFSFFYSYIQINPERLAKNFEKSG
RFIPGIRPGNDTEKHIARVLIRINFIGAPFLTVIAIIPYIVSYFIRLPNSLSLGGTGIII
IVTAVVEFISALRSAATATNYQQLRRNLAIEVQQTAKQDSLEQLQKEAPGIGNLW

MG071

>sp|P47317|ATCL_MYCGE Probable cation-transporting P-type ATPase OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=pacL PE=3 SV=1
MNSWTGLSEQAAIKSRQEHGANFLPEKKATPFWLLFLQQFKSLVVILLLLASLLSFVVAI
VSGLRSNWNFNHDLIIEWVQPFIILLTVFANSLIGSIQEFKAQKSASALKSLTKSFTRVF
RNGELISINVSEVVVGDIIFVDAGDIIPADGKLLQVNNLRCLESFLTGESTPVDKTIDSN
EKATILEQTNLVFSGAQVVYGSGVFQVEAVGIKTQVGKIAKTVDDSVTKLSPLQQKLEKI
GKWFSWFGLGLFAVVFLVQTALLGFDNFTNNWSIALIGAIALVVAIIPEGLVTFINVIFA
LSVQKLTKQKAIIKYLSVIETLGSVQIICTDKTGTLTQNQMKVVDHFCFNSTTQTDLARA
LCLCNNASISKDANKTGDPTEIALLEWKDRSQLDLKTYYRVYEKAFDSIRKLMTVVVQKD
NRFIVIVKGAPDVLLPLCNNVQNEVKNIENLLDQSAGQGLRTLAVALKVLYKFDQNDQKQ
IDELENNLEFLGFVSLQDPPRKESKEAILACKKANITPIMITGDHLKTATVIAKELGILT
LDNQAVLGSELDEKKILDYRVFARVTPQQKLAIVSAWKEAGFTVSVTGDGVNDAPALIKS
DVGCCMGITGVDIAKDASDLIISDDNFATIVNGIEEGRKTFLTCKRVLLNLFLTSIAGTV
VVLLGLFILGQVFKTNLLQQGHDFQVFSPTQLLIINLFVHGFPAVALAVQPVKEKLMVGS
FSTKNLFYNRQGFDLIWQSLFLSFLTLLFYSLGIIYAINNRDLQTSGDLINRAGSTCGFF
ILGASAALNSLNLMVDKPLLMTNPWFFKLVWIGSLASILVFLLIIFINPLGLVFNVLQDL
TNHPVLISYSFGGVILYMGMNEVVKLIRLGYGNI

MG078

>sp|P47324|OPPC_MYCGE Oligopeptide transport system permease protein OppC OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=oppC PE=3 SV=1
MDRNKSFDPNLFKRVDINLLKRNDQLIGKPTTNSIEIIKRLFQNKWAILFFLLIVVIVLL
AIIVPLTSPFSAVTPVSTNALAQNLPPRYLWHKPGDILVHKITARSIAEISQASGVLVGT
LPSANSNPLATNVQYDIAPFQLQELRNYFPLLGTNGLGIDIWTLLWASVAKSLWIAVVVA
IIAMVFGTIYGAVAGSFVGHMADNIMSRIIEIIDIVPSILWIIVLGATFRFGGVKQFDDS
VVIFTLIFVFWTWPATTTRIYILKNKDTEYIQAAKTLGAHQIRIIFVHMLPVVFGRLAVV
FVSLIPAVIGYEASLVFLGLKPATDIGLGALLNQVTSSDNVALILSSIVSFAVLTVAART
FANALNDAIDPRVVKR

MG188

>sp|P47434|Y188_MYCGE Probable ABC transporter permease protein MG188 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG188 PE=3 SV=1
MFKWLLKHHNQPHSLQLGLLDQPLPFWKPFLLFLPALLTTILFTIIPFFLSLQKGFSANS
DLYDLSSQSFSLRTFQDLFSESNFVLGLRNSFLYSLISLPFSIIIAIVIASAIVFVYKKL
LRGFWQTVFFLPYVTSGVAISIAFVYIFDSASGILNTVFNVNTKWLDSGSRDTFNALWAI
LIFGVWKNLAFNVLIISTAMLSVNPQLYKVASLDSANPVRQFFKITLPSIRPTLIFLTTL
LILGGMQVFPLALFENKPEEAVANGGNSILLYIFQQIQSGNTNLAGAATLVLFVLGVCYG
LVLRNGFYLIEWLQWKIKQLYVQKQLTLY

MG014

>sp|P47260|Y014_MYCGE Putative ABC transporter ATP-binding protein MG014 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG014 PE=3 SV=1
MGLVLKEFNNKIRTALILAPFFTFAQIVIDLIIPSFLASAISVVFSIDKLKQDESGGKTI
SVDFIGGANINFANVREAQIVLATTVILLALCGLFFGLISIYCASYVSANTSFLLRKKIF
AKLMRITTPSHDHYGSSTLLVRLTNDVYLMEVIAFDFLRLIIRAPLLFIGGLVFAVTTNQ
DMSISLLITFPLILLVIGILNRKSIPLFKENQKSVDKINERVEEDVSGYKVIQSFNLHSF
TNNKFKIANEGWKKNSTSSLFINSLNIPFTFFLSSLTIIIALLLVFQLDSSVSVDPLPQD
AAIRPNIFAFFQYNFYIVLGFILTSLTMVNFNRSRVALGRIKDILSQPEIKTITNKDQKE
LLPTLEFRNISFGLGNKNNNNFLQNLSFKFEAYKTYGIVGPTGSGKSLIANIIGGLYEPN
EGEILIGGEKIQSIDSLYLSEMIGIVFQQNILFKGTISSNIKIGIETRSDWKNQSDLQKN
EAMKNAAKIACADTFIEKFSDSYDHNVEQLGKNLSGGQKQRVAIARTLITKPRILVFDDS
MSALDALTEKKVRENIENDLKLTTKIIISQNINSIKHADKILVIDNGRIVGFDSDQKLMK
NCSLYQKMKESQKDLGGDFDAVN

MG015

>sp|P47261|Y015_MYCGE Putative ABC transporter ATP-binding protein MG015 OS=Mycoplasma genitalium (strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG015 PE=3 SV=1
MEGSWSYKLLYVFLCIVLGILYGIANPILLAQGLGFIFPITSSNGRAVDSIYSLIYPTNL
NVFIRLTIVSVTVFVAYALIFVFNVAQNYVGIKLYQQTCATLRWKAYLKMQSMSTSFFDT
QNNGDLMSRLTNDMYNIDNLFTQAGGQAIQSLFNILTTSVLIFLLSPVIALISLSILATL
ITFSFAFLKKSKTSYSQVQNNLGDMSGYIEEVLTNHKVVHVLKLQEIMIKDFDQYNKSMI
KPTVRGNTYSIFLFSWFGFISNITYLVSISIATAFSVNSIPSFGISVINYSFMLSYIASL
RQITLALDQIFTLWNLVQLGVVSAERVFKVLDLNVEKDTATIDKLPDIKGNIRFENVAFG
YNKDKPTLTGINFSVKHGDVVAIVGPTGAGKSTIINLLMKFYKPFEGKIYMDNFEISDVT
KKAWREKISIVLQDSFLFSGTIKENIRLGRQDATDDEIIAACKTANAHDFIMRLPKGYDT
YISNKADYLSVGERQLLTIARAVIRNAPVLLLDEATSSVDVHSEKLIQESIGRLMKNKTS
FIISHRLSIIRDATLIMVINDGKVLEMGNHDQLMKQNGFYARLKQSSVR

MG146

```
>sp|Q49399|Y146_MYCGE UPF0053 protein MG146 OS=Mycoplasma genitalium
(strain ATCC 33530 / G-37 / NCTC 10195) OX=243273 GN=MG146 PE=3 SV=1
MDSAPSGLTLTVIILSIILLAFISTVVSAYETAITSLTPYRWKNYIKTNNKQDKLSTKII
NHFQNHYSSCLITILITNNIVAIMVSNILFLALEQTIKNELLSSVLNLVVSGVLIVSFCE
ILPKTLGRINVIRTLVLFAYLVYFFYLIFWPITKLTSLILKKYENPLPVSRKDVYYFIDE
IEQNGLFSKEDSLLIKKTLIFDQVLVKKVMIKWKKVAYCYLNDSINLIAKQFLQRQFSRM
PVVDKTTNKIVGFIHLKDFFTAKEANPKSLDLNQLLYPVVLVQDSTPIKQALRQMRLNRA
HLAVVNDKHEKTIGIVSMEDIIEELVGEIYDEHDDIQPIQVLDENVWLVLPNVKAAYFFN
KWIKPDLVKSKNITIQHYLASLDNDSFACQNKLDTPLFSVEVIADSEDKTKILYEIRKKS
DVIA
```

7.2 R Scripts

Obtain protein sequence in fasta format from RefSeq Accession

```
#Call the proteomes files
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/Inicials")
nm <- list.files(pattern="*.txt")
for (i in 1:length(nm)) assign(nm[i], read.delim(nm[i]))

#Plot protein number in each file
p<-c()
for(i in 1:length(nm)){
  L<-get(nm[i])
  x<-nrow(L)
  p<-c(p,x)
}
barplot(p, xlim=c(0, 60), ylim=c(0, 800), col=c(rep("orange", 5),
rep("blue", 49)))
title(main = "Proteins per strain", ylab="Proteins")
legend("bottomright", legend = c("M.genitalium", "M.pneumoniae"), fill
= c("orange","blue"))

#Obtain FASTA and add as a new column
for(x in nm){
  Fasta = c()
  organisme<-get(x)
  for(i in 1:nrow(organisme)){
    a=entrez_fetch(db="Protein", id=organisme$Protein.product[i],
rettype="fasta")
    a=gsub("\n", "", a)
    a=gsub("]","]\n", a)
    Fasta[i] = a
  }
  v <-  cbind(organisme, Fasta)
  m <-paste(paste("Proteomes/Amb Fasta/",x,sep=""),".csv",sep="")
  write.csv(v, m, row.names = FALSE)
}

#Call the new files
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/Amb Fasta")
```

```
nm <- list.files(pattern="*.csv")
for (i in 1:length(nm)) assign(nm[i], read.csv(nm[i]))

#Fasta file of each proteome's strain
for(x in nm){
  proteoma<-get(x)
  x <-gsub(".txt.csv", "", x)
  m <-paste(paste("Proteomes/Fasta/",x,sep=""),".fasta",sep="")
  write.table(proteoma$Fasta, m, row.names = FALSE, col.names = FALSE,
quote = FALSE)
}
```

Change the identifier from each cds sequence

```
#Call cds files of each strain
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/cds")
nm <- list.files(pattern=".fna")
for (i in 1:length(nm)) assign(nm[i], read.fasta(nm[i],as.string =
TRUE, set.attributes = TRUE, forceDNAtolower = FALSE))

#Create cds table
for(i in 1:length(nm)){
  L<-get(nm[i])
  n <- length(L[[1]])
  df <- structure(L, row.names = c(NA, -n), class = "data.frame")
  df <- as.data.frame(t(df))
  df <- cbind(rownames(df), data.frame(df, row.names=NULL))
  colnames(df)<-c("protein_id","cds_fasta")
  f<-df$protein_id
  f<-gsub("^([^_]*_[^_]*_[^_]*)_([^_]*_[^_]*)_.*$", "\\2", f) #obtain
only Refseq
  df <- cbind(f,df)
  df <- df[ ,-2]
  colnames(df)<-c("protein_id","sequence")
  del <- grep("lcl", df$protein_id) #we lose some cds, since they do
no have Refseq
  if (length(del) > 0) {
    df <- df[-del,]
  }
  x <-gsub("Mycoplasma ", "M.", nm[i])
  x <-gsub(" ", "_", x)
  x <-gsub(".fna", "", x)
  m <-paste(paste("Proteomes/taulescds/",x,sep=""),".csv",sep="")
  write.csv(df, m, row.names = FALSE)
}
```

Merge all the tables into one for each strain and obtain those that are in the membrane

```
#Call datasets with protein sequence
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/Amb Fasta")
nm1 <- list.files(pattern=".csv")
for (i in 1:length(nm1)) assign(nm1[i], read.csv(nm1[i]))
```

```
#Call files from TOPCONS
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/TOPCONS")
nm2 <- list.files(pattern="*.txt")
for (i in 1:length(nm2)) assign(nm2[i], read.delim(nm2[i], header =
FALSE))

#Call cds tables
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/taulescds")
nm3 <- list.files(pattern="*.csv")
for (i in 1:length(nm3)) assign(nm3[i], read.csv(nm3[i]))

#Merge and delete unnecesary columns
M.pneumoniae_M129.txt.csv <- select(M.pneumoniae_M129.txt.csv, -
COG.s.)
for(i in 1:length(nm1)){
  Mdata<-get(nm1[i])
  Mdata<-Mdata[order(Mdata$Protein.product), ]
  Mdata<-Mdata[ ,-c(1,6,7)]
  Soca<-Mdata$Fasta
  x <-gsub(".txt.csv", "", nm1[i])
  n <-paste(paste(">",x,sep=""),"_",sep="")
  Soca<-gsub(">", n, Soca)
  Mdata<-cbind(Mdata,Soca)
  TOPCONS<-get(nm2[i])
  TOPCONS<-TOPCONS[order(TOPCONS$V7), ]
  TOPCONS<-TOPCONS[ ,-c(1,2,5,6,8)]
  names(TOPCONS) = c("numTM","SignalPeptide","Identifier")
  cds<-get(nm3[i])
  cds<-cds[order(cds$protein_id), ]
  names(cds) = c("Protein.product","cds")
  TOPCONS <-cbind(Mdata,TOPCONS)
  TOPCONS <-merge(TOPCONS, cds, by="Protein.product", all.x = TRUE)
  TOPCONS$Identifier <-
paste(paste(">",x,sep=""),TOPCONS$Identifier,sep="_")
  TOPCONS$Identifier <-gsub(" ", "_", TOPCONS$Identifier)
  cols<-  c("Identifier", "cds")
  TOPCONS$cds_fasta <- do.call(paste, c(TOPCONS[cols], sep="\n"))
  TOPCONS <-subset(TOPCONS, TOPCONS$numTM >0)
  m <-paste(paste("Proteomes/SoquesMem/",x,sep=""),".csv",sep="")
  write.csv(TOPCONS, m, row.names = FALSE)
}
```

Obtain protein sequence and cds files in fasta format of the membrane proteomes and create a dataset of all the membrane proteins

```
#Call the membrane proteome datasets
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Proteomes/SoquesMem")
nm <- list.files(pattern=".csv")
for (i in 1:length(nm)) assign(nm[i], read.csv(nm[i]))

#Plot of the protein number in each dataset
p<-c()
```

```
for(i in 1:length(nm)){
  L<-get(nm[i])
  x<-nrow(L)
  p<-c(p,x)
}
barplot(p, xlim=c(0, 60), ylim=c(0, 150), col=c(rep("orange", 5),
rep("blue", 49)))
title(main = "Membrane proteins per strain", ylab="Proteins")
legend("bottomright", legend = c("M.genitalium", "M.pneumoniae"), fill
= c("orange","blue"))

#Create the cds and aa sequence files
for(x in 1:length(nm)){
  proteoma<-get(nm[x])
  x <-gsub(".csv", "", (nm[x]))
  proteoma$Soca <-gsub(" ", "_", proteoma$Soca)
  m <-paste(paste("Proteomes/TMFasta/",x,sep=""),".fasta",sep="")
  write.table(proteoma$Soca, m, row.names = FALSE, col.names = FALSE,
quote = FALSE)
  m <-paste(paste("Proteomes/fastacds/",x,sep=""),".fasta",sep="")
  write.table(proteoma$cds_fasta, m, row.names = FALSE, col.names =
FALSE, quote = FALSE)
}

#Merge all the strains into one dataset
Mptot <-c()
for(i in 1:length(nm)){
  proteoma<-get(nm[i])
  Mptot<-rbind(Mptot, proteoma)
}
write.csv(Mptot, "Proteomes/Fullmem/Mptot.csv", row.names = FALSE)
```

Delete orthogroups that do not contain Mycoplasma genitalium proteins

```
Orthogroups<-
read_tsv("Orthology/Results_May18/Orthogroups/Orthogroups.tsv")
Orthogroups_M.genitalium<-
Orthogroups[!with(Orthogroups,is.na(M.genitalium_G37)&
is.na(M.genitalium_M2288)& is.na(M.genitalium_M2321)&
is.na(M.genitalium_M6282)& is.na(M.genitalium_M6320)),]

write.csv(Orthogroups_M.genitalium,
"Orthology/Mycoplasma_genitalium_Orthogroups.csv", row.names = FALSE)
```

Orthogroup conversion from aa to cds

```
#Call the orthogroups
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Orthology/prank/alignment")
nm <- list.files(pattern=".fa.best.fas")
for (i in 1:length(nm)) assign(nm[i], read.fasta(nm[i],as.string =
TRUE, set.attributes = TRUE, forceDNAtolower = FALSE))

#Plot of the protein number in each orthogroup
p<-c()
for(i in 1:length(nm)){
```

```
  L<-get(nm[i])
  x<-length(L)
  p<-c(p,x)

}
barplot(p, ylim=c(0, 110), col=c(rep("cadetblue3", 88), rep("orange",
9)))
title(main = "Proteins per orthogroups", ylab="Proteins")

#This code is a bit convoluted since it seems that several
orthogroups, in specifics OG0, OG2, OG8 and OG12, need modifications
to output the correct file. The code needs to run once and then apply
the specific modification for each of the 4 groups. The bucle
Mptot<-read.csv("Proteomes/Fullmem/Mptot.csv")
#Mptot$Identifier <-gsub("-", "_", Mptot$Identifier) #for OG0
#Mptot$Identifier <-gsub("\\(", "_", Mptot$Identifier) #for OG12
#Mptot$Identifier <-gsub("\\)", "_", Mptot$Identifier) #for OG12
#i=1 for OG0, 3 for OG2, 9 for OG8 or 13 for OG12
for(i in 1:length(nm)){ #this part needs to be ignored when running
the code after the first time
  L<-get(nm[i])
  n <- length(L[[1]])
  df <- structure(L, row.names = c(NA, -n), class = "data.frame")
  df <- as.data.frame(t(df))
  df <- cbind(rownames(df), data.frame(df, row.names=NULL))
  colnames(df)<-c("protein_id","sequence")
  f<-df$protein_id
  f<-paste(">", df$protein_id, sep = "")
  u<-c()
  for(y in 1:length(f)){
    o<-Mptot[Mptot$Identifier %in% f[y], ]
    #o<-o[!duplicated(o),] #for OG2 i OG8
    u<-rbind(u,o)
  }
  u$cds_fasta <-gsub("-", "_", u$cds_fasta)
  x <-gsub(".fa.best.fas", "", nm[i])
  m <-
paste(paste("Orthology/Orthogrups/Orthogroup_cds/",x,sep=""),".fasta",
sep="")
  write.table(u$cds_fasta, m, row.names = FALSE, col.names = FALSE,
quote = FALSE)
  m <-
paste(paste("Orthology/Orthogrups/Orthogroups_Taules/",x,sep=""),".csv
",sep="")
  write.csv(u, m, row.names = FALSE)
} #this part needs to be ignored when running the code after the first
time
```

Plots of the gaps and indentity per alignment calculated by trimAl

```
p<-
c(3,29,2,2,0,5,2,2,2,2,1,1,1,2,1,2,2,0,3,1,1,2,0,1,1,2,1,2,0,2,0,1,2,1
,0,1,1,1,2,0,1,2,2,2,1,0,2,1,1,2,1,2,2,2,1,1,2,2,3,1,2,0,0,1,1,1,2,1,1
,2,0,0,1,2,1,2,2,1,4,1,1,2,1,1,11,2,1,2,0,2,1,0,0,0,2,0,0)
barplot(p, ylim=c(0, 30), col=c(rep("cadetblue3", 88), rep("orange",
```

```
9)))
title(main = "Gaps per orthogroups", ylab="Gaps")

p<-
c(0.7692,0.5687,0.9407,0.9461,0.9808,0.9716,0.9181,0.9231,0.9593,0.958
0,0.9501,0.9550,0.9477,0.9534,0.9465,0.9307,0.9417,0.9341,0.9126,0.961
1,0.9410,0.9220,0.9521,0.9345,0.9542,0.9054,0.9376,0.9561,0.9310,0.958
8,0.9727,0.9109,0.9454,0.9695,0.9592,0.9499,0.9301,0.8993,0.9258,0.946
5,0.9129,0.9534,0.9182,0.9462,0.9259,0.9387,0.9474,0.8958,0.8825,0.905
6,0.9625,0.9234,0.9179,0.9284,0.9584,0.8956,0.9464,0.9531,0.8827,0.978
7,0.9297,0.9415,0.9558,0.9459,0.9663,0.9561,0.9410,0.9751,0.9487,0.954
9,0.9243,0.9543,0.9370,0.9297,0.9728,0.9615,0.9580,0.9734,0.9128,0.975
7,0.9501,0.9635,0.9593,0.9740,0.8707,0.9389,0.9823,0.8894,0.9122,0.561
1,0.8609,1.0000,0.9910,0.9973,0.2283,1.0000,0.9862)
barplot(p, ylim=c(0, 1), col=c(rep("cadetblue3", 88), rep("orange",
9)))
title(main = "Average Identity per orthogroups", ylab="Identity")
```

Plot of the protein number inputted in ete3-evol

```
setwd("C:/Users/guill/Desktop/Estudis/UOC/TFM i
Practiques/TFM/Orthology/prank/alignment")
nm <-
c("OG0000003.fa.best.fas","OG0000004.fa.best.fas","OG0000005.fa.best.f
as","OG0000009.fa.best.fas","OG0000010.fa.best.fas","OG0000011.fa.best
.fas","OG0000012.fa.best.fas","OG0000013.fa.best.fas","OG0000014.fa.be
st.fas","OG0000019.fa.best.fas","OG0000022.fa.best.fas","OG0000027.fa.
best.fas","OG0000029.fa.best.fas","OG0000030.fa.best.fas","OG0000033.f
a.best.fas","OG0000034.fa.best.fas","OG0000035.fa.best.fas","OG0000039
.fa.best.fas","OG0000041.fa.best.fas","OG0000043.fa.best.fas","OG00000
50.fa.best.fas","OG0000062.fa.best.fas","OG0000066.fa.best.fas","OG000
0067.fa.best.fas","OG0000115.fa.best.fas")
for (i in 1:length(nm)) assign(nm[i], read.fasta(nm[i],as.string =
TRUE, set.attributes = TRUE, forceDNAtolower = FALSE))

p<-c()
for(i in 1:length(nm)){
  L<-get(nm[i])
  x<-length(L)
  p<-c(p,x)

}
barplot(p, ylim=c(0, 60), col="cadetblue3")
title(main = "Proteins per orthogroups", ylab="Proteins")
```

Create table with the information from HPIDB 3.0 about M.pneumoniae interactions with Homo sapiens

```
ly8gew.txt<- read_delim("Orthology/HPIDB3.0/ly8gew.txt", "\t",
escape_double = FALSE, trim_ws = TRUE)
ly8gew <-
read_delim("Orthology/HPIDB3.0/e85u3y1559317147_s_list.mitab_plus.txt"
, "\t", escape_double = FALSE, trim_ws = TRUE)
ly8gew <-ly8gew[ ,c(26)]#delete unnecessary columns
ly8gew.txt <-cbind(ly8gew.txt,ly8gew)
```

```
ly8gew.txt<-subset(ly8gew.txt, !duplicated(ly8gew.txt$protein_xref_2))
ly8gew.txt <-ly8gew.txt[ ,c(2,16)]#delete unnecessary columns

#Add M.genitalium proteins obtained from BLASTp
protein_xref_3 <-
c("uniprotkb:P13927","uniprotkb:P47514","uniprotkb:P47543","uniprotkb:
P47515","uniprotkb:P47458","uniprotkb:P47516","uniprotkb:P47572","unip
rotkb:P47513","uniprotkb:P47698","uniprotkb:P47647","uniprotkb:P47312"
,"uniprotkb:P47669")
protein_3_cover<- c(100,100,100,100,100,100,4,99,100,97,100,99)
protein_3_identity<-
c(96.70,71.43,82.20,88.99,78.15,88.27,39.29,72.81,82.69,79.73,71.76,71
.43)

ly8gew.txt <-
cbind(ly8gew.txt,protein_xref_3,protein_3_cover,protein_3_identity)
colnames(ly8gew.txt) <-c("M.pneumoniae_protein",
"M.pneumoniae_display_id", "M.genitalium_protein", "Query cover",
"Identity")

write.csv(ly8gew.txt,
"Orthology/HPIDB3.0/Mycoplasma_genitalium_HPIDB3.0.csv", row.names =
FALSE)
```

Create table with the information from HPIDB 3.0 obtained from internal BLAST with the membrane proteome of M.genitalium G37

```
pr_unq <- read_tsv("Orthology/HPIDB3.0/4tju7p1559309256_pr_res.tsv")
pr_unq<-subset(pr_unq, !duplicated(pr_unq$HPIDB_Host_ID))
pr_unq<-subset(pr_unq, pr_unq$HPIDB_Host_Taxon_Category=="ANIMAL" &
!pr_unq$HPIDB_Host_Taxon=="bovin|Bos taurus (Bovine)") #only those
that interact with humans
pr_unq <-pr_unq[ ,-c(6,8,11,12,13,14,16)]#delete unnecessary columns

f<-pr_unq$Pathogen_Input_ID
pr_unq$Pathogen_Input_ID<-gsub("^([^_]*_[^_]*_[^_]*_[^_]*)_.*$",
"\\1", f)

write.csv(pr_unq,
"Orthology/HPIDB3.0/Mycoplasma_genitalium_membrane_HPIDB3.0.csv",
row.names = FALSE)
```

7.3 Python scripts

PRANK

```
from os import scandir, getcwd, system

def ls(ruta = getcwd()):
    return [arch.name for arch in scandir(ruta) if arch.is_file()]

def prog_prank(input, output):

#PRANK
    print ('I am aligning with this input',input, 'and this output',output)
```

```
    cmd = 'prank -F -d='+input+' -o='+output+' -showanc -showxml'
    system(cmd)

file_list=ls("Orthogroup_Sequences")
for i in range(len(file_list)):
    inp="Orthogroup_Sequences/"+file_list[i]+""
    outp="prank/"+file_list[i]+""
    prog_prank(inp,outp)
    print (i)
```

trimAl

```
from os import scandir, getcwd, system

def ls(ruta = getcwd()):
    return [arch.name for arch in scandir(ruta) if arch.is_file()]
def prog_trim(input, output):
#pal2nal
    print ('I am analizing with this input',input1, 'and this output',output)
    cmd = 'trimal -in '+input+' -sgt -sident>'+output+''
    system(cmd)

file_list=ls("Orthogroup_Sequences")
for i in range(len(file_list)):
    inp1="prank/"+file_list[i]+".best.fas"
    outp="trimAl/"+file_list[i]+".fasta"
    prog_trim(inp,outp)
    print (i+1)
```

pal2nal

```
from os import scandir, getcwd, system

def ls(ruta = getcwd()):
    return [arch.name for arch in scandir(ruta) if arch.is_file()]
def prog_pal2nal(input1, input2, output):
#pal2nal
    print ('I am analizing with this input',input1, 'and this output',output)
    cmd = 'pal2nal.pl '+input1+' '+input2+' -codontable 4 -output fasta > '+output+''
    system(cmd)
```

```
file_list=ls("Orthogroup_Sequences")
for i in range(len(file_list)):
    inp1="prank/"+file_list[i]+".best.fas"
    inp2="Orthogrups/Orthogroup_cds/"+file_list[i]+"sta"
    outp="pal2nal/"+file_list[i]+".fasta"
    prog_pal2nal(inp1,inp2,outp)
    print (i+1)
```

Ete3-evol

```
#!/usr/bin/env python3

from os import system
def prog_ete3(input1, input2, output1, output2):
#ete3
    print ('I am analizing with this input',input1, 'and this output',output1)
    cmd = 'ete3 evol -t '+input2+' --alg '+input1+' -o '+output1+' --internals --leaves --cpu
2 --models fb -i '+output2+' --resume --clean_layout &'
    print (cmd)
    print (i+1)
    system(cmd)

file = open("Orthogroups.txt", "r")
text = file.read()
file_list = text.split("\n")
for i in range(len(file_list)):
    inp1="pal2nal/"+file_list[i]+".fa.fasta"
    inp2="tree/"+file_list[i]+".nwk"
    outp1="ete3/"+file_list[i]+"/"
    outp2= ""+file_list[i]+"_tree.pdf"
    prog_ete3(inp1,inp2,outp1,outp2)
```