Genomic comparison of mycoplasma species isolated from commercial chickens in South Africa

by

Amanda Beylefeld

submitted in fulfilment in accordance with the requirements for the degree

Doctor of Philosophy (PhD)

in Production Animal Studies

in the Faculty of Veterinary Sciences,
University of Pretoria

October 2018

Supervisor: Prof Celia Abolnik

DECLARATION

I, Amanda Beylefeld, student number 25285573 hereby declare that this dissertation, “Genomic comparison of mycoplasma species isolated from commercial chickens in South Africa”, is submitted in accordance with the requirements for the Doctor of Philosophy (PhD) degree at University of Pretoria, is my own original work and has not previously been submitted to any other institution of higher learning. All sources cited or quoted in this research paper are indicated and acknowledged with a comprehensive list of references.

Amanda Beylefeld

31 October 2018

ETHICS STATEMENT

The author, whose name appears on the title page of this thesis, has obtained, for the research described in this work, the applicable research ethics approval. The author declares that she has observed the ethical standards required in terms of the University of Pretoria's Code of ethics for researchers and the Policy guidelines for responsible research.

ACKNOWLEDGEMENTS

First and foremost, I would like to give my praise and thanks to my father God, who put this opportunity on my path and walked this road with me.

I would also like to give a very special thanks to Prof. Celia Abolnik for her outstanding leadership, invaluable advice, guidance, support and patience and for granting me this amazing opportunity.

I would also like to take the time to express my sincere gratitude to the following people, without whom this thesis would not be possible:

Mr. Johan Gouws and Ms. Pamela Wambulawaye for their help with the mycoplasma isolations and identification

Dr. Buks Wandrag for his help, support and invaluable talks about the poultry industry

The whole poultry group for their support, friendship and encouragement, in no particular order Thandeka Phiri, Dr. Dauda Bwala, Dr. Adeyemi Modupeore, Dr Agnes Oladokun and the late Dr Ponman Solomon

The Technology Innovation Agency-Tshwane Animal Health Innovation Cluster Initiative, the Welfare Sector Education and Training Authority, University of Pretoria and the National Research Foundation for their financial support.

Then also my parents, Arnold and Sharon van Tonder for all their love and support and for giving me all the opportunities to become the successful young adult I am today

And my grandparents for their love, support and financial aid throughout my studies

Then to my amazing husband, Jaco Beylefeld for all his love, encouragement, support, never-ending patience and invaluable help with the command line programs (or Linux headaches, as would like to call it)

This has been a great journey, without many ups and downs, but I enjoyed every minute of it and I am very grateful for all the love, support and patience provided by all my friends and family

ABSTRACT

Mycoplasma gallisepticum and *Mycoplasma synoviae* are listed by the World Organisation for Animal Health (OIE) as notifiable avian diseases agents and are considered as economically important species affecting the South African poultry industry. A decreased number of cases identified by culture with growth inhibition of *M. gallisepticum* and *M. synoviae* and an increase in unidentified mycoplasma species were observed at the University of Pretoria diagnostic laboratory. Samples were isolated from chickens displaying typical signs associated with pathogenic mycoplasma infection, thus necessitating a closer look at the lesser known mycoplasma species found in poultry. The aim of this study was to use second generation sequencing to identify and compare mycoplasma species isolated from commercial chickens in South Africa. Mycoplasma samples were isolated, sequenced and *de novo* assembled for identification by 16S rRNA phylogeny. A total of 124 samples were received between 2003 and 2015, 44 of these samples contained multiple species resulting in 179 isolates identified as *M. gallisepticum* (24.58%), *M. synoviae* (9.50%), *M. gallinaceum* (24.02%), *M. gallinarum* (24.58%), *M. pullorum* (13.97%) and *M. iners* (2.79%) and one *Acholeplasma laidlawii* (0.56%). Antimicrobial resistance (AMR) to chlortetracycline, enrofloxacin, tylosin and tiamulin and the genes involved in AMR were studied. Three *M. gallinaceum* samples showed possible multidrug resistance and novel point mutations associated with AMR in *M. gallinaceum* and *M. gallinarum* were identified. The first complete genome of *M. pullorum* was assembled, annotated and published. Draft genome assemblies for axenic strains were constructed and candidate genes that can be tested for novel diagnostic and vaccine targets were identified.

Key terms: Mycoplasma, poultry, antimicrobial resistance, 16S rRNA, *Mycoplasma gallisepticum, M. synoviae, Mycoplasma pullorum, Mycoplasma gallinaceum*, sequencing, whole genome sequencing

LIST OF ABBREVIATIONS

AIC	-	Akaike information criteria
AMR	-	Antimicrobial resistance
ATCC	-	American Type Culture Collection
BIC	-	Bayesian information criteria
CDS	-	coding sequences
CI	-	Consistency index
COG	-	Clusters of Orthologous Group
CRD	-	Chronic respiratory disease
crmA	-	Cytokine response modifier A gene
cysP	-	cysteine proteases gene
DIVA	-	Differentiating infected from vaccinated animals
DOE-JGI	-	Department of Energy Joint Genome Institute
DNA	-	Deoxyribonucleic acid
dNTP	-	deoxynucleoside triphosphate
DVTD	-	Department of Veterinary Tropical Diseases
ELISA	-	Enzyme-linked immunosorbent assays
gapA	-	Glyceraldehyde-3-phosphate dehydrogenase A gene
GTS	-	Gene-targeted sequencing
*gyr*A	-	DNA gyrase subunit A gene
*gyr*B	-	DNA gyrase subunit B gene
HGT	-	Horizontal gene transfer
HI	-	Haemagglutination inhibition
IBV	-	Infectious bronchitis virus
IgG	-	immunoglobulin G
IGSR	-	Intergenic spacer region
IMG	-	Integrated Microbial Genomes and Microbiomes
LCB	-	Locally collinear block
MAFFT	-	Multiple Alignment using Fast Fourier Transform
MDR	-	Multi-drug resistant
MG	-	*Mycoplasma gallisepticum*
MGAP	-	Microbial Genome Annotation Pipeline
MIC	-	Minimum inhibitory concentration
MS		*Mycoplasma synoviae*
NAD	-	Nicotinamide adenine dinucleotide
NCBI	-	National Centre for Biotechnology Information

ncRNAs	-	non-coding RNAs
NDV	-	Newcastle disease virus
NNI	-	Nearest neighbour interchange
OIE	-	World Organisation for Animal Health
osmC	-	osmotically inducible protein C gene
OLC	-	Overlap-layout-consensus
ORF	-	Open reading frames
ohr	-	organic hydroperoxide resistance gene
*par*C	-	Topoisomerase IV subunit A gene
parE	-	Topoisomerase IV subunit B gene
PCR	-	Polymerase chain reaction
PGAP	-	Prokaryotic Genome Annotation Pipeline
PGM	-	Personal genome machine
PTS	-	phosphotransferase system
PvpA	-	Phase-variable putative adhesin
qPCR	-	Real-time PCR
QRDR	-	Quinolone resistance-determining resistance regions
RAPD	-	random amplification of polymorphic DNA
RAST	-	Rapid Annotation using Subsystem Technology
RFLP	-	Restriction fragment length polymorphism
RI	-	Retention index
*rpl*D	-	ribosomal protein L4 gene
r*pl*V	-	ribosomal protein L22 gene
rRNA	-	Ribosomal ribonucleic acid
RSA	-	Rapid serum agglutination
rt-PCR	-	reverse transcription PCR
SGS	-	Second-generation sequencing
SMRT	-	Single molecule real time
SNP	-	Single-nucleotide polymorphisms
SOLiD	-	Sequencing by oligonucleotide ligation and detection
SPR	-	Subtree pruning and regrafting
TBR	-	Tree bisection and reconnection
tRNA	-	transfer RNA
UP	-	University of Pretoria
VlhA	-	Variable lipoprotein haemagglutinin
WGS	-	Whole genome sequencing
w/v	-	Weight per volume

TABLE OF CONTENTS

DECLARATION .... **I**

ETHICS STATEMENT .... **II**

ACKNOWLEDGEMENTS .... **III**

ABSTRACT .... **IV**

LIST OF ABBREVIATIONS .... **V**

TABLE OF CONTENTS .... **VII**

LIST OF FIGURES .... **X**

LIST OF TABLES .... **XV**

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW **1**

1.1. INTRODUCTION 1

1.2. MYCOPLASMA 2

1.2.1. General characteristics 2

1.2.2. Poultry mycoplasmas 3

1.2.3. Role of poultry mycoplasmas in disease 4

1.2.4. Diagnosis of poultry mycoplasma 6

1.2.5. Serological methods 7

1.2.6. DNA based methods 8

1.2.7. Treatment and prevention of poultry mycoplasma 9

1.2.8. Mycoplasma genomics 13

1.3. GENOME SEQUENCING 16

1.3.1. Brief overview 16

1.3.2. Sample processing 18

1.3.3. DNA sequencing 19

1.3.4. Data analysis 21

1.4. AIM OF THE RESEARCH 24

1.5. PURPOSE OF THE RESEARCH 24

CHAPTER 2: IDENTIFICATION OF MYCOPLASMA SPECIES FROM SOUTH AFRICAN POULTRY FARMS AND ASSESSMENT OF ANTIMICROBIAL RESISTANCE................ 25

2.1. INTRODUCTION... 25

2.2. MATERIALS AND METHODS 26

2.2.1. Sample collection..26

2.2.2. Mycoplasma isolation by culture and identification by growth inhibition 29

2.2.3. Mycoplasma DNA isolation 30

2.2.4. 16S rRNA gene phylogeny 31

2.2.5. Minimum inhibitory concentration (MIC) assays...32

2.2.6. Antimicrobial resistance genes...33

2.3. RESULTS... 34

2.3.1. Mycoplasma identification by growth inhibition 34

2.3.2. Mycoplasma DNA isolation 34

2.3.3. Mycoplasma identification by 16S rRNA gene identification 35

2.3.4. Comparison of culture with 16S rRNA gene identification...............................40

2.3.5. Minimum inhibitory concentration (MIC) assays...40

2.3.6. Antimicrobial resistance genes...42

2.4. DISCUSSION 49

CHAPTER 3: GENOME ASSEMBLY AND ANNOTATION OF MYCOPLASMA PULLORUM, ISOLATED FROM DOMESTIC POULTRY IN SOUTH AFRICA.................. 53

3.1. INTRODUCTION... 53

3.2. MATERIALS AND METHODS 54

3.2.1. Sample collection, isolation and identification...54

3.2.2. Quality control..54

3.2.3. Sequence assembly 54

3.2.4. Genome annotation and viewing 56

3.3. RESULTS... 58

3.3.1. Quality control..58

3.3.2. Sequence assembly 59

3.3.3. Genome annotation and viewing 62

3.4. DISCUSSION 67

CHAPTER 4: COMPARATIVE GENOME ANALYSIS OF MYCOPLASMA SPP ISOLATED FROM SOUTH AFRICAN POULTRY 71

4.1. INTRODUCTION 71

4.2. MATERIALS AND METHODS 72

4.2.1. Sample collection, isolation and identification 72

4.2.2. Quality control 73

4.2.3. Sequence assembly 73

4.2.4. Genome comparison 75

4.3. RESULTS 76

4.3.1. Whole genome sequencing and assembly 76

4.3.2. Genome annotation 81

4.3.3. Draft genome assembly comparison 87

4.3.4. Candidate gene identification 95

4.4. DISCUSSION 108

CHAPTER 5: CONCLUSION AND FUTURE PERSPECTIVES 114

APPENDIX A: RESULTS 117

A.1. DNA CONCENTRATION OF ISOLATED MYCOPLASMA DNA 117

A.2. GEL ELECTROPHORESIS 122

APPENDIX B: SEQUENCE ALIGNMENTS 125

B.1. M. GALLINARUM VLPD GENE ALIGNMENT 125

B.2. M. GALLINARUM RIBOSOMAL PROTEIN L4 ALIGNMENT 126

REFERENCES 127

ANIMAL ETHICS COMMITTEE 143

LIST OF FIGURES

Figure 1-1: a) Cell morphology of mycoplasma as viewed on transmission electron microscope. b) “Fried egg” morphology of mycoplasma on an agar plate. Image used with permission (Citti and Blanchard, 2013) 3

Figure 1-2: Left-hand side depicts targets of antimicrobials in a cell and right-hand side depicts mechanisms of antimicrobial resistance. Reused under Creative Commons Attribution 2.0 Generic (CC BY 2.0) License (Wright, 2010). 10

Figure 1-3: Bacterial horizontal gene transfer. a) Naked DNA is released during cell lysis and taken up by another bacterial cell though transformation. b) Bacteriophages transfer genes between different bacterial cells during bacterial transduction. c) Two bacteria from a mating bridge resulting in exchange of genetic material through conjugation. Figure used with permission from Furuya and Lowy (2006) 12

Figure 1-4: Circular representation of the MG strain R_{low} genome. Used with permission from Papazisi et al. (2003)..14

Figure 1-5: Workflow for DNA sequencing. DNA extraction from various sources, fragmented and amplified by PCR. The DNA fragments are then separated and immobilized depending on sequencing technology used and then sequenced in parallel. Used with permission from (Goldman and Domschke, 2014). 18

Figure 1-6: DNA amplification during library preparation for a) Ion Torrent sequencing using emulsion PCR and b) Illumina sequencing using Solid-phase bridge amplification adapted with permission from (Goodwin et al., 2016b) 19

Figure 1-7: Reversible terminator sequencing of Illumina sequencing. A mixture containing primers, DNA polymerase and fluorophore-labelled terminally blocked nucleotides are added to the DNA template. Only one nucleotide is added per cycle, due to the terminal block group. Each of the four nucleotides have different cleavable fluorophores attached that will emit different colours using fluorescent microscopy. The fluorophore is then cleaved, and the blocked nucleotide repaired before restarting the cycle for attachment of the next nucleotide. Adapted with permission from (Metzker, 2010, Goodwin et al., 2016b) 20

Figure 1-8: Ion Torrent sequencing using semiconductor-based detection. Beads containing amplified DNA template are arrayed into a microtiter plate with only one bead per well. One nucleotide type is added at a time. When a nucleotide binds to the DNA a single H^+ ions is released resulting in a change in the pH of 0.02 units. This change is detected by a metal-oxide semiconductor and ion-sensitive field-effect transistor device. Unbound nucleotides are washed away, and the next nucleotide type is added. Adapted with permission from (Goodwin et al., 2016b) 20

Figure 2-1: Mycoplasma identification in cell culture by growth inhibition with hyperimmune sera. 35

Figure 2-2: Phylogenetic tree of the 16S rRNA gene of avian mycoplasmas, representatives of the isolates from this study are shown. Values on branch refer to maximum parsimony, Bayesian inference and maximum likelihood bootstrap values, respectively. Bootstrap values below 50 not shown. 37

Figure 2-3: Mycoplasma identification by 16S rRNA gene identification. 38

Figure 2-4: Mycoplasma identification in South African provinces. 38

Figure 2-5: Euler diagram showing multi-infection in mycoplasma samples.........................39

Figure 2-6: Mycoplasma species isolated per time period. Top (dark coloured) values of each column is representative of axenic cultures and bottom (light coloured) sections of columns are indicative of mixed sample cultures. 39

Figure 2-7: Antibiotic sensitivity of mycoplasma strains to chlortetracycline, enrofloxacin, tylosin and tiamulin. In (a) total number of strains that are either resistant, susceptible or in the intermediate range is depicted, with relative proportions given in (b). 48

Figure 3-1: Workflow for manual joining of contigs using the Genome finishing tool of CLC Genomics Workbench 57

Figure 3-2: Genome annotation pipeline that forms part of the standard operating procedures of IMG genome analysis which include i) a quality control step, ii) structural annotation of genome and iii) functional annotation of genome 58

Figure 3-3: Per base sequence quality as presented by FastQC quality control. A) Ion Torrent sequencing data B) Paired and trimmed Illumina sequencing data C) Untrimmed forward Illumina reads file D) Untrimmed reverse Illumina reads file. The y-axis shows the quality scores and is divided into a green, orange and red block representing very good, reasonable and poor quality calls, respectively. The x-axis the position in a read bp. The box-whisker plot represents each nucleotide position with the central red line indicating the median value, the yellow box representing the inter-quartile range (25-75%), the upper and lower whiskers representing the 10 and 90% range and the blue line indicates the mean quality across each base position...60

Figure 3-4:COG categories of annotations generated by the a) IMG-MGAP (2017), b) NCBI-PGAP (2017) and c) NCBI-PGAP (2016)..64

Figure 3-5: Correlation of COG categories produced by IMG and NCBI-PGAP pipelines. COG categories found in M. pullorum are C - Energy production and conversion, D - Cell cycle control, cell division, chromosome partitioning, E - Amino acid transport and metabolism, F - Nucleotide transport and metabolism, G - Carbohydrate transport and metabolism, H - Coenzyme transport and metabolism, I - Lipid transport and metabolism, J - Translation, ribosomal structure and biogenesis, K - Transcription, L - Replication, recombination and repair, M - Cell wall/membrane/envelope biogenesis, N - Cell motility, O - Posttranslational modification, protein turnover, chaperones, P - Inorganic ion transport and metabolism, R - General function prediction only, S - Function unknown, T - Signal transduction mechanisms, U - Intracellular trafficking, secretion, and vesicular transport, V - Defense mechanisms and X - Mobilome: prophages, transposons 65

Figure 3-6: *M. pullorum* strain B359-6 complete genome map visualised in Gview. Rings represent the following starting from the inner ring 1) genome position in genomes, 2) GC content on forward (green) and reverse (purple) strand of genomes 3) Genes on the positive strand (blue) 4) Genes on negative strand (red) and 5) Protein coding genes using the COG functional categories. 66

Figure 4-1: Representation of aligning contigs to reference in using the “align contigs” tool of the CLC Genomics Workbench. The table in the top section lists the contig matches to the reference genome with metrics used to evaluate which contigs to join. The bottom section is a linear representation of how the contigs align to the reference. Green sequences indicate that the contig aligned to the reference in a forward direction and red sequence indicate that the contig aligned in the reverse direction. 75

Figure 4-2: Amount of missamblies compared to reference genome as measured by Quast for each strain analysed for a) *Mycoplasma gallisepticum* b) *M. synoviae* c) *M. gallinaceum* and d) *M. pullorum*. 85

Figure 4-3: Comparison of RAST SEED category classification for the annotated CDS produced by the RAST annotation pipeline for each strain in each species. 90

Figure 4-4: Comparison of RAST seed category classification for annotated CDS of strains as percentage of the total amount of classified CDS per strain. 91

Figure 4-5: Comparison of RAST seed category classification for annotated CDS of strains excluding the proteins classified as hypothetical for a better view of the protein categories with assigned protein function. 92

Figure 4-6: Comparison of RAST seed category classification for annotated CDS of strains as percentage of the total amount of classified CDS per strain, excluding the hypothetical protein category to for better definition of other categories 93

Figure 4-7: Linear view of locally collinear block (LCB) colour scheme alignment of *M. synoviae* strains as determined by the progressiveMauve algorithm of the Mauve software. Coloured blocks represent segments of conserved DNA between the strains that are internally free from genome arrangements. Black and white blocks represent the genes identified in each strain. Blocks facing upwards are found on the sense strand and blocks facing downwards are found on the antisense strand. Strains in order are a) MS strain 53 b) B2214-07 c) B1064-14-H3 d) B1064-14-H4 e) B1064-14-H5 f) B1393-14-10 g) B1394-14-2 h) B1394-14-5 i) B458-15-11 j) B458-15-5M and k) B458-15-6. 96

Figure 4-8: Linear view of locally collinear block (LCB) colour scheme alignment of *M. pullorum* strains as determined by the progressiveMauve algorithm of the Mauve software. Coloured blocks represent segments of conserved DNA between the strains that are internally free from genome arrangements. Black and white blocks represent the genes identified in each strain. Blocks facing upwards are found on the sense strand and blocks facing downwards are found on the antisense strand. Strains in order are a) B359-15-6 b) B2096-14-3 c) B293-15-12 d) B293-15-13 e) B293-15-15 f) B293-15-17 and g) B359-15-15. 97

Figure 4-9: Linear view of backbone colour scheme of progressiveMauve alignment algorithm of *M. synoviae* strains. Mauve coloured lines indicate conserved regions present in all aligned strains and other coloured regions indicate conserved regions only present in some of the strains. Strains in order are a) MS strain 53 b) B2214-07 c) B1064-14-H3 d) B1064-14-H4 e) B1064-14-H5 f) B1393-14-10 g) B1394-14-2 h) B1394-14-5 i) B458-15-11 j) B458-15-5M and k) B458-15-6. 98

Figure 4-10: Linear view of backbone colour scheme of progressiveMauve alignment algorithm of *M. pullorum* strains. Mauve coloured lines indicate conserved regions present in all aligned strains and other coloured regions indicate conserved regions only present in some of the strains. Strains in order are a) B359-15-6 b) B2096-14-3 c) B293-15-12 d) B293-15-13 e) B293-15-15 f) B293-15-17 and g) B359-15-15. 99

Figure 4-11: BLAST matrix comparing *M. gallisepticum* strains. Green blocks depict homology between strains, and red blocks depict homology (paralogs) within each strain. Homology is determined by pairwise alignment and significance of the alignment is dependent on a 50% identify between the proteins as well as at least 50% of the longest proteins has to be aligned (Vesth et al., 2013). 100

Figure 4-12: BLAST matrix comparing *M. synoviae* strains. Green blocks depict homology between strains, and red blocks depict homology (paralogs) within each strain. Homology is determined by pairwise alignment and significance of the alignment is dependent on a 50% identify between the proteins as well as at least 50% of the longest proteins have to be aligned (Vesth et al., 2013). 101

Figure 4-13: BLAST matrix comparing *M. pullorum* strains. Green blocks depict homology between strains, and red blocks depict homology (paralogs) within each strain. Homology is determined by pairwise alignment and significance of the alignment is dependent on a 50% identify between the proteins as well as at least 50% of the longest proteins has to be aligned (Vesth et al., 2013). 102

Figure 4-14: Pan genome analysis of *M. gallisepticum* depicting the core and pan genome of the strains. 103

Figure 4-15: Pan genome analysis of *M. synoviae* depicting the core and pan genome of the strains. 103

Figure 4-16: Pan genome analysis of *M. pullorum* depicting the core and pan genome of the strains. 104

Figure 4-17: Pan genome analysis of *M. gallinaceum* depicting the core and pan genome of the strains. 104

Figure 4-18: Pan genome analysis of all species depicting the core and pan genome of the all strains tested. 105

Figure 4-19: COG categories assigned to core genes for a) *M. gallisepticum* strains, b) *M. synoviae* strains and c) *M. pullorum* strains..106

Figure 4-20: COG categories assigned to core genes for a) *M. galllinaceum* strains and b) all the strains combined. 107

LIST OF TABLES

Table 1-1: Mycoplasma species found in avian species. Table used with permission from Ferguson-Noel (2013) and adapted. 4

Table 1-2: Comparison of general characteristics of poultry mycoplasma........ 15

Table 1-3: Characteristics, strengths and weaknesses of commonly used sequencing platforms[a] (Besser et al., 2017). 17

Table 2-1: South African Mycoplasma species analysed in this study........ 27

Table 2-2: Position of nucleotide and amino acid substation of *Mycoplasma* species 34

Table 2-3: Correlation between Mycoplasma identification methods in axenic samples. 40

Table 2-4: MIC's and resistance mutations of Mycoplasma strains. 43

Table 2-5: Minimum Inhibitory Compound breakpoints. 46

Table 2-6: Minimum inhibitory compound values (µg/ml) distribution of Mycoplasma strains........ 47

Table 3-1: Basic read statistics of sequencing files collected from FastQC quality control statistics and CLC quality control statistics 59

Table 3-2: *De novo* assembly statistics of datasets used in this study generated in Quast........ 61

Table 3-3: General features produced by NCBI-PGAP and IMG-MGAP 63

Table 3-4: Table of correlation between COG categories of proteins found in the annotations generated 65

Table 4-1: Strain sequencing information........ 78

Table 4-2: Summary of comparative analysis of genome assembly to reference genomes as determined by Quast. Colour scale indicates best (blue) and worst (red) values of each measure in each species‡. 82

Table 4-3: Annotation results for draft mycoplasma genomes. Colours rank values in column from highest to lowest for each species 88

Table 4-4: List of core genes for all mycoplasma species with the *M. gallinaceum* protein accession number. 105

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1. Introduction

The South African poultry industry in 2016 grossed over R 48 billion from meat and egg production which is 18% of the total agricultural income and 39% of the animal product income of South Africa. It is estimated that 47 000 people are directly employed and 59 000 people are indirectly employed by the South African poultry industry (SAPA, 2016, DAFF, 2017). One of the biggest contributing factors that influence the productivity and viability of this industry is disease. A common respiratory disease found in poultry is mycoplasmosis which causes a reduction in the rate of growth of broilers, egg production and hatchability in layers and breeders, as well as carcass downgrading. Along with the influence on production yields, the costs involved in biosecurity measures, disease screening and diagnosis, treatment of this disease and vaccination strategies have made mycoplasmosis an economically important disease of poultry (Bradbury, 2005).

Mycoplasmosis is mostly associated with respiratory disease and symptoms can include coughing, rales, airsacculitis, depression and swollen joints (Whithear, 1993, Stipkovits and Kempf, 1996). Mycoplasmosis is caused by species from the genus *Mycoplasma*, with *M. gallisepticum* (MG), and *M. synoviae* (MS) considered the most economically important pathogenic mycoplasmas infecting chickens and being listed by the World Organisation for Animal Health (OIE) as notifiable avian diseases agents. MG and MS are usually identified using culture with growth inhibition, serological and immunological methods, such enzyme-linked immunosorbent assays (ELISA) (Kleven, 2008), and polymerase chain reaction (PCR) and real-time PCR (qPCR) in recent years (Kleven, 2008). Little research is available on mycoplasma species considered as “non-pathogenic” commonly found in poultry, such as *M. gallinarum* and *M. gallinaceum*. Standard assays to identify these species are not common but 16S ribosomal ribonucleic acid (rRNA) and 16S-23S rRNA intergenic spacer region (IGSR) species-specific PCR have been used to distinguish these species.

Treatment of mycoplasmas include the use of antibiotics, but the standard practise of use of antimicrobial agents in feed, not only for treatment but also for enhanced growth and improved production worldwide, including in the South African poultry industry, has contributed to the growing international concern about antimicrobial resistance (AMR). How AMR is acquired is currently gaining more interest in research, including the possibility of acquired resistance from other bacteria.

In this study full genome sequencing of numerous mycoplasma species isolated from South African poultry farms were analysed for acquired antimicrobial resistance, and possible future diagnostic and vaccine targets.

1.2. Mycoplasma

1.2.1. General characteristics

Mycoplasma forms part of the class *Mollicutes* from the Latin words "*mollis*" and "*cutis*" meaning "soft" and "skin" respectively, which contains some of the smallest known bacteria (Bradbury, 2005). Species from this class are characterised by the absence of a bacterial cell wall; instead, it is encapsulated by a thin tri-laminar cell membrane, consisting mainly of 60 to 70% proteins and 20 to 30% lipids (Bradbury, 2005, Razin and Hayflick, 2010). There are more than 200 proteins in the membrane which play a role in antigenic variation, host cell adhesion, motility and transport of nutrients (Raviv and Ley, 2013). Due to the lack of a bacterial cell wall, mycoplasmas are resistant to antibiotics that target the cell wall, such as penicillin, but this also comes at a cost as these bacteria are sensitive to osmotic shock and detergents (Razin and Hayflick, 2010, Bradbury, 2005). Mycoplasmas have various cell shapes, including pear-shaped, filaments, flask-shaped filaments and spheres. Spherical shapes (Figure 1-1a) ranging in diameters of between 0.3 and 0.8 µm, are the most prevalent of these, but it has been noted that many of the pathogenic species are flask-shaped with a terminal tip structure that can aid in motility and host cell interaction (Razin and Hayflick, 2010, Baseman and Tully, 1997). Mycoplasmas with sizes of $5\text{x}10^8$ Da with less than 300 genes have been recorded, can pass through filters that normally block bacteria, and are also regularly found as contaminants of eukaryotic cell lines (Stipkovits and Kempf, 1996).

Mycoplasmas evolved from low G+C content, Gram-positive bacteria through reductive evolution, retaining the genes involved in replication and sacrificing the genes involved in biosynthesis and cell wall synthesis (Bradbury, 2005). Reductive evolution resulted in some interesting characteristics of mycoplasma, such as 1) a low G+C content of between 23 and 40%, 2) the requirement of sterols for growth and membrane function, as well as a host for many of their nutrients 3) very small genomes, with sizes ranging from 580 to 1350 kb, 4) a modification in the mycoplasma genetic code, where the UGA codon is translated to a tryptophan instead of a stop codon (Baseman and Tully, 1997, Bradbury, 2005, Ferguson-Noel, 2013) and 5) mycoplasmas are slow growing and difficult to cultivate, requiring specialised media. Mycoplasmas can take a few days to grow and form colonies characteristically shaped like a "fried egg" (Figure 1-1b) (Citti and Blanchard, 2013, Razin and Hayflick, 2010).

There are currently over 200 mycoplasmas, infecting a wide variety of hosts from humans, mammals, and reptiles to fish, plants and even insects (Razin and Hayflick, 2010). Most mycoplasma species are host specific, but some species have been found in multiple hosts (Pitcher and Nicholas, 2005). *M. bovis*, *M. agalactiae* and *M. mycoides* are known to infect sheep, goats and cattle; *M. gateae* infects both cats and dogs and *M. gallisepticum* has been found in various bird species, just to name a few (Pitcher and Nicholas, 2005).

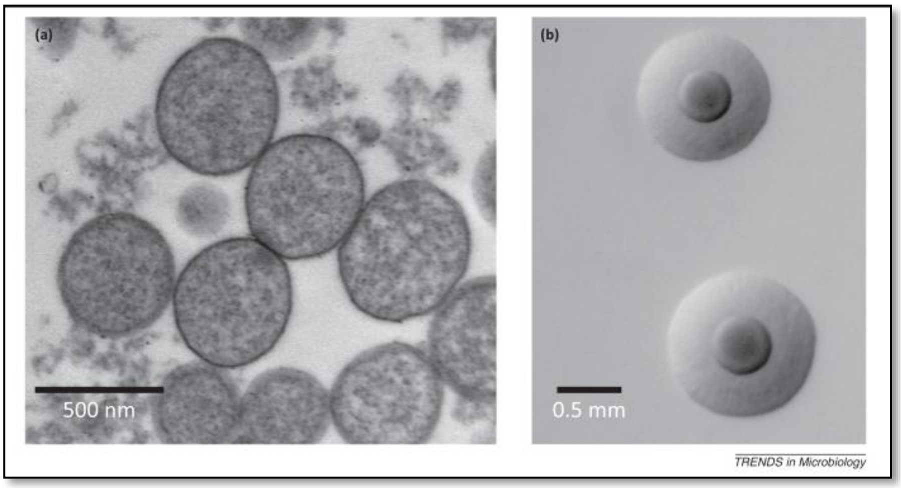


Figure 1-1: a) Cell morphology of mycoplasma as viewed on transmission electron microscope. b) "Fried egg" morphology of mycoplasma on an agar plate. Image used with permission (Citti and Blanchard, 2013)

1.2.2. Poultry mycoplasmas

Currently, there are over 23 avian mycoplasmas found in avian species (Table 1-1), four of which are considered as economically important pathogens; namely MG, MS, *M. iowae* and *M. meleagridis*. The latter two are primarily important in the turkey industry. MG and MS are the only species listed by the OIE as notifiable avian diseases of chickens and research has been aimed mainly at these two pathogens.

MG is considered economically important due to a reduction in weight gain (20-30%) and feed-conversion (10-20%) and an increase in mortality rate (5-10%) and carcass downgrading (10-20%) in broilers, as well as reduced egg production (10-20%) and an increase in embryo mortality (5-10%) in both breeders and layer (Stipkovits and Kempf, 1996). The economic impact of MS infection results from a decrease in growth rate and egg production (5-10%), egg hatchability (5-7%) and an increase in chic mortality rate (5%) as well as carcass condemnation (1.88%-15%) (Stipkovits and Kempf, 1996, King et al. 1973, Sentíes-Cué et al. 2005).

Most mycoplasma species were discovered between 1960 and the 1990's, and apart from MG and MS, little research has been done on the other mycoplasma species isolated from chickens since they were first described. For most of these species, only the basic biochemical characteristics, used at the time of discovery for description of a new species, is known, such as its ability to ferment glucose, or hydrolyse arginine (Table 1-1). Other mycoplasmas species commonly isolated from chickens include *M. gallinaceum; M. gallinarum, M. glycophilum, M. iners, M. iowae, M. lipofaciens*, and *M. pullorum* (Ferguson-Noel, 2013). Other species have also been isolated from chickens, such as *M. cloacale* and *M. imitans*, but are not commonly found in chickens (Benčina et al., 1987). *M. imitans* is an interesting case, as it is closely related to MG and could easily be misdiagnosed as MG, therefore the true occurrence of *M. imitans* is not known. These species are generally considered as non-pathogenic.

Table 1-1: Mycoplasma species found in avian species. Table used with permission from Ferguson-Noel (2013) and adapted.

Species	Main host	Energy source	Reference
M. anatis	Duck, goose	Ferment glucose	(Roberts, 1964)
M. anseris	Goose	Hydrolyse arginine	(Bradbury et al., 1988)
M. buteonis	Buzzard, Buteo hawk	Ferment glucose	(Poveda et al., 1994)
M. cloacale	Turkey, goose	Hydrolyse arginine	(Bradbury and Forrest, 1984)
M. columbinasale	Pigeon	Hydrolyse arginine	(Jordan et al., 1982)
M. columbinum	Pigeon	Hydrolyse arginine	(Shimizu et al., 1978)
M. columborale	Pigeon	Ferment glucose	(Shimizu et al., 1978)
M. corogypsi	Vulture	Ferment glucose	(Panangala et al., 1993)
M. falconis	Falcon	Hydrolyse arginine	(Poveda et al., 1994)
M. gallinaceum	Chicken, pheasant, partridge	Ferment glucose	(Jordan et al., 1982)
M. gallinarum	Chicken, turkey	Hydrolyse arginine	(Freundt, 1955)
M. gallisepticum	Chicken, turkey, pheasant, partridge, songbird, etc	Ferment glucose	(Edward and Kanarek, 1960)
M. gallopavonis	Turkey	Ferment glucose	(Jordan et al., 1982)
M. glycophilum	Chicken, pheasant, partridge	Ferment glucose	(Forrest and Bradbury, 1984)
M. gypis	Vulture	Hydrolyse arginine	(Poveda et al., 1994)
M. imitans	Goose, duck, partridge	Ferment glucose	(Bradbury et al., 1993)
M. iners	Chicken, turkey, pheasant, partridge	Hydrolyse arginine	(Edward and Kanarek, 1960)
M. iowae	Turkey, chicken	Both	(Jordan et al., 1982)
M. lipofaciens	Chicken, turkey	Both	(Bradbury et al., 1983)
M. meleagridis	Turkey	Hydrolyse arginine	(Yamamoto et al., 1965)
M. nasistruthionis	Ostrich	?	(Botes et al., 2005, Langer, 2009)
M. pullorum	Chicken, pheasant, partridge	Hydrolyse arginine	(Jordan et al., 1982)
M. sturni	Starling	Ferment glucose	(Forsyth et al., 1996)
M. struthionis	Ostrich	?	(Botes et al., 2005, Langer, 2009)
Ms02	Ostrich	?	(Botes et al., 2005)
M. synoviae	Chicken, turkey	Ferment glucose	(Olson et al., 1964, Jordan et al., 1982)

1.2.3. Role of poultry mycoplasmas in disease

MG causes respiratory disease in broiler, breeder and commercial layer chickens, symptoms include, airsacculitis, conjunctivitis, coughing, nasal discharge, rales and sinusitis (Whithear, 1993, Stipkovits and Kempf, 1996). MG is also associated with chronic respiratory disease (CRD) when it forms part of a multifactorial disease complex with other pathogens, usually a respiratory virus, such as Newcastle disease virus (NDV) or avian infectious bronchitis virus (IBV) and other bacteria species, of which *Escherichia coli* is the most common (Stipkovits and Kempf, 1996, Raviv and Ley, 2013). MS causes the same respiratory symptoms as MG but is also associated with synovitis; symptoms include depression, swelling in the hock joint and bursae and also pale face and comb (Stipkovits and Kempf, 1996, Whithear, 1993). MS has also been involved in multifactorial disease complexes, mostly with Newcastle disease virus and IBV (Stipkovits and Kempf, 1996).

Mycoplasmas have adapted various efficient ways to cause persistent infection with multifactorial pathogenesis mechanisms, not all of which are known or properly understood yet. Mycoplasmas can be transmitted by direct contact through inhalation or vertically from hen to chick through eggs and target the epithelial surfaces of the respiratory tract (Umar et al., 2017, Raviv and Ley, 2013). MG can also enter through the conjunctiva and MS can also target the joints (Bradbury, 2005). MG has a terminal tip structure that through gliding motility aids in reaching the target tissue and is one of the first methods of evading the host immune system, by bypassing the mucociliary clearance mechanism of the host (Indikova et al., 2014). Indikova et al. (2014) found through mutation studies that the *Mgc2*, glyceraldehyde-3-phosphate dehydrogenase A (*gapA*) and cytokine response modifier A (*crmA*) genes play a role in gliding motility and MG morphology. Furthermore, co-expression of these has been shown to play a role in cytadherence and colonization of host tissue as well aiding in evading the host immune response through phase variation (Papazisi et al., 2002).

Various other genes encoding proteins that play a role in cell adherence, adaption to the host environment, and systemic infection as well as in evading the host immune system have been found. These include the large pMGA gene family; renamed variable lipoprotein haemagglutinin A (*vlhA*) by Papazisi et al. (2003), and the phase-variable putative adhesin (*pvpA*) gene (Raviv and Ley, 2013, Papazisi et al., 2003). The *vlhA* gene family of MG plays a role in attachment to host cell as well as evading the host immune system by phase variation (Papazisi et al., 2003). In MG this gene family is made up of 43 genes, but most of the time only one protein is expressed at a time, but some studies have shown low levels of expression of a second or third *vlhA* gene. This expression is controlled by DNA slippage, a molecular switch that influences the length of an upstream GAA repeat region (Citti et al., 2010).

The *mga1142* gene was first thought to encode the osmotically inducible protein C (*OsmC*)-like adhesion but was later shown to encode an organic hydroperoxide resistance (Ohr) protein (Jenkins et al., 2007, Jenkins et al., 2008). This gene is a cell surface protein shown to play a possible role in adhering to and invading host cells, as well as evasion of the host immune system through its ability to detoxify the peroxidase immune response (Jenkins et al., 2007, Jenkins et al., 2008). Mycoplasmas can produce hydrogen peroxidase which is hypothesized to not only play a role in host cell entry by compromising the host cell membrane, but also play a role in activation of genes for adhesion and invasion and SpxA protein (Matyushkina et al., 2016). Matyushkina et al. (2016) showed that the SpxA protein plays a large role in the adaption of mycoplasma to the intracellular environment of the host cells. The cysteine proteases gene (*cysP*) encodes CysP that have been shown to digest chicken immunoglobulin G (IgG) (Cizelj et al., 2011). There are numerous other genes that play a possible role in pathogenesis of mycoplasmas, and with advances in the technologies and genomic information more genes are expected to be characterised, each one a step closer to fully understanding the complex nature of mycoplasma pathogenesis.

Some MG strains also have the ability to form a biofilm that plays a role in adapting to the host environment and causes systemic infection (Wang et al., 2017). Wang et al. (2017) identified 10 genes that play a possible role in biofilm production; the genes identified encode the following proteins: S-adenosylmethionine synthetase, phosphomannomutase, ABC transporter ATP-binding protein, ABC transporter permease, *vlhA* gene family, methionyl-transfer RNA (tRNA) synthetase, phosphotransferase system (PTS) fructose-specific enzyme EIIABC, pyruvate dehydrogenase E1 subunit alpha, enolase and an unknown conserved hypothetical membrane protein.

Of the species considered as non-pathogenic, *M. gallinaceum* and *M. gallinarum,* are most frequently isolated from poultry flocks. An association study by Welchman et al. (2002) found some correlation between *M. gallinaceum* and conjunctivitis as well as airsacculitis in pheasants (Welchman et al., 2002). However, Adeyemi et al. (2018) showed that *M. gallinaceum* is non-pathogenic in chickens but can aggravate disease caused by IBV as well as increase its proliferation (Adeyemi et al., 2017). The potential to become pathogenic in the presence of a respiratory virus, such as IBV were shown for both *M. gallinarum* and *M. imitans* (Kleven et al., 1978, Ganapathy and Bradbury, 1999). The pathogenic nature of *M. pullorum, M. iner* and *M. lipofaciens* in chicken embryos have been shown (Moalic et al., 1997, Wakenell et al., 1995, Lierz et al., 2007). *M. iowae* is considered highly pathogenic in turkeys and has been found in chickens and is furthermore associated with increase mortality in chicken embryos (Bradbury and McCarthy, 1983).

1.2.4. Diagnosis of poultry mycoplasma

There are several ways to diagnose mycoplasmas ranging from culture to deoxyribonucleic acid (DNA) methods using mucosal swabs, blood or tissue samples (Feberwee et al., 2005). The OIE compiled a manual of internationally accepted laboratory methods for the diagnosis of various diseases, including mycoplasmas, described briefly below.

Isolation and culture

Mycoplasmas are difficult to cultivate, requiring media rich in proteins and between 10-15% animal serum, from swine, horse or avian species, there are species that have additional requirements, such as MS, that require nicotinamide adenine dinucleotide (NAD) for growth (Ferguson-Noel, 2013). Some mycoplasma species are slow growing, MG generally takes between 3 and 10 days, at 37 to 38˚C to form colonies, but can take up to 3 weeks to grow (Raviv and Ley, 2013, Ferguson-Noel, 2013). Other species such as *M. gallinarum* and *M. gallinaceum* are fast growing, where colonies can appear after one day, easily overgrowing MG making a proper diagnosis of MG difficult (Raviv and Ley, 2013).

Culturing mycoplasmas on agar instead of in broth allows time for the slow-growing mycoplasmas to form colonies, aiding in the diagnosis of MG (OIE, 2008). Phenol red is normally used in broth cultures to grow MG and MS, because some mycoplasmas can ferment glucose as a source of energy, which results in acid production and a drop in pH, causing the phenol red to change to an orange or yellow colour (Ferguson-Noel, 2013). Mycoplasma species that hydrolyse arginine as a source of energy do not cause a colour change and can prevent this pH change if mixed infections are present (OIE, 2008). If no colour change is observed after 7 to 10 days, samples are plated onto agar for MG and MS identification, see table 1 for a list of avian mycoplasma species that either ferment glucose or hydrolyse arginine (Ferguson-Noel, 2013, OIE, 2008). The most sensitive method to identify mycoplasma by culture is to first incubate in broth followed by plating onto an agar plate (Raviv and Ley, 2013). Biochemical reactions can also be used to identify mycoplasma species, but this method is not specific enough and not widely used (OIE, 2008).

After cultivation mycoplasma species can be identified using methods that target the immunological properties of mycoplasma and include direct and indirect immunofluorescence assays with or without immunoperoxidase, growth inhibition and metabolism inhibition (Raviv and Ley, 2013). Growth inhibition test identifies MG or MS using species-specific hyperimmune monoclonal anti-serum but is most effective when pure cultures are used (OIE, 2008). Immunofluorescence techniques use polyclonal antibodies prepared in rabbits for simpler, faster, more sensitive and specific identification of mycoplasma species and can be used when mixed infections are present (OIE, 2008).

1.2.5. Serological methods

Serological methods of mycoplasma identification include the commonly used ELISA, haemagglutination inhibition (HI), rapid serum agglutination (RSA) and lesser known microimmunofluorescence-assay and radioimmunoassay (OIE, 2008). Even though the latter two tests have high sensitivity, low specificity of RSA and time constraints of HI has resulted in ELISA as the serological test of choice. Numerous commercial ELISAs are available, with high sensitivity and specificity, but non-specific reactions can influence the results (Levisohn and Kleven, 2000). The use of antimicrobials can cause a delay in the immune response of the chicken serological response and the use of vaccines requires tests that can distinguish vaccine strains from field strains (Levisohn and Kleven, 2000). These serological assays are widely used for flock monitoring programs, rather than individual diagnosis (OIE, 2008).

Although the above-described culture and serological methods are valuable research tools, they have mostly been replaced by DNA based methods which are generally less time-consuming, more accurate and more sensitive (OIE, 2008).

1.2.6. DNA based methods

DNA based methods are widely used to identify mycoplasma species as well as distinguish different strains within a species. Hybridization using DNA or rRNA gene probes can be used to detect MG or MS effectively, even though this method is faster than culture methods, it is a very difficult assay and was quickly replaced by simpler, more sensitive PCR-based methods (Razin, 1994). PCR methods include the 1) conventional PCR, which is the amplification of a target DNA sequence; 2) reverse transcription PCR (rt-PCR), which is the amplification of complement DNA (cDNA) after reverse transcription from RNA of expressed gene; 3) real-time PCR (qPCR), which monitors normal PCR to quantify the amount of DNA molecules using fluorescent dyes or probes, as well as 4) multiplex PCRs, where multiple primers sets are used in either a single conventional or q-PCR to identify multiple sequences simultaneously (Raviv and Ley, 2013). The products of these PCR methods can then be used in various ways to identify mycoplasmas, distinguish between different species of mycoplasmas or even differentiate between different strains within a species, including between field and vaccine strains. The most widely used methods include 1) Electrophoresis of species-specific PCR product directly to identify species of interest 2) Restriction fragment length polymorphism (RFLP), where the PCR product is digested with restrictions enzyme and separated by gel electrophoresis as a DNA fingerprinting technique; and 3) gene-targeted sequencing, where PCR products are sequenced and analysed (Levisohn and Kleven, 2000). Another type of PCR that has been widely used as a DNA profiling technique is random amplification of polymorphic DNA (RAPD), but this method has some reproducibility issues (OIE, 2008). DNA microarrays using various genes, such as the 23S rRNA and *tuf* gene have been used with similar sensitivity to rt-PCR and was able to identify multiple mycoplasma infection in samples, but MG was difficult to recognize (Schnee et al., 2012).

Mycoplasma species differentiation is normally done using the widely accepted bacterial 16S rRNA gene primers, but some exceptions do exist. The most notable of which is the closely related species of MG and *M. imitans* that share high sequence homology in the 16S rRNA gene (Kempf, 1998). RFLP analysis of the 16S rRNA gene PCR product is one method that has been used successfully to differentiate these species. Harasawa et al. (2004) found a putative transposon gene in the 16S-23S rRNA intergenic spacer region (IGSR) that could also be used to distinguish these species (Harasawa et al., 2004). The 16S-23S rRNA IGSR has also been used to differentiate between the other mycoplasma species (Kleven, 2008). Various genes encoding surface proteins have been used with gene-targeted sequencing (GTS) to distinguish between strains of MG and MS, such as the *mgc2, gapA, pvpA* or MGA_0309 genes for MG strains or the *vlhA* gene for MS strains (OIE, 2008, García et al., 2005, Hong et al., 2004). Diagnosis of mycoplasma infection is the first step in the control of mycoplasmas.

1.2.7. Treatment and prevention of poultry mycoplasma

Control and treatment

Mycoplasmas cause chronic infections and are difficult to eradicate. Kleven et al. (2008) divided poultry mycoplasma control into three facets, namely prevention, medication and vaccination. Starting with mycoplasma free stock followed by good biosecurity measures should be the best method of mycoplasma control (Kleven, 2008). However, there are numerous factors that can affect the efficiency of this method, such as mycoplasma infection in nearby poultry farms, especially in the intensive farming areas; human error resulting from failure in executing repetitive steps of biosecurity measures. Thus, maintaining flocks free of mycoplasma infection is difficult, and a good monitoring system is required, which normally includes screening methods using methods discussed in section 1.2.4. In the South African poultry industry, monitoring is done using commercially available MG and MS-specific ELISAs (D.B.R. Wandrag, Personal communication). Antimicrobial agents are used as treatment after mycoplasma infection is confirmed.

Antimicrobial treatment

Mycoplasmas are sensitive to antimicrobials that target protein synthesis, such as macrolides, tetracyclines and pleuromutilin and nucleic acid synthesis, such as fluoroquinolones. Macrolides are a group of natural and semi-synthetic antimicrobials that targets the 50S ribosomal subunit to inhibit protein synthesis, see Figure 1-2 (Guardabassi and Courvalin, 2006). Members of this class have a lactone ring attached to a deoxy sugar. Example of macrolides include tilmicosin, spiramycin, kitasamycin, josamycin, erythromycin and tylosin, the latter is the most popular antimicrobial used in poultry and along with tetracyclines is mostly used for preventing egg transmission and respiratory disease (Guardabassi and Courvalin, 2006, Umar et al., 2017). Tetracyclines which are natural antimicrobials produced by *Streptomyces* spp. is the class of antimicrobials most often used in animal health and includes oxytetracycline and chlortetracycline (Guardabassi and Courvalin, 2006). Antimicrobials from this class target the 30S ribosomal subunit inhibiting protein synthesis (Guardabassi and Courvalin, 2006). Compounds in the pleuromutilin class of antimicrobial agents, such as tiamulin and valnemulin, are mainly produced by *Basidiomycetes*, but some semi synthetic compounds also exist. These compounds also target the 50S ribosomal subunit to inhibit protein synthesis (Guardabassi and Courvalin, 2006). Fluoroquinolones are a group of synthetic compounds that inhibit nucleic acid synthesis by binding to type II (also known as DNA gyrase) and IV topoisomerases. These enzymes play a key role in unzipping DNA for translation (Guardabassi and Courvalin, 2006). This class of antimicrobials is the class that is used the least in animal health, with enrofloxacin as the most widely used. The use of antimicrobials has proven to decrease egg transmission and clinical disease symptoms resulting in lower economic losses (Umar et al., 2017). However, this is only a short-term solution as antimicrobial resistance is a continuous concern and problem, and other long-term methods, such as vaccination are required.

Acquired antimicrobial resistance in poultry mycoplasma

The lowest concentration required of an antimicrobial agent to visibly inhibit growth or impair the metabolism of an organism is known as the minimum inhibitory concentration (MIC) and is used as a measure of the efficacy of an antimicrobial agent. Breakpoints are specific concentrations of antimicrobial agents that are used to classify a bacterial species as sensitive, resistant or intermediately sensitive to the antimicrobial tested. International standards for breakpoints of avian mycoplasmas have not been published yet, however Hannan (2000) published guidelines on assays and suggested breakpoints that can be used to evaluate various antimicrobial agents. Antimicrobial resistance has been observed in chickens against macrolides, tetracyclines and fluoroquinolones in both *in vitro* and *in vivo* studies.

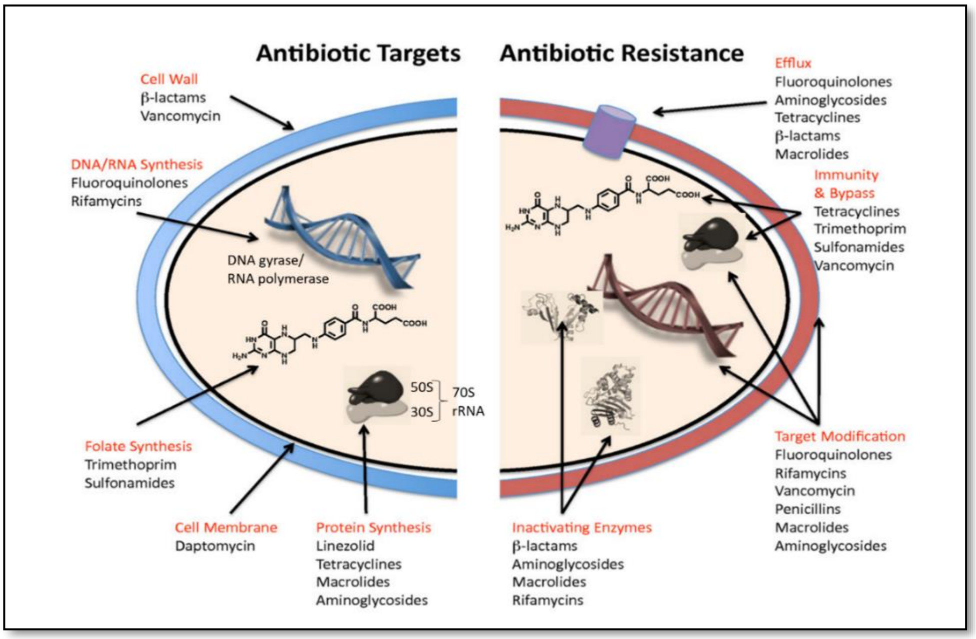


Figure 1-2: Left-hand side depicts targets of antimicrobials in a cell and right-hand side depicts mechanisms of antimicrobial resistance. Reused under Creative Commons Attribution 2.0 Generic (CC BY 2.0) License (Wright, 2010).

The *in vitro* studies done by Zanella et al. (1998) found that resistance developed quicker to erythromycin than to tylosin and enrofloxacin, where resistance developed approximately at the same rate. Resistance to chlortetracycline developed more slowly over time. A similar *in vitro* study was done by Gautier-Bouchardon et al. (2002) and showed similar results with resistance developing quickly to erythromycin and tylosin, however resistance to enrofloxacin developed slower over time, and no resistance to tiamulin and oxytetracycline resistance could obtained in MG or MS. *In vivo* testing for antimicrobial sensitivity found resistance to enrofloxacin, tylosin and tilmicosin in Israeli poultry flocks (Gerchman et al., 2011) and an increase in observed resistance in Jordanian poultry flocks of various antimicrobials including erythromycin, tylosin, enrofloxacin, chlortetracycline, doxycycline, and oxytetracycline (Gharaibeh and Al-Rashdan, 2011).

Bacteria acquire antimicrobial resistance primarily in two ways, namely through mutations in the antimicrobial target (Figure 1-2) or exchange of genetic material between species by horizontal gene transfer (HGT) (Munita and Arias, 2016). Mechanisms of antimicrobial resistance resulting from mutations include, 1) activation of an efflux pump to remove harmful substances from the cell, 2) bypassing the immune response by binding of proteins to antimicrobials preventing binding to target, 3) modification of antimicrobials by enzymes preventing binding to targets or 4) modification of target genes by mutation in genes (Wright, 2010). In poultry mycoplasma species the latter mechanism of antimicrobial has been mostly observed.

Bacterial HGT can occur in one of three methods, 1) bacterial transformation, 2) bacterial transduction or 3) bacterial conjugation (Figure 1-3) (Furuya and Lowy, 2006). Even though transformation is the easiest method, conjugation occurs at a much high rate (Munita and Arias, 2016). Acquired antimicrobial resistance through HGT has not been shown in avian mycoplasmas to date, but it has been observed in other mycoplasma species, of which the most well-known example is the *tetM* gene acquired through conjugation from *Streptococcus* spp. by *M. hominis* (Roberts et al., 1985). However, HGT was hypothesized by Vasconcelos et al. (2005) to be possible between MG and MS that share almost identical genes.

The molecular mechanism of acquired antibiotic resistance has been studied in both field and *in vitro* induced antimicrobial resistant MG and MS mutants. Acquired macrolide resistance was associated with point mutations G2057A, A2058G or A2059G in one or both of the 23S rRNA genes of MG and MS (Ammar et al., 2016, Gerchman et al., 2011, Lysnyansky et al., 2015, Wu et al., 2005). As mentioned, fluoroquinolones inhibit nucleic acid synthesis by binding to the quinolone resistance-determining resistance regions (QRDRs) of DNA gyrase and topoisomerase IV enzymes (Guardabassi and Courvalin, 2006). Both of these enzymes have a tetrameric structure made up of two sets of subunits; DNA gyrase is made up of the GyrA and GyrB subunits, encoded by *gyrA* and *gyrB* genes and topoisomerase IV of an A and B subunits, encoded by *parC* and *parE*, respectively. Enrofloxacin resistance was linked to amino acid substitutions in the of GyrA and GyrB gene and the ParC and ParE genes of MG and MS and in MG it appears as though mutations in GyrA plays the biggest role in acquired fluoroquinolone resistance (Reinhardt et al., 2002a, Reinhardt et al., 2002b, Lysnyansky et al., 2013).

Pleuromutilin resistance was studied by Li et al. (2010) who found a correlation between a combination of point mutations in the 23S rRNA gene at two or more of the following positions: 2058, 2059, 2061, 2447, and 2503. Tetracycline resistance has not been characterised in poultry mycoplasma, however as discovered for human mycoplasmas, the *tetM* gene obtained by HGT confers tetracyclines resistance, and in bovine mycoplasma species point mutations in the 16S rRNA gene have been associated with acquired tetracycline resistance (Roberts et al., 1985, Amram et al., 2014).

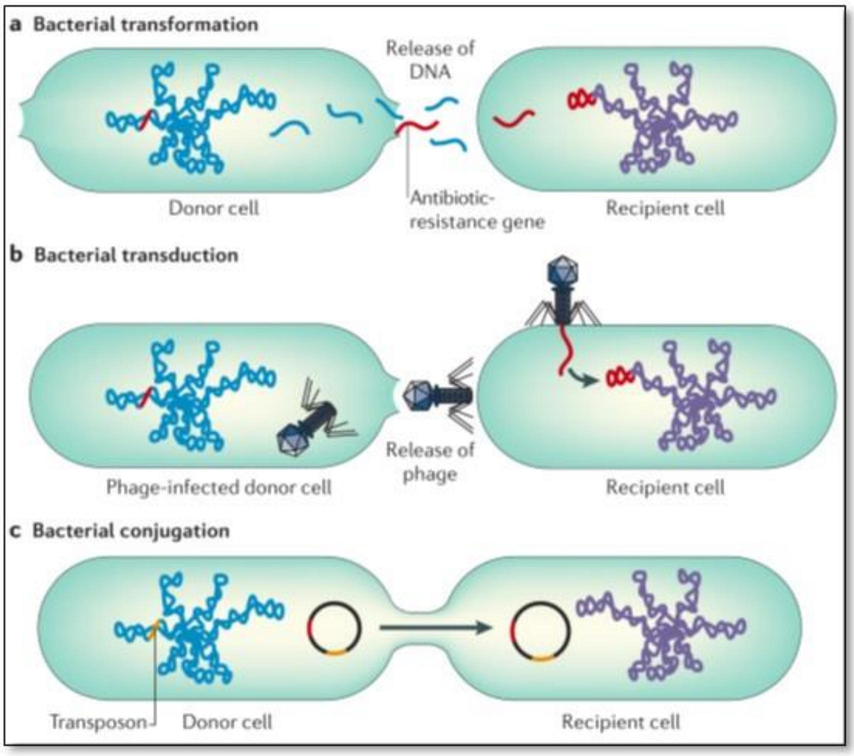


Figure 1-3: Bacterial horizontal gene transfer. a) Naked DNA is released during cell lysis and taken up by another bacterial cell though transformation. b) Bacteriophages transfer genes between different bacterial cells during bacterial transduction. c) Two bacteria from a mating bridge resulting in exchange of genetic material through conjugation. Figure used with permission from Furuya and Lowy (2006)

Vaccination

Various types of vaccines have been developed to protect against mycoplasma infection, mainly against MG, including; 1) inactivated oil-emulsion bacterins, 2) live attenuated vaccines, and 3) recombinant vector vaccines expressing specific MG antigens (Kleven, 2008). The biggest advantage of inactivated oil-emulsion bacterins are that there is no chance of reversion to virulence as non-infectious agents are used making it safer to use than live vaccines, but these vaccines are expensive to produce and must be administered individually to each chicken. Bacterins have been reported to reduce clinical signs of respiratory infection, egg transmission and egg production losses, but variable efficacy has been observed and no protection was observed if vaccinated before 1-2 weeks of age (OIE, 2008).

Commercially available MG live attenuated vaccines are F-strain, ts-11 and MG strain 6/85; but other vaccine strains are also commonly used, such as MG-K strain and MS-H strain (Vaxsafe MS) (Umar et al., 2017). Vaccine strains ts-11 and 6/85 are avirulent, thus safer to use than F-strain, but a lower immune response has also been observed compared to F-strain (Kleven, 2008). F-strain has shown mixed virulence in chickens and can cause infection, persist longer in the upper respiratory tract of the chicken and can be transmitted, but this vaccine is more protective than ts-11 and 6/85 (Kleven, 2008). Although F-strain has been widely used in various countries for decades, this strain was only registered for use in South Africa in 2015 (Bwala et al., 2018).

Recombinant vaccines use avirulent bacteria or virus strains to express antigens from the species of interest to elicit an immune response in the host. Fowlpox-virus (FPV) is currently the most common vector used, such as in the commercially available vaccine recombinant-FPV-MG vaccine that uses the 40k and *mgc* genes (Armour and García, 2014). Vaccine strategies to date have had variable success and research is still ongoing. The rapid advances in molecular techniques will be useful in understanding mycoplasmas better and aid in future strategies of control and prevention.

1.2.8. Mycoplasma genomics

Some general characteristics of mycoplasma genomes have been described above, but another interesting characteristic of mycoplasma genomes is the presence of one, two or three copies of the rRNA genes compared to the five to ten copies generally observed in bacterial species (Dybvig and Voelker, 1996). To date the complete genomes of only five mycoplasmas commonly found in poultry species have been published, namely MG, MS, *M. cloacale, M. gallinaceum* and *M. pullorum*. The latter is an output of this dissertation and is the topic of Chapter 3.

The first complete genome of MG was published by Papazisi et al. (2003) (Figure 1-4), since then 11 more genomes for MG have been completed and deposited in Genbank®; a genetic database of publicly available DNA sequences hosted by the National Centre for Biotechnology Information (NCBI). Eight of these strains were isolated from house finches (Tulman et al., 2012), the remaining four strains, vaccine strain F, virulent strain S6, the virulent, low passage strain Rlow and the attenuated, high passage strain Rhigh were isolated from domestic poultry (Szczepanek et al., 2010, Fisunov et al., 2011). The MG strain genomes range in size from 997 kb to 1012 kb with 31.5% G+C content (Table 1-2). Apart from the genes described above that play a role in pathogenesis and antimicrobial resistance and have been used for diagnostic methods and vaccine development other genes have been described that play a role in transport and metabolism. Membrane associated proteins that play a role in the transport of amino acids (PotE), phosphate (Pts), proteins (SecY) and various other biomolecules, such as the large ABC transporter family mentioned above, have also been identified (Papazisi et al., 2003).

Three MS genomes have been completed with sizes ranging from 799 kb to 846 kb, G+C content of around 28% and three 5S rRNA plus 2 sets each of 16S and 23S rRNA. The complete genomes for most of the remaining non-pathogenic species have not been published, although draft assemblies containing scaffolds or contigs from whole genome shotgun sequencing methods are available on Genbank® for *M. imitans*, *M. glycophilum*, *M. iowae, M. meleagridis, M. cloacale, M. gallinarum, M. lipofaciens* and *M. iners* (Table 1-2). Advances in genomics, proteomics, transcriptomics and metabolomics (also referred to as “omics”), along with a comparative analysis of these three fields will provide a better understanding of these complex pathogens and their interaction with the host, environment and other microorganisms.

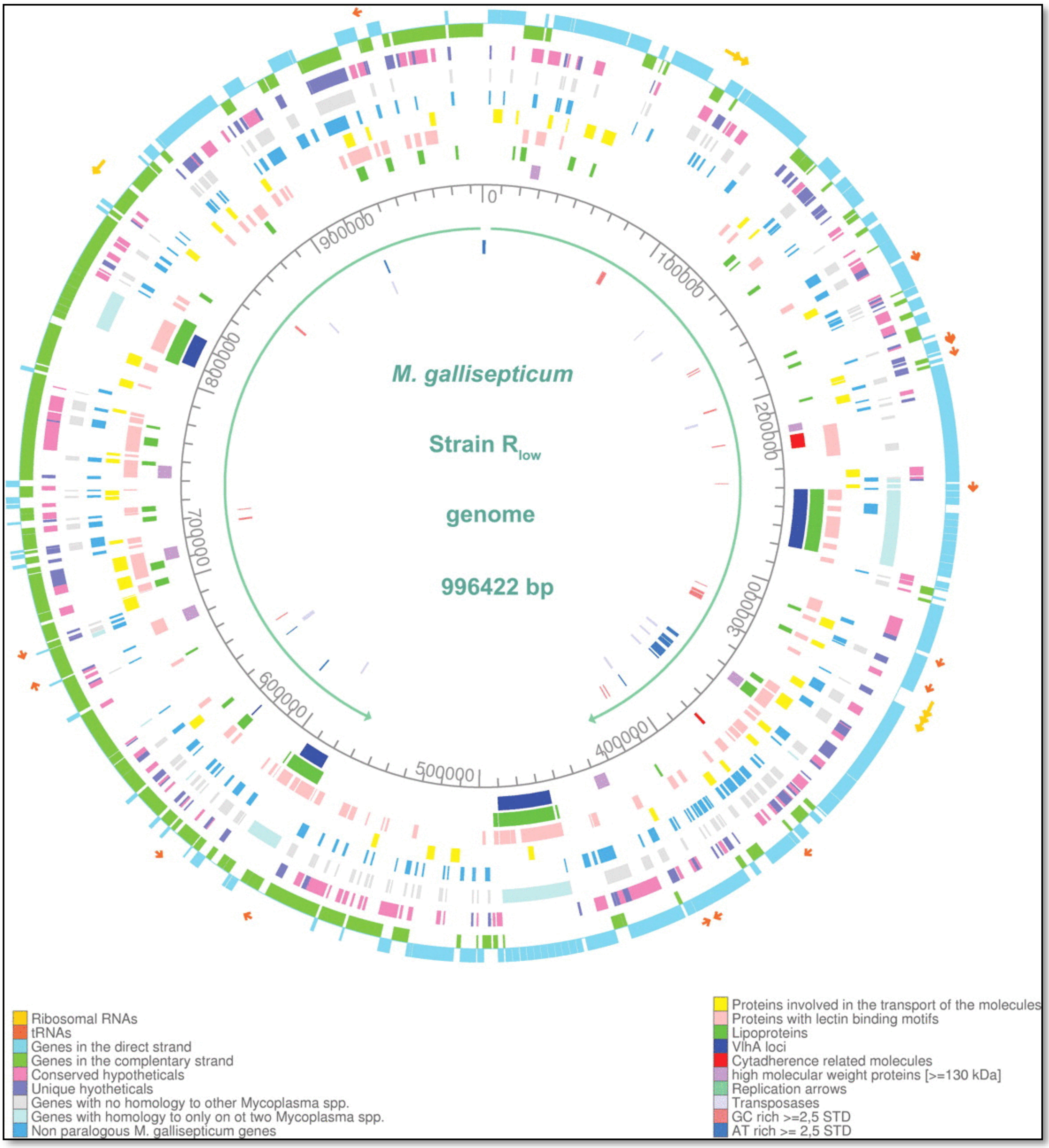

M. gallisepticum
Strain R$_{low}$
genome
996422 bp
0
100000
200000
300000
400000
500000
600000
700000
800000
900000
Ribosomal RNAs
tRNAs
Genes in the direct strand
Genes in the complentary strand
Conserved hypotheticals
Unique hyotheticals
Genes with no homology to other Mycoplasma spp.
Genes with homology to only on ot two Mycoplasma spp.
Non paralogous M. gallisepticum genes
Proteins involved in the transport of the molecules
Proteins with lectin binding motifs
Lipoproteins
VlhA loci
Cytadherence related molecules
high molecular weight proteins [>=130 kDa]
Replication arrows
Transposases
GC rich >=2,5 STD
AT rich >= 2,5 STD

Figure 1-4: Circular representation of the MG strain R$_{low}$ genome. Used with permission from Papazisi et al. (2003).

Table 1-2: Comparison of general characteristics of poultry mycoplasma

Mycoplasm species (Accession number)	Genome size (bp)	% GC	# Genes (CDS)	rRNA (5S, 16S, 23S)	tRNA	ncRNA	Pseudo genes	Reference
MG strain Rlow (NC_004829)	1 012 800	31.5	823 (733)	2, 2, 2	32	3	49	Papazisi et al. (2003) updated by Szczepanek et al. (2010)
MG strain Rhigh (NC_017502)	1 012 027	31.5	822 (729)	2, 2, 2	32	3	52	Szczepanek et al. (2010)
MG strain S6 (NC_023030)	985 433	31.5	814 (701)	2, 2, 2	33	3	71	Fisunov et al. (2011)
MG strain F (NC_017503)	977 612	31.4	808 (722)	2, 2, 2	32	3	45	Szczepanek et al. (2010)
MS ATCC 25204 (NC_CP011096)	846 495	28.3	761 (676)	3, 2, 2	34	3	41	May et al. (2015)
MS strain 53 (NC_007294)	799 476	28.5	725 (651)	3, 2, 2	34	3	30	Vasconcelos et al. (2005)
MS strain MS-H (NZ_CP021129)	818 848	28.2	727 (643)	3, 2, 2	33	3	41	Genbank®
M. cloacale NCTC 10199 (NZ_CP030103.1)	659 552	27.0	579 (539)	2, 1, 1	31	3	2	Genbank®
M. gallinaceum B2096 8B (CP011021)	845 307	28.4	631 (571)	5 total	17	0	-	Abolnik and Beylefeld (2015)
M. gallinarum DSM 19816* (NZ_JHZE00000000.1)	833 494	26.4	704 (652)	2, 2, 2	33	3	10	Yacoub et al. (2016)
M. glycophilum ATCC 35277* (NZ_JHYE00000000.1)	893 830	28.5	710 (626)	2, 2, 2	33	2	43	Genbank®
M. lipofaciens ATCC 35015* (NZ_JMKY00000000.1)	775 211	25.1	692 (639)	3, 3, 4	34	3	6	Genbank®
M. meleagridis IZSVE/2944/9/2011* (NZ_LOHQ00000000.1)	645369	25.8	572 (510)	2, 2, 1	34	3	20	Rocha et al. (2016)
M. imitans ATCC 51306* (NZ_JADI00000000.1)	919 667	30.5	765 (671)	2, 3, 3	31	3	52	Genbank®
M. iners ATCC 19705* (NZ_JNJW00000000.1)	766 027	28.2	652 (580)	2, 3, 3	34	3	27	Genbank®
M. iowae strain 695* (NZ_AGFP00000000.1)	1 195 147	24.4	980 (906)	1, 1, 1	29	3	39	Wei et al. (2012)
M. iowae strain DK-CPA* (NZ_AWQU00000000.1)	1 184 115	24.5	975 (888)	1, 1, 1	29	3	52	Pritchard et al. (2014)

CDS – coding sequences
ncRNA – non-coding RNA
*Whole genome shotgun sequencing project containing scaffolds and contigs

1.3. Genome sequencing

1.3.1. Brief overview

Numerous strategies for sequencing DNA have been developed since the early 1970's, here follows a brief description of the most notable sequencing technologies developed to date with more emphasis on the technologies used in this study in the subsequent chapters. The first method of DNA sequencing commonly used was Sanger sequencing, first developed by Frederick Sanger and based on a chain-termination (Liu et al., 2012). Improvements on this method along with the introduction of PCR and other molecular techniques led to the development of the first automated DNA sequencing machine utilising capillary electrophoresis by Applied Biosystems in 1987 (Liu et al., 2012). Until 2005 Sanger sequencing was the sequencing method of choice, producing a single long read of 500-1000bp with high accuracy which could be used in a shot-gun sequencing strategy to assemble genes or genomes. However, the cost and low throughput of this technology along with the introduction of second-generation sequencing (SGS) technologies has shifted the use of this technology to mainly gene sequences, and other gene-level studies.

The search continued for faster, high-throughput sequencing methods resulting in the development of SGS technologies, the most notable of which are Roche 454 pyrosequencing (discontinued in 2013), Illumina Solexa and sequencing by oligonucleotide ligation and detection (SOLiD) platform from Applied Biosystems (discontinued in 2016) and Ion Torrent sequencing (Schadt et al., 2010, Heather and Chain, 2016). These technologies have similar workflows from constructing a library using by shearing extracted DNA into smaller pieces, preparing a template of these DNA pieces and sequencing using various biochemical techniques. These methods produce billions of short read sequences that can be assembled using *in silico* methods (Besser et al., 2017). SGS technologies are used for a wide variety of applications, including *de novo* microbial whole genome sequencing (WGS) to produce draft genomes, mapping, targeted re-sequencing, characterization of the transcriptome, metagenomics, mutations, insertions, deletions and even for gene expression studies using RNA sequencing, but for the purposes of this thesis focus will be on *de novo* WGS and mapping (Besser et al., 2017, Glenn, 2011, Loman et al., 2012).

SGS technologies produce very short reads and the need for longer reads sequencing without the need for DNA amplification resulted in third generation sequencing technologies, such as the single molecule real time (SMRT) platform from Pacific Biosciences, which is the most widely used third generation technology, and nanopore-based sequencing from Oxford Nanopore Technologies (Heather and Chain, 2016, Feng et al., 2015). Current third generation sequencing technologies are faster and capable of producing longer sequencing reads in real time, however these are still relatively new technologies with high error rates and low output compared to Illumina and Ion Torrent sequencing (Table 1-3) (Bleidorn, 2016).

WGS has multiple application including species identification, antibiotic resistance, virulence and comparative genomics (Besser et al., 2017). Comparative genomics is a popular field where genomes of various bacterial species or strains are compared to view phylogenetic relationships and identify conserved and unique genes between species and strains to aid in bacterial diagnosis, treatment and prevention (Touchman, 2010). The main workflow of WGS with comparative genomics is characterised by the following steps: 1) sample processing; 2) sequencing and 3) data analysis, described in more detail below.

Table 1-3: Characteristics, strengths and weaknesses of commonly used sequencing platforms[a] (Besser et al., 2017).

Platform/Instrument	Throughput range (Gb)[b]	Read length (bp)	Strength	Weakness
Sanger sequencing				
ABI 3500/3730	0.0003	Up to 1 kb	Read accuracy and length	Cost and throughput
Illumina				
MiniSeq	1.7–7.5	1×75 to ×150	Low initial investment	Run and read length
MiSeq	0.3–15	1×36 to 2×300	Read length, scalability	Run length
NextSeq	10–120	1×75 to 2×150	Throughput	Run and read length
HiSeq (2500)	10–1000	×50 to ×250	Read accuracy, throughput,	High initial investment
NovaSeq 5000/6000	2000–6000	2×50 to ×150	Read accuracy, throughput	High initial investment
IonTorrent				
PGM	0.08–2	Up to 400	Read length, speed	Throughput, homopolymers[d]
S5	0.6–15	Up to 400	Read length, speed,	Homopolymers[d]
Proton	10–15	Up to 200	Speed, throughput	Homopolymers[d]
Pacific BioSciences				
PacBio RSII	0.5–1[c]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate
Sequel	5–10[c]	Up to 60 kb	Read length, speed (Average 10 kb, N50 20 kb)	High error rate
Oxford Nanopore				
MInION	0.1–1	Up to 100 kb	Read length, portability	High error rate Run length,

[a] Used under CC BY-NonCommercial-No Derivatives 4.0 International (CC BY NC ND 4.0) licence

[b] The throughput ranges are determined by available kits and run modes on a per run basis. As an example of a 15-Gb throughput, thirty-five 5-MB genomes can be sequenced to a minimum coverage of 40× on the Illumina MiSeq using the v3 600 cycle chemistry.

[c] Per one single-molecule real-time cell.

[d] Results in increased error rate (increased proportion of reads containing errors among all reads) which in turn results in false-positive variant calling.

1.3.2. Sample processing

Sample processing starts with two key steps, DNA extraction and library preparation which will determine the quality of data generated by sequencing. DNA extraction can be done using commercial kits or standardized laboratory procedures, but the quality and purity of the DNA has to be checked using the spectrophotometric A260/280 and A260/230 ratios and gel electrophoresis for degradation (Haridas et al., 2011). DNA extraction is followed by library preparation where DNA is fragmented, amplified and immobilized depending on the sequencing technology used, Figure 1-5 (Goldman and Domschke, 2014).

The two main SGS technologies in use currently are Illumina and Ion Torrent. For Ion Torrent sequencing DNA amplification is done by emulsion PCR. Briefly primers that are bound to beads, DNA template, deoxynucleoside triphosphate (dNTPs) and polymerase are loaded into micelle droplets. Then amplification by PCR is performed in each droplet, resulting in beads with multiple copies of the same DNA template bound that is ready for sequencing (Figure 1-6) (Goodwin et al., 2016b, Metzker, 2010). For Illumina sequencing DNA is amplified by solid-phase bridge amplification, first single stranded template ligated to an adapter sequence is bound to primer bound to patterned flow cell, then the free ends can bind to other primers forming bridges, which are amplified by PCR, resulting in clusters of forward and reverse strands (Figure 1-6) (Goodwin et al., 2016b, Metzker, 2010).

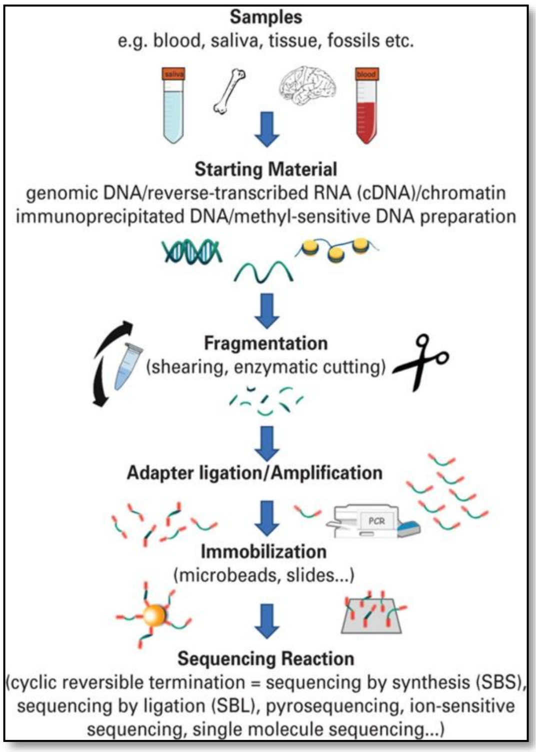

Samples
e.g. blood, saliva, tissue, fossils etc.
saliva
blood
Starting Material
genomic DNA/reverse-transcribed RNA (cDNA)/chromatin immunoprecipitated DNA/methyl-sensitive DNA preparation
Fragmentation
(shearing, enzymatic cutting)
Adapter ligation/Amplification
PCR
Immobilization
(microbeads, slides...)
Sequencing Reaction
(cyclic reversible termination = sequencing by synthesis (SBS), sequencing by ligation (SBL), pyrosequencing, ion-sensitive sequencing, single molecule sequencing...)

Figure 1-5: Workflow for DNA sequencing. DNA extraction from various sources, fragmented and amplified by PCR. The DNA fragments are then separated and immobilized depending on sequencing technology used and then sequenced in parallel. Used with permission from (Goldman and Domschke, 2014).

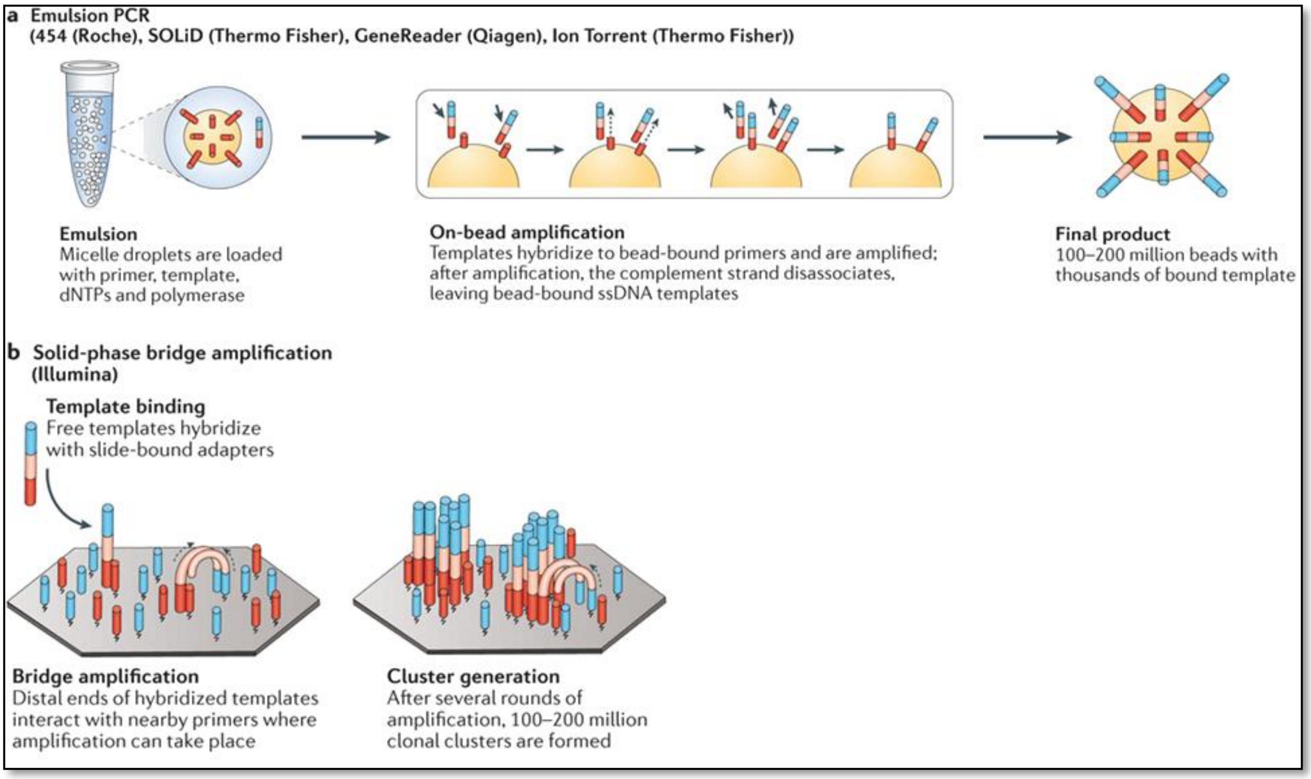

a Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))
Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase
On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates
Final product
100–200 million beads with thousands of bound template
b Solid-phase bridge amplification (Illumina)
Template binding
Free templates hybridize with slide-bound adapters
Bridge amplification
Distal ends of hybridized templates interact with nearby primers where amplification can take place
Cluster generation
After several rounds of amplification, 100–200 million clonal clusters are formed

Figure 1-6: DNA amplification during library preparation for a) Ion Torrent sequencing using emulsion PCR and b) Illumina sequencing using Solid-phase bridge amplification adapted with permission from (Goodwin et al., 2016b)

1.3.3. DNA sequencing

DNA sequencing using Illumina sequencing technology is based on reversible dye terminator technology (Figure 1-7). Four different cleavable fluorophores emitting different colours in a fluorescent microscope are attached to each of the nucleotide types and the nucleotides are also blocked at the 3' end to prevent elongation after one nucleotide has bound (Goodwin et al., 2016b, Metzker, 2010). When a nucleotide binds to the amplified DNA a signal is emitted and recorded according to the colour registered. Only one nucleotide binds and is recorded per cycle. Illumina sequencing produce high throughput data at low cost, but sample concentration plays a large role as overloading can result in low quality data (Van Dijk et al., 2014).

Ion Torrent sequencing takes advantage of the basic chemistry of DNA elongation where a proton is released when nucleotides bind resulting in an increase in pH that is measured by the ion sensors n a semiconductor chip (Figure 1-8) (Goodwin et al., 2016a, Van Dijk et al., 2014). Torrent sequencing is less common than Illumina sequencing but unlike Illumina optical scanning and fluorescence is not necessarily due to the semi-conductor technology measuring basic chemistry (Van Dijk et al., 2014). Another advantage of Ion Torrent sequencing is fast runs with longer read lengths at a lower cost than most of the other SGS technologies, however homopolymers are difficult to measure and can cause high error rates (Besser et al., 2017).

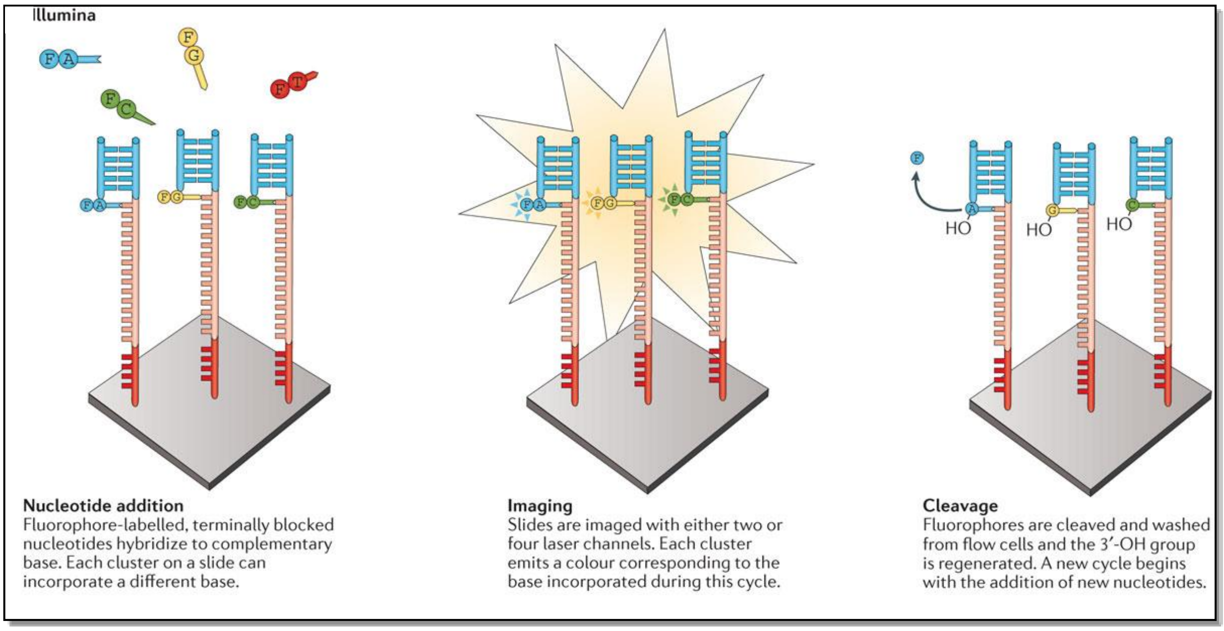

Illumina
F A
F G
F C
F T
HO
Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.
Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.
Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Figure 1-7: Reversible terminator sequencing of Illumina sequencing. A mixture containing primers, DNA polymerase and fluorophore-labelled terminally blocked nucleotides are added to the DNA template. Only one nucleotide is added per cycle, due to the terminal block group. Each of the four nucleotides have different cleavable fluorophores attached that will emit different colours using fluorescent microscopy. The fluorophore is then cleaved, and the blocked nucleotide repaired before restarting the cycle for attachment of the next nucleotide. Adapted with permission from (Metzker, 2010, Goodwin et al., 2016b)

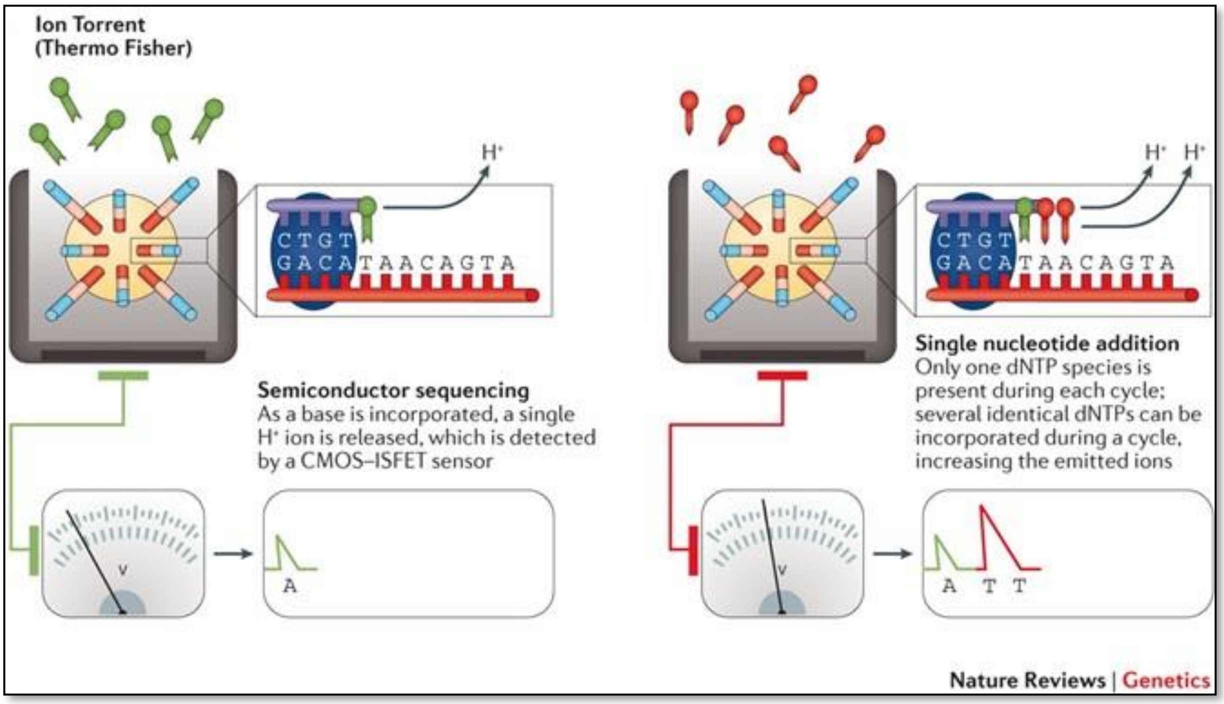

Ion Torrent
(Thermo Fisher)
H⁺
CTGT
GACA TAACAGTA
H⁺ H⁺
Semiconductor sequencing
As a base is incorporated, a single H⁺ ion is released, which is detected by a CMOS–ISFET sensor
Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions
V
A
A T T
Nature Reviews | Genetics

Figure 1-8: Ion Torrent sequencing using semiconductor-based detection. Beads containing amplified DNA template are arrayed into a microtiter plate with only one bead per well. One nucleotide type is added at a time. When a nucleotide binds to the DNA a single H^+ ions is released resulting in a change in the pH of 0.02 units. This change is detected by a metal-oxide semiconductor and ion-sensitive field-effect transistor device. Unbound nucleotides are washed away, and the next nucleotide type is added. Adapted with permission from (Goodwin et al., 2016b)

Another consideration for sequencing is how the DNA sequence should be sequenced, i.e. whether paired-end or single-end sequencing is selected and if paired-end sequencing is chosen should the sequencing be mate-paired or just paired-end? Generally, SGS produce single-end reads, where the DNA is only sequenced in one direction and this can be either in the forward or reverse direction (Glenn, 2011). With paired-end sequencing the DNA fragments are sequenced from both ends and for mate-paired reads during library preparation the fragmentation step is modified to produce larger fragments that are circularized and fragmented further for sequencing, resulting in long-insert paired-end reads, however mate-paired reads are more expensive and time-consuming than paired-end reads (Glenn, 2011). The choice of reads is dependent on the downstream application, even though both paired-end and single-end reads can be used for genome assembly, paired-end reads can be used more efficiently due to additional information on direction of the reads, but producing paired-end reads is more time-consuming and can be more expensive.

1.3.4. Data analysis

After DNA sequencing comes the daunting task of data analysis. SGS technologies produce large raw data sets containing millions of sequences, known as reads, in gigabyte size files, and high computational capability is required. An ever-increasing list of bioinformatic tools is available to aid in different steps of data analysis, ranging from free online tools, to downloadable software packages and even some pipelines where some of the tools were combined into a single process. Numerous articles are available comparing some of these tools, and a few tools have been mentioned more than others, but the algorithms of these tools are being enhanced on a continuous basis. A detailed discussion of all the available strategies is beyond the scope of this thesis and only a basic overview of requirements for WGS will be discussed briefly. For WGS the basic steps include 1) sequencing quality, 2) genome assembly, 3) genome annotation followed by the 4) research dependent applications.

Sequence quality

The quality of the sequencing data is a critical first step in determining the quality of assembly. During this step the quality of the raw reads are assessed using various characteristics including, base quality scores, read length and quantity, G+C content, sequence duplication and contaminants, such as adapters and primer-dimers (Schmieder and Edwards, 2011). In high-throughput sequencing the quality score is measured by the Phred quality score which indicates the probability that a base call is incorrect (Bokulich et al., 2013). Low quality reads and bases must be removed to improve the quality of the assembly, along with possible sequencing adapters that were added to the sequence during library preparation. After quality control the reads can be assembled. Another important aspect in sequencing to consider is depth and breadth of coverage, the depth of coverage. (Equation 1) is a measure of the average number of times a base is sequenced in the genome and breadth of coverage is a measure of the percentage of the genome that is covered by the reads at a specific depth (Sims et al., 2014). The depth of coverage required for WGS is dependent on the objectives of the study, for example, for mapping assembly lower coverage of around 10x to 30x is required, but for *de novo* assembly around 100x coverage is required.

Equation 1: Depth of coverage

$$Depth\ of\ coverage\ (C) = \frac{Number\ of\ reads\ (N) \times\ Average\ read\ length\ (L)}{Length\ of\ genome\ (G)}$$

Genome assembly

There are two approaches to assemble the short reads, i.e. *de novo* and mapping. *De novo* assembly reconstructs a genome using only the reads produced by the SGS technology which is much more time-consuming and memory intensive compared to mapping assembly where a closely related reference genome is used as a guide to reconstruct the genome (Pop, 2009). Three main strategies have been applied to *de novo* assemble the short-read sequences, the string-based method Greedy-extension, and the graph-based methods *De Bruijn* graph and overlap-layout-consensus (OLC). The *De Bruijn* graph method is most suited for large datasets containing millions of short reads, as is produced by SGS technologies (Zhang et al., 2011). *De novo* assembly can't reconstruct a complete sequence, rather reads are assembled into several contigs (from contiguous). A contig is a set of overlapping DNA sequences representative of a consensus part of the complete genome. It is recommended to use more than one assembly program so that the metrics can be compared, and the best assembly method can be chosen for further analysis (Del Angel et al., 2018). The order of these contig sequences can be determined using a reference genome or a closely related genome or information from paired-end reads, if used. Next, using a combination of two methods known as scaffolding, the contigs can be linked, either using a string of ambiguous bases (N) or gap filling, where overlapping regions between the contigs, to produce a draft genome, that is ready to be annotated (Pop, 2009, Del Angel et al., 2018).

Annotation

The final step of WGS is annotation, which include the identification and location transposable elements, structural and functional annotation (Del Angel et al., 2018). The first step is finding genes by identifying open reading frames (ORFs); followed by non-coding RNA, including tRNA, rRNA and small nucleolar RNA and regulatory regions, some of which can be identified by motifs in nucleotide sequences; then repetitive elements, such as the 16S rRNA followed by 23S rRNA and then a 5S rRNA, and segmental-duplication and finally variations among individuals of a species, identified using a comparative gene or genome approach, the most notable of which single-nucleotide polymorphisms (SNPs) (Stein, 2001, Binnewies et al., 2006). Once these elements have been identified in the nucleotide sequence, the proteins and their functions can be inferred, the genes for which the function or name is not known yet are classified as hypothetical proteins.

Comparative genomics

Comparison of bacterial species can include the genome, transcriptome and proteome level, for the scope of this thesis only genome level alignments will be considered. Genomic bacterial comparisons include the general features of a genome, such as genome size, GC content and number of genes, as well as the presence of specific genes and the order of genes (Binnewies et al., 2006). The loss, gain or change of genes can be used to infer evolution of bacterial species and have been used to show the possibility of HGT between species in the same environment, recombination between closely related species and the occurrence of genetically uniform microbes.

1.4. Aim of the research

The general aim was to assemble, annotate and compare draft genomes of mycoplasma species isolated from South African poultry farms to identify novel genes that can be used to aid in the diagnosis and treatment of poultry mycoplasma species in South Africa.

1.5. Purpose of the research

Some of the benefits that can arise from this thesis include:

- Construction of full genomes for previously unsequenced mycoplasmas
- Insights into comparative genome organisation of mycoplasmas
- Identification of the molecular markers for antimicrobial resistance
- Identification of potential virulence genes
- Identification of new genetic targets for differentiating infected from vaccinated animals (DIVA) test as well as for strain identification

CHAPTER 2: IDENTIFICATION OF MYCOPLASMA SPECIES FROM SOUTH AFRICAN POULTRY FARMS AND ASSESSMENT OF ANTIMICROBIAL RESISTANCE

Content from this chapter was published as a research article in Volume: 84 Issue 21 (2018) of *Applied and Environmental Microbiology* by A Beylefeld, P Wambulawaye, DG Bwala, JJ Gouws, OM Lukhele, DBR Wandrag and C Abolnik, entitled "**Evidence for multidrug resistance in non-pathogenic Mycoplasma species isolated from South African poultry**".

2.1. Introduction

Mycoplasma gallisepticum (MG) causes chronic respiratory disease in chickens, with symptoms that include coughing, sneezing, rales and nasal discharge (Raviv and Ley, 2013). *M. synoviae* (MS) is associated with upper respiratory infections or synovitis, symptoms include lameness, a pale comb, retarded growth and swelling around the joints (Ferguson-Noel and Noormohammadi, 2013).

Poultry flocks are screened for the presence of mycoplasmas using serological tests, such as serum plate agglutination or enzyme linked immunosorbent assays (ELISA) (Kleven, 2008). Culturing is the golden standard for diagnosing mycoplasmas, but MG is slow growing and can take anything from 72-96 hours to grow, and some isolates can take up to 4 weeks to grow. MG can therefore easily be overgrown by faster growing mycoplasmas, such as the prevalent non-pathogenic mycoplasma species *M. gallinarum* and *M. gallinaceum* (Kleven, 2008). In recent years culture-based identification have been largely replaced by DNA-based methods, such as MG- and MS-specific polymerase chain reaction (PCR) or real-time PCR, not only due to their sensitivity and speed, but also because clinical samples can be used directly without culture (Sprygin et al., 2010, Feberwee et al., 2005). Identification using the 16S ribosomal ribonucleic acid (16S rRNA) gene is considered a standard for bacterial identification and is a useful tool for the characterisation of mycoplasma species (Johansson et al., 1998). Phylogenetic analysis of the 16S rRNA gene has been used successfully to classify 50 *Mycoplasma* species into 5 phylogenetically distinct groups (Weisburg et al., 1989).

Medication and vaccination have been used to treat and control mycoplasma infection with variable success. Mycoplasmas are intrinsically resistant to antimicrobial agents that target the bacterial cell wall, but susceptible to antimicrobial agents that target protein synthesis. In the poultry industry, commonly used classes of antimicrobial agents used include tetracyclines, quinolones, macrolides and pleuromutilins such as oxytetracycline; enrofloxacin; tylosin and tiamulin, respectively. Antimicrobial resistance (AMR) is a global health threat and AMR in animal production, including poultry is a contributing source. One of the biggest contributing factors is the practice of administering antimicrobials in feed not only for the purpose of medication, but also for enhanced growth and productivity (Nhung et al., 2017).

Numerous studies have been done on acquired AMR in human mycoplasmas, but only a few studies have been done in poultry mycoplasmas. Tetracyline resistance in *M. hominis* has been associated with the acquisition of the *tetM* gene from *Streptococcus* sp., and in *M. bovis* with point mutations in the 16S rRNA gene, but this has not been shown in poultry mycoplasmas yet (Roberts et al., 1985, Lysnyansky and Ayling, 2016). In poultry mycoplasmas, quinolone resistance has been associated with mutations in the quinolone resistance determining regions (QRDR) of the Topoisomerase II and IV proteins (Gerchman et al., 2011, Lysnyansky et al., 2013). Acquired macrolide resistance has been associated with mutations in the 23S rRNA gene in MG, MS, and the ribosomal protein L4 and L22 have also been linked to resistance in *M. bovis* (Ammar et al., 2016, Lysnyansky et al., 2015, Lysnyansky and Ayling, 2016). Pleuromutlin resitance has also been associated with mutations in the 23S rRNA gene (Li et al., 2010).

In many South African poultry flocks mycoplasmosis caused by MG and MS, remains a persistent problem and in-feed medication is a common practise. Routine diagnostic tests at the Bacteriology laboratory in the Department of Veterinary Tropical Diseases (DVTD) of University of Pretoria has involved culturing with growth inhibition for the identification of mycoplasmas as MG, MS or *M.* spp, for the unidentified mycoplasma species. In this study samples were collected between 2003 and 2015 and tested to determine the diversity of mycoplasma species found in the South African poultry industry using 16S rRNA gene phylogeny. Axenic samples were then tested for their minimum inhibitory concentration (MIC) values against the commonly used antibiotics used to treat mycoplasma-infected flocks, chlortetracycline, enrofloxacin, tylosin and tiamulin. I also explored the gene and protein sequences for point mutations known to be involved in antimicrobial resistance to these antibiotics, 23S rRNA and 16S rRNA genes, ribosomal proteins L4 and L22, as well as topoisomerase II subunit A and B, and topoisomerase IV subunit A and B.

2.2. Materials and Methods

2.2.1. Sample collection

Sampling was undertaken between June 2014 and November 2015. Predominantly commercial layer chickens, but also some breeder and broiler poultry farms in the geographically separated, poultry-intensive regions of the Gauteng and Western Cape Provinces in South Africa were targeted. Poultry flocks that through routine screening with commercially available MG- and MS-specific ELISAs showed a history of mycoplasma infection were selected. Ten dry tracheal swabs were collected per house, from chickens showing typical signs of MG or MS infection, and submitted by veterinarians within 24 hours to the Bacteriology laboratory of the DVTD in the Faculty of Veterinary Science at the University of Pretoria (UP). Some of the flocks had a history of MG vaccination, therefore antibiotic treatment had been lifted at least one week prior to sampling. Passive sampling from mycoplasma-positive samples submitted to the Poultry Section of the Department of Production Animal Studies in the Faculty of Veterinary Science, UP for post-mortem examination was also sent to the Bacteriology laboratory at DVTD. These 106 samples along with 18 archived mycoplasma isolates collected from 2003 to 2013 resulted in a total of 124 samples used for this study (Table 2-1).

Table 2-1: South African Mycoplasma species analysed in this study

			Species identification	
Year	Strain	Province	Culture	16 rRNA sequencing
2003	B1102-03	N/A	MG	MG
2005	B313-05	N/A	MG	*M. gallinaceum*
2005	B733-05	N/A	MG	*M. gallinaceum*
2006	B1102-06	N/A	MG	MG
2006	B726-06	N/A	MG	MG
2006	B852-06	N/A	MG	MG
2006	B943-06	N/A	MG	MG
2007	B1028-07	N/A	MG	MG
2007	B2214-07	N/A	MS	MS
2007	B04-09-07	N/A	MG	*M. gallinarum*, MG
2008	B1072-08	North West	MG	*M. gallinarum*, MG
2008	B642-08	N/A	MG	MG
2008	B758-08	Limpopo	MG	MG
2009	B730-09	N/A	MG	MG, MS
2013	B2076-13-3	N/A	MS	MG, MS
2013	B2159-13	N/A	MG	MG
2013	B2176-13	N/A	MG	*M. gallinaceum*
2013	B2888-13-1A	N/A	*M.* spp	*M. gallinaceum*, MG
2014	B1064-14-H3	Gauteng	MG	MS
2014	B1064-14-H5	Gauteng	MS	MS
2014	B1101-14-10	Gauteng	MG	*M. gallinarum, M. pullorum*
2014	B1101-14-6	Gauteng	MG	*M. gallinarum*
2014	B1101-14-7	Gauteng	*M.* spp	*M. gallinaceum*
2014	B1101-14-8	Gauteng	MG	*M. gallinarum*
2014	B1101-14-9	Gauteng	MG	*M. gallinarum*
2014	B1064-14-H4	Gauteng	MS	MS
2014	B1173-14-2a	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-2b	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-4a	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-4b	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-5b	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-6b	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-7b	Western Cape	MG	*M. gallinaceum*
2014	B1173-14-8b	Western Cape	MG	*M. gallinaceum*
2014	B1342-14-10	Western Cape	*M.* spp	*M. gallinaceum*
2014	B1342-14-13	Western Cape	*M.* spp	*M. gallinaceum*
2014	B1342-14-18	Western Cape	*M.* spp	*M. gallinarum, M. gallinaceum*
2014	B1342-14-14	Western Cape	*M.* spp	*M. gallinaceum*
2014	B1342-14-4	Western Cape	MG	*M. gallinarum, M. gallinaceum*
2014	B1342-14-9	Western Cape	MG	*M. gallinarum*, MG
2014	B1342-14-8	Western Cape	*M.* spp	*M. gallinaceum*
2014	B1393-14-10	Gauteng	MS	MS

2014	B1393-14-4	Gauteng	MS	*M. gallinarum*, MS
2014	B1394-14-5	Gauteng	MS	MS
2014	B1394-14-2	Gauteng	MS	MS
2014	B1395-14-1	Gauteng	MG	MG
2014	B1395-14-2	Gauteng	*M.* spp	*M. gallinaceum*
2014	B1395-14-5	Gauteng	MS	*M. gallinaceum, M. pullorum*
2014	B1396-14-6	Gauteng	MG	*M. gallinarum*, MG
2014	B1396-14-7	Gauteng	*M.* spp	*M. gallinaceum*
2014	B1396-14-8	Gauteng	*M.* spp	*M. gallinaceum*
2014	B1396-14-9	Gauteng	*M.* spp	*M. gallinaceum*
2014	B1412-14-18	Gauteng	*M.* spp	*Acholeplasma laidlawii*
2014	B1414-14-1	Western Cape	*M.* spp	*M. gallinaceum*
2014	B1552-14-19	Gauteng	MG	MG
2014	B2096-14-2	Gauteng	*M.* spp	*M. gallinaceum*
2014	B2096-14-3	Gauteng	*M.* spp	*M. pullorum*
2014	B2096-14-4	Gauteng	*M.* spp	*M. gallinaceum*
2014	B2096-14-7	Gauteng	*M.* spp	*M. gallinaceum*
2014	B2096-14-8	Gauteng	*M.* spp	*M. gallinaceum*
2014	B2771-14-1A	Gauteng	MG	MG
2014	B2771-14-1B	Gauteng	MG	MG
2014	B2771-14-15A	Gauteng	MG	MG, *M. pullorum*
2014	B878-14-L3	Gauteng	MG	MG
2014	B878-14-M1	Gauteng	*M.* spp	*M. gallinaceum*
2014	B878-14-M2	Gauteng	MG	M. gallinarum, MG
2014	B878-14-M3	Gauteng	MG	*M. gallinarum*
2014	B878-14-M4	Gauteng	*M.* spp	*M. gallinaceum*
2014	B878-14-M5	Gauteng	M. spp	*M. gallinaceum*
2015	B1931-15-6A	Gauteng	MG	*M. gallinarum*, MG
2015	B1932-15-2	Gauteng	MG	*M. gallinarum*, MG
2015	B2053-15-1	Gauteng	*M.* spp	*M. gallinarum*, MG
2015	B2053-15-2	Gauteng	*M.* spp	*M. gallinarum*
2015	B2053-15-3	Gauteng	*M.* spp	*M. gallinarum*, MG
2015	B2053-15-5	Gauteng	*M.* spp	*M. gallinarum*, MG
2015	B2063-15-3	North West	MG	*M. gallinarum*, MG
2015	B2772-15-1	Gauteng	*M.* spp	*M. gallinarum*
2015	B2777-15A-7	Gauteng	MS	*M. gallinarum*, MS
2015	B2777-15A-8	Gauteng	*M.* spp	*M. gallinarum*, MG
2015	B293-15-10	Gauteng	MG	*M. gallinarum*
2015	B293-15-11	Gauteng	*M.* spp	*M. gallinarum*
2015	B293-15-12	Gauteng	*M.* spp	*M. pullorum*
2015	B293-15-13	Gauteng	*M.* spp	*M. pullorum*
2015	B293-15-14	Gauteng	*M.* spp	*M. gallinarum, M. pullorum, M. iners*
2015	B293-15-15	Gauteng	*M.* spp	*M. pullorum*
2015	B293-15-16	Gauteng	*M.* spp	*M. gallinaceum*
2015	B293-15-17	Gauteng	*M.* spp	*M. pullorum*
2015	B293-15-18	Gauteng	*M.* spp	MG, *M. pullorum*
2015	B293-15-4	Gauteng	MG	*M. gallinarum*
2015	B293-15-6	Gauteng	*M.* spp	*M. gallinarum*
2015	B293-15-7	Gauteng	*M.* spp	*M. gallinarum*, MS
2015	B293-15-8	Gauteng	MG	*M. gallinarum*

2015	B293-15-9	Gauteng	MG	*M. gallinarum*, MG
2015	B3381-15-1	Gauteng	*M.* spp	*M. gallinaceum*
2015	B3381-15-2	Gauteng	*M.* spp	*M. gallinaceum*
2015	B3381-15-3	Gauteng	*M.* spp	*M. gallinaceum*
2015	B3381-15-4	Gauteng	*M.* spp	*M. gallinaceum*
2015	B3381-15-5	Gauteng	*M.* spp	*M. gallinaceum*
2015	B3443-15-1	Gauteng	*M.* spp	*M. gallinarum*, MG
2015	B3443-15-2	Gauteng	*M.* spp	*M. gallinarum*, MG, *M. pullorum*
2015	B3443-15-3	Gauteng	*M.* spp	*M. gallinarum, M. pullorum*
2015	B3443-15-4	Gauteng	*M.* spp	*M. gallinarum*, MG, *M. pullorum*
2015	B3443-15-5	Gauteng	*M.* spp	*M. gallinarum*, MG, *M. pullorum*
2015	B3443-15-6	Gauteng	*M.* spp	*M. gallinarum*, MG, *M. pullorum*
2015	B3443-15-7	Gauteng	*M.* spp	M. *gallinarum, M. pullorum*
2015	B3443-15-8	Gauteng	*M.* spp	*M. gallinarum*, MG, *M. pullorum*
2015	B359-15-2	Gauteng	*M.* spp	*M. gallinaceum, M. pullorum, M. iners*
2015	B359-15-3	Gauteng	*M.* spp	*M. pullorum, M. iners*
2015	B359-15-4	Gauteng	*M.* spp	*M. pullorum, M. iners*
2015	B359-15-5	Gauteng	*M.* spp	*M. pullorum*
2015	B359-15-6	Gauteng	*M.* spp	*M. pullorum*
2015	B359-15-8	Gauteng	*M.* spp	*M. gallinaceum, M. pullorum, M. iners*
2015	B457-15-3	Gauteng	MG	*M. gallinarum*, MG
2015	B457-15-5	Gauteng	MG	MG
2015	B458-15-1	Gauteng	MS	MS
2015	B458-15-10	Gauteng	MG	*M. gallinarum*, MG
2015	B458-15-11	Gauteng	MS	MS
2015	B458-15-5	Gauteng	*M.* spp	*M. gallinarum*, MG
2015	B458-15-5M	Gauteng	MS	MS
2015	B458-15-6	Gauteng	MS	MS
2015	B464-15-3	Gauteng	MS	MG, MS
2015	B540-15-2	Gauteng	MG	*M. pullorum*
2015	B540-15-4	Gauteng	*M.* spp	*M. gallinarum*, MG, *M. pullorum*
2015	B540-15-5	Gauteng	*M.* spp	*M. gallinarum, M. gallinaceum*, MG

MG-*M. gallisepticum*; MS-*M. synoviae*; *M.* spp-unidentified Mycoplasma species

2.2.2. Mycoplasma isolation by culture and identification by growth inhibition

Mycoplasma isolation by culturing and identification by culture with growth inhibition was done by Johan Gouws and Pamela Wambulawaye. The standard method of *Mycoplasma* culturing and identification in the Bacteriology laboratory of the DVTD is as follows: the collected swab samples were plated directly onto Frey's agar medium, before the tip of each swab was swirled in a 5ml tube of Frey's broth medium (Frey et al., 1968). Agar plates were incubated in a 5% CO_2 in air atmosphere and examined daily for the presence of colonies under a stereomicroscope at 40x magnification. Cultures are reported as negative for mycoplasma if no growth is observed after 21 days. Morphologically distinct colonies were subcultured on a plate by cutting out a piece of agar with one isolated colony and rubbing it face-down on a new agar plate. Plates were incubated in the same manner as mentioned above and examined daily for development of colonies. The broth cultures were observed daily, and if a colour change was observed, the broth was also sub-cultured onto agar plates.

Culture identification was performed by a growth inhibition test on agar using mono-specific antisera (Clyde Jr, 1983). Monospecific antisera were prepared in-house by hyper-immunisation of rabbits with American Type Culture Collection (ATCC) cultures of MG strain NCTC10115 and MS strain ATCC25204 as described (Ruhnke and Rosendal, 1989, Clyde Jr, 1983). An isolate was identified as the specific species when a clear zone of inhibition of growth was observed around a well in the agar filled with the homologous mono-specific antiserum.

2.2.3. Mycoplasma DNA isolation

Mycoplasma DNA was isolated using the PureLink® Genomic DNA mini kit (Invitrogen). A Mycoplasma cell lysate was first prepared by harvesting cells from 100 ml of culture by centrifugation at 14 175 xg for 1h at 4˚C (Eppendorf 5804R Centrifuge with a 6 x 85ml High-speed fixed-nagle rotor). The supernatant was carefully removed with a micropipette and the pellet resuspended in 360 µl PureLink® Genomic Digestion Buffer before 20 µl of Proteinase K was added. The sample was briefly vortexed (Labnet Vortex mixer) and incubated at 55˚C for 1h (Labnet AccuBlock Digital Dry Bath), with another brief vortex after 30 min incubation. The sample was vortexed again after 20 µl of RNase A was added and then incubated at room temperature for 2 min, before 200 µl PureLink® Genomic Lysis/Binding Buffer and 200 µl Ethanol (96-100%) was added followed with another brief vortex.

The cell lysate was then added to a spin column and centrifuged at 10 000 x g for 1 min at room temperature (Sigma 1-14 centrifuge). The collection tube was discarded and the spin column with bound DNA placed into a clean PureLink® Collection Tube. The DNA was first washed with 500 µl Wash Buffer 1 and centrifuged at 10 000 x g for 1 min at room temperature and then again with 500 µl Wash Buffer 2 and centrifuged at 12 470 xg for 3 min at room temperature, discarding the collection tube after each step and placing the spin column into a clean collection tube. The DNA was eluted with 50 µl PureLink® Genomic Elution buffer, incubated at room temperature for 1 min and centrifuged at 12 470 xg for 1,5 min at room temperature into a 1,5 ml Eppendorf tube.

The quality and quantity of the DNA was checked with a Nanodrop 2000 spectrophotometer (Thermo Scientific). The samples were also checked by gel electrophoresis on a 1% (w/v) agarose gel (Seakem LE agarose) in 1x TAE buffer (40 mM Tris-acetate; 20 mM glacial acetic acid and 1 mM Ethylenediaminetetraacetic acid (EDTA), pH 8.0) and 0.175µg/ml ethidium bromide (Merck) for DNA visualization and the Gene Ruler 1kb Plus DNA ladder (Thermo Scientific). DNA samples were prepared for loading by mixing 5 µl of DNA with 1 µl DNA loading dye (Thermo Scientific). Samples were then electrophoresed at 100V for 1h and visualized under UV light. The samples were stored at -20˚C. Archives samples were analysed by Illumina MiSeq whole genome sequencing (Inqaba Biotech (Pty) Ltd, Pretoria) and samples isolated in 2014 and 2015 were analysed by Ion Torrent personal genome machine (PGM) whole genome sequencing (University of Pretoria). At the sequencing facility of the University of Pretoria samples are also subjected to a quality control check of the concentration on a Qubit fluorometer before sequencing.

2.2.4. 16S rRNA gene phylogeny

Illumina MiSeq whole genome sequencing

The paired-end MiSeq Illumina reads were first trimmed using the Nextera library and then *de novo* assembled into contigs using the default settings in CLC Genomics Workbench version 8.5.1 (CLC Bio-Qiagen, Aarhus, Denmark). The contigs was uploaded to the online RNAmmer 1.2 Server (http://www.cbs.dtu.dk/services/RNAmmer/) (Lagesen et al., 2007) and the contig(s) containing the 16S rRNA gene(s) for each sample was extracted into a single FASTA file.

Ion Torrent PGM whole genome sequencing

The number of Ion torrent reads was reduced with digital normalization, using Khmer (version 2.0) (Brown et al., 2012, Crusoe et al., 2015) for submission to the integrated genome analysis platform for Ion Torrent sequence data (IonGAP) online platform: (http://iongap.hpc.iter.es/iongap) using the default settings for the Genome Assembly and Bacterial Classification and Annotation modules (Baez-Ortega et al., 2015). The Genome Assembly module uses FastQC: A quality control tool for high throughput sequencing data (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Andrews, 2010) for quality analysis of the ion torrent reads and MIRA (Chevreux et al., 1999) for genome assembly. The Bacterial Classification and Annotation module uses BLAST (Altschul et al., 1990) with the National Centre for Biotechnology Information Nucleotide (NCBI) 16S rRNA database (ftp://ftp.ncbi.nih.gov/blast/db) for taxonomic classification of submitted data. Ion torrent reads were also *de novo* assembled using the default settings in CLC Genomics Workbench and saved for downstream analysis.

The 16S rRNA genes from both the Illumina and Ion Torrent data was checked for chimeric sequences using the UCHIME which forms part of the USEARCH program version 10 (http://drive5.com/usearch/) (Edgar er al., 2011).

Phylogenetic relationships

The 16S rRNA genes of all the avian mycoplasma species available from NCBI were collected and an avian mycoplasma 16S rRNA reference database was created. The online tool Multiple Alignment using Fast Fourier Transform (MAFFT) version 7.304 (http://mafft.cbrc.jp/alignment/server/) was used to align the 16S rRNA sequences of the archived mycoplasma samples to the reference database (Katoh et al., 2002). The resulting alignment was edited with BioEDIT Sequence Alignment Editor (version 7.2.5) and saved in a FASTA format for downstream analysis (Appendix B.1.).

The phylogenetic program PAUP (version 4) was used to perform a parsimony analysis (Swofford, 2003). A heuristic search with tree bisection and reconnection (TBR) was performed to find the best fit phylogenetic tree. Tree length, consistency index (CI), retention index (RI) and homoplasy index tree scores was used to determine homoplasy and the reliability of the phylogenetic tree was constructed by bootstrap analysis with 1000 resamplings.

For further analysis the best fitting nucleotide substitution model was determined with Akaike and Bayesian information criteria (AIC and BIC) using Jmodeltest 2.1.7 v20150530 (Darriba et al., 2012). The resulting best fit model with nucleotide substitution rates was used for the maximum-likelihood and Bayesian inference analysis. PhyML (version 3.0) was used to perform a maximum-likelihood analysis using nearest neighbour interchange (NNI) and subtree pruning and regrafting (SPR) to determine the best phylogenetic tree typology and bootstrap analysis with 1000 resamplings (Guindon et al., 2010). MrBayes (version 3.2.6) was used to perform a Bayesian inference analysis (Ronquist and Huelsenbeck, 2003).

The program FigTree (1.4.3) was used to view the resulting phylogenetic trees. The phylogenetic relationship between the samples and the reference strains were inferred and used to identify each sample.

Differentiation of MG and M. imitans species

The 16S-23S rRNA intergenic spacer region (IGSR) of samples identified by 16S rRNA gene identification as MG were extracted from the *de novo* assembled contigs. For the axenic samples the 16S rRNA and 23S rRNA results obtained from the RNAmmer 1.2 server were used to extract the 16S-23S rRNA IGSR. For the mixed samples the only the contigs containing the 16S rRNA and 23S rRNA gene identified as MG was used to extract the 16S-23S rRNA IGSR. The resulting 16S-23S rRNA (IGSR) of the samples were aligned to reference strains MG (accession number: AB098504) and *M. imitans* (accession number: AB098503) and compared as described by Harasawa et al. (2004).

The 16S rRNA and 16S-23S rRNA IGSR sequences determined in this study were deposited in the Genbank® genetic sequence database hosted by the NCBI (http://www.ncbi.nlm.nih.gov/Genbank/index.html) under accession numbers MH538971-MH539148 and MH571894-MH571937, respectively.

2.2.5. Minimum inhibitory concentration (MIC) assays.

MIC assays were performed at the Bacteriology laboratory of the DVTD by Johan Gouws according to the method published by Hannan et al. (1997). The method was modified slightly by replacing M–Broth with Frey's broth, both as culture medium and diluent (Hannan et al., 1997). Briefly, microtitre plates were coated with the antibiotic of interest, namely chlortetracycline, enrofloxacin, tylosin or tiamulin in a dilution series. The medium was adjusted to a final pH of 7.6 using 0.5M sodium hydroxide. For each sample Frey's broth was inoculated with mycoplasma sample and added to the plate. Uninoculated broth was used as an end point control and inoculated broth without antibiotic was used as positive controls. Glucose was used as fermentation substrate with phenol red as pH indicator. Adhesive tape was used to seal the plates before incubation at 36˚C and examined daily for colour change from red to yellow on the positive control. The result was reported as the lowest concentration of antibiotic where no colour change was observed (Matros L, 2001).

2.2.6. Antimicrobial resistance genes

The *de novo* assembled contigs for axenic samples for which reference genomes are available were aligned to their respective reference genomes using the CLC Genome Finishing Tool version 1.5.4. Reference genomes used were *M. gallisepticum* strain R(low) (accession no. AE015450), *M. synoviae* strain 53 (accession no. AE017245), *M. pullorum* strain B359_6 (accession no. CP017813) and *M. gallinaceum* strain B2096 8B (accession no. CP011021). The 23S rRNA, ribosomal protein L4 (*rpl*D), ribosomal protein L22 (r*pl*V), DNA gyrase subunit A (*gyr*A), DNA gyrase subunit B (*gyr*B), Topoisomerase IV subunit A (*par*C) and Topoisomerase IV subunit B (*par*E) genes were extracted. *M. gallinarum* does not have a reference genome yet, therefore *de novo* assembled contigs of two samples were submitted to the RAST prokaryotic genome annotation server (http://rast.nmpdr.org) for annotation (Aziz et al., 2008, Brettin et al., 2015, Overbeek et al., 2013). The genes of interest were extracted from the annotated contigs and used as reference to align the *de novo* assembled contigs of the remaining *M. gallinarum* samples. The 23S rRNA, genes for each sample were aligned to the reference genes of their respective species using CLC genomic workbench and compared. Nucleotide sequence of the *rpl*D*, rpl*V*, gyr*A*, gyr*B*, par*C and *par*E genes were translated to their respective proteins using the genetic code table 4. The protein sequences for each species were aligned and compared. The reference genes and proteins were also aligned to the respective reference gene or protein for *Escherichia coli* to find the gene and amino acid positions described by Gerchman et al. (2011), Lysnyansky et al. (2013) and Lysnyansky et al. (2015) (Gerchman et al., 2011, Lysnyansky et al., 2015, Lysnyansky et al., 2013) (Table 2-2).

The 23S rRNA, *rplD, rplV, gyrA, gyrB, parC* and *parE* sequences determined in study were deposited in the Genbank® genetic sequence database under accession numbers MH540196 to MH540321, MH548710 to MH548772, MH548647 to MH548709, MH548523 to MH548584, MH548585 to MH548646, MH548834 to MH548895 and MH548773 to MH548833, respectively.

Table 2-2: Position of nucleotide and amino acid substation of *Mycoplasma* species

Species	*E. coli*	MG	MS	*M. gallinarum*	*M. pullorum*	*M. gallinaceum*
Enrofloxacin						
Topoismerase II-A	D87	E97	N143	E148	E150	E154
Topoismerase II-B	A401/ G402	A416/ S417	S415/ S416	G406/ S407	G417/ G418	G415/ G416
Topoismerase IV-A	D79/ S80/ A81/ E84	D80/ S81/ S82/ E85	D84/ T85/ S86/ D89	D90/ S91/ S92/ E95	D84/ S85/ S86/ E89	D84/ S85/ S86/ E89
Topoismerase IV-B	D420/ E454	D428/ D462	D427/ D461	D426/ E460	D425/ D459	D428/ D462
Tylosin						
23S rRNA gene	G748A A2058G A2059G A2503T	G780 A2068 A2069 A2513	G789 A2053 A2054 A2499	G791 A2058 A2059 A2503	G797 A2068 A2069 A2513	G792 A2059 A2060 A2505
L4 protein	NS	NS	NS	NS	NS	NS
L22 protein	NS	NS	NS	NS	NS	NS
Tiamulin						
23S rRNA gene	A2503U G2447A	A2513	A2499	A2503	A2513	A2505

NS – Not specified

2.3. Results

2.3.1. Mycoplasma identification by growth inhibition

Active and passive sampling of chickens showing clinical signs commonly associated with MG and MS infection, e.g. coughing, sneezing, rales and nasal discharge, resulted in 124 mycoplasma samples identified by the Bacteriology laboratory as MG, MS or neither, i.e. unidentified mycoplasma species (*M.* spp) (Table 2-1). Identification in culture by growth inhibition with hyperimmune sera resulted in 50 (40.32%) mycoplasma-positive samples identified as MG, 15 (12.10%) as MS and 59 (47.58%) as *M.* spp (Figure 2-1).

2.3.2. Mycoplasma DNA isolation

Genomic DNA was extracted from mycoplasma-positive samples and prepared for NGS sequencing. Archived mycoplasma samples were isolated and submitted for Illumina MiSeq whole genome sequencing in the initial stages of the investigation, but later samples were sequenced using Ion Torrent technology due to improved coverage, cost and turnaround time. For Ion torrent sequencing 3-5 µg of amplified DNA at a concentration of at least 50-100 ng/µl of DNA is required (N. Olivier, personal communication). Gel electrophoresis was used in conjunction with nanodrop readings to determine purity of the isolated DNA (Appendix A.1 and Appendix A.2). Both RNA and DNA absorb at 260nm, proteins absorb at 280nm and EDTA and carbohydrates absorb close to 230nm, phenol and other contaminates also absorb at one of these wavelengths. These characteristics can be used to aid in determining the purity of a sample, through the 260/230 and

260/280 ratios. The 260/280 ratio is a measure of nucleic acid purity and a sample is considered to contain pure DNA, RNA or protein and other contaminants if the ratio is close to 1.8, 2.0, but this is just a rule of thumb as only value that are significantly lower than 1.8 are considered as low quality. The 260/230 ratio is normally used as a secondary measure of the purity of nucleic acids in general and a value in the range of 2.0 to 2.2 is generally expected for this metric (Thermo Scientific, 2010). Of the 2014 and 2015 samples isolated 81/106 were considered as pure, with only a few samples having possible impurities or RNA contamination that could interfere with sequencing (Appendix A.1). Some samples failed quality control at the sequencing facility and were re-isolated and sent to the sequencing facility as quickly as possible to prevent any loss of DNA that might occur during transit and long-term storage.

Sequencing data received from sequencing facility were checked for quality before *de novo* assembly by assessing the quality distribution and per base analysis. *De novo* assembly was also quality checked by assessing the depth of coverage and N50 values (data not shown). The N50 values is a metric used to evaluate the quality of sequence assemblies and is defined as the minimum contig length covering 50% of the genome size. These results will be discussed in more depth in Chapter 3 and 4. For the purpose of this Chapter, only one sample, B293-15-11, had a low assembly quality and was excluded from analysis after 16S rRNA identification. Some of the results files are too large to include in the thesis and is available upon request.

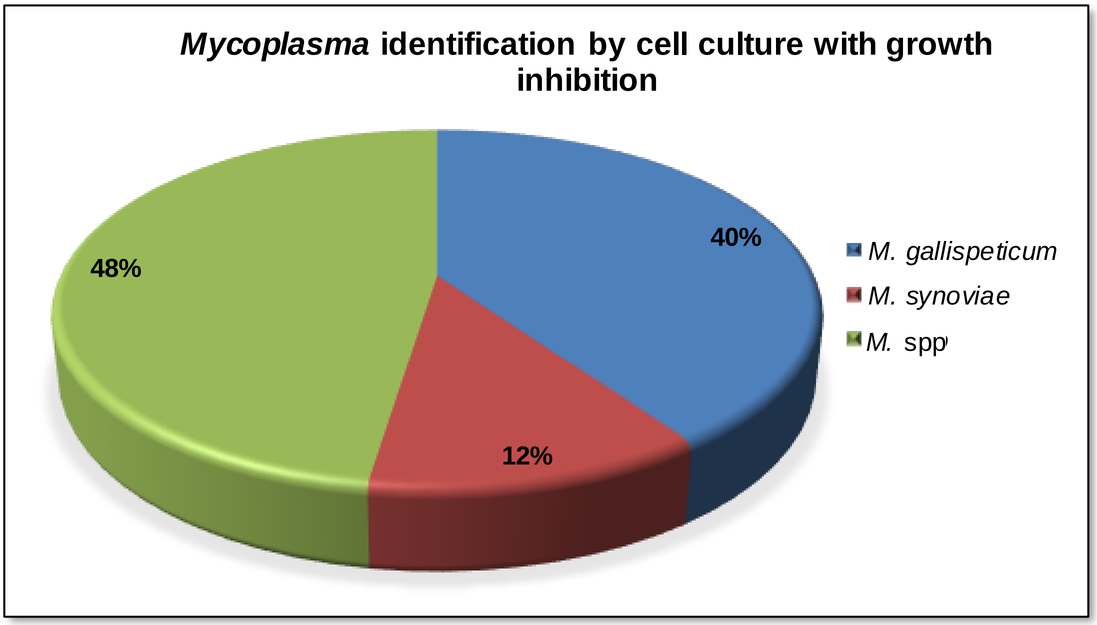


Figure 2-1: Mycoplasma identification in cell culture by growth inhibition with hyperimmune sera.

2.3.3. Mycoplasma identification by 16S rRNA gene identification

Samples were identified by combining 16S rRNA gene information from RNAmmer and the longap bacterial classification results (Figure 2-2, Table 2-1). No chimeric sequences were found using the UCHIME program. The RNAmmer results revealed more than one 16S rRNA gene for 89/124 mycoplasma samples, 44 of these contained the 16S rRNA genes of mulitple species, resulting in a total of 178 mycoplasma species identified as follows: 45 (25.28%) as MG, 17 (9.55%) as MS, 44 (24.72%) as *M. gallinarum*, 25 (14.04%) as *M. pullorum*, 41 (23.03%) as *M. gallinaceum*, 5 (2.81%) as *M. iners* and 1 sample was identified as *Acholeplasma laidlawii* (Figure 2-3). The use of the 16S rRNA gene is considered a standard for bacterial identification. Some exceptions do exist as in the case where the 16S rRNA of MG and *M. imitans* that are nearly identical and other means of distinguishing these species are required.

A putative transposase gene was found by Harasawa et al. (2004) in the 16S-23S rRNA IGSR of *M. imitans* (Harasawa et al., 2004). The 16S-23S IGSR of one MG sample (B293-15-18) could not be found in the sequencing data due no read coverage in most parts of this region, even though the sample had a depth of coverage of 846x, *de novo* assembly resulted in 12,448 reads with N50 value of only 1,822bp. The 16S-23S IGSR of the remaining 44 MG isolates were determined and aligned to MG and *M. imitans*. None of the isolates contained the putative transposase gene (Data not shown).

Proportionally more samples were received from the Gauteng province (134/178) compared to the Western Cape (19/178), North West (4/178) and Limpopo (1/178) provinces. No location information was available for the remaining 20 samples (Table 2-1)(Figure 2-4).

In the 44 samples containing two or more mycoplasma species, 98 mycoplasma species were identified. With *M. gallinarum,* the highest frequency of co-infection occurred with MG (Figure 2-5). MG was also found in co-infection with MS, *M. pullorum,* and *M. gallinaceum*, whereas MS was only found in co-infection with MG and *M. gallinarum*. *M. gallinaceum* was found in co-infections with MG, *M. gallinarum*, *M. pullorum* and *M. iners*. *M. iners*, detected at the lowest frequency in only five of the isolates, was only found in co-infections with *M. gallinaceum* and *M. pullorum* (Figure 2-5).

In 2014 and 2015 the species most frequently isolated were *M. gallinaceum* and *M. gallinarum*, respectively (Figure 2-6), but *M. gallinarum* tended to occur more frequently in co-infections in 33/44 (75.00%) cases, whereas *M. gallinaceum* co-infections were only detected in 7/44 (15.91%) cases. Overall more co-infections were observed from the 2015 samples than from the 2003-2014 samples. Both *M. gallinaceum* and *M. gallinarum* are fast-growing species, with growth on agar visible at 48 hours, whereas growth of MG and MS are usually only visible after 72-96 hours (1, 2). It is standard laboratory procedure to incubate plates for longer to allow any small colonies that are possibly MG or MS to grow.

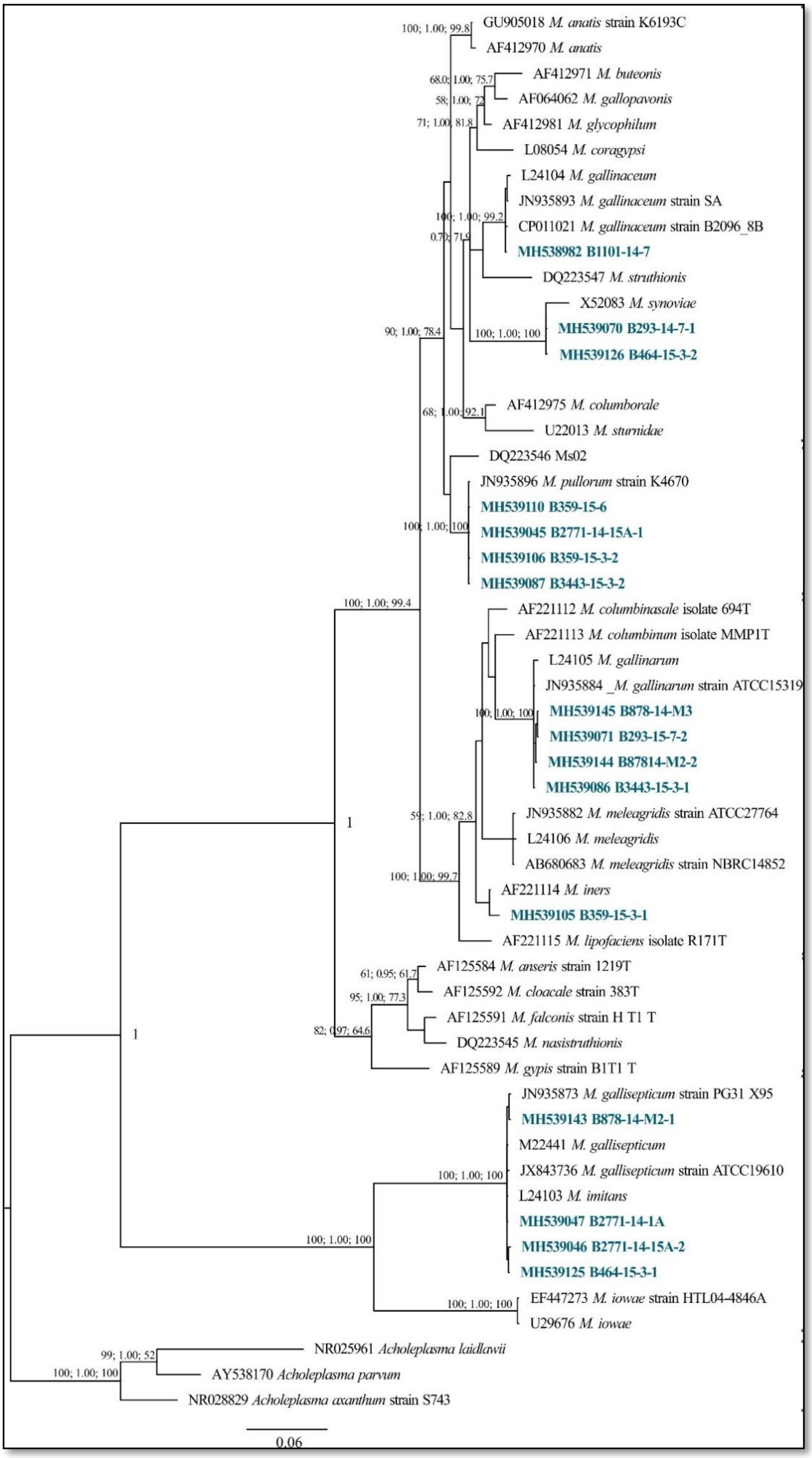


Figure 2-2: Phylogenetic tree of the 16S rRNA gene of avian mycoplasmas, representatives of the isolates from this study are shown. Values on branch refer to maximum parsimony, Bayesian inference and maximum likelihood bootstrap values, respectively. Bootstrap values below 50 not shown.

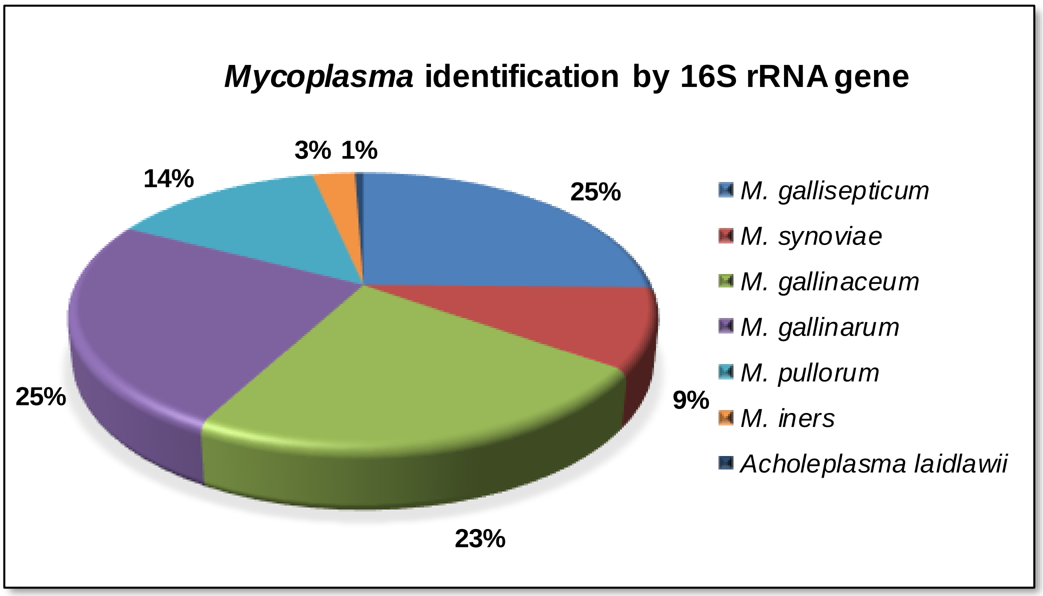


Figure 2-3: Mycoplasma identification by 16S rRNA gene identification.

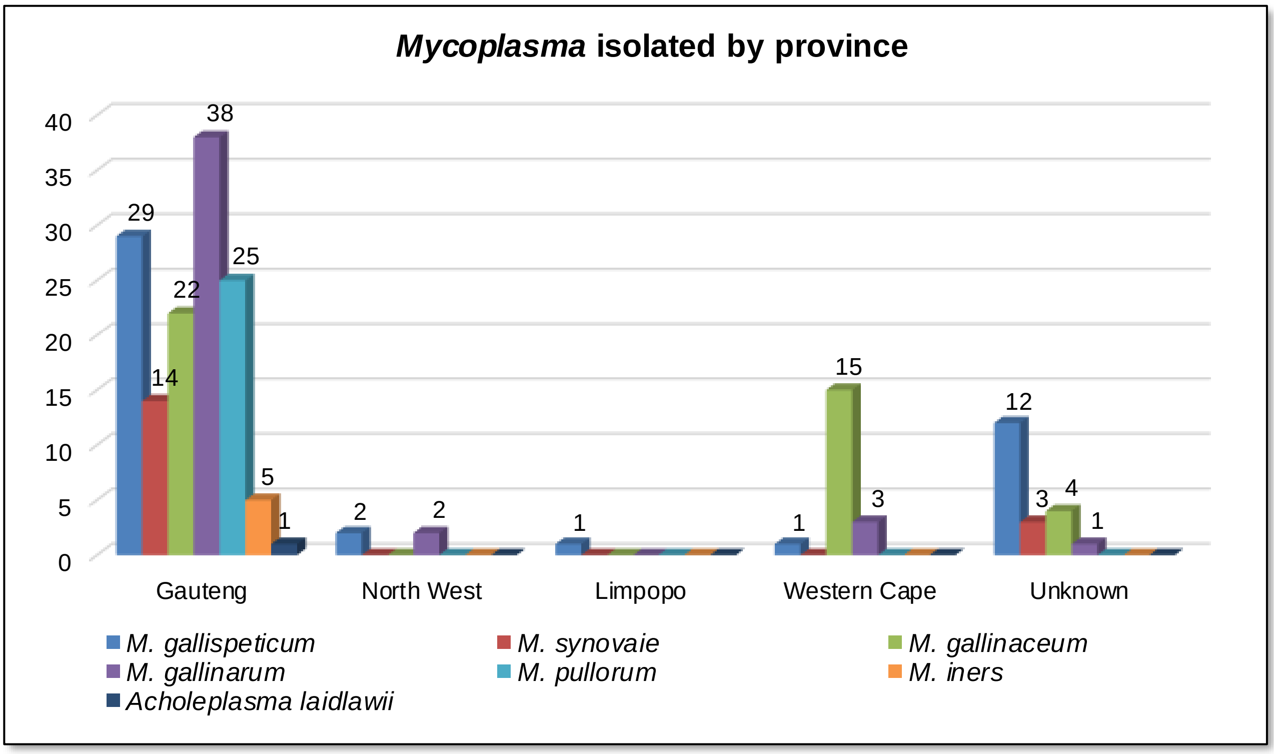


Figure 2-4: Mycoplasma identification in South African provinces.

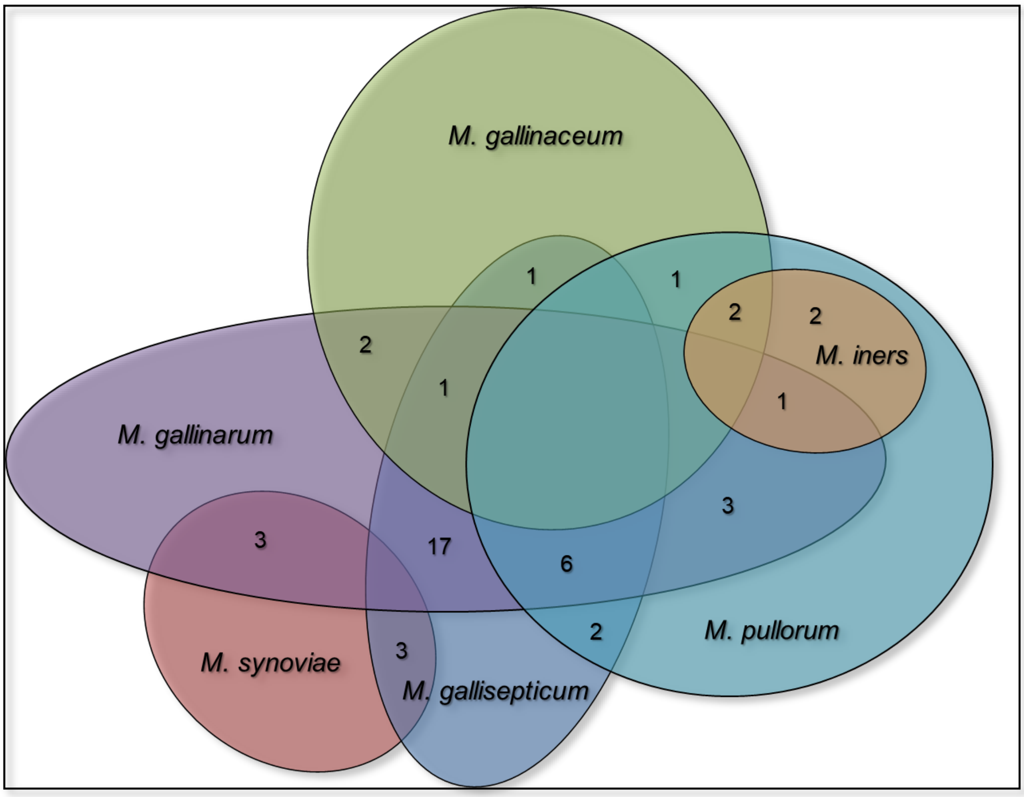


Figure 2-5: Euler diagram showing multi-infection in mycoplasma samples.

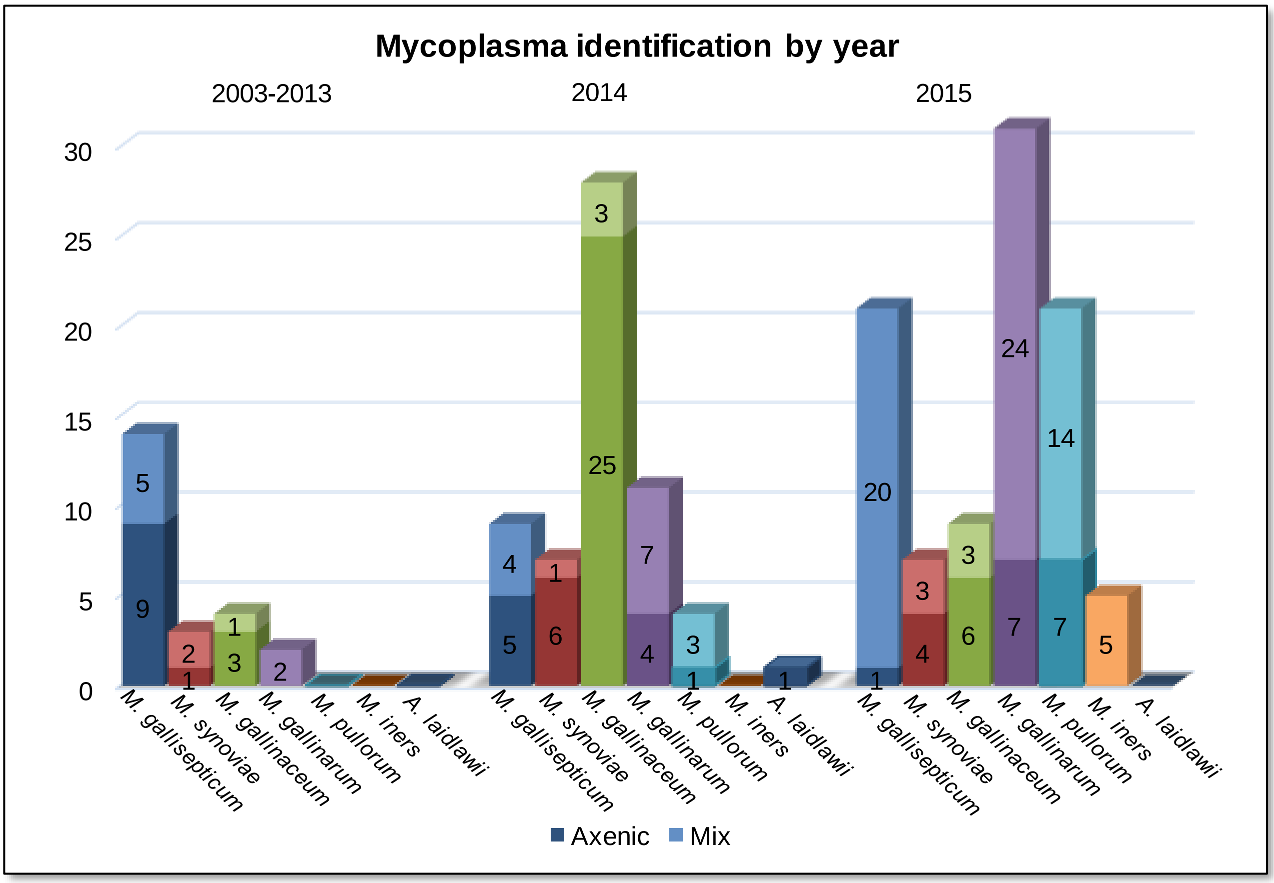


Figure 2-6: Mycoplasma species isolated per time period. Top (dark coloured) values of each column is representative of axenic cultures and bottom (light coloured) sections of columns are indicative of mixed sample cultures.

2.3.4. Comparison of culture with 16S rRNA gene identification

Error! Not a valid bookmark self-reference. compares species classification from culture identification against 16S rRNA sequencing, for the 80 axenic cultures identified out of 124 samples by 16S rRNA sequencing. Results identified as MS positive with culture were 98.75% accurate whereas the accuracy of identifying samples as MG or *M.* spp was lower at 75% and 76.25%, respectively. The specificity for MS and *M.* spp were both 100%. Samples are only classified as *M.* spp after a negative result with both MG-and MS- specific hyperimmune sera are obtained.

Culture with growth inhibition using MG-specific hyperimmune serum was the most inaccurate. Although all of MG isolates were correctly identified resulting in high sensitivity, a positive result using this antiserum was only able to correctly identify MG in 15/35 (42.86%) of cases and at a low specificity. The highest rate of misdiagnosis using MG-specific antisera due to cross-reactions were obtained with *M. gallinarum*, followed by *M. gallinaceum*, *M. pullorum* and MS. This caused a low sensitivity for MS and *M.* spp identification.

Table 2-3: Correlation between Mycoplasma identification methods in axenic samples.

16S rRNA sequence	**Growth Inhibition by Mono-Specific Antiserum**		
identification	**MG**	**MS**	***M.* spp**
M. gallisepticum (n=15)	15 (100%)	0	0
M. synoviae (n=11)	1 (9.09%)	10 (90.91%)	0
M. gallinarum (n=11)	7 (63.64%)	0	4 (36.36%)
M. pullorum (n=8)	1 (12.50%)	0	7 (87.50%)
M. gallinaceum (n=34)	11 (32.35%)	0	23 (67.65%)
Acholeplasma laidlawii (n=1)	0	0	1 (100%)
Total:	35	11	35
Accuracy:	0.75 (75.00%)	0.99 (98.75%)	0.76 (76.25%)
Sensitivity:	1.00 (100%)	0.91 (90.91%)	0.65 (64.81%)
Specificity:	0.69 (69.23%)	1.00 (100%)	1.00 (100%)

2.3.5. Minimum inhibitory concentration (MIC) assays

MIC analysis was performed on the axenic isolates and results are listed in (Table 2-4). The MICs for 16/80 isolates could not be determined, because isolates were difficult to culture after prolonged storage at minus 80˚C. No international standards for *in vitro* susceptibility testing criteria are currently available for poultry mycoplasmas. Hannan (2000) suggested breakpoints for various mycoplasma species that could be used as guidelines, the proposed breakpoints used in this study for tylosin, enrofloxacin and tiamulin are shown in Table 2-5. A breakpoint for chlortetracycline is not available, however, in MIC studies of human enteric isolates the microbiological activity of chlortetracycline and oxytetracycline was found to be comparable, thus the breakpoints for oxytetracycline was used for chlortetracycline (EMEA, 2009, Hannan, 2000). The distribution of MIC values, concentrations of compounds to which 50% or 90% of the isolates are susceptible (MIC_{50} or MIC_{90}) as well the percentage of resistance in each species is listed in Table 2-6. MIC_{50} or MIC_{90} values are usually only quoted for sample sizes of 10 or more (Hannan, 2000). The amount and relative proportion of sensitive, intermediate sensitive and resistant reaction to each antibiotic is depicted in Figure 2-7

The MG strains showed variance of MICs for both chlortetracycline, range 1 to 64 µg/ml and tylosin, range <0.01 to 16 µg/ml (Table 2-6). The MIC$_{50}$ value for chlortetracycline was 4 µg/ml indicating that at least half of the MG isolates remained sensitive to chlortetracycline, whereas the MIC$_{50}$ value for tylosin was 10 µg/ml indicating that more than half of the MG isolates are resistant to tylosin. Only MG strains showed resistance to chlortetracline (2/10) and tylosin (6/10), and intermediate susceptibility to both chlortetracycline (3/10) and enrofloxacin (2/10). The two chlortertracycline resistant MG strains, B758-08 and B943-06, also showed resistance to tylosin and intermediate sensitivity to enrofloxacin. Two of the three MG strains that showed intermediate susceptibility to chlortetracycline also showed resistance to tylosin. Four MG strains were resistant to one antimicrobial, and two were resistant to two antimicrobials. (Table 2-4 and Figure 2-7).

The variance in MS MICs for chlortetracycline ranged from 2 to 32 µg/ml, and for tylosin the range was 0.02 to 10 µg/ml (Table 2-6). The MIC$_{50}$ value for chlortetracycline was 16 µg/ml indicating a high percentage of resistance, MIC$_{50}$ values for the other antibiotics showed less than 50% resistance. MS strains had resistance either to chlortetracycline (6/11), tylosin (3/11) or enrofloxacin (1/11). Acquired intermediate susceptibility to chlortetracycline (3/11) and enrofloxacin was also observed (4/11) Only B1394-14-5 was resistant to both chlortetracycline and tylosin. Two other tylosin resistant MS strains, B1394-14-2 and B1393-14-10, also showed intermediate sensitivity to chlortetracycline and three chlortetracycline resistant strains, B1394-14-5, B1064-14-H5 and B1064-14-H3 showed intermediate resistance to enrofloxacin. (Table 2-4 and Figure 2-7)

For *M. gallinarum,* resistance was observed to chlortetracycline (6/9) and tylosin (6/9) as well as intermediate susceptibility to chlortetracycline (3/9) and enrofloxacin (5/9). Three of the nine chlortetracycline-resistant strains, B2053-15-2, B2772-15-1 and B293-15-6, also showed resistance to tylosin with sample B2053-15-2 also showing intermediate susceptibility to enrofloxacin (Table 2-4 and Figure 2-7). Two of the six *M. pullorum* strains, B293-15-12 and B293-15-15, showed resistance to tylosin and chlortetracycline. Two more strains showed resistance to chlortetracycline. Two strains showed intermediate susceptibility to enrofloxacin. (Table 2-4 and Figure 2-7). MIC$_{50}$ and MIC$_{90}$ values were not determined for these two species.

Proportionately more *M. gallinaceum* strains were isolated, with 28 used for MIC analysis. MIC ranges were 0.5 to >20µg/ml and 1 to >20 µg/ml, for chlortetracycline and tylosin respectively (Table 2-6). The MIC$_{50}$ for chlortetracycline of 10 µg/ml and MIC$_{90}$ of >20 µg/ml indicates that the majority of *M. gallinaceum* strains were resistant to chlortetracycline, in fact only six remained fully susceptible (Table 2-4, Table 2-6 and Figure 2-7). The MIC$_{50}$ and MIC$_{90}$ for tylosin was 10 and >20 µg/ml, respectively, with almost all the strains showing resistance to tylosin (24/28). Only one strain was fully sensitive, and three samples had intermediate sensitivity to tylosin (Table 2-4, Table 2-6 and Figure 2-7). The MICs for enrofloxacin ranged from 0.08-10µg/ml, with an MIC_{50} of 0.32 and an MIC_{90} of 5µg/ml (Table 2-6). Five of the strains were resistant to enrofloxacin, and five more samples showed intermediate resistance. Resistance to both chlortetracycline and tylosin was observed in 14 strains, one of these *M. gallinaceum* strains, B1173-14-4a, showed intermediate sensitivity to tiamulin. One strain showed resistance to both enrofloxacin and tylosin and three strains, B1101-14-7, B1342-14-10 and B1342-14-13, were resistant to chlortetracycline, enrofloxacin and tylosin.

Proportionally between species, only 20% of the MG strains were resistant to chlortetracycline compared to 55%, 67%, 71% and 48% observed for MS, *M. gallinarum, M. pullorum* and *M. gallinaceum,* respectively (Table 2-6, Figure 2-7(b)). For tylosin, only 27% of MS samples were resistant compared to 60%, 67%, 43% and 86% for MG, *M. gallinarum, M. pullorum* and *M. gallinaceum,* respectively (Table 2-6, Figure 2-7(b)). Enrofloxacin resistance was only detected in *M. gallinarum* (33%), *M. gallinaceum* (18%) and MS (9%) (Table 2-6, Figure 2-7(b)).

2.3.6. Antimicrobial resistance genes

The rplD, rplV, gyrA, gyrB, parC and parE genes of every strain for each species were extracted from *de novo* assembled contigs and translated to their respective protein sequences. Ribosomal proteins L4 and L22, DNA gyrase subunits A and B and topoisomerase IV subunits A and B; were also aligned and compared (Mulitmedia). The 23S rRNA gene was also aligned and compared for each species. Specific point mutation in the 23S rRNA gene have been associated with acquired resistance to tylosin and tiamulin, as well as amino acid substitutions in ribosomal proteins L4 and L22 (Gerchman et al., 2011, Lysnyansky et al., 2015, Lysnyansky and Ayling, 2016, Li et al., 2010). Amino acid substitutions in DNA gyrase subunit A and B and topoisomerase IV subunit A and B have been associated with quinolone resistance (Gerchman et al., 2011, Lysnyansky et al., 2013). *E. coli* numbering used throughout, except for ribosomal protein L4 and L22 (Table 2-2).

A comparison of the 23S rRNA gene showed that 5/6 tylosin resistant MG strains had point mutation A2059G on one or both 23S rRNA genes (Table 2-4). No amino acid substitutions were observed in the L4 and L22 ribosomal proteins; DNA gyrase subunits A and B; and DNA topoisomerase IV subunit A and B proteins, respectively. Both 23S rRNA genes of the three tylosin resistant MS strains had acquired mutation A2059G, and no amino acid substitution was observed in the L4 and L22 ribosomal proteins. All MS strains had amino acid substitutions N89D and D461E of the Topoisomerase IV subunits A and B proteins when compared to the MS reference strain, but the only enrofloxacin resistant strain, B1394-14-5, also had amino acid substitution D420N (Table 2-4).

Table 2-4: MIC's and resistance mutations of Mycoplasma strains.

Strain	MIC (µg/ml)				Macrolide resistance genes			Quinolone resistance genes	
	Chlortetracycline[a,b]	Enrofloxacin[a]	Tylosin[a]	Tiamulin[a]	23SrRNA[c]	*rpl*D	*rpl*V	*par*C[c]	*par*E[c]
M. gallisepticum									
NCTC 10115 (Control)	1.250	0.160	0.160	0.160					
USDA 56 (Control)	2.500	0.160	0.160	0.160					
B1102-03	1	0.250	0.125	0.060					
B1102-06	1	0.250	0.125	0.060					
B726-06	4	0.250	**16**	0.250	A2059G[d]				
B943-06	**16**	1	**16**	2	-				
B1028-07	8	0.250	**16**	0.250	A2059G[d]				
B758-08	**64**	1	**16**	1	A2059G[d]				
B2159-13	4	0.250	**16**	0.120	A2059G[d]				
B1395-14-1	10	0.080	**10**	0.160	A2059G				
B878-14-L3	10	0.040	0.010	0.010					
B457-15-5	2	0.250	0.125	0.060					
M. synoviae									
NCTC 10124 (Control)	5	**2.500**	0.080	2.500				N84D	D454E
ATCC 25204 (Control)	2.500	**5**	0.080	2.500	N/D	N/D	N/D	N/D	N/D
B2214-07	2	**2**	0.125	0.500				N84D	D420N, D454E
B1064-14-H4	10	0.640	0.020	2.500				N84D	D454E
B1064-14-H3	**20**	0.640	0.040	2.500				N84D	D454E
B1064-14-H5	**20**	0.640	0.040	2.500				N84D	D454E
B1394-14-2	10	0.080	**10**	2.500	A2059G[d]			N84D	D454E
B1393-14-10	10	0.320	**10**	2.500	A2059G[d]			N84D	D454E
B1394-14-5	**20**	0.640	**10**	2.500	A2059G[d]			N84D	D454E
B458-15-1	4	0.250	0.125	0.120				N84D	D454E
B458-15-5	**16**	0.250	0.125	0.120				N84D	D454E
B458-15-6	**32**	0.250	0.125	0.250				N84D	D454E
B458-15-11	**32**	0.500	0.125	0.250				N84D	D454E

Strain	MIC (µg/ml)				Macrolide resistance genes			Quinolone resistance genes	
	Chlortetracycline[a,b]	Enrofloxacin[a]	Tylosin[a]	Tiamulin[a]	23SrRNA[c]	*rpl*D	*rpl*V	*par*C[c]	*par*E[c]
M. gallinarum									
B1101-14-6	**20**	0.640	0.040	2.500	G2059A				
B1101-14-8	**20**	0.640	0.040	2.500					
B1101-14-9	**20**	0.640	0.040	2.500	G2059A				
B878-14-M3	10	0.320	**>20**	1.250	G745A, G2059A	I196T	H91K		
B2053-15-2	**16**	1	**> 16**	0.500	G2059A	I196T			
B2772-15-1	**16**	0.250	**> 16**	0.250	G2059A	I196T			
B293-15-10	8	0.250	**> 16**	0.500	G745A, G2059A	I196T	H91K		
B293-15-11	8	1	**> 16**	0.500	N/D	N/D	N/D	N/D	N/D
B293-15-6	**16**	0.250	**> 16**	1	G745A, G2059A	I196T	H91K		
M. pullorum									
B293-15-12	**64**	0.250	**4**	0.250					
B293-15-15	**32**	0.250	**4**	0.250					
B293-15-17	**16**	0.250	0.250	0.060					
B359-15-5	4	1	0.125	0.250					
B359-15-6	1	1	0.125	0.500					
B540-15-2	**32**	0.250	0.125	0.060	G748A			S81P	
M. gallinaceum									
B313-05	**16**	0.250	**> 16**	1	G748A[d]				
B733-05	**16**	1	**8**	1	G748A[d]				
B1101-14-7	**20**	**10**	**>20**	5	G748A[d]			S80L	
B1173-14-2a	2.500	0.160	**5**	0.640	G748A[d]				
B1173-14-2b	10	0.320	**10**	1.250	G748A[d]				
B1173-14-4a	>20	0.320	**>20**	10	G748A[d]				
B1173-14-4b	**20**	0.320	**>20**	5	G748A[d]				
B1173-14-5b	**20**	0.320	**>20**	5	G748A[d]				
B1173-14-6b	**20**	0.160	**20**	1.250	G748A[d]				
B1173-14-7b	10	0.160	**10**	1.250	G748A[d]				

Strain	MIC (µg/ml)				Macrolide resistance genes			Quinolone resistance genes	
	Chlortetracycline[a,b]	Enrofloxacin[a]	Tylosin[a]	Tiamulin[a]	23SrRNA[c]	*rpl*D	*rpl*V	*par*C[c]	*par*E[c]
B1173-14-8b	**20**	0.160	**>20**	5	G748A[d]				
B1342-14-10	**>20**	**10**	**20**	2.500	G748A[d]				
B1342-14-13	**20**	**2.500**	**10**	1.250	G748A[d]			E84G	
B1342-14-14	10	**5**	**10**	5	G748A[d]				
B1342-14-8	**20**	0.160	**>20**	5	G748A[d]				
B1395-14-2	1.250	0.080	**5**	1.250	G748A[d]				
B1396-14-7	10	0.160	**20**	1.250	G748A[d]				
B1396-14-8	**>20**	0.160	**10**	1.250	G748A[d]				
B1396-14-9	**>20**	0.160	**10**	5	G748A[d]				
B1414-14-1	**20**	N/D	**>20**	2.500	G748A[d]				
B878-14-M1	5	0.160	**>20**	0.320	G748A[d]				
B878-14-M4	5	0.640	**>20**	0.320	G748A[d]				
B878-14-M5	10	0.320	**>20**	0.320	G748A[d]				
B3381-15-1	2	1	2	1					
B3381-15-2	8	0.250	4	0.500	-				
B3381-15-3	2	1	2	1					
B3381-15-4	2	**2**	2	0.500					
B3381-15-5	0.500	1	1	0.500					

[a] Breakpoints according to Hannan (2000) (Table 3) with

Green – Susceptibility to antimicrobial agent

Yellow – Intermediate susceptibility to antimicrobial agent

Red and **Bold face** – Resistance to antimicrobial agent

[b] No breakpoint available, oxytetracycline values used (Table 3).

[c] *E.coli* numbering

[d] Found on both 23S rRNA genes

N/D – Not determined

Table 2-5: Minimum Inhibitory Compound breakpoints.

Class	Antibiotic	Susceptible (µg/ml)	Intermediate (µg/ml)	Resistant (µg/ml)
Tetracyclines	Chlortetracycline[a,b]	≤4	8	≥16
Fluoroquinolones	Enrofloxacin[a]	≤0.5	1	≥2
Macrolides	Tylosin[a]	≤1	2	≥4
Pleuromutilin	Tiamulin	≤8	-	≥16

[a]Breakpoints according to Hannan (2000)
[b]No breakpoint available, oxytetracycline values used

The lack of a reference genome for *M. gallinarum* combined with the presence of contaminating DNA from *Paenibacillus* spp therefore made it difficult to retrieve genes of interest for sample B293-15-11 and was not analysed. Point mutation G2059A was observed in one or both of the 23S rRNA genes of 7/8 *M. gallinarum* strains, but only 5 of these showed resistance to tylosin (Table 2-4). In 3/5 of the tylosin-resistant strains, mutation G745A was present in both 23S rRNAs as well as substitution H91K in ribosomal protein L22 (Table 2-4). All 5 tylosin resistant strains had acquired substitution G354A and T587C in gene rplD which corresponded to no substitution at amino acid position 119 and substitution I196T, respectively ribosomal protein L4 (Table 2-4, Appendix B). One *M. pullorum* strain, B540-15-2, had point mutation G748A in the 23S rRNA and substitution S81P in *par*C of the QRDR, but no correlation with tylosin or enrofloxacin resistance was observed (Table 2-4).

Point mutation G748A was observed in both 23S rRNA genes of 23/24 tylosin resistant *M. gallinaceum* strains (Table 2-4). Comparison of topoisomerase IV subunit A revealed amino acid substitutions S81L and E84G in only two enrofloxacin resistant strains, B1101-14-7 and B1342-14-13, respectively. No further mutations were observed for the *M. gallinaceum* strains.

Table 2-6: Minimum inhibitory compound values (μg/ml) distribution of Mycoplasma strains.

Antimicobial agent	Number of strains with MIC(ug/ml) of																		MIC$_{50}$	MIC$_{90}$	Resis-tance
Chlortetracycline	**0.5**	**1**	**1.25**	**2**	**2.5**	**4**	**5**	**8**	**10**	**16**	**20**	**>20**	**32**	**64**							
M. gallisepticum		2		1		2		1	2	1				1					4	16	20%
M. synoviae				1		1			3	1	3		2						16	0.64	54.6%
M. gallinarum								2	1	3	3										66.7%
M. pullorum		1				1				1			2	1							66.7%
M. gallinaceum	1		1	3	1		2	1	5	2	8	4							10	>20	50%
Enrofloxacin	**0.04**	**0.08**	**0.16**	**0.25**	**0.32**	**0.5**	**0.64**	**1**	**2**	**2.5**	**5**	**10**	**N/D**								
M. gallisepticum	1	1		6				2											0.25	1	0%
M. synoviae		1		3	1	1	4		1										0.50	0.64	9.9%
M. gallinarum				3	1		3	2													0%
M. pullorum				4				2													0%
M. gallinaceum		1	9	2	5		1	4	1	1	1	2	1						0.32	5	21.4%
Tylosin	**0.01**	**0.02**	**0.04**	**0.08**	**0.125**	**0.16**	**0.25**	**0.64**	**1**	**2**	**4**	**5**	**8**	**10**	**16**	**> 16**	**20**	**>20**			
M. gallisepticum	1				3									1	5				10	16	60%
M. synoviae		1	2		5									3					0.125	10	27.3%
M. gallinarum			3													5		1			66.7%
M. pullorum					3		1				2										33.3%
M. gallinaceum									1	3	1	2	1	6	0	1	3	10	10	>20	85.7%
Tiamulin	**0.01**	**0.06**	**0.08**	**0.12**	**0.16**	**0.25**	**0.32**	**0.5**	**0.64**	**1**	**1.25**	**2**	**2.5**	**5**	**10**						
M. gallisepticum	1	3		1	1	2				1		1							0.12	1	0%
M. synoviae				2		2		1					6						2.50	2.50	0%
M. gallinarum						1		3		1	1		3								0%
M. pullorum		2				3		1													0%
M. gallinaceum							3	3	1	4	7		2	7	1				1.25	5	0%

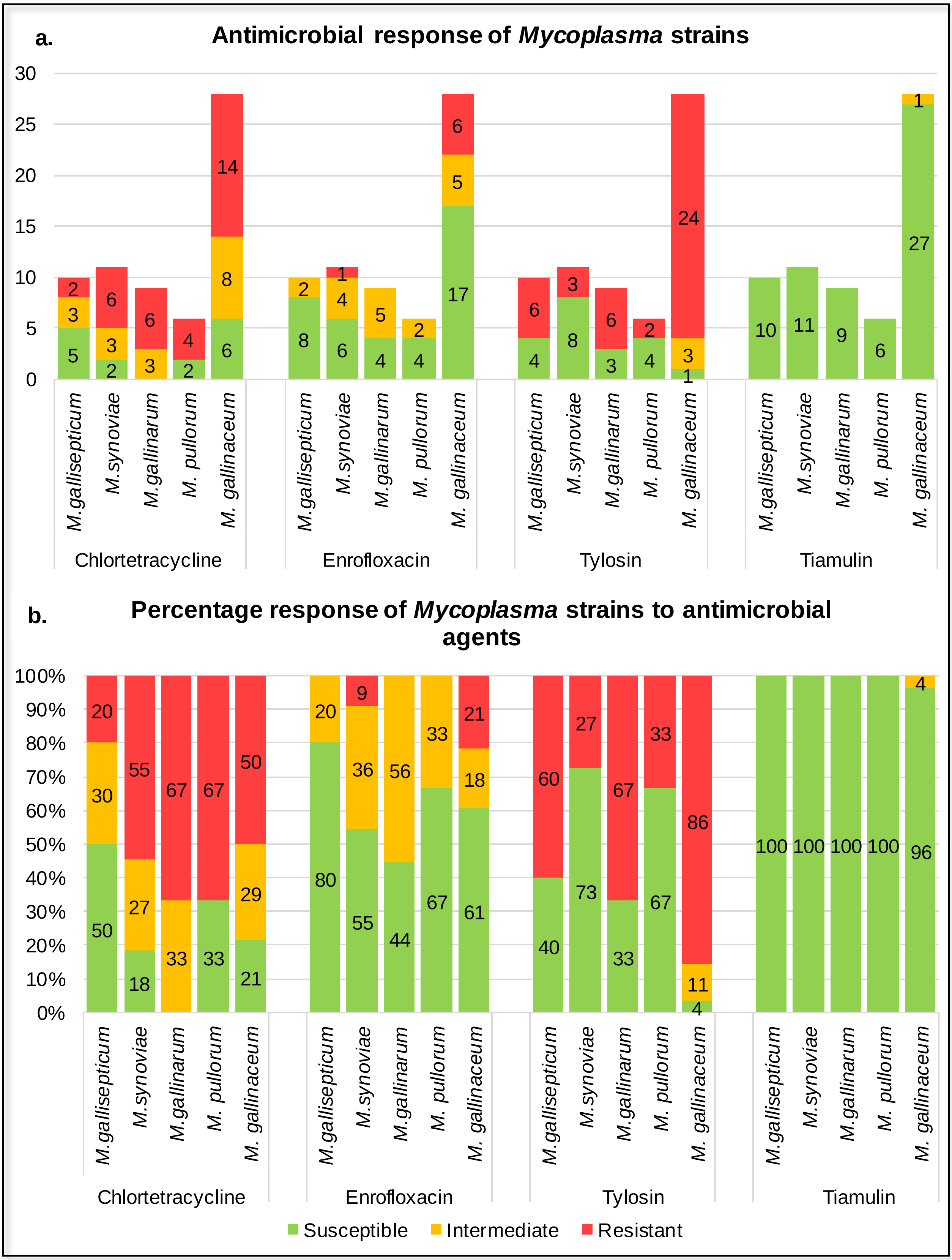


Figure 2-7: Antibiotic sensitivity of mycoplasma strains to chlortetracycline, enrofloxacin, tylosin and tiamulin. In (a) total number of strains that are either resistant, susceptible or in the intermediate range is depicted, with relative proportions given in (b).

2.4. Discussion

For several decades, culture followed by growth inhibition with hyperimmune sera was the only identification method for poultry mycoplasmas in South Africa and diagnostic laboratories only recently implemented ELISA and molecular detection methods. The species identification discrepancy between growth inhibition test and DNA sequencing results has been previously reported by others (Razin, 2012). Even though the shortcomings of identification by culture with growth incubation are well recognised (Razin, 2012), growth inhibition test identification of strains as MS or *M.* spp was reasonably accurate, but where the use of MG hyperimmune sera for identification had a high sensitivity, it had a low accuracy and specificity.

DNA concentration is important for good quality sequencing, but numerous other factors can also affect the quality of sequencing results, such the AT- or GC-rich sequences, or high frequency of repeat sequence, and contaminants (Sims et al., 2014, Schatz et al., 2010). Only one sample, B293-15-11 produced a weak *de novo* assembly and could not be analysed after 16S rRNA identification. The 16S rRNA of contaminating species were found and paired with the lack of a reference genome for *M. gallinarum* made it difficult to separate the different species. The sequencing quality of the remaining samples will be discussed in more detail in Chapter 3 and 4.

Various samples identified as MG and MS by culture with growth inhibition test were succesfully confirmed by 16S rRNA genes extracted from *de novo* assembled contigs of whole genome sequencing data. Additionally *M. gallinaceum*, *M. gallinarum*, *M. pullorum* and *M. iners* were also identified in poultry samples collected between 2003 and 2015 mostly from the Gauteng province, but also from the Western Cape, Limpopo and the North West provinces (Table 2-1). Other poultry mycoplasma species reported globally were not detected, namely, *M. glycophilum*, *M. iowae*, *M. lipofaciens* or *M. meleagridis*, the latter was not unexpected as turkeys are not farmed in South Africa. It is difficult to distinguish *M. imitans* from MG by serological methods, but a putative transposase previously identified in the 16S-23S IGSR is used to distinguish these species (Harasawa et al., 2004). Apart from one co-infection sample that could not be distinguished as discussed, *M. imitans* was not identified in the remaining samples tested.

M. gallinaceum, *M. gallinarum*, *M. pullorum* and *M. iners* are considered as non-pathogenic mycoplasma species but are known to exacerbate respiratory diseases in co-infections (Kleven, 1998). Thus, under certain circumstances the effects of these “non-pathogenic” mycoplasmas on production could be significant. The most common species isolated in 2014 was *M. gallinaceum*, this species has previously been associated with conjunctivitis in pheasants (Welchman et al., 2002), and recently Adeyemi and coworkers (2017) demonstrated the role of *M. gallinaceum* in enhancing infectious bronchitis virus replication *in vivo* (Adeyemi et al., 2017). The most prevalent species isolated in 2015 was *M. gallinarum*, albeit mostly in co-infections with MG. *M. gallinarum* and is known to cause airsacculitis in chickens (Kleven et al., 1978). *M. pullorum* can cause an increase in embryo mortality and *M. iners* can cause lesions in embryos (Wakenell et al., 1995, Moalic et al., 1997). The presence of these non-pathogenic species in flocks should therefore closely monitored for their potential adverse effects on production.

Considering that MG and MS are the listed by the OIE as notifiable disease agents in poultry and remains a problem worldwide, less MG and MS were isolated from the flocks than was expected, and is consider that this could be due to 1) the widespread use of MS and MG vaccines in South Africa; recently, the protective efficacy of ts-11 and 6/85 vaccines were demonstrated in challenge studies with a typical MG field strain, B2159-13, also analysed here (Bwala et al., 2018) or 2) the presence of the fast-growing species *M. gallinarum* and *M. gallinaceum* outgrowing MG and MS in co-infections during culture, although mitigating measures were taken during culture to avoid this 3) the antibiotics in use are still effective against MG and MS field isolates - although insufficient numbers were assessed here to state this conclusively. The joints of the chickens were not swabbed, that may have selectively detected more MS. Accurate diagnosis is vital for effective treatment and control and, the information generated by this study provides a good starting point for future epidemiological surveys, based on differential PCR assays, to assess the incidence and diversity of mycoplasmas in South African poultry flocks.

AMR in poultry microbes is a growing global concern, yet only a few studies could be found that determined MICs for poultry mycoplasma strains, and then only MG and MS were investigated (Nhung et al., 2017). Resistance to chlortetracycline, enrofloxacin or tylosin were observed in some strains of all species cultured in this study. The finding of relatively higher chlortetracycline and tylosin resistance compared to enrofloxacin for MG and MS was not unexpected as long-term use of oxytetracycline, as practised over decades in South Africa, is known to cause resistance (Pakpinyo and Sasipreeyajan, 2007, Eagar et al., 2012). Furthermore, *in vitro* studies have shown that tylosin resistance develops quickly, compared to enrofloxacin resistance that develops slowly over time (Gautier-Bouchardon et al., 2002).

All axenic strains tested were susceptible to tiamulin, except for one *M. gallinaceum* strain that had developed intermediate susceptibility. Tiamulin is a pleuromutilin that in general had been found to be effective in the treatment and control of Mycoplasma spp. *In vitro* studies demonstrated that resistance to tiamulin could not be acquired when MG and MS were passaged up to ten times in the presence of sub-inhibitory concentrations of this drug, whereas the same process resulted in the emergence of resistance against other compounds (Nhung et al., 2017). Continuing comparative genome analysis is expected to provide further insights into how this strain acquired intermediate resistance against tiamulin.

Mycoplasmas have been shown to acquire AMR either by mutations in specific genes or through gene transfer between different species, the latter has not been shown in poultry mycoplasmas so far. Studies on acquired resistance to macrolides in MG and MS indicated that single point mutations in one or both the 23S rRNA genes are responsible (Gerchman et al., 2011, Lysnyansky et al., 2013). As expected, all the MS strains and all but one of the MG strains had A2059G mutations in one or both 23S rRNA genes. Point mutation G745A in the 23S rRNA gene and amino acid substitutions G354A and H91K were found in the L4 and L22 proteins, respectively of *M. gallinarum*. Only the G354A mutation in the L4 protein was found in all tylosin resistant *M. gallinarum* strains which could be the primary marker for acquired macrolide resistance in this species. The mechanism of acquired macrolide resistance for *M. gallinaceum* is possibly linked to mutation G748A in the 23S rRNA gene, as this mutation was present in all but one of the tylosin resistant strains. Only one mutation was observed in the regions of interest for *M. pullorum*, but this was a susceptible strain, as such a mechanism of macrolide resistance could not be inferred. One MG and one *M. gallinaceum* strain did not contain the required mutation A2059G and G748A, respectively suggesting that other mechanisms of macrolide resistance are involved, therefore future studies aimed at proteomic analysis is required (Xia et al., 2015).

Quinolone resistance in MG and MS is acquired by point mutations in the quinolone resistance determining region of the DNA gyrase subunit A and B and Topoisomerase IV subunit A and B proteins (Gautier-Bouchardon et al., 2002, Lysnyansky et al., 2013). All MG, *M. gallinarum* and *M. pullorum* strains tested were either sensitive or intermediately sensitive to enrofloxacin. No mutations in the *gyr*A, *gyr*B*, par*C *or par*E genes were observed, except for one *M. pullorum* strain, with point mutation S81P, however this was not a resistant strain. In the case of *M. gallinaceum* only two of the 6 resistant strains had point mutation in the *par*C, but no other potential markers were observed. Thus, no probable mechanism of resistance could be inferred for these species. The single enrofloxacin resistant strain of MS had a D420N substitution in the *par*E gene which was suggested by Lysnyansky et al. (2013) as one of multiple possible markers for quinolone resistance in MS. All the MS strains also contained D454E and N84D substitutions in the *par*E and *par*C genes, respectively. The latter have shown to be associated with decreased susceptibility to quinolones, which could explain the intermediate susceptible phenotype, but some of the strains were sensitive to enrofloxacin. It is thus possible that point mutation D420N plays a larger role in determining resistance in MS, and further investigation is necessary. To my knowledge, is this the first time that a possible mechanism of acquired resistance to macrolides have been described for the avian mycoplasma species *M. gallinarum* and *M. gallinaeum.* Further investigation is required to identify the mechanism of acquired resistance of *M. pullorum* to macrolides and all three of these species to quinolones.

Bacteria are considered to be multi-drug resistant (MDR) if they acquired resistance to three or more antimicrobial classes (Magiorakos et al., 2012). Three *M. gallinaceum* strains showed multidrug resistance to oxytetracycline (a tetracycline), tylosin (a macrolide) and enrofloxacin (a quinolone). Proportionately more *M. gallinaceum* strains were tested compared to other mycoplasma species, therefore it is possible that other MDR mycoplasmas are circulating too. The *M. gallinarum, M. pullorum* and *M. gallinaceum* samples showed proportionally more AMR compared to MG and MS samples, and the frequent isolation from poultry flocks of non-pathogenic mycoplasma strains that acquired AMR is a cause for concern, especially since they commonly occur in co-infections with MG and MS and no vaccines against these less pathogenic species are available for their control.

Development of antibiotic resistance to oxytetracycline in *in vitro* studies has been difficult, indicating that it is more likely due to the transfer of the *tet*M from other species as has been shown for *M. hominis* (Roberts et al., 1985). This however has not yet been demonstrated in poultry mycoplasmas (Gerchman et al., 2011). Although natural horizontal gene transfer (HGT) between mycoplasma species has not yet been reported, Dordet-Frisoni et al. (2014) recently demonstrated that conjugal transfer, a form of HGT, between mycoplasma species is possible if an integrative conjugate element is present (Dordet-Frisoni et al., 2014). HGT has been put forward as a theory to explain the origin of the pMGA gene found in MG, that is closely related to the *vlh*A gene found in phylogenetically distant MS, but not found in other phylogenetically close mycoplasma species (Markham et al., 1999, Vasconcelos et al., 2005). Investigating the ability of poultry mycoplasmas for inter-and intra-specie AMR gene transferral, or even the uptake or transferral of AMR genes between mycoplasmas and other bacterial species in the same environment should be prioritized.

CHAPTER 3: GENOME ASSEMBLY AND ANNOTATION OF MYCOPLASMA PULLORUM, ISOLATED FROM DOMESTIC POULTRY IN SOUTH AFRICA

The content of this chapter was published as a *Genome Announcement* by A Beylefeld and C Abolnik, entitled "**Complete Genome Sequence of *Mycoplasma pullorum* Isolated from Domestic Chickens**" published online 23 February 2017.

3.1. Introduction

Various strategies have been used in the last couple of decades to assemble complete genomes for various species. The first strategies involved shearing the DNA into smaller sizes and sequencing each piece by Sanger-sequencing either randomly (also known as shot-gun sequencing) or directed using primer walking, but these methods are labour intensive and time-consuming. The first complete mycoplasma genome sequenced, *M. genitalium*, was sequenced using shot-gun sequencing with capillary electrophoresis. The first completed poultry mycoplasma genome, *M. gallisepticum* strain R_{low}, was also sequenced using shot-gun sequencing with capillary electrophoresis, but was followed by primer walking to close gaps (Fraser et al., 1995, Papazisi et al., 2003).

The introduction of second generation sequencing technologies (SGS) has made it possible to sequence the complete DNA complement of an organism in a single experiment, however this strategy produces large datasets containing billions of short read sequences that require computational resources for assembly of a complete genome (Besser et al., 2017). Strategies to assemble these reads are mainly by *de novo* assembly or mapping to a closely-related reference genome (Pop, 2009). The whole genome sequence for *M. gallinaceum* was completed using only *de novo* assembly of high-throughput Illumina data (Abolnik and Beylefeld, 2015). However, factors such as sequencing errors known to occur in SGS technologies, repeat regions and other factors influence the data output, resulting in a draft genome consisting of multiple scaffolds, rather than complete genomes (Pop, 2009, Ekblom and Wolf, 2014). Experimental methods, such as primer walking can be used to close the gaps to produce better quality genomes but are still time consuming and expensive. Hybrid methods combining data from different sequencing technologies have also been introduced with some success, however every organism is different, and the optimal strategy will depend on genomic characteristics, such as size, GC content and repetitive regions, and other external factors including budget and available resources (Ekblom and Wolf, 2014).

Mycoplasma genomes have a low GC content, contain numerous repeat region, and utilise a different genetic code, making producing complete genomes for species from this genus very difficult. *De novo* assembly strategies usually result in numerous contig sequences that will be too time consuming and expensive to assemble into a complete genome. The aim of this study was to assembly a complete genome for the previously uncompleted *M. pullorum* using whole genome sequencing data and *in silico* methods.

3.2. Materials and Methods

3.2.1. Sample collection, isolation and identification

Poultry mycoplasma samples were collected, isolated and identified as described in Chapter 2. Briefly, samples were collected by veterinarians from chickens using swabs and sent to the Bacteriology laboratory of the DVTD for the identification of mycoplasma species by culture with growth inhibition by Johan Gouws and Pamela Wambulawaye. The DNA of the mycoplasma-positive samples were isolated as described and sent for Ion Torrent PGM whole genome sequencing at UP before samples were identified using the 16S rRNA gene. The remainder of the samples were frozen at -20˚C for future downstream analysis. Sample B359-6 was also sent to Inqaba Biotech (Pty) Ltd, Pretoria for Illumina MiSeq whole genome sequencing.

3.2.2. Quality control

The fastq sequencing files produces by Ion Torrent PGM whole genome sequencing and Illumina MiSeq whole genome sequencing were submitted to the FASTQC program (version 0.11.5), (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to produce a quality control report and assess the amount and quality of reads and the presence of adapters (Andrews, 2010). The sequencing files were imported into CLC Genomics Workbench version 8.5.1 (CLC Bio-Qiagen, Aarhus, Denmark) using the platform specific import function. Low quality reads were trimmed and filtered, and sequencing adapters trimmed using the default settings of the Trim Sequences function of CLC Genomics Workbench with the Nextera Trim Adapter Library. The trimmed files were analysed again with FASTQC for quality control.

3.2.3. Sequence assembly

Single-end reads produced by the Ion Torrent sequencing platform were assembled *de novo* in CLC Genomics Workbench (version 8.5.1) using the default settings, and a minimum contig length of 500 bp. The reads were also mapped back to the contigs using the default settings with global alignment and saved for downstream analysis. *De novo* assembly of the Ion Torrent data was performed twice in CLC genomics workbench. As described in Chapter 2, Ion Torrent reads were also subjected to digital normalization using Khmer (version 2.0) (Brown et al., 2012, Crusoe et al., 2015) to decrease the amount of reads and submitted to the IonGAP server twice (available at http://iongap.hpc.iter.es/iongap) (Baez-Ortega et al., 2015), the first time using the Genome assembly and Bacterial classification module and a second time using only the Genome assembly module.

Paired-end reads produced by the Illumina MiSeq sequencing platform were also assembled *de novo* in CLC Genomics Workbench (version 8.5.1) using the default settings with the “include the paired-end reads to detect paired distances and perform scaffolding” option activated and produce only contigs with a minimum length of 500 bp. Illumina reads were also mapped back to the contigs using the default settings with global alignment and saved for downstream analysis.

The quality of each assembly was assessed and compared using Quast, a Quality Assessment Tool for Genome Assemblies from the Center for Algorithmic Biotechnology (available at http://quast.bioinf.spbau.ru/) (Gurevich et al., 2013). The complete genome assembly of sample B359-15-6 identified as *M. pullorum* was completed *in silico* using different strategies.

Strategy 1: De novo assembly with manual contig joining

The *de novo* assembled Ion Torrent contigs were aligned using the input contigs as reference with the default settings of the “align contigs” tool of the Genome Finishing Module (version 1.5.4) of CLC Genomics Workbench to produce a contig match table file. Starting with the largest contig, contigs were joined manually dependent on overlapping contigs at the 3’ and 5’ ends using the following parameters: 1) minimum contig match identity of above 95% and 2) minimum contig overlap length of 20 bp. Where multiple contigs aligned, the best fit was chosen for the join. When contigs could not be joined further the minimum contig match identity was lowered to above 80% to reduce the number of contigs to a single contig representing the whole genome. This process was continued until no more contigs could be joined.

Strategy 2: De novo assembly with manual contig joining from multiple genome assembly platforms

The *de novo* assembled Ion Torrent contigs produced were aligned and joined as described for strategy 1. Before the minimum contig match identity was lowered to 80%, the LargeContigs.fasta file produced by the IonGAP server was imported into CLC Genomics Workbench and the contigs added to the contig match table file. The joined contigs were extended using the contigs produced by the IonGAP server using the same parameters described above. As with strategy 1, when contigs could not be joined further, the minimum contig match identity was lowered to above 80% to reduce the number of contigs to a single contig representing the whole genome. This process was continued until no more contigs could be joined

Strategy 3: Hybrid genome de novo assembly with manual contig joining using multiple sequencing platforms and multiple genome assembling platforms with stepwise addition of each data set

The workflow shown in Figure 3-1 was followed starting with the largest contig until the full genome was assembled. Briefly the 5’ or 3’ end of the contigs were viewed to assess the possible matches for one of the four scenarios 1) when one possible match existed, the contigs are joined and the newly joined contig analysed again, 2) when multiple matches existed the matches were first compared to each other to determine if a) the matches are the same: the longest contig was then joined and the remaining matches were notes as part of the particular join b) the contig matches were not similar a copy of the file was made and every distinct contig match evaluated using each of the above scenarios, 3) when the 5' end matched to the 3' end the size of the contig was evaluated for possible completion of genome or noted as a possible repeat sequence and saved for resolve by downstream analysis, 4) when no matches were possible the contig was saved for resolution by downstream analysis. If multiple matches were possible the contig was not elongated and saved for downstream analysis.

Strategy 4: Hybrid genome de novo assembly with manual contig joining using multiple sequencing platforms

All three sets of *de novo* assembled contigs produced in CLC Genomics Workbench for Illumina and Ion Torrent sequencing data were pooled with the contigs produced by the two IonGap assemblies and a contig match table produced in CLC Genomics Workbench using the default settings of BLAST word size of 20 and minimum match size of 100. The workflow described in Figure 3-1 was followed. The end was determined using scenario (3) where the 5' end matched to the 3' end and the genome size was in range with the expected size of the mycoplasma genome as determined by the total length of combined contigs obtained from the genome assembly statistics. The remaining contigs were analysed for the following scenarios 1) if there is a high contig match percentage, the contig was removed, 2) no contig matches and short contig length and low read coverage contigs were removed, 3) no contig match and high coverage, the contig was exported and submitted to the National Centre for Biotechnology Information (NCBI) nucleotide BLAST webtool (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi) (Zhang et al., 2000) for identification.

The final contig was exported from the contig match table and the Illumina and Ion Torrent reads were mapped onto the contig separately and a report generated for evaluation.

3.2.4. Genome annotation and viewing

The genome was then submitted for annotation to the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016). The resulting Genbank® file was downloaded from Genbank®® and a complete circular genome was viewed using the custom analysis pipeline of the online server G-view: a circular and linear genome viewer, see appendix B for the style sheet (available at https://server.gview.ca/#) (Petkau et al., 2010). A complete genome analysis was also produced by the US Department of Energy Joint Genome Institute (DOE-JGI) in collaboration with the user community and presented on the Integrated Microbial Genomes and Microbiomes (IMG) website (available at https://img.jgi.doe.gov/) (Chen et al., 2017). The resulting genome statistics and results from the DOE-JGI Microbial Genome Annotation Pipeline (MGAP) were viewed (Figure 3-2) (Huntemann et al., 2015).

The genome was also reanalysed using the NCBI-PGAP pipeline by the NCBI team in 2017 and annotated as a reference sequence. The protein files for the two NCBI-PGAP annotations were exported and submitted to the online webserver WebMGA (available at http://weizhong-lab.ucsd.edu/webMGA/server/) for functional analysis using the Clusters of Orthologous Group (COG) categorisation of proteins (Wu et al., 2011). The COG classification of proteins generated for each annotation were compared and results correlated.

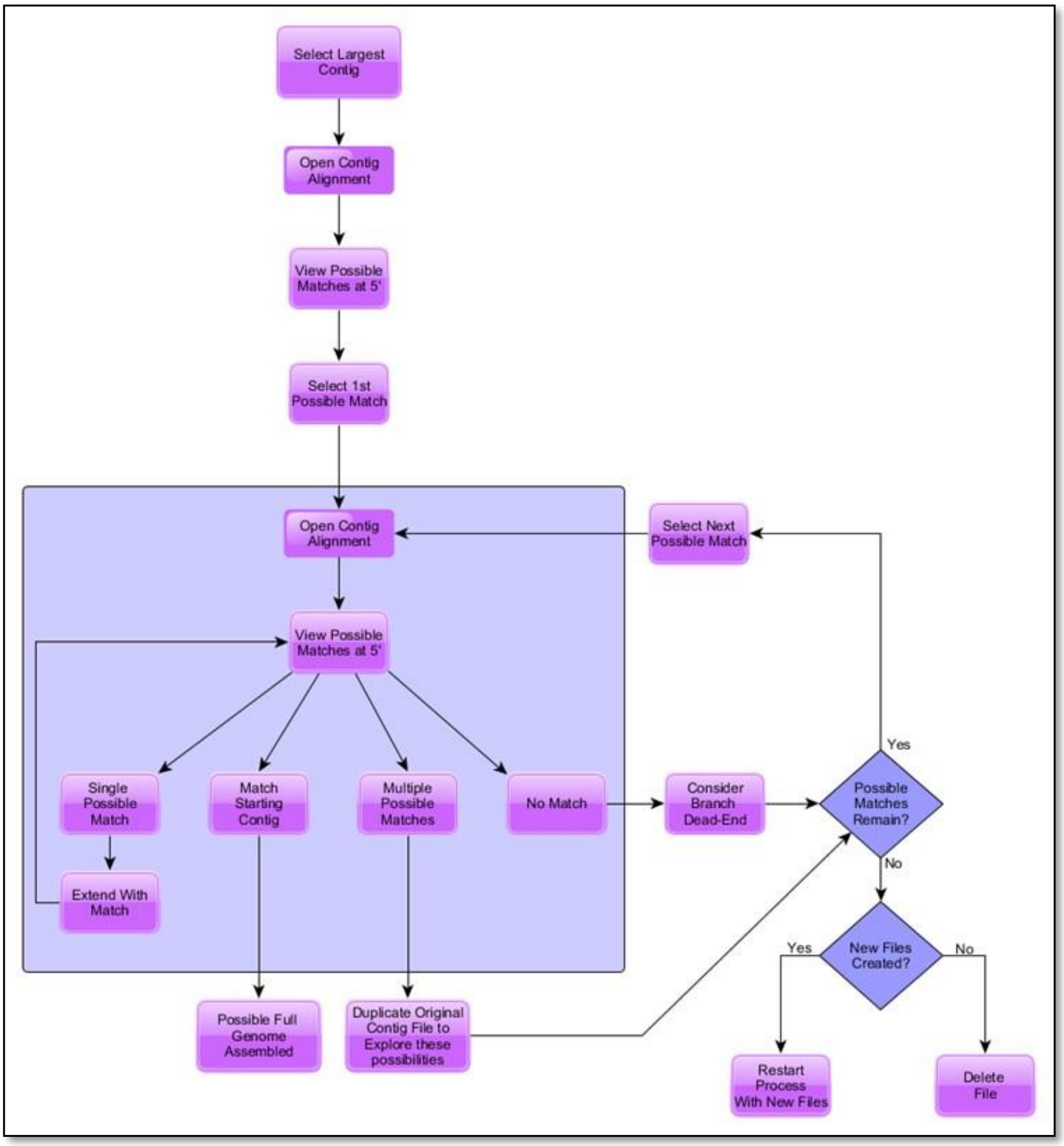

Select Largest Contig
Open Contig Alignment
View Possible Matches at 5'
Select 1st Possible Match
Open Contig Alignment
View Possible Matches at 5'
Select Next Possible Match
Single Possible Match
Match Starting Contig
Multiple Possible Matches
No Match
Consider Branch Dead-End
Possible Matches Remain?
Yes
No
Extend With Match
New Files Created?
Yes
No
Possible Full Genome Assembled
Duplicate Original Contig File to Explore these possibilities
Restart Process With New Files
Delete File

Figure 3-1: Workflow for manual joining of contigs using the Genome finishing tool of CLC Genomics Workbench

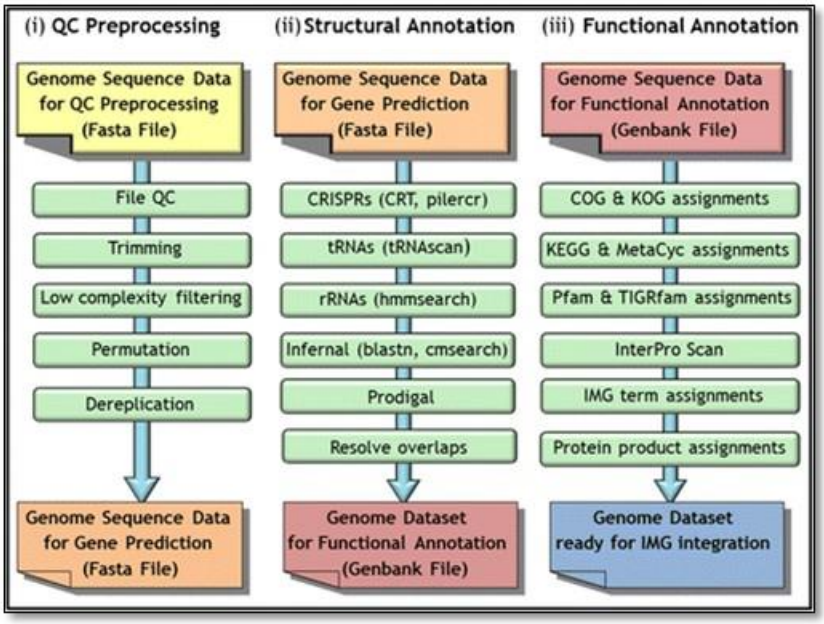

(i) QC Preprocessing
(ii) Structural Annotation
(iii) Functional Annotation
Genome Sequence Data for QC Preprocessing (Fasta File)
File QC
Trimming
Low complexity filtering
Permutation
Dereplication
Genome Sequence Data for Gene Prediction (Fasta File)
Genome Sequence Data for Gene Prediction (Fasta File)
CRISPRs (CRT, pilercr)
tRNAs (tRNAscan)
rRNAs (hmmsearch)
Infernal (blastn, cmsearch)
Prodigal
Resolve overlaps
Genome Dataset for Functional Annotation (Genbank File)
Genome Sequence Data for Functional Annotation (Genbank File)
COG & KOG assignments
KEGG & MetaCyc assignments
Pfam & TIGRfam assignments
InterPro Scan
IMG term assignments
Protein product assignments
Genome Dataset ready for IMG integration

Figure 3-2: Genome annotation pipeline that forms part of the standard operating procedures of IMG genome analysis which include i) a quality control step, ii) structural annotation of genome and iii) functional annotation of genome

3.3. Results

3.3.1. Quality control

The Illumina and Ion Torrent sequencing data files received after sequencing were submitted to the FastQC application for quality control. The basic statistics, shown in Table 3-1, shows that Ion Torrent sequencing produced 4 048 281 reads with an average read length of 176 bp and assuming an expected genome size of 1 Mbp resulted in a depth of coverage of about 707 times. The FastQC program is optimised for Illumina data, and it appears as though the Ion Torrent data had reasonable or poor quality calls (Figure 3-3A, the rest of the report is available upon request due to size of files), however the average quality score of Ion Torrent reads is generally accepted to be distributed in the range of 17 to 30 compared to Illumina data distribution range between 20 and 35. The average quality scores across each base position range between 18 and 23, no overrepresented sequences or adapters were found, thus quality of the Ion torrent reads were considered as good and no trimming and filtering performed. To reduce the number of reads for submission to the IonGap pipeline, the Ion Torrent reads were subjected to digital normalization producing 898 956 reads with length ranging from 32 – 305 bp and a GC content of 28%.

Illumina paired-end sequencing files are released as two separate files, a set of forward reads and a set of reverse reads. Each of these files were checked for quality control in FastQC (Figure 3-3C and D). The forward reads gave a warning for the per base sequence quality, due to the last base position falling into the adequate and poor quality range, and the reverse reads failed the per base sequence quality with the last almost 100 bp positions falling in the adequate and poor quality range.

During import of the Illumina paired reads files into CLC Genomics Workbench the file is automatically combined into a single paired file. Possible adapter sequences, poor quality reads and quality bases were trimmed and filtered using the Nextera adapter library and the default settings in CLC Genomics Workbench to produce a paired-trimmed sequencing file. The quality of the trimmed file was checked again in FastQC (Figure 3-3B). The trimmed sequences passed the per base sequence quality and no overrepresented sequences or adapter content were present. All the quality control reports indicated either a warning or failure of the per base sequence quality and per sequence GC content determinations (available upon request). Illumina sequencing produced two read files of 166 659 bp each with an average read length of 185 bp and depth of coverage of 30.6x, combined 332 828 bp with an average read length of 185 and depth of coverage of 61.1 times was produced (Table 3-1).

Table 3-1: Basic read statistics of sequencing files collected from FastQC quality control statistics and CLC quality control statistics

Sequencing Platform	**Ion Torrent reads**	**Illumina reads before trimming (Forward)**	**Illumina reads before trimming (Reverse)**	**Illumina reads after trimming**
Total sequences	4 048 281	166 659	166 659	332 828
Sequences length (bp)	25 – 313	35 – 301	35 – 301	3 – 301
Average read length (bp)	176	185	185	185
%GC	29	34	34	34
Depth of coverage	707.4x	30.6x	30.6x	61.1x

3.3.2. Sequence assembly

An overview of the *de novo* assembly results used in this study is given in Table 3-2. Ion Torrent reads were *de novo* assembled in CLC genomic workbench twice and produced 143 (Ion Torrent 1) and 150 (Ion Torrent 2) contigs with combined lengths of 1 024 434 bp and 1 041 976 bp, respectively, the largest contigs were 136 443 bp and 73 130 bp, respectively. The minimum contig length covering 50% of the genome (N50) was found to be 31 904 bp and 35 243 bp, respectively. *De novo* assembly of the Illumina reads resulted in 147 contigs with a combined length of 964 189 bp, the largest contig was 47 419 bp and the N50 15 299 bp (Table 3-2).

De novo assembly in IonGap was also performed twice, a first time for identification as discussed in Chapter 2 and a second time only using the genome assembly module, both submissions generated assemblies that were used in this study. The IonGap pipeline uses the assembly program MIRA that exports only the large contigs, for downstream analysis. The first submission resulted in 41 large contigs (IonGap1) with a combined size of 1 022 315 bp and the largest contig 189 531 bp, which is also the largest contig of all assembly data sets. The second submission resulted in 34 large contigs (IonGap2) with a combined length of 1 019 999 bp, and the largest contig 140 233 bp.

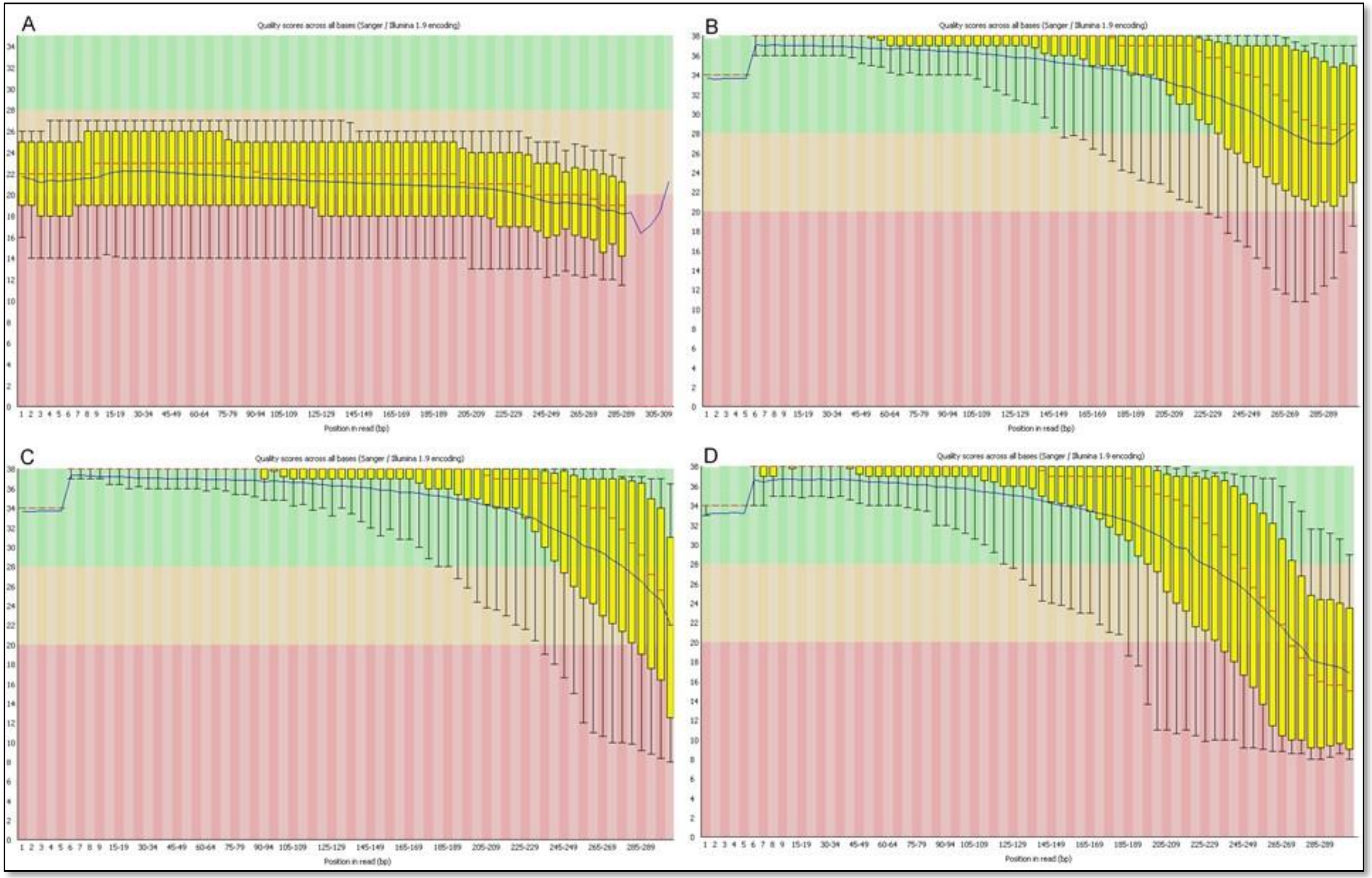


Figure 3-3: Per base sequence quality as presented by FastQC quality control. A) Ion Torrent sequencing data B) Paire d and trimmed Illumina sequencing data C) Untrimmed forward Illumina reads file D) Untrimmed reverse Illumina reads file. The y-axis shows the quality scores and is divided into a green, orange and red block representing very good, reasonable and poor quality calls, respectively. The x-axis the position in a read bp. The box-whisker plot represents each nucleotide position with the central red line indicating the median value, the yellow box representing the inter-quartile range (25-75%), the upper and lower whiskers representing the 10 and 90% range and the blue line indicates the mean quality across e ach base position

Assembly of the Illumina reads resulted in the smallest combined size and less large contigs, producing no contigs of 50 000 bp or larger compared to 4 and 6, respectively from the Ion Torrent 1 and 2 assemblies and 4 and 7 contigs, respectively of the IonGap1 and 2 assemblies. The four assemblies utilising the Ion Torrent sequencing data had an average GC content of 28.87% compared to 29% from Illumina data (Table 3-2).

Table 3-2: *De novo* assembly statistics of datasets used in this study generated in Quast

Assembly	Ion Torrent 1	Ion Torrent 2	Illumina	IonGap1	IonGap2
Number of contigs	143	150	147	41	34
Number of contigs (>= 1000 bp)	61	66	111	29	28
Number of contigs (>= 5000 bp)	34	37	58	21	20
Number of contigs (>= 10000 bp)	23	24	35	18	18
Number of contigs (>= 25000 bp)	13	16	7	13	14
Number of contigs (>= 50000 bp)	4	6	0	4	7
Total length (>= 0 bp)	1024434	1041976	964189	1022315	1019999
Total length (>= 1000 bp)	963890	979910	938070	1012794	1015899
Total length (>= 5000 bp)	914148	924295	802460	990108	996360
Total length (>= 10000 bp)	833157	827187	651788	970188	982323
Total length (>= 25000 bp)	654775	716010	218198	889088	919267
Total length (>= 50000 bp)	328750	364365	0	570146	669731
Largest contig	136443	72130	47419	189531	140233
Total length	1024434	1041976	964189	1022315	1019999
GC content (%)	28.88	28.86	29	28.88	28.84
N50*	31904	35243	15299	95808	66725
N75*	16379	13645	6813	34044	41551
L50#	9	10	22	4	5
L75#	19	21	45	9	10
Number of N's per 100 kbp	0	0	1.56	0.78	1.18

*N50 and N75 – minimum contig size representing 50% or 75%, respectively of the of the genome size
#L50 and L75 – smallest number of contigs whose size represents 50% or 75%, respectively of the genome size

Various strategies were attempted to assemble the complete genome for *M. pullorum* using only *in silico* methods. The data for the first three strategies was unfortunately lost, so the failed strategies were only mentioned briefly, and the main focus of the discussion was on the successful strategy used for the reconstruction of the genome for *M. pullorum.* The first strategy only used the contigs produced by Ion Torrent 1 assembly, joining the contigs, based first on a 95% contig overlap match identity and was decreased to above 80%. When multiple contigs could be joined only the best fit was joined, a single contig could not be obtained reliably and the strategy was re-evaluated (data not shown). The second strategy involved using assembly data from 2 assembly programs, CLC genomics workbench and the IonGap pipeline that uses the MIRA assembler. First the Ion Torrent 1 and Ion Torrent 2 contigs were joined using the CLC genome finishing module based on 95% contig overlap match identity and a single possible match. The IonGap1 and IonGap2 contigs were added to the pool of contigs to extend the joined contigs based on 95% contig overlap match identity and a single possible match. This decreased the number of joined contigs, but a single contig could not be reliably obtained and the strategy was re-evaluated again (data not available).

The third strategy started with the Ion Torrent data sets, the IonGap data sets were when contigs could not be joined with a high level of confidence, followed by the Illumina data set. The paired-end information was used as a guide for determining the correct direction of the joined contig, however this strategy was very time consuming and the multiple files generated for each possible match became a cumbersome and resource intensive task and the strategy was abandoned for a more condensed fourth and final strategy (data not available). This strategy involved combining all five of the assembly data sets, Ion Torrent 1, Ion Torrent 2, Illumina, IonGap1 and IonGap2, using the Genome finishing module in CLC genomics workbench and systematically joining the contigs. The contig match table started with the 515 combined contigs and after 98 contig joinings, 416 contigs were left with the final large contig (Data not shown). The remaining contigs were checked against the final contig and to each other in two separate contig match tables. 162 contigs could not be matched to the final contig, upon closer inspection 140 of these contigs had an average contig length of 788 bp and average read coverage of 2.88x and omitted from further analysis. The remaining contigs were submitted to the NCBI nucleotide BLAST webtool using the default settings. Eleven of the contigs had no hits and 8 of the 11 remaining contigs matched on less than 3% of the contig lengths and were also omitted from further analysis. The remaining three contigs, Ilumina contig 42 (3828 bp), Illumina contig 74 (3998 bp) and Illumina contig 138 (size 758 bp), matched 31.81%, 50.87% and 100% of the respective contig lengths to a putative *M. edwardii* gene, DNA gyrase subunit B gene of *M. edwardii* and a putative gene of the Equine encephalosis virus. Initial analysis of the first two contigs in 2016 matched less than 3% of the contigs to genes in other mycoplasma species (data not shown), at the time of writing this thesis in 2018 the analysis was repeated, and a more significant hit was found to the newly released complete genome of *M. edwardii*. As these hits were only found in the Illumina reads, it was considered as contamination at the sequencing facility and was omitted from further analysis. Some of the results files are too large to include in the thesis and is available upon request.

The final contig was 1 007 271 bp in length and 29.1% GC content and 95.02% and 95.96% of the Ion Torrent and Illumina reads mapped to the contig, respectively. The contig was exported and annotated by the NCBI-PGAP pipeline and published under accession number CP017813.

3.3.3. Genome annotation and viewing

With initial submission for annotation in the NCBI-PGAP pipeline a total of 814 genes were found (Table 3-3). Ten pseudo genes and 763 coding genes produced a total of 773 coding sequences (CDS) (94.96%). The remaining 41 genes (5.03%) were RNA genes that were made up of one 5S rRNA gene, two of each 16S and 23S rRNA genes, 34 transfer RNAs (tRNAs) and 2 non-coding RNAs (ncRNAs). The file was automatically submitted to the IMG-MGAP pipeline in March of 2017 and the analysis produced 822 genes, consisting of 783 coding genes (95.26%) and 39 rRNAs (4.74%) (Table 3-3). The NCBI Genbank® file was also automatically updated in April 2017 and produced 825 genes consisting of 784 coding genes (95.03%) and 41 RNA genes (4.97%) and updated on the NCBI Genbank® database as a reference sequence under accession number NZ_CP017813. The latter two automatic annotations also produced one 5S rRNA gene, two of each 16S and 23S rRNA genes, 34 tRNAs, but only the updated NCBI-PGAP produced the additional 2 ncRNAs present in the original NCBI-PGAP analysis (Table 3-3).

Table 3-3: General features produced by NCBI-PGAP and IMG-MGAP

	NCBI-PGAP (2016)[1]	**NCBI-PGAP (2017)[2]**	**IMG-MGAP (2017)[3]**
Genome size	1007172	1007172	1007172
Genes (Total)	814	825	822
CDS (Total)	773 (94.96%)	784 (95.03%)	783 (95.26%)
CDS (Coding)	763	758	783
Pseudo genes	10	26	Not reported
COG classification	437	438	488
RNA	41 (5.04%)	41 (4.97%)	39 (4.74%)
5S rRNA	1	1	1
16S rRNA	2	2	2
23S rRNA	2	2	2
tRNA	34	34	34
other RNA	2	2	Not reported

1 – Original NCBI-PGAP analysis upon submission of genome in 2016
2 – Updated NCBI-PGAP analysis performed in 2017
3 – IMG-MGAP analysis performed by DOE-JGI in 2017

Using the COG database, 437 (57.27%), 438 (57.78%) and 488 (62.32%) of the total coding CDS each of the three annotations were assigned to a COG category (Table 3-3). The most proteins were assigned to the COG category J: Translation, ribosomal structure and biogenesis; category G: Carbohydrate transport and metabolism and category L: Replication, recombination and repair for all three annotations (Figure 3-4). The least amount of proteins was assigned to category M: Cell wall/membrane/envelope biogenesis, category N: Cell motility and category X: Mobilome: prophages, transposons for all three annotations (Figure 3-4). A correlation was drawn for the COG categorization between the three annotations sets (Figure 3-5). The NCBI-PGAP (2016) annotation had a correlation coefficient of 0.999991 and 0.998 with the NCBI-PGAP (2017) and IMG_MGAP annotations, respectively, and the correlation coefficient between the NCBI-PGAP (2017) and IMG-MGAP annotations were 0.998 (Table 3-4).

A genome map of *M. pullorum* B359_6 was drawn in Gview from the updated NCBI-PGAP analysis, depicting the location and direction of CDS, the prokaryotic COG protein categories and %GC content of the genome (Figure 3-6). Of the 773 CDS identified were 400 annotated as hypothetical proteins.

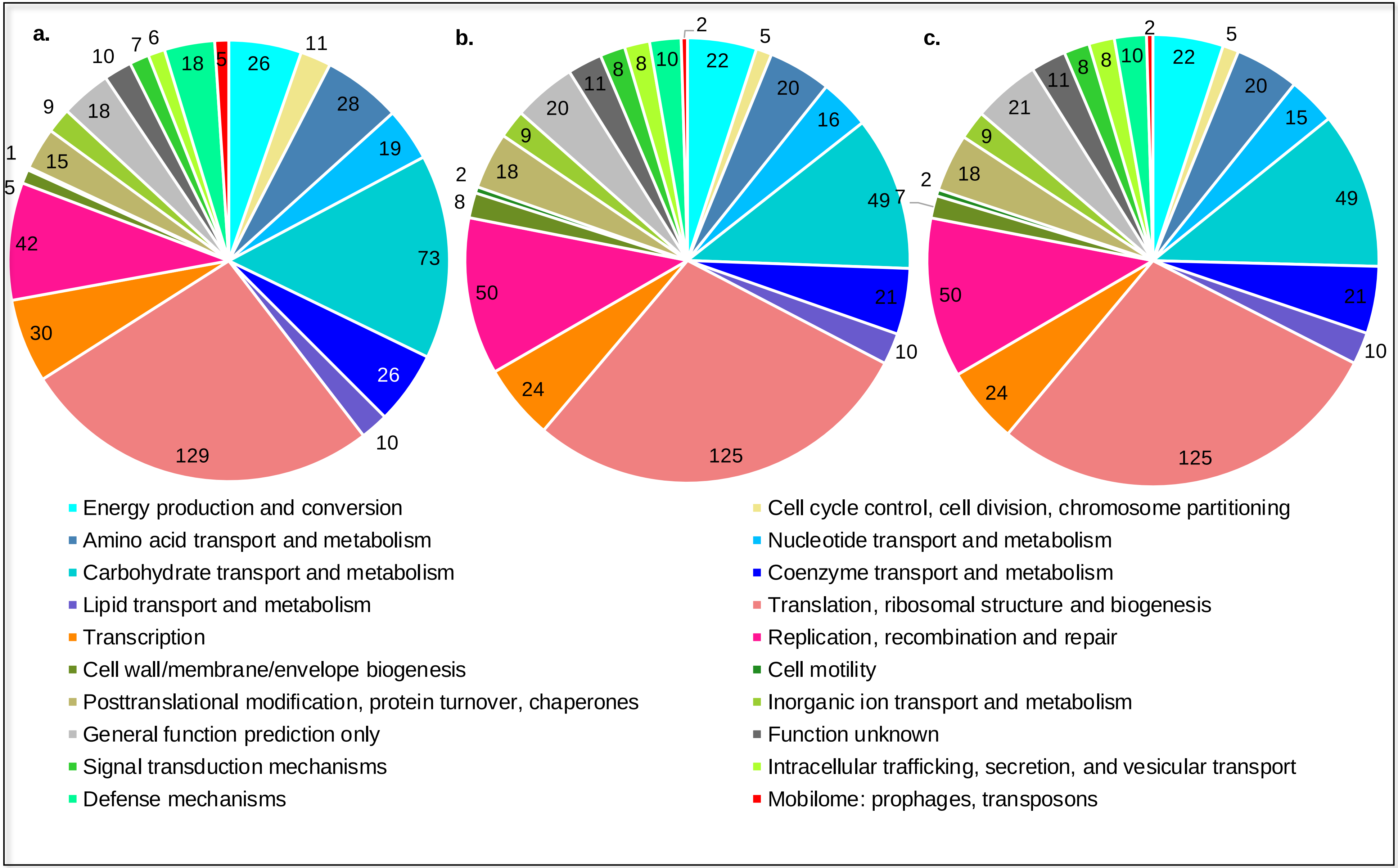


Figure 3-4:COG categories of annotations generated by the a) IMG-MGAP (2017), b) NCBI-PGAP (2017) and c) NCBI-PGAP (2016).

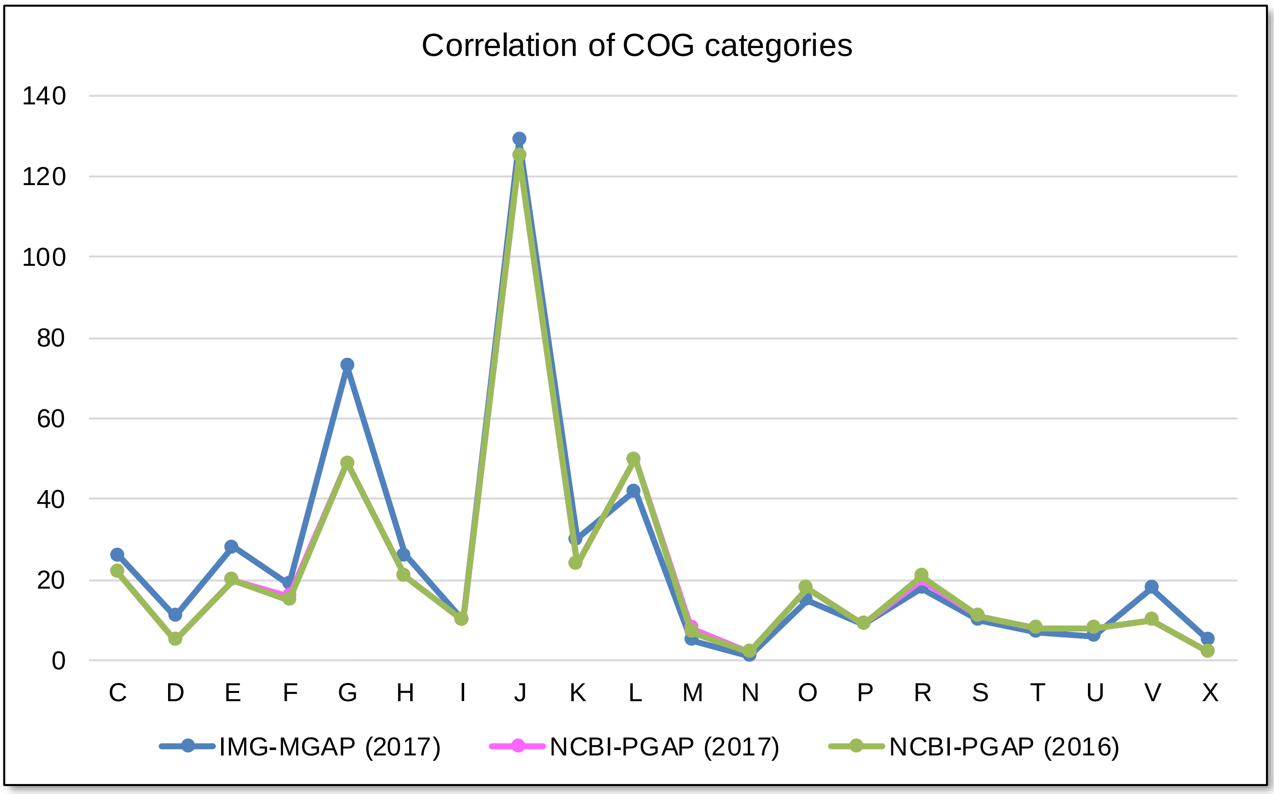


Figure 3-5: Correlation of COG categories produced by IMG and NCBI-PGAP pipelines. COG categories found in M. pullorum are C - Energy production and conversion, D - Cell cycle control, cell division, chromosome partitioning, E - Amino acid transport and metabolism, F - Nucleotide transport and metabolism, G - Carbohydrate transport and metabolism, H - Coenzyme transport and metabolism, I - Lipid transport and metabolism, J - Translation, ribosomal structure and biogenesis, K - Transcription, L - Replication, recombination and repair, M - Cell wall/membrane/envelope biogenesis, N - Cell motility, O - Posttranslational modification, protein turnover, chaperones, P - Inorganic ion transport and metabolism, R - General function prediction only, S - Function unknown, T - Signal transduction mechanisms, U - Intracellular trafficking, secretion, and vesicular transport, V - Defense mechanisms and X - Mobilome: prophages, transposons

Table 3-4: Table of correlation between COG categories of proteins found in the annotations generated

	IMG-MGA	**NCBI PGAP (2017)**	**NCBI (2016)**
IMG	1		
NCBI-PGAP (2017)	0.998089	1	
NCBI-PGAP (2016)	0.998063	0.999991	1

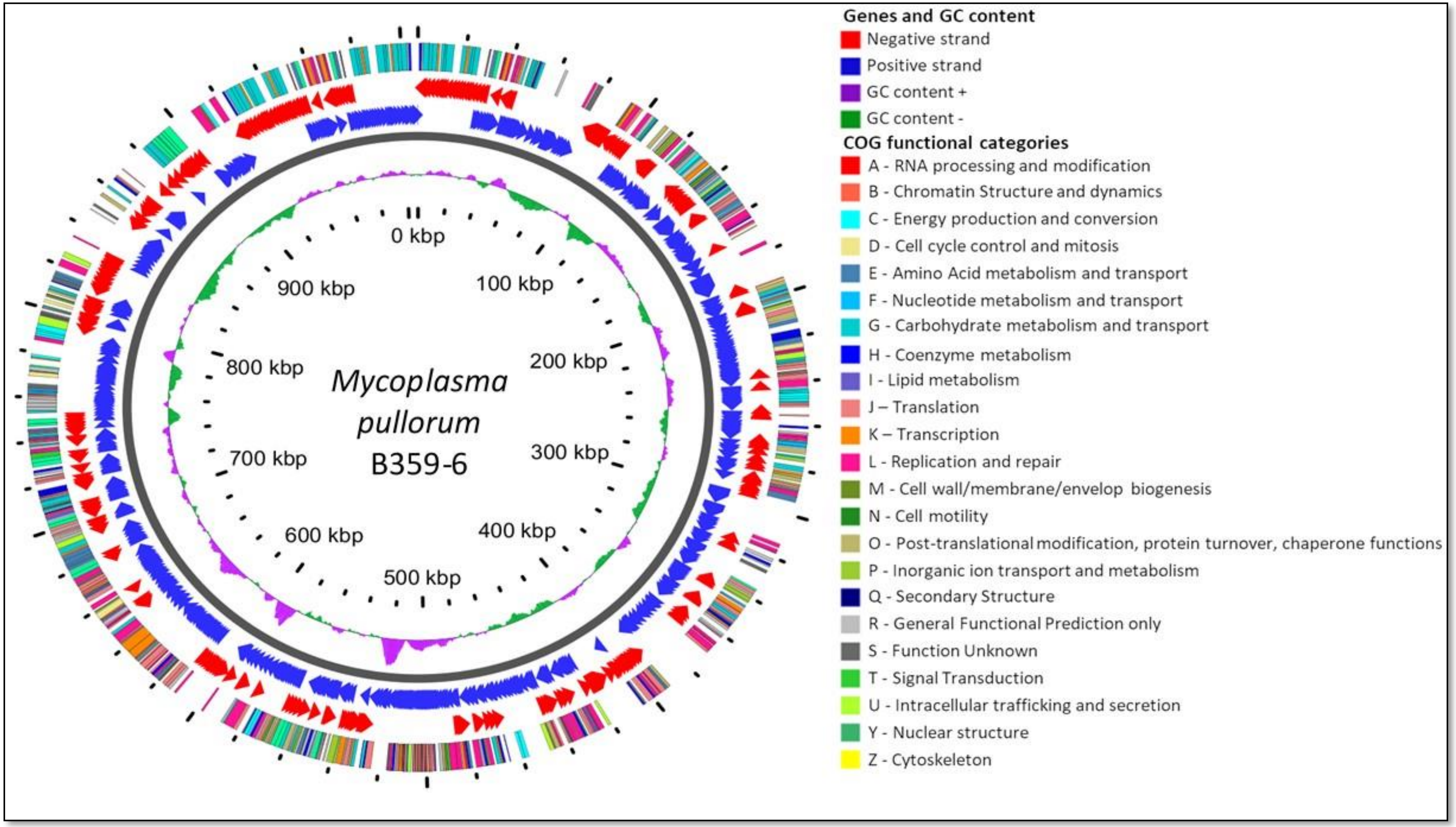

Genes and GC content
Negative strand
Positive strand
GC content +
GC content -
COG functional categories
A - RNA processing and modification
B - Chromatin Structure and dynamics
C - Energy production and conversion
D - Cell cycle control and mitosis
E - Amino Acid metabolism and transport
F - Nucleotide metabolism and transport
G - Carbohydrate metabolism and transport
H - Coenzyme metabolism
I - Lipid metabolism
J – Translation
K – Transcription
L - Replication and repair
M - Cell wall/membrane/envelop biogenesis
N - Cell motility
O - Post-translational modification, protein turnover, chaperone functions
P - Inorganic ion transport and metabolism
Q - Secondary Structure
R - General Functional Prediction only
S - Function Unknown
T - Signal Transduction
U - Intracellular trafficking and secretion
Y - Nuclear structure
Z - Cytoskeleton
0 kbp
100 kbp
200 kbp
300 kbp
400 kbp
500 kbp
600 kbp
700 kbp
800 kbp
900 kbp
Mycoplasma
pullorum
B359-6

Figure 3-6: ***M. pullorum*** **strain B359-6 complete genome map visualised in Gview. Rings represent the following starting from the inner ring 1) genome position in genomes, 2) GC content on forward (green) and reverse (purple) strand of genomes 3) Genes on the positive strand (blue) 4) Genes on negative strand (red) and 5) Protein coding genes using the COG functional categories.**

3.4. Discussion

Various bioinformatic programs, both open-source and proprietary, have been developed for use with SGS technologies. Illumina sequencing has been used more frequently than Ion Torrent sequencing and most open-source bioinformatics tools have been optimised for use with Illumina data. One of the biggest advantages of proprietary programs is the 1) ability to handle data from different platforms, 2) run on most operating systems and 3) a user-friendly interface, compared to open-source software that are mostly Linux-based (Smith, 2015). Proprietary software also has several plug-ins available that allow for multiple functions in a single program using workflows and pipelines (Smith, 2015). However, most of the open-source tools are powerful and numerous studies are available comparing these tools to determine the best tool for handling platform specific data and the problems associated with these platforms, the problematic characteristics of some genomes, as well as for specific applications.

Illumina and Ion Torrent sequencing technologies are completely different, with Illumina sequencing based on reversible dye terminator technology and Ion Torrent sequencing on semi-conductor technology measuring basic chemistry (Goodwin et al., 2016b, Van Dijk et al., 2014). This has an influence on how results from these programs should be interpreted, starting with the quality control parameters used in FastQC, a commonly used quality control program for SGS data (Del Angel et al., 2018). Quality score for Illumina and Ion Torrent sequencing technologies are determined based on different factors, which will influence the Phred scores reported by programs, such as FastQC. Generally the Phred scores results are lower for Ion Torrent reads than for Illumina reads, an example of this is shown in a study done by Utturkar et al. (2015) where Ion Torrent and Illumina sequencing data for the same species was compared.

The FastQC results for Ion Torrent data can thus be re-evaluated using less stringent parameters and Phred scores in the range of 17 to 30 are considered as good quality and is in range of what was obtained in this study. Other factors, such as the genome characteristics of the organism of interest also play a role in how results from any bioinformatic analysis should be interpreted. One of the most notable characteristics of most mycoplasma genomes is a low GC content, therefore the warning and failure of the per base sequence content and per sequence GC content was not unexpected and could be ignored when evaluating the quality control report. Thus, the Ion Torrent reads passed quality control and no trimming or filtering was required before *de novo* assembly. The Illumina data passed quality control after adapter sequences were trimmed and low-quality sequence reads and bases were filtered and trimmed. Ion Torrent sequencing produced the most reads and highest coverage (707x) compared to Illumina data (60x), but the coverage for both technologies are more than enough for assembling a complete microbial genome.

Various measures are used to assess the assembly of reads into contigs, including N50, contig count and size. However, there are no clear guidelines to determine if an assembly is good, especially if a reference genome is not available (Wajid and Serpedin, 2014, Del Angel et al., 2018). Quast is a widely used assembly assessment tool to evaluate an assembly and compare assemblies from different platforms and programs (Del Angel et al., 2018). The assembler used in CLC genomic workbench uses the *De Bruijn* graph method to assemble reads into contigs and the program uses multithreading to speed up the assembly, thus running the same data set with the same settings can result in different results (Qiagen, 2016). The two assembly runs were similar, but the major differences are found in the length of the largest contigs, which could aid in joining contigs that were not extended in the assembly process due to possible repeat regions.

The IonGap pipeline uses the MIRA assembler for the Genome assembly module, which is the recommended assembler for Ion Torrent reads. MIRA uses a combination of *De Bruijn* graphs and OLC graphs and uses specific algorithms dependent on the read coverage information to determine the minimum contig size for contigs that are recommended for inclusion as large contigs and downstream application. The *de novo* assemblies generated by the IonGap pipeline are overall the best when comparing the N50 value, largest contig size and the L50 values, however these metrics were calculated using the large contig output and not the complete list of contigs generated. Comparing the assemblies generated in CLC genomics workbench indicates that the assembly of the Ion Torrents reads were better than the assembly of the Illumina reads, even though the amount of contigs were almost the same, the N50 for the Ion Torrent reads were more than double the size and the largest contigs were 3 times and 1.5 times larger, respectively. The information contained in the paired-end was used to aid in the correct direction of contigs when joined (Glenn, 2011).

Four strategies to assemble a complete genome using only *in silico* methods were attempted. The first two methods were based on the most frequently used strategy for genome assembly, utilising only data from one sequencing platform, the main difference being that the second strategy uses data from two assemblers (Edwards and Holt, 2013). For bacteria and viruses, these strategies have the potential to produce complete genomes, but various factors, including genetic repeats and GC content can be difficult for the current assemblers to handle and mainly results in draft genome assemblies consisting of multiple contigs or scaffolds. Different assemblers have different strengths and weaknesses, and utilising multiple assemblers can aid in producing better assemblies. The quality of the draft assemblies produced using these strategies are generally good enough for downstream applications such as genome comparisons and variant calling, if the regions of interest have been successfully sequenced (Edwards and Holt, 2013). The last two strategies were hybrid assemblies using data from different sequencing platforms as well as different assemblers. The main difference between these strategies was how the data of each assembly is added, starting with all the datasets from the beginning was less time and resource consuming. Using the fourth strategy, a complete genome for *M. pullorum* was assembled, even though two sets of CLC-generated and IonGap-generated assemblies were used, one of each should also be enough with the Illumina dataset for use as a good strategy for assembling the low GC, highly repetitive mycoplasma genomes in a less resource intensive and time-consuming manner than previous assembly strategies.

The complete genome was annotated using one of the most widely used automatic annotation pipelines, the NCBI-PGAP pipeline, developed by NCBI for annotation of prokaryotic genomes with quality checks for minimum standards of complete genomes, i.e. 1) at least one copy of 5S, 16S and 23S rRNAs, 2) at least one copy of tRNA for each amino acid, 3) ratio of protein coding regions to the genome size divided by 1000 close to 1, 4) No gene is completely contained in another and 5) no partial features are present (Tatusova et al., 2016). If a submission is considered as a reference sequence the genome is reannotated and given a new accession. The complete genome of *M. pullorum* strain B359-6 was 1 007 271 bp in length with a GC content of 29.1% and was successfully annotated and published online under accession number CP017813. As this is the first time a complete has been published for *M. pullorum*, and met the prerequisites for classification as a reference sequence, this entry was reannotated and added to the NCBI Reference Sequence Database RefSeq under accession number NZ_CP017813.1.

Reference data submitted to the NCBI's Genbank® database is also automatically imported to the IMG-MGAP pipeline for a complete genome analysis and annotation by the DOE-JGI group. The aim of this group is to provide a resource to the scientific community for the analysis, annotation and comparison of genomic and metagenomic data, as quickly as possible to aid in the development of science (Chen et al., 2017). The data analysis provided by this platform is very extensive and lays the ground work for various analysis, including comparative genome analysis. Comparing the COG categorization of these three annotations showed a positive correlation between the annotations. Only about 57% of the coding CDS could be assigned a function associated with a COG category. As expected from the general characteristics of mycoplasma genomes already discussed and a study comparing COG categories of similar genes in various mycoplasmas species, the proteins that play a role in translation and ribosomal structure, as well as in replication, recombination and repair contained were the most abundant and conserved, while proteins involved in cell membrane biogenesis were the least abundant and conserved. The NCBI-PGAP pipeline annotated 52% of the identified genes as hypothetical proteins, thus the functions of more than half of the genes of *M. pullorum* is unknown and laboratory methods are required to study these genes in order to identify their function and produce a more complete annotation of the *M. pullorum* genome.

To the best of my knowledge is this the first time that a hybrid genome assembly strategy, using different assemblers and sequencing platforms, has been used to assemble a poultry mycoplasma genome. However, there are some caveats to this approach, i.e. it is still fairly time consuming, computationally complex as well prone to the inherited problems associated with second generation technologies, such as accuracy in homopolymer regions and repeat regions. In future studies, this strategy can be combined with long read sequencing technologies, such as PacBio or nanopore sequencing in a true hybrid genome assembly strategy to assemble complete genomes of the remaining poultry species for which a full genome is not available yet, such as *M. gallinarum*, as well as improve on current available genomes.

CHAPTER 4: COMPARATIVE GENOME ANALYSIS OF MYCOPLASMA SPP ISOLATED FROM SOUTH AFRICAN POULTRY

4.1. Introduction

Poultry mycoplasmas are usually associated with chronic respiratory disease that is difficult to eradicate. Prevention strategies include a good monitoring system for early detection of infection in poultry flocks as well as vaccination. Commercially available enzyme-linked immunosorbent assays (ELISAs) are commonly used to monitor *Mycoplasma gallisepticum* (MG) and *M. synoviae* (MS) infection in the South African poultry industry, but serologic cross-reaction and other factors can influence the sensitivity of these assays. Culture or DNA-based methods are required to confirm infections (Feberwee et al., 2005, OIE, 2008). Advances in molecular techniques that are less time-consuming and generally more sensitive and can be used to distinguish different strains within a species has driven mycoplasma diagnostic research towards more DNA-based methods.

Various genes have been used in PCR-based methods to distinguish mycoplasma species, such as the 16S rRNA and 16S-23S intergenic spacer region (IGSR). Genes encoding surface proteins and that play a role in pathogenesis of mycoplasmas, such as the *mgc2*, *gapA*, *pvpA* or MGA_0309 have been used to distinguish between MG strains and the *vlhA* gene have been used to distinguish MS strains (OIE, 2008, García et al., 2005, Hong et al., 2004). However, genotypic differences in MG isolates from different countries, such as the Southern African countries South Africa and Zimbabwe can also influence the sensitivity and specificity of these test and was noted to be a possible cause for the reduced efficacy observed in current vaccine strategies (Moretti et al., 2013). Numerous other genes have been found to play a role in mycoplasma pathogenesis as discussed in Chapter 1, and numerous more genes have been found in mycoplasma genomes that have been classified as hypothetical proteins for which the functions are not yet known, whereas other genes are yet to be discovered as the complete pathogenesis of mycoplasmas are not completely understood. Therefore, are these uncharacterised genes potential candidates for future diagnostic assays and vaccine development.

Comparative genome analysis is a useful tool for various applications and usually start with an analysis of the basic characterisation and annotation of each genome, followed by the arrangement of genes in the genome and conservation of these genes. Arrangement of genes in bacterial species are influenced by numerous events, such as mutations, causing insertion, deletions (also known as indels) or duplications or horizontal gene transfer that can result in recombination events that can result in inversions, translocations or relocations (Darmon and Leach, 2014). Types of recombination events between species include events such as homologous and illegitimate recombination. Homologous recombination is the exchange of identical or nearly identical DNA between species, resulting in gene rearrangements, while illegitimate recombination is caused by DNA slippage during replication or annealing of single strands induced by a DNA double strand break (Darmon and Leach, 2014).

Repeat sequences, such as those associated with variable protein expression, can result in a single bacterial species with multiple phenotypes in the same environment (Rocha and Blanchard, 2002). This has made it possible for bacteria to adapt quickly and efficiently to changing environments (Rocha and Blanchard, 2002). These repeat sequences have resulted in numerous rearrangement, deletion and multiplication events (Rocha and Blanchard, 2002). The large number of repeats known to occur in mycoplasma genomes as well as the high mutation rates observed in mycoplasma genomes show the potential for large scale recombination events, such as homologous or illegitimate recombination events and have been shown for *M. pulmonis* as well as *M. pneumoniae* and *M. genitalium*, respectively (Rocha and Blanchard, 2002).

Genome analysis can aid in the *in silico* identification of possible candidate genes and comparative genome analysis can aid in decreasing the amount of possible candidate genes based on different criteria dependent on the use. Finding candidate genes for diagnostic purposes is in theory not too difficult, with the most important criteria being that the gene should be highly conserved and be able to distinguish the species of interest from other species or strains, dependent on the purpose of the assay. For vaccine development, finding candidate genes that elicit an immune response as well as confer protection can be difficult and costly, good vaccine candidates are usually secretion proteins from various secretion pathways or surface-localised proteins, such as those described above for MG (Curtiss, 2002).

In this chapter, draft genomes for axenic mycoplasma isolates with reference genomes available were assembled. All draft genome assemblies of a species were first compared to each and then to the other species using comparative genome analysis tools. Comparison of mycoplasma genomes is expected to aid in finding novel genes that can be used for developing improved diagnostic tools and aid in future vaccine development.

4.2. Materials and Methods

4.2.1. Sample collection, isolation and identification

Poultry mycoplasma samples were collected, isolated and identified as described in Chapter 2. Briefly, samples were collected by veterinarians from chickens using a swab and sent to the Bacteriology laboratory of the Department of Veterinary Tropical Diseases (DVTD) for identification of mycoplasma species by culture with growth inhibition. The DNA of the mycoplasma positive samples were isolated as described. Samples collected prior to 2014 were sent to Inqaba Biotech (Pty) Ltd, Pretoria for Illumina MiSeq whole genome sequencing and samples collected in 2014 and 2015 were sent for Ion Torrent PGM whole genome sequencing at UP. Samples were identified using the 16S rRNA gene as described in Chapter 2. The remainder of the sample was frozen at -20˚C for future downstream analysis.

4.2.2. Quality control

Quality control for each sample was performed as described in Chapter 3. Briefly the fastq sequencing file received for each sample was submitted to the FASTQC program (version 0.11.5), (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to produce a quality control report (Andrews, 2010). The sequencing files were imported into CLC Genomics Workbench version 8.5.1 (CLC Bio-Qiagen, Aarhus, Denmark). Low quality reads and sequencing adapters were trimmed and filtered using the default settings of the Trim Sequences function of CLC Genomics Workbench with the Nextera Trim Adapter Library. Trimmed files were analysed again with FASTQC for quality control.

4.2.3. Sequence assembly

Ion Torrent sequencing reads

Single-end reads produced by the Ion Torrent sequencing platform were *de novo* assembled in CLC Genomics Workbench (version 8.5.1) twice using the default settings, changing the minimum contig length to 1) 500 bp and 2) 200bp. Reads were also mapped back to the contigs with global alignment and saved for downstream analysis.

As described in Chapter 2, Ion Torrent reads were also subjected to digital normalization using Khmer (version 2.0) (Brown et al., 2012, Crusoe et al., 2015) to decrease the amount of reads for submission to the IonGAP server (Baez-Ortega et al., 2015) (available at http://iongap.hpc.iter.es/iongap) twice, the first time using the Genome assembly and Bacterial classification module and a second time using the Genome assembly and Comparative genomics modules using reference genomes dependent on 16S rRNA identification of each sample. Reference genomes used were *M. gallisepticum* strain R(low) (accession no. AE015450), *M. synoviae* strain 53 (accession no. AE017245), *M. pullorum* strain B359-6 (accession no. CP017813) and *M. gallinaceum* strain B2096 8B (accession no. CP011021).

Illumina sequencing reads

Paired-end reads produced by the Illumina MiSeq sequencing platform were also *de novo* assembled in CLC Genomics Workbench (version 8.5.1) twice using the default settings, as described above. The paired-end reads were used to detect paired distances and perform joined contiging and the reads were also mapped back to the contigs using the default settings with global alignment and saved for downstream analysis.

Illumina sequencing reads were also *de novo* assembled with the SPAdes assembler (version 3.12.0) (Bankevich et al., 2012, Nurk et al., 2013) using the default settings of the "- -careful" pipeline option with on a desktop computer running the Linux-4.4.0-31-generic-x86_64-with-Ubuntu-16.04-xenial operating system. After successful completion of the pipeline, the resulting joined contigs file was imported into CLC Genomics Workbench (version 8.5.1) for downstream analysis.

Draft genome assemblies

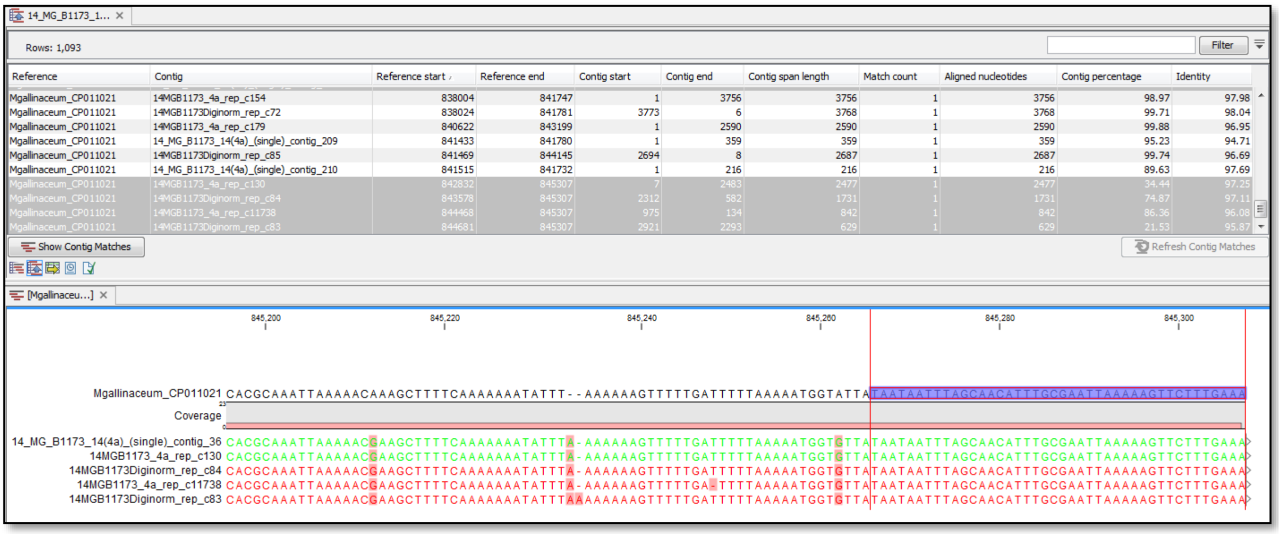
Draft assemblies for each sample was done *in silico* using the following reference-guided *de novo* strategy:

Contigs produced by CLC Genomics Workbench (version 8.5.1) were pooled with contigs produced by the IonGap pipeline (Ion Torrent PGM sequenced samples) and SPAdes (Illumina MiSeq sequenced samples) and then aligned to refence genome(s) using the "align contigs" tool of the CLC Genome Finishing Module (version 1.5.4). Contigs were systematically joined using the reference genome as a guide for direction and order of contigs. Overlapping contigs were joined based on the percentage of contig match to the reference, and contig match identity above 90%. Exceptions included 1) contig matches at the 3' and 5' ends of the reference genomes where contigs were joined dependent on a combined contig match percentage and identity above 90 at the 3' and 5' ends, 2) rearrangements in the genome indicated by a split in the contig match across genome and 3) lower contig match percentage and identity were used in variable regions of the reference genome, see Figure 4-1 for a representation of contigs aligned to reference genome. Where the distance between contigs was below 10 bp when compared to the reference genome, contigs were extracted and extended using the "extend contigs" tool of the CLC Genome finishing module and added back to the contig match table for contig joining. When no more contigs could be joined with confidence when compared to the reference genome, the order of the joined contigs were noted. Joined contigs were extracted and aligned with the "align contigs" tool using the contigs themselves as reference. The order of the contigs that was noted earlier was used as a guide for joining contigs that could not be joined when aligned to the reference genome, and the amount of joined contigs were decreased. Contigs that matched either at the 3' or 5' end of a contig, were in the correct order and had a contig match identity of at least 90% were joined. The resulting joined contigs were extracted and saved for downstream analysis.

Reference genomes used as described above including the additional reference genomes available for *M. gallisepticum* strain F (accession no. NC_017503); strain R(high) (accession no. NC_017502) and strain S6 (accession no. NC_023030) as well as for *M. synoviae* strain ATCC 25204 (accession no. NZ_CP011096).

Draft genome quality control

The quality of each draft genome was assessed and the basic sequencing statistics compared using Quast, a Quality Assessment Tool for Genome Assemblies from the Center for Algorithmic Biotechnology (available at http://quast.bioinf.spbau.ru/) (Gurevich et al., 2013). Briefly, all the fasta files containing joined contigs were uploaded to the online Quast server and compared to each other and the respective reference genome.



Reference	Contig	Reference start	Reference end	Contig start	Contig end	Contig span length	Match count	Aligned nucleotides	Contig percentage	Identity
Mgallinaceum_CP011021	14MGB1173_4a_rep_c154	838004	841747	1	3756	3756	1	3756	98.97	97.98
Mgallinaceum_CP011021	14MGB1173Diginorm_rep_c72	838024	841781	3773	6	3768	1	3768	99.71	98.04
Mgallinaceum_CP011021	14MGB1173_4a_rep_c179	840622	843199	1	2590	2590	1	2590	99.88	96.95
Mgallinaceum_CP011021	14_MG_B1173_14(4a)_(single)_contig_209	841433	841780	1	359	359	1	359	95.23	94.71
Mgallinaceum_CP011021	14MGB1173Diginorm_rep_c85	841469	844145	2694	8	2687	1	2687	99.74	96.69
Mgallinaceum_CP011021	14_MG_B1173_14(4a)_(single)_contig_210	841515	841732	1	216	216	1	216	89.63	97.69
Mgallinaceum_CP011021	14MGB1173_4a_rep_c130	842832	845307	7	2483	2477	1	2477	34.44	97.25
Mgallinaceum_CP011021	14MGB1173Diginorm_rep_c84	843578	845307	2312	582	1731	1	1731	74.87	97.11
Mgallinaceum_CP011021	14MGB1173_4a_rep_c11738	844468	845307	975	134	842	1	842	86.36	96.08
Mgallinaceum_CP011021	14MGB1173Diginorm_rep_c83	844681	845307	2921	2293	629	1	629	21.53	95.87

```
                 Mgallinaceum_CP011021 CACGCAAATTAAAAACAAAGCTTTTCAAAAAAATATTT--AAAAAAGTTTTTGATTTTTAAAAATGGTATTATAATAATTTAGCAACATTTGCGAATTAAAAAGTTCTTTGAAA
14_MG_B1173_14(4a)_(single)_contig_36 CACGCAAATTAAAAACGAAGCTTTTCAAAAAAATATTTA-AAAAAAGTTTTTGATTTTTAAAAATGGTGTTATAATAATTTAGCAACATTTGCGAATTAAAAAGTTCTTTGAAA
                 14MGB1173_4a_rep_c130 CACGCAAATTAAAAACGAAGCTTTTCAAAAAAATATTTA-AAAAAAGTTTTTGATTTTTAAAAATGGTGTTATAATAATTTAGCAACATTTGCGAATTAAAAAGTTCTTTGAAA
             14MGB1173Diginorm_rep_c84 CACGCAAATTAAAAACGAAGCTTTTCAAAAAAATATTTA-AAAAAAGTTTTTGATTTTTAAAAATGGTGTTATAATAATTTAGCAACATTTGCGAATTAAAAAGTTCTTTGAAA
               14MGB1173_4a_rep_c11738 CACGCAAATTAAAAACGAAGCTTTTCAAAAAAATATTTA-AAAAAAGTTTTTGA-TTTTAAAAATGGTGTTATAATAATTTAGCAACATTTGCGAATTAAAAAGTTCTTTGAAA
             14MGB1173Diginorm_rep_c83 CACGCAAATTAAAAACGAAGCTTTTCAAAAAAATATTTAAAAAAAAGTTTTTGATTTTTAAAAATGGTGTTATAATAATTTAGCAACATTTGCGAATTAAAAAGTTCTTTGAAA
```

Figure 4-1: Representation of aligning contigs to reference in using the "align contigs" tool of the CLC Genomics Workbench. The table in the top section lists the contig matches to the reference genome with metrics used to evaluate which contigs to join. The bottom section is a linear representation of how the contigs align to the reference. Green sequences indicate that the contig aligned to the reference in a forward direction and red sequence indicate that the contig aligned in the reverse direction.

Draft genome annotation

For joined contig order, the "move contigs" tool of the multiple genome alignment program Mauve (version 20150226 build 10) (Darling et al., 2004) was used to order the contigs to a reference genome. The resulting fasta file was submitted to the Rapid Annotation using Subsystem Technology (RAST) (version 2.0) (Aziz et al., 2008, Overbeek et al., 2013, Brettin et al., 2015) online server for genome annotation. The resulting Genbank®, protein FASTA, GFF3 feature and tab delimited features files were downloaded and saved for downstream applications. The GFF3 file was used to extract gene information from all strains using a script written by Jaco Beylefeld (available from https://github.com/jj-beylefeld/g-annotation-analyser) the gene content and functions were compared between all strains of the same species as well as the SEED-viewer classification of protein function (Overbeek et al., 2005).

4.2.4. Genome comparison

Intraspecies genome comparison

Annotated draft genomes for all strains produced in this study for the same species were aligned using the "align with progressive mauve" function of the MAUVE program to view rearrangements and the overall linear alignment (Darling et al., 2004).

Various tools of the CMG-biotools package (version 2.1) (available at http://www.cbs.dtu.dk/biotools/CMGtools/) (Vesth et al., 2013) were used to compare strains of each species, i.e. 1) “BLAST matrix” tool was used to produce a BLAST matrix representing a pairwise comparison of the proteins identified in each draft genome; 2) A “pancoreplot” analysis was used to produce pan-and core-genome plots representing a set of shared genes and a set of conserved gene families between the draft genomes of each species, respectively. The list of core genes was extracted and the resulting text file containing protein sequences was submitted to the BLASTp suite (version 2.8.1) (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi#) of NCBI for protein identification using the default settings (Altschul et al., 1997, Altschul et al., 2005). The protein file was also submitted to the online tool WebMGA (available at http://weizhong-lab.ucsd.edu/webMGA/server/) for functional annotation based on the COG families (Wu et al., 2011).

Interspecies genome comparison

The GFF3 file produced by the RAST server was used to extract gene information, i.e. amount of CDS, amount of CDS that are categorised, uncategorised and hypothetical, amount of tRNAs and amount of RNAs subdivided into the amount of 5S-rRNAs, 16S rRNAs, 23S rRNAs and ncRNAs as well as a more detailed information on the RAST annotation subsystem classification of the CDS in a tabular format from all strains using a script written by Jaco Beylefeld (available from https://github.com/jj-beylefeld/g-annotation-analyser). Using the extracted information, the gene content and functions were compared between all draft genomes.

Various tools of the CMG-biotools package (version 2.1) (available at http://www.cbs.dtu.dk/biotools/CMGtools/) (Vesth et al., 2013) was used to compare strains of each species, i.e. 1) “BLAST matrix” tool was used to produce a BLAST matrix representing a pairwise comparison of the proteins identified in each draft genome; 2) A “pancoreplot” analysis was used to produce pan-and core-genome plots representing a set of shared genes and a set of conserved gene families between the draft genomes of each species, respectively. The list of core genes, containing the protein sequences was extracted submitted to the BLASTp suite (version 2.8.1) of NCBI for protein identification using the default settings (Altschul et al., 1997, Altschul et al., 2005). The protein file was also submitted to the online tool WebMGA for COG protein function annotation (Wu et al., 2011).

4.3. Results

4.3.1. Whole genome sequencing and assembly

As discussed in Chapter 2, 80/124 samples were found to be axenic isolates. One isolate was identified as *Acholeplasma laidwalli* and excluded from the further analysis along with 11 *M. gallinarum* isolates for which a reference genome is not available. The remaining 68 mycoplasma isolates consisted of 15 MG, 11 MS, 8 *M. pullorum* and 34 *M. gallinaceum* strains. Taking the differences between the Ion Torrent and Illumina sequencing technologies and data outputs discussed in Chapter 3 into consideration, the quality control of all fastq sequencing files was performed using FASTQC. All Illumina samples were trimmed for both Nextera trim adapters and low quality sequences and the quality of the sequencing files improved. The sequence quality for all Ion Torrent sequencing files were in the expected range of 17 to 30 and no trimming was required. The number of files generated during this study is too large to include fully in this thesis and is available in multimedia format on request.

Depth of coverage for each strain was determined using Equation 1. Of the 15 MG strains, 9 were sequenced using Illumina sequencing technology, resulting a read length of 300bp and coverage ranging between 143 to 765 times, and the remaining 6 MG strains were sequenced using Ion Torrent sequencing technology resulting in an average read length pf 158 bp and coverage range of between 178 and 777 times (Table 4-1). All but one of the MS strains were sequenced using Ion Torrent sequencing technology resulting in an average read length of 162 bp and coverage ranged between 648 and 1505 times, the single MS strain sequenced using Illumina sequencing technology had 178 times coverage and read length of 300 bp (Table 4-1). All the *M. pullorum* strains were sequenced using Ion Torrent sequencing resulting in a coverage range of 133 to 1259 times and an average read length of 150 bp (Table 4-1). Only four *M. gallinaceum* strains were sequenced using Illumina sequencing and three produced reads of 300bp length with a coverage range of between 155 and 255 times, sample B2096-14-8 (synonymous with the reference strain *M. gallinaceum* B2096-8) produced reads of 133 bp with a coverage of 980 times. The remaining 30 *M. gallinaceum* were sequenced using Ion Torrent sequencing and produced reads with an average read length of 160 bp and coverage in the range of 227 and 1923 times (Table 4-1).

Equation 1: Depth of coverage

$$Depth\ of\ coverage\ (C) = \frac{Number\ of\ reads\ (N) \times\ Average\ read\ length\ (L)}{Length\ of\ genome\ (G)}$$

Draft genomes for all 68 samples were assembled in CLC Genomics Workbench and compared to their respective reference genomes using Quast (Table 4-2). As reference genomes are available, additional metrics were used to assess the quality of the assemblies and address some of the problems associated with the N50 metric, i.e. the contig length that represents 50% of the refence genome is represented by the NG50 metric and is a more standardized method of assessing different assemblies from the same species, the length of an aligned contig block that represent 50% of the assemble genome is represented the NA50 metric and the length of the aligned contig block that represents 50% of the reference genome is represented by the NGA50 metric. The latter two metrics are determined by aligning the contigs to the reference and where misassemblies are found the contigs are broken into smaller contigs changing the size of the contigs (Gurevich et al., 2013).

Table 4-1: Strain sequencing information

Strain	Year	Technology	Reads (N)	Avg read length (L)	Coverage (C)
MG	Genome size (G) = 1 012 800 bp				
B1102-03	2003	Illumina MiSeq	481546	300	143
B1102-06	2006	Illumina MiSeq	656392	300	194
B726-06	2006	Illumina MiSeq	728532	300	216
B852-06	2006	Illumina MiSeq	643158	300	191
B943-06	2006	Illumina MiSeq	558976	300	166
B1028-07	2007	Illumina MiSeq	609104	300	180
B642-08	2008	Illumina MiSeq	601924	300	178
B758-08	2008	Illumina MiSeq	707270	300	209
B2159-13	2013	Illumina MiSeq	623756	300	185
B1395-14-1	2014	Ion Torrent	4349673	178.02	765
B1552-14-19	2014	Ion Torrent	1190475	149.21	175
B2771-14-1A	2014	Ion Torrent	3143611	157.3	488
B2771-14-1B	2014	Ion Torrent	1460193	149.85	216
B878-14-L3	2014	Ion Torrent	2945203	157.03	457
B457-15-5	2015	Ion Torrent	4692281	157.05	728
MS	Genome size (G) = 799 477 bp				
B2214-07	2007	Illumina MiSeq	474834	300	178
B1064-14-H3	2014	Ion Torrent	6876966	142.99	1230
B1064-14-H4	2014	Ion Torrent	3041488	170.23	648
B1064-14-H5	2014	Ion Torrent	3804018	163.45	778
B1393-14-10	2014	Ion Torrent	3268902	174.81	715
B1394-14-2	2014	Ion Torrent	4808154	174.41	1049
B1394-14-5	2014	Ion Torrent	3509183	173.52	762
B458-15-1	2015	Ion Torrent	4574435	148.68	851
B458-15-11	2015	Ion Torrent	4361647	164.48	897
B458-15-5M	2015	Ion Torrent	5472464	153.72	1052
B458-15-6	2015	Ion Torrent	7477157	160.95	1505
M. pullorum	Genome size (G) = 1 007 172 bp				
B359-15-6	2015	Ion Torrent	898956	149.1	133
B2096-14-3	2014	Ion Torrent	5959103	150.71	892
B293-15-12	2015	Ion Torrent	5347189	146.66	779
B293-15-13	2015	Ion Torrent	6204621	143.04	881
B293-15-15	2015	Ion Torrent	5473616	152.88	831
B293-15-17	2015	Ion Torrent	5759868	140.93	806
B359-15-5	2015	Ion Torrent	6565843	155.76	1015
B540-15-2	2015	Ion Torrent	7684024	164.97	1259
M. gallinaceum	Genome size (G) = 845 307 bp				
B2096-14-8	2014	Illumina	6229108	133.03	980
B313-05	2005	Illumina	523440	300	186
B733-05	2005	Illumina	717620	300	255
B2176-13	2013	Illumina	435492	300	155
B1101-14-7	2014	Ion Torrent	1886415	172.18	384
B1173-14-2a	2014	Ion Torrent	4652152	149.46	823
B1173-14-2b	2014	Ion Torrent	1621017	154.09	295

B1173-14-4a	2014	Ion Torrent	9393060	173.09	1923
B1173-14-4b	2014	Ion Torrent	3774934	168.68	753
B1173-14-5b	2014	Ion Torrent	3336615	170.36	672
B1173-14-6b	2014	Ion Torrent	5404736	158.78	1015
B1173-14-7b	2014	Ion Torrent	3734540	172.18	761
B1173-14-8b	2014	Ion Torrent	3181126	154.63	582
B1342-14-10	2014	Ion Torrent	3718455	169.46	745
B1342-14-13	2014	Ion Torrent	1586740	150.76	283
B1342-14-14	2014	Ion Torrent	1557093	152.53	281
B1342-14-8	2014	Ion Torrent	1398030	156.37	259
B1395-14-2	2014	Ion Torrent	3580126	174.22	738
B1396-14-7	2014	Ion Torrent	1239974	154.51	227
B1396-14-8	2014	Ion Torrent	7038053	146.34	1218
B1396-14-9	2014	Ion Torrent	4054902	128.96	619
B1414-14-1	2014	Ion Torrent	3501393	170.81	708
B2096-14-2	2014	Ion Torrent	5662880	168.95	1132
B2096-14-4	2014	Ion Torrent	1204072	160.63	229
B2096-14-7	2014	Ion Torrent	4002756	154.87	733
B878-14-M1	2014	Ion Torrent	2499782	155.88	461
B878-14-M4	2014	Ion Torrent	2013452	165.31	394
B878-14-M5	2014	Ion Torrent	3942490	170.99	797
B293-15-16	2015	Ion Torrent	6620990	161.75	1267
B3381-15-1	2015	Ion Torrent	6678557	163.68	1293
B3381-15-2	2015	Ion Torrent	6477326	156.61	1200
B3381-15-3	2015	Ion Torrent	5975741	152.41	1077
B3381-15-4	2015	Ion Torrent	6165309	163.99	1196
B3381-15-5	2015	Ion Torrent	6418220	164.65	1250

M. gallisepticum strain R(low) (accession number AE015450) was used as a reference for comparing the MG strains and is 1 012 800 bp in length with a GC content of 31.47%. Draft assemblies of MG strains ranged in total length from 815 882 bp for sample B1395-14-1 to 1 219 214 bp for sample B2771-14-1B, and an average GC content of 34.46% was observed between all MG strains. Samples B1395-14-1 and B1552-14-19 only covered 73.4 and 79.8% of the reference genome, respectively. The other 13 MG strains covered 87.5 to 91.5% of the reference genome. The N50 metric for all MG strains was equal to the NG50, except for strain B1395-14-1 and strain B2771-14-1B for which the N50 was higher and lower, respectively compared to the NG50. All but one of the observed misassemblies between MG draft genomes and the reference genome were classified by the Quast as relocations and one misassembly between MG strain B1028-07 was classified as an inversion (Figure 4-2). Relocations in QUAST are defined as a a breakpoint in a sequence where one part of a sequence is more than 1kbp downstream from other part in comparison to their location next to each other in the reference genome and inversion are defined as a breakpoint in the sequence align on the opposite strand of the reference genome. Taking the misassemblies into account the NA50 and NGA50 values were equal for 9/15 of the MG strains. Four strains, B1102-03, B726-06, B1395-14-1 and B1552-14-19, had higher NA50 values and two strains, B2771-14-1B and B457-15-5, had lower NA50 values than their respective NGA50 values (Table 4-2).

M. synoviae strain 53 (accession no. AE017245), was used as a reference for the MS strains with a genome size of 799 476 bp and GC content of 28.5%. The largest MS draft genome was strain B 1393-14-10 with 781 849 bp and the smallest draft MS genome was strain B458-15-11 with 624 730 bp. The average GC content across all MS draft genomes was 28.1% and all but one covered in the range of 88.4 to 90.3% of the reference genome, whereas MS strain B458-15-11 only covered 75.3% of the reference genome. All N50 values of each MS strain were equal to their respective NG50 values. All misassemblies observed were classified as relocations (Figure 4-2). For 6/11 MS strains the NA50 was equal to the NGA50 and for the remaining five strains, B1064-14-H3, B1394-14-5, B458-15-1, B458-15-11 and B458-15-5M the NA50 metric was higher than the NGA50 metric (Table 4-2).

M. pullorum strain B359-6 (accession no. CP017813) was used as the reference genome for *M. pullorum* strains with a size of 1 007 172 bp and GC content of 29.1 %. The largest draft M. pullorum strain was strain B2096-14-3 with a size of 1 054 672 bp and the smallest draft genome was strain B540-15-2 with a size of 677 833 bp. The average GC content across *M. pullorum* strains analysed was 29.1% and all strains covered in the range of 91.6 to 100% of the reference genome, except for sample B540-15-2 that only aligned to 63.5% of the reference genome. All misassemblies were classified as relocations except for two from strain B293-15-15 that was classified as inversions (Figure 4-2). For all but one of the *M. pullorum* strains the N50 was equal to the NG50 metric and the NA50 was equal to the NGA50, strain B540-15-2 had a higher N50 and NA50 than NG50 and NGA50 respectively (Table 4-2)

M. gallinaceum strain B2096 8B (accession no. CP011021) was used as the reference genome with a size of 845,307 bp and GC content of 28.38%. Strain B1173-14-8b and B733-03 were the largest and smallest draft genome assemblies and were 1 028 596 bp and 816 164 bp in length, respectively. The average GC content was 28.4% and strains covered between 76.1 and 100% of the reference genome. All but one of the missassemblies were classified as relocations and one misassembly from strain B1101-14-7 was classified as an inversion (Figure 4-2). The N50 metric was equal to NG50 metric for 23 *M. gallinaceum* strains, N50 was smaller than NG50 for 10 strains and one strain, B733-05 had a higher N50 than NG50. The NA50 metric was equal to the NGA50 metric for only 8 strain, for 23 strains the NA50 was smaller than the NGA50 and for strain B733-05 the NA50 was higher than the NGA50 metric (Table 4-2).

All of the above mentioned are draft genome assemblies that were obtained from a reference-guided *de novo* assembly, and the files are available upon request. The completion and deposition in the NCBI database of full genomes is a future objective, and will possibly involve long-read sequencing to confirm the sequence in the complex genome regions.

4.3.2. Genome annotation

All the draft genomes were submitted to the online RAST server for genome annotation. The reference genomes downloaded from NCBI Genbank® was also loaded into the RAST annotation pipeline for consistency across gene and protein names and used for comparison of the basic genome annotations and downstream annotation. The script produced by Jaco Beylefeld was used to extract basic genome annotation information from the GGF3 files generated by the RAST server (Table 4-3). The MG reference genome consisted of 836 coding sequences (CDS), with 32 tRNAs and two copies each of the 5S, 16S and 23S rRNA genes. In the MG strains analysed an average of 811 CDS were identified with MG strains B1395-14-1 and B2771-14-1B having the lowest (701) and highest (1016) CDS count observed, respectively. On average 387 of the CDS were assigned a gene name expressing proteins with a known function in the SEED-based subsystem categorisation used by the RAST server. The remaining CDS were either uncategorised or annotated as hypothetical proteins (Overbeek et al., 2005). Additionally, an average of 34 tRNAs were identified in the MG strains, strain B457-15-5 had the highest count with 47 tRNAs and strain B1552-14-19 had the lowest with 31 tRNAs. Two copies each of the 5S, 16S and 23S rRNA genes were identified in 11/15 MG strains, two copies of the 5S and 23S rRNA and only one copy of the 16S rRNA gene were identified in strain B852-06, only one copy of each gene was identified in two strains, B1395-14-1 and B1552-14-19. Four copies each of the 5S and 23S genes and 3 copies of the 16S rRNA gene were identified in strain B 2771-14-1B.

The annotation of the MS reference genome by the RAST server included 727 CDS of which 314 were categorised, 34 tRNAs, three 5S rRNA gene copies and two copies each of the 16S and 23S rRNA genes. Annotation of the MS strains identified an average count of 720 CDS per draft genome with 332 CDS assigned to protein categories. Strain B458-15-11 had the least amount of CDS with 632 (302 categorised) and strain B458-15-1 the most with 819 CDS (386 categorised). An average of 33 tRNAs were identified, strain B458-15-11 and strain B458-15-6 had the lowest and highest count with 26 and 40 tRNAs, respectively. All MS strains analysed in this study had three 5S rRNA gene copies and two copies each of the 16S and 23S rRNA genes.

Table 4-2: Summary of comparative analysis of genome assembly to reference genomes as determined by Quast. Colour scale indicates best (blue) and worst (red) values of each measure in each species[‡].

Strain	# contigs	Total length	GC (%)	N50*	NG50*	Genome fraction (%)	Aligned length	# mismatches	# indels	NA50■	NGA50■
MG											
B1102-03	16	989903	31.43	123348	123348	91.456	949605	12614	1037	97414	89768
B1102-06	7	957211	31.49	633149	633149	90.215	917978	11741	968	91435	91435
B726-06	10	986867	31.56	577654	577654	90.848	931691	12111	726	87160	85521
B852-06	15	989573	31.43	182568	182568	90.022	962848	12780	1090	80757	80757
B943-06	15	937397	31.5	283719	283719	89.122	907836	11388	699	79207	79207
B1028-07	11	935226	31.43	214181	214181	89.561	912834	11321	695	89768	89768
B642-08	14	967457	31.52	171815	171815	91.293	934075	12649	1129	90477	90477
B758-08	9	962838	31.5	199100	199100	90.935	926048	12714	815	88426	88426
B2159-13	14	1030086	31.57	211555	211555	90.765	970527	13116	1037	86273	86273
B1395-14-1	17	815882	31.41	219978	100793	73.369	793276	10551	653	93439	92001
B1552-14-19	10	845452	31.4	202948	202948	79.848	834874	10684	662	87111	76434
B2771-14-1A	12	985763	31.46	203672	203672	91.034	955206	12730	873	90076	90076
B2771-14-1B	23	1219214	31.33	211835	289030	91.394	1182196	15878	1031	85342	89729
B878-14-L3	6	894230	31.43	224551	224551	87.515	887622	10866	722	88769	88769
B457-15-5	16	1065674	31.49	223928	223928	90.272	1033706	13344	914	76412	90023
MS											
B2214-07	12	747877	28.03	470038	470038	90.126	721118	5349	394	145225	145225
B1064-14-H3	12	755632	28.14	202957	202957	90.25	727410	5917	621	185999	130720
B1064-14-H4	10	737788	28.02	159097	159097	88.369	714768	5309	418	130780	130780
B1064-14-H5	4	747869	28.02	528503	528503	89.393	718564	5532	558	132067	132067
B1393-14-10	8	781849	28.07	469458	469458	89.604	744697	5808	513	145251	145251
B1394-14-2	6	735711	28.05	714131	714131	88.8	714955	5334	410	145227	145227
B1394-14-5	9	744284	28.01	194151	194151	89.212	713250	5429	485	136812	131671
B458-15-1	11	745513	28.09	160137	160137	89.398	716575	5495	728	97383	90903
B458-15-11	9	624730	28.32	187670	187670	75.259	601507	4501	433	95906	87196
B458-15-5M	7	738033	28.05	203867	203867	88.792	711648	5313	670	133702	132003
B458-15-6	6	728526	28.07	504129	504129	88.357	707711	5220	753	132008	132008

M. pullorum											
B359-15-6	1	1007172	29.07	1007172	1007172	100	1007172	0	0	1007172	1007172
B2096-14-3	1	1054672	28.98	1054672	1054672	100	1054672	522	227	1007073	1007073
B293-15-12	3	1012892	29.07	689328	689328	99.755	1005987	519	206	689280	689280
B293-15-13	2	1001534	29	986572	986572	95.68	1001381	650	239	963559	963559
B293-15-15	3	1026436	29.04	836823	836823	95.527	1013047	866	204	787080	787080
B293-15-17	1	935521	29	935521	935521	91.584	930147	605	220	648094	648094
B359-15-5	1	972098	29.02	972098	972098	96.038	967299	560	199	967299	967299
B540-15-2	8	677833	29.25	356031	102078	63.473	639035	11141	522	349754	37822
M. gallinaceum											
B2096-14-8	1	845307	28.38	845307	845307	100	845307	0	0	845307	845307
B313-05	11	966687	28.36	150847	170455	88.246	746818	2630.88	63.54	35872	40078
B733-05	19	816164	28.6	87526	81355	88.349	746987	2508.38	56.51	29201	28424
B2176-13	10	971900	28.25	140768	140768	77.84	683005	2650.82	61.86	18918	26115
B1101-14-7	18	870215	28.25	78490	78490	91.315	781297	2760.48	72.16	36116	36856
B1173-14-2a	12	976615	28.44	96567	477841	92.143	802086	2839.55	97.06	46625	79642
B1173-14-2b	8	968231	28.16	311795	311795	91.781	778791	2747.63	71.28	35887	46652
B1173-14-4a	4	907869	28.45	349349	349349	92.557	785218	2749.01	79.88	46652	46652
B1173-14-4b	11	968009	28.41	138234	138234	92.204	825852	2875.78	73	35880	46645
B1173-14-5b	5	1001801	28.41	265011	449789	92.573	837222	2899.71	70.54	46644	46646
B1173-14-6b	15	957829	28.37	110666	110666	91.726	802473	2850.79	77.13	32898	35879
B1173-14-7b	1	1000532	28.32	1000532	1000532	76.124	790984	3422.61	86.09	46641	89519
B1173-14-8b	5	1028596	28.35	636942	636942	92.513	813082	2831.25	77.62	35897	80481
B1342-14-10	7	969278	28.25	545511	545511	92.405	785995	2767.74	69.13	35882	72993
B1342-14-13	12	955097	28.39	132788	132788	92.882	796082	2803.34	73.75	35569	46493
B1342-14-14	8	959000	28.38	194253	465345	90.339	769278	2749.31	65.87	46648	80492
B1342-14-8	18	914526	28.32	171857	171857	86.916	735830	2786	73.23	35696	35883
B1395-14-2	9	948426	28.41	239155	239155	92.306	780324	2738.8	68.57	46643	65681
B1396-14-7	15	958672	28.54	116837	116837	92.449	818686	2833.62	70.76	46642	64810
B1396-14-8	10	921208	28.31	236279	236279	89.395	786122	2864.39	82.05	46651	65251
B1396-14-9	11	999745	28.43	156626	178196	92.394	811592	2853.74	75.29	35569	46646

B1414-14-1	11	889682	28.34	115057	128959	92.149	780013	2415.32	65.47	35948	36154
B2096-14-2	1	854538	28.4	854538	854538	99.585	844878	23.52	21.03	292678	292678
B2096-14-4	1	864492	28.39	864492	864492	99.995	856317	24.02	9.46	615926	615926
B2096-14-7	1	977232	28.59	977232	977232	100	878432	33.6	4.14	845310	845310
B878-14-M1	2	875056	28.36	583767	583767	80.423	685960	2739.85	71.05	32877	32877
B878-14-M4	3	904065	28.39	639897	639897	79.427	678223	2712.67	68.07	34918	35898
B878-14-M5	4	892443	28.59	303535	303535	77.663	659884	2695.09	64.89	28507	32874
B293-15-16	22	888154	28.41	103381	112524	92.799	796100	2699	78.53	35496	35887
B3381-15-1	7	961517	28.47	210339	225311	89.194	810938	2673.75	124.28	35870	47576
B3381-15-2	8	884790	28.26	217378	217378	88.752	776213	2569.48	102.9	34688	34688
B3381-15-3	13	854311	28.4	144728	144728	89.207	756734	2519.53	122.14	31000	31000
B3381-15-4	8	976338	28.46	286711	286711	88.659	772545	2518.01	109.28	33413	34682
B3381-15-5	9	933382	28.46	162745	208980	88.517	762090	2500.27	89.28	34826	35880

‡ - Color range is a heatmap scale comparing the results of each genome subjectively from the weakest value (Dark red) to the best value (Dark blue) in the dataset compared to the reference genome.

* - N50 – contig length equal to and higher than value representing 50% of assembled genome.
- NG50 – contig length equal to and higher than value representing 50% of reference genome

■ - NA50 – length of aligned blocks equal to and higher than value representing 50% of assembled genome
- NGA50 – length of aligned blocks equal to and higher than value representing 50% of reference genome

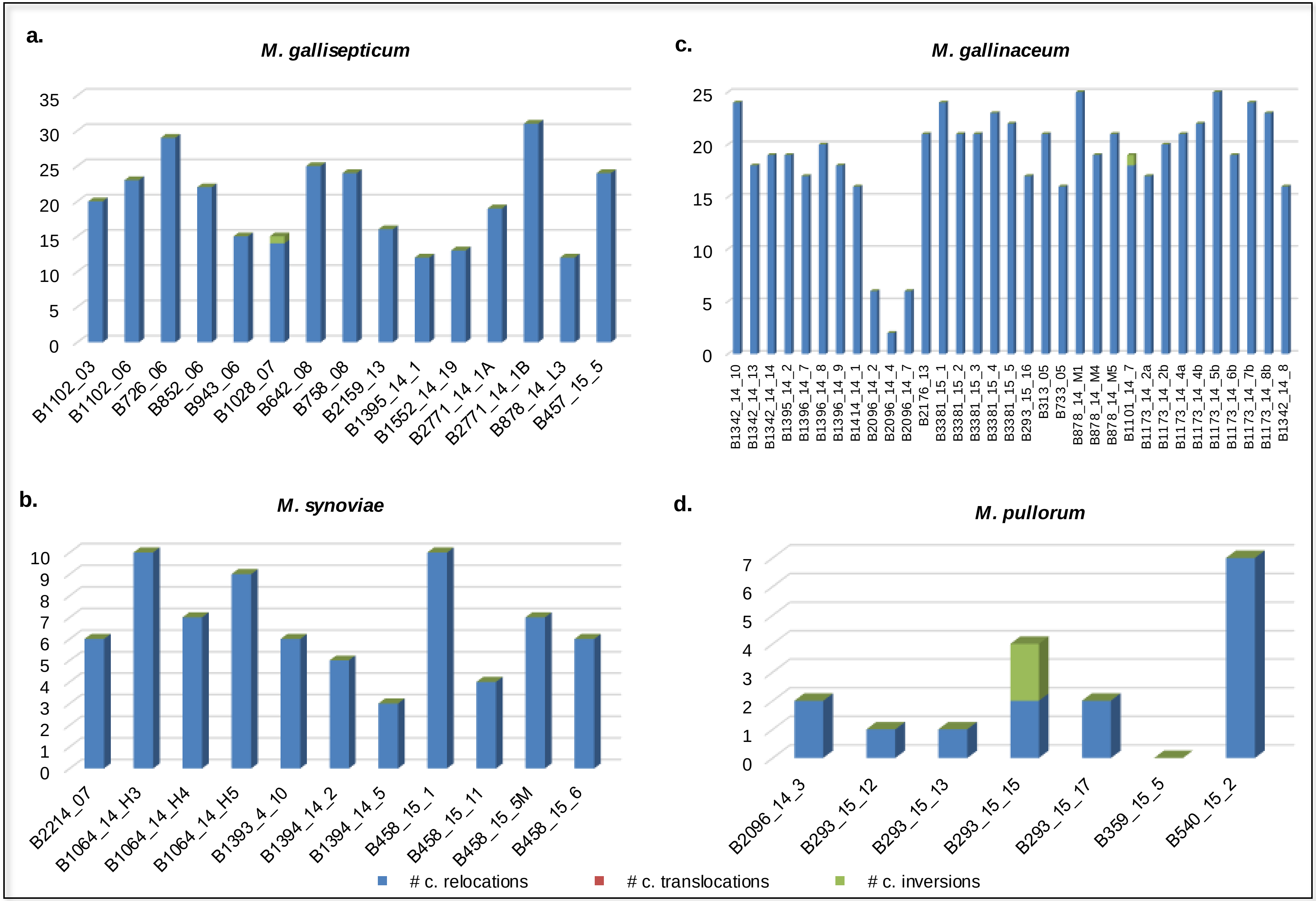


Figure 4-2: Amount of missamblies compared to reference genome as measured by Quast for each strain analysed for a) *Mycoplasma gallisepticum* b) *M. synoviae* c) *M. gallinaceum* and d) *M. pullorum*.

The annotated reference genome of *M. pullorum* strain B359-5 (synonymous with strain B359-15-5) contained 800 CDS of which 319 were assigned to RAST protein categories, 33 tRNAs, one copy of the 5S rRNA and two copies each of the 16S and 23S rRNAs. RAST identified an average of 805 CDS of which 327 CDS are assigned to protein categories, strain B540-15-2 had the least amount of CDS at 633 (256 categorised) and strain B2096-14-3 the most with 875 CDS (361 categorised). An average of 32 tRNAs were identified among the *M. pullorum* strains, strain B540-15-2 only had 26 and four *M. pullorum* strains had 34 tRNAs. Five *M. pullorum* strains had one copy each of the 5S rRNA and two copies each of the 16S and 23S rRNAs, strain B293-15-17 had two copies of each of these genes and strain B540-15-2 had 2 copies each of the 16S and 23S rRNA, but no 5S rRNA genes were identified.

The reference strain B2096-8 contained 621 CDS of which 272 CDS have been assigned to RAST protein categories. Only 17 tRNAs, three copies of the 5S rRNA and one copy each of the 16S and 23S rRNA were also annotated. Between the *M. gallinaceum* strains assembled an average of 750 CDS were annotated with an average of 307 CDS assigned to categories. Sample B3381-15-4 and B733-05 had the most with 948 CDS (384 categorised) and least with 602 (270 categorised) CDS, amount of annotated CDS, respectively. An average of 21tRNAs were annotated, three strains, B878-14-M1, B878-14-M4 and B3381-15-5 had the most with 30tRNAs and four strains, B733-05, B1414-14-1, B2096-14-2 and B2096-14-4 had the least with 17 tRNAs which is also the same amount as the reference strains.

Proteins assigned were assigned to the following SEED categories: 1) amino acids and derivatives, 2) carbohydrates, 3) cell division and cell cycle 4) cell wall and capsule 5) clustering-based categories 6) cofactors, vitamins, prosthetic groups, pigments 6) DNA metabolism 7) fatty acids, lipids, and isoprenoids 8) iron acquisition and metabolism 9) membrane transport and miscellaneous 10) motility and chemotaxis 11) nucleosides and nucleotides 12) phages, prophages, transposable elements, plasmids 13) phosphorus metabolism 14) potassium metabolism 15) protein metabolism 16) regulation and cell signalling 17) respiration 18) RNA metabolism 19) stress response 20) sulphur metabolism 21) virulence, disease and defence and 22) uncategorised.

The amount of classified CDS divided into SEED categories as well as percentage of each category are shown graphically in Figure 4-3 and Figure 4-4, respectively. An average of 235/497 (47%) CDS were characterised as uncategorised proteins, so to get a clearer picture of the other categories the uncategorised CDS were left out of the amount of categorised CDS and percentage of CDS in each categories shown in Figure 4-5 and Figure 4-6, respectively. Categorisation of CDS were similar across all species, with most CDS classified in the protein metabolism, carbohydrates and DNA metabolism categories. Some interesting observation that was made include the annotation of on average one CDS in the iron acquisition and metabolism across the MG strains not found in any of the other strains, also more MG CDS (2%) were assigned to phosphorous metabolism compared to the other species (0.35%). More of the MS CDS (8%) were assigned to the cell wall and capsule category compared to the average assignment of 3% across the other three species.

The SEED subsystem classification divides the categories into subcategories and subsystems. Taking a closer look at the categories mentioned above, the protein metabolism category includes genes involved in protein degradation, biosynthesis, folding processing and modification; the carbohydrate category includes genes involved in among others carbohydrate metabolism, CO_2 fixation, and fermentation and the DNA metabolism category involve genes that play a role in DNA recombination, repair and replication. The Iron acquisition and metabolism category is linked to the Campylobacter iron metabolism subsystem, and the phosphorous metabolism category include genes that play a role in phosphate metabolism and transport. The cell wall and capsule category involves genes that are involved in capsular and extracellular polysaccharides.

4.3.3. Draft genome assembly comparison

The draft genomes of each species were aligned with the “progressiveMauve” algorithm to their respective reference. Coloured blocks represent segments of conserved DNA, i.e. homologous, that are free of internal genome arrangements also referred to as Locally Collinear Blocks (LCBs). The MG and *M. gallinaceum* strains produced large images and is available in multimedia format. MG strains showed some differences in the order of the LCB coloured blocks (Multimedia). Mauve is an interactive program and selecting an LCB region aligns all the same regions respective to the chosen region, from this it was observed that the arrangements of the LCB regions for strains B1102-06, B726-06, B1028-07, B758-08, B2271-14-1A, B878-14-L3 and B457-15-5 are similar to strain R(low), strains B1102-03, B852-06 and B1552-14-19 had similar profiles to each other and strains B943-06, B642-08 and B2159-13 had similar profiles. Of the samples for which information was available, none were isolated in the same batch or from the same locations (Chapter 2)

MS strains were divided into 4 groups with similar arrangements, and only one or two rearrangements within each group (Figure 4-7). Strains B1064-14-H5, B1394-14-2 and B458-15-6 were similar to the reference. Strains B2214-07, B1393-14-10, B1394-14-5, B458-15-1 and B458-15-5M had similar profiles, and strains B1064-14-H3 and B1064-14-H4 were different from each and from the rest of the strains. Strains B1064-14-H3, B1064-14-H4 and B1064-14-H5 were isolated from the same farm in the same batch, the same is true for strains B1394-14-2 and B1394-14-5 as well as for strains B458-15-1, B458-15-5M and B458-15-6 (Chapter 2).

Table 4-3: Annotation results for draft mycoplasma genomes. Colours rank values in column from highest to lowest for each species

Strain	CDS	CDS Known (Categorised)	CDS Known (Uncategorised)	CDS Hypothetical	tRNA	5S rRNA	16S rRNA	23S rRNA
MG								
R(low)	836	388	230	375	32	2	2	2
B1102-03	803	374	225	354	33	2	2	2
B1102-06	777	372	214	341	32	2	2	2
B726-06	808	370	214	373	32	2	2	2
B852-06	807	375	221	358	34	2	1	2
B943-06	776	381	215	334	32	2	2	2
B1028-07	765	373	208	334	32	2	2	2
B642-08	793	365	218	355	32	2	2	2
B758-08	768	369	219	329	32	2	2	2
B2159-13	831	383	229	376	33	2	2	2
B1395-14-1	701	328	193	316	33	1	1	1
B1552-14-19	754	386	221	307	31	1	1	1
B2771-14-1A	897	439	257	381	32	2	2	2
B2771-14-1B	1016	479	272	457	39	4	3	4
B878-14-L3	780	395	222	322	32	2	2	2
B457-15-5	897	410	268	382	47	2	2	2
MS								
53	727	314	275	254	34	3	2	2
B2214-07	685	314	286	200	33	3	2	2
B1064-14-H3	727	325	302	219	33	3	2	2
B1064-14-H4	695	326	291	200	33	3	2	2
B1064-14-H5	734	333	301	220	33	3	2	2
B1393-14-10	717	319	294	220	33	3	2	2
B1394-14-2	689	323	280	204	33	3	2	2
B1394-14-5	701	325	289	209	33	3	2	2
B458-15-1	819	386	370	213	33	3	2	2
B458-15-11	632	302	262	175	26	3	2	2
B458-15-5M	756	342	331	207	33	3	2	2
B458-15-6	767	359	317	225	40	3	2	2
M. pullorum								
B359-15-6	804	323	250	358	34	1	2	2
B2096-14-3	875	361	264	401	34	1	2	2
B293-15-12	836	350	260	365	34	1	2	2
B293-15-13	846	338	255	386	33	1	2	2
B293-15-15	859	349	267	383	34	1	2	2
B293-15-17	783	312	232	362	32	2	2	2
B359-15-5	800	319	246	360	33	1	2	2
B540-15-2	633	256	184	289	26	0	2	2
M. gallinaceum								
B2096-14-8	621	272	183	277	17	3	1	1
B313-05	722	300	224	315	21	3	1	1
B733-05	602	270	182	257	17	3	1	1

B2176-13	767	281	220	369	22	3	2	1
B1101-14-7	663	275	188	308	20	3	2	1
B1173-14-2a	862	340	243	407	21	3	2	1
B1173-14-2b	778	300	215	382	20	3	1	1
B1173-14-4a	728	302	200	346	20	3	1	1
B1173-14-4b	754	316	210	353	27	3	1	1
B1173-14-5b	783	322	225	361	21	4	2	1
B1173-14-6b	766	308	201	378	27	4	1	1
B1173-14-7b	740	294	207	359	18	1	1	1
B1173-14-8b	822	340	231	387	21	3	2	1
B1342-14-10	760	300	213	365	20	3	2	1
B1342-14-13	744	307	213	345	19	2	1	1
B1342-14-14	761	311	211	360	21	3	1	1
B1342-14-8	730	291	197	357	20	1	2	1
B1395-14-2	745	310	209	345	21	3	1	1
B1396-14-7	748	313	208	348	20	3	2	2
B1396-14-8	782	309	206	388	19	3	1	1
B1396-14-9	791	327	230	359	22	3	1	1
B1414-14-1	669	266	200	316	17	2	1	1
B2096-14-2	686	289	198	316	17	3	1	1
B2096-14-4	656	281	193	295	17	3	1	1
B2096-14-7	692	305	231	278	18	3	1	1
B878-14-M1	687	271	188	333	20	2	2	1
B878-14-M4	709	279	220	312	30	3	1	1
B878-14-M5	686	296	211	292	30	2	1	1
B293-15-16	722	284	201	346	18	3	2	1
B3381-15-1	914	366	258	429	26	5	1	2
B3381-15-2	797	320	215	390	19	3	1	1
B3381-15-3	840	350	226	395	20	2	2	1
B3381-15-4	948	384	282	426	30	3	1	1
B3381-15-5	809	353	215	376	28	3	1	1

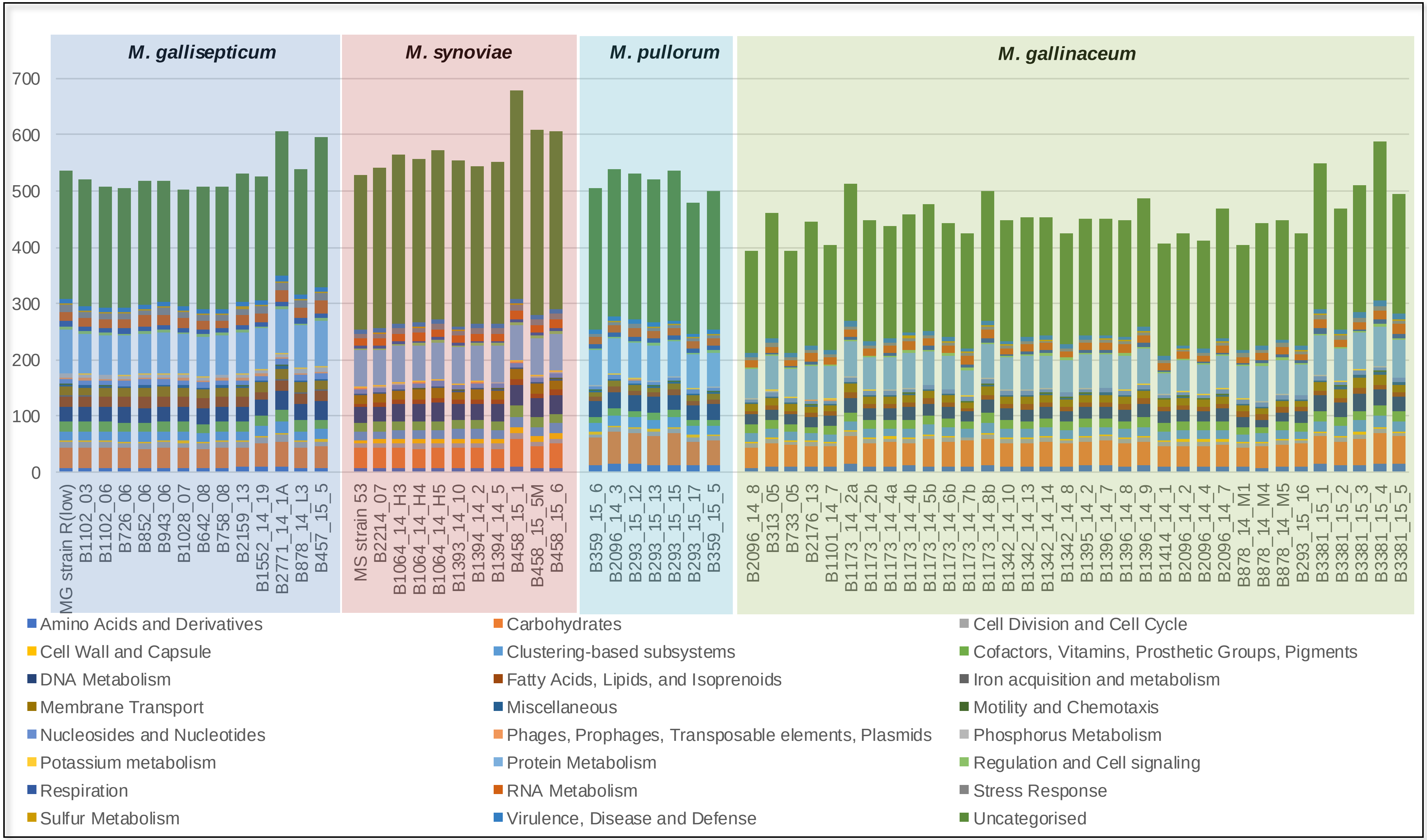


Figure 4-3: Comparison of RAST SEED category classification for the annotated CDS produced by the RAST annotation pipeline for each strain in each species.

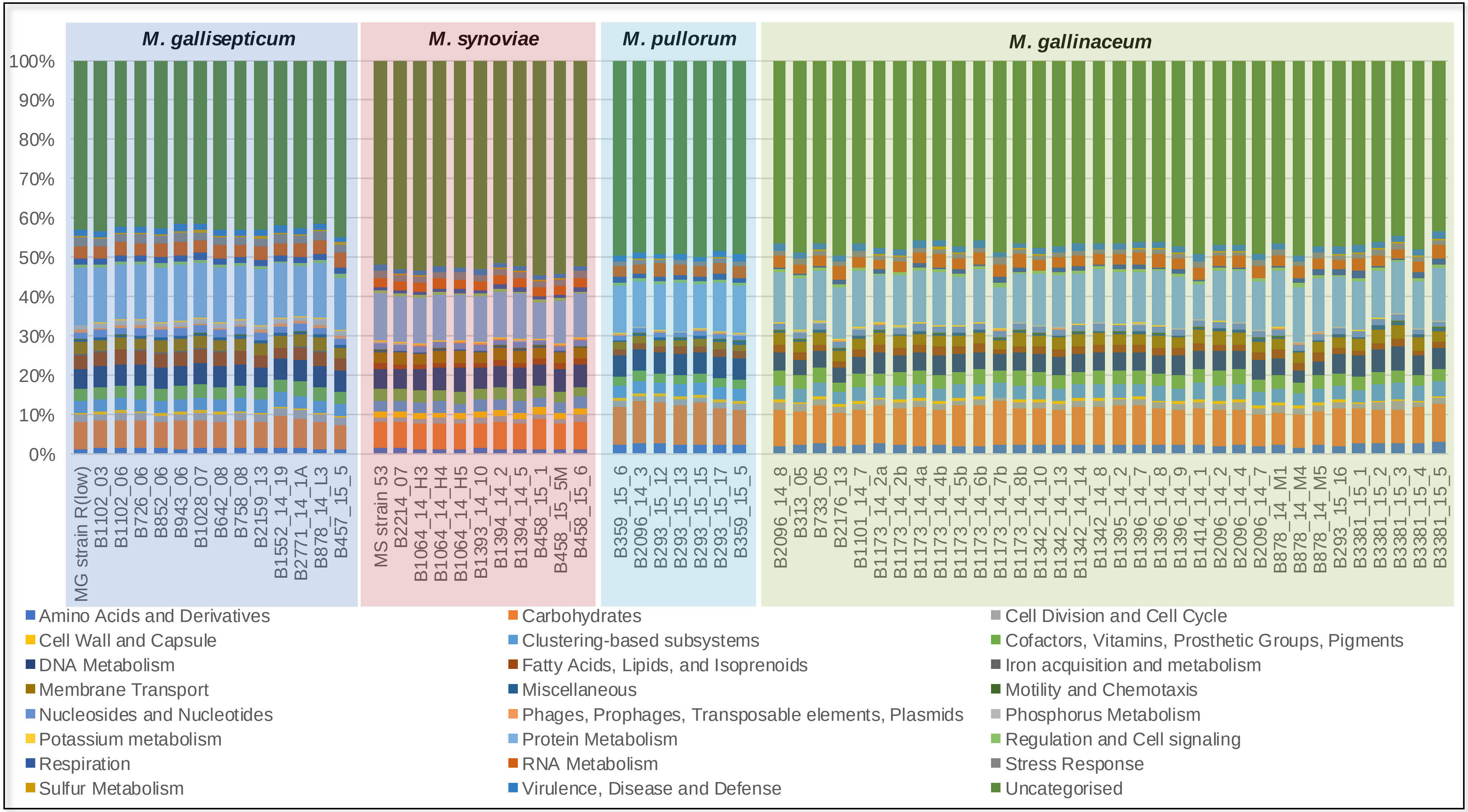


Figure 4-4: Comparison of RAST seed category classification for annotated CDS of strains as percentage of the total amount of classified CDS per strain.

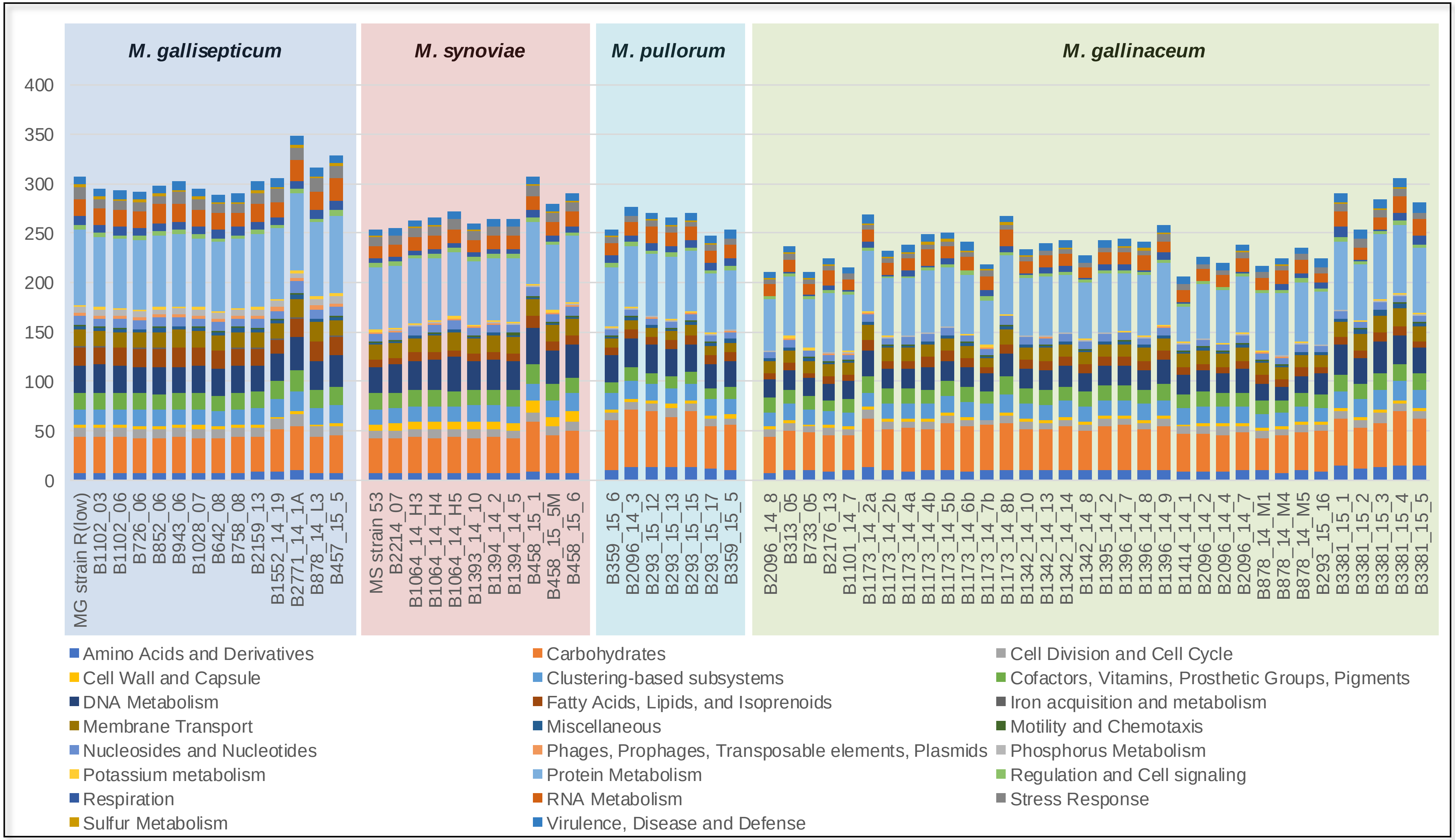


Figure 4-5: Comparison of RAST seed category classification for annotated CDS of strains excluding the proteins classified as hypothetical for a better view of the protein categories with assigned protein function.

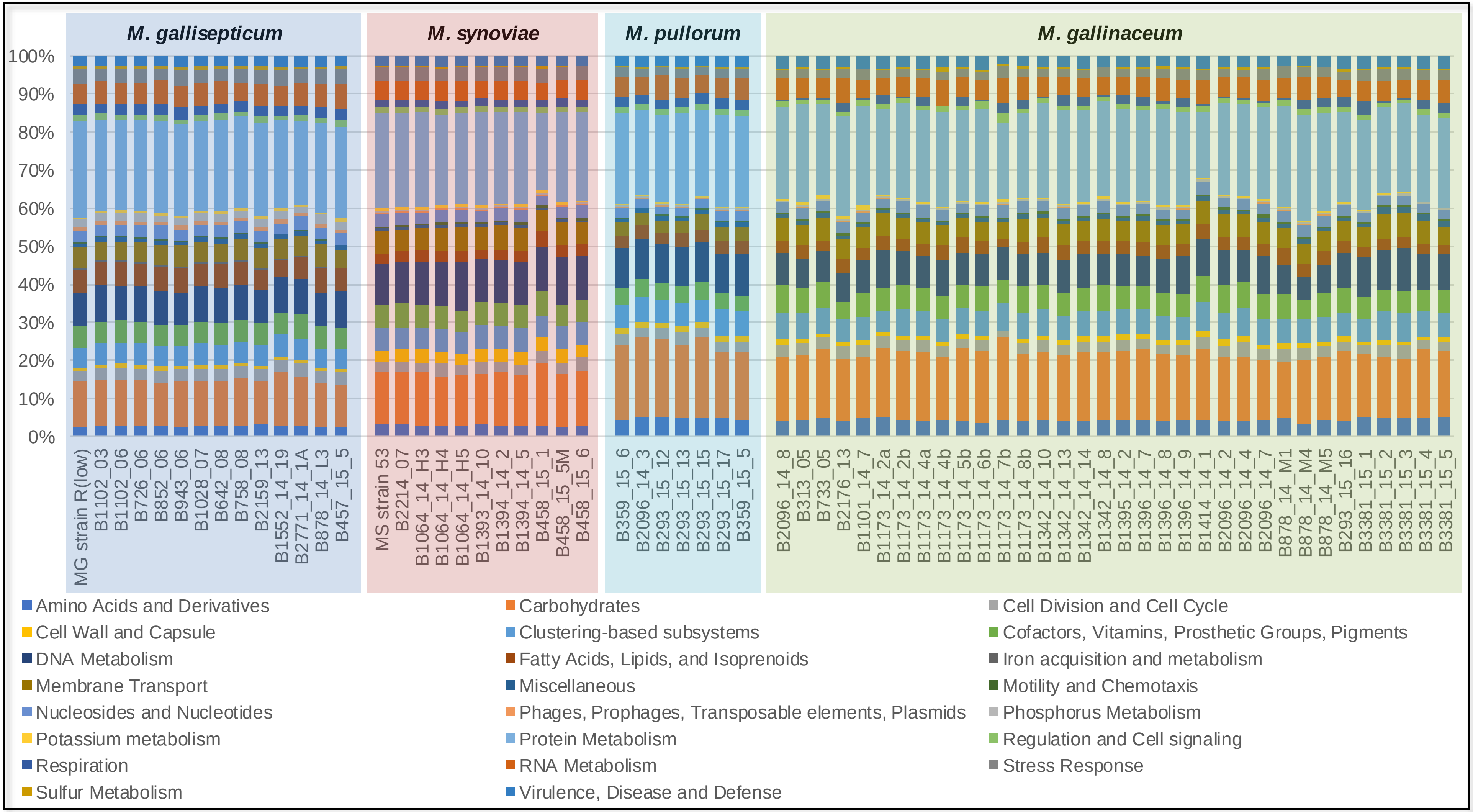


Figure 4-6: Comparison of RAST seed category classification for annotated CDS of strains as percentage of the total amount of classified CDS per strain, excluding the hypothetical protein category to for better definition of other categories

M. pullorum strains showed three main segments of conservation (Figure 4-8). Some rearrangement was observed for strain B293-15-12 and B293-15-15 and possibly strain B293-15-17. Strains B293-15-12, B293-15-13, B293-15-15 and B293-15-17 were isolated from the same batch and location, the same is true for strains B359-15-5 and B359-15-6 (Chapter 2). The length of conserved regions observed in *M. gallinaceum* strains were smaller than the other strains, and due to the size of the sample set and small regions of conservation, it was difficult to observe distinct patterns of rearrangement in the strains (Multimedia).

The backbone view colours conserved regions based on in their presence in each of the strains, for example conserved regions found in all strains are coloured mauve, other colours are used to show regions only present is some of the strains. As with the LCB colour view, the figures for MG and *M. gallinaceum* are only available in multimedia format due to their size. Most of the conserved regions of MG were found in all MG strains tested, one conserved region was found in all strains except strain B457-15-15 and two conserved regions were only observed in the reference and strains B1102-03, B726-06, B943-06, B1028-07, B642-08 and B2159-13. Most of the conserved regions was observed to be present across all MS strains and no large conserved regions was observed in only some strains (Figure 4-9). Most of the conserved regions observed for the *M. pullorum* strains were also found across all the genomes, and strains B293-15-13, B293-15-15 and B293-17 had a conserved region pattern not found in the other strains (Figure 4-10). Among the *M. gallinaceum strains,* only about half of the strains were observed to contain the same conserved regions, almost half of the other major conserved regions were only observed in some of the strains.

The BLAST matrix tool of the CMG package was used to produce a pairwise comparison of the proteins between the strains. When two proteins align at least 50% identity and the alignment length is at least 50% of the longest protein, the protein alignment is considered as significant and the proteins grouped in the same protein family. In the context of this program package, the amount of shared protein families between two strains determines the homology and the homologies within each strain are defined as paralogs, which in turn is determined by finding matching protein families within the genome itself (Vesth et al., 2013). Homologies observed between the MG proteomes were in the range of 69 and 96.8%, with the highest homology observed between strain B1102-03 and B1102-06 (Figure 4-11). Homology observed within each strain ranged from 2.2 to 12.1%, strain B457-15-5 and B1102-06 had the highest and lowest level of similarity within the respective strains (Figure 4-11).

MS strains showed homology between strains in the range of 63 to 92.3%, with the highest homology shown between strains B1393-14-10 and B1395-14-5 (Figure 4-12). Homology within each strain was observed in the range of 2.6 to 7%, with strains B1393-14-10 showing the highest degree of homology (Figure 4-12). The shortest range of homology was observed between the *M.*

pullorum strains, ranging from 82.5 to 91.4%, with the highest homology observed between strain B359-15-5 and B359-15-6 (Figure 4-13). The observed range of homology between the *M. pullorum* strains were between 1.2 and 7.8% with highest internal homology observed for strain B2096-14-3 and the lowest for strain B293-15-17. The BLAST matrix generated for the *M. gallinaceum* strains was too large and is available as part of the multimedia. The range of homology observed between these strains was the longest of all the species, ranging from 38.4 to 93.3%. The highest homology observed was between strain B2096-14-4 and the reference strain B2096-14-8. Homology within the *M. gallinaceum* strains was the lowest and highest of all the species in the range of 0.6 and 14.4%, respectively with strain B1173-14-7 showing the highest and strain B3381-15-3 the lowest homology within their genomes.

A pan- and core genome plot was also generated for each species as well as for all species combined. The pan genome is an accumulation of all the genes present across the submitted dataset, and the core genome is a set of genes that is conserved among the submitted dataset. The same parameters as the pairwise analysis of the BLAST matrix that 50% of the alignment must be identical and 50% of the longest protein must be aligned was used. The pan genome of the MG strains contained 997 genes and the core genome were made up of 573 genes (Figure 4-14). Across the MS strains 990 genes formed part of the pan genome and 544 genes made up the core genome (Figure 4-15). The pan genome of the *M. pullorum* strains consist of 951 genes and the core genome contained the most genes with 673 genes (Figure 4-16). Analysis of the *M. gallinaceum* strains resulted in a pan genome containing 1580 genes and core genome containing only 220 genes (Figure 4-17). Comparing the genes in all the samples resulted in a pan genome of 3218 genes and core genome of only 26 genes (Figure 4-18 and Table 4-4).

4.3.4. Candidate gene identification

The protein file containing the list of candidate core genes were submitted to WebMGA for assigning COG categories to each of the proteins. Of the 573 core genome candidates found for MG, were 388 assigned to 21 COG categories and 113 were classified as hypothetical proteins the remaining candidates could not be assigned to a functional category (Figure 4-19(a)) The core genome for MS consisted of 544 genes, 383 of these were assigned to 21 COG categories and 119 as hypothetical proteins (Figure 4-19(b)). Of the 673 core *M. pullorum* genes could 390 be assigned to 20 COG categories, and an additional 213 were classified as hypothetical proteins, the remaining genes could not be classified to a functional category (Figure 4-19(c)). Of the 220 *M. gallinaceum* core genes were 157 assigned to 19 COG categories and 96 were classified as hypothetical proteins (Figure 4-20(a)). All 26 of the core genes across all species could be assigned to 8 COG categories (Figure 4-20(b)).

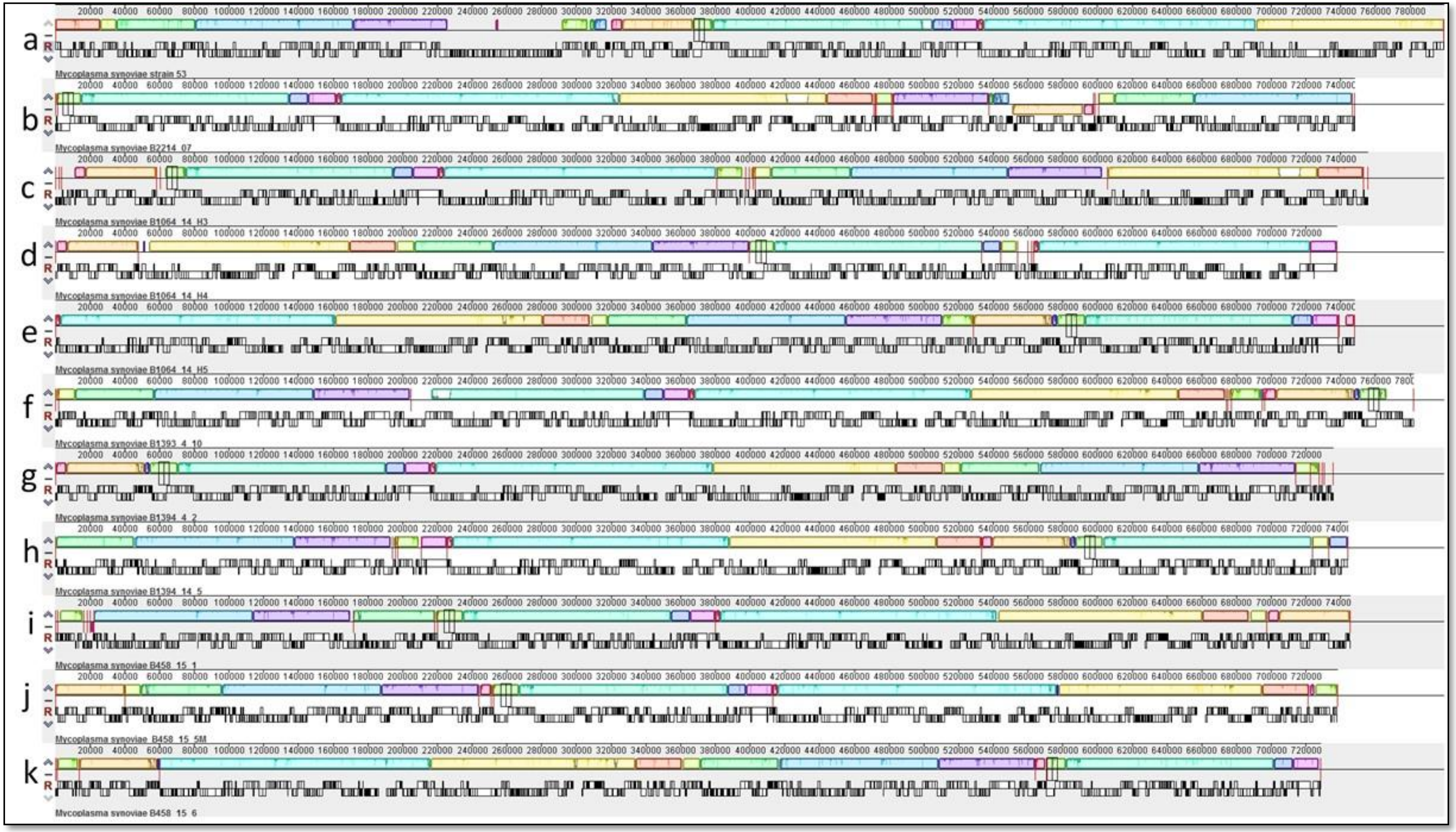


Figure 4-7: Linear view of locally collinear block (LCB) colour scheme alignment of *M. synoviae* strains as determined by the progressiveMauve algorithm of the Mauve software. Coloured blocks represent segments of conserved DNA between the strains that are internally free from genome arrangements. Black and white blocks represent the genes identified in each strain. Blocks facing upwards are found on the sense strand and blocks facing downwards are found on the antisense strand. Strains in order are a) MS strain 53 b) B2214-07 c) B1064-14-H3 d) B1064-14-H4 e) B1064-14-H5 f) B1393-14-10 g) B1394-14-2 h) B1394-14-5 i) B458-15-11 j) B458-15-5M and k) B458-15-6.

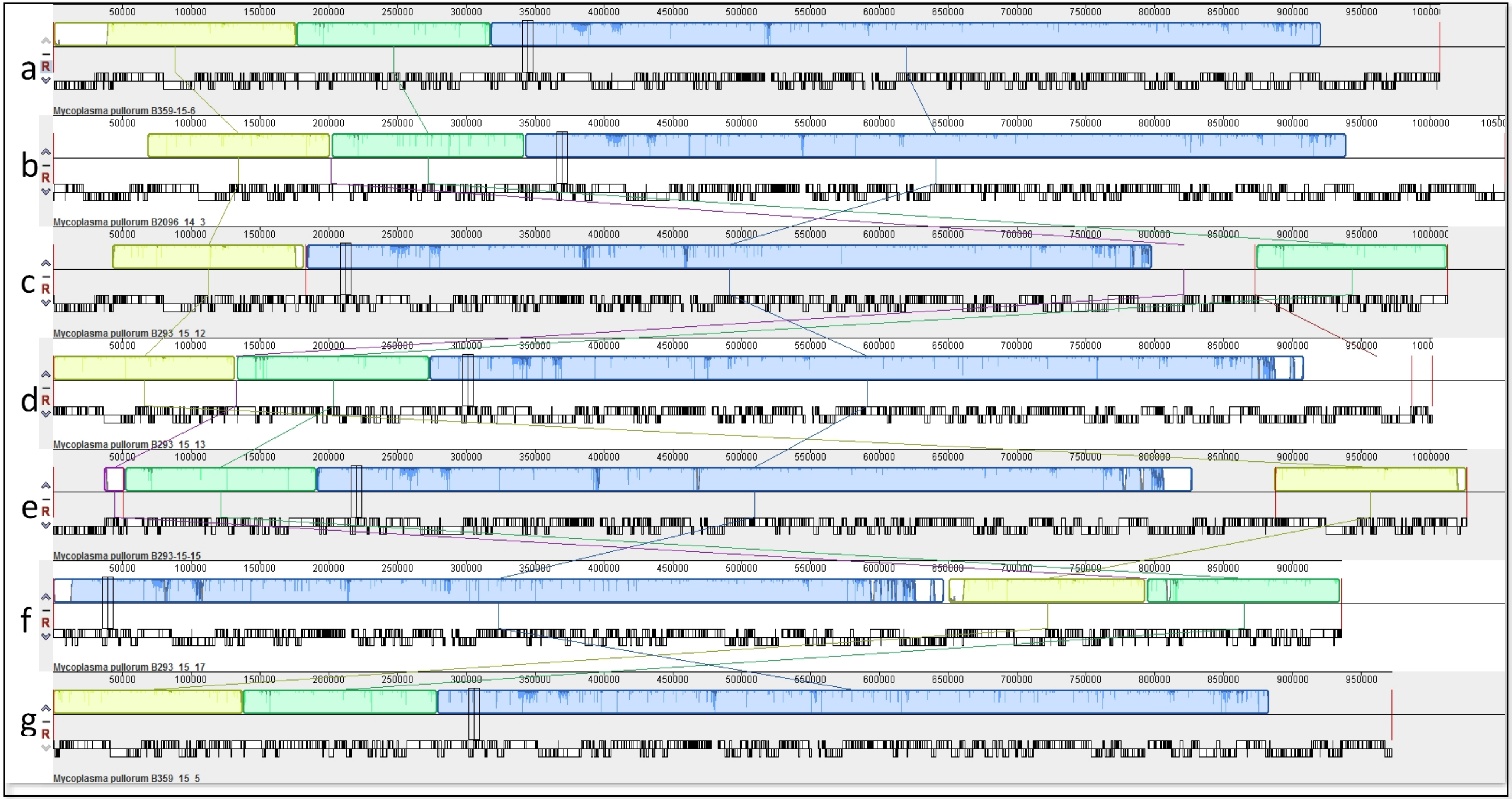


Figure 4-8: Linear view of locally collinear block (LCB) colour scheme alignment of *M. pullorum* strains as determined by the progressiveMauve algorithm of the Mauve software. Coloured blocks represent segments of conserved DNA between the strains that are internally free from genome arrangements. Black and white blocks represent the genes identified in each strain. Blocks facing upwards are found on the sense strand and blocks facing downwards are found on the antisense strand. Strains in order are a) B359-15-6 b) B2096-14-3 c) B293-15-12 d) B293-15-13 e) B293-15-15 f) B293-15-17 and g) B359-15-15.

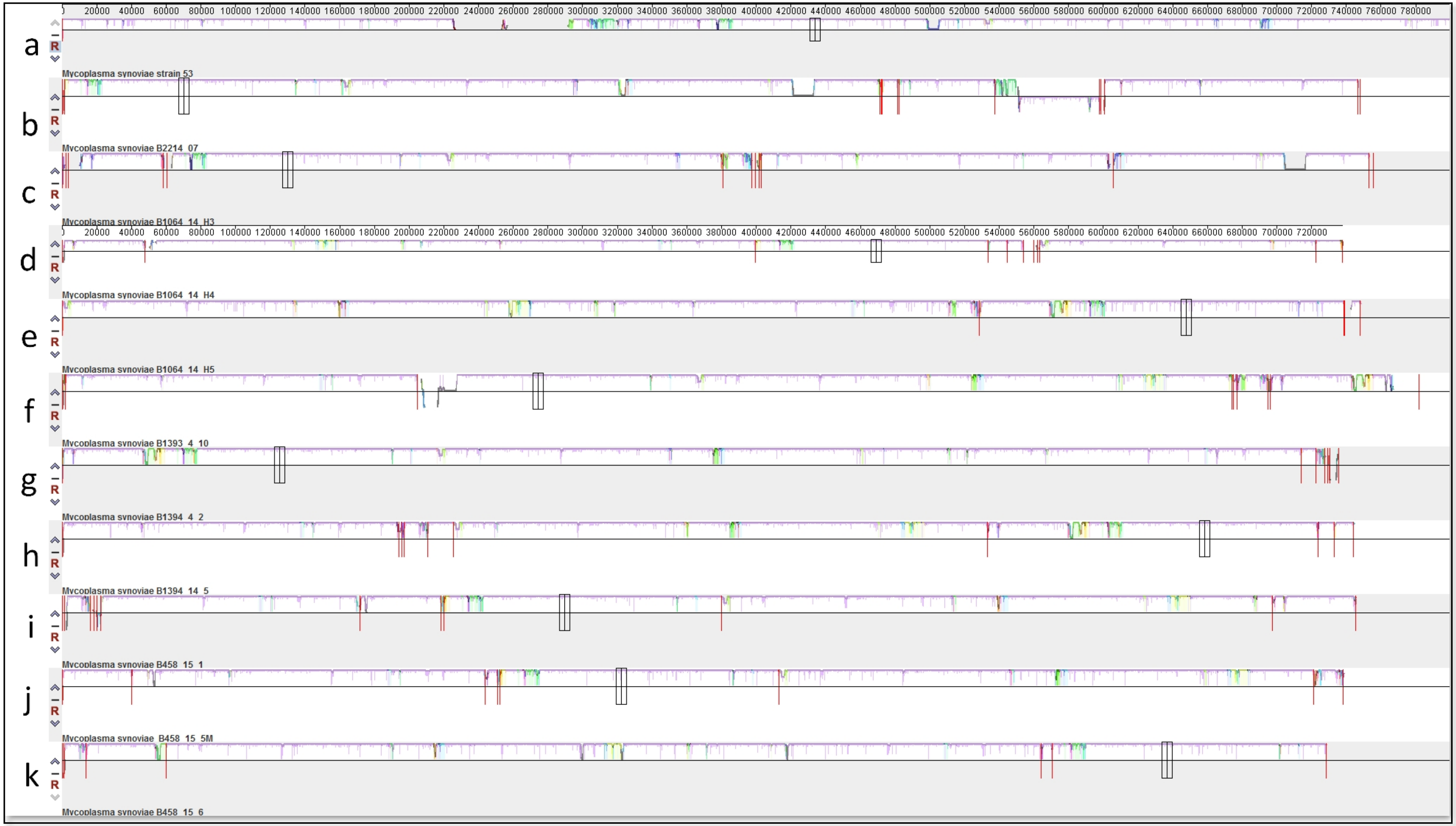


Figure 4-9: Linear view of backbone colour scheme of progressiveMauve alignment algorithm of *M. synoviae* strains. Mauve coloured lines indicate conserved regions present in all aligned strains and other coloured regions indicate conserved regions only present in some of the strains. Strains in order are a) MS strain 53 b) B2214-07 c) B1064-14-H3 d) B1064-14-H4 e) B1064-14-H5 f) B1393-14-10 g) B1394-14-2 h) B1394-14-5 i) B458-15-11 j) B458-15-5M and k) B458-15-6.

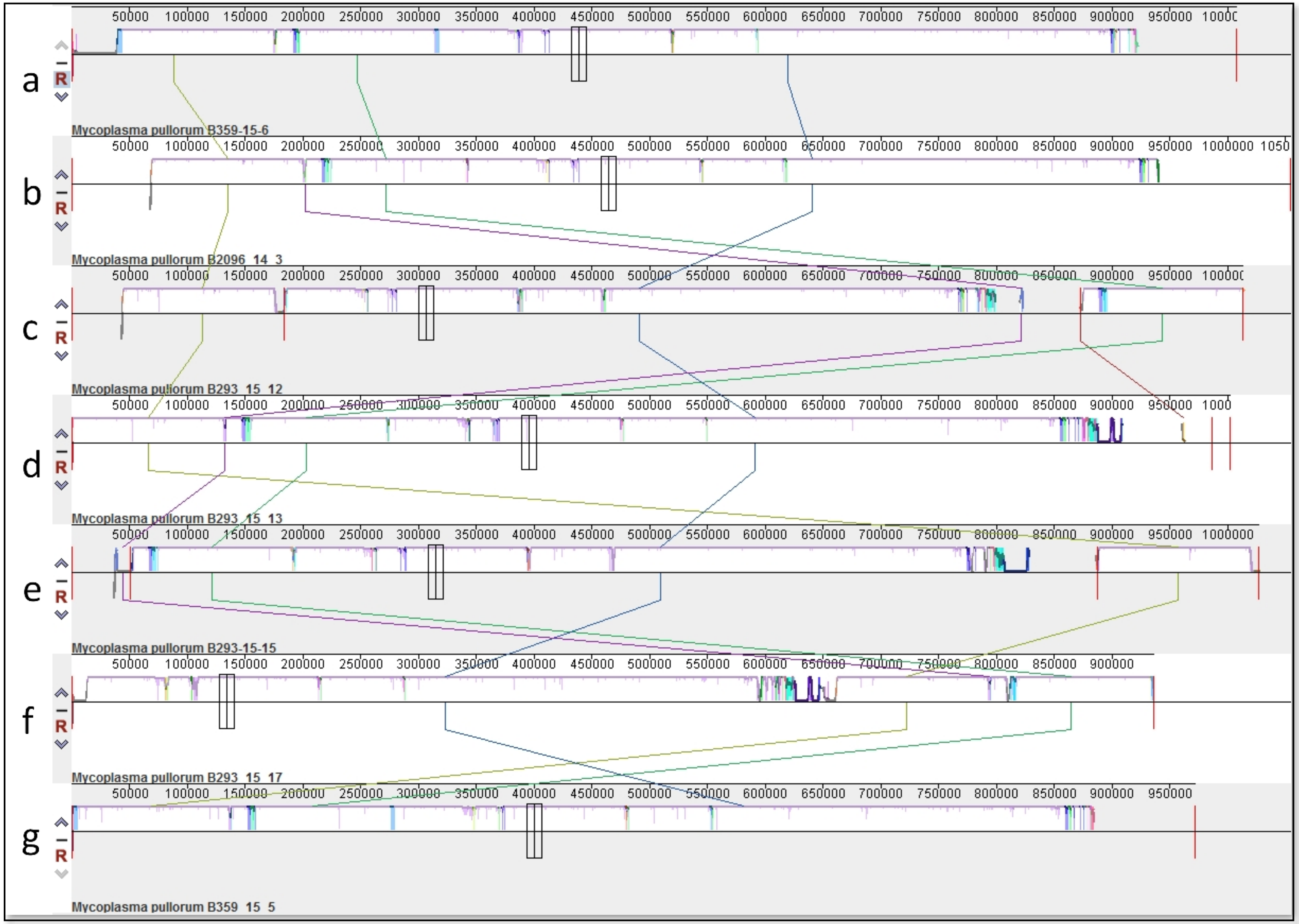


Figure 4-10: Linear view of backbone colour scheme of progressiveMauve alignment algorithm of *M. pullorum* strains. Mauve coloured lines indicate conserved regions present in all aligned strains and other coloured regions indicate conserved regions only present in some of the strains. Strains in order are a) B359-15-6 b) B2096-14-3 c) B293-15-12 d) B293-15-13 e) B293-15-15 f) B293-15-17 and g) B359-15-15.

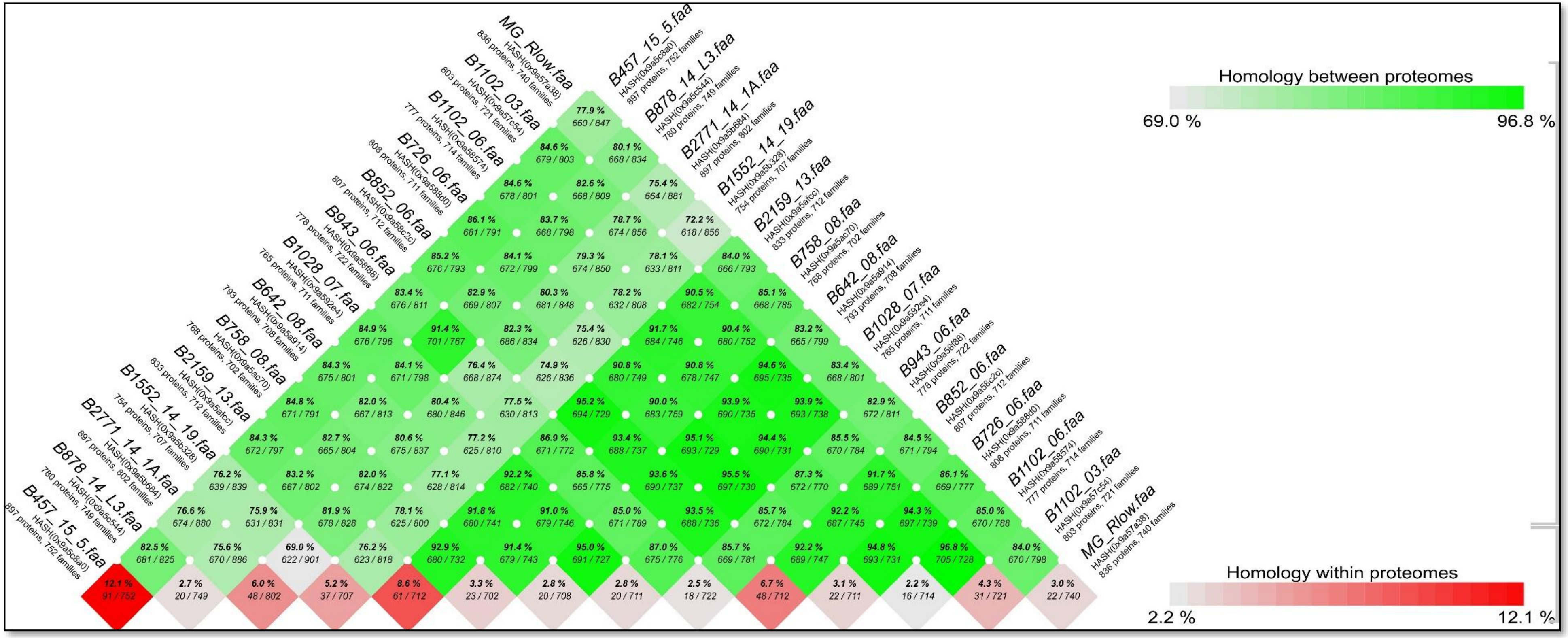


Figure 4-11: BLAST matrix comparing *M. gallisepticum* strains. Green blocks depict homology between strains, and red blocks depict homology (paralogs) within each strain. Homology is determined by pairwise alignment and significance of the alignment is dependent on a 50% identify between the proteins as well as at least 50% of the longest proteins has to be aligned (Vesth et al., 2013).

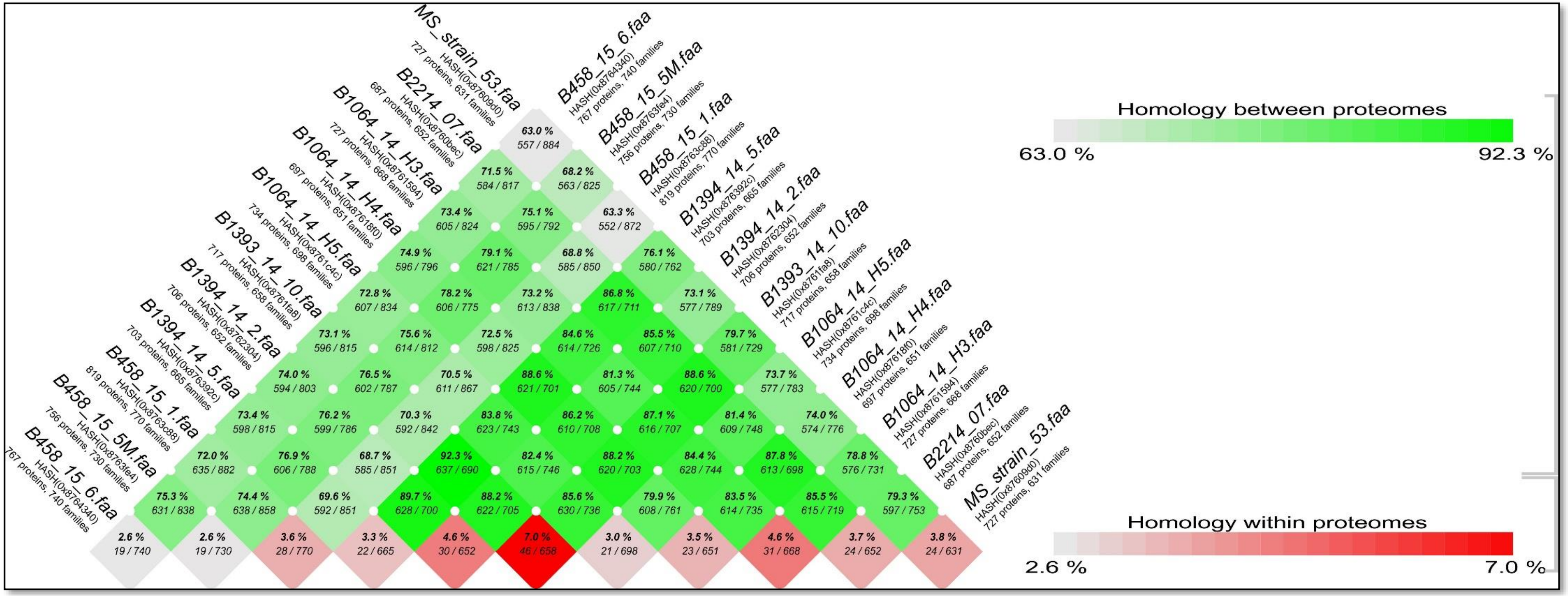


Figure 4-12: BLAST matrix comparing *M. synoviae* strains. Green blocks depict homology between strains, and red blocks depict homology (paralogs) within each strain. Homology is determined by pairwise alignment and significance of the alignment is dependent on a 50% identify between the proteins as well as at least 50% of the longest proteins have to be aligned (Vesth et al., 2013).

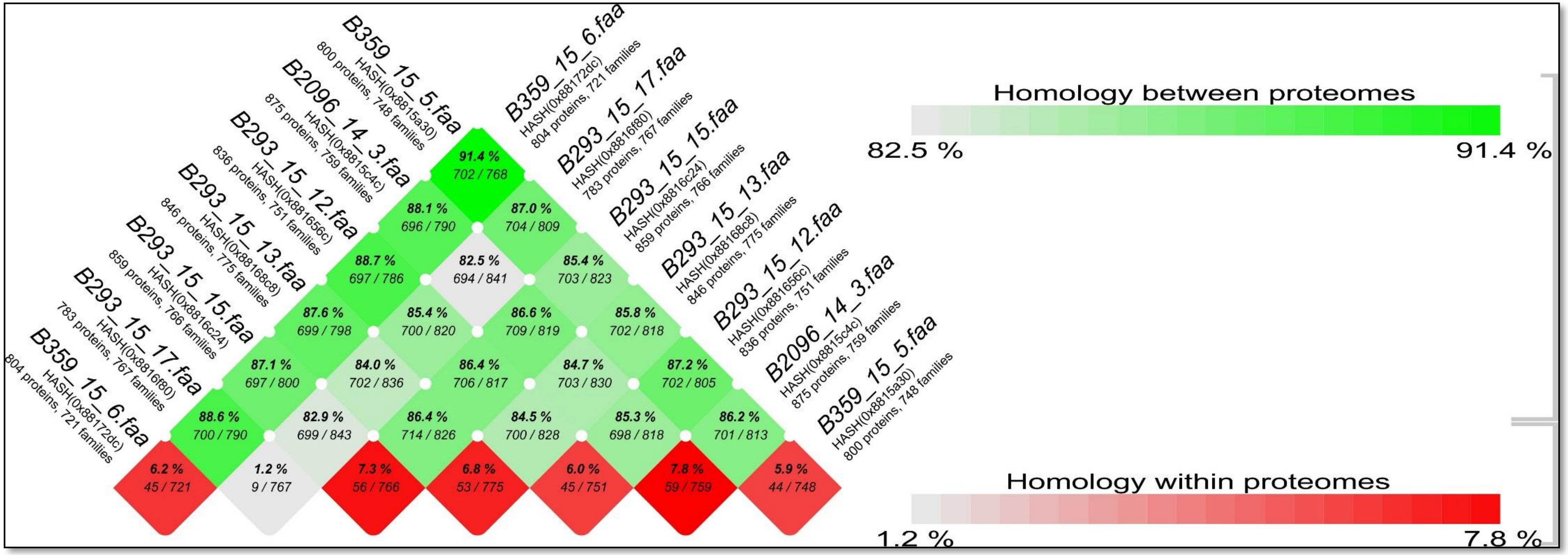


Figure 4-13: BLAST matrix comparing *M. pullorum* strains. Green blocks depict homology between strains, and red blocks depict homology (paralogs) within each strain. Homology is determined by pairwise alignment and significance of the alignment is dependent on a 50% identify between the proteins as well asat least 50% of the longest proteins has to be aligned (Vesth et al., 2013).

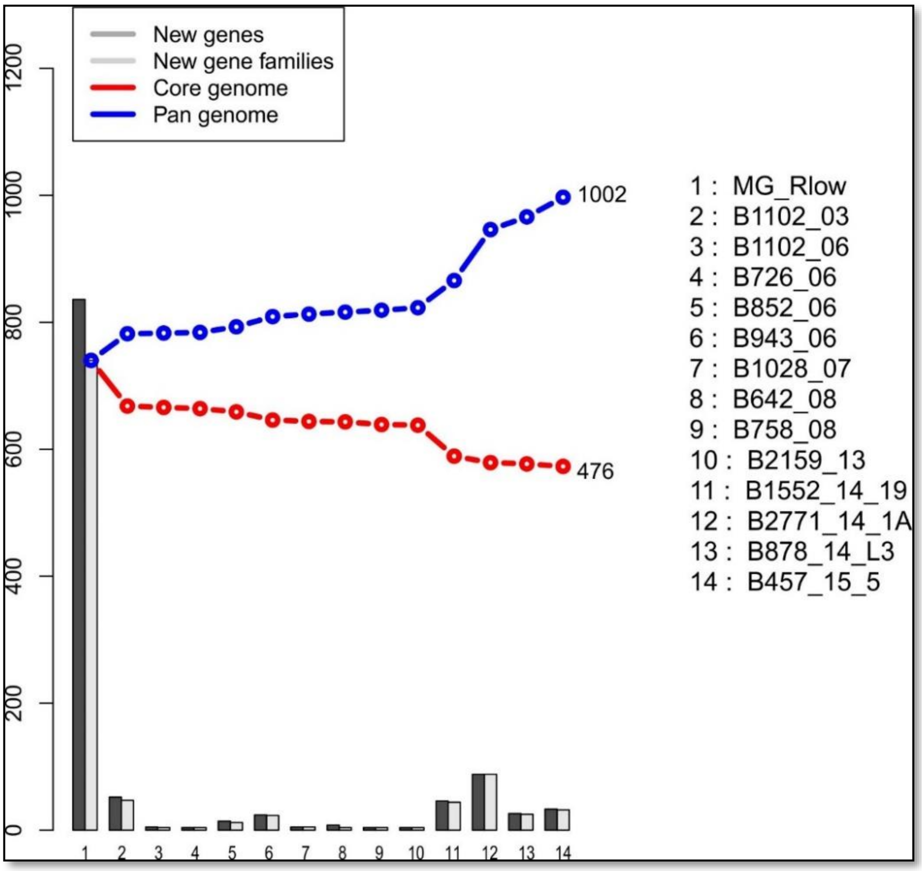


Figure 4-14: Pan genome analysis of *M. gallisepticum* depicting the core and pan genome of the strains.

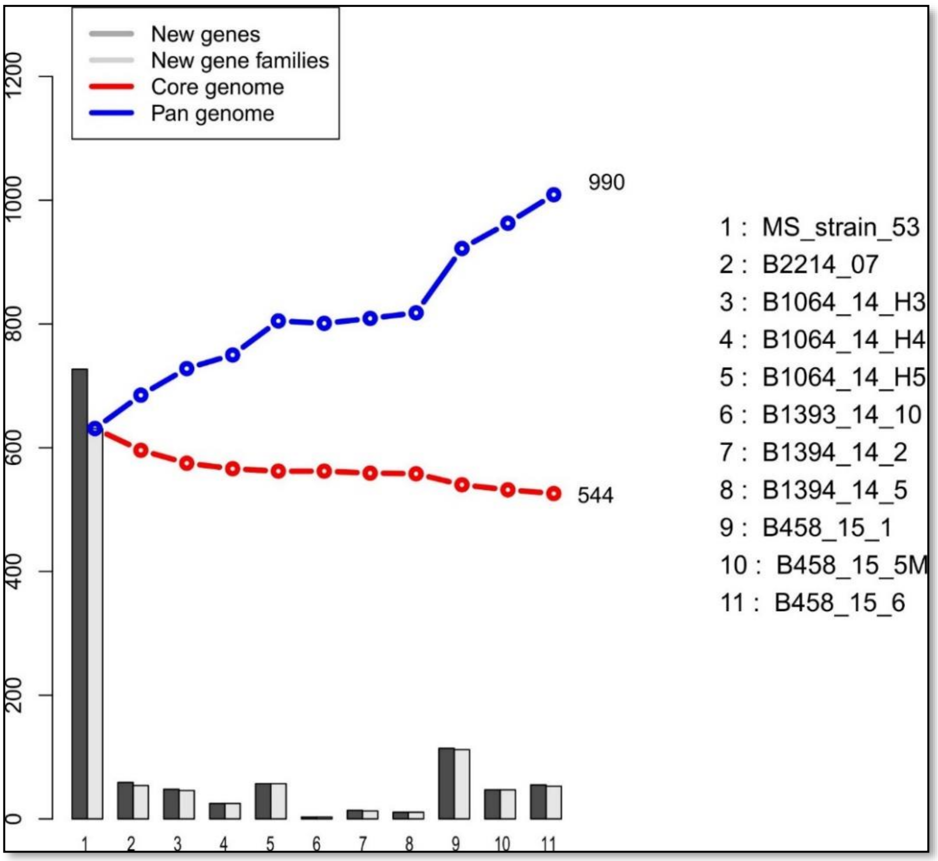


Figure 4-15: Pan genome analysis of *M. synoviae* depicting the core and pan genome of the strains.

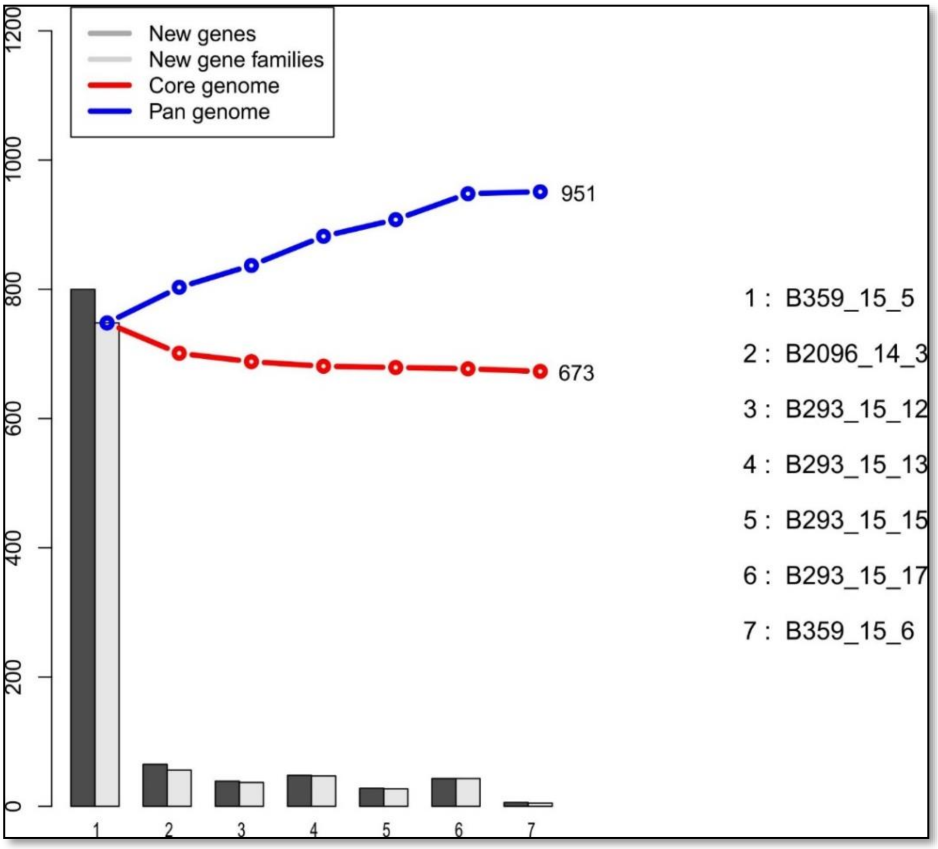


Figure 4-16: Pan genome analysis of *M. pullorum* depicting the core and pan genome of the strains.

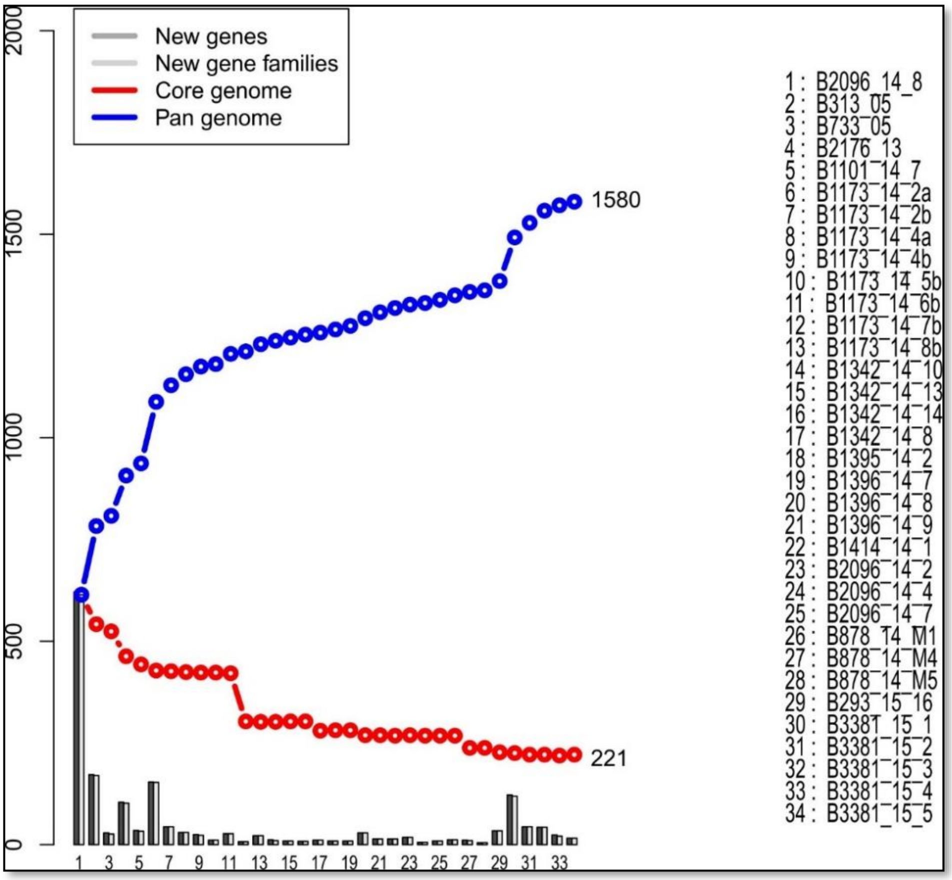


Figure 4-17: Pan genome analysis of *M. gallinaceum* depicting the core and pan genome of the strains.

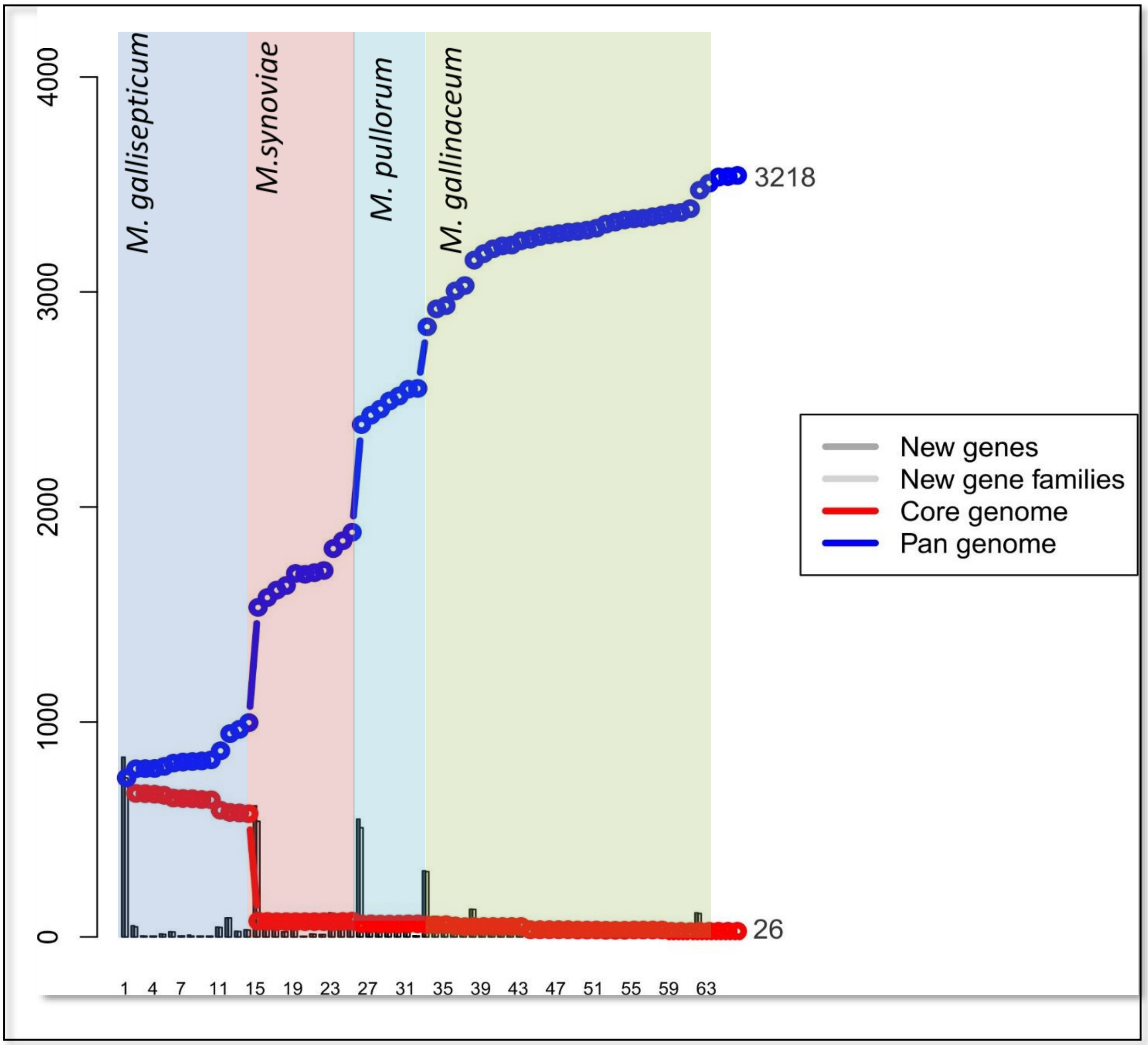


Figure 4-18: Pan genome analysis of all species depicting the core and pan genome of the all strains tested.

Table 4-4: List of core genes for all mycoplasma species with the *M. gallinaceum* protein accession number.

Hit Definition	Accession	Hit Definition	Accession
30S ribosomal protein S11	AKA49980	Fructose-bisphosphate aldolase	AKA49871
30S ribosomal protein S12	AKA49959	Glucose-inhibited division protein A	AKA49686
30S ribosomal protein S13	AKA49979	Glycyl-tRNA ligase	AKA50218
30S ribosomal protein S7	AKA49960	Inorganic pyrophosphatase	AKA50022
30S ribosomal protein S9	AKA49911	Molecular chaperone DnaK	AKA49816
50S ribosomal protein L1	AKA49905	PAP phosphatase	AKA49818
50S ribosomal protein L11	AKA49906	Phosphocarrier protein HPr	AKA49708
50S ribosomal protein L19	AKA49992	Phosphoglycerate mutase	AKA49902
50S ribosomal protein L20	AKA49919	Preprotein translocase subunit SecA	AKA49915
50S ribosomal protein L33	AKA49877	Ribosome-binding ATPase YchF	AKA49996
ABC transporter permease	AKA49896	Translation initiation factor IF-1	AKA49977
DNA gyrase subunit A	AKA49729	Translation initiation factor IF-2	AKA50004
Elongation factor G	AKA49961	Transposase	AKA50233

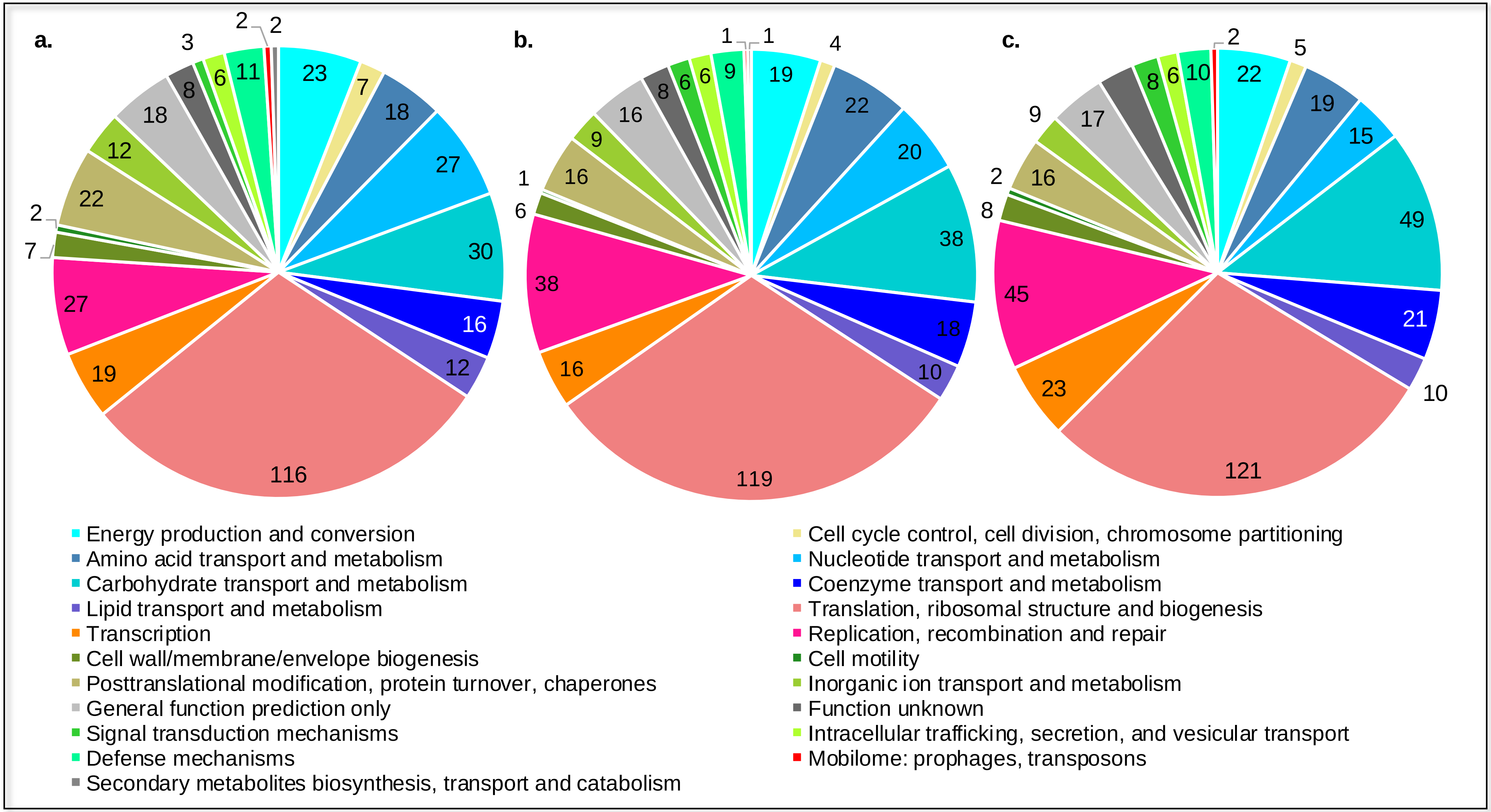


Figure 4-19: COG categories assigned to core genes for a) *M. gallisepticum* strains, b) *M. synoviae* strains and c) *M. pullorum* strains.

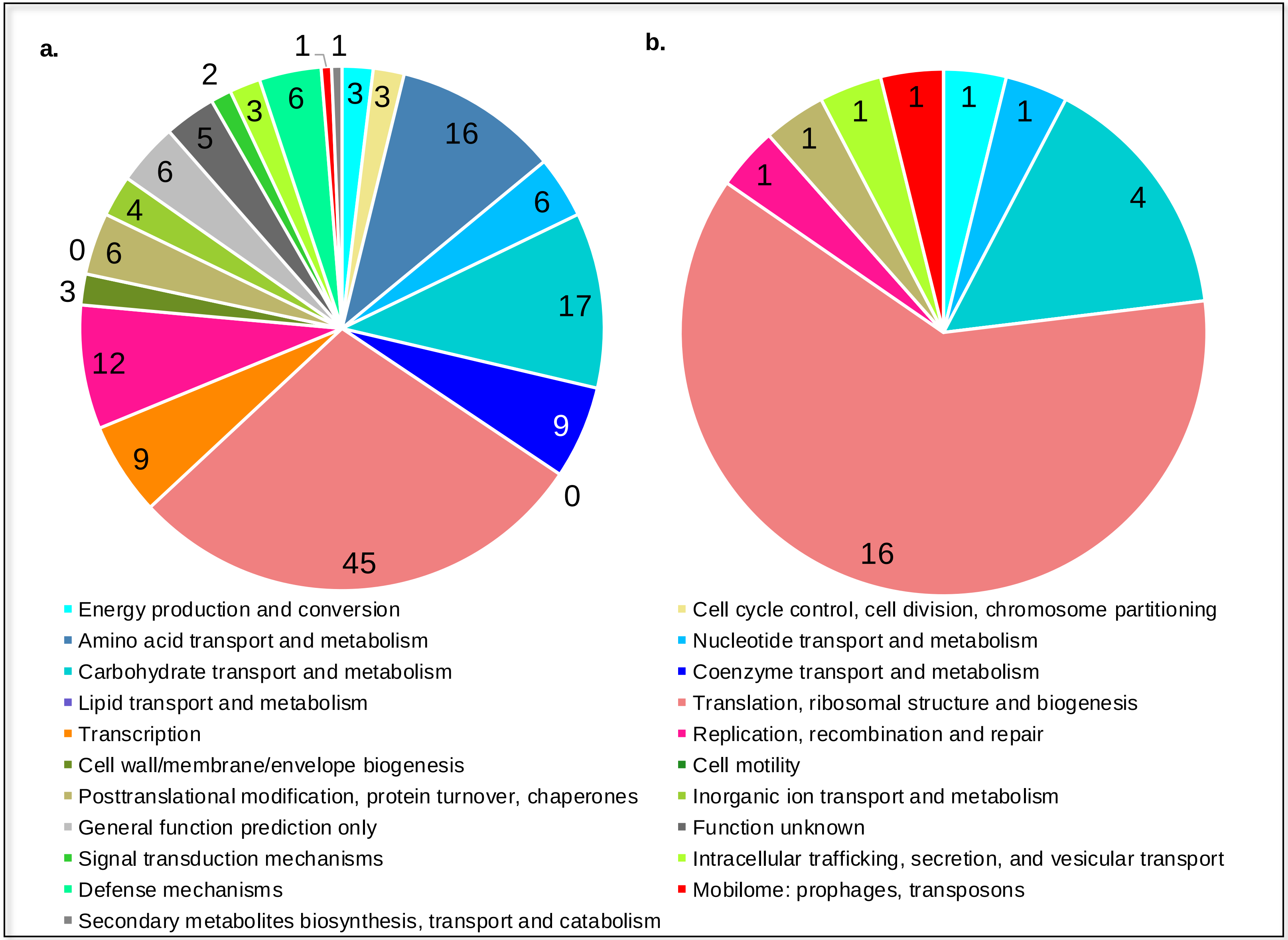


Figure 4-20: COG categories assigned to core genes for a) *M. galllinaceum* strains and b) all the strains combined.

4.4. Discussion

During this study, 124 samples in total were analysed, of which 80 samples were found to be axenic isolates. One of these isolates was identified as *Acholeplasma laidwalli* and as the focus of this study was mycoplasma genomes, this sample was excluded from further analysis along with 11 isolates that were identified as *M. gallinarum* which does not have a reference genome available yet. The remaining 68 axenic mycoplasma isolates consisted of 15 MG, 11 MS, 8 *M. pullorum* and 34 *M. gallinaceum* strains that were sequenced using predominantly Ion Torrent sequencing technology. Samples received prior 2014 were sequenced using Illumina MiSeq sequencing with high depth of coverage ranging from 133 to 1923 times and 143 to 216 times, respectively which is much higher than the minimum recommended depth of coverage for whole genome assembly of at least 30 times.

Each strain was independently *de novo* assembled using CLC Genomics Workbench for both Ion Torrent and Illumina reads as well as the recommended assemblers MIRA through the IonGap pipeline for Ion Torrent reads and SPAdes for the Ilumina reads. As with MIRA and CLC Genomics Workbench, SPAdes also uses *De Bruijn* graphs but also uses variations of the *A-Bruijn* graphs to remove bubbles and chimeric reads (Bankevich et al., 2012). Resulting contigs were then aligned to a reference genome and overlapping contigs joined to produce draft assemblies of less than 25 contigs. Mixed methods for genome assembly, such as reference guided *de novo* assembly are also useful for genomes with a high occurrence of repeat sequences as is the case with mycoplasma genomes (Sims et al., 2014). Highly variable regions were difficult to align, and laboratory methods, such as primer walking or third and fourth generation sequencing technologies that produces larger reads can be used to close the gaps if required. For this study, draft assemblies were adequate as good candidate genes include genes present in all strains or species being studies. Draft genome assemblies were successfully annotated by the RAST server. Different annotation programs use different algorithms for determining CDS assign different naming conventions to the identified genes. The reference genomes were submitted to the RAST server for reannotated and consistency in gene naming across all strains, so difference between these and the genome annotations available from the NCBI server are expected.

Additional metrics are used to assess the quality of genome assemblies if a reference is available, to address some of the problems associated with the N50 metric, as discussed in Chapter 2. The N50 metric is subjective as it can be manipulated by choosing a cut-off for contig size and is dependent on the total length of the assembled contigs using the NG50 metric instead the assembly metrics are standardized to the length of the reference genome for a better comparison (Earl et al., 2011, Elin Videvall, 2017). Misassemblies can occur when contigs are joined without the reference genome, as was the case in the final steps of reference-guided *de novo* draft assembly method used in this study. Quast introduced additional metrics to address possible problems associated with misassemblies by aligning the contigs to the reference and splitting the contigs that align on different regions of the reference genome and the N50 and NG50 values are recalculated based on the split contig lengths and presented as NA50 and NGA50 (Gurevich et al., 2013). These misassemblies are further classified by Quast as relocations (parts of contig align on different regions of reference genome), translocations to a different chromosome (this classification is not relevant to the single circular double stranded DNA of mycoplasma genomes) or inversion (contigs align in reverse order).

Comparing the metrics reported by Quast of the MG strains draft genome assembly of strain B1395-14-1 was the shortest assembly compared to the reference, and also the worst assembly, this was unexpected as the depth of coverage and quality analysis of this sample was very good. In my opinion there are two possible reasons for this either the genome of this strain is very divergent from the reference strain used, or errors occurred during sequencing including that some regions of the genome was sequenced at a much higher rate than expected than other regions of the genomes, and some regions of genome might not have been sequenced at all. When the reads were mapped to the reference genome, large regions of the reference genome had very low or no read coverage (data not shown), but further analysis is required to study the poor-quality assembly. The same argument could also be made for the weakest assemblies observed for MS strain B458-15-11 and *M. pullorum* strain B540-15-2 (data not shown).

Misassemblies can be due to assembly errors or can represent true genome variation (Gurevich et al., 2013). Differences between mycoplasma species isolated from different countries, paired with observations made during assembly of the draft genomes indicate that the misassemblies reported by the Quast analysis is due to true genome variation that could be caused by numerous events such as rearrangement or large indels (Gurevich et al., 2013). MG (20) strains had the most misassembled contigs on average, followed by *M. gallinaceum* (18), MS (7) and *M. pullorum* (2).

Divergence of MG strains in South Africa from other countries have been shown (Moretti et al., 2013, Bwala, 2017). As most of the MG samples were isolated from archived samples, little information was available on the area these samples were collected, and as these samples were all collected between 2003 and 2015 the amount of observed misassemblies due to relocations in these strains were not unexpected. The progressiveMauve and pairwise alignment of these strains showed high protein indentity and a high occurrence of conservation between strains, however the contigs were joined using the reference strain as a guide, so the observed arrangements might not be a true reflection of the genome arrangement for each, but for the purpose of identifying conserved genes for downstream diagnostic and vaccine development applications are these draft assemblies more than adequate.

Genome statistics and annotations for ten of the MG strains were in range what was expected and considered as very good draft genomes. The five remaining MG strains analysed differed from the reference genome with regards to the amount of rRNAs identified in each. Only one 16S rRNA gene was not found in strain B852-06 and one of each copy of the rRNA genes were not found in strain B1552-14-19, when comparing all the genome statistics analysed these two strains were still considered as good assemblies and the lack of the rRNAs was attributed to a failure to duplicate in the genome assembly step. Sample B2771-14-1B and B1395-14-1 had less and more rRNAs, respectively than was expected and a comparison of the other genome statistics these strains were considered as low-quality assemblies and for the purpose of this study excluded from further analysis. The complete genome of MG was the first of all the poultry mycoplasma species to be published and was assembled using shotgun sequencing with capillary electrophoresis and primer walking (Papazisi et al., 2003). Most of the MG strains assembled in this study was the same size or smaller than the reference genome and is probably close to the correct genome.

Of the eleven MS strains analysed, only strain B458-15-11 was considered as a low-quality assembly for this study and excluded from downstream analysis. MS showed some divergence from the reference genome and MS strains could possibly be more conserved between countries, but further analysis is required. Sample collection information show that the samples were mainly collected from three farms, however the four rearrangement profiles observed shows differences between the samples collected on the same farms and are good candidates for planned future studies on recombination in MS strains. The high level of conservation and protein identity observed between the MS strains also show that the draft genome assemblies are adequate for candidate gene identification. The reference genome for MS was sequenced using shotgun sequencing and the gaps closed using *in silico* methods as well as PCR assisted contig extensions (Vasconcelos et al., 2005). Some of the strains was however longer than the reference genome and it is possible that the reference genome is not complete or strains in South Africa are larger and a more updated reference genome can be assembled using current technologies for future analysis.

Eight *M. pullorum* strains, including the reference genomes, were analysed and only strain B540-15-2 was excluded from further analysis due to an observed low-quality assembly compared to the other strains and for the purpose of this study. As the reference genome of *M. pullorum* was sequenced from a South African strain in the same time period and mainly from two farms in the Gauteng province, little or no divergence was observed as expected between these strains. Possible rearrangement observed for between samples from the same farms also makes these strain good candidates for planned future genome recombination studies of *M. pullorum*. The degree of protein identity observed between these draft genome assemblies are also good for candidate gene finding. The reference genome of *M. pullorum* was assembled using only *in silico* methods using both Illumina and Ion torrent data. Considering that it was the largest among all the *M. pullorum* strains, is this genome probably close to the correct genome, but can still be validated using laboratory and long sequence read technologies, such as Pac-Bio methods.

M. gallinaceum strains were isolated from at least four farms in the Gauteng province and the three farms in the Western Cape and was also the species isolated the most among all strains. These strains grow faster than MG and MS strains which along with the geographical spread of this organism could account for the high level of divergence observed for this species. A higher degree of protein identity was observed between strain isolated from the same farms and provinces than between different farms and provinces and no obvious patterns of conservation could be observed between the strains. The *M. gallinaceum* reference genome had the least amount of tRNAs across all species with only 17 tRNAs which is less than the recommend amount of at least one for every amino acid, i.e 21, indicating that the reference genome of *M. gallinaceum* might not be completed, which is possible as this genome was assembled using only *in silico* methods with Illumina data.

More than 50% of all CDS were assigned to categories with known protein function with the most assigned to genes involved in protein metabolism, including protein degradation, biosynthesis, folding processing and modification; followed by carbohydrates, which includes carbohydrate metabolism, CO_2 fixation, and fermentation, followed by DNA metabolism that includes DNA recombination, repair and replication. The amount of CDS assigned to DNA metabolism is not unexpected as mycoplasma evolved by reductive evolution, and genes involved in replication were prioritised, however at the expense of genes involved in biosynthesis of lipids, amino acids and co-factors (Bradbury, 2005, Razin et al., 1998). Closer inspection of the genes identified in the protein metabolism category, showed that most of these genes are involved in biosynthesis of ribosomal proteins that play a key role in protein translation and the genes involved in carbohydrate metabolism are required for ATP synthesis as was expected (Arraes et al., 2007).

A number of interesting differences were observed in the classification of proteins across all species. MG strains contained CDS encoding proteins involved in iron metabolism, which has been shown in *Campylobacter jejuni* to play a role in iron storage as well as protection against an high iron environments and the resulting oxidative stress caused by these environments (Wai et al., 1996). Most of the cell wall and capsule categorised CDS identified for MS play a role in sialic acid metabolism which has been shown to play a role in bacterial pathogenesis of *Escherichia coli, Pasteurella multocida* and numerous other bacteria (Li and Chen, 2012).

The pan- and core genome analysis produced a theoretical representation of the total gene complement and conserved gene complement of each dataset, respectively. The MG strains theoretically had 573 candidate genes for future analysis, 388 of these could be assigned to COG categories and 113 were hypothetical proteins. The MS strains have a theoretical candidate gene list of 544 genes, 383 could be assigned to COG categories for easy evaluation in the future, 119 were hypothetical proteins for which protein IDs still need to be assigned. In the *M. pullorum* strains studied, the candidate genome list for this species was the largest with 673 genes, of which 390 could be assigned to COG categories and 213 were hypothetical proteins this is probably due to the high level of conservation observed between these strains. *M. gallinaceum* strains had the least conservation and protein identity between its strains and produced the smallest list of candidate genes with 220 genes. COG categories could be assigned to 157 of these and 96 were hypothetical proteins. These are still large lists of genes, and the discriminatory power of each gene still needs to be evaluated.

The largest COG class across all the species contain genes involved in translation, ribosomal structure and biogenesis, and these mostly contain ribosomal proteins that are generally highly conserved within a species and might not give good discriminatory power to distinguish strains from each other. But an alignment of the genes that encode these proteins will aid studying these genes and study their potential as candidate genes for multiplex diagnostic assays and vaccine development. Genes that play possible roles in pathogenesis, such as the genes involved in metabolism, motility and defence mechanism categories are good categories to start with for both diagnostic and vaccine development targets. As the list of candidates produced during the pan-core genome analysis test for 50% identity, there is a high probability of finding good candidate genes with good discriminatory power that can be used in laboratory assays to differentiate between different species or strains.

A comparative analysis across all strains and reference genomes used in this study resulted in a list of candidate genes of only 26 genes. Ribosomal proteins are known to be conserved within and between species, and the discriminatory power of the 16S rRNA gene is well known. Sixteen genes form part of this COG category of which ten encode various ribosomal proteins. Due to the highly conserved nature of these ribosomal proteins are these good candidates for vaccine development. Nine of the candidate putative genes encoding DNA gyrase subunit A, elongation factor G, glucose-inhibited division protein A, molecular chaperone DnaK, ribosome-binding ATPase UYchF, tranposase, preprotein translocase subunt SecA and translation initation factors IF-1 and IF-2 also play various roles in DNA replication, tRNA modification, protein translation and intracellular protein transport. Two more of the putative genes encode fructose-bispohate aldolase and phospoglycerate mutase that play a role in glycolysis. One putative gene encodes the phospocarrier protein HPr that forms part of the phospoenolpyruvate-dependent sugar phosphotransferase system (PTS) and three putative genes encode proteins glycyl-tRNA ligase, inorganic pyrophosphatse and PAP phosphatase that play a role in amino acide metabolism, lipid metabolism and nucleotide metabolism, respectively. The last putative candidate gene encodes the ABC transporter permease protein UgpE that play a role carbohydrate transport and metabolism. Little information is available on most of these putative genes and a more in-depth study on their functions as well as suitability as candidate genes is required.

The use of reference-guided *de novo in silico* assembly methods for genome assembly can be time consuming, depending on the characteristics of the bacterial genome of interest; as well as computationally resource-intensive, but is an effective starting point for finding conserved regions between species of the same genus as well as different strains of a species. Comparative genome analysis using various available bioinformatic tools is furthermore an effective method of identifying candidate genes for vaccine and diagnostic assay development, as well as shortlist candidates without the need for expensive, time-consuming laboratory methods. This is however just theoretical and laboratory methods, such as DNA microarray and two-dimensional gel electrophoresis are still required to analyse the characterise these candidate genes, i.e. determine whether these genes are expressed and to analyse the protein products which can then be used to narrow the list down even more and could aid in assigning functions to the large amount of hypothetical proteins. Transcriptome analysis using RNA-sequencing (RNA-Seq), instead of DNA microarrays is also a useful tool to study the RNA expression, but this technology is very expensive and not yet feasible for a large comparative omics project, such as the current project.

CHAPTER 5: CONCLUSION AND FUTURE PERSPECTIVES

Second generation sequencing (SGS) technologies produces high throughput DNA sequencing data that have changed the face of genome sequencing, changing the rate of whole genome sequencing (WGS) from months and years to days and weeks as well as decreasing the cost involved. This led to major growth in the field of the bioinformatics and the size of scientific studies. Even though *in silico* methods can never completely replace laboratory methods for studying biological systems, it can shorten the time requirements for answering various biological questions. However, as I have learned in this study, things are not always so easy and various factors can influence the effectiveness with which these SGS datasets can be processed, such as 1) the choice of sequencing technology and the effect this has on how the data is interpreted and processed, 2) the amount of data generated is computationally very heavy requiring strong computer capabilities, 3) the sequence quality and depth of coverage influence what type of information can be generated from the data set and 4) the genomic characteristics of a species of interest, such as G+C content and frequency of repeats can influence the quality of data generates, as well as the amount of processing times required to analyse the data (Schatz et al., 2010).

WGS using SGS has various applications and in this study WGS data for 124 mycoplasma positive samples isolated from commercial poultry in South Africa was generated. The dataset has not only been used to identify 178 *Mycoplasma* isolates in both axenic and mixed samples using 16S rRNA phylogeny, but also for a pioneering study on the mutations associated with antimicrobial resistance (AMR) in non-pathogenic *Mycoplasma* species. *Mycoplasma* species isolated included *M. gallisepticum, M. synoviae, M. iners, M. pullorum* and *M. gallinarum*. *M. gallisepticum, M. gallinaceum* and *M. gallinarum* were the most prevalent species found in South African poultry. Novel mutations associated with macrolide resistance were discovered in the L4 protein (I196T) and 23S (G748A) rRNA gene for *M. gallinarum* and *M. gallinaceum*, respectively.

The problems associated with the golden standard for mycoplasma identification assays, i.e. culture with growth inhibition were reiterated in these results, it is thus important to find more sensitive, specific and accurate diagnostic assays. However, genetic differences within and between species especially between countries also has an influence on the effectiveness of both culture and DNA-based methods as any mutation in the target antigen or genes will influence the specificity or sensitivity of these assays (Moretti et al., 2013). In this study the well-used, proven 16S rRNA gene was used as an alternative, and as expected was an effective target for differentiating between mycoplasma species. This gene just has one limitation in that *M. imitans* can't be distinguished from *M. gallisepticum* which can luckily be mitigated by studying 16S-23S intergenic region of the strains identified as *M. gallisepticum* strains for the presence of the putative transposon gene unique to *M. imitans* (Harasawa et al., 2004). Various other genes have been used for differentiation, but these were mainly aimed at distinguishing the different known strains of the important pathogenic mycoplasmas, *M. gallisepticum* and *M. synoviae*. Even though the other species are considered as non-pathogenic, their potential for causing disease has been shown (Moalic et al., 1997), and the high occurrence of the these species found in South African poultry emphasizes the need for monitoring these species, which will require improving diagnostics not just for *M. gallisepticum* and *M. synoviae* strain differentiation, but also for identifying all the poultry mycoplasmas.

Hybrid whole genome assembly strategies were developed in this study for handling the low GC content and high occurrence of repeats characteristic of mycoplasma genomes. The first successful hybrid strategy as described in Chapter 3 used sequencing data from both Illumina and Ion torrent sequencing platforms, as well as different assemblers to produce the first 1 007 271 bp complete genome of a previously unpublished mycoplasma, *M. pullorum*, which was annotated and published under accession number CP017813. However, this method was still time consuming due to problems caused by repeat sequences. Third generation sequencing technologies, such as single molecule real time (SMRT) platform from Pacific Biosciences, also referred to as PacBio sequencing can produce longer read sequences, that can aid with elucidating repeat sequence regions. This technology has only recently been offered in South Africa and could aid in developing an improved complete genome assembly strategy for other unpublished mycoplasmas, such as *M. gallinarum*.

A reference-guided *de novo* assembly strategy was used in Chapter 4, where *de novo* and reference guided assemblies were combined with different assemblers to produce draft assemblies of 78 *Mycoplasma* strains for which reference genome were available, i.e. *M. gallisepticum, M. synoviae, M. pullorum* and *M. gallinaceum*. A basic comparative genomics study of these draft assemblies produced large list of candidates for each *Mycoplasma* species than can be used for novel intraspecies diagnostic strategies and candidates for vaccine development. As expected, proteins from the translation, ribosomal structure and biosynthesis category encompassed the largest part of these lists. Within this class are ribosomal proteins the most prevalent, and the conserved nature of these proteins may be good diagnostic candidates, as the 16S rRNA and 23S rRNA genes have shown. The list of candidates can for each species can also be shortened by aligning the genes for each of the strains available and evaluating their discriminatory power.

The main aim of this study was to identify novel genes that can aid in diagnosis and treatment of *Mycoplasma* species in South Africa using comparative genome analysis. A pan- and core genome analysis of *M. gallisepticum, M. synoviae, M. pullorum* and *M. gallinaceum* produced a shortlist of 26 candidates that can be studied for their potential to differentiate between multiple mycoplasma species without the requirement for additional assays. A large part of these genes were ribosomal proteins and the next step will be to align each of these genes for all the species to assess the possible discriminatory power of these genes.

This is just a drop in the water of all the information contained in the database I have created from the WGS of the mycoplasmas from South African poultry. This database of draft assemblies, and the availability of the raw sequencing data for the axenic and mixed infection will provide data for numerous future studies, including:

- A more in-depth study of AMR and the genetic changes associated with AMR. This is a main priority as the treatment with antimicrobials is currently the standard method of mycoplasma control in South Africa. Understanding the mechanisms behind AMR and identifying genetic markers for AMR can aid in recommending appropriate antibiotics for use not only in treatment, but also for the practise of enhanced growth. This will decrease the risk of spreading AMR faster through inappropriate antibiotic use as well as transfer of the AMR genes between species.
- Further studies on the differences and similarities between and within poultry *Mycoplasma* species can be used to direct *in vivo* and *in vitro* studies horizontal gene transfer studies. This can aid in improving our understanding of mycoplasma survival strategies.
- Complete genome assembly annotation of the first *M. gallinarum* genome.
- Combining data from other species known to interact with mycoplasma species, such as the multifactorial disease complex between *M. gallisepticum* and *Escherichia coli* and infectious bronchitis virus, can improve our knowledge of how this microorganism interactions with the host and environment to cause disease or negatively affect poultry production, which will eventually aid in improved diagnostic and preventative measures.
- PacBio sequencing can be used to improve hybrid assembly strategy for completing the other unpublished mycoplasma genomes.
- PacBio sequencing will be used to improve the draft assemblies of a few representative strains, to produce higher quality, more complete reference genomes and improve on the currently method of identifying core genes between and within species producing a better set of candidate genes for the development of novel diagnostic assays and future vaccine development.

APPENDIX A: RESULTS

A.1. DNA concentration of isolated *Mycoplasma* DNA

Number on agarose gel[a]	Sample	Concentration (ng/µl)	A260	A280	A260/A280	A260/230
	B1102-03	24.2	N/A	N/A	N/A	N/A
	B313-05	53.1	N/A	N/A	N/A	N/A
	B733-05	16.4	N/A	N/A	N/A	N/A
	B1102-06	33.3	N/A	N/A	N/A	N/A
	B726-06	30.8	N/A	N/A	N/A	N/A
	B852-06	19.6	N/A	N/A	N/A	N/A
	B943-06	35.4	N/A	N/A	N/A	N/A
	B1028-07	21.8	N/A	N/A	N/A	N/A
	B2214-07	12	N/A	N/A	N/A	N/A
	B04-09-07	81.5	N/A	N/A	N/A	N/A
	B1072-08	12.8	N/A	N/A	N/A	N/A
	B642-08	42.4	N/A	N/A	N/A	N/A
	B758-08	59.7	N/A	N/A	N/A	N/A
	B730-09	27.7	N/A	N/A	N/A	N/A
	B2076-13-3	29.6	N/A	N/A	N/A	N/A
	B2159-13	53	N/A	N/A	N/A	N/A
	B2176-13	20.6	N/A	N/A	N/A	N/A
	B2888-13-1A	19.8	N/A	N/A	N/A	N/A
50	B1064-14-H3	96.7	1.934	1.082	1.79	1.49
48	B1064-14-H5	105.2	2.104	1.048	2.01	1.56
71	B1101-14-10	118.1	2.362	1.134	2.08	1.54
46	B1101-14-6	15.7	0.313	0.215	0.29	1.46
5	B1101-14-7	175.2	3.504	1.942	2.27	1.8

47	B1101-14-8	59.4	1.189	0.657	1.81	2.24
31	B1101-14-9	35.2	0.705	0.435	1.62	0.95
9	B1064-14-H4	103.4	2.068	1.14	2.28	1.81
21	B1173-14-2a	135.3	2.705	1.591	1.7	0.65
28	B1173-14-2b	61.9	1.239	0.753	1.65	0.95
14	B1173-14-4a	43.3	0.866	0.5	1.75	1.73
10	B1173-14-4b	48.7	0.973	0.545	2.07	1.79
15	B1173-14-5b	32.1	0.641	0.359	1.92	1.78
29	B1173-14-6b	100.4	2.008	1.078	1.86	0.83
6	B1173-14-7b	28.9	0.579	0.313	1.73	1.85
26	B1173-14-8b	41.9	0.838	0.466	1.8	1.52
7	B1342-14-10	29.5	0.59	0.341	1.72	1.73
39	B1342-14-13	50.2	1.005	0.574	2.11	1.75
22	B1342-14-18	53.4	1.067	0.668	1.6	0.89
38	B1342-14-14	25.9	0.519	0.291	2.13	1.78
30	B1342-14-4	12.3	0.247	0.142	1.74	1.47
85	B1342-14-9	159.8	3.195	1.729	2.4	1.85
32	B1342-14-8	17.9	0.357	0.212	1.69	0.75
16	B1393-14-10	45	0.9	0.502	2.09	1.79
23	B1393-14-4	51.4	1.029	0.556	1.85	1.57
20	B1394-14-5	36.7	0.735	0.412	1.21	1.76
11	B1394-14-2	34.5	0.69	0.396	1.87	1.74
12	B1395-14-1	62.3	1.246	0.69	3.54	1.81
19	B1395-14-2	156.7	3.134	1.754	2.05	1.79
65	B1395-14-5	100.4	2.008	1.121	0.93	1.79
86	B1396-14-6	272.6	5.453	2.774	1.97	0.99
40	B1396-14-7	45.9	0.918	0.55	1.53	1.67
49	B1396-14-8	88.1	1.763	1.065	1.65	0.73

66	B1396-14-9	70.7	1.415	0.752	1.88	1.21
34	B1412-14-18	26	0.52	0.297	1.5	1.75
18	B1414-14-1	84	1.679	0.954	1.99	1.76
35	B1552-14-19	35.7	0.714	0.389	1.78	1.84
89	B2096-14-2	57.8	1.157	0.648	1.79	1.61
51	B2096-14-3	253.2	5.063	2.569	1.97	1.35
41	B2096-14-4	80.6	1.611	1.018	1.58	0.73
42	B2096-14-7	26.4	0.528	0.298	1.78	1.46
52	B2096-14-8	40	0.799	0.601	1.33	0.34
36	B2771-14-1A	42	0.84	0.454	1.9	1.85
37	B2771-14-1B	33.1	0.661	0.362	1.54	1.83
53	B2771-14-15A	182.6	3.652	2.032	1.8	0.67
33	B878-14-L3	44	0.881	0.52	1.7	1.71
27	B878-14-M1	75.1	1.503	0.992	1.52	0.68
25	B878-14-M2	28.6	0.572	0.334	1.72	1.94
24	B878-14-M3	22	0.439	0.229	1.92	0.98
8	B878-14-M4	214.1	4.282	2.354	2.33	1.82
13	B878-14-M5	113.8	2.276	1.245	2.32	1.83
87	B1931-15-6A	88	1.759	1.029	1.71	1.08
88	B1932-15-2	241.5	4.83	2.608	1.85	2.03
109	B2053-15-1	121.9	2.439	1.307	1.87	1
110	B2053-15-2	213.6	4.273	2.497	1.71	0.75
111	B2053-15-3	115.9	2.317	1.265	1.83	1.88
112	B2053-15-5	177.7	3.554	1.954	1.82	1.07
113	B2063-15-3	118.7	2.373	1.333	1.78	1.7
107	B2772-15-1	198	3.961	1.866	2.12	1.42
105	B2777-15A-7	296.4	5.928	3.373	1.76	0.72
106	B2777-15A-8	151.4	3.028	1.79	1.69	0.79

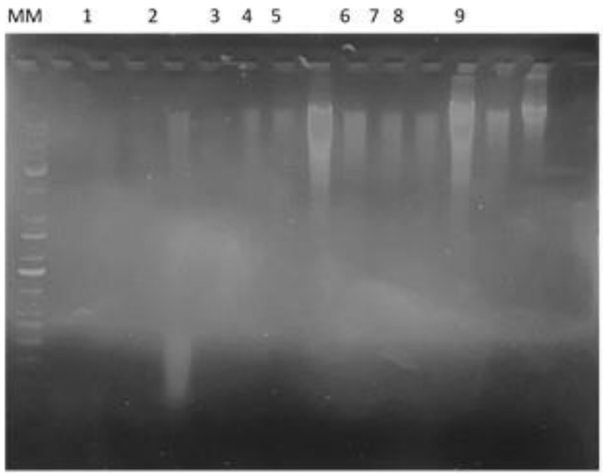
55	B293-15-10	139.2	2.783	1.597	1.74	1.24
60	B293-15-11	77.2	1.544	0.924	1.67	0.94
61	B293-15-12	173.5	3.47	1.796	1.93	1.09
90	B293-15-13	150.4	3.007	1.666	1.8	0.76
45	B293-15-14	86.5	1.729	0.972	1.78	0.69
43	B293-15-15	100.7	2.014	1.089	1.85	0.91
58	B293-15-16	123.2	2.463	1.524	1.62	0.55
62	B293-15-17	57.7	1.154	0.668	1.73	0.63
44	B293-15-18	127.5	2.55	1.425	1.79	0.74
63	B293-15-4	131.7	2.633	1.59	0.53	1.66
56	B293-15-6	239.3	4.786	2.581	1.85	1.66
57	B293-15-7	83.1	1.662	0.967	1.72	1.07
54	B293-15-8	93.8	1.876	0.997	0.73	1.88
59	B293-15-9	140.4	2.808	1.564	1.8	1.73
100	B3381-15-1	71	1.419	0.772	1.84	1.8
101	B3381-15-2	138.1	2.762	1.512	1.83	1.7
102	B3381-15-3	191.4	3.827	2.113	1.81	1.85
103	B3381-15-4	141	2.821	1.578	1.79	0.77
104	B3381-15-5	163.8	3.275	1.892	1.73	0.63
92	B3443-15-1	126.8	2.536	1.396	1.82	2.48
93	B3443-15-2	273.8	5.475	3.015	1.82	1.58
94	B3443-15-3	151.3	3.026	1.679	1.8	1.66
95	B3443-15-4	172.6	3.453	1.866	1.85	1.58
96	B3443-15-5	177.2	3.544	1.914	1.85	2.46
97	B3443-15-6	111	2.219	1.186	1.87	1.62
98	B3443-15-7	225.3	4.505	2.462	1.83	1.52
99	B3443-15-8	158.1	3.163	1.758	1.8	1.39
4	B359-15-2	54.4	1.089	0.561	1.52	1.94

2	B359-15-3	380.4	7.607	3.542	2.34	2.15
70	B359-15-4	143.4	2.867	1.525	1.88	0.98
64	B359-15-5	143.8	2.876	1.596	1.8	1.85
1	B359-15-6	156.1	3.122	1.542	1.9	2.05
3	B359-15-8	84.2	1.684	0.854	1.36	1.97
67	B457-15-3	90	1.8	1.127	1.6	0.55
69	B457-15-5	24	0.479	0.319	1.5	0.27
76	B458-15-1	61.8	1.236	0.663	1.86	1.08
73	B458-15-10	139.5	2.791	1.505	1.85	1.5
78	B458-15-11	83.6	1.672	0.902	1.85	0.73
68	B458-15-5	55.3	1.105	0.64	1.73	1.61
74	B458-15-5M	93.5	1.87	0.984	1.9	1.23
77	B458-15-6	52.9	1.059	0.61	1.74	0.69
75	B464-15-3	107.1	2.142	1.153	1.86	0.91
82	B540-15-2	107	2.141	1.177	1.82	1.52
83	B540-15-4	201.7	4.034	2.141	1.88	1.29
84	B540-15-5	130.4	2.608	1.446	1.8	0.91

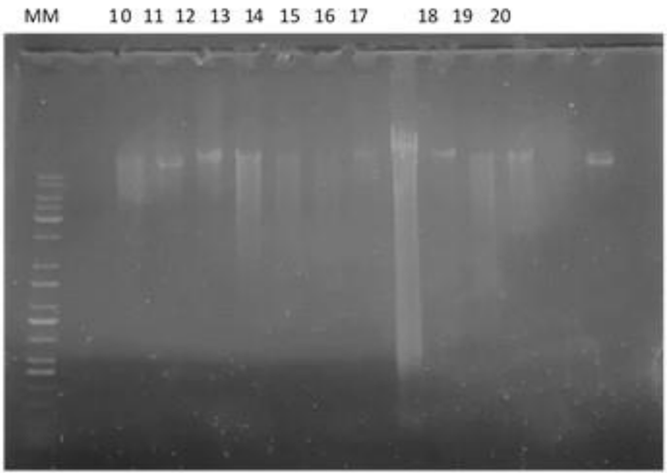
N/A – Not available – sequenced prior to start of this PhD project.
a – see gel photo’s Appendix A.2

A.2. Gel electrophoresis

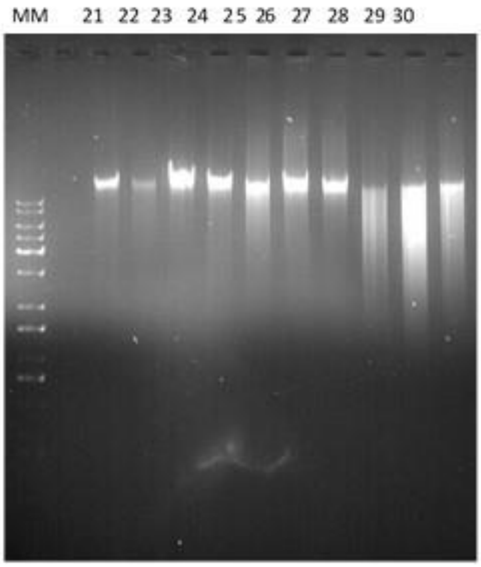
DNA agarose gel 1:



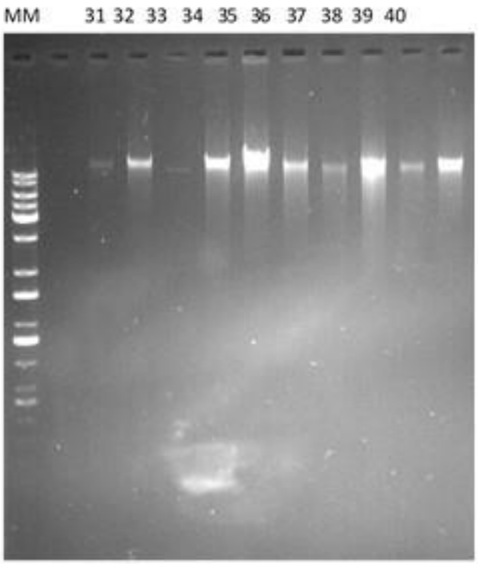
DNA agarose gel 2:



DNA agarose gel 3:

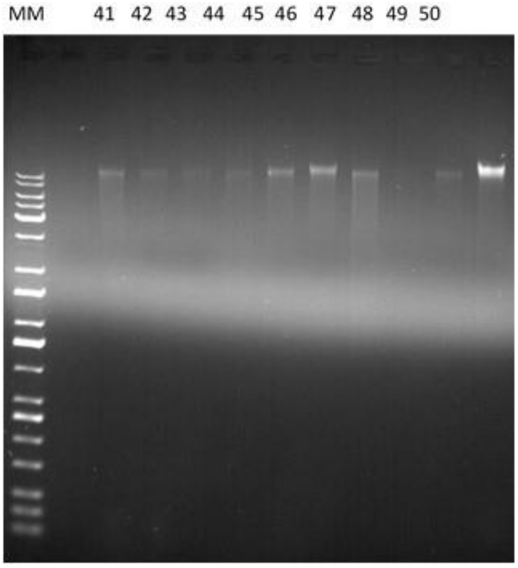


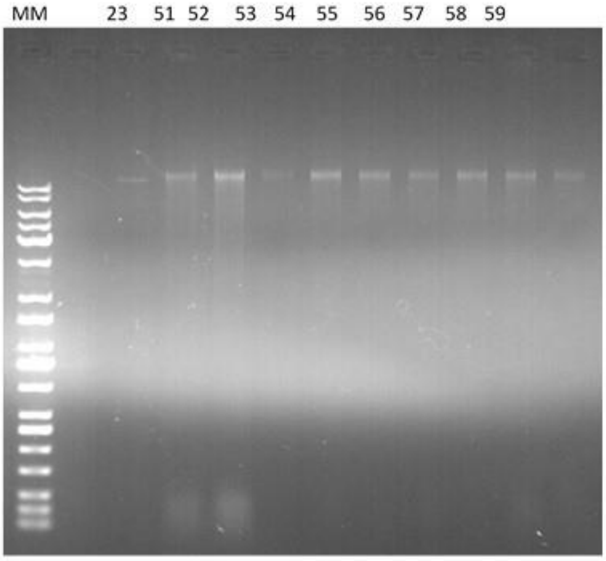
DNA agarose gel 4:



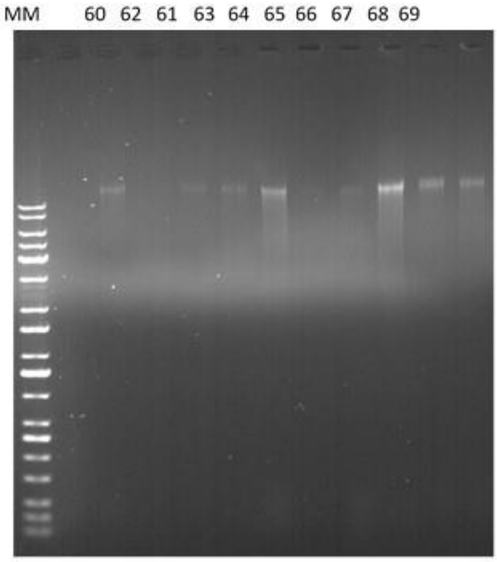
DNA agarose gel 5:

DNA agarose gel 6:

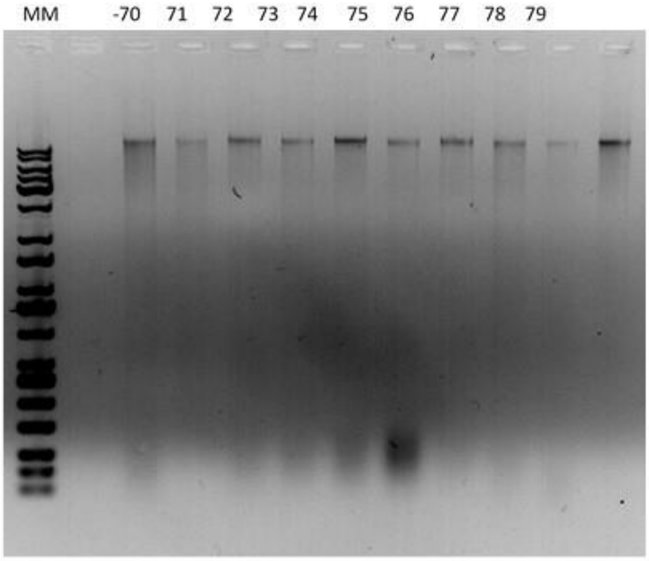




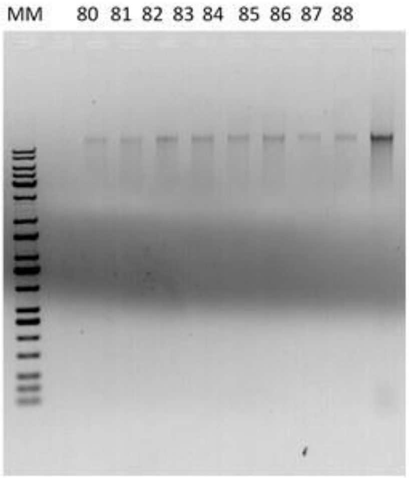
DNA agarose gel 7:



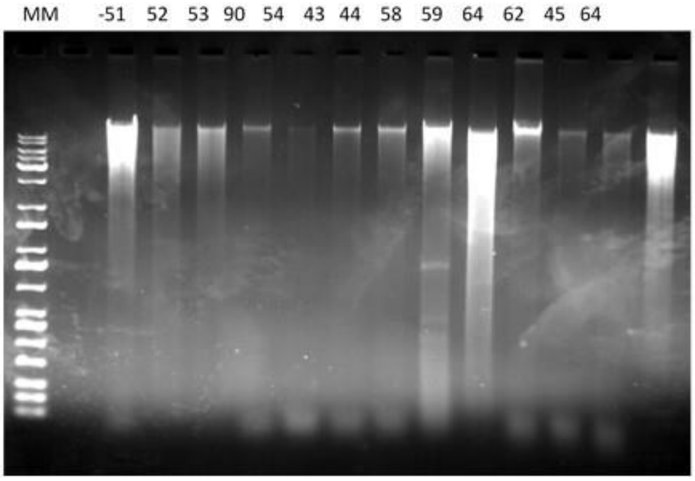
DNA agarose gel 8:



DNA agarose gel 9:

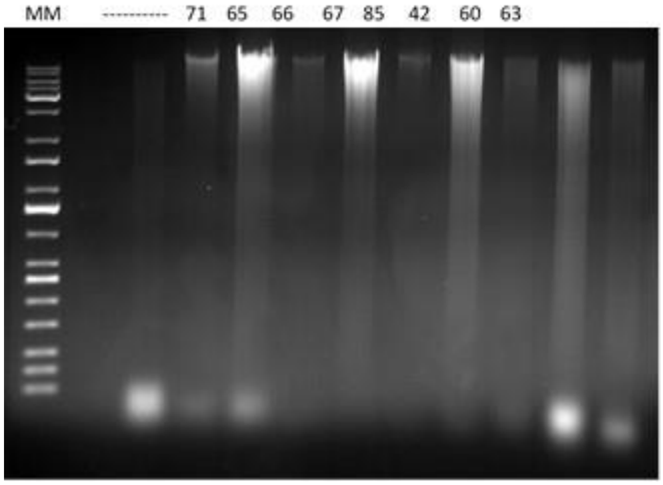


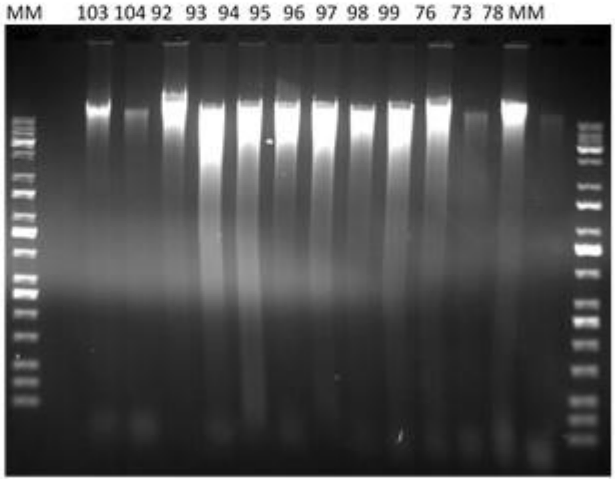
DNA agarose gel 10:



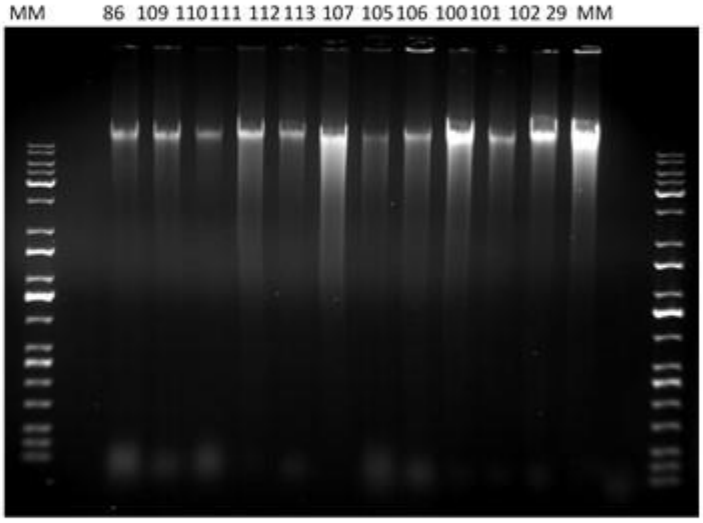
DNA agarose gel 11:

DNA agarose gel 12:

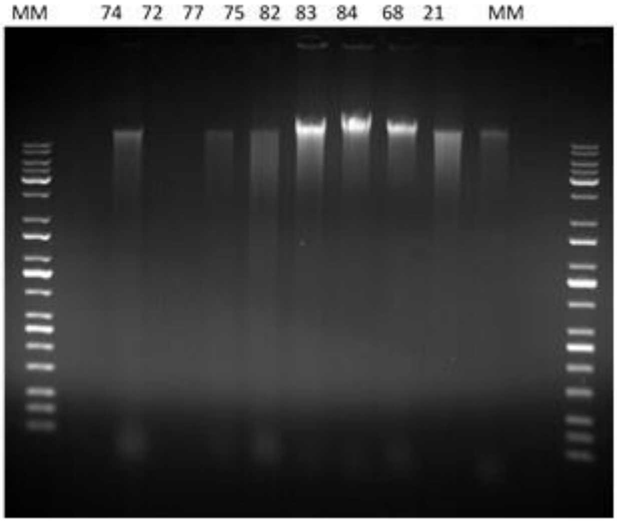




DNA agarose gel 13:

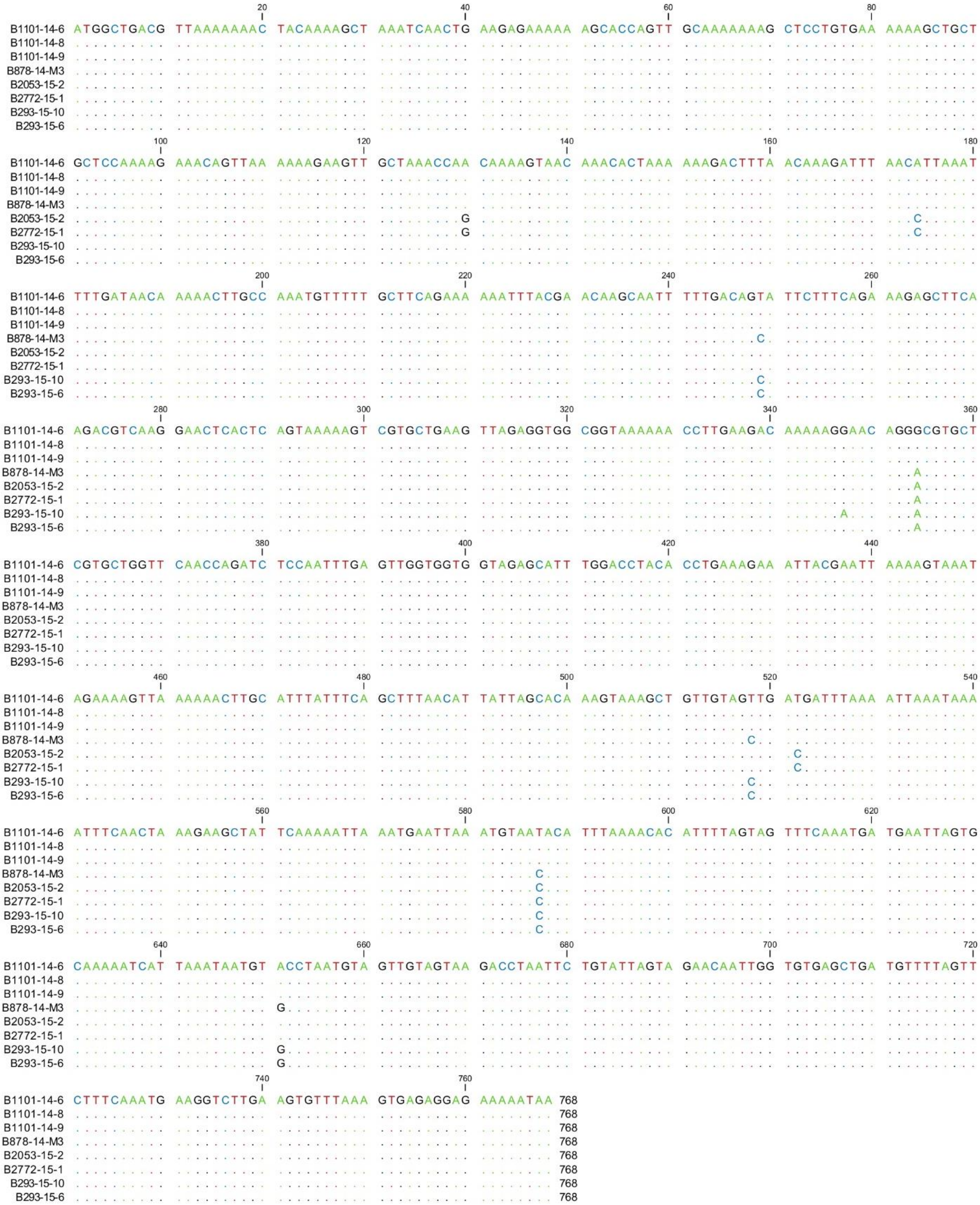


DNA agarose gel 14:



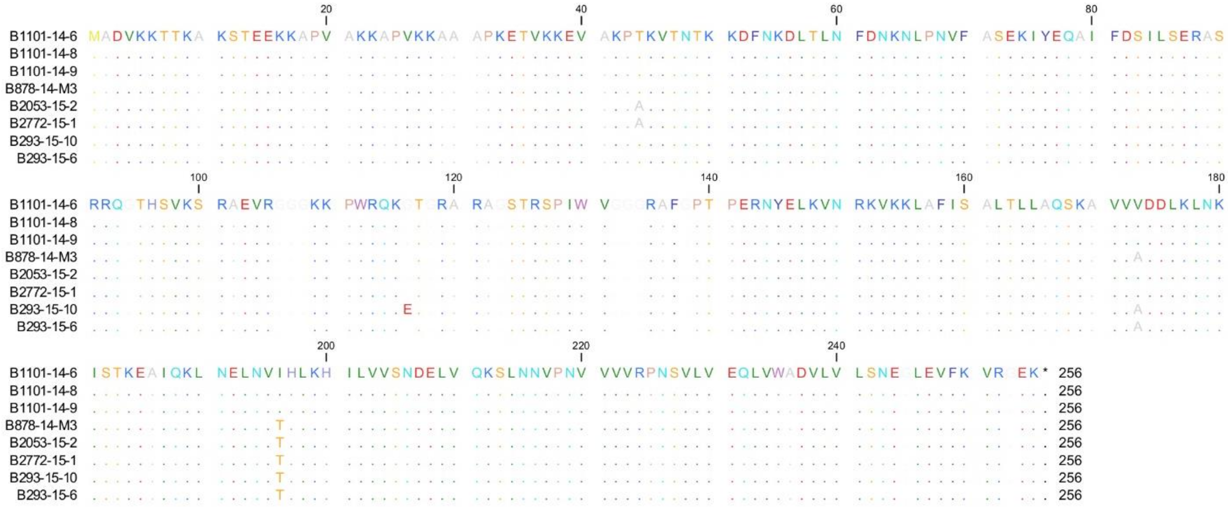
APPENDIX B: SEQUENCE ALIGNMENTS

B.1. *M. gallinarum vlpD* gene alignment



```
                              20                    40                    60                    80
B1101-14-6 ATGGCTGACG TTAAAAAAAC TACAAAAGCT AAATCAACTG AAGAGAAAAA AGCACCAGTT GCAAAAAAAG CTCCTGTGAA AAAAGCTGCT
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2053-15-2 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2772-15-1 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B293-15-10 .......... .......... .......... .......... .......... .......... .......... .......... ..........
 B293-15-6 .......... .......... .......... .......... .......... .......... .......... .......... ..........
                   100                   120                   140                   160                   180
B1101-14-6 GCTCCAAAAG AAACAGTTAA AAAAGAAGTT GCTAAACCAA CAAAAGTAAC AAACACTAAA AAAGACTTTA ACAAAGATTT AACATTAAAT
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2053-15-2 .......... .......... .......... .........G .......... .......... .......... .......... ...C......
B2772-15-1 .......... .......... .......... .........G .......... .......... .......... .......... ...C......
B293-15-10 .......... .......... .......... .......... .......... .......... .......... .......... ..........
 B293-15-6 .......... .......... .......... .......... .......... .......... .......... .......... ..........
                              200                   220                   240                   260
B1101-14-6 TTTGATAACA AAAACTTGCC AAATGTTTTT GCTTCAGAAA AAATTTACGA ACAAGCAATT TTTGACAGTA TTCTTTCAGA AAGAGCTTCA
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... .......... .......... ........C. .......... ..........
B2053-15-2 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2772-15-1 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B293-15-10 .......... .......... .......... .......... .......... .......... ........C. .......... ..........
 B293-15-6 .......... .......... .......... .......... .......... .......... ........C. .......... ..........
                   280                   300                   320                   340                   360
B1101-14-6 AGACGTCAAG GAACTCACTC AGTAAAAAGT CGTGCTGAAG TTAGAGGTGG CGGTAAAAAA CCTTGAAGAC AAAAAGGAAC AGGGCGTGCT
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... .......... .......... .......... .......... ...A......
B2053-15-2 .......... .......... .......... .......... .......... .......... .......... .......... ...A......
B2772-15-1 .......... .......... .......... .......... .......... .......... .......... .......... ...A......
B293-15-10 .......... .......... .......... .......... .......... .......... .......... ......A... ...A......
 B293-15-6 .......... .......... .......... .......... .......... .......... .......... .......... ...A......
                              380                   400                   420                   440
B1101-14-6 CGTGCTGGTT CAACCAGATC TCCAATTTGA GTTGGTGGTG GTAGAGCATT TGGACCTACA CCTGAAAGAA ATTACGAATT AAAAGTAAAT
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2053-15-2 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2772-15-1 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B293-15-10 .......... .......... .......... .......... .......... .......... .......... .......... ..........
 B293-15-6 .......... .......... .......... .......... .......... .......... .......... .......... ..........
                   460                   480                   500                   520                   540
B1101-14-6 AGAAAAGTTA AAAAACTTGC ATTTATTTCA GCTTTAACAT TATTAGCACA AAGTAAAGCT GTTGTAGTTG ATGATTTAAA ATTAAATAAA
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... .......... .......... .......C.. .......... ..........
B2053-15-2 .......... .......... .......... .......... .......... .......... .......... .C........ ..........
B2772-15-1 .......... .......... .......... .......... .......... .......... .......... .C........ ..........
B293-15-10 .......... .......... .......... .......... .......... .......... .......C.. .......... ..........
 B293-15-6 .......... .......... .......... .......... .......... .......... .......C.. .......... ..........
                              560                   580                   600                   620
B1101-14-6 ATTTCAACTA AAGAAGCTAT TCAAAAATTA AATGAATTAA ATGTAATACA TTTAAAACAC ATTTTAGTAG TTTCAAATGA TGAATTAGTG
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... .......... .......... ......C... .......... .......... .......... ..........
B2053-15-2 .......... .......... .......... .......... ......C... .......... .......... .......... ..........
B2772-15-1 .......... .......... .......... .......... ......C... .......... .......... .......... ..........
B293-15-10 .......... .......... .......... .......... ......C... .......... .......... .......... ..........
 B293-15-6 .......... .......... .......... .......... ......C... .......... .......... .......... ..........
                   640                   660                   680                   700                   720
B1101-14-6 CAAAAATCAT TAAATAATGT ACCTAATGTA GTTGTAGTAA GACCTAATTC TGTATTAGTA GAACAATTGG TGTGAGCTGA TGTTTTAGTT
B1101-14-8 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3 .......... .......... G......... .......... .......... .......... .......... .......... ..........
B2053-15-2 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2772-15-1 .......... .......... .......... .......... .......... .......... .......... .......... ..........
B293-15-10 .......... .......... G......... .......... .......... .......... .......... .......... ..........
 B293-15-6 .......... .......... G......... .......... .......... .......... .......... .......... ..........
                              740                   760
B1101-14-6 CTTTCAAATG AAGGTCTTGA AGTGTTTAAA GTGAGAGGAG AAAAATAA 768
B1101-14-8 .......... .......... .......... .......... ........ 768
B1101-14-9 .......... .......... .......... .......... ........ 768
B878-14-M3 .......... .......... .......... .......... ........ 768
B2053-15-2 .......... .......... .......... .......... ........ 768
B2772-15-1 .......... .......... .......... .......... ........ 768
B293-15-10 .......... .......... .......... .......... ........ 768
 B293-15-6 .......... .......... .......... .......... ........ 768
```

B.2. *M. gallinarum ribosomal protein L4 alignment*



```
                                   20                    40                    60                    80
                                    |                     |                     |                     |
B1101-14-6  MADVKKTTKA KSTEEKKAPV AKKAPVKKAA APKETVKKEV AKPTKVTNTK KDFNKDLTLN FDNKNLPNVF ASEKIYEQAI FDSILSERAS
B1101-14-8  .......... .......... .......... .......... .......... .......... .......... .......... ..........
B1101-14-9  .......... .......... .......... .......... .......... .......... .......... .......... ..........
B878-14-M3  .......... .......... .......... .......... .......... .......... .......... .......... ..........
B2053-15-2  .......... .......... .......... .......... ...A...... .......... .......... .......... ..........
B2772-15-1  .......... .......... .......... .......... ...A...... .......... .......... .......... ..........
B293-15-10  .......... .......... .......... .......... .......... .......... .......... .......... ..........
B293-15-6   .......... .......... .......... .......... .......... .......... .......... .......... ..........
                     100                   120                   140                   160                   180
                       |                     |                     |                     |                     |
B1101-14-6  RRQGTHSVKS RAEVRGGGKK PWRQKGTGRA RAGSTRSPIW VGGGRAFGPT PERNYELKVN RKVKKLAFIS ALTLLAQSKA VVVDDLKLNK
B1101-14-8  ... ...... .....   .. ..... . .. .. ....... .   ... .. .......... .......... .......... ..........
B1101-14-9  ... ...... .....   .. ..... . .. .. ....... .   ... .. .......... .......... .......... ..........
B878-14-M3  ... ...... .....   .. ..... . .. .. ....... .   ... .. .......... .......... .......... ..A.......
B2053-15-2  ... ...... .....   .. ..... . .. .. ....... .   ... .. .......... .......... .......... ..........
B2772-15-1  ... ...... .....   .. ..... . .. .. ....... .   ... .. .......... .......... .......... ..........
B293-15-10  ... ...... .....   .. .....E. .. .. ....... .   ... .. .......... .......... .......... ..A.......
B293-15-6   ... ...... .....   .. ..... . .. .. ....... .   ... .. .......... .......... .......... ..A.......
                                   200                   220                   240
                                     |                     |                     |
B1101-14-6  ISTKEAIQKL NELNVIHLKH ILVVSNDELV QKSLNNVPNV VVVRPNSVLV EQLVWADVLV LSNEGLEVFK VRGEK* 256
B1101-14-8  .......... .......... .......... .......... .......... .......... .... ..... .. ... 256
B1101-14-9  .......... .......... .......... .......... .......... .......... .... ..... .. ... 256
B878-14-M3  .......... .....T.... .......... .......... .......... .......... .... ..... .. ... 256
B2053-15-2  .......... .....T.... .......... .......... .......... .......... .... ..... .. ... 256
B2772-15-1  .......... .....T.... .......... .......... .......... .......... .... ..... .. ... 256
B293-15-10  .......... .....T.... .......... .......... .......... .......... .... ..... .. ... 256
B293-15-6   .......... .....T.... .......... .......... .......... .......... .... ..... .. ... 256
```

REFERENCES

Abolnik, C. & Beylefeld, A. 2015. Complete genome sequence of Mycoplasma gallinaceum. *Genome announcements,* 3.

Adeyemi, M. A., Bwala, D. G. & Abolnik, C. 2017. Comparative evaluation of the pathogenicity of Mycoplasma gallinaceum in chickens. *Avian Diseases*.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* 25**,** 3389-402.

Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schaffer, A. A. & Yu, Y. K. 2005. Protein database searches using compositionally adjusted substitution matrices. *Febs j,* 272**,** 5101-9.

Ammar, A. M., El-Aziz, N. K. A., Gharib, A. A., Ahmed, H. K. & Lameay, A. E. 2016. Mutations of domain V in 23S ribosomal RNA of macrolide-resistant *Mycoplasma gallisepticum* isolates in Egypt. *The Journal of Infection in Developing Countries,* 10**,** 807-813.

Amram, E., Mikula, I., Schnee, C., Ayling, R., Nicholas, R., Rosales, R., Harrus, S. & Lysnyansky, I. 2014. 16S rRNA gene mutations associated with decreased susceptibility to tetracycline in Mycoplasma bovis. *Antimicrobial agents and chemotherapy***,** AAC. 03876-14.

Andrews, S. 2010. *FastQC: a quality control tool for high throughput sequence data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc [Accessed].

Armour, N. K. & García, M. 2014. Current and future applications of viral-vectored recombinant vaccines in poultry. *The poultry informed professional. Department of Population Health, University of Georgia, Athens, GA***,** 1-9.

Arraes, F., Carvalho, M. J. A. d., Maranhão, A. Q., Brígido, M. M., Pedrosa, F. O. & Felipe, M. S. S. 2007. Differential metabolism of Mycoplasma species as revealed by their genomes. *Genetics and Molecular Biology,* 30**,** 182-189.

Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M. & Kubal, M. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC genomics,* 9**,** 75.

Baez-Ortega, A., Lorenzo-Diaz, F., Hernandez, M., Gonzalez-Vila, C. I., Roda-Garcia, J. L., Colebrook, M. & Flores, C. 2015. IonGAP: integrative bacterial genome analysis for Ion Torrent sequence data. *Bioinformatics,* 31**,** 2870-2873.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. & Prjibelski, A. D. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology,* 19**,** 455-477.

Baseman, J. B. & Tully, J. G. 1997. Mycoplasmas: sophisticated, reemerging, and burdened by their notoriety. *Emerging infectious diseases,* 3**,** 21.

Benčina, D., Dorrer, D. & Tadina, T. 1987. Mycoplasma species isolated from six avian species. *Avian Pathology,* 16**,** 653-664.

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L. & Trees, E. 2017. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., Hampson, D. J., Bellgard, M., Wassenaar, T. M. & Ussery, D. W. 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Functional & integrative genomics,* 6**,** 165-185.

Bleidorn, C. 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and biodiversity,* 14**,** 1-8.

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A. & Caporaso, J. G. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods,* 10**,** 57.

Botes, A., Peyrot, B., Olivier, A., Burger, W. & Bellstedt, D. 2005. Identification of three novel mycoplasma species from ostriches in South Africa. *Veterinary microbiology,* 111**,** 159-169.

Bradbury, J. 2005. Poultry mycoplasmas: sophisticated pathogens in simple guise. *British poultry science,* 46**,** 125-136.

Bradbury, J. M., Abdul-Wahab, O. M. S., Yavari, C. A., Dupiellet, J.-P. & Bové, J. M. 1993. Mycoplasma imitans sp. nov. is related to *Mycoplasma gallisepticum* and found in birds. *International Journal of Systematic and Evolutionary Microbiology,* 43**,** 721-728.

Bradbury, J. M. & Forrest, M. 1984. Mycoplasma cloacale, a new species isolated from a turkey. *International Journal of Systematic and Evolutionary Microbiology,* 34**,** 389-392.

Bradbury, J. M., Forrest, M. & Williams, A. 1983. Mycoplasma lipofaciens, a new species of avian origin. *International Journal of Systematic Bacteriology,* 33**,** 329-335.

Bradbury, J. M., Jordan, F., Shimizu, T., Stipkovits, L. & Varga, Z. 1988. Mycoplasma anseris sp. nov. found in geese. *International Journal of Systematic and Evolutionary Microbiology,* 38**,** 74-76.

Bradbury, J. M. & Mccarthy, J. D. 1983. Pathogenicity of Mycoplasma iowae for chick embryos. *Avian Pathology,* 12**,** 483-496.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B. & Pusch, G. D. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports,* 5**,** 8365.

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv preprint arXiv:1203.4802*.

Bwala, D. G. 2017. *Mycoplasma gallisepticum infection dynamics and vaccine protection in South African poultry.* Doctoral disseration, University of Pretoria, South Africa.

Bwala, D. G., Solomon, P., Duncan, N., Wandrag, D. B. & Abolnik, C. 2018. Assessment of *Mycoplasma gallisepticum* vaccine efficacy in a co-infection challenge model with QX-like infectious bronchitis virus. *Avian Pathology,* 47**,** 261-270.

Bwala, D. G., Solomon, P., Duncan, N., Wandrag, D.B.R., Abolnik, C In press. Assessment of *Mycoplasma gallisepticum* vaccine efficacy in a co-infection challenge model with QX-like infectious bronchitis virus.

Chen, I. M. A., Markowitz, V. M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Andersen, E., Huntemann, M., Varghese, N., Hadjithomas, M., Tennessen, K., Nielsen, T., Ivanova, N. N. & Kyrpides, N. C. 2017. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research,* 45**,** D507-D516.

Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. German conference on bioinformatics, 1999. Hanover, Germany, 45-56.

Citti, C. & Blanchard, A. 2013. Mycoplasmas and their host: emerging and re-emerging minimal pathogens. *Trends in microbiology,* 21**,** 196-203.

Citti, C., Nouvel, L.-X. & Baranowski, E. 2010. Phase and antigenic variation in mycoplasmas. *Future microbiology,* 5**,** 1073-1085.

Cizelj, I., Berčič, R. L., Dušanić, D., Narat, M., Kos, J., Dovč, P. & Benčina, D. 2011. *Mycoplasma gallisepticum* and *Mycoplasma synoviae* express a cysteine protease CysP, which can cleave chicken IgG into Fab and Fc. *Microbiology,* 157**,** 362-372.

Clyde Jr, W. A. 1983. Growth inhibition tests. *In:* RAZIN, S. A. T., J. G. (ed.) *Methods in Mycoplasmology.* Academic Press.

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G. & Fay, S. 2015. The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research,* 4.

Curtiss, R. 2002. Bacterial infectious disease control by vaccine development. *The Journal of Clinical Investigation,* 110**,** 1061-1066.

Daff 2017. Economic Review of the South African Agriculture - 2016/17. *In:* DEPARTMENT OF AGRICULTURE, F. A. F. (ed.). Republic of South Africa: Directorate: Statistics and Economic Analysis.

Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research,* 14**,** 1394-1403.

Darmon, E. & Leach, D. R. 2014. Bacterial genome instability. *Microbiol Mol Biol Rev,* 78**,** 1-39.

Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods,* 9**,** 772-772.

Del Angel, V. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., Amselem, J., Bouri, L., Bocs, S. & Klopp, C. 2018. Ten steps to get started in Genome Assembly and Annotation. *F1000Research,* 7.

Dordet-Frisoni, E., Sagné, E., Baranowski, E., Breton, M., Nouvel, L. X., Blanchard, A., Marenda, M. S., Tardy, F., Sirand-Pugnet, P. & Citti, C. 2014. Chromosomal transfers in mycoplasmas: when minimal genomes go mobile. *MBio,* 5**,** e01958-14.

Dybvig, K. & Voelker, L. L. 1996. Molecular biology of mycoplasmas. *Annual Reviews in Microbiology,* 50**,** 25-57.

Eagar, H., Swan, G. & Van Vuuren, M. 2012. A survey of antimicrobial usage in animals in South Africa with specific reference to food animals. *Journal of the South African Veterinary Association,* 83**,** 15-23.

Earl, D. A., Bradnam, K., John, J. S., Darling, A., Lin, D., Faas, J., Yu, H. O. K., Vince, B., Zerbino, D. R. & Diekhans, M. 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research***,** gr. 126599.111.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince & C., Knight, R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics,* 27, 2194-2200.

Edward, D. G. & Kanarek, A. 1960. Organisms of the pleuropneumonia group of avian origin: their classification into species. *Annals of the New York Academy of Sciences,* 79**,** 696-702.

Edwards, D. J. & Holt, K. E. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation,* 3**,** 2.

Ekblom, R. & Wolf, J. B. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications,* 7**,** 1026-1042.

Elin Videvall, A. P., and Alexey Gurevich. 2017. The N50 filtering problem. *The Molecular Ecologist* [Online]. [Accessed August 2018].

Emea. 2009. *The European Agency for the Evauluation of Medicinal Products, Committee for veterinary medicinal products- oxytetracycline, tetracycline, chlortetracycline summary report* [Online]. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Maximum_Residue_Limits_-_Report/2009/11/WC500015378.pdf [Accessed 2/10/2017].

Feberwee, A., Mekkes, D., De Wit, J., Hartman, E. & Pijpers, A. 2005. Comparison of culture, PCR, and different serologic tests for detection of *Mycoplasma gallisepticum* and *Mycoplasma synoviae* infections. *Avian diseases,* 49**,** 260-268.

Feng, Y., Zhang, Y., Ying, C., Wang, D. & Du, C. 2015. Nanopore-based fourth-generation DNA sequencing technology. *Genomics, proteomics & bioinformatics,* 13**,** 4-16.

Ferguson-Noel, N. 2013. Mycoplasmosis. *In:* SWAYNE, D. E., GLISSON, J. R., MCDOUGALD, L. R., NOLAN, L. K., SUAREZ, D. L. & NAIR, V. (eds.) *Diseases of poultry.* Ames,: Wiley-Blackwell.

Ferguson-Noel, N. & Noormohammadi, A. H. 2013. *Mycoplasma synoviae* Infection. *In:* SWAYNE, D. E., GLISSON, J. R., MCDOUGALD, L. R., NOLAN, L. K., SUAREZ, D. L. & NAIR, V. (eds.) *Diseases of poultry.* Ames,: Wiley-Blackwell.

Fisunov, G. Y., Alexeev, D. G., Bazaleev, N. A., Ladygina, V. G., Galyamina, M. A., Kondratov, I. G., Zhukova, N. A., Serebryakova, M. V., Demina, I. A. & Govorun, V. M. 2011. Core proteome of the minimal cell: comparative proteomics of three mollicute species. *PLoS One,* 6**,** e21964.

Forrest, M. & Bradbury, J. M. 1984. Mycoplasma glycophilum, a new species of avian origin. *Microbiology,* 130**,** 597-603.

Forsyth, M., Tully, J., Gorton, T., Hinckley, L., Frasca Jr, S., Van Kruiningen, H. & Geary, S. 1996. Mycoplasma sturni sp. nov., from the conjunctiva of a European starling (Sturnus vulgaris). *International Journal of Systematic and Evolutionary Microbiology,* 46**,** 716-719.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G. & Kelley, J. M. 1995. The minimal gene complement of Mycoplasma genitalium. *science,* 270**,** 397-404.

Freundt, E. 1955. The classification of the pleuropneumonia group of organisms (Borrelomycetales). *International Bulletin of Bacteriological Nomenclature and Taxonomy,* 5**,** 67-78.

Frey, M., Hanson, R. & Andrson, D. 1968. A medium for the isolation of avian mycoplasmas. *American journal of veterinary research,* 29**,** 2163.

Furuya, E. Y. & Lowy, F. D. 2006. Antimicrobial-resistant bacteria in the community setting. *Nature Reviews Microbiology,* 4**,** 36.

Ganapathy, K. & Bradbury, J. 1999. Pathogenicity of Mycoplasma imitans in mixed infection with infectious bronchitis virus in chickens. *Avian pathology,* 28**,** 229-237.

García, M., Ikuta, N., Levisohn, S. & Kleven, S. 2005. Evaluation and comparison of various PCR methods for detection of *Mycoplasma gallisepticum* infection in chickens. *Avian diseases,* 49**,** 125-132.

Gautier-Bouchardon, A., Reinhardt, A., Kobisch, M. & Kempf, I. 2002. In vitro development of resistance to enrofloxacin, erythromycin, tylosin, tiamulin and oxytetracycline in *Mycoplasma gallisepticum*, Mycoplasma iowae and *Mycoplasma synoviae*. *Veterinary microbiology,* 88**,** 47-58.

Gerchman, I., Levisohn, S., Mikula, I., Manso-Silván, L. & Lysnyansky, I. 2011. Characterization of in vivo-acquired resistance to macrolides of *Mycoplasma gallisepticum* strains isolated from poultry. *Veterinary research,* 42**,** 90.

Gharaibeh, S. & Al-Rashdan, M. 2011. Change in antimicrobial susceptibility of *Mycoplasma gallisepticum* field isolates. *Veterinary Microbiology,* 150**,** 379-383.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources,* 11**,** 759-769.

Goldman, D. & Domschke, K. 2014. Making sense of deep sequencing. *International Journal of Neuropsychopharmacology,* 17**,** 1717-1725.

Goodwin, S., Mcpherson, J. D. & Mccombie, W. R. 2016a. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17**,** 333.

Goodwin, S., Mcpherson, J. D. & Mccombie, W. R. 2016b. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17**,** 333.

Guardabassi, L. & Courvalin, P. 2006. Modes of antimicrobial action and mechanisms of bacterial resistance. *In:* AARESTRUP, F. M. (ed.) *Antimicrobial resistance in bacteria of animal origin.* American Society of Microbiology.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology,* 59**,** 307-321.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-1075.

Hannan, P., Windsor, G., De Jong, A., Schmeer, N. & Stegemann, M. 1997. Comparative susceptibilities of various animal-pathogenic mycoplasmas to fluoroquinolones. *Antimicrobial agents and chemotherapy,* 41**,** 2037-2040.

Hannan, P. C. 2000. Guidelines and recommendations for antimicrobial minimum inhibitory concentration (MIC) testing against veterinary mycoplasma species. *Veterinary research,* 31**,** 373-395.

Harasawa, R., Pitcher, D. G., Ramírez, A. S. & Bradbury, J. M. 2004. A putative transposase gene in the 16S–23S rRNA intergenic spacer region of Mycoplasma imitans. *Microbiology,* 150**,** 1023-1029.

Haridas, S., Breuill, C., Bohlmann, J. & Hsiang, T. 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. *Journal of microbiological methods,* 86**,** 368-375.

Heather, J. M. & Chain, B. 2016. The sequence of sequencers: the history of sequencing DNA. *Genomics,* 107**,** 1-8.

Hong, Y., García, M., Leiting, V., Benčina, D., Dufour-Zavala, L., Zavala, G. & Kleven, S. H. 2004. Specific detection and typing of *Mycoplasma synoviae* strains in poultry with PCR and DNA sequence analysis targeting the hemagglutinin encoding gene *vlhA*. *Avian Diseases,* 48**,** 606-616.

Huntemann, M., Ivanova, N. N., Mavromatis, K., Tripp, H. J., Paez-Espino, D., Palaniappan, K., Szeto, E., Pillay, M., Chen, I. M. A., Pati, A., Nielsen, T., Markowitz, V. M. & Kyrpides, N. C. 2015. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Standards in Genomic Sciences,* 10**,** 86.

Indikova, I., Vronka, M. & Szostak, M. P. 2014. First identification of proteins involved in motility of *Mycoplasma gallisepticum*. *Veterinary research,* 45**,** 99.

Jenkins, C., Geary, S. J., Gladd, M. & Djordjevic, S. P. 2007. The *Mycoplasma gallisepticum* OsmC-like protein MG1142 resides on the cell surface and binds heparin. *Microbiology,* 153**,** 1455-1463.

Jenkins, C., Samudrala, R., Geary, S. J. & Djordjevic, S. P. 2008. Structural and functional characterization of an organic hydroperoxide resistance protein from *Mycoplasma gallisepticum*. *Journal Of Bacteriology,* 190**,** 2206-2216.

Johansson, K.-E., Heldtander, M. U. & Pettersson, B. 1998. Characterization of mycoplasmas by PCR and sequence analysis with universal 16S rDNA primers. *Mycoplasma protocols***,** 145-165.

Jordan, F., Ernø, H., Cottew, G., Hinz, K. & Stipkovits, L. 1982. Characterization and taxonomic description of five mycoplasma serovars (serotypes) of avian origin and their elevation to species rank and further evaluation of the taxonomic status of *Mycoplasma synoviae*. *International Journal of Systematic and Evolutionary Microbiology,* 32**,** 108-115.

Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research,* 30**,** 3059-3066.

Kempf, I. 1998. DNA amplification methods for diagnosis and epidemiological investigations of avian mycoplasmosis. *Avian Pathology,* 27**,** 7-14.

King, D., Kleven, S., Wenger, D., & Anderson, D. 1973. Field studies with *Mycoplasma synoviae*. *Avian Diseases.* 17(4), 722-726.

Kleven, S. 1998. Mycoplasmas in the etiology of multifactorial respiratory disease. *Poultry science,* 77**,** 1146-1149.

Kleven, S. 2008. Control of avian mycoplasma infections in commercial poultry. *Avian diseases,* 52**,** 367-374.

Kleven, S., Eidson, C. & Fletcher, O. 1978. Airsacculitis induced in broilers with a combination of Mycoplasma gallinarum and respiratory viruses. *Avian diseases***,** 707-716.

Lagesen, K., Hallin, P., Rødland, E., Stærfeldt, H. & Ussery, D. 2007. RT: RNammer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res,* 35**,** 3100-3108.

Langer, S. 2009. *Proposal for molecular tools for the epidemiology of contagious bovine pleuro pneumonia and classification of unknown mycplasma sp. isolated from struthio camelus.* uniwien.

Levisohn, S. & Kleven, S. 2000. Avian mycoplasmosis (*Mycoplasma gallisepticum*). *Revue Scientifique et Technique-Office International des Epizooties,* 19**,** 425-434.

Li, B.-B., Shen, J.-Z., Cao, X.-Y., Wang, Y., Dai, L., Huang, S.-Y. & Wu, C.-M. 2010. Mutations in 23S rRNA gene associated with decreased susceptibility to tiamulin and valnemulin in *Mycoplasma gallisepticum*. *FEMS microbiology letters,* 308**,** 144-149.

Li, Y. & Chen, X. 2012. Sialic acid metabolism and sialyltransferases: natural functions and applications. *Applied microbiology and biotechnology,* 94**,** 887-905.

Lierz, M., Stark, R., Brokat, S. & Hafez, H. M. 2007. Pathogenicity of Mycoplasma lipofaciens strain ML64, isolated from an egg of a Northern Goshawk (Accipiter gentilis), for chicken embryos. *Avian Pathol,* 36**,** 151-3.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. 2012. Comparison of next-generation sequencing systems. *BioMed Research International,* 2012.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R. & Pallen, M. J. 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology,* 10**,** 599.

Lysnyansky, I. & Ayling, R. D. 2016. Mycoplasma bovis: mechanisms of resistance and trends in antimicrobial susceptibility. *Frontiers in microbiology,* 7**,** 595.

Lysnyansky, I., Gerchman, I., Flaminio, B. & Catania, S. 2015. Decreased susceptibility to macrolide–lincosamide in *Mycoplasma synoviae* is associated with mutations in 23S ribosomal RNA. *Microbial Drug Resistance,* 21**,** 581-589.

Lysnyansky, I., Gerchman, I., Mikula, I., Gobbo, F., Catania, S. & Levisohn, S. 2013. Molecular characterization of acquired enrofloxacin resistance in *Mycoplasma synoviae* field isolates. *Antimicrobial agents and chemotherapy,* 57**,** 3072-3077.

Magiorakos, A. P., Srinivasan, A., Carey, R., Carmeli, Y., Falagas, M., Giske, C., Harbarth, S., Hindler, J., Kahlmeter, G. & Olsson-Liljequist, B. 2012. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clinical microbiology and infection,* 18**,** 268-281.

Markham, P. F., Duffy, M. F., Glew, M. D. & Browning, G. F. 1999. A gene family in Mycoplasma imitans closely related to the pMGA family of *Mycoplasma gallisepticum*. *Microbiology,* 145**,** 2095-2103.

Matros L, W. T. a. t. M. T., IVS Sacramento. 2001. *Microbiology Guide to Interpreting MIC (Minimum Inhibitory Concentration)* [Online]. Available: http://the-vet.net/DVMWiz/Vetlibrary/Lab-%20Microbiology%20Guide%20to%20Interpreting%20MIC.htm [Accessed].

Matyushkina, D., Pobeguts, O., Butenko, I., Vanyushkina, A., Anikanov, N., Bukato, O., Evsyutina, D., Bogomazova, A., Lagarkova, M. & Semashko, T. 2016. Phase transition of the bacterium upon invasion of a host cell as a mechanism of adaptation: a *Mycoplasma gallisepticum* model. *Scientific reports,* 6**,** 35959.

May, M. A., Kutish, G. F., Barbet, A. F., Michaels, D. L. & Brown, D. R. 2015. Complete Genome Sequence of *Mycoplasma synoviae* Strain WVU 1853T. *Genome Announc,* 3.

Metzker, M. L. 2010. Sequencing technologies—the next generation. *Nature reviews genetics,* 11**,** 31.

Moalic, P.-Y., Kempf, I., Gesbert, F. & Laigret, F. 1997. Identification of two pathogenic avian mycoplasmas as strains of Mycoplasma pullorum. *International Journal of Systematic and Evolutionary Microbiology,* 47**,** 171-174.

Moretti, S. A., Boucher, C. E. & Bragg, R. R. 2013. Molecular characterisation of *Mycoplasma gallisepticum* genotypes from chickens in Zimbabwe and South Africa. *South African Journal of Science,* 109**,** 1-4.

Munita, J. M. & Arias, C. A. 2016. Mechanisms of Antibiotic Resistance. *Microbiology spectrum,* 4**,** 10.1128/microbiolspec.VMBF-0016-2015.

Nhung, N. T., Chansiripornchai, N. & Carrique-Mas, J. J. 2017. Antimicrobial Resistance in Bacterial Poultry Pathogens: A Review. *Frontiers in veterinary science,* 4**,** 126.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A. & Sirotkin, Y. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology,* 20**,** 714-737.

OIE. 2008. *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2017* [Online]. Available: http://www.oie.int/en/animal-health-in-the-world/information-on-aquatic-and-terrestrial-animal-diseases/ [Accessed Access 2008].

Olson, N., Kerr, K. & Campbell, A. 1964. Control of infectious synovitis 13. The antigen study of three strains. *Avian Diseases,* 8**,** 209-214.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., De Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., Mchardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O. & Vonstein, V. 2005. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research,* 33**,** 5691-5702.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B. & Shukla, M. 2013. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research,* 42**,** D206-D214.

Pakpinyo, S. & Sasipreeyajan, J. 2007. Molecular characterization and determination of antimicrobial resistance of *Mycoplasma gallisepticum* isolated from chickens. *Veterinary microbiology,* 125**,** 59-65.

Panangala, V. S., Stringfellow, J. S., Dybvig, K., Woodard, A., Sun, F., Rose, D. L. & Gresham, M. M. 1993. Mycoplasma corogypsi sp. nov., a new species from the footpad abscess of a black vulture, Coragyps atratus. *International Journal of Systematic and Evolutionary Microbiology,* 43**,** 585-590.

Papazisi, L., Frasca Jr, S., Gladd, M., Liao, X., Yogev, D. & Geary, S. 2002. GapA and CrmA coexpression is essential for *Mycoplasma gallisepticum* cytadherence and virulence. *Infection and immunity,* 70**,** 6839-6845.

Papazisi, L., Gorton, T. S., Kutish, G., Markham, P. F., Browning, G. F., Swartzell, S., Madan, A., Mahairas, G. & Geary, S. J. 2003. The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain Rlow. *Microbiology,* 149**,** 2307-2316.

Petkau, A., Stuart-Edwards, M., Stothard, P. & Van Domselaar, G. 2010. Interactive microbial genome visualization with GView. *Bioinformatics,* 26**,** 3125-3126.

Pitcher, D. & Nicholas, R. 2005. Mycoplasma host specificity: fact or fiction? *The Veterinary Journal,* 170**,** 300-306.

Pop, M. 2009. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics,* 10**,** 354-366.

Poveda, J., Giebel, J., Flossdorf, J., Meier, J. & Kirchhoff, H. 1994. Mycoplasma buteonis sp. nov., Mycoplasma falconis sp. nov., and Mycoplasma gypis sp. nov., Three Species from Birds of Prey. *International Journal of Systematic and Evolutionary Microbiology,* 44**,** 94-98.

Pritchard, R. E., Prassinos, A. J., Osborne, J. D., Raviv, Z. & Balish, M. F. 2014. Reduction of hydrogen peroxide accumulation and toxicity by a catalase from Mycoplasma iowae. *PLoS One,* 9**,** e105188.

Qiagen, A. 2016. *CLC genomics Workbench 8.5.1 User Manual*.

Raviv, Z. & Ley, D. H. 2013. *Mycoplasma gallisepticum* Infection. *In:* SWAYNE, D. E., GLISSON, J. R., MCDOUGALD, L. R., NOLAN, L. K., SUAREZ, D. L. & NAIR, V. (eds.) *Diseases of poultry.* Ames,: Wiley-Blackwell.

Razin, S. 1994. DNA probes and PCR in diagnosis of mycoplasma infections. *Molecular and cellular probes,* 8**,** 497-511.

Razin, S. 2012. *Methods in Mycoplasmology V1: Mycoplasma Characterization*, Elsevier Science.

Razin, S. & Hayflick, L. 2010. Highlights of mycoplasma research—an historical perspective. *Biologicals,* 38**,** 183-190.

Razin, S., Yogev, D. & Naot, Y. 1998. Molecular biology and pathogenicity of mycoplasmas. *Microbiology and Molecular Biology Reviews,* 62**,** 1094-1156.

Reinhardt, A., Bébéar, C., Kobisch, M., Kempf, I. & Gautier-Bouchardon, A. 2002a. Characterization of mutations in DNA gyrase and topoisomerase IV involved in quinolone resistance of *Mycoplasma gallisepticum* mutants obtained in vitro. *Antimicrobial agents and chemotherapy,* 46**,** 590-593.

Reinhardt, A., Kempf, I., Kobisch, M. & Gautier-Bouchardon, A. 2002b. Fluoroquinolone resistance in *Mycoplasma gallisepticum*: DNA gyrase as primary target of enrofloxacin and impact of mutations in topoisomerases on resistance level. *Journal of antimicrobial chemotherapy,* 50**,** 589-592.

Roberts, D. 1964. The isolation of an influenza A virus and a mycoplasma associated with duck sinusitis. *Vet. Rec,* 76**,** 470-473.

Roberts, M., Koutsky, L., Holmes, K., Leblanc, D. & Kenny, G. 1985. Tetracycline-resistant Mycoplasma hominis strains contain streptococcal tetM sequences. *Antimicrobial agents and chemotherapy,* 28**,** 141-143.

Rocha, E. P. C. & Blanchard, A. 2002. Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. *Nucleic acids research,* 30**,** 2031-2042.

Rocha, T. S., Bertolotti, L., Catania, S., Pourquier, P. & Rosati, S. 2016. Genome Sequence of a Mycoplasma meleagridis Field Strain. *Genome Announc,* 4.

Ronquist, F. & Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19**,** 1572-1574.

Ruhnke, H. & Rosendal, S. 1989. *Useful protocols for diagnosis of animal mycoplasmas,* Ames, Iowa State University Press.

Sapa. 2016. *South African poultry Association 2016 Industry Profile* [Online]. Available: https://www.sapoultry.co.za/pdf-docs/sapa-industry-profile.pdf [Accessed Access 2016].

Schadt, E. E., Turner, S. & Kasarskis, A. 2010. A window into third-generation sequencing. *Human molecular genetics,* 19**,** R227-R240.

Schatz, M. C., Delcher, A. L. & Salzberg, S. L. 2010. Assembly of large genomes using second-generation sequencing. *Genome research***,** gr. 101360.109.

Schmieder, R. & Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27**,** 863-864.

Schnee, C., Schulsse, S., Hotzel, H., Ayling, R. D., Nicholas, R. A., Schubert, E., Heller, M., Ehricht, R. & Sachse, K. 2012. A novel rapid DNA microarray assay enables identification of 37 Mycoplasma species and highlights multiple Mycoplasma infections. *PloS one,* 7**,** e33237.

Sentíes-Cué, G. Shivaprasad, H. L. & chin, R. P. 2005. Systemic *Mycoplasma synoviae* infection in broiler chickens. *Avian Pathology.* 34(2), 137-142.

Shimizu, T., Erno, H. & Nagatono, J. 1978. Isolation and characterization of Mycoplasma columbinum and Mycoplasma columborale, two new species from pigeons. *International Journal of Systematic and Evolutionary Microbiology,* 28**,** 538-546.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics,* 15**,** 121.

Smith, D. R. 2015. Buying in to bioinformatics: an introduction to commercial sequence analysis software. *Briefings in Bioinformatics,* 16**,** 700-709.

Sprygin, A., Andreychuk, D., Kolotilov, A., Volkov, M., Runina, I., Mudrak, N., Borisov, A., Irza, V., Drygin, V. & Perevozchikova, N. 2010. Development of a duplex real-time TaqMan PCR assay with an internal control for the detection of *Mycoplasma gallisepticum* and *Mycoplasma synoviae* in clinical samples from commercial and backyard poultry. *Avian pathology,* 39**,** 99-109.

Stein, L. 2001. Genome annotation: from sequence to biology. *Nature reviews genetics,* 2**,** 493.

Stipkovits, L. & Kempf, I. 1996. Mycoplasmoses in poultry. *Revue scientifique et technique (International Office of Epizootics),* 15**,** 1495-1525.

Swofford, D. L. 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0 b10.

Szczepanek, S. M., Tulman, E. R., Gorton, T. S., Liao, X., Lu, Z., Zinski, J., Aziz, F., Frasca, S., Jr., Kutish, G. F. & Geary, S. J. 2010. Comparative genomic analyses of attenuated strains of *Mycoplasma gallisepticum*. *Infect Immun,* 78**,** 1760-71.

Tatusova, T., Dicuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M. & Ostell, J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research***,** gkw569.

Thermo Scientific. 2010. T042--Technical Bulletin NanoDrop Spectrophotometers. 260/280 and 260/230 ratios.

Touchman, J. 2010. Comparative genomics. *Nature Education Knowledge,* 3**,** 13.

Tulman, E. R., Liao, X., Szczepanek, S. M., Ley, D. H., Kutish, G. F. & Geary, S. J. 2012. Extensive variation in surface lipoprotein gene content and genomic changes associated with virulence during evolution of a novel North American house finch epizootic strain of *Mycoplasma gallisepticum*. *Microbiology,* 158**,** 2073-2088.

Umar, S., Munir, M., Ur-Rehman, Z., Subhan, S., Azam, T. & Shah, M. 2017. Mycoplasmosis in poultry: update on diagnosis and preventive measures. *World's Poultry Science Journal,* 73**,** 17-28.

Utturkar, S. M., Klingeman, D. M., Bruno-Barcena, J. M., Chinn, M. S., Grunden, A. M., Köpke, M. & Brown, S. D. 2015. Sequence data for Clostridium autoethanogenum using three generations of sequencing technologies. *Scientific Data,* 2**,** 150014.

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. 2014. Ten years of next-generation sequencing technology. *Trends in genetics,* 30**,** 418-426.

Vasconcelos, A. T. R., Ferreira, H. B., Bizarro, C. V., Bonatto, S. L., Carvalho, M. O., Pinto, P. M., Almeida, D. F., Almeida, L. G., Almeida, R. & Alves-Filho, L. 2005. Swine and poultry pathogens:

the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of *Mycoplasma synoviae*. *Journal of bacteriology,* 187**,** 5568-5577.

Vesth, T., Lagesen, K., Acar, Ö. & Ussery, D. 2013. CMG-biotools, a free workbench for basic comparative microbial genomics. *PloS one,* 8**,** e60120.

Wai, S. N., Nakayama, K., Umene, K., Moriya, T. & Amako, K. 1996. Construction of a ferritin-deficient mutant of Campylobacter jejuni: contribution of ferritin to iron storage and protection against oxidative stress. *Molecular microbiology,* 20**,** 1127-1134.

Wajid, B. & Serpedin, E. 2014. Do it yourself guide to genome assembly. *Briefings in functional genomics,* 15**,** 1-9.

Wakenell, P., Damassa, A. & Yamamoto, R. 1995. In ovo pathogenicity of Mycoplasma iners strain Oz. *Avian diseases***,** 390-397.

Wang, Y., Yi, L., Zhang, F., Qiu, X., Tan, L., Yu, S., Cheng, X. & Ding, C. 2017. Identification of genes involved in *Mycoplasma gallisepticum* biofilm formation using mini-Tn4001-SGM transposon mutagenesis. *Veterinary microbiology,* 198**,** 17-22.

Wei, S., Guo, Z., Li, T., Zhang, T., Li, X., Zhou, Z., Li, Z., Liu, M., Luo, R., Bi, D., Chen, H., Zhou, R. & Jin, H. 2012. Genome sequence of Mycoplasma iowae strain 695, an unusual pathogen causing deaths in turkeys. *J Bacteriol,* 194**,** 547-8.

Weisburg, W., Tully, J., Rose, D., Petzel, J., Oyaizu, H., Yang, D., Mandelco, L., Sechrest, J., Lawrence, T. & Van Etten, J. 1989. A phylogenetic analysis of the mycoplasmas: basis for their classification. *Journal of bacteriology,* 171**,** 6455-6467.

Welchman, D. B., Bradbury, J., Cavanagh, D. & Aebischer, N. 2002. Infectious agents associated with respiratory disease in pheasants. *The Veterinary Record,* 150**,** 658-664.

Whithear, K. 1993. Avian mycoplasmosis. *Australian standard diagnostic techniques for animal diseases***,** 1-12.

Wright, G. D. 2010. Q&A: Antibiotic resistance: where does it come from and what can we do about it? *BMC biology,* 8**,** 123.

Wu, C.-M., Wu, H., Ning, Y., Wang, J., Du, X. & Shen, J. 2005. Induction of macrolide resistance in *Mycoplasma gallisepticum* in vitro and its resistance-related mutations within domain V of 23S rRNA. *FEMS Microbiology Letters,* 247**,** 199-205.

Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. 2011. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics,* 12**,** 444.

Xia, X., Wu, C., Cui, Y., Kang, M., Li, X., Ding, S. & Shen, J. 2015. Proteomic analysis of tylosin-resistant *Mycoplasma gallisepticum* reveals enzymatic activities associated with resistance. *Scientific reports,* 5**,** 17077.

Yacoub, E., Sirand-Pugnet, P., Barré, A., Blanchard, A., Hubert, C., Maurier, F., Bouilhol, E. & Ben Abdelmoumen Mardassi, B. 2016. Genome Sequences of Two Tunisian Field Strains of Avian Mycoplasma, M. meleagridis and M. gallinarum. *Genome Announcements,* 4.

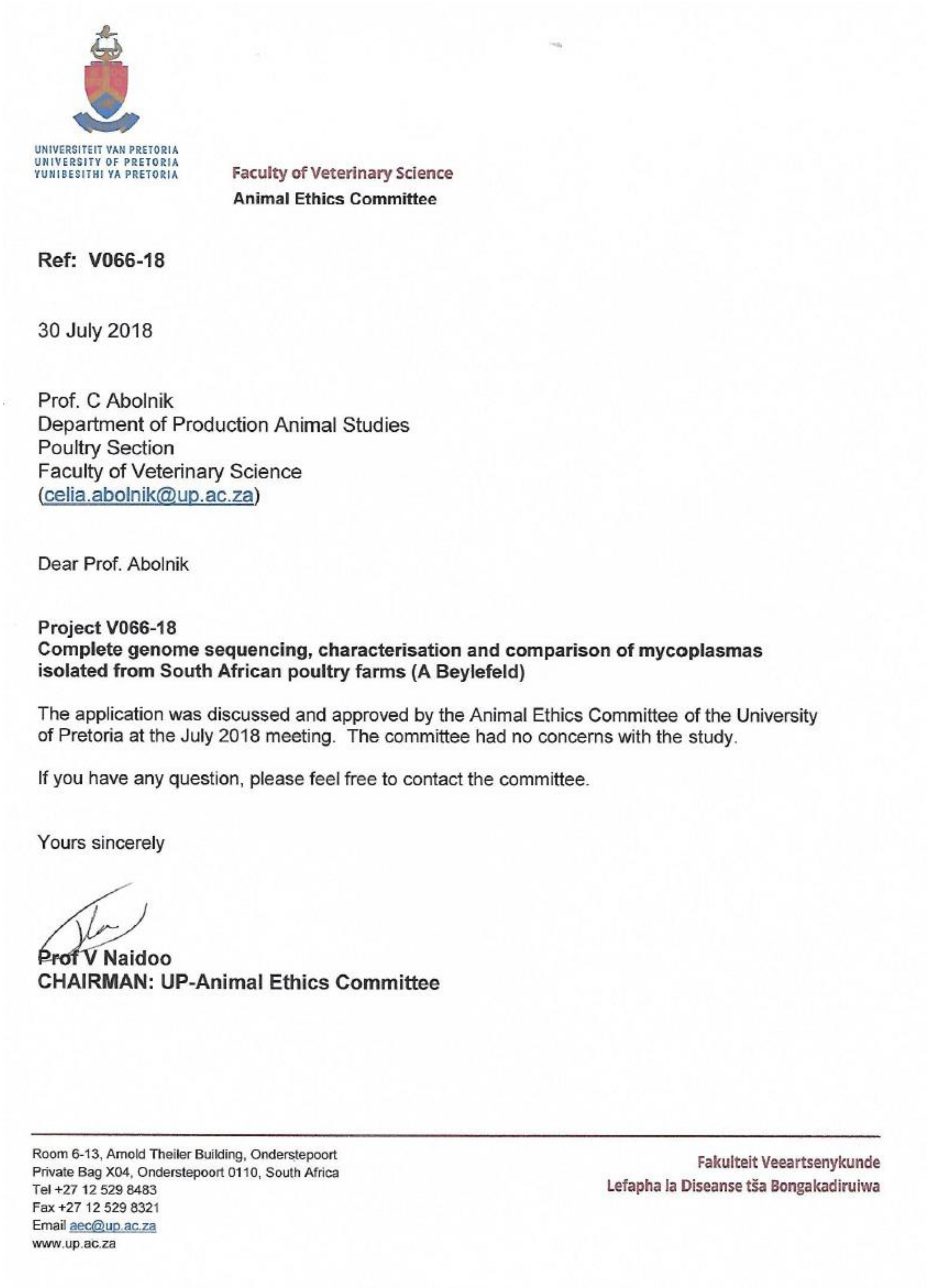
Yamamoto, R., Bigland, C. & Ortmayer, H. 1965. Characteristics of Mycoplasma meleagridis sp. n., isolated from turkeys. *Journal of bacteriology,* 90**,** 47-49.

Zanella, A., Martino, P., Pratelli, A. & Stonfer, M. 1998. Development of antibiotic resistance in *Mycoplasma gallisepticum* in vitro. *Avian Pathology,* 27**,** 591-596.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. & Shen, B. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one,* 6**,** e17915.

Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol,* 7**,** 203-14.

ANIMAL ETHICS COMMITTEE



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Veterinary Science
Animal Ethics Committee

Ref: V066-18

30 July 2018

Prof. C Abolnik
Department of Production Animal Studies
Poultry Section
Faculty of Veterinary Science
(celia.abolnik@up.ac.za)

Dear Prof. Abolnik

Project V066-18
Complete genome sequencing, characterisation and comparison of mycoplasmas isolated from South African poultry farms (A Beylefeld)

The application was discussed and approved by the Animal Ethics Committee of the University of Pretoria at the July 2018 meeting. The committee had no concerns with the study.

If you have any question, please feel free to contact the committee.

Yours sincerely

Prof V Naidoo
CHAIRMAN: UP-Animal Ethics Committee

Room 6-13, Arnold Theiler Building, Onderstepoort
Private Bag X04, Onderstepoort 0110, South Africa
Tel +27 12 529 8483
Fax +27 12 529 8321
Email aec@up.ac.za
www.up.ac.za

Fakulteit Veeartsenykunde
Lefapha la Diseanse tša Bongakadiruiwa