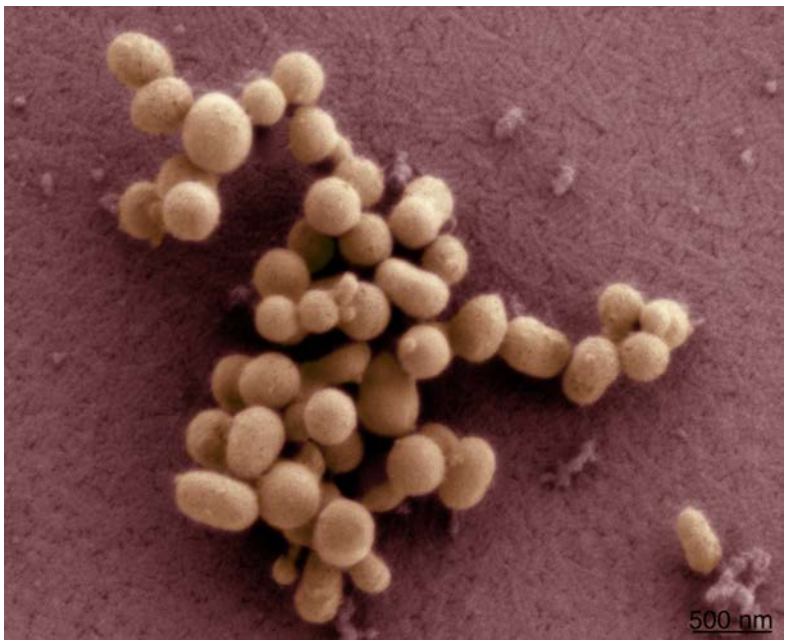
Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Comparative Analysis of the "Mycoplasma mycoides cluster"

Hadrien Gourlé



Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics
482
Uppsala 2015

Master's Thesis, 30 hp

Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Comparative Analysis of the "Mycoplasma mycoides cluster"

Hadrien Gourlé

Supervisors:
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics

Examiner:
Tomas Bergström, SLU, Department of Animal Breeding and Genetics

Credits: 30 hp
Course title: Degree project in Biology
Course code: EX0578
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2015
Cover picture: Scanning electron micrograph of M. mycoides JCVI-syn1.
Image courtesy of Thomas Deerinck and Mark Ellisman, NCMIR, and John Glass, JCVI
Name of series: Examensarbete / Swedish University of Agricultural Sciences,
Department of Animal Breeding and Genetics, 482
On-line publicering: http://epsilon.slu.se

Key words: '*Mycoplasma*', Comparative analysis,Pathogenicity, Bioinformatics, Core genome

I wish to acknowledge Dr. Erik Bongcam-Rudloff and Dr. Joerg Jores for entrusting me with this project and for their advice and expertise.

I would also like to thank Dr. Elise Schieck and Dr. Juliette hayer for their help in writing and reviewing this thesis, as well as the INRA and JCVI teams for their help in the genome sequencing and assembly.

Finally, I'd like to thank the SGBC team as a whole for their encouragement and support during the entire duration of my thesis.

Abbreviations

Mmm: Mycoplasma mycoides subsp. *mycoides*

Mmc: Mycoplasma mycoides subsp. *capri*

Mcc: Mycoplasma capricolum subsp. *capricolum*

Mccp: Mycoplasma capricolum subsp. *capripneumoniae*

CBPP: Contagious Bovine Pleuropneumonia

CCPP: Contagious Caprine Pleuropneumonia

OIE: World Organization of Animal Health (formerly Office International des Epizooties)

MLST: Multi Locus Sequence Tag

COG: Clusters of Orthologous Groups

Introduction:

The so-called "*Mycoplasma mycoides* cluster" comprises a group of bacteria that is quite unusual phylogenetically speaking within the class *Mollicutes*. It contains five closely related pathogens that are all infecting ruminants.

These five taxa, *Mycoplasma mycoides* subsp. *mycoides* (*Mmm*), *Mycoplasma mycoides* subsp. *capri* (*Mmc*), *Mycoplasma capricolum* subsp. *capricolum (Mcc)*, *Mycoplasma capricolum* subsp. *capripneumoniae* (*Mccp*) and *Mycoplasma leachii* are characterized by general Mycoplasma features such as small size (about 0,1 μm in length or diameter), their lack of a cell wall and therefore their lack of a definite shape, and a small genome size of about one Mbp, which make them one of the smallest self-replicating bacterial organisms. They probably have evolved from their ancestors, the *Firmicutes*, gram-positive bacteria, by deletions of genes. Their low GC content (24%) and their relatively high amount of insertion sequences are also worth mentioning.

Two members of the "*Mycoplasma mycoides* cluster" are considered of the utter importance: *Mycoplasma mycoides* subsp. *mycoides* (*Mmm*) and *Mycoplasma capricolum* subsp. *capripneumoniae* (*Mccp*) causative agents of the Contagious Bovine Pleuropneumonia (CBPP) and the Contagious Caprine Pleuropneumonia (CCPP), respectively.

CBPP:

Contagious Bovine Pleuropneumonia (CBPP) is a cattle disease, notifiable to the World Organization of Animal health (formerly Office International des Epizooties, OIE) and is caused by *Mycoplasma mycoides* subsp. *mycoides* [1].

CBPP can be present as acute or chronic disease. After an incubation period of up to six weeks, acutely affected animals develop symptoms such as fever, depression and respiratory distress. The mortality rate of CBPP can be as high as 60% for the most virulent strains when introduced into naïve herds. Once the first symptoms are noticeable the animal either dies of pleuropneumonia, or the symptoms gradually disappear after several weeks. Clinically recovered cattle may transit into a chronic phase of the disease. In that case clinical signs are emaciation and coughing, and the lungs may contains lesions, called *sequestra*, from where live bacteria have been isolated. These chronically effected animals may be infectious and may play a role in the epidemiology of the disease [2].

Since the eradication of Rinderpest, CBPP is the most important cattle disease in Africa. It is widespread in sub-Saharan Africa and has been suspected in certain parts of Asia. CBPP threatens livestock production, limits trade exchange and is therefore of huge economic concern in affected countries.

CBPP was clearly described for the first time by B de Haller in 1773 but it may have been documented as soon as in the 17th century [3]. It is believed that CBPP was exported from Europe through cattle trade [4]. CBPP reached a worldwide distribution during the second half of the 19th century. It has been eradicated from most continents by strict stamping-out policies: from Australia in the 1970's and in Europe in the beginning of the 20th century. A last epidemiologically unexplained outbreak occurred in Portugal, Spain France and Italy in the 1980 and 1990 but was contained and eradicated in 1993 [5][6].

Figure 1. Cow infected by CBPP (picture: Joerg Jores)

The OIE advices the use of vaccination for control of the disease but eradication works only on slaughter and control of movements. The vaccines that are now used are live vaccines based on the strain T1/44. This vaccine strain, isolated in 1951 has been attenuated by 44 passages in embryonated eggs [2]. The vaccine, although attenuated has shown to rarely trigger severe post-vaccinal reactions and is known to be still virulent. The vaccine also provides immunity for a rather short timespan and requires annual revaccination. Antibiotic treatment is not recommended since it may produce resistant strains and suppress the development of clinical signs, postponing the recognition of the disease [7].

Vaccination and antibiotic treatments are however used in the control of the disease in Africa, since movement control is difficult to achieve, and slaughter campaigns require considerable resources to compensate and restock the owners. Annual and well-planned campaigns of vaccination are successful in reducing CBPP outbreaks but eradication remains impossible without other policies [8].

CCPP:

CCPP, or Contagious Caprine Pleuropneumonia is a disease that affects goats. First described in Algeria in 1873, the disease is caused by *Mycoplasma capricolum* subsp. *capripneumoniae* [9].

The first symptoms are reluctance to walk, followed by extreme fever (around 41°C). Respiratory symptoms become gradually worse, with violent coughing and lesions concentrated in the thoracic cavity. Dead usually comes within a few days but the animal may survive for up to a month, or even recover. The mortality rate varies from 60% to 100%. A chronic form of the

disease is also present where CCPP is endemic, presenting a milder version of the symptoms [10].

CCPP is also notifiable to the OIE and is responsible of huge economic losses for goat producers in Africa, the Middle East and Western Asia.

Other members of the Cluster:

Two other members of the cluster are also caprine pathogens: *Mycoplasma mycoides* subsp. *capri* and *Mycoplasma* subsp. *capricolum*. They are both known to cause various forms of clinical disease such as mastitis, pneumonia, septicemia and arthritis. *M. Leachii*, the last and more recently classified member of the cluster, is a bovine pathogen causing polyarthritis, mastitis and abortion [11].

Phylogeny:

Lineage: *Bacteria, Tenericutes, Mollicutes, Mycoplasmataceae, Mycoplasma, "Mycoplasma mycoides cluster".*

The "*Mycoplasma mycoides* cluster" is an extremely monomorphic group of closely related taxa of the genus *Mycoplasma,* the class *Mollicutes* (phylum *Tenericutes*) [12].

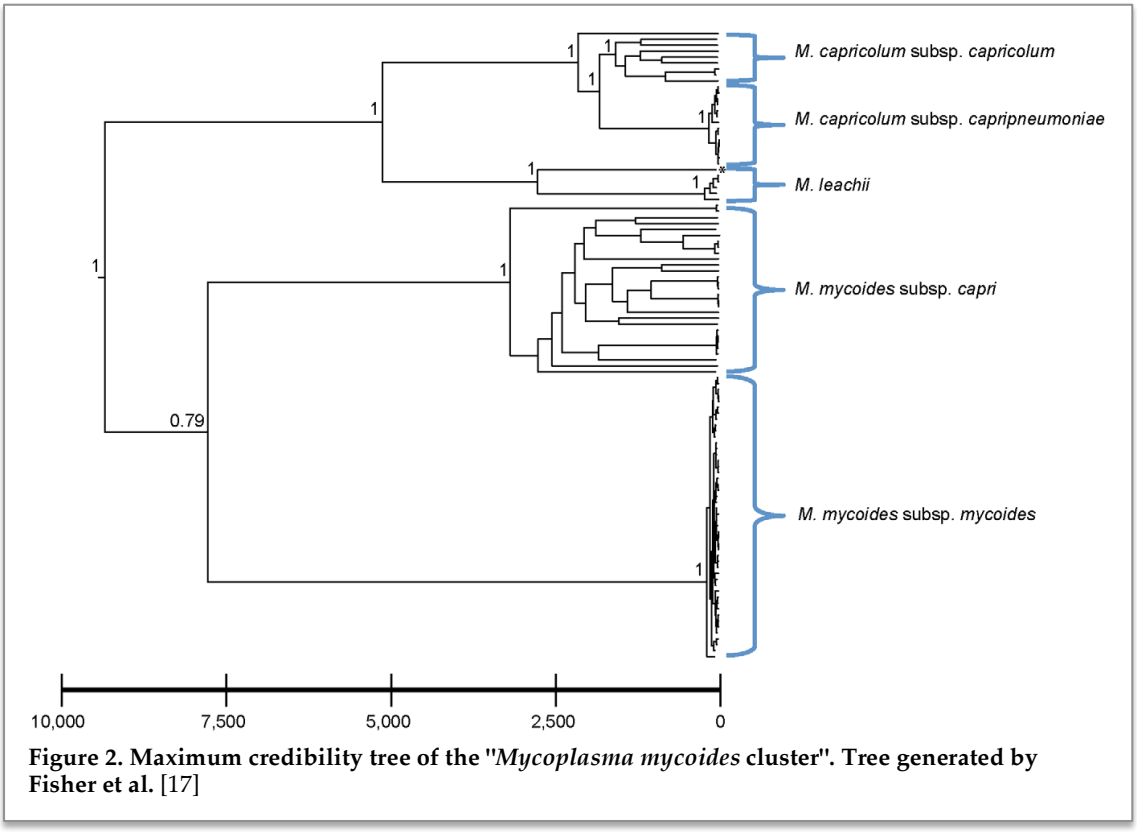


Figure 2. Maximum credibility tree of the "*Mycoplasma mycoides* cluster". Tree generated by Fisher et al. [17]

The genus contains around 120 species, which are all obligate parasites. They are found in a wide spectrum of hosts (human, animals and plants) [13]. Within the genus, the "*Mycoplasma mycoides* cluster" is a tight phylogenetic clade of ruminant pathogens, varying in disease and severity. The phylogeny of the cluster has been difficult to establish [14][15][16], the organisms being too close to efficiently differentiate their rRNA. MLST (Multiple Locus Sequence Tag) has been used to resolve the phylogeny of the cluster in 2012 [17] (Figure 2).

It has been found that the origin of the cluster could be traced to the beginning of the domestication of ruminants, 10,000 years ago. It has been established that *Mmm*, and therefore CBPP, has emerged around 1700 [3]. It coincides with the first description of the disease in 1773. *Mmm* probably adapted to a new host from small ruminants [4]. Another study aiming at establishing the evolutionary history of *Mycoplasma mycoides* subsp. *mycoides* effectively retraced the spread of CBPP from Europe in the 19th century, through cattle trade routes.

Metabolism and Pathogenicity:

The physiology and the pathogenicity with its host-pathogen interactions of the members of the "*Mycoplasma mycoides* cluster", is not well understood. Hypotheses have been made but few have been verified experimentally [18].

No known virulence factors such as toxins and adhesions have been described and *Mycoplasma* is believed to rely on components of the outer cell surface [19] and intrinsic metabolic functions.

First, membranes proteins and lipoproteins show phase variation, by mutations in poly(TA) tract-containing promoters, leading to surface diversification, hence theoretically allowing the *Mycoplasmas* to escape host immune response and more generally to modulate its interaction with the host [20].

H2O2 produced by glycerol metabolism has also been proposed as a virulence factor. It cannot however be considered as the sole factor, since vaccine strains such as T1/44 have shown to release important amount of H2O2 as well [21].

Finally, polysaccharides have been recently proposed as key virulence factors. *Mycoplasma*s from the "*Mycoplasma mycoides* cluster" are known to produce two polysaccharides: a capsular polysaccharide (CPS), galactan, and an exopolysaccharide (EPS), that has been shown to circulate in the blood stream of the host [19][22].

Objectives:

Our hypothesis is that all *Mycoplasma mycoides* share a core set of genes for general anabolic and catabolic pathways. The pan-genome of the cluster is likely to include genes that code for virulence traits and host-specificity.

The objective of this thesis is to identify candidate molecules that are involved in pathogenicity and host tropism in *Mycoplasmas* of the "*M. myoides* cluster". The output of this work will present global public goods that will inform the research community and foster to a better understanding of *Mycoplasma* genomes.

Materials and Methods:

Overview:

31 genomes were used in that study: 13 strains of *Mmm*, 2 of *M. leachii*, 4 of *M. capricolum* subsp. *capricolum (Mcc)*, 6 of *M. capricolum subsp. capripneuomiae (Mccp)* and 6 of *Mmc* (Table 1). The first objective was to identify the core and pan genome of the following set of species or subspecies:

1. The entire "*M. mycoides* cluster"
2. Bovine Pathogens of the "*M. mycoides* cluster"
3. Caprine Pathogens of the "*M. mycoides* cluster"
4. *Mmm*
5. *Mmm* + *Mmc* (*M. mycoides*)
6. *M. capricolum*
7. *M. capricolum* subsp. *capripneumoniae*

Table 1. List of *Mycoplasma* strains studied

Species	Strain Designation	4 letters code used	Genome accession no.	Country
M. mycoides subsp. *mycoides*	Gladysdale	GLAD	NC_021025	Australia
	5713 (IZSAM)	5713	NZ_CP010267.1	Italy
	95014*	9501		Tanzania
	Afadé	AFAD	NZ_LAEX00000000.1	Cameroon
	B237	B237	NZ_LAEW01000001.1	Kenya
	B66*	B66		Kenya
	C11*	C11		Chad
	Fatick*	FATI		Senegal
	L2*	L2		Italy
	Matapi*	MATA		Namibia
	PG1	PG1	NC_005364.2	
	T1/44*	T144		Tanzania
	V5*	V5		Australia
M. mycoides subsp. *capri*	Capri L*	Capr		France
	G1313*	G131		Germany
	GM12	GM12	NZ_CP001668.1	USA
	LC95010	LC95	NC_015431.1	France
	PG3	PG3	NZ_ANIV00000000.1	Turkey
	Y-goat*	YGOA		Australia
M. capricolum subsp. *capripneumoniae*	87001	8700	NZ_CP006959.1	China
	99108	9910	NZ_JMJI00000000.1	Eritrea
	Abomsa	ABOM	NZ_LM995445.1	Ethiopia
	F38	F38	NZ_LN515398.1	Kenya
	ILRI181	ILRI	NZ_LN515399.1	Kenya
	M1601	M160	NZ_CM001150	China
M. capricolum subsp. *capricolum*	14232	1423	NZ_JFDO00000000.1	France
	14DL	14DL	NZ_LBMF00000000.1	Germany
	California Kid	ATCC	NC_007633.1	USA
	GM508D	GM50	NZ_JXQB00000000.1	USA
M. Leachii	99014	9901	NC_017521.1	Australia
	PG50	PG50	NC_014751.1	Australia

Sampling:

Out of the 31 genomes used for the study, 20 were publicly available and 11 were sequenced by project partners (indicated with a * in Table 1). Briefly, liquid cultures of Mycoplasma (in PPLO medium) were filtered and plated on PPLO agar in different dilutions. After 3 to 4 days of incubation at 37°C, a single colony was picked and used to inoculate 4ml of PPLO medium, which was stored at -80°C.

Filter cloned Mycoplasma were grown overnight in 100 ml PPLO medium at 37°C. Before entering the stationary growth phase the culture was centrifuged at 2.862 g for 1h, and the pellet was resuspended in 2.5 ml of TNE buffer. Samples were treated with 50/50 SDS/Protein kinase-K for 2h at 37°C. 100 mM PMSF were added to the samples and they were incubated for 15min in room temperature followed by addition of RNase A and additional 1 hour incubation. Sodium acetate/phenol saturated buffered were added and samples centrifuged at ~16,000xg after mixing. Top phase were removed and subjected to Phenol:Chloroform:Isoamyl extraction and isopropanol precipitation.

Sequencing and Assembly:

The genomes were sent for sequencing to INRA, France. The genomes were sequenced using Illumina HiSeq with two Mate Pairs libraries of 2*200bp and one Paired End library of 2*100bp. All samples were found to have long, high identity matches to *M. mycoides* with no evidence of *E. coli* or phage contamination. Between 93 and 96% of the reads were found unique before kmer normalization.

GC peaks were found in the FastQC [23] analysis and were confirmed by high prevalence of matches to TruSeq and illumina adapters sequences. The adapters were removed using CLC [24].

Different read correction methods, verified by a quick assembly against Gladysdale, were tried on the T1/44 strain. The best results were achieved by using kmer normalization and exact de-duplication followed by trimming the reads by quality.

Different assembly methods were evaluated, still using the strain T1/44 as a test case: Overlap-layout-consensus (Newbler [25], Celera [26]), de Bruijn Graphs (Velvet [27], SOAP [28], Allpaths [29]) and simulated multi-De-Bruijn (SPAdes [30], IDBA [31], Velvet-SC [32]). The coverage was reduced to 40x and 60x using targeted bin selection (NeatFreq [33]) for the OLC (overlap-layout-consensus) methods (Figure 3). SPAdes and Newbler showed the best results (better N50 and better mapping against Gladysdale) and were chosen for the assembly of the 10 remaining strains.

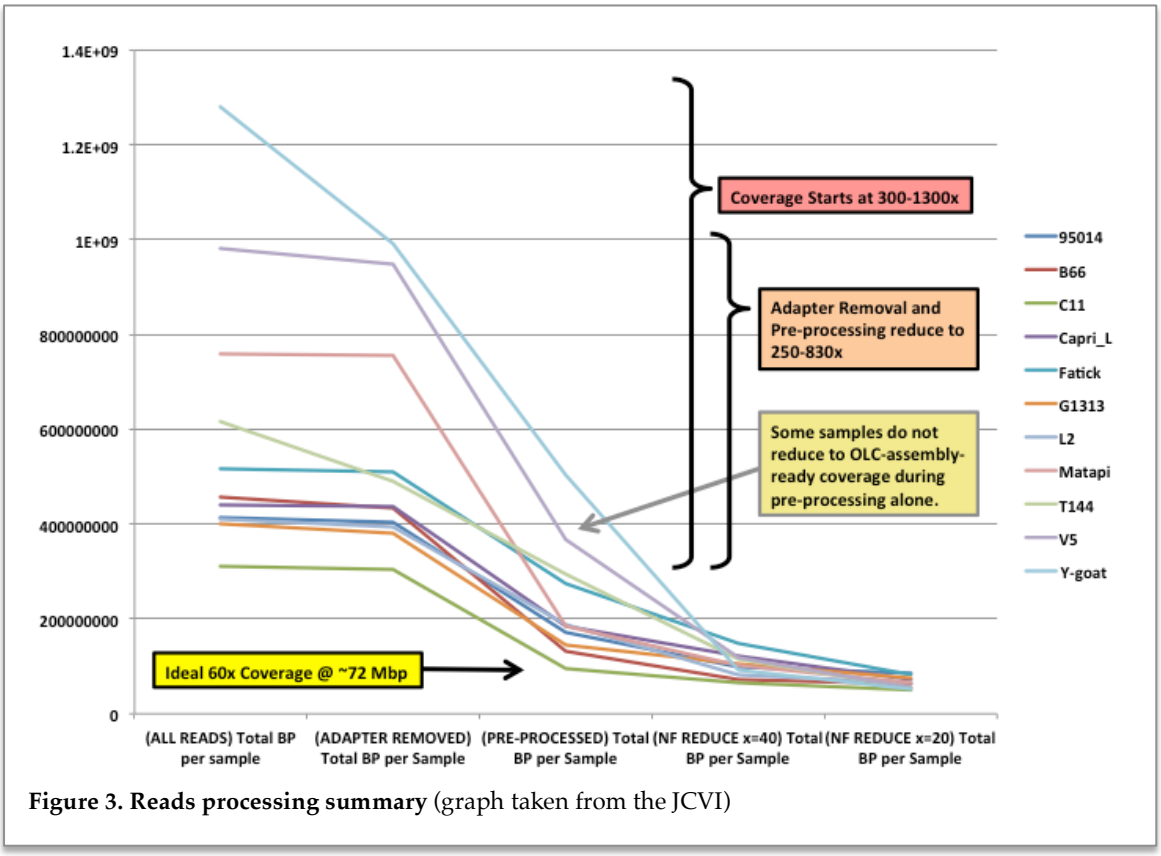


Figure 3. Reads processing summary (graph taken from the JCVI)

Annotation:

The best assemblies were selected and the genome sequences were added to the pool of 20 genomes already available. The 31 genomes were then annotated or re-annotated using Prokka v1.10 [34].

Prokka uses Aragorn [35] to find tRNAs, prodigual [36] was used for CDS predictions. Prodigual simply uses a log-likelihood function [37] of signal to background to predict CDS across the genome. Un-annotated CDS are then compared to custom databases (RefSeq [38] Mycoplasma, Bacteria) using Blastp [39]. Remaining un-annotated CDS were searched against Pfam [40] using HMMER3 [41].

Core and Pan genome characterization:

The annotated genomes were divided into the 7 datasets previously mentioned, a few out of these were overlapping. OrthoMCL [42] was used for the clustering part of the analysis. OrthoMCL generates clusters of proteins where each cluster consists of orthologs or "recent" paralogs from at least two species.

The procedure starts with an all-against-all Blastp comparison of the set of proteins from the genomes present in the dataset. An e-value cutoff was set to 1e-5.

Next, putative orthologous relationships were converted into a graph, which is represented by a similarity matrix, given to the MCL software [43]. MCL, using a Markov Cluster algorithm, considers all the relationships in the graph globally and simultaneously, separating orthologs mistakenly assigned based on weak reciprocal best hits.

An important parameter in the MCL algorithm is the inflation value, regulating the cluster tightness (granularity). That parameter was set to 1.5. The output of OrthoMCL was divided into core and pan clusters. The division, as well as basic statistics and a summary of the analyses were all obtained using a custom python script.

The division into core and pan clusters was done using the following definition: the core genome of a bacterial group (e.g. members of a subspecies, species or genus) consists of those sequences, which are conserved among members of that species [44]. This strict definition of the core genome was used for the clustering. Therefore for a dataset containing *n* organisms, a core cluster is a cluster containing at least one protein for each of the *n* organisms. A pan cluster will contain proteins for maximum *n-1* organisms. The pan genome is the content of the genomes of a group to be tested minus the core genome.

Functional Characterization:

COG terms (for Clusters of Ortholog Groups [45]) were assigned to each proteins using rpsblast [46] with an e-value cutoff of 1e-5. The blast results were then parsed and the best hit was assigned to each protein using a custom python script. The COG categories and subcategories were plotted for both the pan genomes and the core genomes using R.

Scripting:

All the scripts used, from the clustering to the functional assignment and the plotting, were compiled into a pipeline. The main motivation was the lack of comprehensive software to interpret the output of OrthoMCL. The pipeline, mainly written in bash and python, performed the following steps:

1. Created and Configured a MySQL database for OrthoMCL to use
2. Did run the all-against-all blastp and OrthoMCL
3. Separated the groups produced by OrthoMCL into core and pan genome
4. Retrieved the annotated functions of the proteins present in each cluster. Computed statistics about each cluster as well as general statistics for the genomes present in the analysis.
5. Downloaded and installed the COG database
6. Did run rpsblast against the COG database
7. Assigned the best COG hits to the proteins present in the cluster
8. Produced plots of the COG categories and subcategories for both the core and the pan genome.

The pipeline including all scripts generated is available on Github at https://github.com/HadrienG/OrthoMCLAnalyser

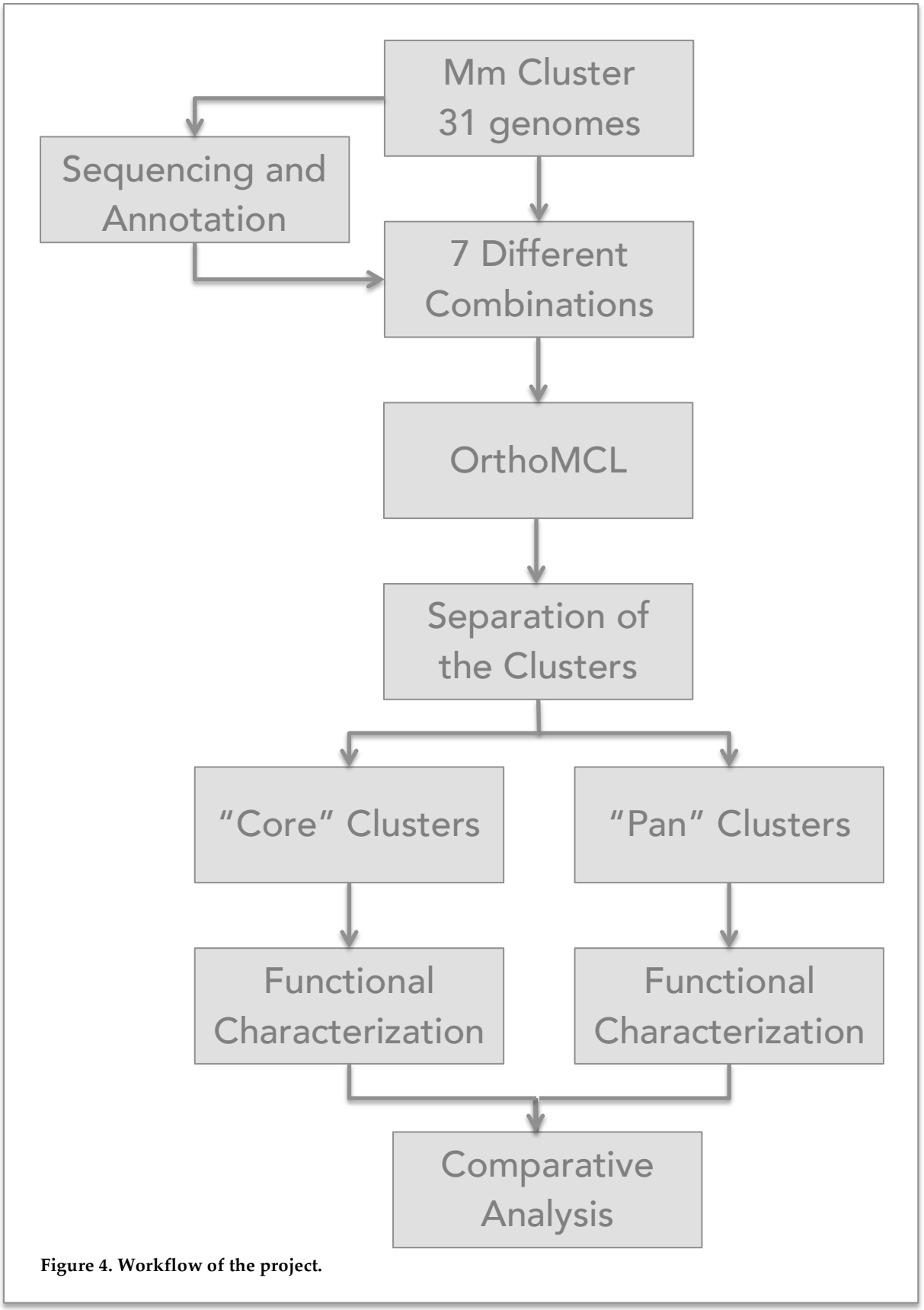


Figure 4. Workflow of the project.

Results

Sequencing, Assembly and Annotation

11 draft assemblies were obtained from J. Graig Venter Institute. 8 out of the 11 were from various strains of *Mycoplasma mycoides* subsp. *mycoides*. The strains were isolated from various African countries excepted for the strains L2 and V5, respectively from Italy and Australia. The 8 assemblies resulted in 117 to 210 contigs (length >200bp) with a N50 from 16,249 to 25,087, for a total length from 984,029bp to 1,070,522bp (mean: 1,035,367bp) and GC content from 23.72% to 24.53% (Table 2).

The three other assemblies were from three strains of *Mycoplasma mycoides* subsp. *capri*, namely YGoat, capriL, and G1313, isolated from Australia, France and Germany, respectively. The Assemblies contained 58 to 283 contigs, for a N50 of 34,917 to 113,501 and a total length from 1,058,262bp to 1,219,757bp. The GC content varied from 23.87% to 24.2% (table 2).

Table 2. Assemblies statistics

Species	Strain name	Assembler	Coverage	Contigs > 200	N50	Sum	GC
M. mycoides subsp.	95014	Spades v2.3.0	66.8	180	23643	1,070,522	24.53
mycoides	B66	Newbler v2.8	49.2	178	17965	988,252	24.02
	C11	Spades v2.3.0	43.8	141	19437	1,047,421	23.72
	V5	Newbler v2.8	29	191	16249	984,029	23.86
	Fatick	Spades v2.3.0	97.4	210	24724	1,068,375	24.73
	T144	Spades v2.3.0	251.5	173	20424	1,050,457	24
	L2	Spades v2.3.0	59.5	117	25087	1,040,980	23.98
	Matapi	Spades v2.3.0	31.6	210	20204	1,032,902	24.3
M. mycoides subsp.	Y-goat	Newbler v2.8	29	159	34917	1,088,983	24.2
capri	Capri L	Newbler v2.8	132.1	283	51667	1,210,757	23.93
	G1313	Spades v2.3.0	50	58	113501	1,058,262	23.87

Those draft assemblies were added to the poll of 20 genomes already available. All 31 genome sequences of this study were annotated using Prokka. The annotation revealed an average of 949 coding sequences (CDS) per genome. The subspecies with the least CDS was *M. capricolum* subsp. *capricolum* with an average of 830 CDS. *M. mycoides* subsp. *mycoides* has most CDS with an average of 1,000 CDS per genome. Between 27 and 40 tRNAs were identified per genome, with an average of 29 (Table 3).

Table 3. Annotations statistics

Species	Strain Designation		CDS	tRNA
M. mycoides subsp. *mycoides*	Gladysdale		1,102	31
	5713		1,101	31
	95014		942	19
	Afade		1,122	31
	B237		1,128	31
	B66		879	27
	C11		954	22
	Fatick		953	28
	L2		926	40
	Matapi		900	26
	PG1		1,153	31
	T144		963	23
	V5		875	23
		Average:	1,000	28
M. mycoides subsp. *capri*	Capri L		961	21
	G1313		875	23
	GM12		880	31
	LC95010		952	31
	PG3		781	24
	Y-Goat		863	31
		Average:	885	27
M. capricolum subsp. *capripneumoniae*	87001		1,008	31
	99108		977	31
	Abomsa		1,002	31
	F38		1,002	31
	ILRI181		1,001	31
	M1601		1,000	31
		Average:	998	31
M. capricolum subsp. *capricolum*	14232		848	30
	14DL		776	31
	ATCC		845	31
	GM508D		850	31
		Average:	830	31
M. Leachii	99014		904	31
	PG50		890	31
		Average:	897	31

Core and Pan genome characterization

The core and pan-genomes were determined for the seven different subsets of the "*Mycoplasma mycoides* cluster" presented in the methods section. Between 992 (for *Mmm*) and 1417 (for the whole cluster) clusters of Proteins were identified (Figure 5). The proportion of pan-genome clusters varied from 3.16% (σ = 0.71) for *Mccp* to 32.84% (σ = 4.32) for the entire M. mycoides cluster (Table 4, Annexes 1-7).

It should be kept in mind that we are not considering the core and pan-genomes as in a definition of "housekeeping" and "accessory" genome but rather as the core-genome being a pool of shared genes between members of a group of

microorganisms and the pan-genome being the pool of genes specific to a fraction of the members.

A clusters belonging to the core-genome contains at least one protein coming from each genome in the dataset. On the contrary, a cluster that belongs to the pan-genome contains maximum *n-1* number of genomes, where *n* is the total number of genomes present in the dataset.

A cluster can also contain several proteins coming from the same genome. Due to the high level of insertion sequences and the above average level of lipoproteins, these two elements often end up in big clusters, containing - per example - all the insertion sequences from a same family.

Table 4. Percentage of core and pan genome

Dataset	Core (in %)	Pan (in %)	σ (St. dev.)
All	67.16	32.84	4.33
Bovine pathogens	81.05	18.95	6.54
Caprine pathogens	74.61	25.39	3.15
M. mycoides	70.9	29.1	5.53
Mmm	86.84	13.36	7.23
M. capricolum	84.3	15.7	2.94
Mccp	96.83	3.16	0.71

Table 4 represents the mean percentage of coding sequences from each genome that is comprised in the core and the pan genome, for the seven subset of the "*M. mycoides* cluster". The right column is the standard deviation, also in %.

Figure 5 also represents the proportion of core and pan genome, but in total number of clusters.

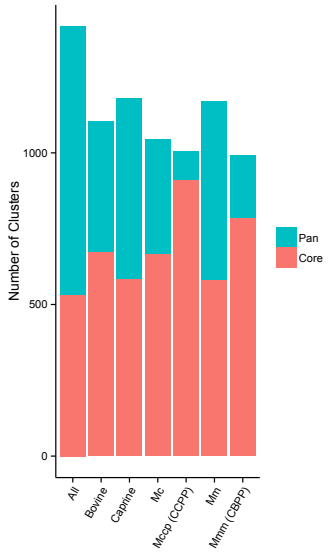


Figure 5. Proportion of core and pan genome in the 7 datasets

Functional characterization

All the proteins were functionally characterized using NCBI database of Clusters of Orthologous Groups of proteins (COGs). The database currently contains more than 5,000 COGs. While each COG has a specific functional description, it may also have one or more general category letter associations. We grouped subcategories into four categories: (a) cellular processes, (b) signaling, (c) information storage and processing, and metabolism (Table 5). Also, the subcategory "Mobilome: prophages and transposons" has not been assigned to any of the four categories

It can be noticed that many proteins, especially in the pan-genome, appear not to have matched with any COG and are therefore labeled as "not in COG database". For clarity, they have been removed from the graphs displaying the general COG categories. The category "poorly characterized" contains only the two-subcategories "General public prediction only" and "function unknown".

Table 5. List of COG categories and subcategories

COG category	COG subcategory	Code
Information storage and processing	Translation, ribosomal structure and biogenesis	J
	RNA processing and modification	A
	Transcription	K
	Replication, recombination and repair	L
	Chromatin structure and dynamics	B
Cellular processes and signaling	Cell cycle control, cell division, chromosome partitioning	D
	Nuclear structure	Y
	Defense mechanisms	V
	Signal transduction mechanisms	T
	Cell wall/membrane/envelope biogenesis	M
	Cell motility	N
	Cytoskeleton	Z
	Extracellular structures	W
	Intracellular trafficking, secretion, and vesicular transport	U
	Posttranslational modification, protein turnover, chaperones	O
Metabolism	Energy production and conversion	C
	Carbohydrate transport and metabolism	G
	Amino acid transport and metabolism	E
	Nucleotide transport and metabolism	F
	Coenzyme transport and metabolism	H
	Lipid transport and metabolism	I
	Inorganic ion transport and metabolism	P
	Secondary metabolites biosynthesis, transport and catabolism	Q
Poorly characterized	General function prediction only	R
	Function unknown	S
Not categorized	Mobilome: prophages, transposons	X

The 7 datasets contained an average of 28.77% ($\sigma = 5.74$) of protein-encoding genes not present in the COG for their core-genomes, with a maximum of 38.74% for *Mcc*, the causative agent of CCPP. An average of 62% ($\sigma = 8.57$) of the protein-encoding genes of the pan-genomes did not match any COG. Again, the maximum number was observed in *Mccp* with 76.22% of the protein-endcoding genes not matching to any COG (Table 6).

Table 6. percentage of proteins not in COG database

Dataset	Core	Pan
All	23.58	56.44
Bovine	28.39	61.12
Caprine	24.69	60.86
Mm	24.34	57.04
Mmm (CBPP)	34.44	53.36
Mc	27.22	71.82
Mccp (CCPP)	38.74	76.62
Average	28.77	62.47
Sdev	5.74	8.57

Insertion sequences (IS) dominated the pan-genome of the bovine pathogens of the "*M. mycoides* cluster", and particularly *M. mycoides* subsp. *mycoides*, with about 30% of the pan-genome are IS elements (Annex 12). The proportion of IS in the caprine pathoegns was less than 5% (Annex 10).

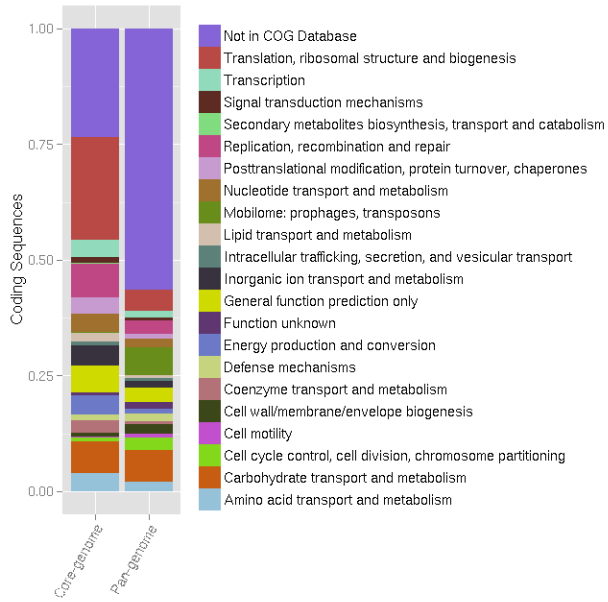


Figure 6. Distribution of COG subcategories for the entire "*M. mycoides* cluster".

The plots for the other experiments can be found in the annexes

Overall, the subcategory most present in the core genomes was "Translation, ribosomal structure and biogenesis" (Figure 6, Annex 8). On the other hand, "Carbohydrate transport and metabolism" dominated the pan genomes overall. We also noticed the following enrichments in the pan genomes: "Replication, recombination and repair" in all the caprine pathogens (Annex 10), "Defense mechanisms" in *Mccp* (CCPP) (Annex 14), "Cell wall/membrane/envelope biogenesis" in the bovine pathogens (Annex 9) and "Inorganic ion transport and metabolism" in *Mmm* (CBPP) (Annex 12). The core genomes presented a similar structure regardless of the experiment.

Trends in categories were also identified. The category "Cellular processes and signaling " was enriched in the pan-genomes, especially for the bovine pathogens (Figure 7). This was less so for the caprine pathogens, *Mycoplasma capricolum* subsp. *capricolum* having absolutely no enrichment of this category compared to its core-genome.

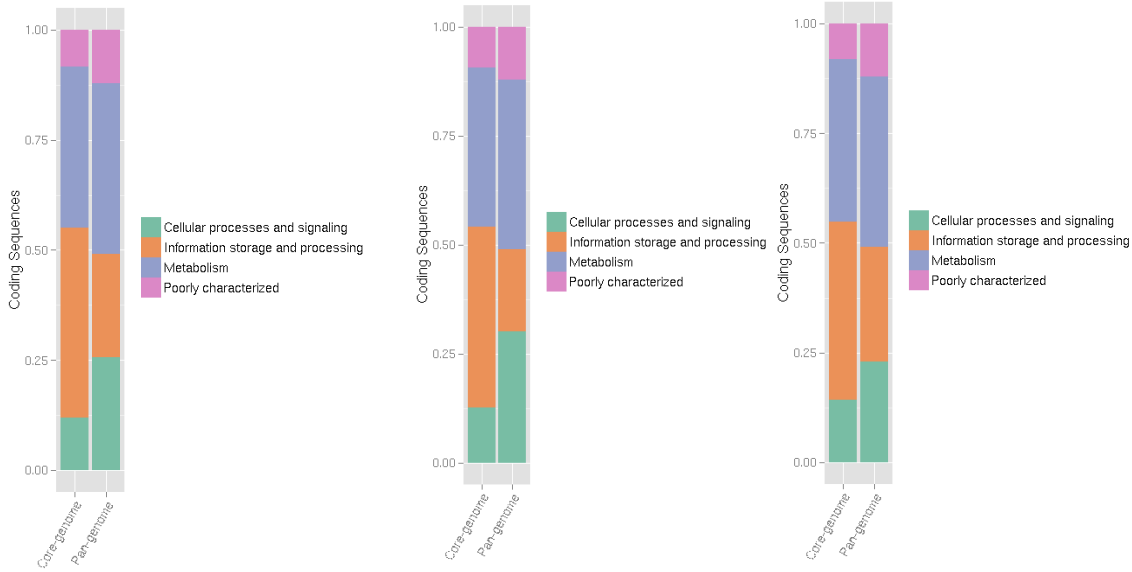


Figure 7, 8 and 9. Distribution of COG categories for the datasets "all", "bovine pathogens" and "caprine pathogens", respectively.

Discussion and Perspectives

Sequencing, Assembly and Annotation

The assemblies produced did all pass minimal standards for Genome Announcements publications. They must albeit be considered draft genomes and are subject to improvements. They also had a high amount repetivive sequences such as Insertion sequences, that influenced especially the ability to reduce the number of contigs in the *Mmm* dataset. Further experiments using long reads such as Pacbio sequencing will help to improve those genome sequences [47].

All the genomes included in the analysis were annotated, even those for which an annotation was already publicly available. This step was crucial to avoid bias generated by different annotation tools and settings as well as differences in manual curation. By re-annotating all the genomes with the same pipeline, using the same database, we ensured that our dataset was consistent and ready for comparative analysis.

Core and pan-genome characterization

As expected if more genomes were added to the dataset the smaller the core-genome was. This makes perfect sense since the core genome shrinks in favor of the pan genome. The more distantly related organisms were included in a group the smaller was the core genome.

The core genomes of *Mmm* and *Mccp*, the causative agents of CBPP and CCPP, respectively were of particular interest. By subtracting the core-genome of a pathogen by the core genome of the subset including its closest relative, we intended to be able to identify genes encoding proteins that are responsible for host tropism and pathogenicity/virulence in CBPP and CCPP.

Our analysis narrowed down to 207 candidates for host tropism and pathogenicity/virulence in *Mmm*. These candidate genes belonged to the core genome of *Mycoplasma mycoides* subsp. *mycoides* but not to the core genome of *Mycoplasma mycoides (*both subspecies*)*. 244 candidate genes were identified for *Mccp*, not belonging to the core-genome of *Mycoplasma capricolum* while being present in the core-genome of *Mycoplasma capricolum* subsp *capripneumoniae*. These candidates, likely to encode proteins specific to host tropism and pathogenicity/virulence should be subjected to laboratory experiments such as in vivo or in vitro experiments that compare wild type strains with mutant strains that lack specific genes. If a role of such protein encoding genes has been confirmed they are candidate molecules for new vaccines against CBPP/CCPP.

Table 6 shows the standard deviation for the clustering of *Mmm* to be higher than in the other groups. The genome size of the 11 sequenced *Mycoplasma* strains was on averagely smaller than the genomes publicly available (1,035,367 for the new genomes, 1,198,410 for the published ones).

This is likely to be attributed to the absence of the entire genome sequences in the draft assembly. The difference of observed and real genome size influenced our analysis in that it underestimated the real number of clusters present in the dataset. Therefore the core genome of *Mmm* is very likely to be larger than estimated. As a result core genes missing in the 11 sequenced strains may have been assigned to the pan genome. Validating the analysis with only finished genomes would have improved our analysis to a) confirm or infirm the current size of the core genome and b) produce more complete *Mycoplasma* genomes to strengthen and confirm this study and further *Mycoplasma* comparisons. Loosing up the definition of pan genome, i.e. making a cluster belonging to the pan genome at *n-1* number of strains present could be a solution as well. On the other hand this would have resulted in a low number of strains tested and therefore resulted in a small pool of input genomes.

Another observation is that *M. mycoides* subsp. *capri* has a smaller core genome with its other subspecies that infects cattle in contrast to the other caprine pathogens (580 clusters vs. 586 in the core genomes). It is consistent with our claim that the core genome of a subspecies of interest contains the genes that encode pathogenicity, virulence and host tropism.

Functional characterization

The transposon category is overly represented in *Mmm*. This confirmed the phylogeny and evolutionary history of the "*Mycoplasma myoicdes* cluster": *Mmm* evolved from a small ruminant pathogen to a bovine-only, lung-specific pathogen [4]. The amount of insertion sequences correlated with the recent adaptation the a new bovine host [48].

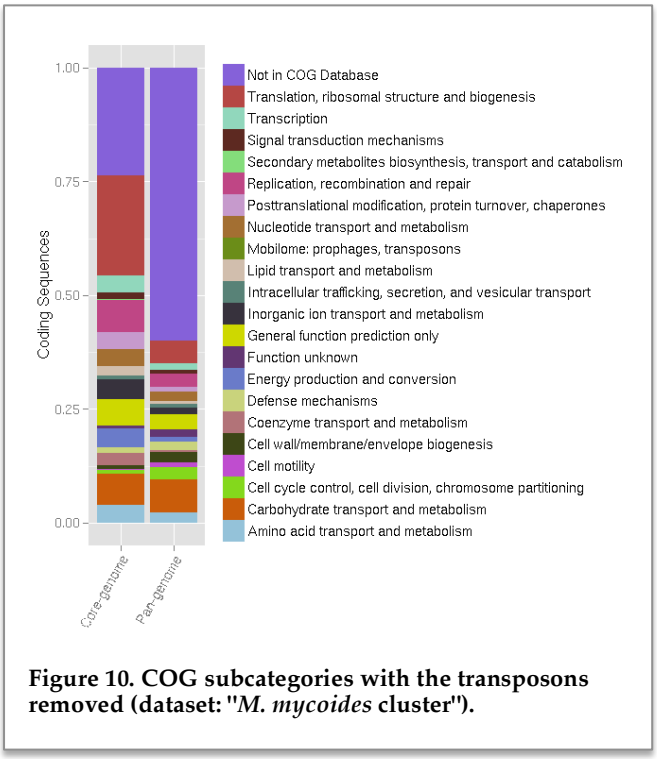


Figure 10. COG subcategories with the transposons removed (dataset: "*M. mycoides* cluster").

However, In the context of developing new vaccines against CBPP and CCPP, transposons are of limited value as vaccine targets. They do not code for virulence factors, or host-specificity; at best they contributes to genome plasticity and regulatory elements. It will be beneficial to exclude IS elements from future analyses (Figure 10).

Proteins not matching to the COG database are considered of importance despite being mostly hypothetical proteins. We can not follow the same logic as for transposons, as little is known about the metabolism

of the '*Mycoplasma mycoides* cluster' and therefore there is no reason to rule out hypothetical proteins for pathogenicity or host-specificity, especially with the current state of genome annotation and databases [49].

While, as previously explained, the core-genomes are interesting to investigate, but in silico analysis should also focus on membrane molecules such as lipoproteins. The host-pathogen interactions of the *Mycoplasma are* suspected to be driven by lipoproteins. Lipoproteins however can be differentially expressed due to their phase variation. In the functional characterization, they seem not to have matched with any COG at many occasions. It is likely that lipoproteins are underrepresented in the COG database. A library of lipoproteins should be constructed using specialized tools for their detection before any further research.

Concluding Remarks

The core and pan genomes of the *Mycoplasma mycoides* cluster have been successfully characterized. The code used is available online and can be useful for analysis future orthoMCL outputs in the context of eukaryotic or prokaryotic comparative analyses.

The work produced here provides a solid baseline for future research on the *Mycoplasma mycoides* cluster. Genes candidate for host tropism and pathogenicity/virulence in *Mmm* and *Mccp* have been discovered; those candidates will be subjected to in vitro and in vivo experiments.

References:

1. Westberg J. The Genome Sequence of Mycoplasma mycoides subsp. mycoides SC Type Strain PG1T, the Causative Agent of Contagious Bovine Pleuropneumonia (CBPP). Genome Res. 2004;14: 221–227. doi:10.1101/gr.1673304

2. Thiaucourt F, Van Der Lugt J, Provost A. Contagious Bovine Pleuropneumonia.

3. Dupuy V, Manso-Silván L, Barbe V, Thebault P, Dordet-Frisoni E, Citti C, et al. Evolutionary History of Contagious Bovine Pleuropneumonia Using Next Generation Sequencing of Mycoplasma mycoides Subsp. mycoides "Small Colony." PLoS One. 2012;7. doi:10.1371/journal.pone.0046821

4. Thiaucourt F, Manso-Silvan L, Salah W, Barbe V, Vacherie B, Jacob D, et al. Mycoplasma mycoides, from "mycoides Small Colony" to "capri". A microevolutionary perspective. BMC Genomics. BioMed Central Ltd; 2011;12: 114. doi:10.1186/1471-2164-12-114

5. Gosney F, Corrò M, Iob L, McAuliffe L, Nicholas R a J. Variable number tandem repeat (VNTR) typing of strains of Mycoplasma mycoides subspecies mycoides small colony isolated from the north-eastern regions of Italy between 1990 and 1993. Vet Microbiol. 2011;147: 220–222. doi:10.1016/j.vetmic.2010.06.009

6. Orsini M, Krasteva I, Marcacci M, Ancora M, Ciammaruconi A, Gentile B, et al. Whole-Genome Sequencing of Mycoplasma mycoides subsp. mycoides Italian Strain 57/13, the Causative Agent of Contagious Bovine Pleuropneumonia. genome A. 2015;3: 2014–2015. doi:10.1128/genomeA.00197-15.Copyright

7. OIE. Contagious Bovine Pleuropneumonia [Internet]. 2014 [cited 2 Jun 2015]. Available: http://www.oie.int/fileadmin/Home/eng/Health_standards/tahm/2.04.09_CBPP.pdf

8. Krasteva I, Liljander A, Fischer A, Smith DGE, Inglis NF, Scacchia M, et al. Characterization of the in vitro core surface proteome of Mycoplasma mycoides subsp. mycoides, the causative agent of contagious bovine pleuropneumonia. Vet Microbiol. Elsevier B.V.; 2014;168: 116–123. doi:10.1016/j.vetmic.2013.10.025

9. Chu Y, Gao P, Zhao P, He Y, Liao N, Jackman S, et al. Genome sequence of Mycoplasma capricolum subsp. capripneumoniae strain M1601. J Bacteriol. 2011;193: 6098–9. doi:10.1128/JB.05980-11

10. Thiaucourt F, Bölske G. Contagious caprine pleuropneumonia and other pulmonary mycoplasmoses of sheep and goats. Rev Sci Tech. 1996;15: 1397–1414.

11. Wise KS, Calcutt MJ, Foecking MF, Madupu R, DeBoy RT, Röske K, et al. Complete genome sequences of Mycoplasma leachii strain PG50T and the pathogenic Mycoplasma mycoides subsp. mycoides small colony biotype strain Gladysdale. J Bacteriol. 2012;194: 4448–4449. doi:10.1128/JB.00841-12

12. Cottew GS, Breard A, DaMassa AJ, Ernø H, Leach RH, Lefevre PC, et al. Taxonomy of the Mycoplasma mycoides cluster. Isr J Med Sci. 1987;23: 632–5. Available: http://www.ncbi.nlm.nih.gov/pubmed/3312102

13. Thomas A, Linden A, Mainil J, Bischof DF, Frey J, Vilei EM. Mycoplasma bovis shares insertion sequences with Mycoplasma agalactiae and Mycoplasma mycoides subsp. mycoides SC: Evolutionary and developmental aspects. FEMS Microbiol Lett. 2005;245: 249–55. doi:10.1016/j.femsle.2005.03.013

14. Manso-Silván L, Perrier X, Thiaucourt F. Phylogeny of the Mycoplasma mycoides cluster based on analysis of five conserved protein-coding sequences and possible implications for the taxonomy of the group. Int J Syst Evol Microbiol. 2007;57: 2247–58. doi:10.1099/ijs.0.64918-0

15. Kim K-S, Ko KS, Chang M-W, Hahn TW, Hong SK, Kook Y-H. Use of rpoB sequences for phylogenetic study of Mycoplasma species. FEMS Microbiol Lett. The Oxford University Press; 2003;226: 299–305. doi:10.1016/S0378-1097(03)00618-9

16. Nwankpa ND, Manso-silvan L, Lorenzon S, Yaya a., Lombin LH, Thiaucourt F. Variable Number Tandem Repeat (VNTR) analysis reveals genetic diversity within Mycoplasma mycoides mycoides Small Colony isolates from Nigeria. Vet Microbiol. Elsevier B.V.; 2010;146: 354–355. doi:10.1016/j.vetmic.2010.05.020

17. Fischer A, Shapiro B, Muriuki C, Heller M, Schnee C, Bongcam-Rudloff E, et al. The origin of the "mycoplasma mycoides cluster" coincides with domestication of ruminants. PLoS One. 2012;7: 3–8. doi:10.1371/journal.pone.0036150

18. Pilo P, Frey J, Vilei EM. Molecular mechanisms of pathogenicity of Mycoplasma mycoides subsp. mycoides SC. Vet J. 2007;174: 513–21. doi:10.1016/j.tvjl.2006.10.016

19. Bertin C, Pau-Roblot C, Courtois J, Manso-Silván L, Thiaucourt F, Tardy F, et al. Characterization of Free Exopolysaccharides Secreted by Mycoplasma mycoides Subsp. mycoides. PLoS One. 2013;8. doi:10.1371/journal.pone.0068373

20. Browning GF, Marenda MS, Noormohammadi AH, Markham PF. The central role of lipoproteins in the pathogenesis of mycoplasmoses. Vet Microbiol. 2011;153: 44–50. doi:10.1016/j.vetmic.2011.05.031

21. Bischof DF, Janis C, Vilei EM, Bertoni G, Frey J. Cytotoxicity of Mycoplasma mycoides subsp. mycoides small colony type to bovine epithelial cells. Infect Immun. 2008;76: 263–9. doi:10.1128/IAI.00938-07

22. Bertin C, Pau-Roblot C, Courtois J, Manso-Silván L, Tardy F, Poumarat F, et al. Highly Dynamic Genomic Loci Drive the Synthesis of Two Types of Capsular or Secreted Polysaccharides within the Mycoplasma mycoides Cluster. Appl Environ Microbiol. 2015;81: 676–687. doi:10.1128/AEM.02892-14

23. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

24. CLC genomic Workbench [Internet]. p. CLC Genomics Workbench 7.0.3 (http://www.clcbio.co.

25. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–80. doi:10.1038/nature03959

26. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science. 2000;287: 2196–204. Available: http://www.ncbi.nlm.nih.gov/pubmed/10731133

27. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–9. doi:10.1101/gr.074492.107

28. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. BioMed Central Ltd; 2012;1: 18. doi:10.1186/2047-217X-1-18

29. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res. 2008;18: 810–20. doi:10.1101/gr.7337908

30. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–77. doi:10.1089/cmb.2012.0021

31. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28: 1420–8. doi:10.1093/bioinformatics/bts174

32. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo M-J, Dupont CL, Badger JH, et al. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nat Biotechnol. Nature Publishing Group; 2011;29: 915–21. doi:10.1038/nbt.1966

33. McCorrison JM, Venepally P, Singh I, Fouts DE, Lasken RS, Methé BA. NeatFreq: reference-free data reduction and coverage normalization for De Novo sequence assembly. BMC Bioinformatics. 2014;15: 357. doi:10.1186/s12859-014-0357-3

34. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–9. doi:10.1093/bioinformatics/btu153

35. Laslett D. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32: 11–16. doi:10.1093/nar/gkh152

36. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

37. Kreutz C, Raue A, Timmer J. Likelihood based observability analysis and confidence intervals for predictions of dynamic models. BMC Syst Biol. 2012;6: 120. doi:10.1186/1752-0509-6-120

38. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35: D61–5. doi:10.1093/nar/gkl842

39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–10. doi:10.1016/S0022-2836(05)80360-2

40. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42: D222–30. doi:10.1093/nar/gkt1223

41. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. Public Library of Science; 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195

42. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13: 2178–89. doi:10.1101/gr.1224503

43. Enright AJ. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30: 1575–1584. doi:10.1093/nar/30.7.1575

44. Ozer EA, Allen JP, Hauser AR. Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic

tools Spine and AGEnt. BMC Genomics. 2014;15: 737. doi:10.1186/1471-2164-15-737

45. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015;43: D261–9. doi:10.1093/nar/gku1223

46. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–402. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

47. Anne Fischer, Ivette Santana-Cruz, Jan Hegermann, Hadrien Gourlé, Elise Schieck, Mathieu Lambert, Suvarna Nadendla, Hezron Wesonga, Rachel A Miller, Sanjay Vashee, Johann Weber, Jochen Meens, Joachim Frey, Joerg Jores. High quality draft genomes of the Mycoplasma mycoides subsp. mycoides challenge strains Afadé and B237. Stand Genomic Sci. (Submitted).

48. Smith NH, Gordon S V, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG. Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol. 2006;4: 670–81. doi:10.1038/nrmicro1472

49. Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics. 2007;23: i41–i48. doi:10.1093/bioinformatics/btm229

Annexes

Annex 1. Clustering statistics for the "*M. mycoides* cluster"

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
5713	23,95	1192498	1039242	87,15	622998	416244	59,95	40,05	24,17	24,24
9501	24,53	1070992	890685	83,16	615222	275463	69,07	30,93	24,14	24,13
AFAD	23,94	1190241	1031712	86,68	617475	414237	59,85	40,15	24,16	24,22
B237	23,95	1203804	1047549	87,02	617724	429825	58,97	41,03	24,16	24,27
B66	24,02	988252	811044	82,07	592953	218091	73,11	26,89	24,09	23,6
C11	23,72	1047736	886191	84,58	632103	254088	71,33	28,67	24,11	23,45
FATI	24,73	1068802	872799	81,66	608004	264795	69,66	30,34	24,14	24,12
GLAD	23,95	1193808	1041309	87,23	618501	422808	59,4	40,6	24,15	24,25
L2	23,98	1041406	878223	84,33	609150	269073	69,36	30,64	24,13	23,67
MATA	24,30	1033254	835998	80,91	595980	240018	71,29	28,71	24,12	23,5
PG1	23,97	1211703	1052265	86,84	619713	432552	58,89	41,11	24,18	24,29
T144	24,00	1050916	867204	82,52	609822	257382	70,32	29,68	24,12	24,05
V5	23,86	984029	798039	81,1	584688	213351	73,27	26,73	24,16	23,49
9901	23,67	1017232	908598	89,32	612804	295794	67,45	32,55	24,1	23,5
PG50	23,75	1008951	901002	89,3	618681	282321	68,67	31,33	24,11	23,69
8700	23,68	1017333	880947	86,59	595743	285204	67,63	32,37	24,2	23,59
9910	23,55	1006326	857724	85,23	592656	265068	69,1	30,9	24,18	23,52
ABOM	23,66	1017293	881253	86,63	597636	283617	67,82	32,18	24,19	23,48
F38	23,67	1016760	880959	86,64	596961	283998	67,76	32,24	24,2	23,51
ILRI	23,67	1017183	880488	86,56	596031	284457	67,69	32,31	24,18	23,56
M160	23,63	1018102	876648	86,11	595224	281424	67,9	32,1	24,21	23,55
1423	23,57	1032226	917187	88,86	628332	288855	68,51	31,49	24,1	23,49
14DL	23,74	964668	872217	90,42	643596	228621	73,79	26,21	24,16	23,47
ATCC	23,77	1010023	908841	89,98	625659	283182	68,84	31,16	24,13	23,8
GM50	23,77	1024448	923988	90,19	639495	284493	69,21	30,79	24,11	23,83
Capr	23,93	1210757	994743	82,16	631206	363537	63,45	36,55	24,04	23,55
G131	23,87	1058351	934713	88,32	627027	307686	67,08	32,92	24,07	23,46
GM12	23,92	1084586	985620	90,88	632679	352941	64,19	35,81	24,05	24,16
LC95	23,82	1153998	1048161	90,83	632115	416046	60,31	39,69	24,06	23,89
PG3	23,67	971239	882681	90,88	629385	253296	71,3	28,7	24,06	23,88
YGOA	24,20	1088983	933030	85,68	623649	309381	66,84	33,16	24,07	23,83

Annex 2. Clustering statistics for the bovine pathogens of the cluster

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
5713	23,95	1192498	1039242	87,15	755478	283764	72,7	27,3	24,12	24,4
9501	24,53	1070992	890685	83,16	745554	145131	83,71	16,29	24,07	24,49
AFAD	23,94	1190241	1031712	86,68	749328	282384	72,63	27,37	24,11	24,38
B237	23,95	1203804	1047549	87,02	750300	297249	71,62	28,38	24,11	24,45
B66	24,02	988252	811044	82,07	716907	94137	88,39	11,61	24,01	23,56
C11	23,72	1047736	886191	84,58	764037	122154	86,22	13,78	24,04	23,19
FATI	24,73	1068802	872799	81,66	737670	135129	84,52	15,48	24,06	24,51
GLAD	23,95	1193808	1041309	87,23	750633	290676	72,09	27,91	24,1	24,43
L2	23,98	1041406	878223	84,33	738477	139746	84,09	15,91	24,06	23,64
MATA	24,3	1033254	835998	80,91	726951	109047	86,96	13,04	24,04	23,29
PG1	23,97	1211703	1052265	86,84	752850	299415	71,55	28,45	24,12	24,47
T144	24	1050916	867204	82,52	738189	129015	85,12	14,88	24,06	24,37
V5	23,86	984029	798039	81,1	712590	85449	89,29	10,71	24,06	23,33
9901	23,67	1017232	908598	89,32	750405	158193	82,59	17,41	23,96	23,66
PG50	23,75	1008951	901002	89,3	759723	141279	84,32	15,68	23,95	24,11

Annex 3. Clustering statistics for the caprine pathogens of the cluster

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
8700	23,68	1017333	880947	86,59	648948	231999	73,66	26,34	24,13	23,64
9910	23,55	1006326	857724	85,23	645084	212640	75,21	24,79	24,11	23,59
ABOM	23,66	1017293	881253	86,63	648813	232440	73,62	26,38	24,09	23,6
F38	23,67	1016760	880959	86,64	648105	232854	73,57	26,43	24,12	23,58
ILRI	23,67	1017183	880488	86,56	649173	231315	73,73	26,27	24,11	23,59
M160	23,63	1018102	876648	86,11	647925	228723	73,91	26,09	24,14	23,59
1423	23,57	1032226	917187	88,86	695253	221934	75,8	24,2	24	23,62
14DL	23,74	964668	872217	90,42	714252	157965	81,89	18,11	24,07	23,57
ATCC	23,77	1010023	908841	89,98	693675	215166	76,33	23,67	24,04	24
GM50	23,77	1024448	923988	90,19	709065	214923	76,74	23,26	24	24,1
Capr	23,93	1210757	994743	82,16	702384	292359	70,61	29,39	23,97	23,59
G131	23,87	1058351	934713	88,32	699858	234855	74,87	25,13	23,96	23,59
GM12	23,92	1084586	985620	90,88	712377	273243	72,28	27,72	23,96	24,41
LC95	23,82	1153998	1048161	90,83	709977	338184	67,74	32,26	23,97	24,03
PG3	23,67	971239	882681	90,88	701868	180813	79,52	20,48	23,98	24,14
YGOA	24,2	1088983	933030	85,68	693321	239709	74,31	25,69	23,96	24,06

Annex 4. Clustering statistics for *Mycoplasma mycoides*

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
5713	23,95	1192498	1039242	87,15	664179	375063	63,91	36,09	24,14	24,29
9501	24,53	1070992	890685	83,16	657459	233226	73,81	26,19	24,11	24,21
AFAD	23,94	1190241	1031712	86,68	658377	373335	63,81	36,19	24,13	24,28
B237	23,95	1203804	1047549	87,02	659331	388218	62,94	37,06	24,13	24,33
B66	24,02	988252	811044	82,07	632217	178827	77,95	22,05	24,08	23,55
C11	23,72	1047736	886191	84,58	675705	210486	76,25	23,75	24,09	23,39
FATI	24,73	1068802	872799	81,66	649515	223284	74,42	25,58	24,12	24,18
GLAD	23,95	1193808	1041309	87,23	659766	381543	63,36	36,64	24,13	24,31
L2	23,98	1041406	878223	84,33	650955	227268	74,12	25,88	24,11	23,65
MATA	24,3	1033254	835998	80,91	640902	195096	76,66	23,34	24,09	23,46
PG1	23,97	1211703	1052265	86,84	663501	388764	63,05	36,95	24,14	24,37
T144	24	1050916	867204	82,52	651153	216051	75,09	24,91	24,11	24,09
V5	23,86	984029	798039	81,1	628152	169887	78,71	21,29	24,12	23,48
Capr	23,93	1210757	994743	82,16	680823	313920	68,44	31,56	24,01	23,54
G131	23,87	1058351	934713	88,32	672939	261774	71,99	28,01	24,03	23,46
GM12	23,92	1084586	985620	90,88	681018	304602	69,1	30,9	24	24,29
LC95	23,82	1153998	1048161	90,83	680967	367194	64,97	35,03	24,01	23,95
PG3	23,67	971239	882681	90,88	675219	207462	76,5	23,5	24,01	24,02
YGOA	24,2	1088983	933030	85,68	672486	260544	72,08	27,92	24,01	23,92

Annex 5. Clustering statistics for *Mycoplasma mycoides* subsp. *mycoides*

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
5713	23,95	1192498	1039242	87,15	813285	225957	78,26	21,74	24,03	24,79
9501	24,53	1070992	890685	83,16	799794	90891	89,8	10,2	23,97	25,59
AFAD	23,94	1190241	1031712	86,68	805401	226311	78,06	21,94	24,02	24,79
B237	23,95	1203804	1047549	87,02	809970	237579	77,32	22,68	24,02	24,84
B66	24,02	988252	811044	82,07	769479	41565	94,88	5,12	23,93	24,58
C11	23,72	1047736	886191	84,58	817941	68250	92,3	7,7	23,94	23,64
FATI	24,73	1068802	872799	81,66	791808	80991	90,72	9,28	23,98	25,64
GLAD	23,95	1193808	1041309	87,23	808371	232938	77,63	22,37	24,01	24,83
L2	23,98	1041406	878223	84,33	792912	85311	90,29	9,71	23,97	24,18
MATA	24,3	1033254	835998	80,91	779109	56889	93,2	6,8	23,95	23,84
PG1	23,97	1211703	1052265	86,84	813672	238593	77,33	22,67	24,01	24,94
T144	24	1050916	867204	82,52	788943	78261	90,98	9,02	23,98	25,35
V5	23,86	984029	798039	81,1	762837	35202	95,59	4,41	23,98	24,09

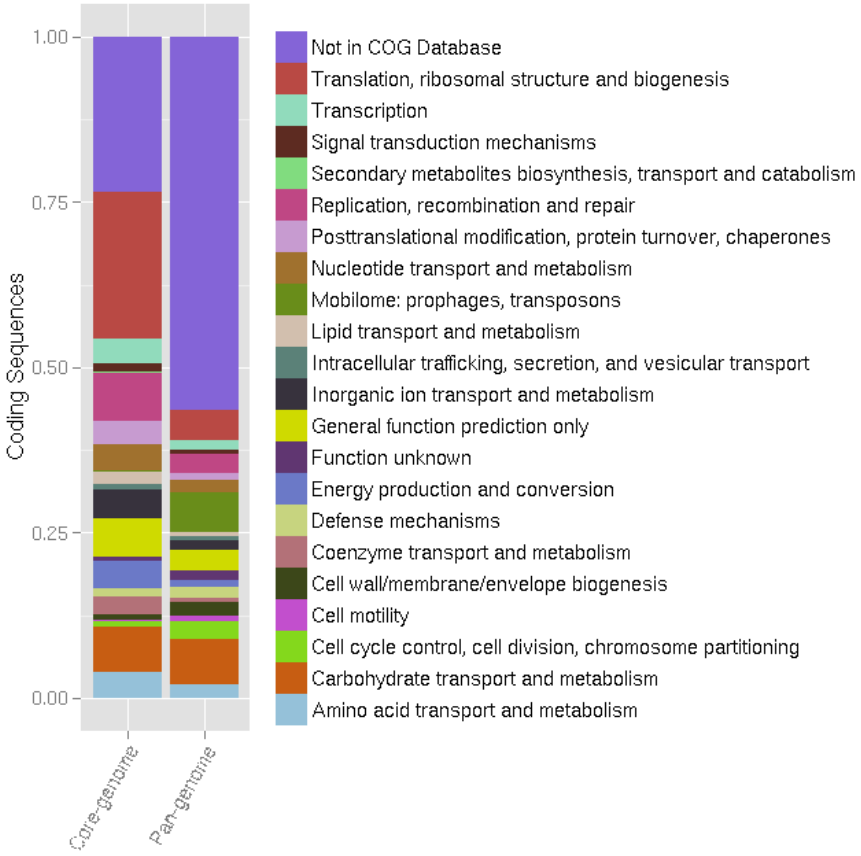
Annex 6. Clustering statistics for *Mycoplasma capricolum*

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
8700	23,68	1017333	880947	86,59	722625	158322	82,03	17,97	24,09	23,61
9910	23,55	1006326	857724	85,23	717684	140040	83,67	16,33	24,06	23,55
ABOM	23,66	1017293	881253	86,63	722526	158727	81,99	18,01	24,03	23,63
F38	23,67	1016760	880959	86,64	722733	158226	82,04	17,96	24,06	23,61
ILRI	23,67	1017183	880488	86,56	723585	156903	82,18	17,82	24,07	23,54
M160	23,63	1018102	876648	86,11	722505	154143	82,42	17,58	24,08	23,62
1423	23,57	1032226	917187	88,86	779460	137727	84,98	15,02	23,94	23,71
14DL	23,74	964668	872217	90,42	801126	71091	91,85	8,15	24	23,73
ATCC	23,77	1010023	908841	89,98	779220	129621	85,74	14,26	23,98	24,33
GM50	23,77	1024448	923988	90,19	795093	128895	86,05	13,95	23,94	24,49

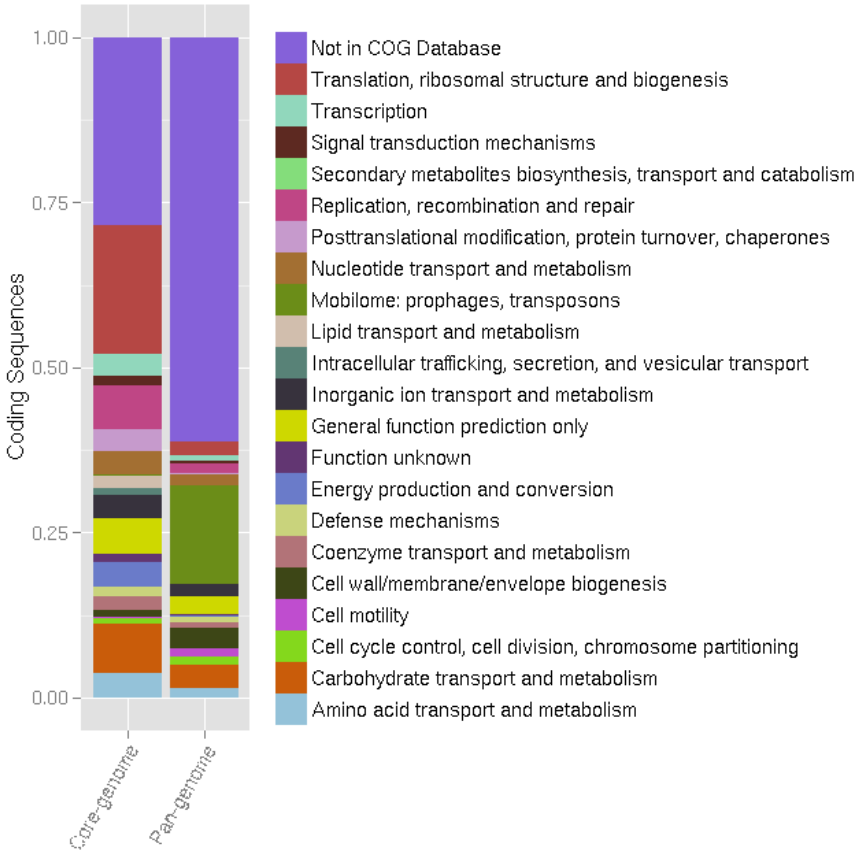
Annex 7. Clustering statistics for *Mycoplasma capricolum* subsp. *capripneumoniae*

Genome	GC	Size	CodingSize	%Coding	CoreSize	PanSize	%Core	%Pan	GC Core	GC Pan
8700	23,68	1017333	880947	86,59	849177	31770	96,39	3,61	24,02	23,51
9910	23,55	1006326	857724	85,23	844086	13638	98,41	1,59	24	22,86
ABOM	23,66	1017293	881253	86,63	849462	31791	96,39	3,61	23,96	23,85
F38	23,67	1016760	880959	86,64	850119	30840	96,5	3,5	23,98	23,76
ILRI	23,67	1017183	880488	86,56	851220	29268	96,68	3,32	24	23,35
M160	23,63	1018102	876648	86,11	847248	29400	96,65	3,35	24,01	23,46

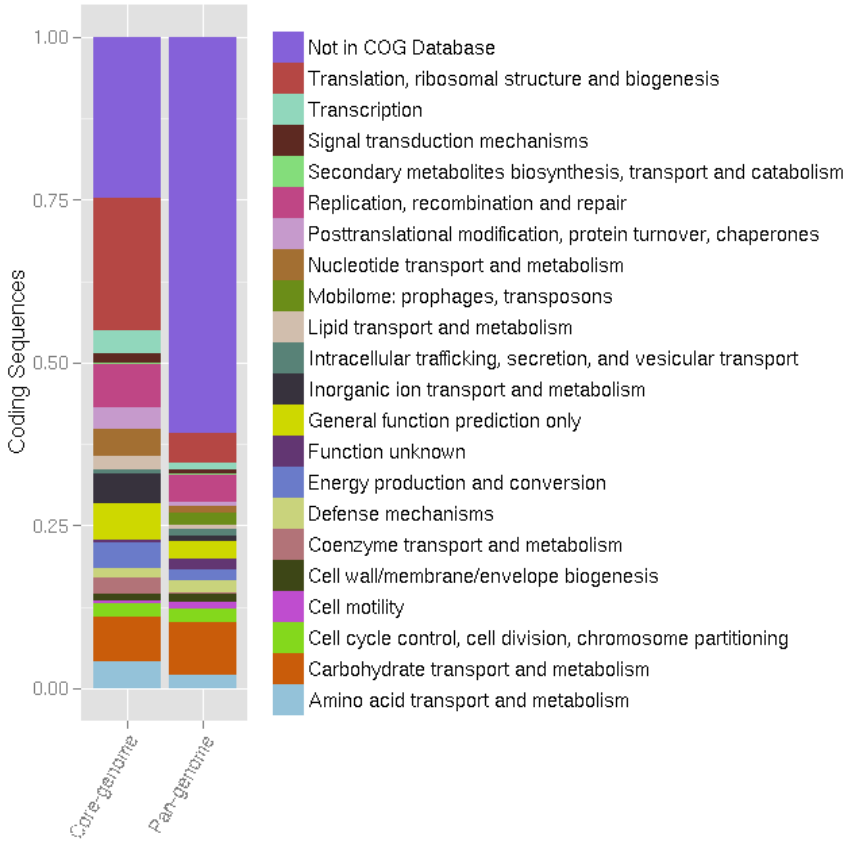
Annex 8. COG subcategories plot for the core and pan genome of the "*Mycoplasma mycoides* cluster"



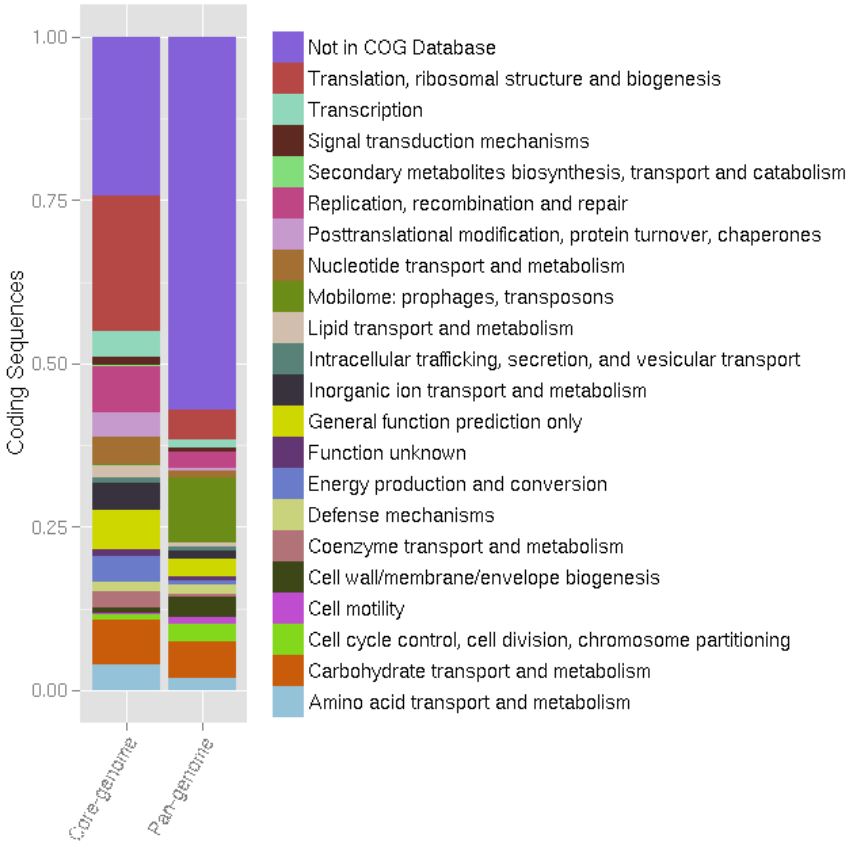
Annex 9. COG subcategories plot for the core and pan genome of the bovine pathogens of the cluster



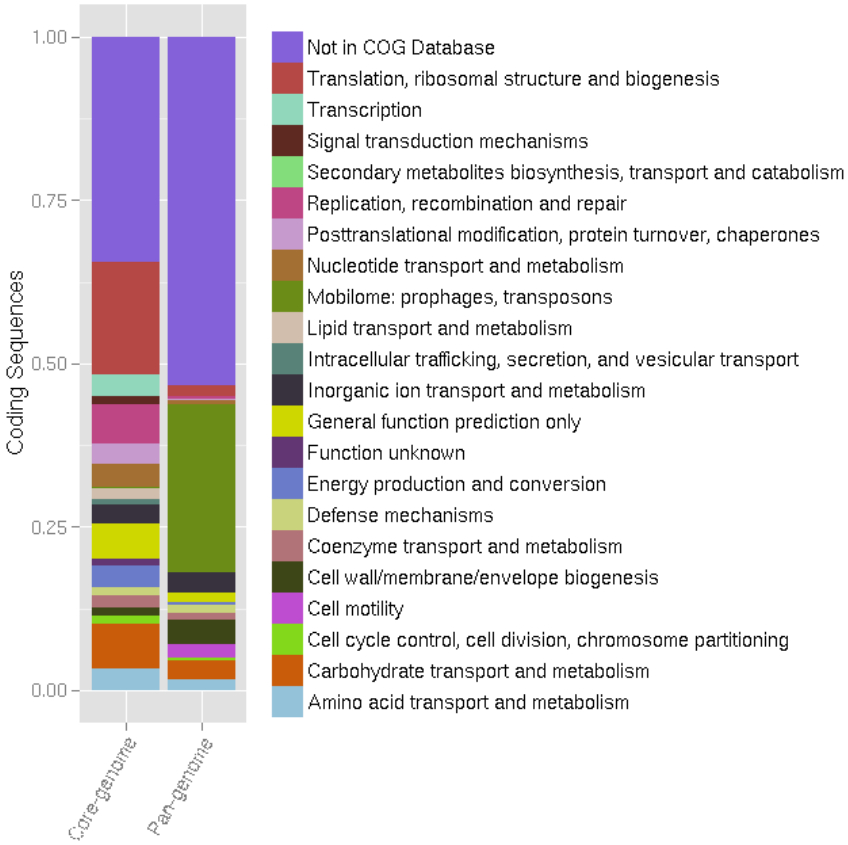
Annex 10. COG subcategories plot for the core and pan genome of the caprine pathogens of the cluster



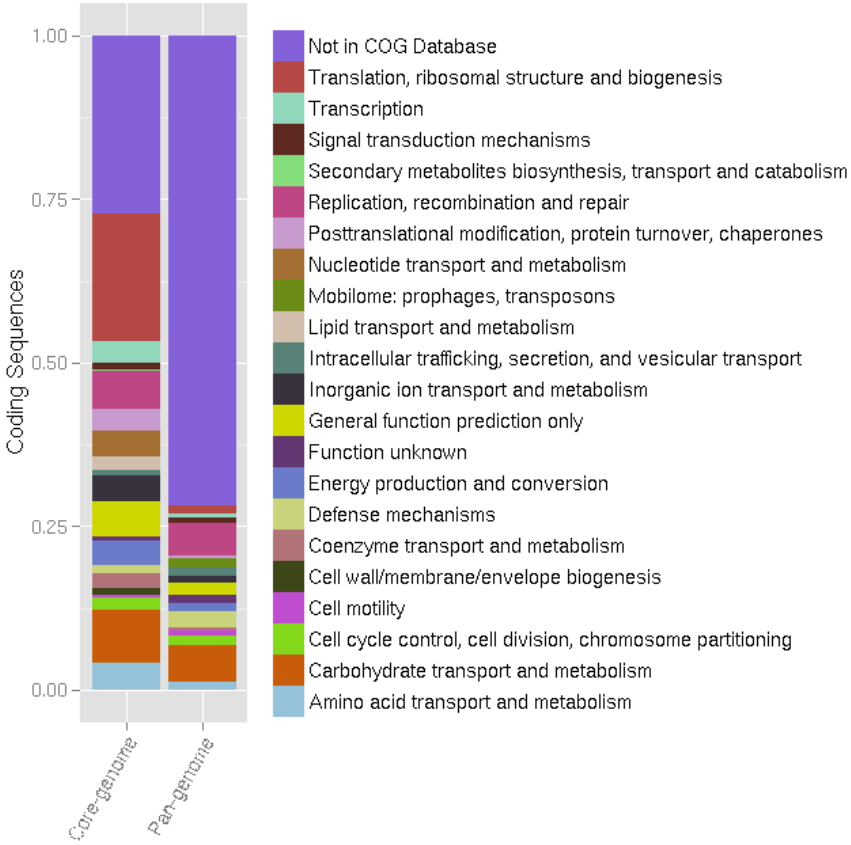
Annex 11. COG subcategories plot for the core and pan genome of *Mycoplasma mycoides*



Annex 12. COG subcategories plot for the core and pan genome of *Mycoplasma mycoides* subsp. *mycoides*



Annex 13. COG subcategories plot for the core and pan genome of *Mycoplasma capricolum*



Annex 14. COG subcategories plot for the core and pan genome of *Mycoplasma capricolum* subsp *capripneumoniae*