University of Nottingham
UK | CHINA | MALAYSIA

Implementation of Machine Learning for the Evaluation of Mastitis and Antimicrobial Resistance in Dairy Cows

Necati Esener, BEng, MSc

A thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

2021

Abstract

Bovine mastitis is one of the biggest concerns in the dairy industry, where it affects sustainable milk production, farm economy and animal health. Most of the mastitis pathogens are bacterial in origin and accurate diagnosis of them enables understanding the epidemiology, outbreak prevention and rapid cure of the disease. This thesis aimed to provide a diagnostic solution that couples Matrix-Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) mass spectroscopy coupled with machine learning (ML), for detecting bovine mastitis pathogens at the subspecies level based on their phenotypic characters.

In Chapter 3, MALDI-TOF coupled with ML was performed to discriminate bovine mastitis-causing *Streptococcus uberis* based on transmission routes; contagious and environmental. *S. uberis* isolates collected from dairy farms across England and Wales were compared within and between farms. The findings of this chapter suggested that the proposed methodology has the potential of successful classification at the farm level.

In Chapter 4, MALDI-TOF coupled with ML was performed to show proteomic differences between bovine mastitis-causing *Escherichia coli* isolates with different clinical outcomes (clinical and subclinical) and disease phenotype (persistent and non-persistent). The findings of this chapter showed that phenotypic differences can be detected by the proposed methodology even for genotypically identical isolates.

In Chapter 5, MALDI-TOF coupled with ML was performed to differentiate benzylpenicillin signatures of bovine mastitis-causing *Staphylococcus aureus* isolates. The findings of this chapter presented that the proposed methodology enables fast, affordable and effective diagnostic solution for targeting resistant bacteria in dairy cows.

Having shown this methodology successfully worked for differentiating benzylpenicillin resistant and susceptible *S. aureus* isolates in Chapter 5, the same technique was applied to other mastitis agents *Enterococcus faecalis* and *Enterococcus faecium* and for profiling other antimicrobials besides benzylpenicillin in Chapter 6. The findings of this chapter demonstrated that MALDI-TOF coupled with ML allows monitoring the disease epidemiology and provides suggestions for adjusting farm management strategies.

Taken together, this thesis highlights that MALDI-TOF coupled with ML is capable of discriminating bovine mastitis pathogens at subspecies level based on transmission route, clinical

outcome and antimicrobial resistance profile, which could be used as a diagnostic tool for bovine mastitis at dairy farms.

Declaration

By this means, I declare that this PhD thesis was conducted following the requirements of the University of Nottingham Regulation and Code of Practice for Research Degree Programmes, and has not been, or will be submitted for any other academic award. The work presented hereby is my own and all work of other authors and material from other sources is properly recognized.

Necati Esener

Acknowledgements

First of all, I would like to thank my supervisors, who made this long journey easier and much more enjoyable; Dr Tania Dottorini for her supervision and enthusiasm, Professor Richard Emes for his guidance and motivation, Professor Martin Green for his invaluable advice and encouragement, and Professor Andrew Bradley for his technical support and recommendations. I have been very lucky to have been mentored by great scholars who are pioneers in their fields.

Secondly, special thanks go to Andrew Warry for his help with bioinformatics and R scripts, Alexandre Maciel Guerra for his help with machine learning, MATLAB and Python scripts, and Aouatif Belkhiri for her help with wet-lab experiments. I am also grateful to Dr Peers Davies, Dr Kat Giebel, all CELE tutors, especially David Bowen, and ADAC members Abril Izquierdo, Dan Lea, Niraj Shah and Adam Blanchard.

I would like to acknowledge the Republic of Turkey Ministry of National Education, and the Ministry of Agriculture and Forestry for providing me with the funding and enabling this opportunity.

I cannot forget the brilliant lunch group members at Vet School: Mary, Ramon, Veronika, Gaurav, Mohammed, Deborah, Meshach, Tosin, Apa, Sophie, Adriano, Chris, Lamyaa, Purba, Lola, Ramzi and many others who participated. I will never forget their genuine friendship and the funny moments we had together.

Lastly but by no means least, I would like to thank all my family members for their encouragement and love. Their belief in me and enormous support meant a lot. Especially important thanks to my partner Gosia, whom I owe a lot for all her effortless support and sacrifices she has made for my benefit during this journey.

Table of Contents

List of Tables

List of Figures

List of Supplementary Figures

List of Abbreviations

ACT	Artemis Comparison Tool
ADASYN	Adaptive Synthetic Sampling Approach
AMR	Antimicrobial Resistance
ANN	Artificial Neural Network
BP	Biological Pathway
BRIG	Blast Ring Image Generator
CC	Cellular Component
CCI	Composite Correlation Indices
CDD	Conserved Domain Database
CFU	Colony Forming Unit
CLSI	Clinical and Laboratory Standards Institute
CV	Cross-Validation
Da	Dalton
DT	Decision Tree
ExPASy	Expert Protein Analysis System
GA	Genetic Algorithm
GO	Gene Ontology
IMI	Intramammary Infection
I-TASSER	Iterative Threading Assembly Refinement
kDa	Kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
K-NN	K-Nearest Neighbour
LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSVM	Linear Support Vector Machine
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization-Time of Flight
MCL	Markov Cluster Algorithm
MF	Molecular Function
MIC	Minimum Inhibitory Concentration
ML	Machine Learning

MLP NN	Multilayer Perceptron Neural Network
MLST	Multilocus Sequence Typing
MRSA	Methicillin-resistant *S. aureus*
MS	Mass Spectrometry
Mw	Molecular Weight
NB	Naïve Bayes
NCV	Nested Cross-Validation
NN	Neural Network
PDB	Protein Data Bank
PFAM	Protein Families Database
PPI	Protein-Protein Interaction
QC	QuickClassifier
QDA	Quadratic Discriminant Analysis
QMMS	Quality Milk Management Services Ltd.
RBF SVM	Radial Basis Function Support Vector Machine
RC	Recognition Capability
ReLU	Rectified Linear Unit
RF	Random Forest
SCC	Somatic Cell Count
SMOTE	Synthetic Minority Oversampling Technique
SNN	Supervised Neural Network
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
TMP/SMX	Trimethoprim/Sulfamethoxazole
VARRS	Veterinary Antimicrobial Resistance and Sales Surveillance
VIDA	Veterinary Investigation Diagnosis Analysis
VISA	Vancomycin-intermediate *S. aureus*
VRE	Vancomycin-resistant enterococci
VSE	Vancomycin-susceptible enterococci
VSSA	Vancomycin-susceptible *S. aureus*
wgSNP	Whole genome single nucleotide polymorphism

CHAPTER 1 INTRODUCTION

1.1 Bovine Mastitis

Bovine mastitis is the inflammation of one or more mammary quarters (Zadoks *et al.*, 2011). In the literature, the term intramammary infection (IMI) is often used instead of bovine mastitis; however, they are not the same according to definitions by the International Dairy Federation. IMI corresponds to the presence of the infection and should be used to talk about bovine mastitis agents rather than the inflammation itself; for instance, IMI cannot be classified as clinical and subclinical but mastitis can (Berry and Meaney, 2006). 137 different agents including bacteria, yeast and algae have been known to cause bovine mastitis, which makes it different from many other diseases (Watts, 1988). However, 75% of the bovine mastitis cases in the UK are caused by bacterial pathogens according to VIDA (Veterinary Investigation Diagnosis Analysis) annual reports between 2012 and 2019 (Surveillance Intelligence Unit, 2020). Hence, the current work has only focused on bacterial agents. To analyse and better understand the disease, mastitis pathogens have been historically categorized based on their transmission routes as contagious and environmental (Blowey and Edmondson, 2010). Contagious mastitis pathogens include *Staphylococcus aureus*, *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, *Mycoplasma* spp., *Corynebacterium bovis* etc. (Fox and Gay, 1993). The main reservoir of the contagious pathogens is the udder of the cow, and they are transmitted between animals mainly during the milking process. Meanwhile, environmental pathogens are acquired from the surrounding habitat of the cows and mainly transmitted during the period between milking sessions (see Figure 1-1). Environmental mastitis pathogens involve *Streptococcus uberis*, *Escherichia coli*, *Klebsiella* spp., *Enterococcus* spp., *Enterobacter* spp., *Serratia* spp., *Pseudomonas* spp. etc. (Smith and Hogan, 1993). However, mastitis pathogens do not always follow the same transmission route. Recently, it has been found that *S. agalactiae*, a longstanding contagious mastitis pathogen, showed environmental characteristics, and *S. uberis*, a longstanding environmental mastitis pathogen, showed contagious transmission (Jørgensen *et al.*, 2016; Davies *et al.*, 2016). There are also other mastitis pathogens such as *Staphylococcus chromogenes*, *Staphylococcus epidermis*, *Staphylococcus similans* etc., whose transmission route has not been elucidated yet.

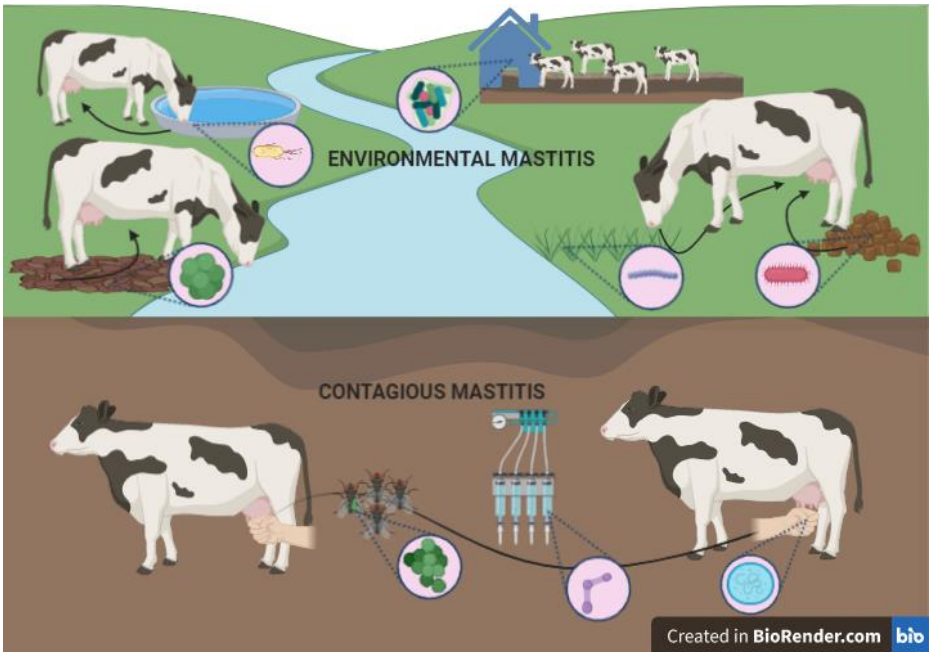


Figure 1-1. Classification of mastitis pathogens due to transmission route as environmental and contagious. *The main reservoir of the environmental mastitis can be anything that surrounds the living habitat of the cows including water, pasture, bedding material, calving pads, manure etc., on the other hand, contagious mastitis is transmitted between cows mostly during the milking through milking equipment or milkers' hand. However, flies are also the vector for carrying the disease between cows. This figure was generated in BioRender.com.*

Another classification of bovine mastitis agents is as major and minor pathogens, where the former cause more severe inflammation and higher somatic cell count (SCC) and are mainly associated with clinical mastitis while the latter cause mild inflammation and lower SCC and are mainly associated with the incidence of subclinical mastitis (Reyher *et al.*, 2012). Moreover, owing to bacteriocins, which are antimicrobial peptides synthesized by bacteria to compete against other bacteria (Nascimento *et al.*, 2005), some beneficial effects of minor pathogens over major pathogens have been observed (Reyher *et al.*, 2012). Minor pathogens include Coagulase-negative staphylococci (CNS), non-aureus staphylococci, *C. bovis* etc.; however, minor pathogen characteristic of CNS has recently been challenged (Pyörälä and Taponen, 2009). Major mastitis pathogens are *E. coli*, *S. aureus*, *S. uberis*, *S. agalactiae*, *S. dysgalactiae* etc., which are responsible for about 80% of the bovine mastitis cases in the UK (Bradley, 2002). In the following sections of this study, detailed information will be provided only for the major mastitis agents *E. coli*, *S. aureus* and *S. uberis;* and for the most commonly isolated enterococci from bedding materials *E. faecalis and E. faecium* (Gagnon *et al.*, 2020).

1.1.1 *Streptococcus uberis*

Streptococcus uberis is a gram-positive, catalase-negative coccus that appears in chains (Oliver, Pighetti and Almeida, 2011). *S. uberis* has been traditionally accepted as an environmental mastitis pathogen; however, contagious transmission outbreaks have also been shown in British and Dutch dairies (Davies *et al.*, 2016; Zadoks *et al.*, 2003). Dairy cow skin, bovine faeces, bedding material and farm environment are major sources of *S. uberis* IMIs (Unnerstad *et al.*, 2009). The incidence of *S. uberis* appears to be higher in tie-stall barns compared to free-stall barns (Riekerink *et al.*, 2008). Moreover, pasture-based systems have also long suffered from *S. uberis* IMI (Lopez-Benavides *et al.*, 2007; Olde Riekerink, Barkema and Stryhn, 2007; Shum *et al.*, 2009; McDougall, 2003). Interestingly, the presence of *S. uberis* in the environment has been shown to positively correlate with the presence of cows which was concluded to result from the shedding of the bacterium in bovine faeces (Zadoks, Tikofsky and Boor, 2005). Prevalence of *S. uberis* contamination is observed to vary between seasons and geographic locations, high during winter but low in summer in New Zealand and Germany (Lopez-Benavides *et al.*, 2007; Tenhagen *et al.*, 2009) but the opposite in Norway and the US (Østerås, Sølverød and Reksen, 2006; Todhunter, Smith and Hogan, 1995; Zadoks, Tikofsky and Boor, 2005).

Although *S. uberis* IMI can occur in any cow from lactating to dry cows or from heifers to multiparous cows (Oliver, Pighetti and Almeida, 2011), it has been reported as more prevalent in older cows, before calving, during lactation and prior to drying off (Jayarao *et al.*, 1999; Tenhagen *et al.*, 2006; Phuektes *et al.*, 2001; Zadoks *et al.*, 2001). However, another study found no significant difference in *S. uberis*-mastitis cases according to the age of the animal (Petrovski *et al.*, 2009). Some of the *S. uberis* IMIs that were acquired during the dry period were seen to develop mastitis in the next lactation (Krömker *et al.*, 2014). The risk of *S. uberis* IMI has also been shown to be higher in those quarters that had already experienced it (Zadoks *et al.*, 2001).

The trend of *S. uberis* IMIs from clinical and subclinical bovine mastitis diagnosed cows in the UK between the years of 2012 and 2019 is shown in Figure 1-2, based on the data from VIDA annual reports (Surveillance Intelligence Unit, 2020). It is seen to fluctuate through these years with an overall average of 19.90% and 20.78% in the latest report (year 2019). It should be noted that VIDA annual reports do not state the prevalence in the UK as submissions were made voluntarily. However, in a comprehensive survey of English and Welsh dairy farms in

2004-2005, *S. uberis* was isolated in 23.5% and 13.8% of clinical and subclinical mastitis cases, respectively (Bradley *et al.*, 2007). Moreover, the prevalence of *S. uberis* was found to be 22.1% and 9.0% in France (Botrel *et al.*, 2010) and Finland (Vakkamäki *et al.*, 2017), respectively, whereas the incidence of *S. uberis* derived clinical mastitis was 18.2% in Belgium (Verbeke *et al.*, 2014).

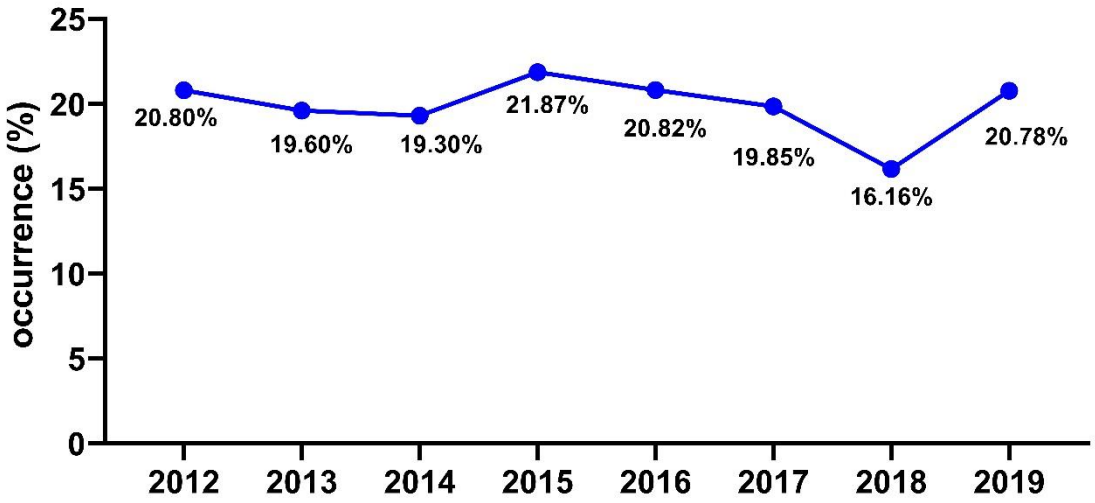


Figure 1-2. The trend of Streptococcus uberis IMIs from clinical and subclinical bovine mastitis diagnosed cows in the UK between the years of 2012 and 2019. *On average, S. uberis was isolated from 19.90% of the dairy cows diagnosed with bovine mastitis in these years. The graph was generated based on the data from VIDA (Veterinary Investigation Diagnosis Analysis) annual reports between 2012 and 2019 (Surveillance Intelligence Unit, 2020), which do not indicate prevalence or incidence. This figure was generated in GraphPad Prism v8.*

Several clonal complexes, each comprised of a central sequence type (ST) and any closely related STs with a few single-locus variants around it, are shown to be related to the clinical and subclinical mastitis outcome (Tomita *et al.*, 2008). Strain-specific pathogenicity of *S. uberis* has been observed across dairy cows; and pathogen factors (i.e. strain type, virulence, antimicrobial-resistance etc.) were concluded to be more important than host factors such as breed, parity, teat anatomy etc. (Tassi *et al.*, 2013). However, differences in terms of gene content between *S. uberis* strains could not be associated with its virulence or clinical outcome (Hossain *et al.*, 2015). *S. uberis* 0140J strain is a well-annotated bovine mastitis model organism (Ward *et al.*, 2009). Genomic differences between 0140J strain and additional twelve *S. uberis* strains isolated from dairy cows diagnosed with either clinical or subclinical bovine mastitis in the UK were compared (Hossain *et al.*, 2015). It was found that there was no obvious gene gain/loss between the strains that cause clinical or subclinical bovine mastitis. Moreover,

the EF20 strain, which was a previously isolated clinical case, was actually shown to be non-virulent. It was concluded that rather than pathogen factors (i.e. virulence) alone, interaction with host factors (i.e. host genetic and immune status) influence the clinical status of bovine mastitis (Hossain *et al.*, 2015), which has recently been shown in vivo experiments as well (Archer *et al.*, 2020).

When the *S. uberis* 0140J strain was first annotated the following proteins were suggested as virulence factors: fibronectin-binding protein, hemolysin-like protein, C5a peptidase precursor, sortase A, lactoferrin binding protein, collagen-like surface-anchored protein, *S. uberis* adhesion molecule (SUAM) and plasminogen activator (PauA) (Ward *et al.*, 2009). However, the genes encoding these proteins were found to be present in the non-virulent EF20 strain as well. Enriched pathway comparison between virulent and non-virulent strains showed that; F0F1-type ATP synthase, fructose and mannose inducible PTS, bacterial checkpoint-control related cluster and phage replication were some of the subsystems (Hossain *et al.*, 2015).

Although the virulence factors were present in both clinical and subclinical strains, genes encoding lactoferrin binding protein and collagen-like surface-anchored protein, whose negative mutant strains were previously shown to lose infection ability (Leigh *et al.*, 2010), were highly variable in terms of DNA sequence alignment. Hence, variation in certain genes may play a role in the infection potential of these strains (Hossain *et al.*, 2015). The hyaluronic acid capsule was another factor that was considered to be related to the infection ability of *S. uberis* 0140J in cattle (Ward *et al.*, 2009). However, other strains missing the hyaluronic acid capsule have already been shown to cause bovine mastitis (Field *et al.*, 2003). Another important difference between virulent 0140J and non-virulent EF-20 strains were bacteriocins which were not present in the latter. This was concluded as a disadvantage for the non-virulent EF20 strain in competing with other bovine mastitis-causing strains.

Several attempts have been made for the prevention of *S. uberis* IMI by killed and live bacterial vaccines; however, these studies have not provided successful immunization against this mastitis pathogen (Finch *et al.*, 1997; Finch *et al.*, 1994). Moreover, researchers have offered several virulence factors that could be potential vaccine targets. Mice vaccinated with fructose-biphosphate aldolase (FBA) and elongation factor Ts (EFTs), which are present both in the cytoplasm and cell wall of the organism, showed significant immunological response against bovine mastitis-causing *S. uberis* (Collado *et al.*, 2016). SUAM and PauA proteins have been suggested as vaccine antigens against *S. uberis* IMIs as their encoding genes were highly

prevalent and conversed. SUAM plays a role in bacterial adhesion, internalization and persistence of the organism in mammary (Almeida *et al.*, 2006) and vaccine studies gave remarkable results *in vitro* conditions (Prado *et al.*, 2011; Almeida *et al.*, 2015). In another study with mice, a subunit vaccine containing SUAM was shown to induce a humoral immune response against *S. uberis* (Perrig *et al.*, 2017). The role of PauA in *S. uberis* has been shown to be associated with colonization in the mammary gland (Ward *et al.*, 2003). Vaccination with PauA protected the cows against *S. uberis* IMIs, which carry this gene; however mastitis-causing *S. uberis* strains with no PauA or modified SUAM encoding genes were also observed (Tassi *et al.*, 2015; Gilchrist *et al.*, 2013; Perrig *et al.*, 2015).

1.1.2 *Escherichia coli*

Escherichia coli is a gram-negative, rod-shaped and facultative anaerobic coliform bacterium (Tenaillon *et al.*, 2010). *E. coli* is accepted as one of the major environmental mastitis pathogens (Smith and Hogan, 1993). *E. coli* is omnipresent in the farm environment and the main reservoirs include dairy manure, bedding material, soil, used pasture etc. (Klaas and Zadoks, 2018). Mastitis control plans such as the Five Point Plan (more details are given in section 1.2) in the UK have focused on contagious mastitis pathogens and have been successful in decreasing clinical cases caused by them (Bradley, 2002). However, this control plan did not affect the environmental transmission route and environmental pathogens have been commonly discovered in well-managed farms (Hogan *et al.*, 1989).

E. coli, like other environmental pathogens, frequently causes clinical mastitis rather than subclinical mastitis. The trend of *E. coli* IMIs from clinical and subclinical bovine mastitis diagnosed cows in the UK between the years of 2012 and 2019 is shown in Figure 1-3, based on data from VIDA annual reports (Surveillance Intelligence Unit, 2020). It is seen to have fluctuated through these years with an overall average of 21.66% and 26.27% in the latest report (year 2019). In an earlier comprehensive survey of English and Welsh dairy farms, *E. coli* was isolated in 19.8% and 3.0% of clinical and subclinical mastitis cases, respectively (Bradley *et al.*, 2007). In Canada, 8.4% of the clinical mastitis cases were found to be *E. coli* originated (Riekerink *et al.*, 2008). US studies also showed *E. coli* as one of the most frequently isolated mastitis pathogen, although coliform mastitis is less common in pasture-based dairy systems like in New Zealand (Oliveira, Hulland and Ruegg, 2013; Compton *et al.*, 2007). Regional differences regarding the prevalence of *E. coli* mastitis have also been observed in a recent Chinese study (Yu *et al.*, 2020).

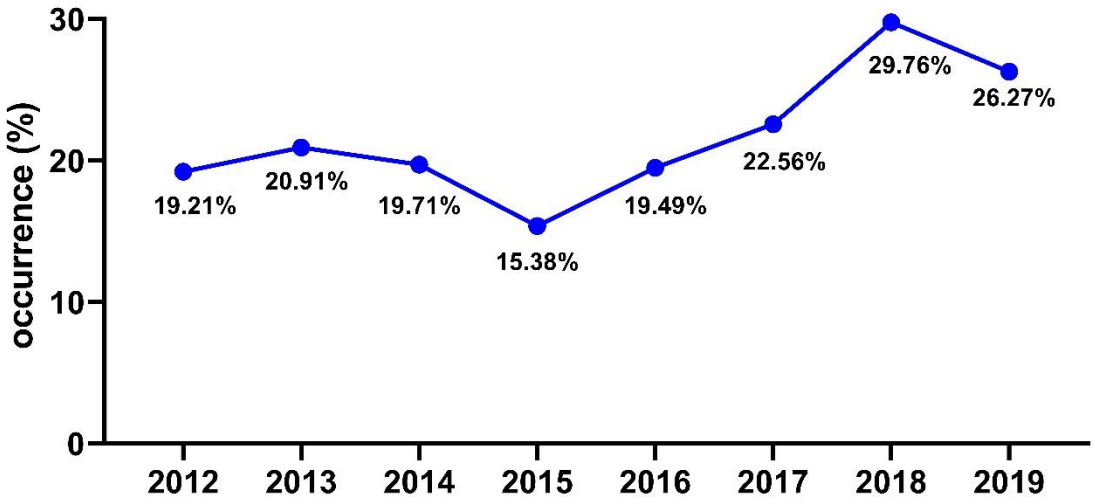


Figure 1-3. The trend of Escherichia coli IMIs from clinical and subclinical bovine mastitis diagnosed cows in the UK between the years of 2012 and 2019. *On average, E. coli was isolated from 21.66% of the dairy cows diagnosed with bovine mastitis in these years was. The graph was generated based on the data from VIDA (Veterinary Investigation Diagnosis Analysis) annual reports between 2012 and 2019 (Surveillance Intelligence Unit, 2020), which do not indicate prevalence or incidence. This figure was generated in GraphPad Prism v8.*

Acquisition of *E. coli* in dairy cows is greater at transition periods which are following drying off, just before and just after calving period due to concentration changes of the immune cells, the incomplete formation of keratin plug, milk cessation, physical changes in the mammary gland and decrease of antimicrobial level through dry cow therapy (Bradley and Green, 2004). Although new intramammary infections were acquired during these periods, clinical signs could be seen through the lactation period. *E. coli* has been widely accepted as a transient organism that is not mammary adapted (Fairbrother *et al.*, 2015). However, the persistence of *E. coli* pathogens for up to 100 days has been shown in other studies (Bradley and Green, 2000; Bradley and Green, 2001a).

E. coli is a predominant mastitis pathogen for well-managed dairy farms with low SCC (Bradley and Green, 2001a; Bradley and Green, 2000; Barkema *et al.*, 1998). Incidence of *E. coli* mastitis causes a significant reduction in milk quantity and yield, especially in high producing cows (Schukken *et al.*, 2012; Gröhn *et al.*, 2004). It has also been shown to result in an increased risk of culling especially in late occurring infections (Gröhn *et al.*, 2005). The outcome of *E. coli* mastitis can differ from mild to severe inflammation as a result of the lipopolysaccharides (LPS) present in the bacterial cell wall (Günther *et al.*, 2017). Some of the clinical *E. coli* cases can be so severe that cow welfare is hugely affected including swollen quarters,

high fever, dehydration, lack of appetite or even death of the animal (Burvenich *et al.*, 2003). The severity of the disease is decided mainly by host factors where the performance of neutrophils play an important role (Burvenich *et al.*, 2003). Most of the time host immune system can eliminate the infection (Ruegg, 2010); however, there are recorded recurrent and persistent mastitis cases caused by *E. coli* (Dogan *et al.*, 2006; Döpfer *et al.*, 1999). During severe inflammation of *E. coli* mastitis, systemic administration of fluoroquinolone or cephalosporins is recommended as there is a serious risk of bacteraemia (Suojala, Kaartinen and Pyörälä, 2013; Erskine, Wagner and DeGraves, 2003; Wenz *et al.*, 2001).

In a study with 82 bovine mastitis associated *E. coli* strains from dairy cows in Switzerland, the most prevalent virulence factors were found to be *traT* (a lipoprotein involved in serum resistance), *fyuA* (ferric yersiniabactin uptake protein) and *iutA* (aerobactin siderophore receptor) (Nüesch-Inderbinen *et al.*, 2019). In another study with 63 bovine mastitis associated *E. coli* strains from Israeli dairy cows, *lpfA* (long polar fimbriae), *astA* (heat-stable enterotoxin 1) and *iss* (increased serum survival) were found to be the most prevalent virulence factors (Blum and Leitner, 2013).

E. coli J5 vaccine, which is made from a mutant *E. coli* strain, was created to combat coliform mastitis in dairy farms. *E. coli* J5 strain is used in vaccine formulation as it stimulates antibody production against a wide variety of coliform bacteria. The exact mechanism of the J5 vaccine is currently not known, although antibody production against LPS has long been suggested (Dosogne, Vangroenweghe and Burvenich, 2002). Vaccination against *E. coli* IMI was shown to reduce the incidence of clinical coliform mastitis cases in early studies (Gonzalez *et al.*, 1989). However, later studies did not observe any significant reduction in the occurrence of *E. coli* mastitis but did show a decrease in the severity of the disease (Wilson *et al.*, 2007a; Gurjar *et al.*, 2013). In the experimental trials from New York state, dairy cows injected with J5 coliform vaccines have been shown to have less milk loss and culling rates and to recover their milk yield performance quicker than uninjected cows (Wilson *et al.*, 2009; Wilson *et al.*, 2007b; Wilson *et al.*, 2008).

1.1.3 *Staphylococcus aureus*

Staphylococcus aureus is a gram-positive, haemolytic, round-shaped, mainly catalase-positive and facultative aerobe bacterium (Masalha *et al.*, 2001). *S. aureus* is accepted as one of the major mastitis pathogens due to its consequences on both cow and bulk milk SCC, milk quantity and quality (Keefe, 2012). The main reservoir of *S. aureus* is the udder of the cow; thus,

there is a highly contagious transmission between the animals of the herd particularly during the milking process (Myllys *et al.*, 1997). Moreover, heifers were also shown to be a significant source of *S. aureus* infection in dairy farms where horn flies were found to spread the infection between animals (Oliver *et al.*, 2005; Anderson *et al.*, 2012). In some studies, the farms with effective fly control measures decreased the risk of *S. aureus* infection (Ryman *et al.*, 2013; Piepers *et al.*, 2011).

The trend of coagulase-positive staphylococci (CPS) IMIs from clinical and subclinical bovine mastitis diagnosed cows in the UK between the years of 2012 and 2019 is shown in Figure 1-4, based on data from VIDA annual reports (Surveillance Intelligence Unit, 2020). It is seen to fluctuate through these years with an overall average of 11.29% and 8.71% in the latest report (year 2019). In VIDA annual reports, the organisms were not specified at the species level, but *S. aureus* is known as the most common CPS isolated from dairy cows with bovine mastitis (Boireau *et al.*, 2018). The exact figures about *S. aureus* can be obtained from a relatively old but comprehensive survey of English and Welsh dairy farms, where it was found in 5.2% and 3.3% of subclinical and clinical mastitis cases, respectively (Bradley *et al.*, 2007). However, the prevalence of *S. aureus* mastitis was quite high in some countries such as 43% in the US (Keefe, 2012), 74% in Canada (Riekerink *et al.*, 2010), 62.6% in Ethiopia (Abebe *et al.*, 2016), 36.3% in Egypt (Algammal *et al.*, 2020). Moreover, it was found to be the main pathogen in relatively early studies from European countries like Switzerland and the Netherlands (Poelarends *et al.*, 2001; Schaellibaum, 1999). Moreover, there were differences in the prevalence of *S. aureus* mastitis between dairy herds of the same country or even the same region which was concluded to be the result of different combinations of virulence factors (Piccinini, Borromeo and Zecconi, 2010).

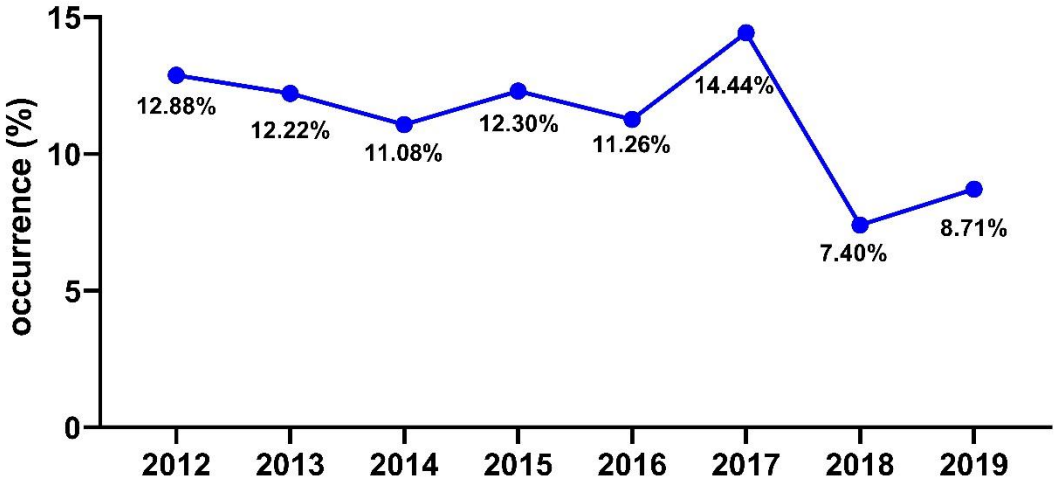


Figure 1-4. The trend of coagulase-positive staphylococci (CPS) IMIs from clinical and subclinical bovine mastitis diagnosed cows in the UK between the years of 2012 and 2019. *On average, CPS, which includes S. aureus, was isolated from 11.29% of the dairy cows diagnosed with bovine mastitis in these years. The graph was generated based on the data from VIDA (Veterinary Investigation Diagnosis Analysis) annual reports between 2012 and 2019 (Surveillance Intelligence Unit, 2020), which do not indicate prevalence or incidence. This figure was generated in GraphPad Prism v8.*

The severity and outcome of the disease were shown to be associated with strain type of *S. aureus* (Le Maréchal *et al.*, 2011; Haveri *et al.*, 2007). Those with the genetic materials for biofilm formation can result in chronic mastitis (Cucarella *et al.*, 2004). The treatment success of *S. aureus* infection relies on the host, pathogen and treatment procedure (Barkema, Schukken and Zadoks, 2006). Parity, days in milk, the infection site (rear or front quarter), SCC, number of the infected mammary quarter count as host factors; strain type and resistance profile as pathogen factors; type, route, initiation and duration of the antimicrobial therapy as treatment procedure were found to be highly affecting the cure rate (Barkema, Schukken and Zadoks, 2006). For instance, multiparity has been found to be negatively correlated with the treatment success in many studies (Deluyker, Van Oye and Boucher, 2005; Sol *et al.*, 2000) and low SCC levels were associated with a higher chance of cure (Deluyker, Van Oye and Boucher, 2005; Dingwell *et al.*, 2003). Moreover, the cure rate was higher in the front quarters (Deluyker, Van Oye and Boucher, 2005; Dingwell *et al.*, 2003) and lower amongst the cows infected by multiple quarters (Østerås, Edge and Martin, 1999). Additionally, several studies (Deluyker, Van Oye and Boucher, 2005; Sol *et al.*, 2000) have proved an increase in the cure rates during long term treatments. *S. aureu*s IMI result in parenchyma deformation; thus, intramammary and/or systemic antimicrobial administration is suggested (Erskine, Wagner and

DeGraves, 2003). It is not wise to follow the same routine for every *S. aureus* infection, because the treatment of some cases is not possible. Chronic *S. aureus* infections generally do not respond to antimicrobial therapy, as the drug cannot be diffused efficiently due to several reasons such as fibrosis of the intramammary tissue, micro-abscesses, penetration of the bacteria with mammary epithelial cells and immune cells (Mullarky *et al.*, 2001; Erskine, Wagner and DeGraves, 2003; Dego, Van Dijk and Nederbragt, 2002). Besides chronicity, the penicillin-resistant *S. aureus* strains and multiparous cows are also not recommended to be treated but culled (Barkema, Schukken and Zadoks, 2006).

Virulence factors and antimicrobial resistance (AMR) genes of *S. aureus* strains isolated from bovine mastitis cases are well studied, these studies were gathered in a very recent review (Pérez *et al.*, 2020a). Biofilm adhesin polysaccharides (*icaA* and *icaD*) play a role in adherence to mammary gland epithelium followed by colonizing and persisting there (Otto, 2013). Staphylococcal enterotoxins (*sea*, *seb*, *sec*, *sed* and *see*) cause inflammation and mammary tissue damage by inducing cytokine secretion (Fang *et al.*, 2019). Toxic shock syndrome toxin 1 (*tst*) also leads to inflammatory reactions in the mammary gland (Kuroishi *et al.*, 2003). *S. aureus* α- and β- hemolysins (*hla* and *hlb*) play a role in the invasion of the mammary gland and causing persistent infection (Dinges, Orwin and Schlievert, 2000). Leukotoxin bicomponent pore-forming complexes (lukMF') fight against host immune defence and helps rapid colonization in the mammary gland (Schlotter *et al.*, 2012). In a recent study with clinical mastitis milk samples from 6 countries (Argentina, Brazil, Germany, Italy, the US and South Africa); *hla*, *hlb* and *sea* were found to be the most prevalent virulence genes with values of 100%, 84.6% and 65.6%, respectively (Monistero *et al.*, 2020).

As a rule of thumb, the prevention of mastitis is more effective than treatment especially for contagious pathogens like *S. aureus*. Several approaches such as live attenuated, inactivated, subunit and toxoid have been applied to develop a vaccine against *S. aureus* IMIs (Pereira *et al.*, 2011). Currently, there are two commercially available vaccines for the prevention of *S. aureus* intramammary infections, which are Lysigin® (Boehringer Ingelheim Vetmedica Inc) - lysed whole-cell vaccine of three most prevalent serotypes of *S. aureus*- in the US and Startvac® (Hipra) – a polyvalent inactivated vaccine - in Europe and Canada (Misra *et al.*, 2018). However, they do not provide full protection for every *S. aureus* strain (Ma, Cocchiaro and Lee, 2004; Scali *et al.*, 2015). Some studies showed significant intramammary reductions in the case of vaccination compared to the control group (Nickerson *et al.*, 1999). In another study, no significant difference was observed in terms of prevention but the severity and duration of

infection were significantly affected (Middleton *et al.*, 2006). Similar results were observed under the field conditions in the UK, where no significant difference in the incidence and prevalence of bovine mastitis was found but a significant reduction in the severity of the clinical cases was seen (Bradley *et al.*, 2015). Another study with Swedish dairy herds also showed that there was no significant difference between commercial polyvalent vaccinated cows and the control group in terms of preventing mastitis problems due to *S. aureus* (Landin *et al.*, 2015).

Control of *S. aureus* is also important for one health approach as livestock-associated methicillin-resistant strains (LA-MRSA) have been detected in dairy farms. The first bovine mastitis related LA-MRSA was detected more than a decade ago in Hungary (Juhász-Kaszanyitzky *et al.*, 2007), then in other European countries such as UK, Denmark, Belgium and Germany (Vanderhaeghen *et al.*, 2010; Spohr *et al.*, 2011; Kreausukon *et al.*, 2012; García-Álvarez *et al.*, 2011). Zoonotic transmission of LA-MRSA is a huge public risk, starting from dairy farmers and their household (Cuny, Wieler and Witte, 2015), on the other hand, the anthroponotic transmission of *S. aureus* should not be discarded as infected livestock were shown earlier (Messenger, Barnes and Gray, 2014; Price *et al.*, 2012). Moreover, the host shift of a clonal complex of MRSA from human to bovine was shown (Sakwinska *et al.*, 2011).

1.1.4 *Enterococcus* spp.

Enterococcus species are gram-positive, catalase and oxidase-negative, non-spore-forming and facultative anaerobic cocci (Ben Braïek and Smaoui, 2019). They cause enteric disorders in animals (Teixeira *et al.*, 2001); moreover, they also cause bovine mastitis. The main *Enterococcus* species isolated from mastitis cases are *Enterococcus faecalis*, *Enterococcus faecium* and *Enterococcus durans* (Rossitto *et al.*, 2002; Cameron *et al.*, 2016). *Enterococcus* species are present in dairy farms especially in organic bedding material, bulk milk tanks or skin of the cows and show environmental transmission route in dairy farms (Rossitto *et al.*, 2002; Cheng and Han, 2020; Petersson-Wolfe *et al.*, 2008). *Enterococcus* species can cause both clinical and subclinical mastitis (Wu *et al.*, 2016).

There is limited information available about pathogenicity, shedding profile and immune response trigger of *Enterococcus* species (Klaas and Zadoks, 2018). Under *in vitro* conditions, bacterial growth differences were observed between *E. faecium* and *E. faecalis* which were collected at various stage of the lactation cycle (Petersson-Wolfe, Wolf and Hogan, 2007).

Dairy cows at early lactation were shown to be more susceptible to *E. faecium* IMI, compared to late lactation cows (Petersson-Wolfe, Wolf and Hogan, 2009).

Enterococcus spp. were found to be the reason for 3.1% and 2.4% of the subclinical and clinical cases in France, respectively (Botrel *et al.*, 2010). Moreover, *Enterococcus* IMIs varied from 3.3% to 15.6% amongst Turkish dairies (Erbas *et al.*, 2016; Gürler *et al.*, 2015). They were recovered from 12.8%, 16.7% and 19.4% of the bovine milk samples in Canada, Czechia and Lithuania, respectively (Cameron *et al.*, 2016; Cervinkova *et al.*, 2013; Klimienė *et al.*, 2011). In a study with Belgian dairy cows, *Enterococcus* (26%) was found to be the second most commonly isolated genus from subclinical intramammary infections (Devriese *et al.*, 1999). Similarly, *Enterococcus* was found to be the predominant agent (28%) related to clinical mastitis in Uganda (Kateete *et al.*, 2013). However, it was not that common in Poland (2.8%), Slovakia (3.1%), Sudan (2.5%) (Krukowski *et al.*, 2020; Ibtisam *et al.*, 2010; Idriss *et al.*, 2014). Two nationwide studies in Norway and Sweden found only one and five positive *Enterococcus* infections, respectively (Østerås, Sølverød and Reksen, 2006; Persson, Nyman and Grönlund-Andersson, 2011). In most of the countries, such as Poland (Różańska *et al.*, 2019), Turkey (Kuyucuoğlu, 2011), Czechia (Cervinkova *et al.*, 2013), Canada (Cameron *et al.*, 2016), Iraq (Hamzah and Kadim, 2018) etc., *E. faecalis* was found to be the most predominant species in *Enterococcus* spp., whereas *E. faecium* was in Uganda (Kateete *et al.*, 2013) and *E. durans* in Lithuania (Klimienė *et al.*, 2011).

The high resistance profile amongst *Enterococcus* species is believed to be a result of horizontal gene transfer (Hershberger *et al.*, 2005) which has been shown *in vitro* (Eaton and Gasson, 2001) and *in vivo* studies (Lester *et al.*, 2006). Pathogenicity of *E. faecalis* is associated with biofilm formation (Elhadidy and Zahran, 2014). It has also been shown that no matter where *E. faecalis* is isolated, in either mastitis cases, milk or bedding material, they can form biofilms and result in persistent infections (Elhadidy and Elsayyad, 2013; Elhadidy and Zahran, 2014). Moreover, endocarditis-specific antigen (*efaA*), enterococcal surface protein (*esp*), gelatinase (*gelE*), hyaluronidase (*hyl*), cytolysin (*cylA*) and collagen-binding cell wall protein (*ace*) were found to be the virulence factors. EfaA is a homolog of adhesin proteins in Streptococci (Waters *et al.*, 2003). Gelatinase plays a role in hydrolyzing biological (e.g. collagen and fibrin) and antibacterial peptides and hence enables bacterial mitigation and spread (Franz *et al.*, 2011; Waters *et al.*, 2003; Schmidtchen *et al.*, 2002). Hyaluronidase has been reported to be associated with adhesion, colonization, tissue damage and spread (Laverde Gomez *et al.*, 2011). Enterococcal surface protein has been shown to participate in adhesion, biofilm formation and

immune response evasion (Tendolkar, Baghdayan and Shankar, 2003; Araújo and Ferreira, 2013). Cytolysin is an enterococci bacteriocin and plays a role in lysing target bacteria cells by forming a pore in their cytoplasmic membrane (Ben Braïek and Smaoui, 2019). Generally, virulence factors are more prevalent in *E. faecalis* than *E. faecium* (Ben Braïek and Smaoui, 2019). The virulence factors of the *E. faecium* strains associated with bovine mastitis were found to be cytolysin (*cylA*), cell wall adhesins (*efaAfm*) and gelatinase (*gelE1*) (Montironi *et al.*, 2020). However, the pathogenicity of *Enterococcus* spp. cannot be explained only with the presence of the virulence genes as AMR genes have also been shown playing a significant role (Franz *et al.*, 2011).

AMR of *Enterococcus* species such as vancomycin resistance is a growing concern within one health approach (Alemayehu and Hailemariam, 2020). Vancomycin-resistant enterococci (VRE) were detected in the late 1990s first in human but subsequently in animals as well (Kühn *et al.*, 2005). Avoparcin usage in the livestock industry was thought to be associated with VRE; hence, animals were considered as the main reservoir of VRE (Kühn *et al.*, 2005). After the ban of avoparcin use in animal farms, the prevalence of VRE colonization in human also dropped (van den Bogaard and Stobberingh, 2000). In Uganda, *Enterococcus* samples collected from dairy cows and farmers showed high similarities of AMR profile; but the zoonotic transmission was not found (Kateete *et al.*, 2013). The multi-resistant profile of *Enterococcus* species, which were isolated from mastitis cases, were found to be highly frequent in Polish dairy farms (Hanna *et al.*, 2019). Similar results were found in studies carried out in the northeast region of China (Gao *et al.*, 2019) and several other regions of China, which were more representative of the country (Yang *et al.*, 2019a). Due to the common multidrug resistance of *Enterococcus* species, the treatment decision should be given carefully. More importantly, control of *Enterococcus* IMI should be focused on the prevention of the infection by proper milking procedures and hygiene rules.

1.2 Mastitis Control

Mastitis prevention has been shown to be more beneficial than the treatment and thus was prioritised in the disease control programs (Dufour *et al.*, 2012). The first mastitis control plan was designed in the UK called “five-point plan” (Dodd and Jackson, 1971). The five-point plan included hygiene of milking equipment, teat disinfection, use of dry cow therapy, prompt treatment of clinical mastitis cases and culling of the chronic cases. The five-point plan successfully reduced the prevalence of the contagious mastitis pathogens, such as *S. agalactiae* and *S.*

aureus, which were a great concern at that time (Bradley, 2002). The broad application of five-point plan decreased the incidence rate of clinical mastitis from 153 cases/100 cows/year to 40 cases/100 cows/year in less than twenty years (Bradley, 2002). However, this was not enough to clear all mastitis cases as the control plan was ineffective against environmental mastitis pathogens, which were the new concern. The cry for the adaptation of a mastitis control programme was partially solved by the ten-point plan of the US National Mastitis Council (NMC), which included management of the dairy environment. Moreover, the UK established its national scheme, Dairy Mastitis Control Plan, in 2009 that focused on individual farms (Down *et al.*, 2016). This plan has achieved a 10% reduction/per year in clinical mastitis in a short time (Green *et al.*, 2012). Other countries such as Canada (Reyher *et al.*, 2011), Australia (Brightling *et al.*, 2009), Netherlands (Lam *et al.*, 2013) and Norway (Østerås and Sølverød, 2009) had their own national mastitis control policies which were similar to the plans outlined above. In terms of outcome, the Norwegian control plan decreased the incidence of clinical mastitis up to 60% in thirteen years (Østerås and Sølverød, 2009). The Dutch control plan was unable to decrease the prevalence of subclinical mastitis significantly (from 23.0% to 22.2%); however, it significantly reduced the incidence of clinical mastitis cases (from 33.5 to 28.1 quarter cases/100 cow-years) (Lam *et al.*, 2013). Although the reduction seems lower in Dutch control plan, it should be noted that its observations were made in a relatively smaller time frame (5 years) than others. Furthermore, there is a growing concern to reduce the antimicrobial usage in dairy farms; thus, blanket dry cow antimicrobial therapy has been replaced with the application of teat sealants and selective antibiotic therapy in many countries (Ruegg, 2017).

1.3 Mastitis Diagnostic Tools for Identification at Strain Level

Strain identification is needed for several reasons which include the epidemiological examination of contagious disease, outbreak investigation, pathogenesis detection and characterisation of the microbial population (van Belkum *et al.*, 2007). Several phenotypic and genotypic methods have been used to identify mastitis pathogens at the strain level. These methods should be assessed case by case based on certain performance criteria such as stability, typeability, discriminatory power, epidemiological convenience, reproducibility, generality on the population, speed, accessibility, easiness and economic cost (van Belkum *et al.*, 2007).

1.3.1 Phenotypic Typing Methods

Phenotypic typing tools categorize the organisms based on the similarities between characteristics (biological or metabolic activities) expressed by them (Emerson *et al.*, 2008). Some of the phenotypic typing methods are biotyping, antimicrobial susceptibility testing, serotyping, multilocus enzyme electrophoresis (MLEE) and MALDI-TOF.

Biotyping is a very conventional typing technique that refers to the biochemical tests measuring different metabolic activities such as hydrolysis of compounds, hemolysis, hemagglutination, sugar fermentation etc. (Maslow, Maury Ellis and Arbeit, 1993). According to the results of these biochemical tests (either positive or negative reaction), the biotype is determined. Although biotyping is a cheap, less laborious and commonly used technique, it is rarely stable and less reproducible (van Belkum *et al.*, 2007). In literature, applications of biotyping can be seen for nearly all mastitis pathogens but most of them are outdated (Myllys *et al.*, 1997; Aarestrup and Jensen, 1996; Nemeth, Muckle and Gyles, 1994). Moreover, the reliability of the biochemical characterisation was found to be low even for species-level identification (Gonano and Winter, 2008).

Antimicrobial susceptibility tests (aka antibiogram-based typing) are based on antimicrobial activity and breakpoints which quantify resistance and susceptibility (Reller *et al.*, 2009). Broth dilution assay, Etest and disk diffusion are currently in use, although traditional applications have generally been replaced with automatic measurements such as Vitek2 (bioMerieux), Sensitre ARIS 2X (Trek Diagnostic Systems), Phoneix (Becton Dickinson Diagnostic Systems), Walk-Away System (Beckman Coulter) and Microscan (Beckman Coulter) (Benkova, Soukup and Marek, 2020). It has been suggested that it should be used to guide treatment decision making for clinical cases (Constable and Morin, 2003). However, the association with antimicrobial susceptibility testing and treatment success depends on the diversity of the population and type of antimicrobials (Petrovski, Laven and Lopez-Villalobos, 2011). There is also a lack of bovine mastitis pathogen-specific protocol, as breakpoints rely on organisms isolated from other species rather than dairy cows, other diseases rather than mastitis or completely different administration route (Hoe and Ruegg, 2005; Apparao *et al.*, 2009).

Serotyping is another phenotypic technique, which is performed by comparison of differing antigens expressed on the cell surface (Jenkins *et al.*, 2017). It has been used since the very early times in microbiology (van Belkum *et al.*, 2007). It has also been used to type several bovine mastitis pathogens such as *S. aureus* and *E. coli* (Ma, Cocchiaro and Lee, 2004;

Fernandes *et al.*, 2011). Although serotyping is fast, easy and reproducible, the discriminatory power of this technique is fair and highly depends on the organism (Maslow, Maury Ellis and Arbeit, 1993).

MLEE is a protein-based technique which is also known as isoenzyme typing (Boerlin, 1997). MLEE performs the typing based on mutations in the bacterial housekeeping enzymes, which can be observed by their electrophoretic migration patterns (Selander *et al.*, 1986). As many enzymes in bacteria are polymorphous, MLEE can provide good discrimination power and highly reproducible results (Maslow, Maury Ellis and Arbeit, 1993). MLEE can be affirmed as the ancestor version of multilocus sequence typing (MLST), which uses housekeeping genes instead of enzymes and enables interlaboratory comparison (Maiden *et al.*, 1998). In literature, there are examples of MLEE usage for the characterisation of bovine mastitis-causing *S. aureus* (Fitzgerald *et al.*, 1997).

MALDI-TOF is another commonly used phenotypic typing technique based on the proteome, which means the set of proteins encoded in an organism (Tyers and Mann, 2003). MALDI-TOF will be discussed comprehensively later in this thesis (section 1.4).

1.3.2 Genotypic Typing Methods

1.3.2.1 Ribotyping

Ribotyping is one of the molecular techniques which identifies bacterial sub-species by using the ribosomal RNA genes' pattern (Caballero, Trugo and Finglas, 2003). In this technique, the genome of the organism is first digested with the family of type-II restriction enzymes and then electrophoresed. After electrophoresis, it is taken to the Southern blot transfer and hybridized with a radiolabelled ribosomal operon probe. The ribotyping profile of the organism is then visualised by autoradiography (Bouchet, Huot and Goldstein, 2008). All bacteria contain unique ribosomal genes, and this differs from each other; however, it provides less discriminatory power at strain level (Bouchet, Huot and Goldstein, 2008). Moreover, the restriction enzyme selection greatly affects the discrimination success of the analysis (Daly *et al.*, 1999). In literature, there are ribotyping technique applications for the identification of bovine mastitis pathogens. These studies were mostly for epidemiological purposes, for instance, the geographical variation of *S. aureus* in Nordic countries (Aarestrup *et al.*, 1997), the distinction of *S. agalactiae* between bovine and human sources (Dogan *et al.*, 2005; Sukhnanand *et al.*, 2005), the distinction of *S. canis* between bovine and cat sources (Tikofsky and Zadoks, 2005),

proportioning of *S. uberis* in environmental sources (Zadoks, Tikofsky and Boor, 2005) or strain isolations of CNS species such as *S. chromogenes*, *S. epidermis*, *S. simulans* from different extramammary sites of the cows (Taponen, Björkroth and Pyörälä, 2008).

1.3.2.2 Pulse Field Gel Electrophoresis (PFGE)

PFGE is another genotypic tool that uses restriction enzymes to cut DNA at several sites and then compares the fragments following electrophoresis (Sharma-Kuinkel, Rude and Fowler, 2016). PFGE has been successfully applied to differentiate strains of bovine mastitis pathogens including *E. coli, S. aureus* and *S. uberis* (Douglas *et al.*, 2000; Blum and Leitner, 2013; Lim *et al.*, 2004). However, interlaboratory reproducibility of PFGE was shown to be low; therefore standardisation of the technique is greatly needed (te Witt *et al.*, 2010). Currently, standardized PFGE protocols are limited to foodborne pathogens only (Gerner-Smidt *et al.*, 2006). PFGE is suggested to be used for studying the outbreaks limited in a geographical area only, not for global epidemiology (Maiden *et al.*, 1998).

1.3.2.3 PCR Based Diagnostic Tools

Randomly amplified polymorphic DNA typing (RAPD) is a PCR technique, but unlike traditional PCR, the DNA segments are randomly amplified as the name suggests (Williams *et al.*, 1990). It has been widely used as a molecular screening tool due to its cheap price, speed and ease to conduct (Munoz and Zadoks, 2007). RAPD is a comparative typing technique; moreover, loci identification for primer binding is not needed. However, the discriminatory power of RAPD-typing was found to be related to the primers used in the analysis (Munoz and Zadoks, 2007). Furthermore, the lack of standardization causes poor reproducibility of the results between laboratories or at different time points (Singh *et al.*, 2009). In terms of screening bovine mastitis pathogens, RAPD has been used for *S. aureus, S. uberis, S. agalactiae, S. dysgalactiae*, *Serratia* spp., *Klebsiella* spp. and *Enterobacter* spp. (Zadoks *et al.*, 2011; Gurjar *et al.*, 2012). It was also employed to show distinct profiles of *S. agalactiae* from the bovine and human origin (Martinez *et al.*, 2000). There were no significant differences found in the profiles of *S. aureus* isolated from bovine and human by using RAPD (Reinoso *et al.*, 2004).

Enterobacterial repetitive intergenic consensus (ERIC), repetitive DNA sequence PCR (rep-PCR) and amplified fragment length polymorphism (AFLP) are other PCR based techniques that have been used to type bovine mastitis-causing *E. coli, Klebsiella* spp. and CNS strains, respectively (Bradley and Green, 2001a; Döpfer *et al.*, 1999; Paulin-Curlee *et al.*, 2007;

Piessens *et al.*, 2011). These comparative typing techniques rely on defining the similar or dissimilar electrophoretic pattern of the isolates. However, they do not provide any genetic information for evolutionary comparisons and cannot be used for wider epidemiological analysis such as comparison of studies no matter when, where and by whom the analysis is performed. To overcome these issues, alternative library typing techniques are needed (Zadoks and Schukken, 2006).

Multiple locus variable-number tandem repeat analysis (MLVA) is another PCR-based typing technique that is based on variabilities on repetitive DNA (van Belkum *et al.*, 2007). Like MLST (section 1.3.2.4), it is a target-specific technique, and similarly, a database was developed but did not become as popular as MLST databases. MLVA was used to strain-type mastitis related *S. aureus* and *S. uberis* (Gilbert *et al.*, 2006b; Gilbert *et al.*, 2006a).

1.3.2.4 Multilocus Sequence Typing (MLST)

MLST is a genotypic technique, which checks the variation of nucleotide sequences in certain sets of housekeeping genes, and was developed as a solution to track the epidemiology of the pathogen (Urwin and Maiden, 2003). Each unique sequence of the housekeeping genes are given allele numbers and a combination of these numbers define the sequence type (ST); which are stored in online databases (https://pubmlst.org/) (Maiden *et al.*, 1998). MLST has been very popular to investigate microbe biology and bacteria evolution, and can commonly be used for the identification of bovine mastitis pathogens at the sub-species level. Online MLST databases are available for the following major mastitis pathogens: *E. coli* (Zhou *et al.*, 2020), *S. uberis* (Coffey *et al.*, 2006), *S. aureus* (Enright *et al.*, 2000), *S. dysgalactiae* (McMillan *et al.*, 2010), *S. agalactiae* (Jones *et al.*, 2003) *E. faecalis* (Ruiz-Garbajosa *et al.*, 2006), *E. faecium* (Homan *et al.*, 2002) etc. Additionally, more than one MLST schemes have been generated, which uses different sets of housekeeping genes, for some species such as *E. coli* and *S. uberis* (Jolley and Maiden, 2010; Zadoks, Schukken and Wiedmann, 2005; Coffey *et al.*, 2006). The STs isolated from mastitis pathogens can be compared within the herds, between countries and with the STs acquired from other sources or infections (Katholm and Rattenborg, 2009; Zadoks, Schukken and Wiedmann, 2005). In this thesis, MLST has been performed in Chapters 3 and 4.

1.3.2.5 Whole-Genome SNP Typing (wgSNP)

SNP represents the single nucleotide variation that occurs in a certain position of a genomic part with respect to a reference, whereas wgSNP refers to the whole genome (Schürch *et al.*,

2018). Hence, the selection of a reference genome is vital for the resolution of the analysis as any genomic information that is absent will be excluded from the comparison analysis (Schürch *et al.*, 2018). One of the disadvantages of reference-based SNP analysis is that comparison with the literature is not possible if studies use a different reference (Schürch *et al.*, 2018). Reference-based SNP mapping can be performed by several workflows such as Snippy, SNVpyl or CSIPhylogeny (Kaas *et al.*, 2014; Katz *et al.*, 2017; Petkau *et al.*, 2017; Seemann, 2015). However, reference-free SNP analysis is also possible (Gardner and Hall, 2013).

With the recent advances and the rising popularity of the next genome sequencing, wgSNP typing has also been applied to mastitis agents. For instance; it has been recently used to explore virulence profiles of *S. aureus* strains isolated from Russian and Danish dairy farms (Fursova *et al.*, 2020; Ronco *et al.*, 2018). Another SNP typing study with bovine mastitis-causing *S. aureus* concluded the importance of variation on the severity of the disease besides virulence genes (Rocha *et al.*, 2019). There were other SNP typing studies with other mastitis pathogens including *E. coli* (Blum *et al.*, 2015; Richards *et al.*, 2015), *S. uberis* (Hossain *et al.*, 2015) and *M. bovis* (Parker *et al.*, 2016). In this thesis, wgSNP typing has been performed in Chapter 4.

1.4 Matrix-Assisted Laser Desorption/Ionization-Time of Flight

Matrix-Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) has been offered as an alternative to biochemical tests and DNA-based techniques due to its straightforward sample preparation, quick analysis, high-throughput capabilities and economical price per run. On average, conventional methods need about one week depending on the biochemical tests (Barreiro *et al.*, 2010) compared to minutes for MALDI-TOF after 1 day of initial bacterial growth (Mellmann *et al.*, 2008). Furthermore, interpretation of the biochemical tests are subjective and mistyping of the mastitis pathogen is common (Bes *et al.*, 2000; Taponen *et al.*, 2006). It was also found to be highly reproducible for the identification of bacterial species by interlaboratory experiments (Mellmann *et al.*, 2009).

MALDI-TOF technique is based on the movement of the sample molecules from either whole cell culture or protein lysate extract mixed with a highly absorbing matrix compound. A laser is used to ionise and desorb the molecules, which are then accelerated by the electromagnetic field in the direction of the detector (Hillenkamp *et al.*, 1991). The arrival time for the particles depends on the molecular weight, as the heavier proteins will arrive later than lighter ones. The pattern of the isolate is generated after the same procedure is repeated multiple times (Coombes, Baggerly and Morris, 2007; Arneberg *et al.*, 2007). The location and intensity of

all peaks are checked against a reference database to detect the best match. Thus, a unique proteome pattern of the analysed organism is found (Ryzhov and Fenselau, 2001).

MALDI-TOF MS was invented through the studies of Franz Hillenkamp and his colleagues on the late 1980s (Karas and Hillenkamp, 1988; Karas *et al.*, 1987). In the second half of the 1990s, bacteria identification using MALDI-TOF MS became possible. Holland and colleagues were the first scientists to show the ability of MALDI-TOF analysis of whole cells for bacteria identification (Holland *et al.*, 1996). In the same year, Krishnamurty and Ross were able to differentiate *Bacillus* sp. at the sub-species level, and then Claydon and colleagues were also able to identify *Staphylococcus* spp. and *E. coli* at species and strain level, respectively (Krishnamurthy and Ross, 1996; Claydon *et al.*, 1996). By 2010, MALDI-TOF was successfully applied to a broad spectrum of bacteria - from gram positives such as *Bacillus*, *Listeria*, *Staphylococcus* and *Streptococcus*; to gram negatives such as *Aeromonas*, *Campylobacter*, *Coxiella*, *Francisella*, *Helicobacter*, *Neisseria* and *Salmonella* - to identify bacteria on species and sub-species level (Murray, 2010).

The promising potential of MALDI-TOF MS technology was also seen by veterinary medicine, where the timely diagnosis of the pathogen is vital for the treatment success of the animal disease (Leitner *et al.*, 2012). By using MALDI-TOF MS, bovine mastitis pathogens including *S. aureus*, *S. agalactiae* and CNS were identified accurately and 8 times quicker than conventional techniques (Barreiro *et al.*, 2010). MALDI-TOF MS were able to identify almost 90% of the *Corynebacterium* spp. isolated from dairy animals diagnosed with subclinical mastitis which are usually hard to specify using conventional methods (Gonçalves *et al.*, 2014; Watts *et al.*, 2000). Other mastitis pathogens *E. faecalis* and *E. faecium* isolates were identified by MALDI-TOF MS as good as phenotypical tests and PCR, but faster and less laborious (Werner *et al.*, 2012). Another study showed the identification success of MALDI-TOF MS for bovine mastitis-causing CNS with an accuracy of 95.4% (Tomazi *et al.*, 2014). In another study with CNS, the accuracy and typeability of MALDI-TOF were increased up to 99.5% and 92.0%, respectively (Cameron *et al.*, 2017). *Enterobacter* spp. which were collected from milk and dairy environment were identified at the species level by using MALDI-TOF (Rodrigues *et al.*, 2017). Specification of *Mycoplasma* spp. from human and animal sources, including mastitic cows, was also shown successfully by using MALDI-TOF MS (Pereyre *et al.*, 2013). In a comprehensive study, MALDI-TOF was shown to have potential of bacteria identification in veterinary applications to alternate current biochemical tests with better accuracy, less laborious and faster (Randall *et al.*, 2015). The higher discriminatory power of MALDI-TOF over other

phenotypic techniques, API and ARIS, as well as quicker analysis time and less reagent usage for the identification of gram-positive and gram-negative bacteria in dairy herds were shown (Savage *et al.*, 2017). However, some examples have shown the insufficient discriminatory power of MALDI-TOF MS for some organisms including dairy-related isolates (Schabauer *et al.*, 2014; Lasch *et al.*, 2014).

Although MALDI-TOF MS, itself, is a quick technique compared to other phenotypic and genotypic diagnosis methods, it still requires the culturing step. Several attempts have been made to exclude culturing, and diagnose the mastitis pathogen directly from the milk. The successful identification of several organisms such as *E. coli*, *E. faecalis*, *S. aureus*, *S. uberis*, *S. agalactiae* and *S. dysgalatiae* were reported (Barreiro *et al.*, 2012; Barreiro *et al.*, 2017). However, at least 10^7 CFU/ml of *E. coli*, 10^6 CFU/ml of *E. faecalis* and *S. aureu*s and 10^8 CFU/ml of *S. uberis*, *S. agalactiae* and *S. dysgalactiae* were needed for direct identification which was way higher than the thresholds to label any milk sample as contaminated with these pathogens (Wisconsin Veterinary Diagnostic Laboratory, 2020).

The working range of the MALDI-TOF MS (<20 kDa) is one of the drawbacks since the entire bacterial proteome cannot be measured in this range (Welker, 2011). The other limitation of MALDI-TOF is that the reference databases are mostly generated with the organisms isolated from human sources rather than animals (i.e. cows diagnosed with mastitis) (Tomazi *et al.*, 2014). This can greatly affect the identification performance of the analysis as identical origins increase the typeability; on the other hand, diverse sources decrease the success of MALDI-TOF MS technology (Mahmmod *et al.*, 2018). Moreover, the accuracy of identification could be improved by adding reference spectra for certain mastitis pathogens (Cameron *et al.*, 2017). Although MALDI-TOF instrument providers (Bruker or Biomeriux) allow the users to customize their databases, there is still a need for a universal database for animal pathogens.

Although MALDI-TOF instruments have been approved only for bacterial identification; they are capable of typing at strain level and predicting antimicrobial susceptibility which have been shown by many studies (van Belkum *et al.*, 2015; Schubert and Kostrzewa, 2017). The proteomic fingerprint of an organism can inform about potential virulence, pathogenicity and antimicrobial profile which can be used to diagnose the disease, estimate the prognosis and even take proper treatment decisions (Coombes, Baggerly and Morris, 2007). Discriminant peaks between different classes can be used to learn more about certain strain or antimicrobial character of an organism (Vrioni *et al.*, 2018).

1.5 Machine Learning Analyses

Several attempts have been made for a proper definition of machine learning (ML). Tom Mitchell described ML as "a computer program is said to learn from experience E concerning some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell, 1997). ML models are generally well-known simple statistic algorithms and, as the definition states, they perform their tasks by learning from the given data unless they are programmed how to process in detail.

ML algorithms are divided into supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and recommender systems (Ayodele, 2010). In this study, as the categorization has been performed on the labelled data only supervised ML has been used so other types will not be described. Supervised ML algorithms train on a set of labelled inputs and learn predicting the correct output to solve mainly classification and regression problems (Sangaiah, 2019). In regression problems, the aim is to predict a continuous numerical outcome whereas in classification problems the aim is to predict the discrete values from a prelabelled list such as binary – yes/no, true/false, 0/1, spam/not spam, positives/negatives etc.- or multiclass, i.e. low, medium and high (Müller and Guido, 2016). In this study, the following supervised learning algorithms were used to solve classification problems: Genetic algorithm (GA), QuickClassifier (QC), Supervised Neural Network (SNN), logistic regression (LR), linear support vector machine (LSVM), radial basis function support vector machine (RBF SVM), multilayer perceptron neural network (MLP NN), decision tree (DT), random forest (RF), AdaBoost, naïve Bayes (NB), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

1.5.1 MALDI-TOF Data Post-Processing Software ClinProTools

ClinProTools is commercial software that was built by Bruker Daltonik GmbH to analyse MALDI-TOF mass spectra. The software calculates the recognition capacity (RC), which measures the discrimination success of the features between different classes, and cross-validation (CV), which estimates the performance of the model by splitting the data into two sections: training, to train the model, and validation to validate it (Bruker Daltonics, 2011). It allows the user to tune data preparation settings such as baseline subtraction, peak definition, recalibration, resolution and normalization. Statistical analysis of the peaks can be calculated by the software. It can be used to generate a model by employing one of the four algorithms:

GA, SNN, QC and SVM. It should be noted that SVM was not available in the software used in this current work. In this study, three algorithms –GA, QC and SNN- were used according to ClinProTools' manual (Bruker Daltonics, 2011).

1.5.1.1 Genetic Algorithm

Genetic algorithm (GA) is the application of natural evolution, which is the idea of survival of the fittest, in computing to replace the brute force approach (Goldberg and Holland, 1988). In this study, the best peak combinations, which are found as the most relevant in discriminating the MALDI profiles of individual classes, are selected by using GA (Bruker Daltonics, 2011). Although the performance of GA is high, the results will always be the closest estimation to the optimal solution, as one cannot guarantee the best peak combination without trying all the peaks (brute force approach).

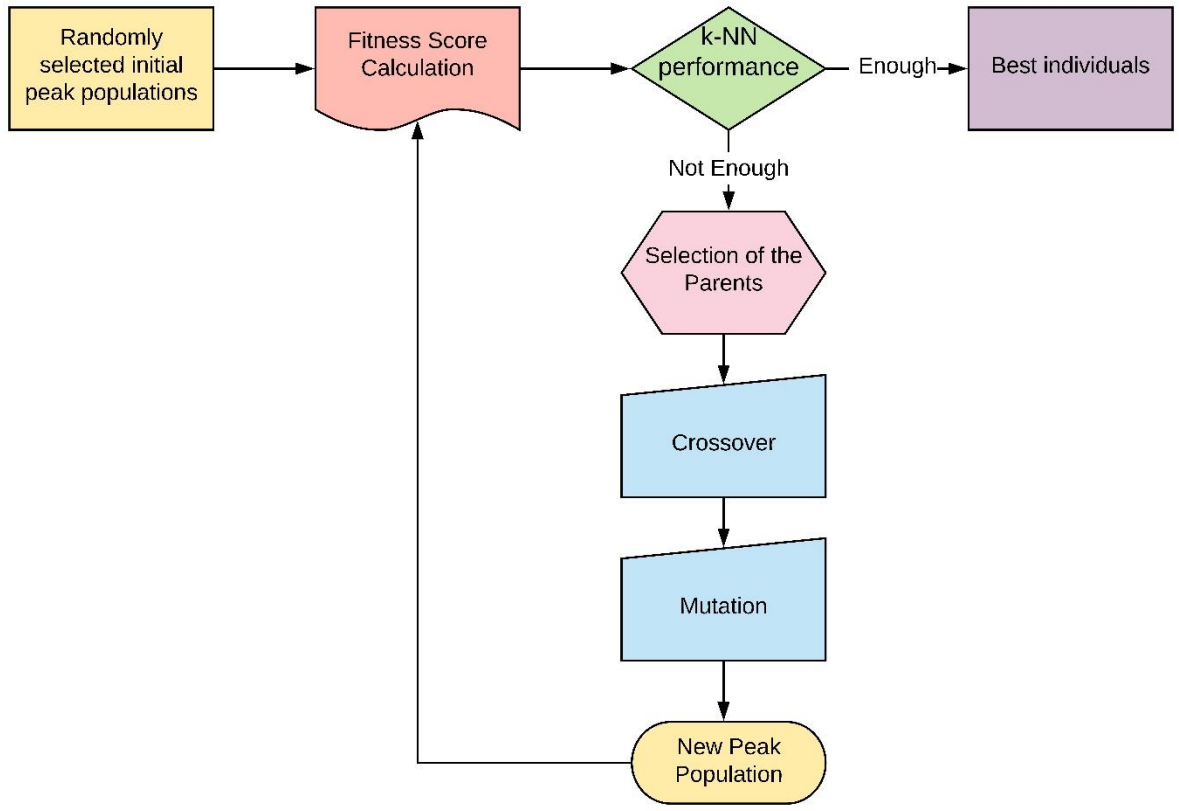


***Figure 1-5. The workflow of Genetic Algorithm**. Initial peak population is generated randomly followed by calculating the fitness scores of each peak. Parents are then selected and altered by crossover and mutation. The performance of the new peak population is rated again by k-NN performance. These steps are repeated until the best performing peak population is achieved. This figure was generated using Lucidchart.com.*

In ClinProTools, GA is coupled with a k-nearest neighbour (k-NN) classifier, which is used to determine the fitness score of the peak combinations. K-NN classifier defines the neighbours of query spectra based on the distance and labels the query with the class membership of the neighbours accordingly. The basic principles of the GA can be summarized as seen in Figure 1-5. The first step is the initiation of the population, which is comprised of random peak combinations. Then, the fitness score is calculated for each peak combination in the population. In the next step, two-parent peak populations are selected based on their fitness scores. These two parents are then used to generate child peak combinations by cross-over, which enables swapping the parts of the parents. Later, the child peak combinations are modified based on predefined mutation probability. Finally, the child peak combinations are added to a new population which takes place of the old population. These steps are repeated until the best results are obtained in the k-NN classifier in terms of RC and CV (Bruker Daltonics, 2011).

1.5.1.2 Supervised Neural Network

Supervised Neural Network (SNN) algorithm identifies unique spectra for individual classes, which are named as prototypes and then performs classification based on these prototypes (Bruker Daltonics, 2011). The determination of the prototypes is a vital process as the query spectra are labelled based on these prototypes only. ClinProTools randomly assigns the predefined number of prototypes and then optimizes them based on their discriminatory power between classes. The determination of the prototypes could be simply explained as shown in Figure 1-6.

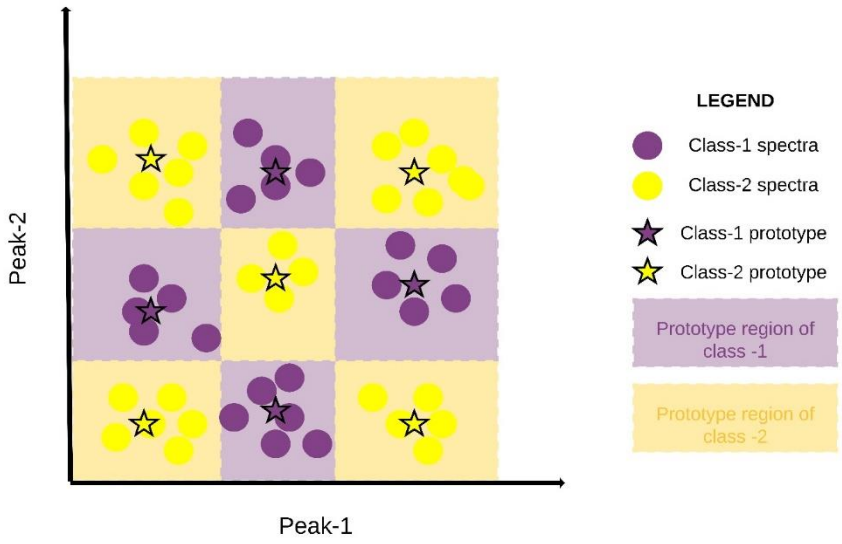


Figure 1-6. Prototype determination of Supervised Neural Network (SNN). *SNN needs to identify proteomic characteristics of the classes to label new (unknown class) spectra. Hence, the prototypes (yellow and purple stars) and the prototype regions for two classes (yellow and purple circles) of the problem were defined. For illustration purposes, only two peaks were used in the axes but data cannot be shown in two-dimensional data space in a real-life problem. The figure was based on Bruker Daltonics (2011) and generated using Lucidchart.com.*

As seen in Figure 1-6, two classes are shown in two-dimensional space where the axis are two different peaks. However, it should be noted that real-life problem cannot be shown just in two-dimensional data space, as there will be generally more than two peaks. Regions are assigned for these two classes according to their prototypes which are the subset of data points from original data. When a query spectrum falls in one of the classes' regions, it is predicted with that class' label (i.e. yellow or purple class).

1.5.1.3 QuickClassifier

QuickClassifier (QC) is a univariate sorting algorithm that separates the classes based on areas and statistical characteristics of the peaks. For classification, the area of each peak is calculated, and the area of all peaks are averaged for each class. These figures are stored together with the sorted values coming from statistical tests such as *t*-test/Anova and Wilcoxon/Kruskal-Wallis. The peaks are determined based on statistical character, and their values are compared to classify the query spectra. QC algorithm enables tracing back of the classification results owing to its simple characteristics. In the case of limited sample size, QC algorithm has been shown to outperform other algorithms in ClinProTools (Bruker Daltonics, 2011). The illustration of QC can be seen in Figure 1-7.

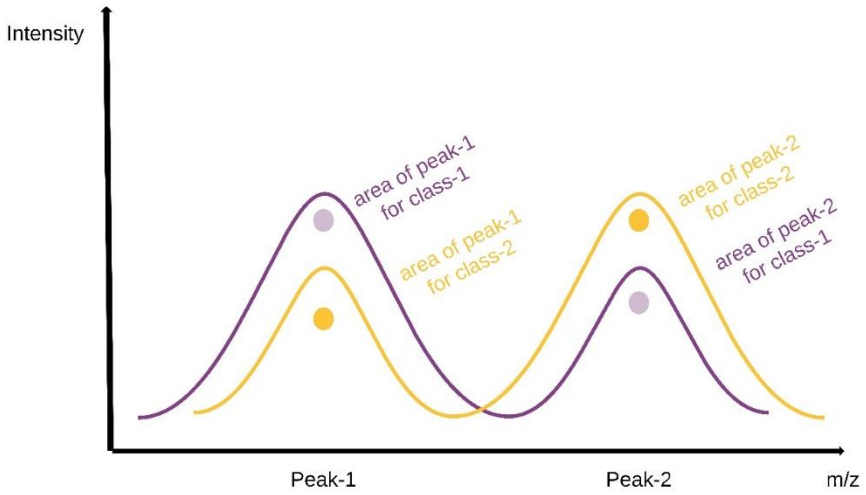


Figure 1-7. Illustration of QuickClassifier (QC). *For each class, the area under the peaks is calculated and stored with values gathered from statistical tests. For illustration purposes, only two peaks were used but in real-life problem data, it would highly likely have more peaks. The area under peak-1 and peak-2 was calculated for class-1 (purple) and class-2 (yellow), respectively. The figure was adapted based on the description from the ClinProTools manual and generated using Lucidchart.com.*

1.5.2 Open-Source Python Environment

To compare with ClinProTools, algorithms from the scikit-learn library in Python were used (Pedregosa *et al.*, 2011). Following ML algorithms were employed in this thesis: LR, LSVM, RBF SVM, MLP NN, DT, RF, AdaBoost, NB, LDA and QDA. In the following sections, the theoretical and mathematical background of the algorithms are briefly explained.

1.5.2.1 Logistic Regression

Logistic regression (LR) is one of the basic but widely used ML algorithm, which often provides solutions for simple problems. LR is one of the initial ML algorithms to apply to complicated problems as it is quick to finalize, requires less computational power, produces easy to interpret results and is possible to run in almost any language environment. It is useful for understanding the key and redundant features to design complicated models (Hosmer Jr, Lemeshow and Sturdivant, 2013). The ability to provide probabilities and classifications for new samples based on continuous and discrete measurements makes LR a popular ML method. LR aims to set the best-fitting model to define the relationship between dependent and independent variables (Yan, Koc and Lee, 2004). This is done by computing weighted sums of input features with bias. Then, the logistic function estimates how probable a query belongs to a certain group (Géron, 2019) (see Figure 1-8).

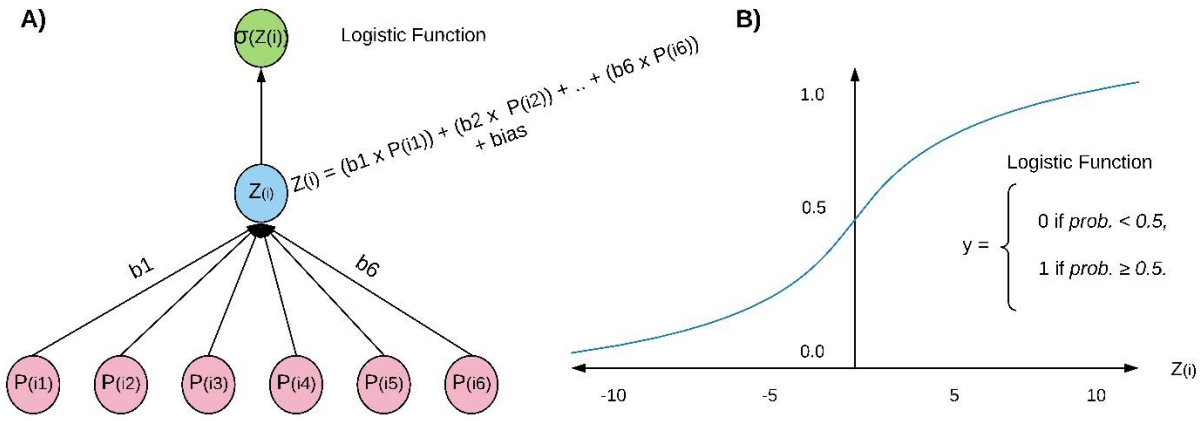


Figure 1-8. Illustration of logistic regression. A) *Basic principles of logistic regression are demonstrated. P(i) represents the peak of (i)th element in the dataset. In this illustration, a total of six peaks are shown. To compute linear predictive model (Z_i); each peak is multiplied by parameter b (b1, b2, b3, b4, b5, b6) correspondingly, which refers to the weight of peak for the prediction, and bias was added to the equation. Then, this equation was converted to a probabilistic equation by logistic function ($\sigma(Z_i)$).* ***B)*** *Logistic function estimates the outcome probability (shown as "prob." in the figure) between 0 and 1; where values equal or greater than 0.5 are appointed to the positive class (class 1), values lower than 0.5 are appointed to the negative class (class 2). The figures are generated based on information from Géron (2019) and Carin (2020). They were generated using Lucidchart.com.*

As the LR has less predictive power than other ML algorithms, complex problems cannot be solved perfectly. However, there are applications of LR on detecting AMR in several bacteria such as *S. aureus* (Rishishwar *et al.*, 2014) and *Enterobacteriaceae* (Pesesky *et al.*, 2016). LR has been widely used in the dairy industry as well and found to outperform DT, SVM and RF on predicting conception success by using insemination records (Hempstalk, McParland and Berry, 2015). In another dairy cow insemination study, LR, DT, RF and NB were employed to predict the insemination outcome of dairy cows and LR was found to be the best performer again (Fenlon *et al.*, 2016).

1.5.2.2 Support Vector Machines (SVMs)

Vladimir Vapnik was the first scientist to come up with the idea of SVMs (Vapnik, 1995), they were then modified and extended to the current version by Vapnik and Cortes (Cortes and Vapnik, 1995). In this study two types of SVMs, linear and radial basis function, are used for binary classification problems (discussed in the next two sections).

SVMs have been previously used both in AMR detection and bovine mastitis diagnosis. SVM was successfully used to differentiate vancomycin-resistant *S. aureus* isolates from susceptible

ones (Rishishwar *et al.*, 2014). In another study, SVM was performed to diagnose bovine mastitis based on animal data such as milk yield, stage of lactation, mastitis history and milk electrical conductivity (Miekley, Traulsen and Krieter, 2013).

1.5.2.2.1 Linear Support Vector Machine (LSVM)

Beside LR, the other most common linear classification algorithm is LSVM (Müller and Guido, 2016). Linear classification models aim to classify training observations by determining the optimal hyperplane (see Figure 1-9). In LSVM, the optimal hyperplane is the one giving the best soft margin (distance between the hyperplane and the training observations on the edge which are called support vectors), which should be as maximum as possible considering the outliers (Géron, 2019). The soft margin is regulated by controlling the "*C*" hyperparameter of LSVM, where smaller values of "*C*" may result in underfitting, and higher values of "*C*" may cause overfitting (Müller and Guido, 2016).

High dimensional problems can be solved efficiently by LSVM. As a subset of training points are used to generate the decision boundary, it is also computationally memory adequate (Pedregosa *et al.*, 2011). LSVM generally work well in the cases of which the number of features is greater than the number of samples (Müller and Guido, 2016); however if the difference is so high it may lead to overfitting (Pedregosa *et al.*, 2011). One of the pitfalls of SVMs is that the probabilities of classification are not provided (Pedregosa *et al.*, 2011).

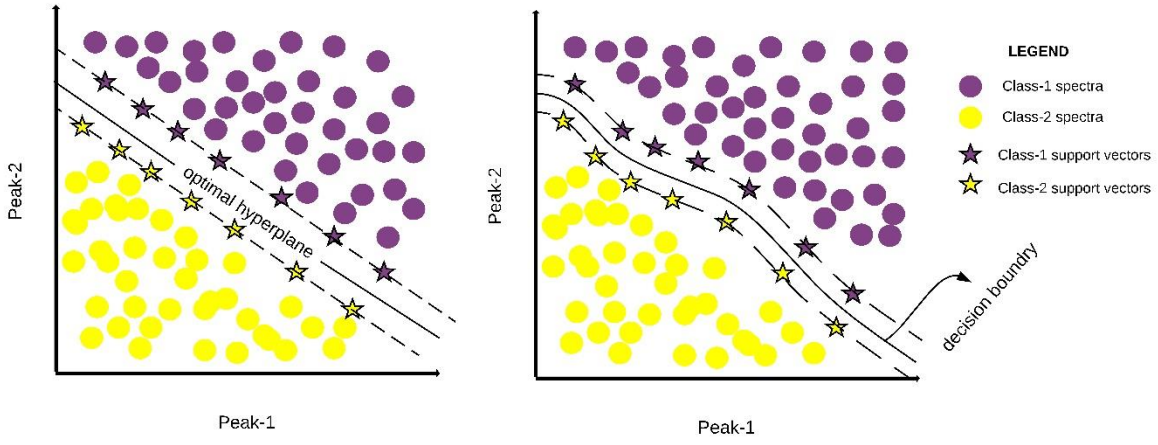


Figure 1-9. Illustration of linear support vector machine (LSVM) (left) and radial basis function support vector machine (RBF SVM) (right). *The support vectors (purple and yellow stars) were determined to give the best separation between the two classes (purple and yellow classes). Decision boundaries were set linearly and non-linearly for LSVM and RBF SVM, respectively. For illustration purposes, only two peaks were used but in real-life problem, data is not always in*

two-dimensional data space. The figure was adapted from Müller and Guido (2016) and generated using Lucidchart.com.

1.5.2.2.2 Radial Basis Function Support Vector Machines (RBF SVM)

Linear models usually do not perform well enough for complex datasets, as the linear boundaries offer limited decision power (Géron, 2019). Hence, LSVMs must be transformed into a much more powerful model by adding nonlinear character such as computing the polynomials of original features. However, this is not an easy task as one cannot know which features to add in advance or the high-dimensional transformations may be computationally demanding (Müller and Guido, 2016). Fortunately, this is achieved by "kernel trick", a computation which provides learning about the model in high-dimensional space by computing the relationship between every pair of the data points without high-dimensional transformations also needing to be made (Géron, 2019). There are several kernel types such as polynomial, sigmoid and radial basis function (RBF) (Pedregosa *et al.*, 2011). RBF kernel can generally show more adaptive characteristics and quicker performance than other kernels (Mueller and Massaron, 2016). As polynomial or sigmoid kernels were not employed, they will not be detailed in this thesis.

RBF kernel is also known as Gaussian kernel and can be briefly explained as the application of all possible polynomial degrees until the best decision boundary is found (Müller and Guido, 2016) (see Figure 1-9). However, it should be noted that as the polynomial degree increases, the importance of the feature decreases (Müller and Guido, 2016). Two hyperparameters need tuning for the good performance of RBF SVM which are "*C*" and "ɣ" (gamma). Here, "*C*" hyperparameter works like the LSVM, when *C* value is small, the data points have less influence so the decision boundary is almost linear; and when *C* value is high, the data points have much more influence which enables bending the decision boundary (Müller and Guido, 2016). Gamma parameter manages the width of the Gaussian kernel and defines the importance of being closer in datasets. When the gamma value is small, the decision boundary will be smooth and increasing the gamma value results in more complex models (Müller and Guido, 2016).

RBF SVM could be used for complex datasets where LSVM fails to perform. RBF SVM could be also employed for cases where the number of features is many or a few. It can handle robustly the overfitting, the noise and outliers. RBF SVM performs well as much as NNs but with a faster analysis time (Kubat, 2017). The computational needs of RBF were also found to be less than other conventional algorithms such as k-NN (Ding and Li, 2009). RBF SVM needs

pre-processing of the datasets and tuning of the parameters. As tree-based models (i.e. RF or gradient boosting) does not require that much pre-processing, RBF SVM is demanded less. Furthermore, it is not so easy to interpret how the prediction is performed (Müller and Guido, 2016).

1.5.2.3 Neural Networks (NNs)

Linear models fail to handle non-linearities in the data; thus, more sophisticated models, such as NNs, are needed to provide flexible decision boundaries. NNs form the base of deep learning where the algorithms are inspired by the structure of the human brain (Géron, 2019). In this thesis, multilayer perceptrons (MLPs), which are the most basic form in the family of NN algorithms, have been used (Müller and Guido, 2016). MLPs are used for classification and regression problems and are the starting point of advanced deep learning. They consist of three main layers: an input layer, one or more hidden layers, and an output layer (Moody and Darken, 1989) (see Figure 1-10). Similar to biological NN; the input layer behaves like a dendrite, which is the input of a neuron, hidden layer process the information just like the cell body and the output layer behaves like an axon which is the output of a neuron (Li *et al.*, 2020).

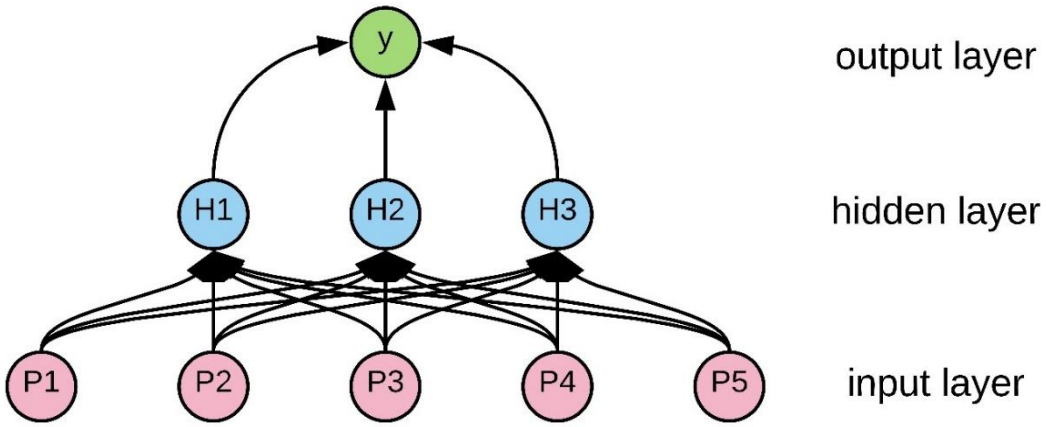


Figure 1-10. Structure of basic multilayer perceptron neural network. *Multilayer perceptron neural network consists of an input layer, one or more hidden layers and an output layer. In this illustration five features of the data (shown as peaks in our study: P1, P2, P3, P4 and P5) are taken as an input, computed in a single hidden layer with three nodes (H1, H2 and H3) and presented as an output (y). The figure was adapted from Müller and Guido (2016) and generated using Lucidchart.com.*

MLP can be briefly described as the improved or extended version of LR (when the sigmoid function is used as an activation factor), where the same working mechanism is applied but not only once (Géron, 2019). The weighted sums of input features are computed like linear models. However, the power of NNs compared to linear models comes from the activation function, which introduces nonlinearity to the model in the hidden layer. The results of the activation function are used in the weighted sum that calculates the output (see Figure 1-11). The errors with delta are calculated and all weights and biases are adjusted by backpropagation (Müller and Guido, 2016).

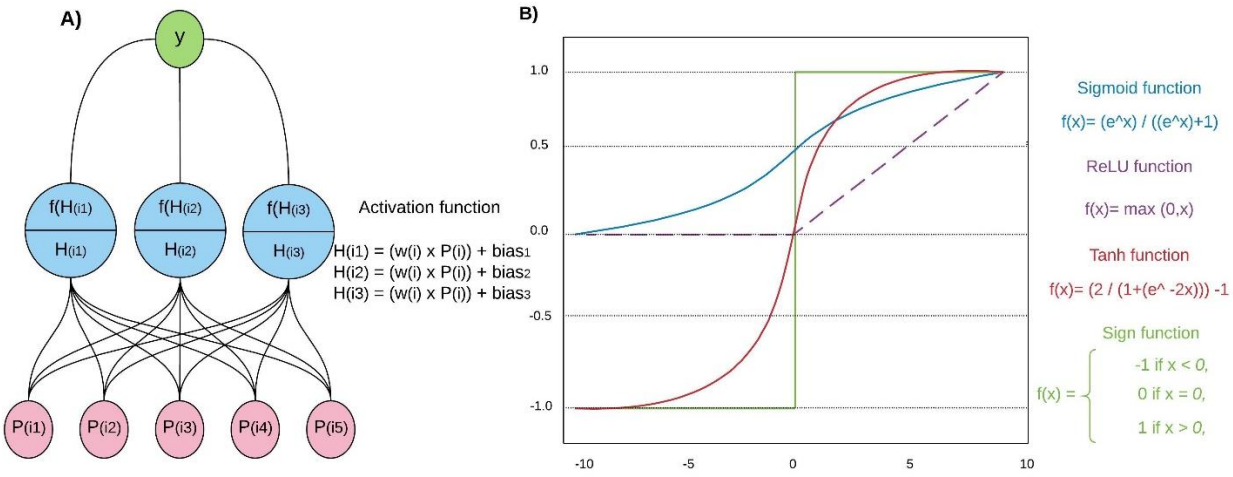


Figure 1-11. Illustration of multilayer perceptron (MLP) neural network. A) *Basic principles of MLP with a single hidden layer are demonstrated. Three layers – input, hidden and output - are shown in pink, blue and green colours, respectively. P(i) represents the peak of (i)th element in the dataset. In this illustration, a total of five peaks are used as input. Each peak is multiplied by its weight w (w1, w2, w3, w4, w5, w6) correspondingly and bias was added to each equation. Then in the hidden layer (blue coloured), non-linearity was introduced based on the selected activation function. Eventually, the values computed in the hidden nodes are summed and output is generated.* ***B)*** *Several activation factors can be used in the hidden layers such as sigmoid, ReLU (rectified linear units), tanh (hyperbolic tangent) and sign function. Sigmoid activation factor is the same function as in logistic regression which outputs a probability between 0 and 1. ReLU activation factor outputs the result if it is greater than 0, or else it outputs 0. Tanh activation factor is like sigmoid function but outputs between -1 and 1 instead. Sign activation factor outputs 1, 0 and -1 for positive, zero and negative values, respectively. The figures are generated based on the information from Géron (2019), and Müller and Guido (2016). The figures were generated using Lucidchart.com.*

The main hyperparameters to tune in NNs can be listed as:

- The number of hidden layers: which is mostly one but could be increased due to the complexity of the problem (Panchal *et al.*, 2011).

- The number of nodes in each hidden layer: there is not a specific answer for this but in literature different criteria are used such as; middle values between input and output nodes when they are drastically different, less than two times of the input nodes in order not to get overfitting or two-thirds of the sum of input and output nodes (Müller and Guido, 2016).
- The types of activation function: the options are sigmoid, ReLU, tanh and sign activation factor (see Figure 1-11) (Müller and Guido, 2016; Pedregosa *et al.*, 2011).
- Learning rate: learning rate affects the speed of outcome as slow learning rates could last hours, days or even weeks while fast learning rates could not give decent results. On the other hand, if model training is performed for a long time, it may result in overfitting and the generalization ability of the model is lost (Géron, 2019; Pedregosa *et al.*, 2011).

It should be noted that NN algorithms start to learn from randomly assigned initial weights. Hence, using the exact parameters with the same datasets may not result in the same outputs. However, this mostly applies to smaller networks whereas larger networks and properly tuned complexity will result in similar accuracy (Müller and Guido, 2016; Pedregosa *et al.*, 2011).

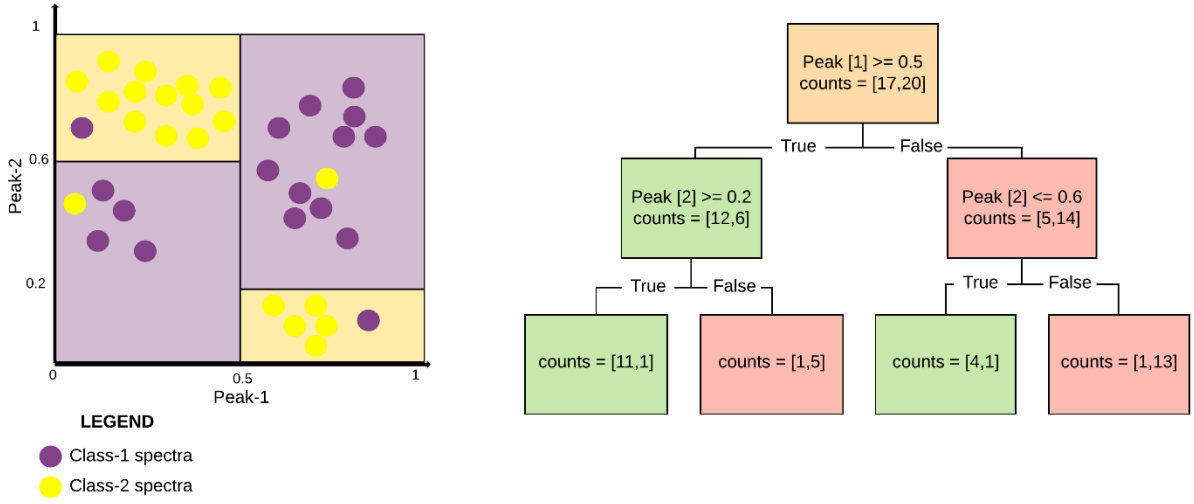
The superiority of NNs over other ML algorithms is that NNs can learn from large datasets and build more complex models when enough computation resources are given. The main drawback of NNs is that some models need a long time to show good performance. For NN algorithms, data processing and hyperparameter tuning are needed to obtain confident results such as kappa values over 85% etc. Just like SVMs, NNs perform the best when the data is homogenous (similar measurements of the features) (Müller and Guido, 2016; Pedregosa *et al.*, 2011).

MLPs are successfully applied to AMR detection (Rishishwar *et al.*, 2014). Application of MLPs in the dairy industry is not rare either, for instance, two hidden layer MLPs have been shown as the best performer to define the cows with artificial insemination difficulties (Grzesiak *et al.*, 2010). MLPs have also been used to detect mastitis in dairy farms by using automatic milking system's data (Sun, Samarasinghe and Jago, 2010; Wang and Samarasinghe, 2005).

1.5.2.4 Decision tree

Decision tree (DT) algorithm, as the name suggests, is a tree-framed diagram and determines a roadmap with a final decision by asking if/else questions (Müller and Guido, 2016). It can be

used for solving both regression and classification problems (Géron, 2019). For a clear understanding of the DT, some terms are vital. The root node is the highest point where the tree starts, whereas the leaf node is the furthest point where no more segregation is possible. The nodes which are not root or leaf node are called internal nodes. The root node is the parent node of all other nodes (leaf and internal nodes) which are also named as a child node. Splitting is the categorization of root and internal nodes according to certain criteria (see Figure 1-12). Pruning is the removal of the branches that are redundant for the performance of the DT. The questions asked during the learning process of the DT are called tests, which should not be confused with the test data (Boehmke and Greenwell, 2019).



*Numbers in the square brackets represent the spectra count of each class in order of Class-1 and Class-2 (Counts = [Class 1, Class 2]).

***Figure 1-12. Illustration of a decision tree**. Decision boundaries are shown on the left and corresponding questions asked and answered to define these borders are shown on the right. Two classes (purple and yellow) are categorized based on the intensities of two peaks (peak-1 and peak-2). Orange block is the parent node, the middle two blocks in the second layer are internal nodes and the bottom blocks are the leaves of the decision tree. In the parent node, data points were separated based on the intensity values of peak-1. In the internal nodes, data points were separated based on the intensity values of peak-2. This figure was derived from Müller and Guido (2016) and generated using Lucidchart.com.*

The tests should be defined accurately for higher performance. DT analyses all the probable tests and defines the most descriptive about the dataset. The classification could be based on binary features such as yes/no and true/false or numeric for continuous data (VanderPlas, 2016). The tests can be decided by calculating different scores such as Gini impurity, information gain, chi-square or variance reduction (for regression) (Mueller and Massaron, 2016). In the current work, Gini impurity, which measures the likelihood of an incorrect classification

(it is 0 when all training data points belong to the same class), was used as it was computationally efficient (Géron, 2019).

The most common problem with DT algorithm is overfitting. Overfitting occurs when the algorithm is trained with the noise in the dataset. It can be prevented by two types of pruning which are pre- and post-pruning. Post-pruning is the erasing of branches that do not contain remarkable information after the model is built. Pre-pruning is the early prevention for the model based on depth in the tree, number of leaves and number of points in each node of the tree. It should be noted that in this thesis, only pre-pruning is applied to DT models as the scikit-learn package is not implemented with post-pruning (Pedregosa *et al.*, 2011).

Graphical representation of the process makes DT models easier to interpret (see Figure 1-12). However, the deeper the tree gets, it will be more confusing to do so. To prevent this, the depth should be defined in the pre-pruning step with the other parameters listed above (Pedregosa *et al.*, 2011; Géron, 2019).

Feature scaling (i.e. normalization and standardisation) is not needed unlike other ML algorithms (e.g. SVM or ANN), and DT performs well when the features are on completely dissimilar scales (Pedregosa *et al.*, 2011).

DTs have been previously used in livestock-related studies, for instance, difficult parturition in Irish dairy cows could be predicted (Fenlon *et al.*, 2017). Milk content data including fat, protein and lactose levels can be provided by automated monitoring devices. This data then can be used handled by ML algorithms for predicting the subclinical mastitis in dairy farms. New Zealander researchers performed classifiers such as NB, LR and DT on such dataset to estimate subclinical mastitis (Ebrahimi *et al.*, 2019; Ebrahimie *et al.*, 2018). In other automated milking data study, DT was employed to diagnose clinical mastitis in Dutch dairy farms (Kamphuis *et al.*, 2010).

1.5.2.5 Random Forest

Ensemble is an ML technique that combines several learners to build a more efficient and accurate model compared to their singular use (Hastie, Tibshirani and Friedman, 2009). These learners are combined with two different methods: bagging and boosting. Random forest (RF) is a bagging algorithm (aka bootstrap aggregating); which means that it is a combination of multiple DT algorithms (Breiman, 2001). RF is offered as a solution for the commonly seen overfitting problem of DTs (Müller and Guido, 2016). The idea is that if more DTs different

from each other employed and performed well, less overfitting will occur by averaging the results of these trees. The first thing is defining the number of trees to include in RF. By the bootstrapping method, random samples are selected from the data points where some of them are selected more than once and typically one-third of them are not even selected once. By doing this, a new dataset as big as the original one is constructed. Then, DTs are grown up from the newly formed dataset which pick a random subset of features and try for the best possible test. This step is controlled by defining the number of the maximum number of features (Mueller and Massaron, 2016; Müller and Guido, 2016; Breiman, 2001). It should be noted that this parameter should be less than the total number of features, otherwise randomness is violated. The lower the number of maximum features is, the higher chance the forest is grown by different trees. Predictions are made in every grown DT and the final verdict is given by aggregating the votes (see Figure 1-13). Bootstrapping and using the aggregate for a decision is termed bagging (Müller and Guido, 2016).

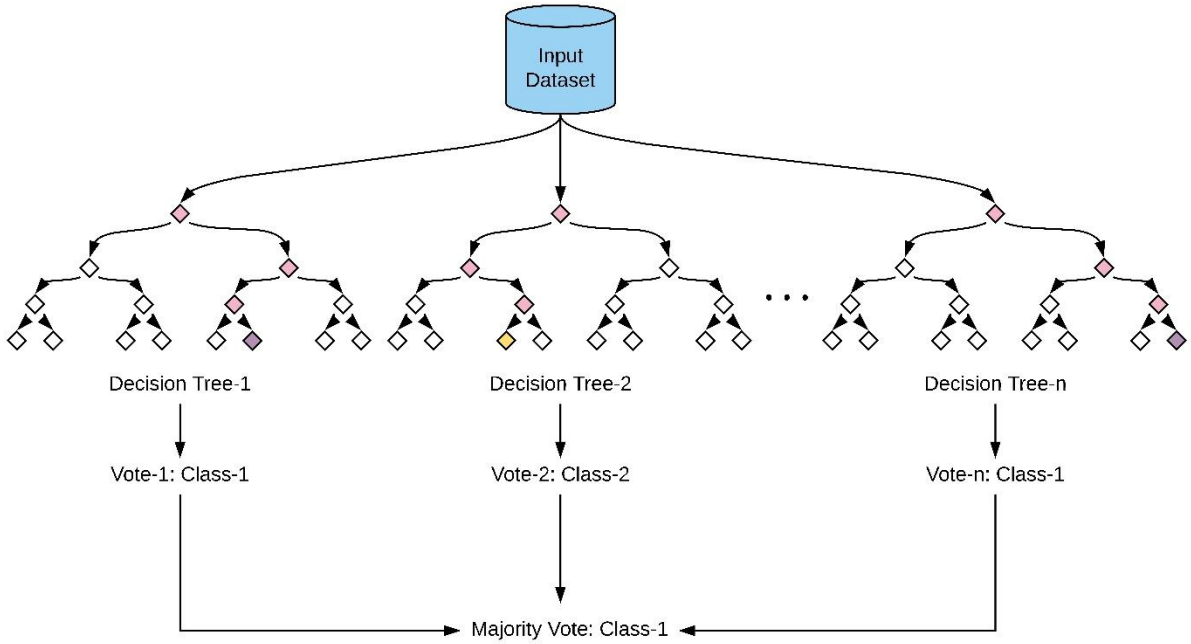


Figure 1-13. Illustration of random forest. *N amount of decision trees are generated by random sampling, and categorization is performed in each one of them separately. Each decision tree has one vote, and the majorly voted class is picked as the final decision. In this particular binary classification example, the majority of the decision trees voted for Class-1 and hence is labelled with that class. This figure was based on Géron (2019) and generated using Lucidchart.com.*

RF is used both for regression and classification problems. It does not require pre-processing steps like scaling or normalization. One of the biggest advantages of RF is that missing values either in the training or test dataset can be tolerated (Mueller and Massaron, 2016). The other

superiority of RF model is that the importance of the features can be calculated as in DT but with more confidence (Géron, 2019).

One of the drawbacks is RF is that visualization of multiple trees is not possible; hence, DT should be selected for visualisation and educational purposes. Moreover, RF models do not perform well on certain characteristics such as high dimensional datasets, sparse datasets and text datasets (Müller and Guido, 2016).

RF has been commonly used in dairy-related studies. Liver fluke exposure in European dairy farms could be predicted by performing RF on bulk milk tank data (Ducheyne *et al.*, 2015). RF was performed on genotype data of Polish dairy cows to predict their susceptibility towards mastitis (Daniel *et al.*, 2016). In another study, RF successfully predicted IMIs between dry and lactation periods (Hyde *et al.*, 2020). In another study, the main behaviours of dairy cows were predicted by performing several algorithms including SVM, RF and AdaBoost where RF had the highest performance (Riaboff *et al.*, 2020). Moreover, RF was found to give better performance than other algorithms including NB and DT, for predicting the insemination outcome in dairy cows (Shahinfar *et al.*, 2014).

1.5.2.6 AdaBoost

Boosting is a method that combines the weak learners sequentially to build a more powerful algorithm (Hastie *et al.*, 2009). There are several boosting methods but the most popular one is adaptive boosting (AdaBoost), which can be combined with linear models or NB but more often with DTs to solve a problem (Mueller and Massaron, 2016). As being an ensemble model of DTs, it may be confused with RF but there are certain differences between them. Each DT is grown fully in RF; however, in AdaBoost generally each tree consists of just a root and two leaves, which is named as “stump”. Stumps are not as good as full trees for making predictions of classification and are technically called weak learners. All features are used to decide in a full-sized tree whereas a stump can use only one feature at a time (Alpaydin, 2020). The other difference between RF and AdaBoost is how they treat their trees or stumps in case of voting for the final verdict. Each tree has the same vote in RF whereas in AdaBoost, some stumps have more, and some have fewer votes for the final decision (Hastie, Tibshirani and Friedman, 2009). AdaBoost creates stumps in order so that the error made in the early stumps affects the following stumps. RF is generated by the independent trees which do not influence each other (Breiman, 2001).

The following steps describe briefly how the AdaBoost works; firstly, the same weight is given to each sample which states that all the samples are equally important. These weights for each sample are then adapted according to the first stump. The first stump is created by deciding the best feature at categorising samples (i.e., Gini impurity). The amount of vote that the first stump will have for the final decision is computed according to its performance of classification (see Figure 1-14). For this, the total error of the stump which indicates the incorrectly classified samples are used in a special formula. The lower the total error is, the higher the vote for the final decision the stump gets. After creating the first stump, the sample weight of incorrectly classified samples is increased while the sample weight of correctly classified samples is decreased. Thus, incorrectly classified samples are ensured to be considered by the following stump. This is the idea behind how the previous stump affect the following stump (Hastie *et al.*, 2009; Mueller and Massaron, 2016; Kubat, 2017).

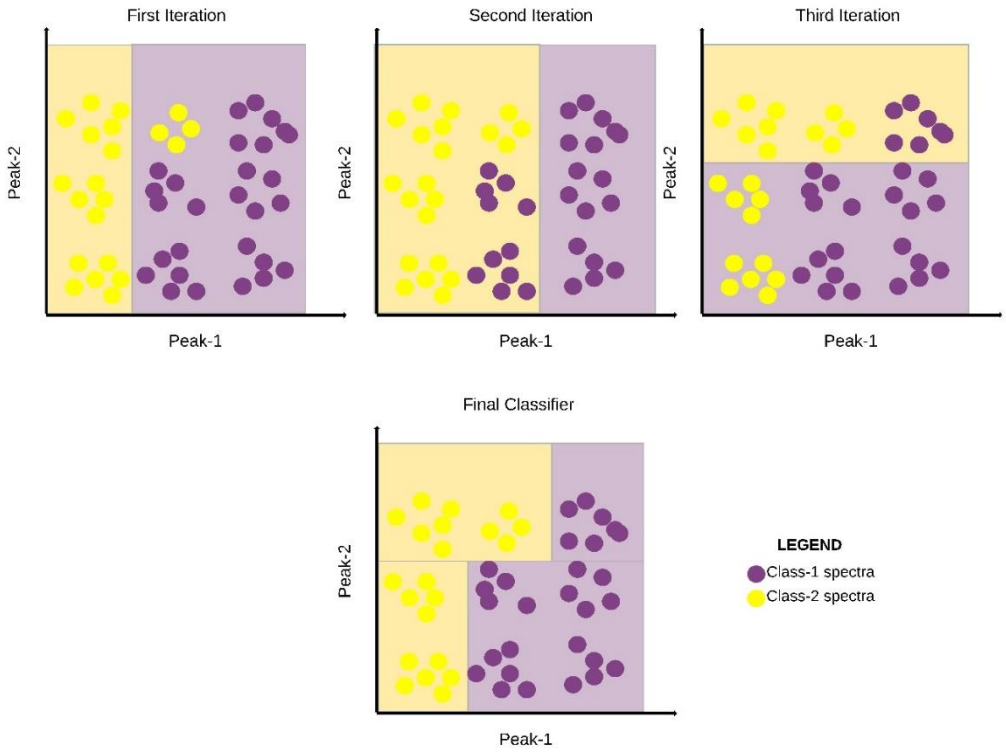


Figure 1-14. Illustration of AdaBoost. *The weak learners try to classify the data subsequently, and the final classification was decided based on the weight given through these iterations. In this example, the decision boundary for the two classes (purple and yellow) was drawn after three iterations. Again, it should be noted that only two peaks were used but in real-life problem, data is not always in two-dimensional data space. The figure was derived from Schapire and Freund (2013) and generated using Lucidchart.com.*

AdaBoost has wide applications in real-life data such as medical decisions (Thongkam *et al.*, 2008), genome-wide association studies (Assareh, Volkert and Li, 2012), AMR detection (Davis *et al.*, 2016) etc. In the Python scikit-learn package, several parameters can be set other than default values but tuning the number of estimators is a good start to cope with the overfitting problem (Pedregosa *et al.*, 2011).

1.5.2.7 Naïve Bayes (NB)

NB is a plain but interestingly efficient model for classification problems (Zhang, 2005). The algorithm gets its name from Bayes' theorem and makes naïve assumption of independence between input features (VanderPlas, 2016). Bayes' theorem states the rational way of revising an existing state of knowledge about a parameter to a new state with given new research data (Green *et al.*, 2008).

Briefly, NB classifier works on the principle of the following steps: firstly, the probabilities of each feature in every class are computed. Then the product of all probabilities related to each resulting class is calculated. These products are normalized, and the class given with the highest probability is chosen as the final verdict (Kubat, 2017). NB learns from the input data by treating each feature independently and calculating basic statistics (Müller and Guido, 2016). The assumption of individual feature is impractical and may not fit in real-world problems. However, NB was shown to have equivalent performance to that of complex ML algorithms (Banko and Brill, 2001).

NB algorithm family consists of three different classifiers in the python sci-kit learn package: Bernoulli, Gaussian and Multinomial (Pedregosa *et al.*, 2011). In this current work as the data was continuous, Gaussian NB was used. Gaussian NB does not have any hyperparameter and is used for baseline classification continuous data by assuming Gaussian distribution (see Figure 1-15) (VanderPlas, 2016).

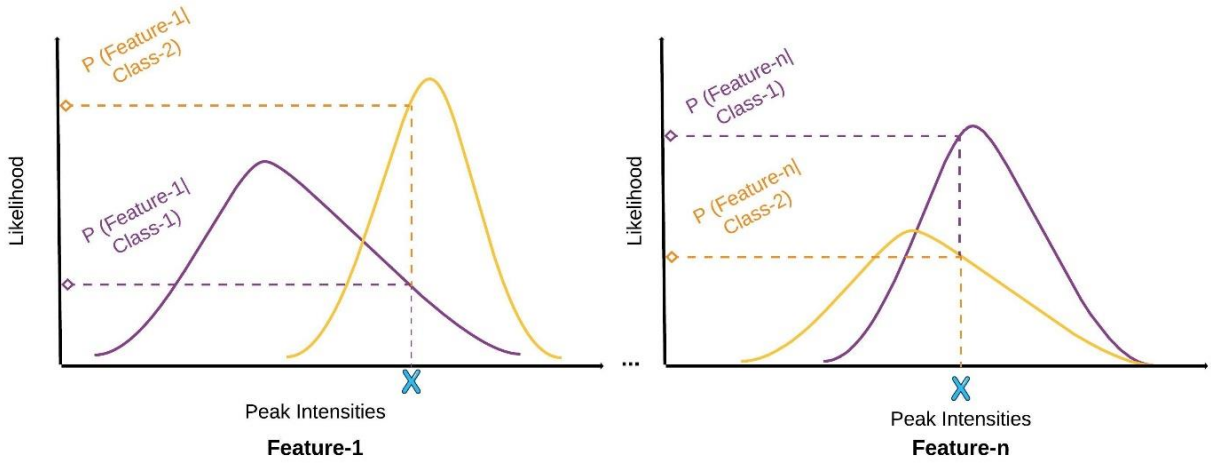


Figure 1-15. Illustration of Gaussian naïve Bayes. *In this illustration, purple and yellow colours represent Class-1 and Class-2, respectively. Gaussian curves are computed for each feature (peak) in the classes based on the mean and standard deviation of the peak intensities. When new data come, the likelihood of the features in this data in each class (Class-1 and Class-2, respectively) are calculated (P(Feature-n|Class-1) and P(Feature-n|Class-2)). The probability of the data belonging to Class-1 and Class-2 is then calculated separately based on Bayes' theorem and higher possibility is defined as the class of the new data (P(new data|Class-1) and P(new data|Class-2)). The figure was adapted from Raizada and Lee (2013) and generated using Lucidchart.com.*

There are several advantages of selecting the NB algorithm over other classifiers. NB does not require huge datasets to train itself. It can handle both continuous and discrete data. It is a quick classifier and thus can be applied to real-time predictions. Even huge datasets can be analysed quicker than other algorithms (Müller and Guido, 2016). It is not restricted to binary classes and can easily process multiple classes. NB does not require detailed pre-processing and is therefore widely used in ML applications in advance of, or instead of, employing more complex classifiers such as NNs (Müller and Guido, 2016).

NB has been previously used in medical diagnosis (Soni *et al.*, 2011), AMR detection (Rishishwar *et al.*, 2014), bacterial colony fingerprinting (Maeda *et al.*, 2018), biomarker discovery (Ralhan *et al.*, 2008) and proteomics (Liu *et al.*, 2009). It was also used to classify clinical mastitis bacteria according to gram status and genus, respectively (Steeneveld *et al.*, 2009).

1.5.2.8 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a linear classification algorithm as the name suggests, which was developed by famous statistician Ronald Fisher in the mid-1930s (Fisher, 1936). LDA uses Bayes' theorem for classification and makes the following assumptions: i) each class

is drawn from a multivariate Gaussian distribution, ii) classes have their own mean vector, iii) variance is common amongst classes (Pedregosa *et al.*, 2011). LDA is commonly mentioned with principal component analysis (PCA), where they both reduce the dimensions but LDA focuses on maximizing the separability between classes while PCA does not have any intention like categorization based on class (Martínez and Kak, 2001). LDA employs all features to create a new axis and outline the data on this new axis in a way to maximize the separation between these classes. A new axis is created by maximizing the distance between means of the classes and minimizing the scatter (variation) within each class (Pedregosa *et al.*, 2011).

LDA can be used for the classification of not only binary classes but also more classes when the measurement of features are continuous (Abdi, 2007). In literature, there are applications of LDA on automatic milking system data to detect mastitis in New Zealander dairy farms (Sun, Samarasinghe and Jago, 2010; Wang and Samarasinghe, 2005). LDA was employed to categorize mastitis pathogens based on farm records and was able to predict with the accuracy between 42% and 57% of the cases depending on the herd characteristics (Heald *et al.*, 2000).

The number of training sample is important for the performance of the algorithm. Although LDA can perform on relatively small datasets, when the number of features is greater than the number of samples, the performance is highly affected. One solution provided in the scikit-learn project is a tool named “shrinkage”. Shrinkage enhances the covariance matrix where the empirical one is not good enough in a situation like described above. Shrinkage parameter could be set a value between 0 and 1, where 0 means no shrinkage (empirical value) and 1 means complete shrinkage (update the covariance matrix by using a diagonal matrix of variances) (Pedregosa *et al.*, 2011).

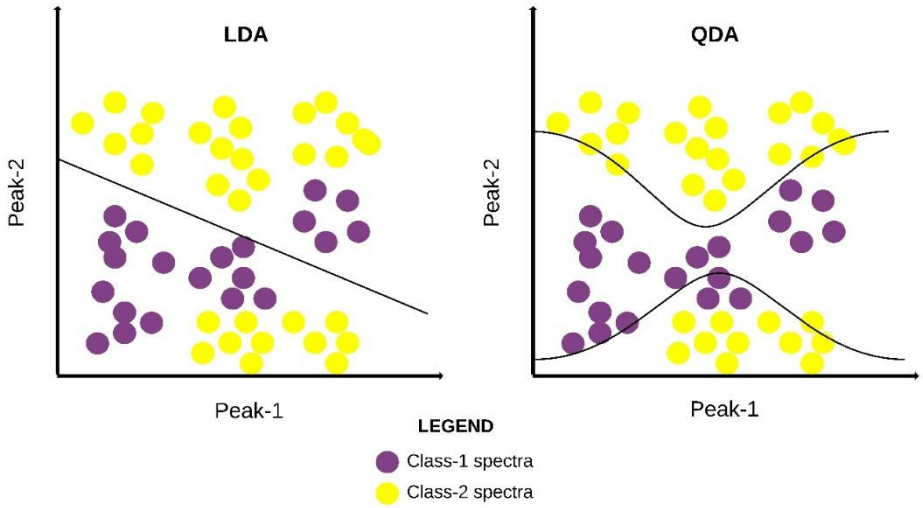


Figure 1-16. Illustration of linear discriminant analysis (LDA) (left) and quadratic discriminant analysis (QDA) (right). *To categorize 2 classes (purple and yellow), decision boundaries were set linearly and non-linearly for LDA and QDA, respectively. For illustration purposes, only two peaks were used but in real-life problem, data is not always in two-dimensional data space. The figure was adapted from Pedregosa et al., (2011) and generated using Lucidchart.com.*

1.5.2.9 Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) also uses Bayes' theorem for classification (Pedregosa *et al.*, 2011). The fundamental difference between LDA and QDA is that LDA is based on a linear function and decision boundary is linear, instead, QDA is a nonlinear function so the decision boundary is not linear but quadratic (Duda, Hart and Stork, 2012) (see Figure 1-16). Both QDA and LDA assumes the Gaussian distribution, but in LDA one common covariance matrix is calculated for all classes whereas the covariance matrix is calculated for every single class in QDA (McLachlan, 2004). In binary classification problems when the covariances of the classes are equal, a linear decision boundary is drawn and LDA will be enough to solve it (Alpaydin, 2020). When the covariances of the classes are not equal, a non-linear boundary will be needed and QDA will outperform the LDA (Hastie, Tibshirani and Friedman, 2009).

QDA performs better if the covariances of the classes are distinct. A handicap of QDA compared to LDA is that QDA cannot be used as a dimension reduction technique (McLachlan, 2004). QDA is more flexible and can handle more variance than LDA. It can be employed to train large datasets where LDA cannot work efficiently (Alpaydin, 2020). QDA and LDA classifiers are commonly used due to their high speed for training the data. The other advantage of these algorithms is that they do not have any hyperparameters to tune (Pedregosa *et al.*, 2011).

1.5.3 ML Applications on MALDI-TOF Data for Bacteria Strain Typing

MALDI-TOF coupled with ML algorithms has been commonly used for bacterial strain typing. DT and SVM models were employed to differentiate strain types of MRSA using MALDI-TOF profiles, providing an alternative to MLST (Wang *et al.*, 2018b). GA, SNN and QC were used to discriminate major serotypes of *S. pneumonia* in Japan (Nakano *et al.*, 2015). In another study which aimed to differentiate sub-species of *Mycoplasma pneumoniae*, GA was able to correctly identify these strains with a sensitivity and specificity of 100% (Xiao *et al.*, 2014). LSVM on MALDI-TOF data of seven *Bacillus* spp. performed accuracy up to 90% (Al-Masoud *et al.*, 2014). SVM was also applied successfully to discriminate *Klebsiella pneumonia* complex members (Rodrigues *et al.*, 2018). Chung and colleagues performed several ML algorithms including RF, DT and SVM to strain type *Staphylococcus haemolyticus* based on MALDI profiles and achieved almost 85% of the area under the curve (AUC) (Chung *et al.*, 2019). On the other hand, not every study of MALDI-TOF coupled with ML gave satisfactory results. For instance, MLPs on *E. faecium* and *S. aureus* isolates were failed to differentiate in strain level by using MALDI-profiles (Lasch *et al.*, 2014).

1.5.4 ML Applications on MALDI-TOF Data for Antimicrobial Susceptibility Testing

ML algorithms have also been successfully employed to classify bacteria according to the antimicrobial profile based on MALDI profiles. Heterogeneous vancomycin-intermediate *S. aureus* (hVISA) could be detected by employing DT, RBF SVM, k-NN and RF, where RBF SVM outperformed the other algorithms (Wang *et al.*, 2018a). Moreover, Asakura and colleagues designed a graphical user interface that uses ML on MALDI-profiles to classify vancomycin susceptible *S. aureus* (VSSA), vancomycin-intermediate *S. aureus* (VISA) and hVISA (Asakura *et al.*, 2018). MALDI-profiles of MRSA and MSSA were also aimed to be differentiated by using GA (Bai *et al.*, 2017), SVM (Sogawa *et al.*, 2017), RF (Tang *et al.*, 2019). *S. aureus* strains used in these studies were all isolated from human patients.

RF, NB, SVM, LR and k-NN algorithms were employed to classify carbapenem-resistant and susceptible *K. pneumoniae* isolates based on MALDI-profiles and RF was found to be the best performer (Huang *et al.*, 2020). MALDI-TOF coupled with SVM was employed to differentiate beta-lactamase-producing isolates of *Enterobacteriaceae* and *Pseudomonas aeruginosa*; however, it could only perform with up to 70% accuracy which was not sufficient enough for routine diagnostics (Schaumann *et al.*, 2012).

MALDI-TOF MS coupled with ML has been recently performed to classify VRE and vancomycin-susceptible enterococci (VSE) on the largest real-world datasets coming from human blood, sterile body fluid, wound, urinary and respiratory tract. In this comprehensive study, three ML algorithms (RF, RBF SVM and k-NN) were used where RF outperformed the other classifiers. The performance of RF was validated by 5-fold cross-validation and two different external data based on time point and location, all resulted in at least 84.00% of AUC. It was concluded that MALDI-TOF MS coupled with ML was more accurate (up to 30%) and quicker (up to 50%) than the traditional approach (Wang *et al.*, 2020).

1.6 Biomarker Characterisation

MALDI-TOF spectral peaks recognized as discriminant by the trained classifiers were cross-matched with proteins (i.e. biomarkers). These proteins were characterized by using bioinformatics analyses such as Gene Ontology, 3D structural modelling and protein-protein interaction, which are explained in the following sections.

1.6.1.1 Gene Ontology

Gene Ontology (GO) was generated to provide a common framework for the functions of gene products across species (Ashburner *et al.*, 2000). GO examines the gene products in three aspects as molecular function, biological process and cellular component (Ashburner *et al.*, 2000). Molecular function represents the activity carried out by the gene product; cellular component states where this activity occurs in the cell; and biological process describes the larger biological objective to which the gene product contributes (Thomas, 2017).

It can be used for several reasons to analyse products of high throughput analysis (Gaudet *et al.*, 2017) such as functional profiling of a subset of the genes (Rhee *et al.*, 2008), evaluating the functional annotations of enzymes (Holliday *et al.*, 2017), or function estimation of unannotated genes (Burge *et al.*, 2012). Performing functional enrichment analysis on proteomics data (i.e. MALDI spectral profile) enables testing systematic measurements of the proteome which provides a better understanding of pathogen metabolism rather than genomics or transcriptomics, as protein biomarkers are more stable for phenotyping (Chen *et al.*, 2020). In the current study, GO analysis was used for two objectives i) to annotate the possible functions of not well known or hypothetical proteins ii) functional enrichment of the sets of gene product including the discriminant proteins and their first interactors.

1.6.1.2 Three-Dimensional (3D) Structural Modelling

3D structural modelling is mainly used to estimate the biological functions of the proteins as the protein structure governs the interaction of it with ligands or other molecules (Lopez *et al.*, 2007). It has been divided into three categories: homology modelling (aka comparative modelling) (Martí-Renom *et al.*, 2000), threading/folding recognition (Jones, Taylort and Thornton, 1992) and *ab initio* modelling (Wu, Skolnick and Zhang, 2007). In this thesis, homology modelling and threading/folding recognition have been used to predict 3D protein structures. Building 3D structure of biomarkers found discriminatory by classifiers provides an understanding of the function of the protein in the cell and estimate the location of the binding sites that can be used as drug targets (Dorn *et al.*, 2014).

1.6.1.3 Protein-Protein Interaction (PPI)

Traditional biochemical techniques centre the characterisation of a single protein, the results of which are well archived and curated at well-known protein databases (e.g. UniProt) (Kaake, Wang and Huang, 2010; UniProt, 2018) However, proteins mainly act as a team rather than individual to perform their biological functions at cellular and system levels such as signal transduction, transportation, DNA regulation and alternative splicing (De Las Rivas and Fontanillo, 2010; Berggård, Linse and James, 2007; Eisenberg *et al.*, 2000).

Analysing the PPI network which is consisted of biomarkers picked by the classifiers can help better understanding the disease-causing mechanism of the pathogen that can eventually play a crucial role in treatment optimization (Arabnia and Tran, 2015). In this study, PPI was used to outline protein complexes and learn their biological pathways in detail.

1.7 Summary of Research Aims

The main aim of this study was to develop a computational diagnostic solution for the rapid and accurate identification of pathogens at the subspecies level causing bovine mastitis, one of the most significant diseases in the dairy world where it causes serious economic and welfare issues. This study focused on the analyses of specific mastitis agents: *S. uberis*, *E. coli*, *S. aureus*, *E. faecalis* and *E. faecium*. The first three pathogens (*S. uberis*, *E. coli* and *S. aureus*) were selected, as they are the most frequently isolated bovine mastitis agents worldwide. *E. faecalis* and *E. faecium* are the other bovine mastitis pathogens that have been increasingly

isolated from recycled manure solids (aka green bedding) and bulk milk tanks (Gagnon *et al.*, 2020; Bradley *et al.*, 2018).

Since each one of the species is featured by various strains which are the significant driving force for the outcome of the disease, this study aimed to develop a tool for the identification of mastitis pathogens at the subspecies level. The current study considered to develop an integrated MALDI-TOF, ML and bioinformatics platform that could provide highly accurate discrimination of the mastitis pathogens at the subspecies level; moreover, to annotate corresponding proteins of the discriminant peaks underlying a phenotype.

MALDI-TOF MS was selected as the method for the identification of the pathogens because this technology has the ability of typing at the subspecies level and to profile AMR profiles. It has been widely used in identifying the proteomic fingerprints of the organisms under certain conditions and enables discrimination by means of multiple biomarkers pattern. Moreover, MALDI-TOF MS coupled with ML is capable of providing fast analysis turnout, economic cost per sample, easy application and resolution ability. Finally, a new bioinformatics pipeline was integrated which enabled the identification and characterisation of the molecular determinants (i.e. their molecular function, biological process and interaction with the rest of bacterial proteome) underlying the studied phenotypes.

The specific aims and objectives of this work were as follows in each chapter.

Chapter 3: i) to investigate MALDI-TOF MS coupled with ML to discriminate between bovine mastitis-causing *S. uberis* isolates with different modes of transmission (contagious and environmental), ii) to compare strain differences within and between dairy farms of the UK, iii) to identify proteins related to the differentiating peaks between transmission routes.

Chapter 4: i) to understand the genotypic and phenotypic characteristics of bovine mastitis-causing *E. coli* strains, ii) to identify genotypic profiles of bovine mastitis-causing *E. coli* isolates by whole-genome sequencing, iii) to investigate MALDI-TOF MS coupled with ML to discriminate between bovine mastitis-causing *E. coli* isolates with different clinical outcome (clinical and subclinical) and disease phenotype (persistent and non-persistent), iv) to identify biomarkers related to the clinical status of bovine mastitis-causing *E. coli* isolates.

Chapter 5: i) to provide a fast and more accurate alternative to standard antimicrobial susceptibility tests, ii) to investigate MALDI-TOF MS coupled with ML to profile multidrug and benzylpenicillin resistance in bovine mastitis-causing *S. aureus* isolates, iii) to identify proteins

related to the differentiating peaks between multidrug-resistant and susceptible, and benzylpenicillin-resistant and susceptible isolates.

Chapter 6: i) to test the power of MALDI-TOF MS coupled with ML for profiling AMR in a more general perspective (several types of antimicrobials and different organisms), ii) to identify proteins related to the differentiating peaks between resistant and susceptible profiles of *E. faecalis*, and between resistant and susceptible profiles of *E. faecium*.

Additionally, in each chapter potential biomarkers related to segregate different classes were targeted to be identified by using bioinformatics tools. These biomarkers were supported with additional information about their PPI, 3D protein structures, GO and KEGG functions and literature mining.

CHAPTER 2 METHODS

2.1 Data Source

The data used in this current work was provided by Quality Milk Management Services (QMMS) Ltd under an awarded Innovate UK grant. The isolates had been originally collected for previous studies. *S. uberis* and *S. aureus* isolates analysed respectively in Chapters 3 and 5 were collected for an intervention study carried out in 52 Welsh and English dairy farms by Green and colleagues (Green *et al.*, 2007). *E. coli* isolates analysed in Chapter 4 were collected for an incidence and aetiology of clinical mastitis study carried out in 6 Somerset (England) dairy farms by Bradley and Green (Bradley and Green, 2001b). *E. faecalis* and *E. faecium* isolates analysed in Chapter 6 were collected in a cross-sectional study of UK farms by Bradley and colleagues (Bradley *et al.*, 2018). The following information is provided from these previous works.

2.2 Sample Preparation

The isolates were kept in a bead-based micro preservation system (Protect, Technical Service Consultants, Heywood, UK) at -80˚C. These isolates had been previously identified by standard classification techniques (Bradley *et al.*, 2007); therefore, direct whole culture MALDI-TOF analysis was carried out to be sure about the identification at the species level.

The samples from the bedding materials were collected by qualified workers who were previously trained about the methodology. Minimum of 10 bedding material samples (75 ml) were taken from different cubicles on each farm. These samples (total of at least 750 ml) were then combined and mixed utterly. The samples (500 ml) from the bulk milk tank, which had 1 or 2 days of milking, were collected on the same visit. On the occasion that multiple bulk milk tanks were employed for the milk collection, samples were taken from all and then combined to represent the proportion of milk in the tanks. The samples were then instantly sent to the laboratory for bacteriological examination.

The bacteriological examination of the samples collected from bedding and bulk milk performed according to the following steps. First, 30 g of mixed bed materials were mixed with 270 ml of the maximum recovery diluent in the lab blender at 100 rpm for 1 min. Serial dilutions of milk and bedding material were taken into plates with Edwards agar and incubated at

around 37˚C between 66 to 72 hours. The thresholds of detection for *Enterococcus* spp. were defined as 10 CFU/g and 1 CFU/ml for bedding and milk samples, respectively.

2.3 Generation of MALDI-TOF Spectra

MALDI-TOF analysis by ethanol-formic acid extraction protocol was performed to generate spectra used in the current work just as previously carried out by Barreiro *et al.* (2010) for mastitis pathogens. Microorganisms were transferred into a tube containing 300 µl of molecular-grade water and vortexed. Then, 900 µl of 100% ethanol was put into the same tube and vortexed again. The sample was then centrifuged at 20,800 x g for 3 min. The supernatant was removed, and the pellet was dried at around 20˚C. Then, 50 µl of 70% formic acid and the same amount of 100% acetonitrile were added, mixed and centrifuged at 20,800 x g for 2 min. After centrifuge, 2 µl of supernatant was spotted onto the MALDI target plate and left drying. After drying, 2 µl of HCCA matrix solution was added onto the target plate and air-dried before MALDI-TOF analysis.

For each isolate, six technical replicate profiles were generated from 240 desorptions (6 x 40 shots). Spectra were compared visually using Biotyper 3.1 (Bruker Daltonics). The technical replicates with an insufficient resolution, low intensity or substantial background noise were removed. Technical replicates were further compared using composite correlation indices (CCI) to remove dissimilar spectra with CCI<0.99 (Arnold and Reilly, 1998). Each isolate with less than three technical replicates was removed from further analyses except Chapter 4, where the isolate counts were relatively small. Isolate recovery, MALDI-TOF analyses and quality control of the spectra were performed by QMMS Ltd.

2.4 Pre-processing of the Data

In Chapter 3, the data pre-processing was performed by ClinProTools (see the Methods section of Chapter 3). In Chapters 4, 5 and 6, it was performed by using MATLAB Bioinformatics Toolbox Release 2017b, The MathWorks, Inc., Natick, Massachusetts, United States. The following parameters were set for pre-processing to extract peak list:

a) Mean computing: the technical replicates of each biological isolate were averaged.

b) Mass range filter: the mass range of the spectra was limited to 2-12kDa.

c) Resampling: the data was up-sampled from 13,740 points to 20,000 points.

d) Baseline correction: window and step sizes of 200Da with pchip regression and quantile method of 0.1 was used for baseline editing.

e) Normalization: the AUC of each spectrum was normalized to the median and post-rescaled such that the maximum intensity was 100.

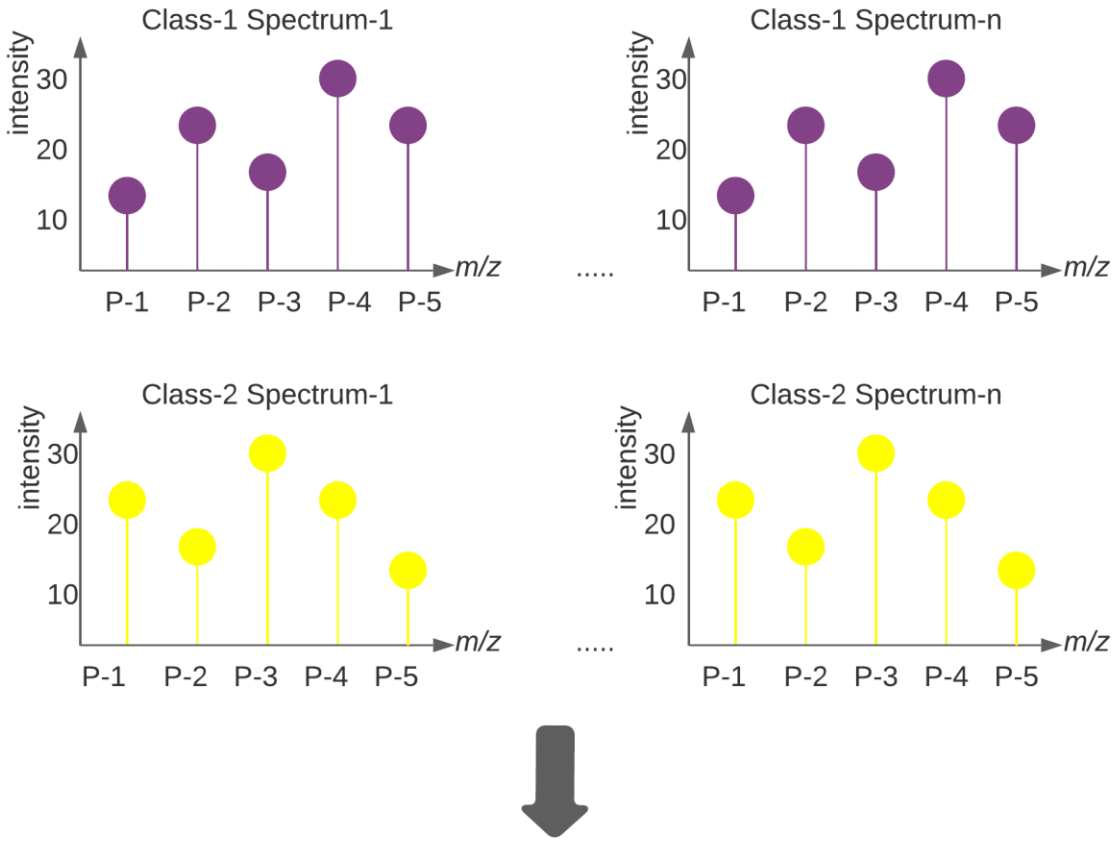
f) Noise reduction: the spectra were denoised using least-squares polynomial with a window of 35Da and a 2-degree polynomial function.

g) Alignment: the spectra were aligned using the peaks found at least 30% of the samples as references.

h) Peak detection: over-segmentation on *m/z* axis set as 20Da. The height filter on intensity was set as 5, 10 and 1 for Chapters 4, 5 and 6, respectively.

2.5 Spectral Features

Statistical analyses (Anderson-Darling test, Welch's *t*-test and Wilcoxon test) similar to ClinProTools 3.0 were performed on the peaks which met the criteria of pre-processing (Bruker Daltonics, 2011). Statistical tests were computed based on the peak intensities of those present in at least 30% of all spectra. The Anderson-Darling test was used to check the normality of data distribution (Anderson and Darling, 1954). Peak selection was performed based on statistical tests as follows: Welch's *t*-test for those showing the normal distribution and Wilcoxon test for others. The data provided to the ML algorithms, after peak selection, look like as shown in Figure 2-1. The selected peaks were further pre-processed to have zero mean and unit variance. Such peaks represented the spectral features used in the classification analysis.



Class	Peak-1	Peak-2	Peak-3	Peak-4	Peak-5
Class-1 Spectrum-1	15.8	22.3	18.5	31.1	22.7
...	...	...	...	...	...
Class-1 Spectrum-n	15.7	22.4	18.4	31.2	22.6
Class-2 Spectrum-1	22.5	18.3	30.7	21.9	14.7
...	...	...	...	...	...
Class-2 Spectrum-n	22.7	18.1	31.0	22.2	14.5

Figure 2-1. Illustration of the MALDI-TOF spectra which were used to train machine learning algorithms.*Binary classification problem is solved by providing the intensity values of selected peaks in each spectrum of the classes (Class-1 and Class-2). In this illustration, five peaks (shown as P-1, P-2, P-3, P-4 and P-5) are selected based on selection criteria and their intensity values are used by the algorithms. These figures are generated using Lucidchart.com.*

2.6 Resampling for the Imbalanced Datasets

Binary classification models are generally not designed for skewed data and perform better in balanced classes (Zhang, 2010; He and Garcia, 2009). In the binary classification of benzylpenicillin/multidrug-resistant and susceptible *S. aureus* isolates (Chapter 5), the datasets in the susceptible isolates holding class is more frequent than the resistant isolates holding one. This may cause a bias towards the class holding susceptible isolates. Therefore, a resampling approach was used to cope with the imbalanced dataset issue. In ML, there are four different approaches to cope with imbalanced datasets, which are undersampling, oversampling, the combination of undersampling and oversampling techniques, and ensemble learning (Lemaître, Nogueira and Aridas, 2017). The undersampling technique decreases the number of samples from the majority class as seen in Figure 2-2-A. Undersampling approaches can be categorized under fixed undersampling and cleaning undersampling. Fixed undersampling is a quick and simple way which only aims to reduce the number of samples from the majority class to ensure an appropriate ratio between classes. In this chapter, the fixed undersampling approach (random under-sampler from imblearn library) (Lemaître, Nogueira and Aridas, 2017) was used to balance benzylpenicillin/multidrug-resistant and susceptible classes of *S. aureus.* Cleaning undersampling instead cleans the data points according to specific empirical criteria such as Tomek's links (Tomek, 1976), edited nearest neighbours (Wilson, 1972), condensed nearest neighbours (Hart, 1968), instance hardness threshold (Smith, Martinez and Giraud-Carrier, 2014) etc. As these techniques have not been used in this study, they will not be discussed in detail.

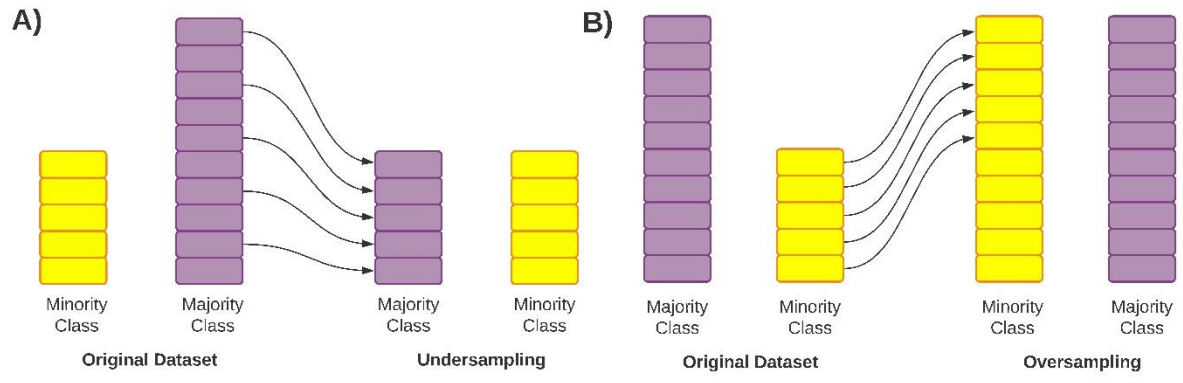


Figure 2-2. Schematic illustration of fixed undersampling and oversampling approaches. ***A)*** *In fixed undersampling, data points of the majority class are selected randomly to ensure the appropriate ratio (1:1) with the minority class. B) In random oversampling, data points of the*

minority class are repeated randomly to ensure the appropriate ratio (1:1) with the majority class. These figures were generated using Lucidchart.com.

In the binary classification of antimicrobial-resistant and susceptible *E. faecalis* and *E. faecium* isolates (Chapter 6), the datasets in the susceptible isolates holding class and the resistant isolates holding one were not balanced. This may cause a bias towards the majority class in the analyses. Therefore, the oversampling approach was used to cope with the imbalanced dataset issue. There are different oversampling algorithms such as random oversampling, oversampling by using ADASYN and SMOTE. The random oversampling technique creates new samples by random repeating the existing dataset in favour of the class that is less frequent (minority class) (Figure 2-2-B). This technique allows representation of the classes (both major and minor) are balanced. Hence, the bias of the decision boundary towards the more frequent class (majority class) is lessened (He *et al.*, 2008). Adaptive synthetic sampling approach (ADASYN) (He *et al.*, 2008) and synthetic minority oversampling technique (SMOTE) (Chawla *et al.*, 2002) offers more advanced oversampling approaches by employing heuristic technique instead of repeating the existing dataset. As these techniques have not been used in this study, they will not be discussed in detail. In Chapter 6, random oversampling (random over-sampler from imblearn library) was used to balance classes (Lemaître, Nogueira and Aridas, 2017).

2.7 Classification Methods

The performance of the classifiers - LR, LSVM, RBF SVM, DT, RF, MLP NN, NB, AdaBoost, LDA and QDA - was investigated using the scikit-learn library in Python (Pedregosa *et al.*, 2011). NB, LDA and QDA do not have hyperparameters. For other classifiers, hyperparameters were optimized by using the following set of values:

- LR: inverse of regularization strength C = [0.001, 0.01, 0.1, 1, 10, 100, 1000].
- LSVM: penalty parameter of the hinge loss error C = [0.001, 0.01, 0.1, 1, 10, 100, 1000].
- RBF SVM: γ (RBF kernel coefficient) = [0.0001, 0.001, 0.01, 0.1] and C (penalty parameter) = [0.001, 0.01, 0.1, 1, 10, 100, 1000].
- DT: maximum depth of the tree = [10, 20, 30, 50, 100].
- RF: number of estimators = [2, 4, 8, 16, 32, 64].
- MLP NN: α (regularization term) = [0.001, 0.01, 0.1, 1, 10, 100], learning rate (initial learning rate used to control the step size in updating the weights with Adam solver) = [0.001, 0.01, 0.1, 1] and hidden layer sizes = [10, 20, 40, 100, 200, 300, 400, 500].

- AdaBoost: number of estimators = [2, 4, 8, 16, 32, 64].

2.8 Prediction Performance

The prediction performance of each classifier was evaluated by considering the following indicators, assuming P and N as the total number of positive and negative classes (see Table 2-1) and using T for true (correct) and F for false (wrong) predictions:

- Sensitivity (True Positive Rate) = TP / P
- Specificity (True Negative Rate) = TN / N
- Accuracy = (TP+TN)/(P+N)
- Cohen's Kappa statistic = $(p_o - p_e)/(1-p_e)$

  where p_o= (TP+TN)/(P+N) and p_e= (P*(TP+FN) + N*(FP+TN)) /$(P+N)^2$

Table 2-1. Positive and negative classes in the analyses.

Analysis	Positive Class	Negative Class
Chapter 3	Contagious *S. uberis*	Environmental *S. uberis*
Chapter 4	Subclinical *E. coli*	Clinical *E. coli*
	Persistent *E. coli*	Non-persistent *E. coli*
Chapter 5	Resistant *S. aureus*	Susceptible *S. aureus*
Chapter 6	Resistant *E. faecalis*	Susceptible *E. faecalis*
	Resistant *E. faecium*	Susceptible *E. faecium*

2.9 Performance Analysis

Nested Cross-validation (NCV) (Cawley and Talbot, 2010) was employed to assess the performance and select the hyperparameters of the proposed classifiers. In NCV, there is an outer loop that splits the data points into test and training sets (see Figure 2-3). For each training set, a grid search (inner loop) is run, to find the best hyperparameters of the classifier by using accuracy. Then, the test set is used to score the best hyperparameters found in the inner loop, showing how well the model performs on unseen data points. Thirty iterations were carried out, wherein each iteration an NCV was employed. The inner loop of the NCV finds the best

hyperparameters of each classifier (when suited) using stratified 3-fold cross-validation; the outer loop measures the accuracy, sensitivity, specificity and kappa by using 5-fold stratified cross-validation (in Chapter 4, 4-fold was used due to data size), to compare all the classifiers (Figure 2-3).

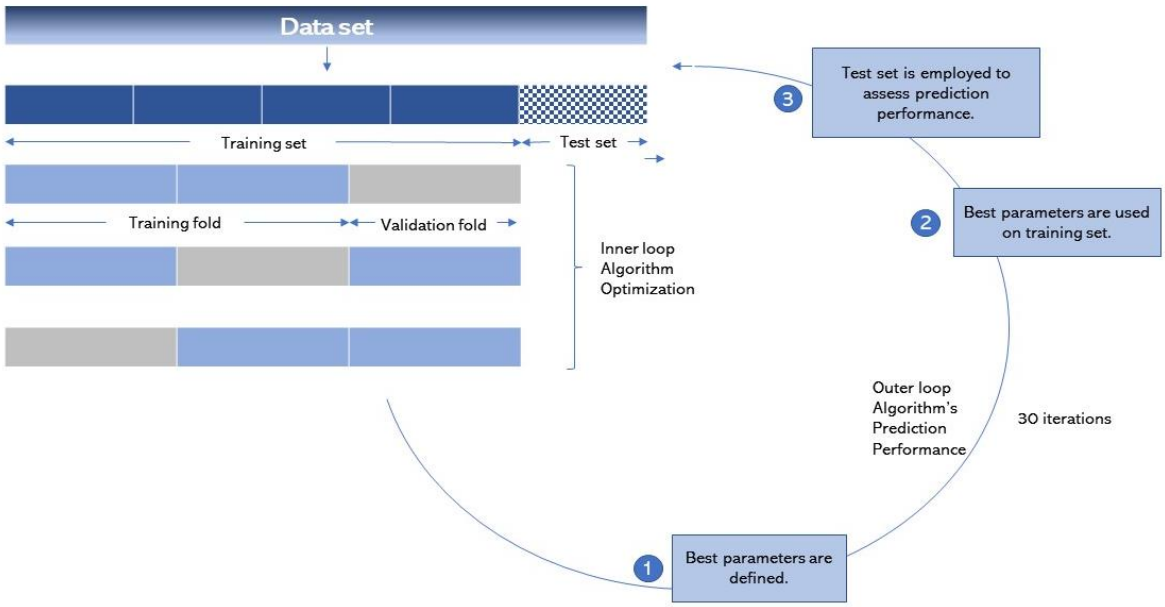


Figure 2-3. Nested Cross-Validation (NCV) loop. *The data points are first divided into 5-folds by the outer loop, where 1-fold is the test set and 4-folds are the training set. The training set is then divided into 3-folds by the inner loop, where 1-fold is the validation and the rest are the training folds. The hyperparameter grid search is realized in the inner loop. Accordingly, the model is trained and tested.*

2.10 Biomarker Characterisation

A dedicated bioinformatics pipeline was developed to find correspondences between individual peaks and actual proteins of organisms. First, amino acid sequences of the proteomes listed in Table 2-2, were retrieved from the PATRIC database in FASTA format. The molecular weights of the proteins were calculated using the Compute pI/Mw tool on ExPASy (Walker, 2005). The proteins were filtered in the range of ± 200Da of the mass of individual peaks as initial methionine cleavage, phosphorylation, additional and removal of molecules or isotopes of elements may change the molecular weight of a protein (Coombes, Baggerly and Morris, 2007; Bonissone *et al.*, 2013). N-terminal methionine cleavage was predicted using the online prediction tool TermiNator (Frottin *et al.*, 2006) and the theoretical molecular weights of the proteins were re-calculated using compute pI/Mw tool according to presence or absence of the initial methionine. Finally, the proteins which fell within a maximum of 0.2% difference as

molecular weight were cross-matched with the individual peaks as in previous studies (Yasui *et al.*, 2003; Zou *et al.*, 2011).

Table 2-2. Model organisms used in the analyses.

Analysis	Organism
Chapter 3	*S. uberis* 0140J (Ward *et al.*, 2009)
Chapter 4	*E. coli* genomes that were sequenced in the current work *E. coli* strain MG1655 for PPI analysis
Chapter 5	*S. aureus* Newbould 305 (ATCC 29740) (Bouchard *et al.*, 2012) *S. aureus* strain NCTC 8325 / PS 47 for PPI analysis
Chapter 6	*E. faecalis* strains isolated from bovine in PATRIC db *E. faecalis* ATCC 29212 (Minogue *et al.*, 2014) for PPI analysis *E. faecium* strains isolated from bovine in PATRIC db *E. faecium* C68 (García-Solache and Rice, 2016) for PPI analysis

To further investigate the function of the identified proteins, PPI was studied as previously described (Esener *et al.*, 2018). The PPI datasets of the organisms were obtained from the STRING database (Szklarczyk *et al.*, 2018) and nodes (proteins) with interaction scores lower than medium confidence level (interaction scores <0.400) were filtered out. The remaining nodes (proteins) were analysed in Cytoscape 3.7.1 (version 3.6.1 was used in Chapter 3) based on the following parameters: the average number of neighbours, clustering coefficient, network density and network heterogeneity (Shannon *et al.*, 2003; Ravasz *et al.*, 2002; Dong and Horvath, 2007).

Resistant genes of the antimicrobial classes that were available in ResFinder v3.1 (Zankari *et al.*, 2012) were obtained and used as queries in a comparative BLAST search against the proteomes in Chapters 5 and 6. Gene functions were annotated by GO terms (biological process, molecular function and cellular component) (Ashburner *et al.*, 2000) and KEGG pathways (Kanehisa and Goto, 2000) in which they were involved. Finally, to gain a more in-depth understanding of the protein functions, homology and threading 3D models for discriminant proteins were built. 3D homology modelling was performed for the proteins with good quality templates (identity ≥30% (Xiang, 2006)) in the Swiss-Model repository (Waterhouse *et al.*, 2018) by using Swiss-PdbViewer (Guex, Peitsch and Schwede, 2009). 3D models of those with

no homology or bad quality templates (identity<30%) were generated by using the threading technique of I-TASSER, where GO functions were predicted as well (Yang and Zhang, 2015). 3D models of all discriminant proteins were visualised and edited in UCSF Chimera (Pettersen *et al.*, 2004).

Homologs of the discriminant proteins were checked in the NCBI database by position-specific iterative basic local alignment tool (PSI-BLAST) (Schäffer *et al.*, 2001). Functional domains were searched against the CDD v3.17-52910 PSSMs (Lu *et al.*, 2020), PFAM (El-Gebali et al., 2018) and SMART databases (Letunic and Bork, 2018). PSORTb v3.0 was used to predict cellular locations of the discriminant proteins (Yu *et al.*, 2010).

CHAPTER 3 DISCRIMINATION OF CONTAGIOUS AND ENVIRONMENTAL STRAINS OF *STREPTOCOCCUS UBERIS* IN DAIRY HERDS BY MEANS OF MASS SPECTROMETRY AND MACHINE LEARNING

This chapter was published in Scientific Reports (https://www.nature.com/articles/s41598-018-35867-6) with the title above by Necati Esener, Martin J. Green, Richard D. Emes, Benjamin Jowett, Peers L. Davies, Andrew J. Bradley & Tania Dottorini. Although this chapter was organised in standard thesis outline (introduction-methods-results-discussion) instead of published format (introduction-results-discussion-methods), most of the content was left untouched. The authors' contributions were as follows: MJG and AJB provided the original data. TD, MJG, RDE and AJB conceived and designed the data analysis procedures. NE, RDE, TD, and BJ carried on the data analysis. TD, NE and PLD wrote the manuscript. All authors reviewed the manuscript.

In **Chapter 3**, the main aim was to inspect MALDI-TOF MS data with ML as a method to discriminate bovine mastitis-causing *S. uberis* isolates based on their transmission dynamics; contagious and environmental. To this end, the commercial software, ClinProTools, was employed for data preparation (MALDI-TOF spectra) which focused on removing 'noise' from the dataset to increase the chance of classification based on solely biological information. Preprocessed data were analysed with three supervised ML algorithms that were available in the ClinProTools software; GA, SNN and QC. These classifiers were run on the MALDI-TOF MS data both coming from individual farms (intra-farm) and overall representation of the country (inter-farm). GA was shown to outperform other classifiers, where the prediction performance was much better for intra-farm analysis rather than inter-farm.

3.1 INTRODUCTION

Clinical mastitis is one of the most important challenges facing the dairy industry, where it reduces productivity, profitability and cow welfare. Considerable progress has been made in understanding the epidemiology and microbiology of mastitis over the past four decades, identifying the physical origins of infection as contagious or environmental (Todhunter, Smith and Hogan, 1995) and temporal origins of the infection e.g. dry period or lactation (Green *et al.*, 2007). Coliform bacteria are almost always environmental, whilst other pathogens such as *S.*

aureus and *S. agalactiae* are typically contagious on the contrary, *S. uberis* can commonly manifest itself in both contagious and environmental forms (Zadoks and Fitzpatrick, 2009; Zadoks *et al.*, 2011), and the ability to diagnose the clinical mastitis transmission pattern (contagious or environmental) is an essential step to identify appropriate and effective management interventions for the control of the disease at a herd level. Limited financial and labour resources are typically available to farmers for mastitis control. A role of the clinician is to identify the mode of transmission as early as possible, to react appropriately before costly production losses have occurred. Currently, clinicians assess the most likely mode of transmission through analysis of historical data (Davies *et al.*, 2016), visual observation of management practices (milking, cleaning, etc.) and knowledge of pathogens, although there is limited evidence that the latter two methods are useful in determining transmission patterns in the modern dairy herd. This leads to inevitable delays and associated losses before a diagnosis of a new, emerging disease pattern can be made. Prompt diagnosis of the likely transmission route in case of an outbreak would allow appropriate control interventions to be implemented earlier and reduce deleterious production and welfare consequences of additional clinical mastitis cases.

Previous studies in this field have used genomic epidemiological techniques to classify individual bacterial strains broadly as contagious or environmental according to their observed patterns of clinical disease within multiple independent herds (Davies *et al.*, 2016). These techniques are useful as research tools to understand the observed patterns but are too costly and laborious to be practical clinical tools for clinicians. The discriminatory ability for genomic techniques such as MLST may also not be appropriate for the classification of bacterial isolates according to their clinical manifestation if those attributes which govern the transmission behaviour of an isolate are determined by epigenetic factors or conferred by mobile genetic elements which are rapidly exchanged between bacteria such as *S. uberis* (Casadesús and Low, 2006).

Evidence of epigenetic strain variation and strain evolution within a bacterial species has been described by several mechanisms, such as differential methylation resulting in phase variation (Seib *et al.*, 2015) as a means for isolates of commensal and pathogenic bacterial species to adapt to new or changing environments. In human cases of Salmonellosis, epigenetic strain variation in virulence and host-pathogen interaction could be demonstrated by proteomic analysis where genomic discrimination of strains was not possible (Badie *et al.*, 2007). When identifying the route of transmission of an individual bacterial isolate from a case of bovine mastitis,

a rapid proteomic technology able to characterise variation in antigenic expression related to virulence, such as surface protein molecules, may be more discriminatory than existing genomic techniques (Baseggio *et al.*, 1997).

Gel-based proteomics techniques for strain differentiation have been used for decades in many species and disease processes including bovine mastitis (Kallow *et al.*, 2006). More recently, highly discriminatory techniques, such as MALDI-TOF MS have increased our ability to investigate the molecular epidemiology and host-pathogen interactions of many bacterial pathogens (Nakano *et al.*, 2015).

In contrast to genomic techniques, MALDI-TOF MS provides a rapid and economic means of identifying bacteria and is capable of strain differentiation within a bacterial species such as *S. pneumoniae, Y. entrocolitica* and *M. pneumoniae* (Xiao *et al.*, 2014; Rizzardi, Wahab and Jernberg, 2013; Barreiro *et al.*, 2010).

Comparison of MALDI-TOF MS proteomic profiles may allow discrimination between bacterial, isolate strains of a pathogen, such as environmental and contagious *S. uberis* strains which have acquired or evolved enhanced survival or colonisation characteristics (genetic or epigenetic) that increase the risk of a cow to cow transmission.

The primary aim of this study was to investigate MALDI-TOF MS data with ML as a method to discriminate between *S. uberis* isolates with different modes of transmission; contagious and environmental. A secondary aim of the study was to compare strain differences within and between farms of the UK. The final aim of the study was to identify proteins related to the differentiating peaks between transmission routes with bioinformatics tools.

3.2 METHODS

3.2.1 Data Source

The data for the present study was obtained from the previous large-scale study (Green et al., 2007) of UK dairy herds within the scope of the control plan of mastitis. The 52 farms were enrolled in the National Milk Records database on the criterion that none of them had an incidence of fewer than 35 cases of clinical mastitis per 100 cows per year. All study farms kept the animals housed during wintertime, while the seasonal and all year long calving herds were distributed equally within the same groups. Therefore, the selection and grouping of the farms represent the characteristics of commercial farms across England and Wales. The incidence

rate of clinical mastitis in these 52 farms were 66 per 100 cows per year as median and 75 per 100 cows per year as mean. On average, 28% of the clinical cases were caused by *S. uberis*, ranging from 7% to 64% in individual farms.

Milk samples of all mastitis cases were collected in a 14-month-time period starting from March 2004 till April 2005. Standardised operating procedure (Green *et al*., 2004) was applied to mastitis diagnosis and sample collection by farm employees. Culturing of the samples were executed by a commercial milk laboratory QMMS, where standard bacterial identification was applied (Bradley *et al*., 2007).

3.2.2 MLST and MALDI-TOF Datasets

The original MALDI-TOF raw spectra had been obtained in the previous study using Bruker Microflex instrument, Flex Control version 3.4 (Davies *et al.*, 2016). To better appreciate the nature of the available data, the following information is provided from that previous work.

MLST: The gDNA was extracted for MLST sequencing following the protocol described in previous literature (Leigh *et al.*, 2010). Clinical cases attributed to isolates of the same MLST occurring in different cows in the same herd within a 42-day time period were classified as contagious, whereas cases attributed to isolates occurring only once in any cow of any herd were classified as environmental.

In this work, to extract each peak list, the following steps were applied in ClinProTools 3.0:

a) baseline subtraction: using the Top Hat baseline (minimal baseline width: 10%) (Serra, 1983);
b) normalization: to the total ion count, leading to spectral intensities in the [0-1] range (Bruker Daltonics, 2011);
c) recalibration of the *m/z* values: using as reference masses those appearing in at least 30% of the spectra and setting 1000 ppm as the maximal peak shift (Bruker Daltonics, 2011);
d) total average spectrum calculation: using weighted contributions from the available replicates (Morris *et al.*, 2005);
e) average peak list calculation: the calculation was applied to the total average spectrum rather than on every single spectrum (Bruker Daltonics, 2011);
f) peak picking on the total average spectrum: using resolution: 800 for spectrum smoothing, signal to noise threshold: 5.00, and 0.000% relative threshold base peak to include all peaks (Bruker Daltonics, 2011);

g) peak normalization: to give the same relevance (weight) to all peaks within the classification/prediction models (Bruker Daltonics, 2011);
h) mass range filter: the mass range of the spectra was limited to 4-10kDa.

3.2.3 Classification Methods

To verify if MALDI-TOF peak lists associated with isolates could be used to predict their contagious or environmental nature, supervised ML technologies were used to implement classifiers, i.e. software systems that, once provided with a spectrum as the input, would respond by predicting the most likely class (i.e. environmental or contagious) for the isolate. Being based on supervised learning, all methods required the availability of training datasets for model building (Russell and Norvig, 2010) (i.e. peak lists with the known associated classification of environmental or contagious from MLST), and validation datasets for assessing the performance of the classifier. The following classification methods available in the Bruker mass spectrometry analysis software ClinProTools 3.0 were tested: GA, SNN and QC.

The GA method uses the training datasets to identify a subset of peaks shared by all of the peak lists (referred to as "peak combination"), acting as the most effective discriminator between contagious and environmental isolates. The performance of each tentative combination in terms of discrimination effectiveness is assessed by evaluating the degree of separation of the clusters formed by the known contagious and environmental isolates, once that combination of peaks is considered. The identification of the best combination is treated as an optimisation problem and solved via GA (Holland, 1975). When the best peak combination is found (end of training), new peak lists submitted to the classifier are predicted as being contagious or environmental by extracting the peak combination and using it to determine which one of the two clusters is closest to the observation using the k-NN metric (Bruker Daltonics, 2011).

The SNN method is based on building a classifier powered by a NN implementing a modified version of the supervised relevance neural gas algorithm (Hammer, Strickert and Villmann, 2005). The peak lists of the training set are investigated by the algorithm, to identify "prototype" lists suitable to act as representative for the corresponding class (contagious or environmental). Once trained, the network can be fed with any new peak list; the prediction will be based on understanding which prototypes are closer to the given list (Bruker Daltonics, 2011).

The QC method is based on grouping the peak lists into two classes (environmental and contagious), generating peak averages representative of each class, and ranking the relevance of

the peaks when acting as discriminators based on statistical testing. Any new observation is then tested on similarity against the weighted averages of each class, the most similar class being elected as the prediction result (Bruker Daltonics, 2011).

3.2.4 Parameters Used for the Classification Methods

GA default parameters (as suggested by the ClinProTools software) were as follows: initial number of peak combinations (INPC): automatic detection; maximal number of peaks (MNP, maximum number of peaks to be included in the combination): 5; maximum number of generations (MNG, number of GA iterations to identify the optimal result): 50; mutation rate (MR): 0.2; crossover rate (CR): 0.5 (mutation and crossover control how new candidate combinations are created starting from those tested in the current generation); number of neighbours for k-NN (number of neighbours considered by the k-NN method to determine the distance of a new observation from an existing class): 5. The optimised set of GA parameters was: INPC: 125, MNP: 19, MNG: 50, MR: 0.2, CR: 0.5 and k-NN: 3.

SNN parameters were MNP: automatic detection; upper limit of cycles (ULC): 1000; number of prototypes (NP): automatic detection.

QC method had only one controllable parameter: MNP was set to automatic detection.

3.2.5 Prediction Performance

The performance of the classifiers was assessed with the following metrics:

- recognition capability (RC): the accuracy obtained when the classifier is trained with the entirety of the dataset and tested on the same data.
- accuracy, kappa and other indicators from cross-validation (CV): the dataset is split into a training subset (containing a percentage of the available spectra) and a testing subset (containing the remaining spectra). The subsets are created by randomly drawing individuals from the same dataset. Accuracy, kappa and any other desired indicator resulting from the confusion matrix are computed. The entire process is repeated n times (training and testing n classifiers), then the final indicators are obtained as the arithmetic means of the values obtained on the n confusion matrices.
- accuracy, kappa and other indicators from external validation (EV): only one classifier is built, by using the entirety of the available dataset for training. Testing is performed using a dataset separates from the training dataset. The indicators are computed on the resulting confusion matrix.

3.2.6 Methods for Cross and External Validation

In the intra-farm analysis, classifiers were developed to operate exclusively within each farm. The spectra of the 19 farms selected as suitable for model building were used to implement 19 separate classifiers (one per farm). Only data pertaining to the specific farm were used to implement and validate each classifier. RC and CV performance indicators were computed for each farm, but no external validation was performed. CV of each classifier was performed using 80% of the available spectra for training and the remaining 20% for validation. The procedure was repeated 10 times, each time randomising the extraction of spectra for the training set.

For inter-farm analysis CV of the global classifier was performed using 80% of the spectra obtained by aggregation of data from the 19 farms for training and the remaining 20% for validation. The procedure was repeated 10 times, each time randomising the extraction of spectra for the training set. For external validation, 100% of the available aggregated spectra were used for training, and the spectra from the 10 holdout farms were used for validation. The entire procedure was repeated 10 times for the GA classifier, because of the random components present in the method.

3.3 RESULTS

3.3.1 Data source

In this work, we first assessed the geographical distribution of clinical cases. The results are shown in Figure 3-1, indicating spread in England and Wales, with a higher concentration towards the south, and no herds in Scotland.

Figure 3-1. Location of the enrolled farms on the map of the United Kingdom. *a) The entire set of 52 farms b) the 19 farms selected for building the model for intra-farm analysis. The red colour represents the environmental isolates of Streptococcus uberis while the green is for contagious ones. The size of the circle indicates the number of Streptococcus uberis isolates in the farms. The figures were generated in R (R Core Team, 2019) using the sp (Pebesma and Bivand, 2005), mapdata (Deckmyn, 2018) and mapplots (Gerritsen, 2014) packages.*

To construct predictors/classifiers, we focused on herds containing both environmental and contagious *S. uberis* isolates. Thus, by looking at the available MLST data, we selected for the study only the 29 farms/herds containing both. The 23 eliminated herds consisted of 2 without any *S. uberis* isolates, 4 containing only contagious, 13 containing only environmental and 4 containing unclassified isolates (see Figure 3-2). Amongst the 29 herds selected, 10 of them were reserved for external validation (holdout group) as they each featured <20 MALDI-TOF spectra, too few to be useful for the generation of effective classifiers. The remaining 19 farms (Figure 3-1b) were considered suitable for the development of classification models.

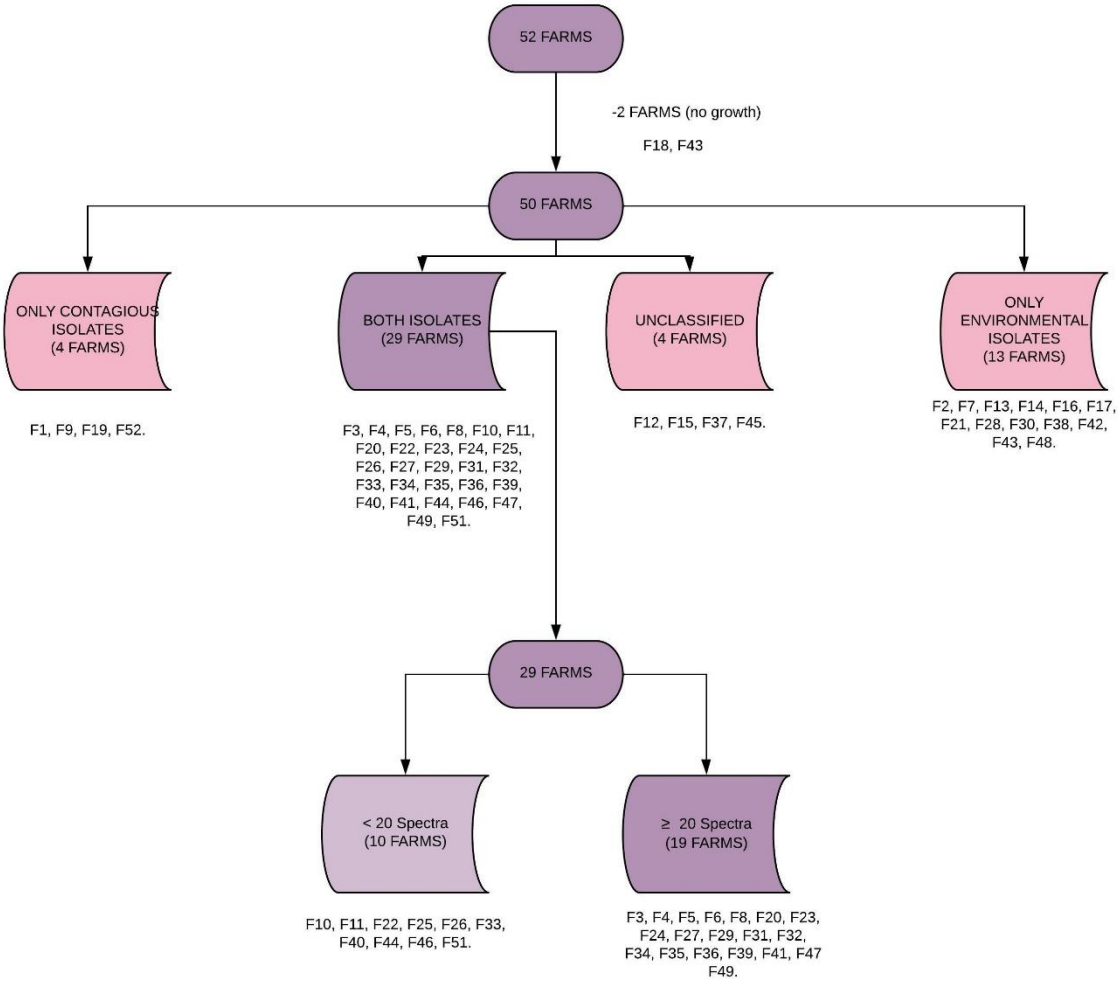


Figure 3-2. Process of initial farm selection and farm codes. *The categorisation was done according to the type and presence of Streptococcus uberis (contagious and environmental) strains and the number of MALDI-TOF spectra. This figure was generated using Lucidchart.com.*

3.3.2 Intra-Farm Analysis

The results of the intra-farm analysis on the 19 selected farms are shown in Figure 3-3. Classifiers were run with default settings as described in the Methods section. RC, CV accuracy and kappa are shown as arithmetic means computed from the individual results of the 19 farms (GA: RC= 100.00%, CV accuracy= 97.81%; CV kappa= 93.72%, sensitivity= 97.13%, specificity= 96.26%. SNN: RC= 84.00%, CV accuracy= 82.17%, CV kappa= 60.20%, sensitivity= 87.22%, specificity= 72.65%. QC: RC= 97.32%, CV accuracy = 91.34%, CV kappa= 80.20%, sensitivity= 92.36%, specificity= 87.20%.).

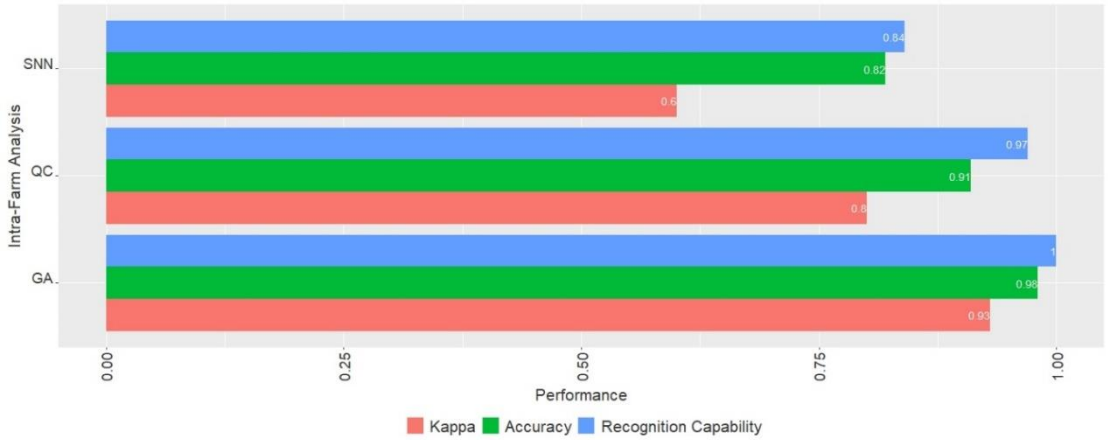


Figure 3-3. Comparison of intra-farm analysis results of 19 farms using Genetic Algorithm (GA), Supervised Neural Network (SNN) and QuickClassifier (QC). *All the results were obtained by adopting the default settings for the classifying methods. This figure was generated using R package ggplot2 (Wickham, 2011).*

3.3.3 Inter-Farm Analysis

In inter-farm analysis, a single classifier was developed and trained on the aggregated data from all the available 19 farms. RC, and performance indicators from CV were computed using the data from the 19 farms. The RC and CV results for the inter-farm analysis are shown in Figure 3-4. Classifiers were run with default settings (GA: RC= 89.02%, CV accuracy= 75.25% CV kappa= 50.59%, sensitivity = 81.75%, specificity =69.16%. SNN: RC=57.51%, CV accuracy= 50.11%, CV kappa= 0.65%, sensitivity= 64.09%, specificity= 36.55%. QC: RC= 57.92%, CV accuracy= 54.02%, CV kappa= 7.71%, sensitivity= 66.97%, specificity= 40.59%).

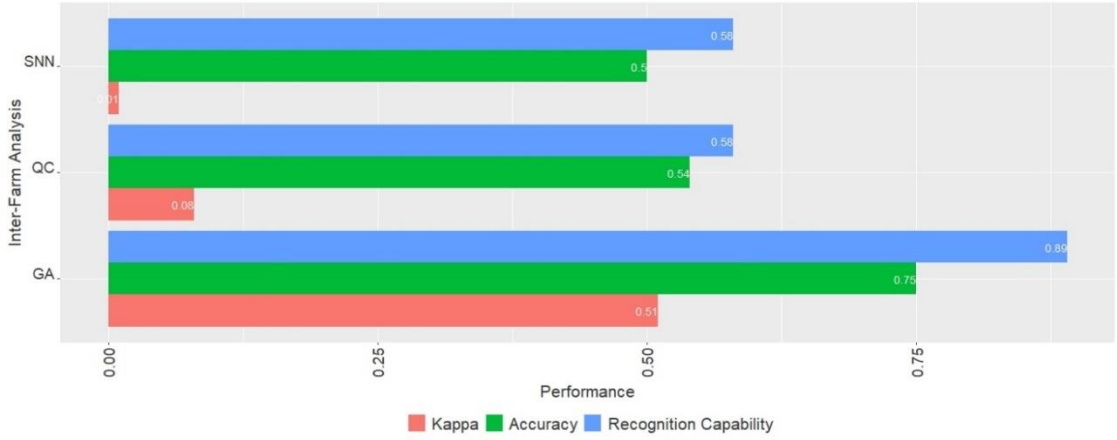


Figure 3-4. Comparison of inter-farm analysis results of 19 farms using Genetic Algorithm (GA), Supervised Neural Network (SNN) and QuickClassifier (QC). *Inter-farm analysis results are the arithmetic mean of the results from nineteen classifiers (one per farm). All the results were obtained by adopting the default settings for the classifying methods. This figure was generated using R package ggplot2 (Wickham, 2011).*

GA was selected for further optimisation, due to better performance. With the optimised settings (see Methods), the performance of the GA classifier went from 89.02% to 99.09% for RC, from 75.25% to 95.88% for CV accuracy, and from 50.59% to 92.62% for CV kappa. The probability distributions of the performance indicators for intra-farm (19 separate classifiers) and inter-farm (single classifier on the aggregated data from 19 farms) analysis are reported in Figure 3-5.

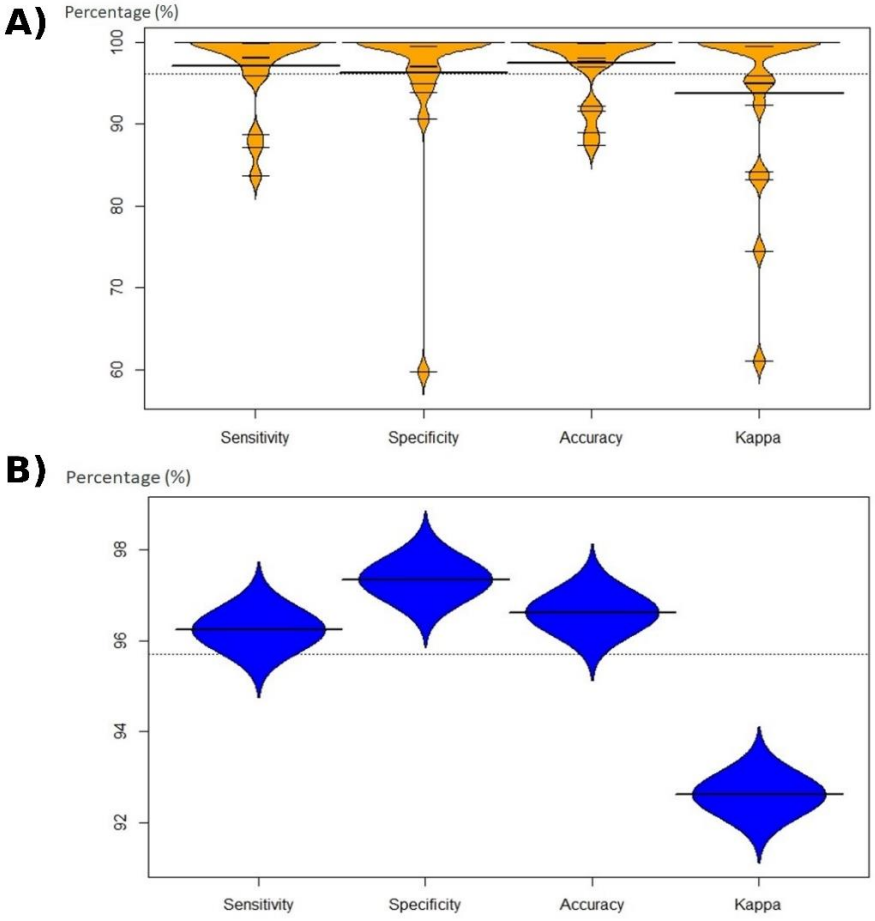


Figure 3-5. Distribution of the performance indicators for analysis performed by Genetic Algorithm. *a) intra-farm cross-validation; b) inter-farm cross-validation. Data from 19 farms. These figures were generated using R package beanplot (Kampstra, 2008).*

External validation was performed on the GA classifier, using the additional 10 farms in the holdout group. The following EV indicators were obtained: sensitivity 82.07%, specificity 50.00%, accuracy 70.67% and Cohen's kappa 33.80%. The probability distributions of the performance indicators for inter-farm external analysis are reported in Figure 3-6.

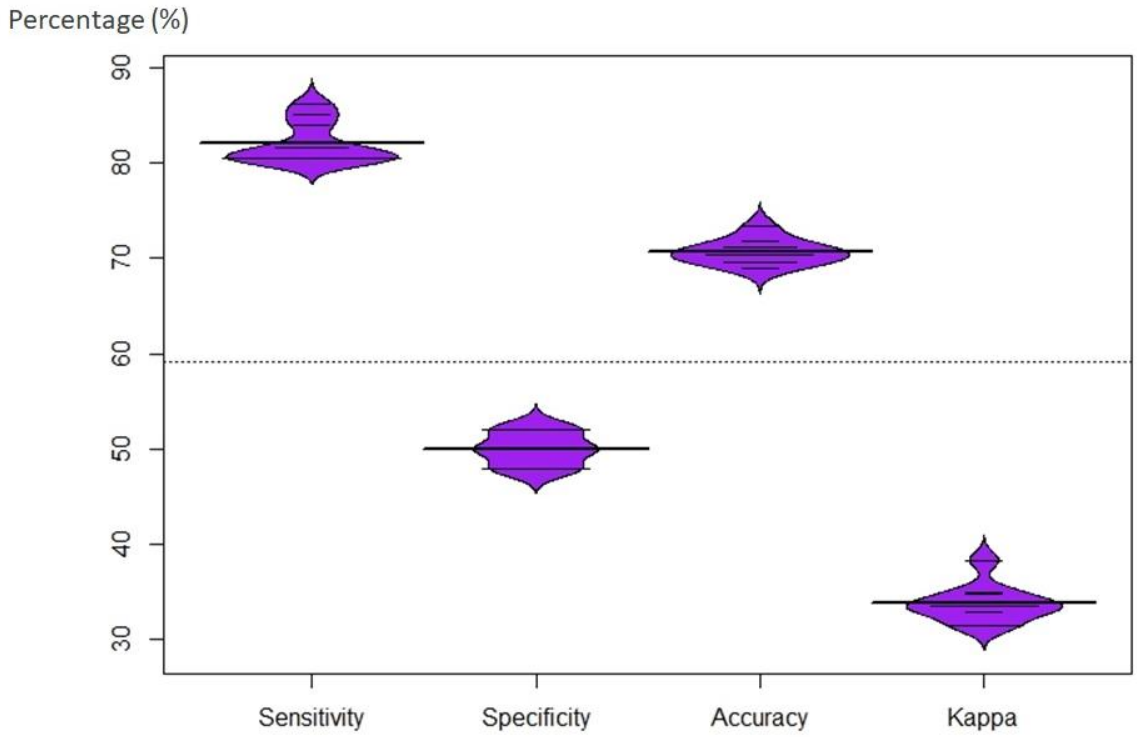


Figure 3-6. Distribution of the performance indicators for inter-farm external validation analysis performed by Genetic Algorithm. *This figure was generated using R package beanplot (Kampstra, 2008).*

3.3.4 Biomarker Characterisation

GA classifier for the intra-farm analysis identified a set of 19 peaks shared by all the isolates, providing optimal discrimination between environmental and contagious isolates. The masses of these peaks were then compared with the molecular weights of *S. uberis* proteins in the NCBI database and 7 out of 19 could be matched to 8 proteins from the proteome (in one case, the molecular weight of a peak was close to two different proteins) (Figure 3-7).

Five of the eight proteins have a known function (according to NCBI): two of them are ribosomal proteins, two of them bacteriocins and one is an ATP synthase protein. The remaining three proteins had unknown functions; two of these were hypothetical and one of them was of an unknown domain. Using the SMART database, the domains of the 8 proteins were found. The predicted three-dimensional models are shown in Figure 3-7.

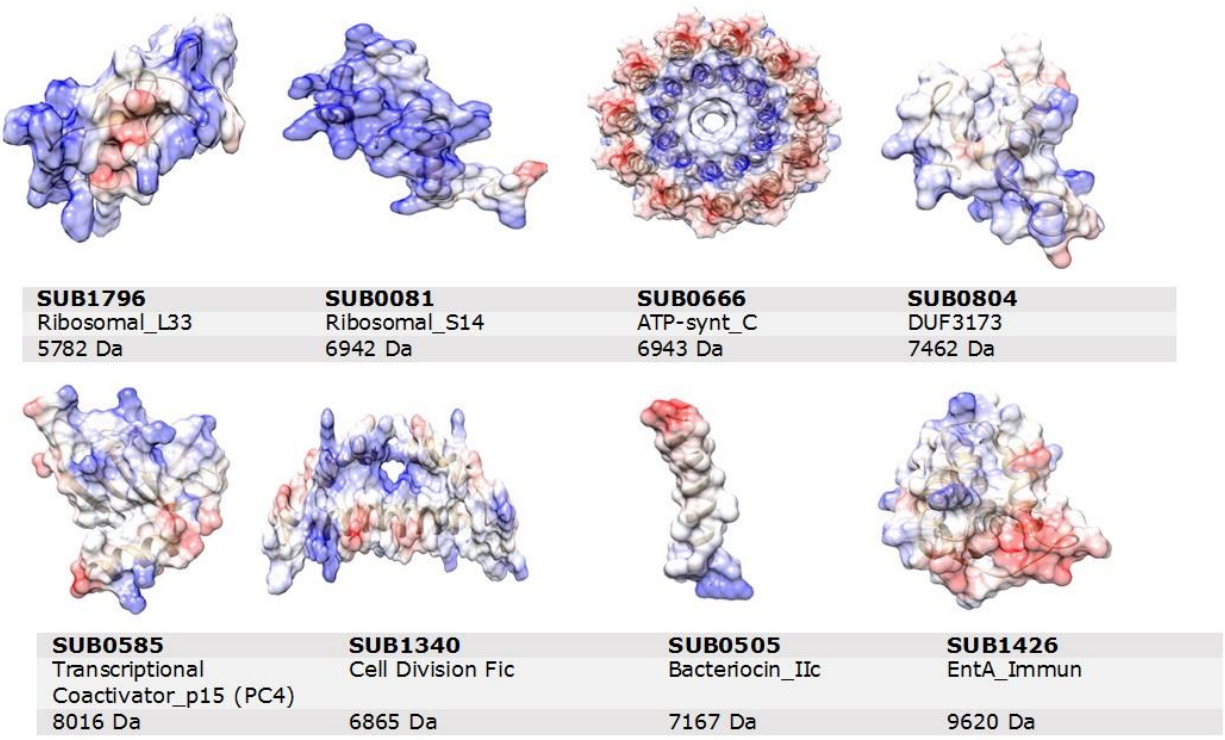


Figure 3-7. Selected proteins of Streptococcus uberis. *Top to bottom: 3D protein structure, protein ID, domain of the protein and molecular weight of the protein. The visualisation was carried out with UCSF Chimera.*

The analysis of the PPI network for the discriminant proteins in *S. uberis* (Figure 3-8) showed that 5 out of 8 proteins (SUB1426, SUB0666, SUB0081, SUB1796 and SUB0585) share common first neighbour proteins. Interestingly, SUB0666, SUB0081 and SUB1796 were also found to interact with each other.

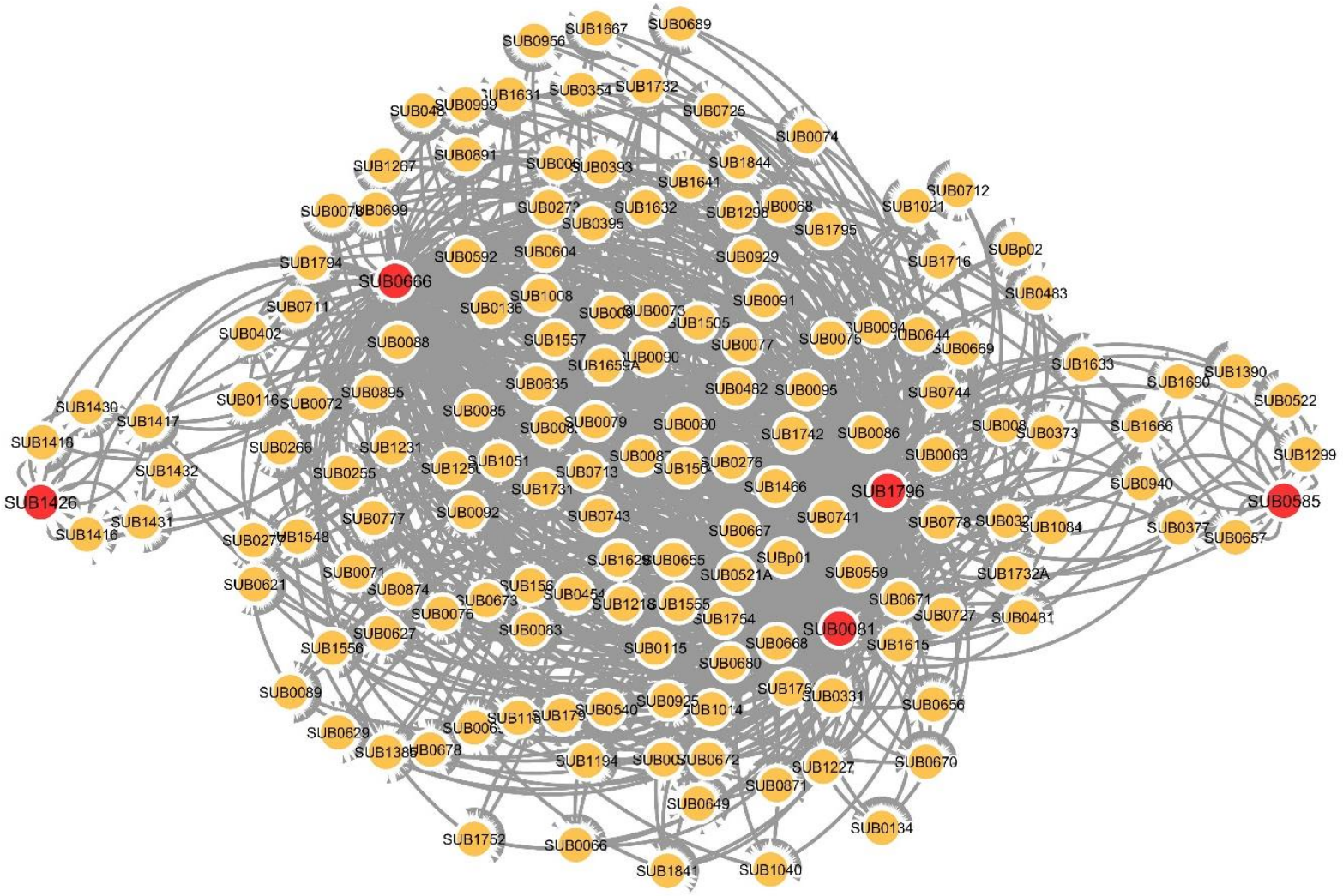


Figure 3-8. The protein-protein interaction (PPI) network showing 153 Streptococcus uberis proteins (yellow) interacting with the 5 discriminant proteins (red). *The visualisation was carried out with Cytoscape.*

GO functions of these 158 proteins (5 of interest and 153 connected to at least two genes of interest) are shown in Figure 3-9 as well as KEGG pathways they are involved in (FDR<0.05).

Figure 3-9. Functional annotation of 158 proteins (5 of interest and 153 interacting with at least two genes of interest) in Streptococcus uberis based on Gene Ontology and KEGG Pathway. *This figure was generated using R package tidyverse (Wickham et al., 2019).*

3.4 DISCUSSION

Our study shows that *S. uberis* isolates classified according to transmission route as either contagious or environmental can be discriminated by MALDI-TOF spectral profiles. The discriminatory power of MALDI-TOF appears to be greater in intra-farm analysis, in particular when using GA, indicating the potential for the development of successful screening solutions at the herd level. The inter-farm classification does not work equally well, indicating that more work

is needed to develop screening solutions applicable at the population level. Such limited classification performance may be due to the divergent evolution of bacterial strains across farms. Contact between dairy farms within the UK is limited, and management practices to control environmental mastitis vary (e.g. different choices of bedding materials, different cleaning practices, different antibiotic treatment protocols). Thus, divergent evolution and hence the emergence of farm-specific populations of *S. uberis* is entirely plausible. Diversity within common bacterial species has been demonstrated, such as for *Campylobacter* populations between broiler farms in Switzerland (Wittwer *et al.*, 2005), and *S. aureus* strains in bovines from different geographical regions in Argentina (Sordelli *et al.*, 2000). At a much larger scale, the divergent evolution of *Helicobacter pylori* has been described in human populations migrated from a common origin to different geographical regions (Falush *et al.*, 2003). Because of such diversity, it is possible that the same combination of discriminant peaks may not work equally well for all the farms.

For the investigated data, classifiers based on the GA method showed better performance over SNN and QC both before and after being optimised. The results showed the presence of proteomic phenotypic differences between contagious and environmental strains of *S. uberis* along with previously demonstrated, identical genotypic characteristics (Leigh *et al.*, 2010). This would appear plausible since selection pressure in the mammary gland is likely to force changes in protein expressions.

In this study, seven of the 19 peaks identified by the GA as the most discriminant between environmental and contagious isolates, where found corresponding to the protein products of eight genes in the reference 0140J genome. While the GA is treating the identification as a purely mathematical optimization problem, it is interesting to see if the mapped proteins have some functional meaning that might explain their differentiating power. This can be done by looking at GO, where protein functions are annotated according to the aspects of biological process, molecular function and cellular component (see Methods).

The SUB1796 protein is the ribosomal protein-L33 (RpmG) is a part of the large ribosomal subunit (Sharp, 1994). In GO, RpmG is annotated as follows: biological process – translation; molecular function – structural of constituent of ribosome; cellular component – intracellular and ribosome. RpmG was shown to have paralogs differing on whether it binds structural zinc or not, which helps bacteria survive in the case of zinc starvation (Panina, Mironov and Gelfand, 2003). This is an important consideration, as most of the ribosomal proteins are

encoded by single and highly conserved genes. Zinc has a crucial role in preserving the immune status of the cow and mammary gland health; its deficiency can result in an increased mastitis incidence (O'Rourke, 2009). RpmG was discovered to be phylogenetically distinct amongst bacterial species and within single genomes such as *Bacillus subtilis*, *Lactococcus lactis*, *Mycoplasma pneumonia*, *M. genitalium*, *Ureaplasma ueralyticum* and *Streptomyces coelicolor* (Makarova, Ponomarev and Koonin, 2001). Moreover, *rpmG* was demonstrated to be a core member of the minimal set of bacterial genes which is essential to maintain cell life (Gil *et al.*, 2004). RpmG was suggested as the putative drug target in three *Streptococcus* species (*S. agalactiae*, *S. pneumoniae*, and *S. pyogenes*) (Georrge and Umrania, 2012), two *Bordetella* species (*B. pertussis* and *B. parapertussis*) (Zhu *et al.*, 2008), *H. pylori* (Rao Reddy Neelapu and Pavani, 2013) and *Mycobacterium tuberculosis* (Fan *et al.*, 2014).

The SUB0081 protein is the ribosomal protein S-14 (RpsN), which involved in the following functions according to GO: biological process – translation, molecular function- metal ion binding, rRNA binding and structural constituent of the ribosome; cellular component - ribosome. RpsN was shown to be essential for growth and deficiency of RpsN resulted in incomplete 30S subunits in *B. subtilis* (Natori *et al.*, 2007; Nanamiya and Kawamura, 2010), *E. coli* (Shoji *et al.*, 2011) and *S. aureus* (Forsyth *et al.*, 2002).

The SUB0666 protein is an ATP synthase (subunit C, atpE). The GO annotation is: biological process - ATP hydrolysis and synthesis, and proton transport; molecular function - hydrogen ion transmembrane transporter and lipid binding; cellular component – plasma membrane. *AtpE* gene was targeted by several drug studies such as R207910 (Koul *et al.*, 2007; Petrella *et al.*, 2006) and Bedaquilline (Aguilar-Ayala *et al.*, 2017) in mycobacterium species. Moreover, subunit C was shown to be the target site of venturicidin in several *E. coli* studies (Sambongi *et al.*, 1999; Fillingame, Oldenburg and Fraga, 1991).

The SUB0585 protein is the transcriptional coactivator p15 (PC4). The GO annotation is: biological process - regulation of DNA-templated transcription; molecular function - DNA binding and transcription coactivator; cellular component- no terms assigned in this category. PC4 is also believed to have an important role in DNA repair of bacterial species since studies with *E. coli* (Wang *et al.*, 2004) and *Leptospira* species (Nascimento *et al.*, 2004) revealed that PC4 protected the DNA during oxidative stress.

The SUB1340 protein contains the Fic domain. No GO terms were assigned for this protein yet. This domain is often found in pathogenic and non-pathogenic bacteria with different

structures, where some families may contain conserved regulatory functions (Roy and Cherfils, 2015). It is suggested that pathogenic bacteria secrete Fic proteins (Roy and Cherfils, 2015); thus, showing similar functionality to toxins in terms of fulfilling some duties in the host cell such as interfering with cytoskeletal, signalling and translation pathways (Roy and Cherfils, 2015). Fic proteins participate in cell division and have been shown to synthesize folate in *E. coli* (Komano, Utsumi and Kawamukai, 1991). In turn, folate is involved in the secretion of enzymes controlling the bacterial pathogenesis in cattle pathogenic bacteria, *Histophilus somni* (Worby *et al.*, 2009).

The SUB0505 protein is a bacteriocin, which has the following GO annotation: biological process - defence response to the bacterium, molecular function and cellular component- no terms assigned in these categories. Many bacteria produce tiny peptides called bacteriocins for the antimicrobial activity to compete with intra and interspecies over limited nutrients in the environment (Eijsink *et al.*, 2002). In the study by Ward *et al.* (2009), six bacteriocin proteins including SUB0505 were found in the 0140J strain, where this redundancy was interpreted as the result of mutations. The study done by Hossain and colleagues (2015) revealed the absence of bacteriocin genes, including SUB0505, in the EF20 strain, which may be correlated to the non-virulent status of the EF20 strains. The EF20 strain of *S. uberis* was shown to be susceptible to phagocytosis by bovine neutrophils in the presence of serum (Leigh and Field, 1991), and mammary gland macrophages were reported to have the capability of killing the EF20 strain in the media containing 50% skimmed milk as the source of opsonin or 10% pooled bovine serum (Grant and Finch, 1997). Moreover, a comparison of the EF20 strain with the reference strain 0140J showed EF20 growing relatively slowly in raw skimmed milk (Leigh, Field and Williams, 1990). In another study (Leigh and Lincoln, 1997), EF20 strains of *S. uberis* performed a high amount of bounding plasmin activity following growth while bovine plasminogen was present in the media. Bacteriocin immunity proteins prevent the bacteria from the toxic effect of its own bacteriocins by forming a stable compound with the receptors (Chang *et al.*, 2009; Kjos *et al.*, 2010). The EntA immune protein (SUB1426) was discovered to guard the particular bacteria against its own class II bacteriocins (Johnsen, Fimland and Nissen-Meyer, 2005). The studies in *Streptococcus* species revealed that immunity proteins play a significant role in antimicrobial sensitivity by regulating quorum-sensing (Wang *et al.*, 2013; Matsumoto-Nakano and Kuramitsu, 2006). In summary, the literature shows that bacteriocins feature high levels of differentiation depending on the environment and host immunity response. This may

justify why we found SUB0505 as a discriminating peak between environmental and contagious strains of *S. uberis*.

Interestingly, the protein network analysis showed that three of the identified proteins (SUB0081, SUB0666 and SUB1796) interact with one another. This may suggest that the functions (ribosome, oxidative phosphorylation and bacterial secretion system) and importantly the expressions of key proteins participating in the differentiation of transmission routes of *S. uberis*, are co-ordinated. The co-ordination implies that the regulatory changes acting on these genes are accumulated over time across strains. Such stringency in constraining expression variance shifts may play an evolutionary role in the definition of different phenotypically related traits.

The results of this study suggest that MALDI-TOF spectral analysis of clinical mastitis isolates could provide a rapid means of diagnosing the likely route of transmission at the early stages of a mastitis outbreak, as previously suggested by Archer and colleagues (Archer *et al.*, 2017). Given that potentially contagious transmission of *S. uberis* has been identified in two-thirds of commercial herds and it has been found as the dominant transmission route in a third of UK herds (Davies *et al.*, 2016), there is a clear need for diagnostic tools capable to discriminate between contagious and environmentally acquired infections. Tools based on MALDI-TOF spectral analysis would enable clinicians to identify the most appropriate control measures promptly, during an outbreak of disease. A diagnosis based on MALDI-TOF spectra has the potential to reduce the incidence of clinical disease, reduce associated production losses, reduce the costs associated with the treatment of clinical mastitis and improve the efficiency of labour and resource allocation on a farm.

In conclusion, the analysis of MALDI-TOF spectral profiles through solutions powered by ML, and in particular GA, was shown to be useful to predict the contagious or environmental nature of *S. uberis* mastitis. Classifiers developed to target individual farms achieved 97.81% CV accuracy (mean over 19 farms), with a mean Cohen's kappa coefficient of 93.72%, clearly indicating the possibility to deploy effective diagnostic solutions capable to distinguish between environmental and contagious *S. uberis* strains within a farm. Prediction performance was still high at cross-validation for a classifier trained and tested on the aggregation of the data available from the same 19 farms (CV accuracy 95.88%, kappa 92.62%) but dropped to accuracy 70.67% (kappa 33.80%) when the predictor was externally validated with data from ten additional farms left as the holdout. It is unclear at the moment if such degradation may be

due to inherent proteomic diversity in *S. uberis* populations between herds, and if performance may be improved by simply increasing the amount of data available to train the predictors, for example by including a larger number of farms, or by focusing on ensuring more variation within the training sets. In any case, elucidating the role of specific proteins that have been found discriminatory between contagious and environmental transmission, may provide insights into the underpinning biology of the pathogen. The protein network analysis has also shown the presence of a protein functional network suggesting the existence of a constrained co-evolution of functional pathways and protein expression in participating in differentiating transmission routes of *S. uberis*.

As a future endeavour, it may be interesting to investigate whether similar solutions based on the analysis of MALDI-TOF spectra by means of ML may be used to develop screening tools to identify early signs of mastitis or related risk factors. Such a research goal has not been covered yet by our studies, as a significantly different approach is required both in terms of planning and executing data collection and for validating the results. Nevertheless, the analysis of MALDI-TOF peaks has proven successful at discriminating between contagious and environmental strains of infected animals, and one may wonder if discrimination between healthy and infected individuals may be carried out by looking at similar sets of peaks.

CHAPTER 4 THE USE OF MALDI-TOF TO DIFFERENTIATE PHENOTYPIC PROFILES OF *ESCHERICHIA COLI* ISOLATES

This manuscript was prepared by Necati Esener, Alexandre M. Guerra, Martin J. Green, Andrew Warry, Richard D. Emes, Andrew J. Bradley and Tania Dottorini to be submitted in an open journal. The authors' contributions were as follows: MJG and AJB provided the original data. TD, MJG, RDE and NE conceived and designed the data analysis procedures. NE, AW and RDE carried on the bioinformatics analysis. NE, AMG and TD carried on the ML analysis. NE wrote the manuscript, RDE and TD contributed with comments and amendments.

In **Chapter 4**, the main aim was to show proteomic differences between bovine mastitis-causing *E. coli* isolates with different clinical outcomes (clinical and subclinical) and disease phenotype (persistent and non-persistent). To this end, first, the persistent isolates were identified by using high throughput sequencing approaches. Data preparation of the MALDI-TOF spectra was performed by an in-house script written in MATLAB platform. Pre-processed data was analysed with ten supervised ML algorithms that were available in the sci-kit learn library in Python: LR, LSVM, RBF SVM, MLP NN, RF, DT, AdaBoost, NB, LDA and QDA. These classifiers were run on the MALDI-TOF MS data to distinguish *E. coli* isolates based on both their clinical outcome and disease phenotype.

4.1 INTRODUCTION

Mastitis can be classified into clinical and subclinical types, where both cases cause more than £200 million loss per year for the UK dairy industry and $2 billion loss per year for the US dairy industry (Bradley *et al*., 2012; Bogni *et al*., 2011). Calculation of economic costs for subclinical and clinical mastitis vary between studies as it is difficult to estimate the indirect effects of subclinical cases. However, in a single study (Huijps, Lam and Hogeveen, 2008), total economic losses per cow per year was found to be €77 and €63 for subclinical and clinical mastitis, respectively. Moreover, subclinical mastitis is more prevalent than clinical mastitis especially in developing countries where the former is up to 40 times higher than the latter (Lakew, Tolosa and Tigre, 2009; Kurjogi and Kaliwal, 2014; Argaw, 2016) and thus is responsible for more of the economic losses in the dairy industry (Aghamohammadi *et al.*, 2018).

Clinical mastitis can show obvious signs both in the animal and the milk; signs include high fever, loss of appetite, dehydration, increased size of the mammary glands whilst in the milk they would include clots, blood, flakes and watery milk appearance (Contreras and Rodríguez, 2011; Petrovski, Buneski and Trajcev, 2006). In the case of subclinical mastitis, although the animal may seem normal, there are possible scenarios that can cause loss to the dairy producer. Subclinical mastitis can result in losing 10-20% of the total milk production as a result of damage in milk-producing cells during neutrophil influx (Petersson-Wolfe *et al.*, 2013; Kumari, Bhakat and Choudhary, 2018). Subclinical mastitis may also result in earlier culling of the cows in a dairy herd (Reksen *et al*., 2006). Cows with subclinical mastitis are also a potential reservoir of IMI for the rest of the herd (Barlow, Zadoks and Schukken, 2013), which is a serious economic issue as the pathogen transmission to healthy cows was shown to be strongly associated with the cost of mastitis (Down, Green and Hudson, 2013). Elimination of subclinical mastitis also provides a decrease in the incidence of clinical mastitis cases (van den Borne *et al.*, 2010).

The clinical outcome of the mastitis pathogen depends on several factors including herd factors such as nutritional and management practices, host factors such as breed, age, lactation period etc., and pathogen factors such as genotype and phenotype of the strain (Veh *et al*., 2015; Contreras and Rodríguez, 2011). However, it is still not known fully to what degree the host, herd and pathogen factors affect the clinical severity of mastitis (Fournier *et al.*, 2008; Rainard *et al.*, 2018). Relation of mastitis pathogen and its clinical outcome was targeted in earlier studies for agents other than *E. coli* (Wolf *et al.*, 2011; Pichette-Jolette *et al.*, 2019). Specific *S. aureus* isolates, which produce a higher amount of bi-component leukocidin that targets bovine neutrophils, were found to cause clinical mastitis rather than subclinical mastitis in dairy cows (Hoekstra *et al.*, 2018). Other studies at different times and geographical locations have also shown the association between strain type of *S. aureus* and clinical severity of bovine mastitis (Haveri *et al*., 2007; Haveri *et al*., 2005; Pichette-Jolette *et al*., 2019). Association between strain type and persistence outcome was shown for *S. uberis* and *S. dysgalactiae*, as well (Phuektes *et al.*, 2001; Zadoks *et al.*, 2003; Oliver, Gillespie and Jayarao, 1998).

Formerly, the dry period used to be ignored as it was thought to be the most resistant time of the cow for new intramammary infections. However dry period has also been shown to include the most susceptible phase for new intramammary infections (Bradley and Green, 2004). On average, 24% of the quarters of dairy cows are infected with new pathogens during at dry period, 67% of which results in clinical mastitis in the following lactation (Arruda *et al.*, 2013;

Pantoja, Hulland and Ruegg, 2009; Cook, Pionek and Sharp, 2005). McDonald and Anderson (1981) observed the new infection of *E. coli* at the dry period which stayed quiescent till early lactation and ended up with clinical mastitis. High concentration levels of lactoferrin and leucocyte, where the former prevents the growth of the bacteria that needs iron by reversibly binding it, may be the reason why clinical mastitis is not often seen during the dry period (Bradley and Green, 2004). Although new infections occur at dry period, they do not generally multiply rapidly and remain subclinical till calving, then flaring up after calving. Subclinical onset followed by clinical flare-up was also observed in other mastitis pathogens *S. uberis* and *S. aureus* (Zadoks *et al.*, 2002; Zadoks *et al.*, 2003). Adaptation of bacteria to new or changing environments has been shown for other diseases as well. For example, the dynamic disease progression of *M. tuberculosis*, from subclinical status to clinical TB, has been recently discovered (Drain *et al.*, 2018) or asymptomatic bacteriuria can develop to symptomatic urinary tract infection in time (Karumanchi, August and Podymow, 2010).

E. coli is the most common mastitis pathogen responsible for up to 80% of the coliform cases (Botrel *et al.*, 2010; Bradley *et al.*, 2007). *E. coli* IMI can lead to different outcomes varying from mild to severe inflammation that can result in quarter loss or even be fatal (Burvenich *et al.*, 2003). However, severe *E. coli* IMIs occur rarely and in case of mild and moderate cases, antimicrobial usage is unnecessary (Suojala, Kaartinen and Pyörälä, 2013).

Currently, dairy producers are dependent on bacteriological analysis which only provides information about the species level. MALDI-TOF MS offers an alternate for conventional techniques by providing speedy and detailed information (at strain level more than species level). This could be used to monitor the herd over time and change strategies around disease treatment and prevention. Hence, unnecessary antimicrobial usage in dairy farms could be also reduced.

This study aims to understand the genotypic and phenotypic characteristics of bovine mastitis-causing *E. coli* strains. To achieve this aim, whole-genome sequencing was used to identify *E. coli* isolates at strain level and then MALDI profiles of these isolates were accordingly compared by using ML. *E. coli* samples were collected from a geographically limited location from the beginning of the dry period to the end of the lactation period. The same quarter was sampled to minimize host factors, environmental conditions and other batch effects. This would clarify the interaction between the host and the pathogen; moreover, any discriminant proteins can be used as potential biomarkers for the diagnosis of the mastitis disease in the subclinical status and/or be used for putative targets for novel drugs for the treatment of mastitis.

4.2 METHODS

4.2.1 Terms Used in the Study

Calving period: The period between the intramammary infusion of dry cow antibiotic and the calving day is called pre-calving; the week starting with the day of calving is called calving week, and the period from 1 week after calving (calving week) to the next intramammary infusion of dry cow antibiotic is called post-calving.

Clinical status: Inflammation of the mammary gland with visible abnormalities in the milk or udder of the cow is called clinical mastitis, and inflammation of the mammary gland with non-visible abnormalities in the milk or udder of the cow is called subclinical mastitis.

Persistence: A strain is named persistent when the samples collected from the same quarter during pre-calving and calving week/post-calving are shown to be genotypically identical. Otherwise, it is called a non-persistent isolate.

Extreme and intermediate isolates: The earliest detected subclinical isolate and the latest detected clinical isolate of each case are called extreme isolates; the other isolates between extreme ones are called intermediate isolates.

4.2.2 Data Source

This study used *E. coli* isolates obtained during a previous study (Bradley and Green, 2000), which was carried out between March 1997 and July 1998 using six farms, named as N, H, S, F, W and B. Four farms, N, H, S and F, were located in Somerset whereas two farms, W and B in North Somerset (Figure 4-1).

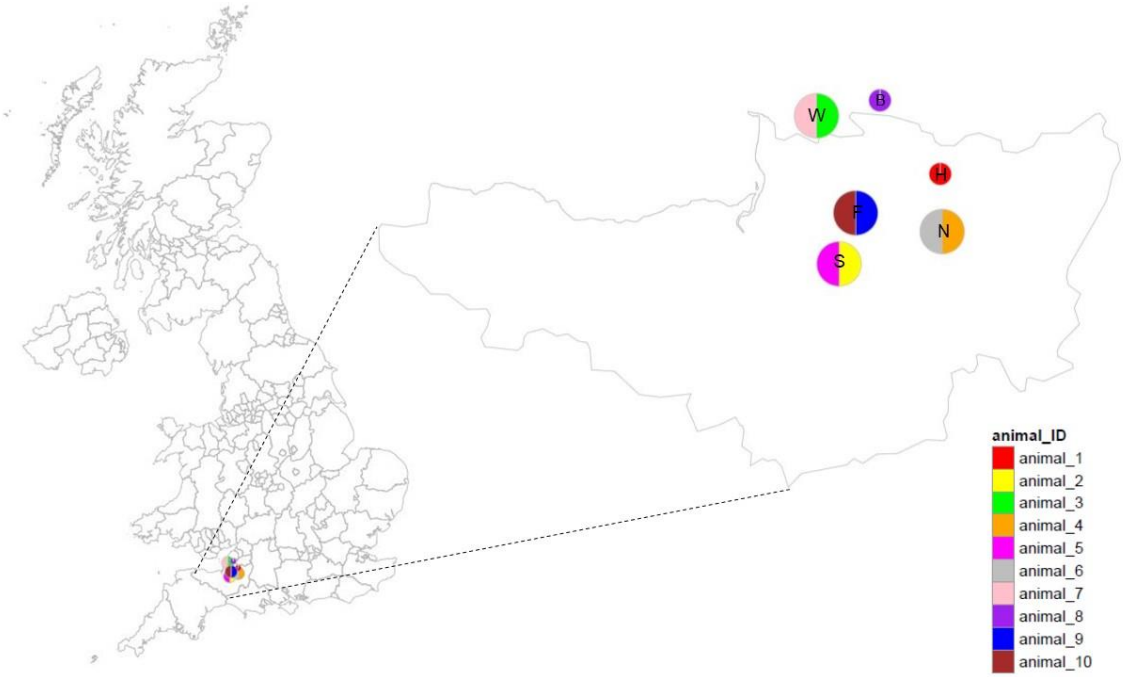


Figure 4-1. Location of the enrolled farms in the United Kingdom. *The circles represent the six farms, named as W, B, F, H, S and N, which were all in Somerset (including North Somerset). Enrolled animals were coloured in circles and the size of the circle indicates the number of cows. This figure was generated in R (R Core Team, 2019) using the sp (Pebesma and Bivand, 2005), mapdata (Deckmyn, 2018) and mapplots (Gerritsen, 2014) packages..*

In the original study (Bradley and Green, 2000), all cows from these six farms were sampled weekly during the dry period but in the following lactation period only when clinical mastitis occurred. In this work, we first assessed the recurrent mastitis in the same animal, same quarter criteria and found ten different cases (in ten different animals) (see Figure 4-2). Extreme isolates of each case (the earliest detected subclinical mastitis isolate and the latest detected clinical mastitis isolate) were sequenced to confirm whether they were persistent infection.

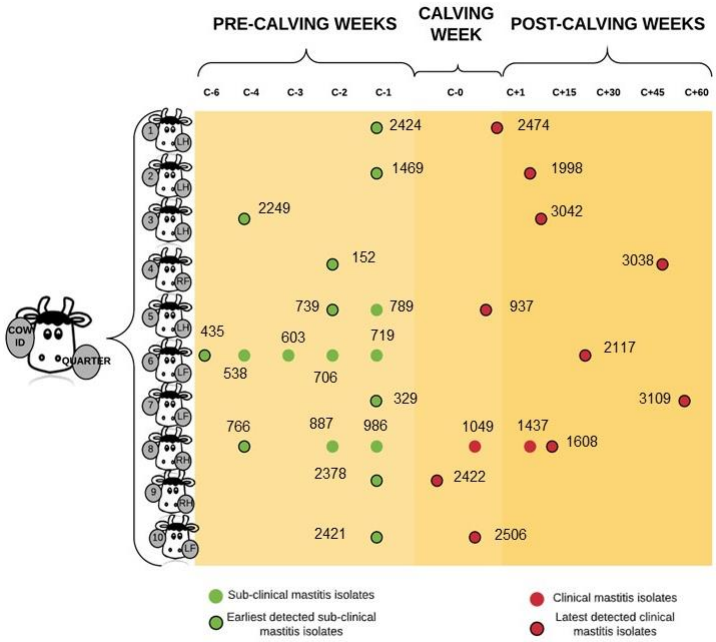


Figure 4-2. Animals and E. coli strains isolated from their quarters. *The green colour represents the subclinical status, while the red colour represents the clinical status. The earliest detected subclinical and the latest detected clinical isolates from each cow are named "extreme isolates" in this study and were sequenced. This figure was generated using Lucidchart.com.*

4.2.3 DNA Extraction

Twenty extreme isolates – defined as the earliest detected subclinical and the latest detected clinical mastitis cases from each animal (total of 10 animals), were grown overnight in Luria-Bertani broth at 37°C in a shaking incubator. The culture was then centrifuged for 5 minutes at 13,000 rpm. DNA was extracted with a commercially available kit (DNeasy Blood-Tissue Kit, Qiagen). Quality and quantity measurements of the samples were performed by using Thermo Scientific NanoDrop spectrophotometer. According to quality control results, the concentration of the samples was normalised at 50 ng/µl with the addition of nuclease-free dH_2O and the volume of 30 µl.

4.2.4 Sequencing

The samples were submitted to Next Generation Sequencing Facility at the University of Leeds. NEBNext Ultra DNA library preparation and sequencing Illumina Miseq 250 bp paired-end lane were conducted.

4.2.5 Bioinformatics Analyses

4.2.5.1 Quality Control of the Miseq Reads

The quality of the raw Miseq sequence reads (coverage 100) was assessed by using FastQC (Andrews, 2014). According to quality control reports generated by FastQC, low-quality reads and adapters were cleaned by using sliding-window trimming and adapter sequence cutting algorithms (number of bases to average across: 4, and average quality required: 28) in Trimmomatic version 0.36.5 (Bolger, Lohse and Usadel, 2014). After the cleaning step, the quality of the reads was checked again to ensure they are high enough (Quality-score>28) for further analysis.

4.2.5.2 *De Novo* Genome Assembly

Trimmed reads were *de novo* assembled by using SPAdes 3.13.1 (Bankevich *et al.*, 2012) with default parameters. Having considered the distribution of the contig length and coverage, contigs with length<500 bp length and coverage<5.0 were removed by using an in-house script. The contig with phiX cloning vector was also removed.

4.2.5.3 Genome Comparison Analyses

4.2.5.3.1 Whole-Genome Average Nucleotide Identity

Whole-genome average nucleotide identity of the assembled *E. coli* genomes, which is the average nucleotide identity of orthologous genes shared between two bacterial genomes, was checked by using FastANI tool (Jain *et al.*, 2018). FastANI was designed by using a novel algorithm Mashmap built on top of Mash technology (Jain *et al.*, 2017) which was able to work well on highly similar genomes and fast compared to BLAST (Jain *et al.*, 2018).

4.2.5.3.2 Genome Comparison by BRIG

BLAST Ring Image Generator (BRIG) v0.95 (Alikhan *et al.*, 2011) was used to compare genotypic similarities and differences between *E. coli* isolates. Genomes of 10 subclinical (pre-calving) isolates were visualised based on reference genome P4.

4.2.5.3.3 Genome Comparison by ACT

Detailed pairwise genome comparison between subclinical and clinical isolates of persistent strains was performed by using the Artemis comparison tool (ACT) (Carver *et al.*, 2005). To

perform genome comparison, input files had to be adapted in a way that ACT would process. First, genome sequences as a multi-FASTA format containing re-ordered contigs by Mauve 2.4.0 (Darling *et al.*, 2004) were converted to a single FASTA format. The single FASTA files of subclinical and clinical isolates of each persistent strain were then blasted against each other and comparison files were generated. These files were then used as inputs for the ACT.

4.2.5.4 Genome Annotation

In this study, assembled genomes were annotated by two different workflows: Prokka and RASTtk.

4.2.5.4.1 Genome Annotation by Prokka workflow

Prokka, which is a rapid prokaryotic genome annotation tool, was used to annotate the assembled genomes, with features such as gene name and functions (Seemann, 2014). Annotated assemblies in gff format were taken into the Roary pan-genome pipeline (Page *et al.*, 2015). Roary was used to construct core-genome alignment as well as statistics on core-genome and accessory-genome. The threshold values were set to 95% of sequence similarity of amino-acid level and must be found in 99% of isolates for a gene to be considered in the core genome. Roary output containing gene presence and absence in comma-separated values (CSV) format were visualised by Phandango (Hadfield *et al.*, 2017).

4.2.5.4.2 Genome Annotation by RASTtk workflow

Annotation of *E. coli* genomes was also performed by using RASTtk (Rapid Annotation using Subsystem Technology toolkit) (Brettin *et al.*, 2015). Functional classification of coding sequence (CDS) for persistent *E. coli* cases was further analysed in the SEED subsystem (Overbeek *et al.*, 2014).

4.2.5.5 Phenotype-Specific Gene Control Analyses

4.2.5.5.1 Persistent-Specific Gene Control Analysis

A CSV file containing the presence/absence of genes generated by Roary was then taken into Scoary tool (Brynildsrud *et al.*, 2016), which was used to find the relationship between the genes and traits. We were able to classify genomes in different groups and generate diagnostic test (sensitivity-specificity) reports as well as statistical reports based on the prevalence of certain genes in these groups.

In the first analysis, persistent strains (isolates 2424&2474, 739&937, 766&1608 and 2421&2506) were labelled as a positive group and non-persistent strains (the remaining twelve isolates) were labelled as a negative group. In this analysis, the probability of presence for persistent-unique genes was checked.

In the second analysis, isolates collected during pre-calving were targeted and the ones that showed persistence till post-calving were labelled as a positive group and the ones that failed to be detected post-calving were labelled as a negative group. Again, the probable genes unique to persistent profiles were checked in a smaller but more accurate subset (the isolates found post-calving could show persistency later).

4.2.5.5.2 Clinical Status-Specific Gene Control Analysis

In this study, OrthoMCL workflow was used to define genes between subclinical and clinical phenotypes of individual persistent strains. OrthoMCL pipeline was available via the Galaxy platform instance of VeuPathDB (Aurrecoechea *et al.*, 2017). The pipeline briefly includes the following steps: firstly, the proteome FASTA file was generated. Then, all-v-all BLAST was run on the filtered proteins file and the file was formatted so that it could be loaded into the OrthoMCL database. Protein pairs were found in the following step and MCL was performed on these pairs to identify ortholog groups. OrthoMCL was set with BLASTp cut-off e-5 and 50% match, leaving other parameters at default (Fischer *et al.*, 2011).

4.2.5.6 Genome Typing Analyses

4.2.5.6.1 MLST Analysis

MLST of *E. coli* relies on several different sets of housekeeping genes which vary according to the technique, for example, the most widely used Achtman scheme screens variation of seven housekeeping genes (*adk*, *fumC*, *gyrB, icd, mdh*, *purA* and *recA*) (Zhou *et al.*, 2020) and Pasteur's scheme sequence typing uses sets of eight housekeeping genes (*dinB, icdA, pabB, polB, putP, trpA, trpB* and *uidA*) (Clermont, Gordon and Denamur, 2015).

4.2.5.6.2 Phylotyping

ClermonTyper (Beghain *et al.*, 2018) and Mash (Ondov *et al.*, 2016), which are *in silico* PCR assay tools, were used to predict the phylogroup of *E. coli* isolates.

4.2.5.6.3 *fumC fimH* (CH) Typing

CH Typing was based on two loci, which were *fumC*, one of the seven household genes from Achtman MLST, and *fimH*, an integral fragment of type 1 fimbrial-adhesin-encoding gene (Roer et al., 2018). In this study, CH types of our 20 *E. coli* isolates were found by using web tool CHTyper 1.0 (https://cge.cbs.dtu.dk/services/CHTyper-1.0/).

4.2.5.6.4 Serotyping

Serotyping of *E. coli* isolates used SerotypeFinder 2.0 web tool, where the minimum length of the match was set to contain 60% of the nucleotide for the serotype gene. The tool checked *wzx*, *wzy*, *wzm* and *wzt* for O-antigen and *fliC*, *flkA*, *fllA*, *flmA* and *flnA* for H antigen (Joensen *et al.*, 2015).

4.2.5.7 Variant Calling

SNPs between the reference genome, mastitis model strain P4, and 20 *E. coli* genomes were found by using Snippy 3.2 (Seemann, 2015). Snippy-core 3.2 (Seemann, 2015) was used to merge Snippy outputs as a core SNP alignment. FastTree (Price, Dehal and Arkin, 2009) was used to generate phylogenetic trees for both whole-genome and core genome of the 20 *E. coli* isolates. Phylogenetic trees were visualised and coloured by ITOL v3 (Letunic and Bork, 2016).

4.2.5.8 Detection of Antimicrobial-Resistant Genes

AMR genes in the *E. coli* genomes were checked using ResFinder v3.1 (Zankari et al., 2012). The ResFinder database contains AMR genes for the following antibiotic classes; aminoglycoside, beta-lactam, colistin, fluoroquinolone, fosfomycin, fusidic acid, glycopeptide, nitroimidazole, oxazolidinone, phenicol, rifampicin, sulphonamide, tetracycline, trimethoprim, macrolide, lincosamide and streptogramin B (Zankari et al., 2012).

4.2.5.9 Detection of Virulence Genes

The virulence genes in the assembled genomes were checked using VirulenceFinder 2.0 database specific to *E. coli* (Joensen *et al.*, 2014). Threshold values for identity and minimum length were set as 90% and 60%, respectively.

4.2.5.10 Detection of Plasmids

PlasmidFinder v2.1 (Carattoli *et al.*, 2014), a web-based tool, was used to perform an *in silico* plasmid detection of a query genome. The PlasmidFinder database holds two different plasmid datasets which occurred mainly in gram-positive bacteria and *Enterobacteriaceae*. In this study, we screened the plasmids in both databases matching with our query genomes at a minimum threshold identity of 95% and minimum threshold coverage of 60%, which were the default values.

4.2.5.11 Detection of Prophages

PHASTER (PHAge Search Tool Enhanced Release) was used to screen the *E. coli* genomes for detection of any possible prophages causing the difference between clinical and subclinical isolates or changing subclinical cases to clinical cases in persistent isolates (Arndt *et al.*, 2016).

4.3 RESULTS

4.3.1 Bioinformatics Analyses

4.3.1.1 *De Novo* Genome Assembly

The quality of each genome assembly before and after the cleaning step was assessed by QUAST (Quality Assessment Tool) (Mikheenko *et al.*, 2018). The statistics such as the number of contigs, GC content, the minimum number of the contigs that produce 50% (L50) and 75% (L75) of the bases in the assembly are shown in Table 4-1.

GC content of the genomes was varied between 50.5 and 50.9. L50 and L75 value of the isolates varied from 4 to 16 and 9 to 32, respectively where the worst in isolate-2117 (L50=16 and L75=32) and best in isolate-1469 (L50=4 and L75=9). The number of uncalled bases was the worst with the count of 495 in isolate-3109. However, the average number of uncalled bases was per 100,000 assembly bases was the worst at 9.82 in isolate 1608.

Table 4-1. Quality assessment of the assembled genome by using QUAST. *The statistics were realized based on the contigs with a length size of 500 bp and above unless otherwise was stated such as the number of contigs (≥0 bp) or the number of contigs (≥1000 bp).*

Assembly	152	329	435	739	766	937	1469	1608	1998	2117	2249	2378	2421	2422	2424	2474	2506	3038	3042	3109
# contigs (≥0 bp)	236	106	127	81	189	72	118	213	174	220	296	144	142	153	162	155	139	165	115	185
# contigs (≥500 bp)	140	49	73	39	108	42	49	129	101	113	127	77	73	82	89	81	76	92	47	95
# contigs (≥1000 bp)	116	41	62	35	94	36	32	99	88	101	100	62	61	60	72	69	60	73	37	84
Largest contig (kbp)	424.2	924.7	320.6	692.3	307.0	692.3	741.5	307.0	363.5	300.8	360.1	490.9	533.9	417.1	454.2	454.2	584.7	406.1	680.9	352.0
Total length (Mbp)	4.9	4.9	4.7	4.8	5.0	4.8	4.7	5.0	4.7	4.7	4.9	5.2	5.2	4.8	5.1	5.1	5.2	4.9	4.8	5.1
GC (%)	50.6	50.6	50.9	50.7	50.6	50.7	50.9	50.6	50.8	50.8	50.8	50.5	50.5	50.8	50.6	50.6	50.5	50.7	50.8	50.7
N50 (kbp)	104.2	217.5	164.1	384.0	127.1	337.4	351.0	127.1	105.1	101.3	132.4	264.4	273.3	176.2	172.7	172.7	273.3	126.8	233.7	119.6
N75 (kbp)	47.5	116.8	76.9	175.2	82.629	175.2	186.1	73.3	57.5	56.4	63.1	111.6	111.6	97.5	93.4	100.7	111.6	79.5	130.8	70.5
L50	14	7	10	5	15	6	4	14	13	16	13	8	7	10	9	9	7	12	6	13
L75	30	14	20	10	26	11	9	26	27	32	26	15	14	19	19	18	13	23	12	26
#N's	97	98	0	195	295	98	0	491	197	298	195	0	99	99	193	188	0	396	0	495
# N's per 100 kbp	1.98	1.99	0	4.05	5.89	2.03	0	9.78	4.22	6.28	3.96	0	1.91	2.05	3.77	3.68	0	8.13	0	9.74

represents “number of”.

N50= the length of the shortest contig in the set of longest contigs that cover at least 50% of the total assembly size.

N75= the length of the shortest contig in the set of longest contigs that cover at least 75% of the total assembly size.

L50= minimum number of the contigs that produce 50% of the bases in the assembly.

L75= minimum number of the contigs that produce 75% of the bases in the assembly.

#N’s= the total number of uncalled bases in the assembly.

#N’s per 100 kbp= the average number of uncalled bases per 100,000 assembly bases.

4.3.1.2 Genome Comparison Analyses

4.3.1.2.1 Whole-Genome Average Nucleotide Identity

Whole-genome average nucleotide analysis was visualised by using pheatmap package in R (Kolde and Kolde, 2015). Achtman MLST results, phenotypes, animal and farm IDs were also included (see Figure 4-3). The isolates collected from the same cow were compared to determine if they are genotypically identical (persistent strain). Four different cases were assigned as “persistent” strains which are isolates 2424&2474 from animal-1, isolates 739&937 from animal-5, isolates 766&1608 from animal-8 and isolates 2421&2506 from animal-10. Isolate 2378, a subclinical isolate of animal-9, was found highly similar to persistent strain (isolates 2421&2506) of animal-10. Animal 9 and 10 were on the same farm. In the rest of the study, cases from animal-1, animal-5, animal-8 and animal-10 were termed persistent while other isolates were termed non-persistent.

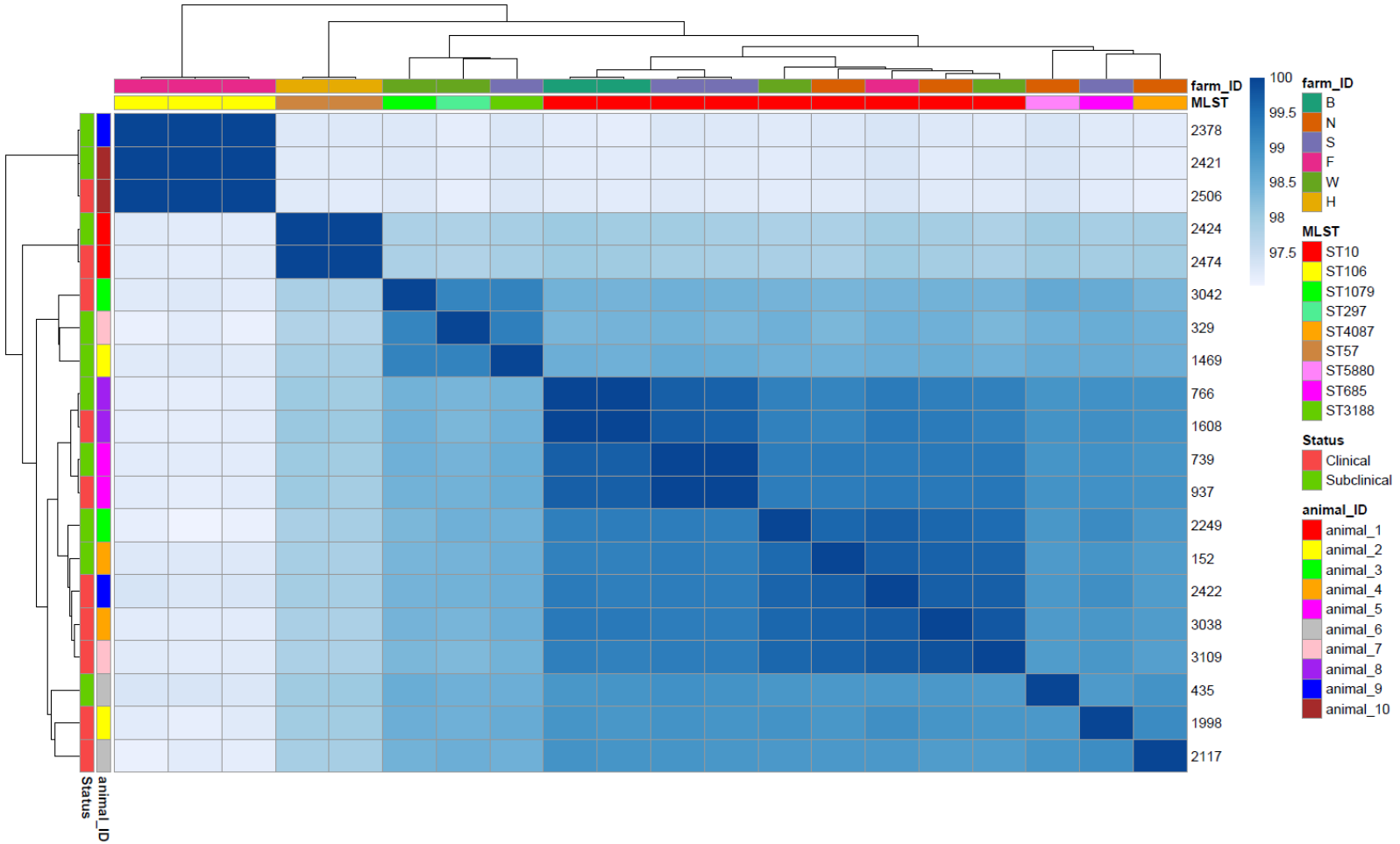


Figure 4-3. Whole-genome average nucleotide identity of 20 E. coli isolates performed by FastANI. *The heatmap represents the similarity between isolates where darkest blue is the most similar (scale from 97.5% to 100.0%). Isolate IDs were shown on the right, animal IDs and phenotypes of the isolates on the left, farm IDs and Achtman MLST schemes of the isolates were shown on the top of the heatmap. Isolates 2424&2474 of animal-1, 739&937 of animal-5, 766&1608 of animal-8 and 2421&2506 of animal-10 were found to be highly similar and assigned as persistent. This figure was generated using R package pheatmap (Kolde and Kolde, 2015).*

4.3.1.2.2 Genome Comparison by BRIG

There were some missing regions in the *E. coli* genomes compared to the P4 model mastitis genome. These regions were found mostly in prophages that the P4 strain contained (see Figure 4-4-A). This was confirmed by phage finder tool PHASTER (see Figure 4-4-B). Overall, no pattern was observed between subclinical isolates of persistent and non-persistent strains.

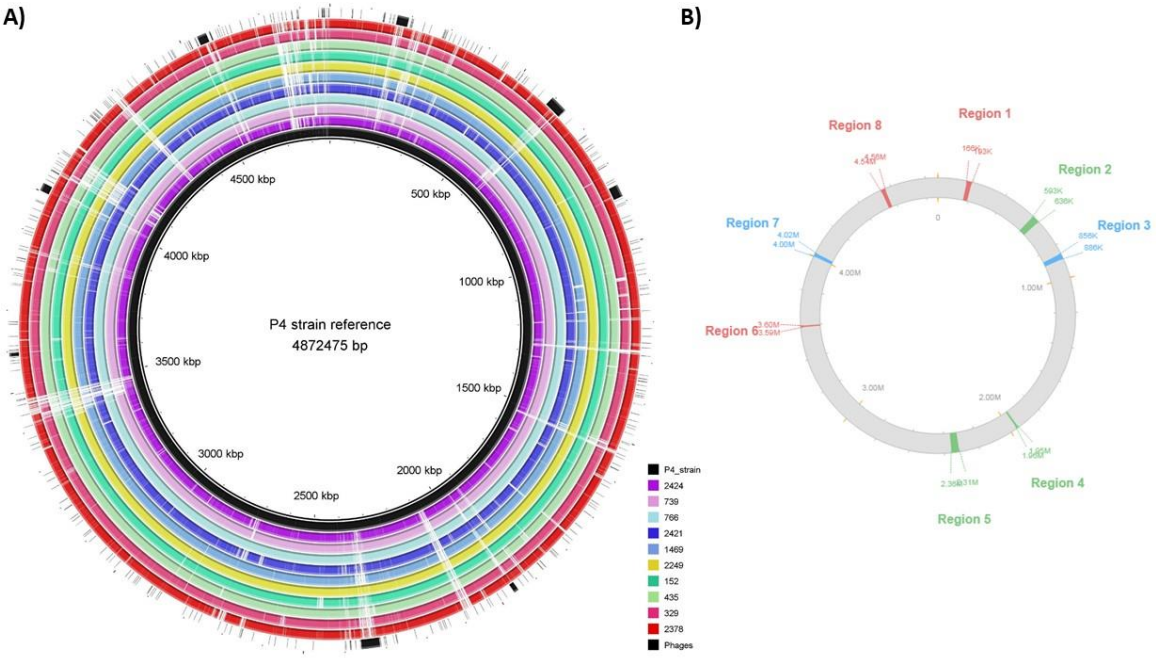


Figure 4-4. Genome comparison of subclinical isolates in a circular diagram created by BRIG. A) *The innermost circle is model mastitis genome P4 which was used as a reference genome to map our E. coli isolates. The outermost circle represents the phages present in the P4 strain.* ***B)*** *The regions of phages that are present in P4 genomes are visualised by PHASTER. Most of the empty regions in our E. coli genomes correspond to these phages of the P4 strain.*

4.3.1.2.3 Genome Comparison by ACT

Genome comparison within persistent cases showed that there were no deletions, and samples showed very high similarity between subclinical and clinical isolates, although there were some inversions, which may be due to errors during short-read sequencing. There were relatively more inversions in persistent strains of animal-1 and animal-8 compared to animal-5 and animal-10 (see Figure 4-5).

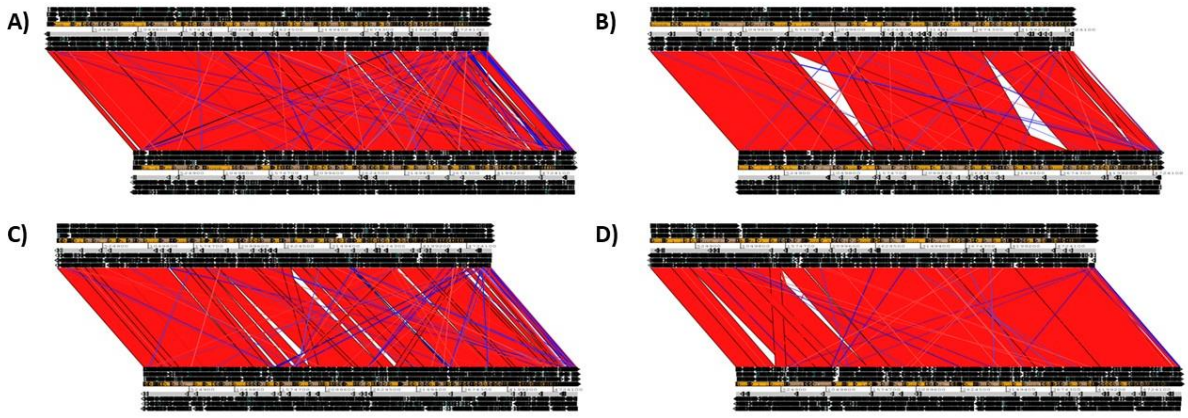


Figure 4-5. Comparison of subclinical and clinical phenotypes of persistent strains by using ACT. *Comparison of* ***A)*** *isolate-2424 with isolate 2474 from animal-1,* ***B****) isolate-739 with isolate-937 from animal-5,* ***C)*** *isolate-766 with isolate-1608 from animal-8 and* ***D)*** *isolate-2421 with isolate-2506 from animal-10 were shown****.*** *Red blocks represent the forward, and blue lines represent the reverse orientation of conserved regions within the isolates. The top section (above red blocks) is the sequence view panel (top three segments are forward and bottom three segments are reverse frame lines) of subclinical isolates and the bottom section (below red blocks) is the sequence view panel (top three segments are forward and bottom three segments are reverse frame lines) of clinical isolates.*

4.3.1.3 Genome Annotation

4.3.1.3.1 Genome Annotation by Prokka

Protein-coding genes detected by Prokka were 4753 in isolate-2424 and 4755 in isolate-2474 of persistent strain in animal-1; 4446 in isolate-739 and 4444 in isolate-937 of persistent strain in animal-5; 4601 in isolate-766 and 4592 in isolate-1608 of persistent strain in animal-8; 4842 in isolate-2421 and 4843 in isolate-2506 of persistent strain on animal-10. These proteins were then detailed in OrthoMCL (section 4.3.2.4.2).

Roary pan-genome analysis showed that 3259 genes were shared in all 20 *E. coli* isolates. Other genes grouped in soft-core genes, shell genes, cloud genes and total genes in all isolates can be seen in Table 4-2.

Table 4-2. The pangenome analysis of 20 E. coli isolates. *Gene counts in the core (core and soft-core genes) and accessory (shell and cloud genes) genome are listed.*

Class	Distribution (%)	Gene Count
Core genes	present in all 20 isolates	3259
Soft-core genes	present in 19 isolates	114
Shell genes	present in 3 ≤ isolates < 19	2629
Cloud genes	present in 0 ≤ isolates < 3	3354
Total genes	present in 0 ≤ isolates ≤ 20	9356

A phylogenetic tree constructed based on pangenome analysis showed that isolates 2424&2474 of persistent strain in animal-1, isolates 739&937 of persistent strain in animal-5, isolates 766&1608 of persistent strain in animal-8, isolates 2421&2506 of persistent strain in animal-10 were clustered within each case (see Figure 4-6). Isolate-2378, which was always the closest to the persistent strain of animal-10 could be separated from those isolates.

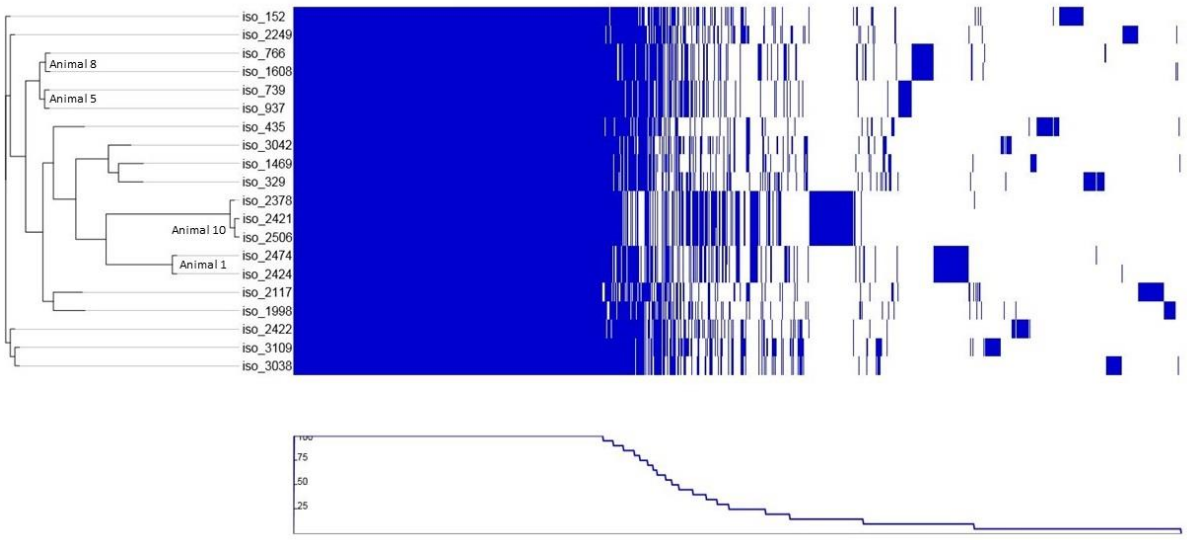


Figure 4-6. Roary pan-genome analysis of 20 E. coli strain visualised by Phandango. *The phylogeny tree on the left is linked to the pan-genome content of the isolates. Blue blocks represent the presence of genes relative to the reference genome of the mastitis model strain P4. The line graph at the bottom represents the percentage of isolates carrying the gene at a certain position. Isolates 2424&2474 of animal-1, isolates 739&937 of animal-5, isolates 766&1608 of animal-8 and isolates 2421&2506 of animal-10 were shown to be clustered within the case. These isolates were proved to be persistent strains in this analysis.*

4.3.1.3.2 Genome Annotation by RASTtk

According to functional classifications based on the SEED subsystem, there was no significant difference between the subclinical and clinical isolates of persistent cases on the animal basis, except the phages, prophages and transposable elements and plasmids subsystem. This was consistent with the findings of OrthoMCL as the difference between the isolates of some persistent strains were also found to be the product of prophages they had. Carbohydrates were the most frequent functional classification category while dormancy and sporulation subsystem was the least common ones (see Figure 4-7). Although the gene count of functional classifications was so close to each other, the following three categories varied the most between each case:

1) Phages, prophages, transposable elements and plasmids:
   166 genes for each isolate (isolate 2424 and 2474) of animal-1,
   33 genes for each isolate (isolate 739 and 937) of animal-5,
   39 and 37 genes for isolate-766 and isolate 1608, respectively, of animal-8,
   167 and 166 genes for isolate-2421 and isolate 2506, respectively, of animal-10.
2) Metabolism of aromatic compounds:
   7 genes for each isolate (isolate 2424 and 2474) of animal-1,
   6 genes for each isolate (isolate 739 and 937) of animal-5,
   21 genes for each isolate (isolate 766 and 1608) of animal-8,
   27 genes for each isolate (isolate 2421 and 2506) of animal-10.
3) Iron acquisition and metabolism:
   66 genes for each isolate (isolate 2424 and 2474) of animal-1,
   39 genes for each isolate (isolate 739 and 937) of animal-5,
   52 genes for each isolate (isolate 766 and 1608) of animal-8,
   76 genes for each isolate (isolate 2421 and 2506) of animal 10.

Particular attention was paid to iron acquisition and metabolism genes as *fec* locus (*fecIRABCDE*) iron (III) dicitrate transport were shown to be significant for pathogenicity of mastitis-causing *E. coli* genomes (Blum *et al.*, 2018). We checked if the genes on this locus cause the different gene counts for this class. However, *fec* locus genes (*fecIRABCDE*) were present in all the genomes and was not the reason for the different count of the genes in this class.

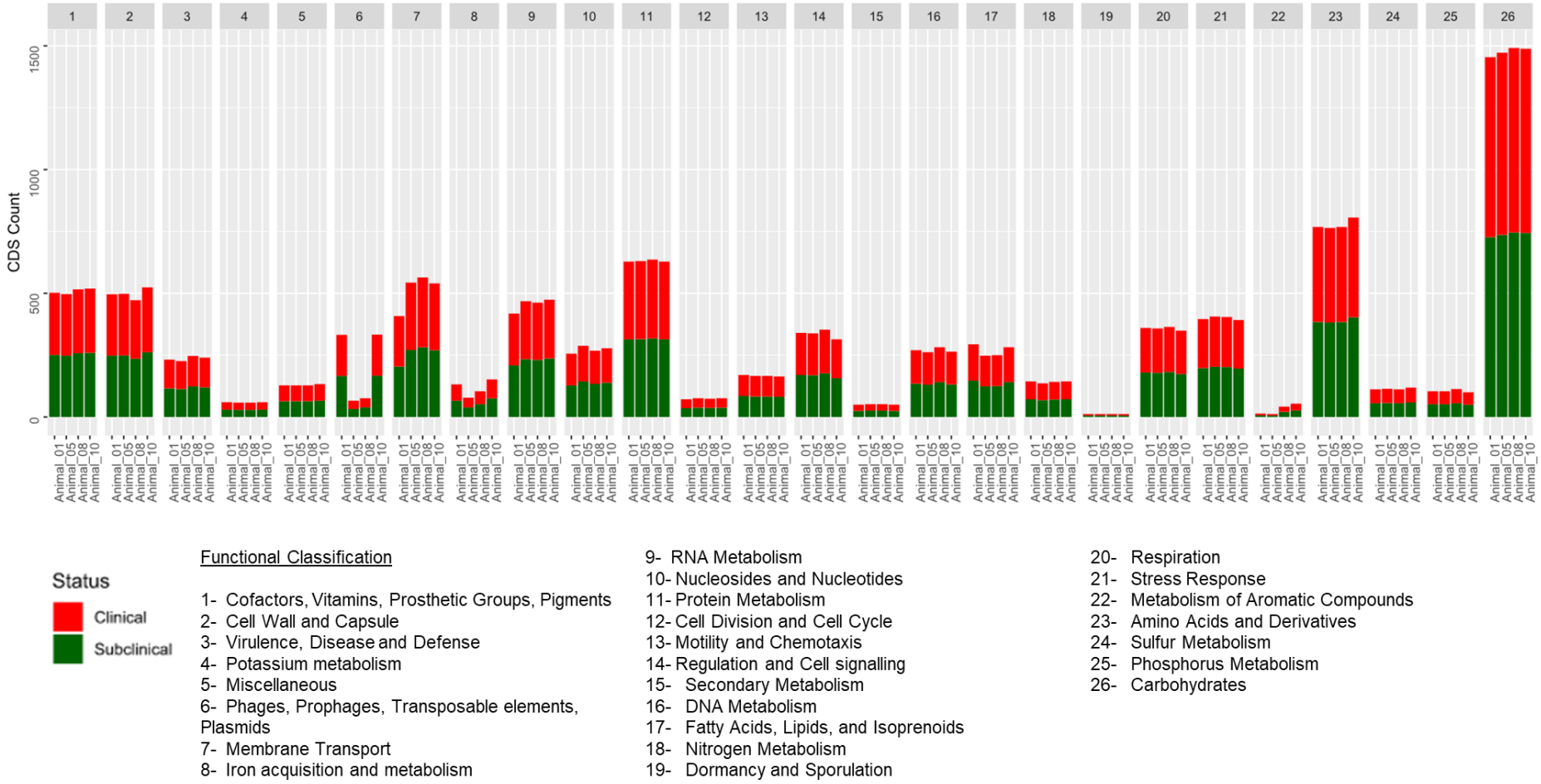


Figure 4-7. CDS counts in the functional classifications based on the SEED subsystem database. *The graph shows the 26 functional categories and the gene counts of 4 persistent cases (based on animals (animal-1, animal-5, animal-8 and animal-10) in these categories. Each subclinical (isolate-2424 (from animal-1), isolate-739 (from animal-5), isolate-766 (from animal-8) and isolate-2421 (from animal-10)) and clinical (isolate-2474 (from animal-1), isolate-937 (from animal-5), isolate 1608 (from animal-8) and isolate-2506 (from animal-10)) isolate in these persistent cases are seen in green and red, respectively. This figure was generated using R package ggplot2 (Wickham, 2011).*

4.3.1.4 Phenotype-Specific Gene Control Analyses

4.3.1.4.1 Persistent-Specific Gene Control Analysis

Persistent-specific gene control analysis was performed based on comparing two different groups as detailed in the Methods section. The reports created by Scoary were examined. There was no gene found with 100% sensitivity and specificity between persistent and non-persistent strains. This was expected as some isolates, i.e. isolate-2378 was consistently found more closely related to the persistent strain of animal-10 (isolates 2421&2506) at phylogenetic trees. Hence, specificity was lowered, and the gene list was examined again. However, there was still no gene found between the evaluated groups in both analyses.

4.3.1.4.2 Clinical Status-Specific Gene Control Analysis

Roary was performed for pan-genome analysis and gave satisfactory results; however, 0.37% proteins of the total proteome could not be appointed correctly, i.e. they appeared in different annotation groups (different annotation name was given), although they are the same. This problem could be solved by supplying a well-annotated reference genome in Roary as suggested by its developers. However, we used another pipeline called OrthoMCL to define missing orthologs and alternate the technique. OrthoMCL runs all-against-all BLASTp (Fischer *et al.*, 2011) and then these BLASTp results are employed to define orthologs and paralogs by performing Markov Clustering (MCL) (Dongen, 2000). The proteins (4.05% of the expected proteome size) which could not be mapped with any known groups (no ID given by OrthoMCL) were filtered out from each proteome and only those mapped with known OrthoMCL-DB groups (~96%) were kept. Common and unique proteins within the isolates of persistent strains from animals were compared (see Figure 4-8).

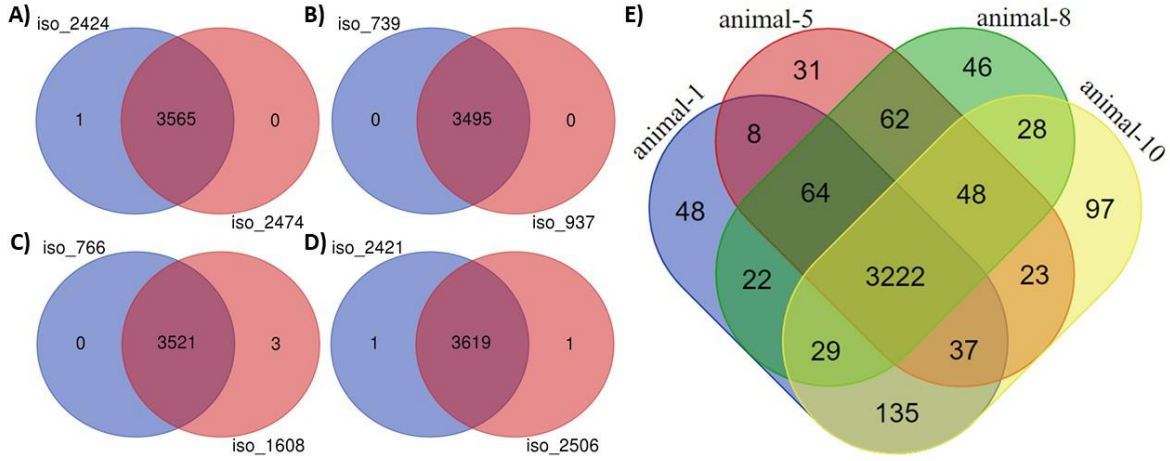


Figure 4-8. Comparison of isolates in terms of common and unique proteins based on orthology. *The count of core and unique genes for **A)** subclinical (iso-2424) and clinical (iso-2474) isolates of animal-1, **B)** subclinical (iso-739) and clinical (937) isolates of animal-5, **C)** subclinical (iso-766) and clinical (iso-1608) isolates of animal-8, **D)** subclinical (iso-2421) and clinical (iso-2506) isolates of animal-10, **E)** persistent strains isolated from each animal. These figures were generated at http://bioinformatics.psb.ugent.be/webtools/Venn.*

A unique protein was found in the subclinical form (isolate-2424) of persistent strain from animal-1 which was not present in its clinical form isolate-2474. There were 3565 common proteins between isolate 2424 and 2474. No unique protein was found between subclinical and clinical forms (isolates 739&937) of persistent strain from animal-5. There were 3495 common proteins between isolate 739 and 937. Three unique proteins were found which were absent in subclinical form (isolate-766) of persistent strain from animal-8 but present in its clinical form (isolate-1608). There were 3521 common proteins between isolate 766 and 1608. One unique protein was found in each form of the isolates (isolates 2421&2506) coming from animal-10. There were common 3619 proteins between isolates 2421 and 2506. Most of the unique proteins were with unknown function as commonly known as hypothetical proteins. These hypothetical proteins were analysed further and seen as the proteins coming from phages as listed in details in Table 4-3.

The common proteins between clinical and subclinical isolates of each persistent strains were also compared with each other. There were 3222 proteins were found common between these persistent strains. 48 unique proteins in animal-1, 31 unique proteins in animal-5, 46 unique proteins in animal-8 and 97 unique proteins in animal-10 were found.

***Table 4-3. Unique proteins found in the isolates by pairwise comparison within the persistent strains*.**

Isolate	Protein	BLASTp Matches			
		Matched Protein	Coverage	E-value	Accession
1608	HokB	type I toxin-antitoxin system toxin HokB	100%	2.00E-26	WP_136750091.1
1608	Hypothetical Protein	IS3-like element IS2 family transposase	100%	0	WP_112842836.1
1608	Insertion element IS6110	IS3 family transposase	100%	5.00E-74	WP_000165309.1
2421	Hypothetical Protein	IS1 family transposase	100%	4.00E-61	WP_000179210.1
2424	Hypothetical Protein	IS1 family transposase	100%	9.00E-64	WP_000951585.1
2506	Hypothetical Protein	Transposase	100%	3.00E-54	ALL91260.1

4.3.1.5 Genome Typing Analyses

4.3.1.5.1 MLST Analysis

According to the Achtman scheme, ST-10 was detected in two persistent strains (isolates 739&937 in animal-5 and isolates 766&1608 in animal-8) as well as six other strains (isolates 152, 1469, 2249, 2422, 3038 and 3109). Other persistent strains were found to be ST-57 and ST-106 for animal-1 (isolates 2424&2474) and animal-10 (isolates 2421&2506), respectively. Other *E. coli* isolates were found as follows: isolate-3042 as ST-1079**,** isolate 329 as ST-297, isolate 685 as ST-685, isolate 2117 as ST-4087 and isolate-435 as ST-5880.

The distance between STs found in our *E. coli* isolates and clonal complexes of these STs based on the Achtman MLST scheme were visualised by using the global optimal eBURST minimum spanning tree (MST) algorithm available in Phyloviz v2.0a (Francisco *et al.*, 2009) (see Figure 4-9).

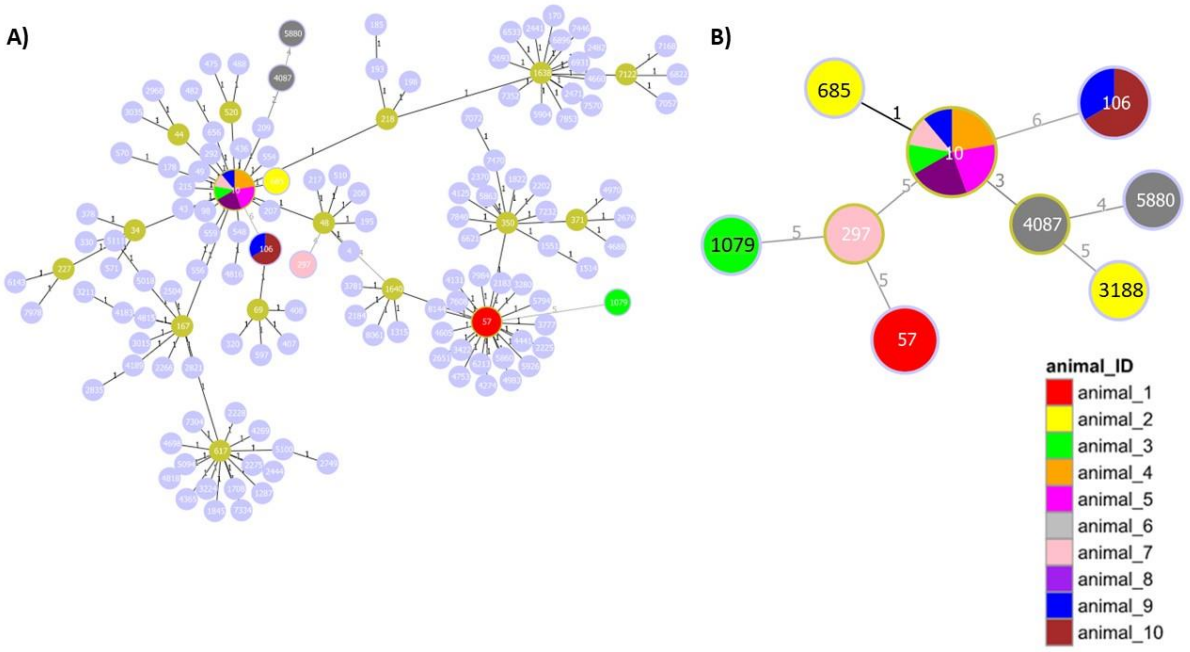


Figure 4-9. Global optimal eBURST (goeBURST) distance analysis of the Achtman MLST scheme. *Numbers in the nodes represent the sequence type (ST), while node size is related to the count of isolate in each ST. A) STs of 20 E. coli genomes of this study with the STs belong to the same clonal complex and B) close-up showing the distance between each other only. The distance labels represent the count of the variant in seven housekeeping loci between STs. These figures were generated at Phyloviz v2.0a.*

By using Pasteur scheme MLST, the most common MLST was detected as ST-2 in this analysis. Persistent strain in animal-5 (isolates 739&937) and persistent strain in animal-8 (isolates 766&1608) were found to be ST-2. However, other isolates (isolate 152, 1469, 2249, 2422, 3038, 3109) which were typed as ST-10 according to the Achtman scheme, were found to have completely different strain types according to the Pasteur scheme; isolate 152 was ST-475. Isolate 1469 was ST-108, isolate 2249 was ST-387, isolate 2422 was ST-818, isolate 3038 and isolate 3109 were ST-383. Other persistent strains in animal-1 (isolates 2424&2474) and animal-10 (isolates 2421&2506) were found to be ST-305 and ST-3 according to the Pasteur scheme, respectively.

4.3.1.5.2 Phylotyping

The phylotyping analyses were consistent between ClermonTyper and Mash tools except for the isolate-435 which gave phylogroup C and A, respectively, as a result of a mutation in one of the binding primers. Persistent strains in animal-1 (isolates 2424&2474), animal-5 (isolates 739&937), animal-8 (isolates 766&1608) and animal-10 (isolates 2421&2506) were found

belonging to phylogroup E, A, A and D, respectively. Amongst the remaining isolates, phylogroup A was the most common as being found in isolate 152, 1998, 2117, 2249, 2422, 3038 and 3109 while three isolates (isolate 329, 1469 and 3042) were found in phylogroup B1 (see Figure 4-10).

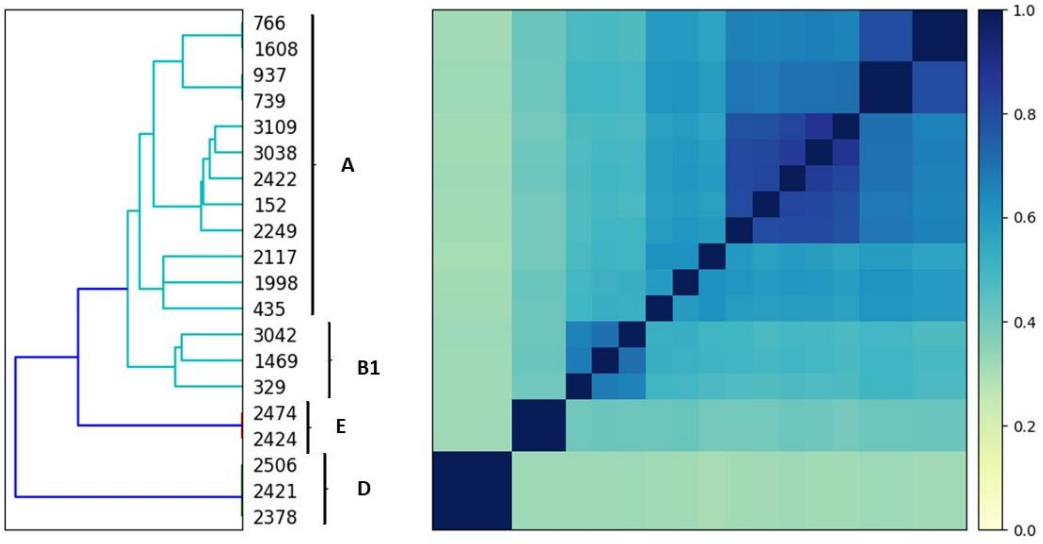


Figure 4-10. Phylogroups of 20 E. coli genomes were found by Mash, genome clustering tool. *The heatmap shows the similarity based on k-mers of genomes, generated by Sourmash tool (Brown and Irber, 2016).*

4.3.1.5.3 *fumC fimH* (CH) Typing

fumC-fimH (CH types) of persistent cases were found as follows; isolates 2424&2474 (in animal-1) as 31-54, isolates 739&937 (in animal-5) as 11-27, isolates 766&1608 (in animal 8) as 11-54, isolates 2421&2506 (in animal 10) as 35-47 (see Table 4-4). Amongst the remaining isolates most common CH type was 11-27 (in 4 more cases; isolates 152, 2422, 3038 and 3109). Interestingly, *fumC*-11 was in 56.25% of the cases in total (9 isolates out of 16) (see Table 4-4).

4.3.1.5.4 Serotyping

Serotypes of persistent isolates were found to be O176-H32 for isolates 2424&2474 (persistent in animal-1), O74-H39 for isolates 739&937 (persistent in animal-5), O9-H17 for isolates 766&1608 (persistent in animal-8) and O17/O44 (100% similarity with *wzy* gene)-O17/O77 (100% similarity with *wzx* gene)-H18 for isolates 2421&2506 (persistent in animal-10). Amongst other strains serotypes of O13 (100% similarity with *wzx* gene), O13/O135 (100% similarity with *wzy* gene) and H11 were found in isolates 3038&3109. As seen in Table 4-4, serotyping of the persistent strain in animal-10 and isolates 3038&3109 resulted in two

different typing outcomes for their O antigens. This kind of outcome was also noticed by the creators of the SerotypeFinder tool and compared with the conventional serotyping technique. According to conventional serotyping technique, typing of O17/O44-O17/O77 was O17 and typing of O13, O13/135 was O135 (Joensen *et al.*, 2015).

Table 4-4. Summary of the genome typing analysis results of 20 E. coli genomes.

ID		Phylogroup		MLST		CH Typing		Serotype	
Isolate	**Animal**	**ClermonType**	**Mash**	**Achtman**	**Pasteur**	***fumC***	***fimH***	**O type**	**H type**
2424	1	E	E	ST-57	ST-305	*fumC*-31	*fimH*-54	O176	H32
2474	1	E	E	ST-57	ST-305	*fumC*-31	*fimH*-54	O176	H32
1469	2	B1	B1	ST-3188	ST-108	*fumC*-23	*fimH*-38	Ont	H16
1998	2	A	A	ST-685	ST-698	*fumC*-11	*fimH*-34	Ont	H34
2249	3	A	A	ST-10	ST-387	*fumC*-11	*fimH*-54	O107	H54
3042	3	B1	B1	ST-1079	ST-360	*fumC*-19	*fimH*-32	O6	H49
152	4	A	A	ST-10	ST-475	*fumC*-11	*fimH*-27	O4	H54
3038	4	A	A	ST-10	ST-383	*fumC*-11	*fimH*-27	O13, O13/O135	H11
739	5	A	A	ST-10	ST-2	*fumC*-11	*fimH*-27	O74	H39
937	5	A	A	ST-10	ST-2	*fumC*-11	*fimH*-27	O74	H39
435	6	C	A	ST-5880	ST-505	*fumC*-153	*fimH*-444	O26	H9
2117	6	A	A	ST-4087	ST-698	*fumC*-11	*fimH*-444	O127	H4
329	7	B1	B1	ST-297	ST-487	*fumC*-65	*fimH*-38	O179	H8
3109	7	A	A	ST-10	ST-383	*fumC*-11	*fimH*-27	O13, O13/O135	H11
766	8	A	A	ST-10	ST-2	*fumC*-11	*fimH*-54	O9	H17
1608	8	A	A	ST-10	ST-2	*fumC*-11	*fimH*-54	O9	H17
2378	9	D	D	ST-106	ST-3	*fumC*-35	*fimH*-47	O17/O44, O17/O77	H18
2422	9	A	A	ST-10	ST-818	*fumC*-11	*fimH*-27	O45	H11
2421	10	D	D	ST-106	ST-3	*fumC*-35	*fimH*-47	O17/O44, O17/O77	H18
2506	10	D	D	ST-106	ST-3	*fumC*-35	*fimH*-47	O17/O44, O17/O77	H18

4.3.1.6 Variant Calling

SNP phylogeny on the core genome and whole-genome showed that subclinical and clinical forms of persistent strains were clustered within the case (see Figure 4-11). Both phylogeny trees gave almost the same results - only isolate 152 was clustered differently- however, using whole genomes were extremely slow. Hence, core genome phylogeny could be an alternative when computer resources are limited. It was interesting to observe that some other isolates were clustered as well based on the same phenotypes; clinical isolate-2117 with clinical isolate-1998, subclinical isolate-1469 with subclinical isolate-329 and clinical isolate-3038 with clinical isolate-3109.

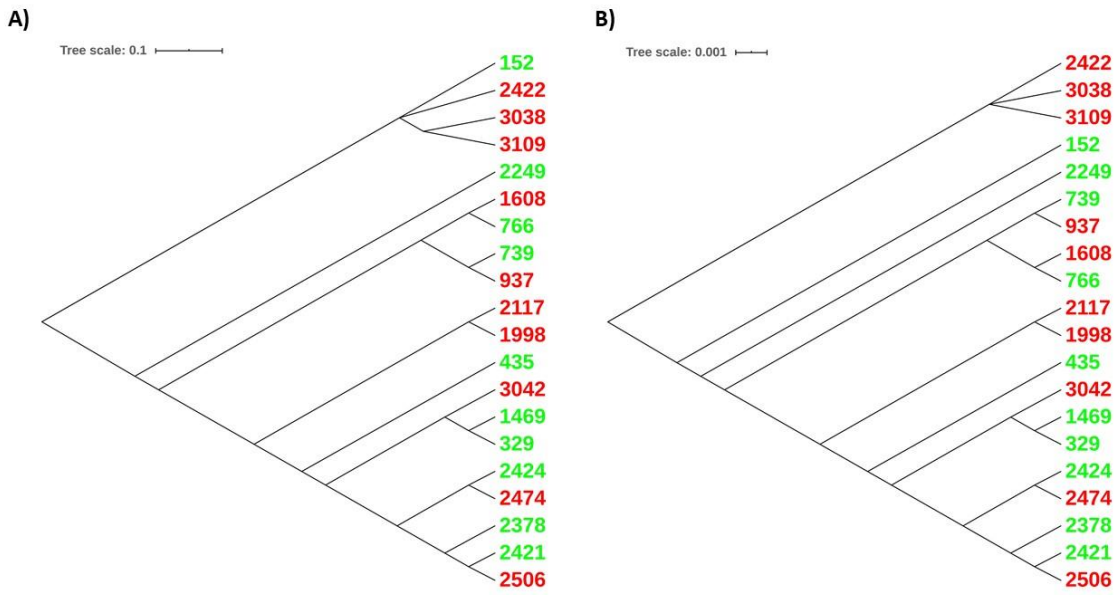


Figure 4-11. SNPs phylogeny analysis of 20 E. coli genomes generated by Snippy pipeline. *The phylogeny was performed based on A) core genome and B) whole genome. The isolates were coloured based on the phenotypes where subclinical ones were green and clinical ones were red. The figures were visualised using iTOL v3.*

4.3.1.7 Detection of Antimicrobial-Resistant Genes

According to ResFinder v3.1, none of the *E. coli* isolates was predicted to be resistant against the following antibiotic classes colistin, fluoroquinolone, fosfomycin, fusidic acid, glycopeptide, nitroimidazole, oxazolidinone and rifampicin (see Figure 4-12). In all *E. coli* isolates multidrug efflux pump *mdf(A)* gene-related with macrolide, lincosamide and streptogramin B was detected. The following AMR genes were detected; *aadA5*, *aph(3')-Ia*, *aph(3'')-Ib* and *aph(6)-Id* as resistant to aminoglycosides; *blaTEM-1B* as resistant to beta-lactams, *sul1* and *sul2* as resistant to sulphonamides, *catA1* as resistant to phenicols, *dfrA16* as resistant to trimethoprim, *tet(B)* as resistant to tetracycline.

The highest amount of AMR genes (11 out of 11 genes) were found in isolates 766 &1608 (persistent strain in animal 8), which had all AMR genes listed. The second-highest (7 genes out of 11) were found in isolate 2378 which had *aph(3')-Ia*, *aph(3'')-Ib* and *aph(6)-Id* genes of aminoglycoside; *blaTEM-1B* of beta-lactams; *sul2* of sulphonamides, *tet(B)* of tetracycline and *mdf(A)* of macrolide, lincosamide and streptogramin B resistance. Isolates 2421&2506 (persistent strain in animal 10) which were found highly similar with isolate 2378 in whole-genome comparison, had all the AMR genes isolate 2378 had except *blaTEM-1B*. The other persistent

strains had the following AMR genes; isolates 2424&2474 (persistent strain in animal 1) and isolates 739&937 (persistent strain in animal 5) happened to carry only *mdf(A)* gene which was detected in the other strains as well.

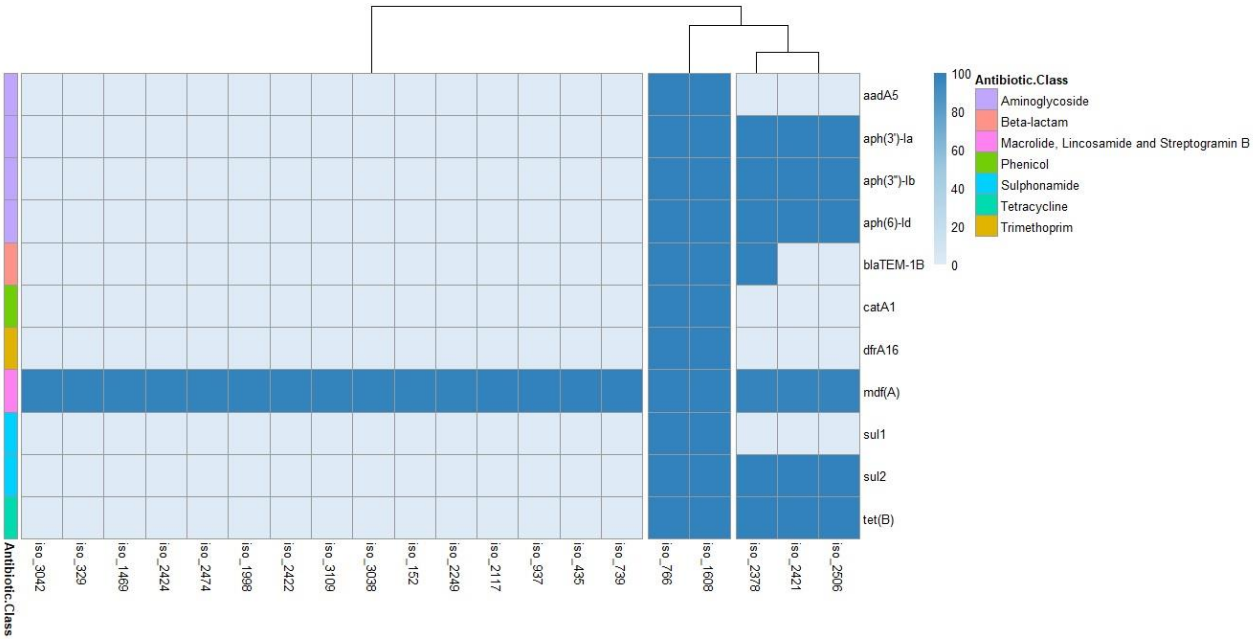


Figure 4-12. AMR genes detected in 20 E. coli isolates against ResFinder v3.1 database. *The shades of blue represent the percentage of identity matches between sequences of AMR gene and the isolate where the brightest is 97.64% and the darkest is 100.00% identity similarity. There were 11 AMR genes found in the E. coli isolates based on ResFinder v3.1 database. This figure was generated using R package pheatmap (Kolde and Kolde, 2015).*

4.3.1.8 Detection of Virulence Genes

There were 11 virulence factors in total (*air*, *astA*, *eilA*, *espP*, *gad*, *iroN*, *iss*, *lpfA*, *mchB*, *mchC* and *mchF*) that were found in our *E. coli* isolates (see Figure 4-13). *Iss* was the most common virulence factor found in the isolates. The highest amount of virulence factors (7 genes) were found in isolates 2424&2474 (persistent strain in animal-1), which had *asta*, *gad*, *iroN*, *iss*, *mchB*, *mchC* and *mchF*. The second-highest amount of virulence factors (6 genes; *air*, *eilA*, *espP*, *gad*, *iss* and *lpfA*) were found in isolates 2421&2506 (persistent strain in animal-10) and isolate 2378. Persistent strain in animal-8 (isolates 766&1608) had 4 virulence factors which were *gad*, *iroN*, *iss* and *mchF*; whereas persistent strain in animal-5 (isolates 739&937) had none of the virulence factors.

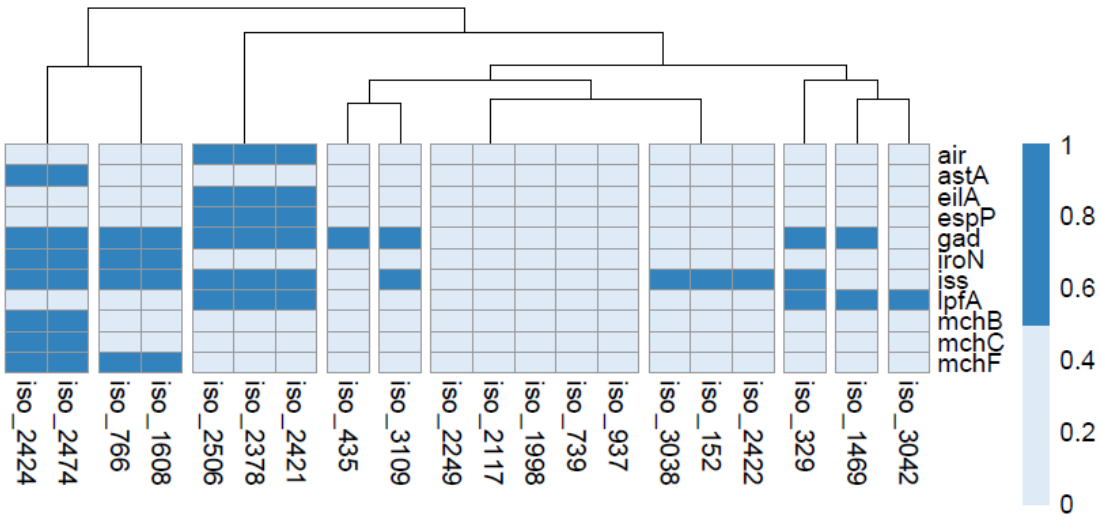


Figure 4-13. Virulence factors detected in 20 E. coli isolates against VirulenceFinder 2.0 database. *The shades of blue represent the presence/absence of the gene where the dark colour means presence, whilst light one means absence. There were a total of 11 virulence factors found in 20 E. coli isolates. This figure was generated using R package pheatmap (Kolde and Kolde, 2015).*

4.3.1.9 Detection of Plasmids

In this study, no hits were found from the gram-positive plasmids database as expected but there were hits in the *Enterobacteriaceae* database. Amongst persistent strains of this study isolates 2424-2474 (animal-1) had no hits, isolates 739&937 (animal-5) had three hits (IncFIA, IncFIB(AP001918), IncFII(29)), isolates 766&1608 (animal-8) had three hits (IncFIB(AP001918), InFIC(FII), Col440II) and isolates 2421&2506 had three hits IncFIA, IncFIB(AP001918), InFIC(pCoo). All persistent strains had the same plasmids with the same identity match between their subclinical and clinical forms (see Table 4-5).

Clinical mastitis-causing isolate 3109 was found to have the most hits with 4 plasmids, 2 of them were the same with subclinical mastitis-causing isolate 329 collected from the same cow (animal 7). Amongst remaining isolates, two isolates (isolates 2249 and 2378) had three plasmid hits, one of them was isolate 2378, which is a closely related sample with persistent strain in animal 10 and had the same plasmids with them (isolates 2421&2506). Two isolates (isolate 152 and 329) had two plasmid hits and three isolates (isolate 2422, 3038 and 3042) had one hit each while four isolates (isolate 435, 1469, 1998 and 2117) had no hits in the plasmid database. IncF plasmid family was the most commonly found one with several types of it (IncFIA(HII), IncFIB(K), IncFIB(pECLA), IncFIB(AP001918), IncFIC(FII), IncFIC(pCoo) and IncFII(29)) in our study.

Table 4-5. List of plasmids that were found in 20 E. coli genomes.

Isolate	Animal	Status	Farm	Plasmid
2424	1	Subclinical	H	-
2474	1	Clinical	H	-
1469	2	Subclinical	S	-
1998	2	Clinical	S	-
2249	3	Subclinical	W	IncFIA(HI1), IncFIB(K), IncY
3042	3	Clinical	W	Col3M
152	4	Subclinical	N	IncFIB(pECLA), p0111
3038	4	Clinical	N	IncY
739	5	Subclinical	S	IncFIA, IncFIB(AP001918), IncFII(29)
937	5	Clinical	S	IncFIA, IncFIB(AP001918), IncFII(29)
435	6	Subclinical	N	-
2117	6	Clinical	N	-
329	7	Subclinical	W	IncFIB(AP001918), IncFIC(FII)
3109	7	Clinical	W	IncFIA, IncFIB(AP001918), IncFIC(FII), IncY
766	8	Subclinical	B	IncFIB(AP001918), InFIC(FII), Col440II
1608	8	Clinical	B	IncFIB(AP001918), InFIC(FII), Col440II
2378	9	Subclinical	F	IncFIA, IncFIB(AP001918), InFIC(pCoo)
2422	9	Clinical	F	IncY
2421	10	Subclinical	F	IncFIA, IncFIB(AP001918), InFIC(pCoo)
2506	10	Clinical	F	IncFIA, IncFIB(AP001918), InFIC(pCoo)

4.3.1.10 Detection of Prophages

Prophages were present in all of the genomes as expected; however, there was no pattern detected in terms of phage types present between clinical and subclinical clusters either within all isolates or persistent strains alone. In persistent cases following count of intact prophage regions were found. Persistent strain in animal-1 had 6 and 8 intact prophage regions found in isolate 2424 and 2474, respectively. Persistent strain in animal-5 had 3 intact prophage regions in both isolate 739 and 937. Persistent strain in animal-8 had 1 intact prophage region in both isolate 766 and 1608. Persistent strain in animal 10 had 7 and 9 intact prophage regions in isolate 2421 and 2506, respectively. The difference counts between the isolates labelled as the same strain in some animals (animal-1 and animal-10) may be a result of incompleteness score based on coverage and sequence quality.

4.3.2 Analysis of Subclinical and Clinical Phenotypes of Persistent Strains

We investigated the possibility to develop a classifier to verify if the MALDI-TOF peak list associated with isolates could be used to predict their phenotype (SCM or CM) of the same

genotype. Classifiers were developed by categorizing persistent 4 subclinical isolates (isolate 2424, isolate 739, isolate 766 and isolate 2421) which had 24 spectra (6 technical replicates of each isolate) as the positive class and persistent 4 clinical isolates (isolate 2474, isolate 937, isolate 1608 and isolate 2506) which had 24 spectra (6 technical replicates of each isolate) as the negative class. The pre-processing led to the identification of two peaks that are statistically different between groups and appear in at least 30% of the total spectra. These two peaks were then used as features to build ten classifiers to develop predictive models for the different phenotypes of persistent strains. 30 runs using NCV was performed.

Amongst ten different classification algorithms LR, LSVM, AdaBoost, NB and LDA gave accuracy over 75.00%. LR gave the best prediction performance values as follows; accuracy: 85.00±5.09%, AUC: 90.83±12.25%, sensitivity: 95.00±10.17%, specificity: 75.00±0.00% and kappa: 70.00±10.71% (see Figure 4-14).

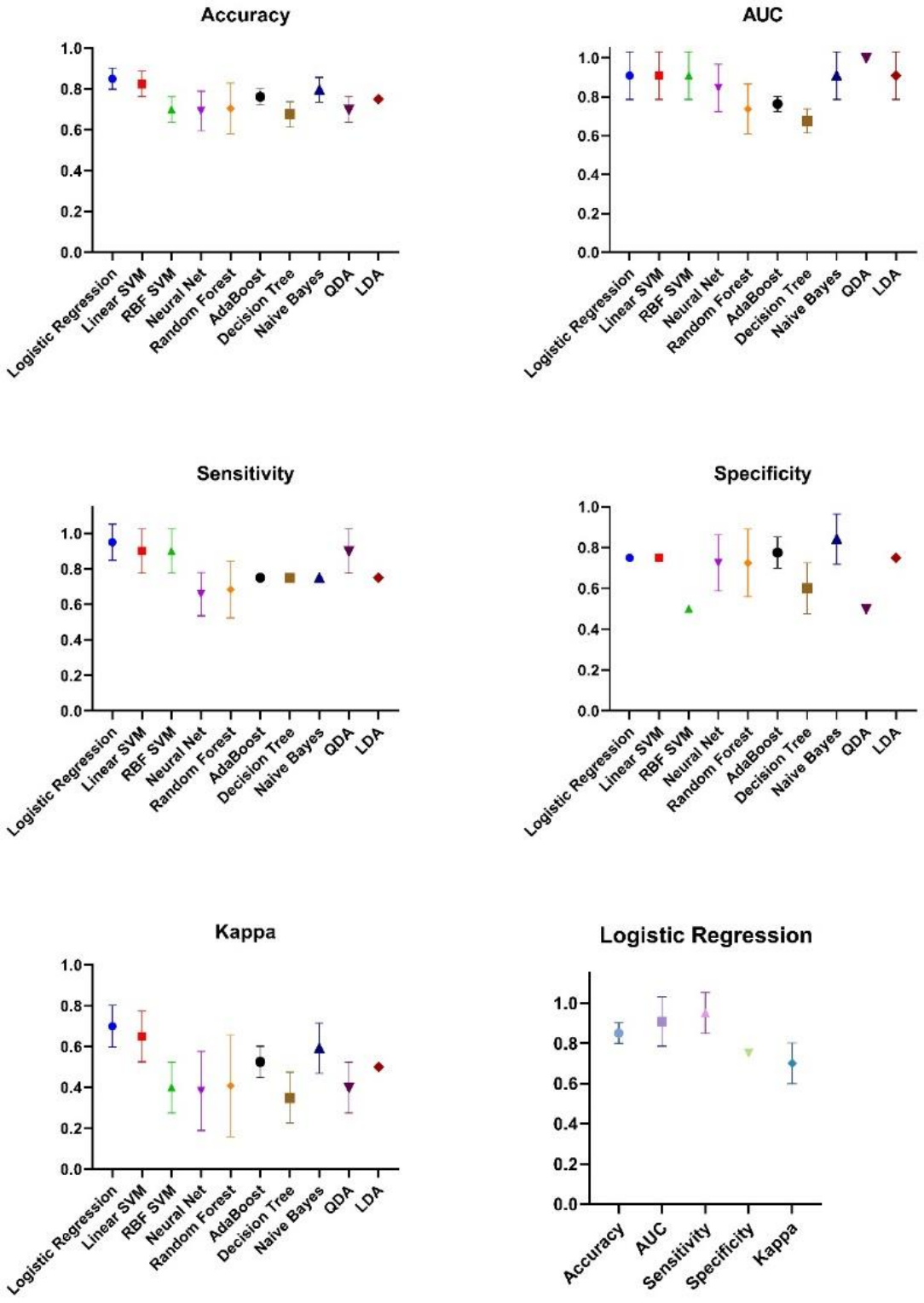


Figure 4-14. Prediction performance results of classifiers for subclinical vs clinical phenotypes of persistent E. coli strains. *Ten different algorithms (X-axis) were used to classify phenotype profiles. Accuracy, AUC, sensitivity, specificity, kappa metrics were calculated for each learner. Moreover, these metrics were shown for logistic regression which was the best performing classifier amongst employed ones. These graphs were generated in GraphPad Prism v8.*

4.3.3 Analysis of Persistent and Non-Persistent *E. Coli* Strains

We, then, investigated the possibility to develop a classifier to verify if the MALDI-TOF peak list associated with isolates could be used to predict persistent and non-persistent *E. coli* strains. Classifiers were developed by categorizing persistent 4 subclinical isolates (isolates 2424, 739, 766 and 2421) which had 24 spectra (6 technical replicates of each isolate) as the positive class and non-persistent 6 subclinical isolates (isolates 1469, 2249, 152, 435, 329 and 2378) which had 36 spectra (6 technical replicates of each isolate) as the negative class. Clinical isolates of persistent and nonpersistent isolates were not involved to prevent the following biases; 1) adding clinical isolates of persistent strains would increase the sensitivity as they are almost identical, 2) clinical isolates of so-called non-persistent strains may have shown persistency which was not detected in during study period. The pre-processing led to the identification of six peaks that are statistically different between groups and appear in at least 30% of all number of spectra. These six peaks were then used as features to build ten classifiers to develop predictive models for persistent and non-persistent strains. 30 runs using NCV was performed.

Amongst ten different classification algorithms LR, LSVM, NB, RBF SVM and RF gave accuracy over 75.00%. Both LR and LSVM gave the best prediction performance values as follows; accuracy: 85.71±4.98%, AUC: 92.50±9.63%, sensitivity: 75.00±0.00%, specificity: 94.17±6.34% and kappa: 68.17±7.46% (see Figure 4-15).

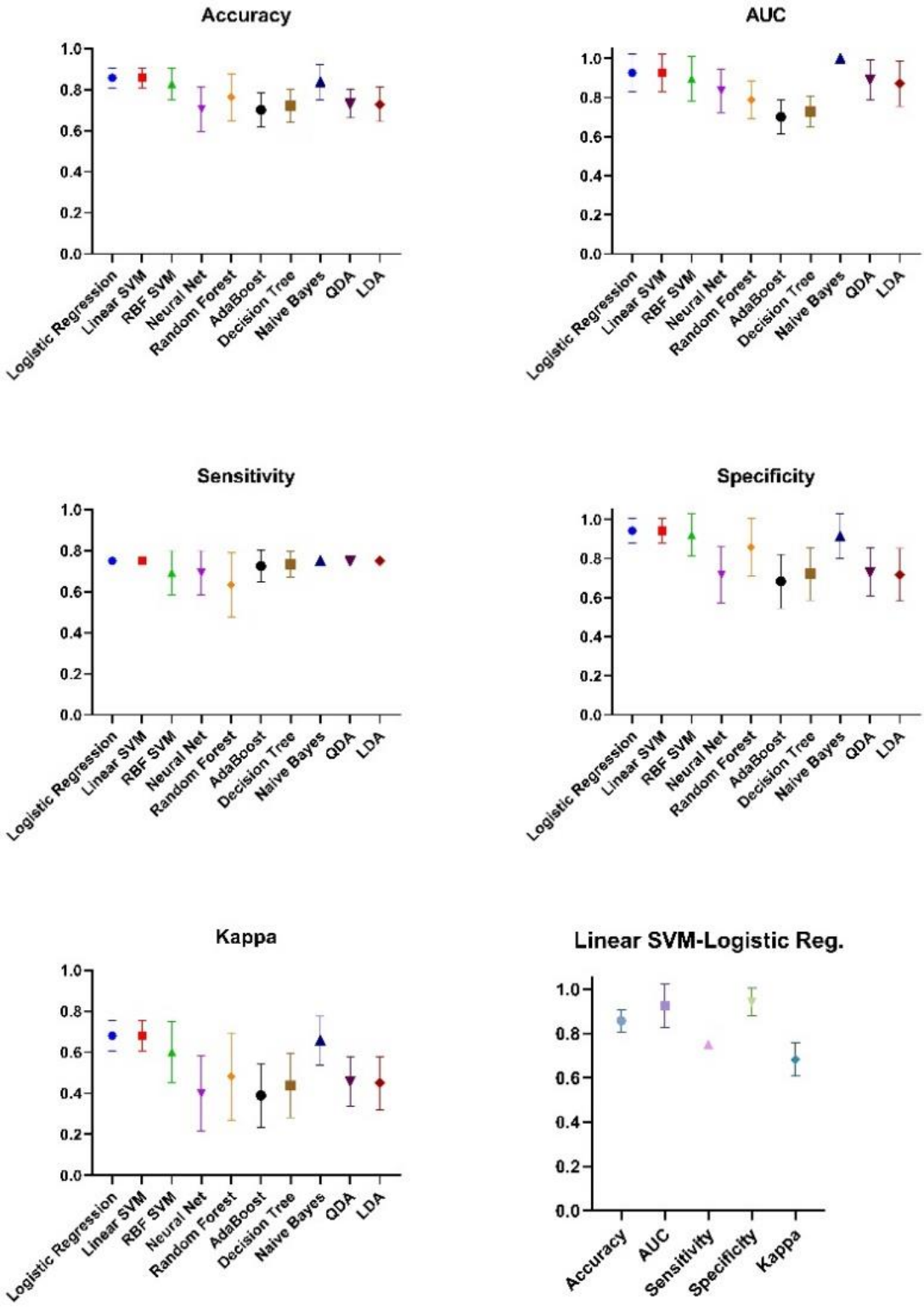


Figure 4-15. Prediction performance results of classifiers for persistent vs non-persistent E. coli strains. *Ten different algorithms (X-axis) were used to classify the persistence profile of E. coli isolates. Accuracy, AUC, sensitivity, specificity, kappa metrics were calculated for each learner. Moreover, these metrics were shown for LSVM and logistic regression which were the best performing classifiers amongst employed ones. These graphs were generated in GraphPad Prism v8.*

4.3.4 Biomarker Characterisation

4.3.4.1 Biomarker Characterisation for Phenotypic Profiles of Persistent Strains

Two peaks identified as providing optimal discrimination between subclinical and clinical phenotypes of persistent strains were further analysed to identify their correspondent *E. coli* proteins. When compared with Prokka annotations of persistent *E. coli* strains, two peaks could be cross-matched with 50S ribosomal protein L35 (RpmI) and DNA gyrase inhibitor (YacG) within a maximum of 0.2% difference as molecular weight (see Table 4-6). 3D models of these discriminant proteins are shown in Figure 4-16.

Table 4-6. Top PSI-BLAST, conserved domain search and cellular location results for the two discriminant proteins between subclinical and clinical phenotypes of persistent E. coli strains.

MALDI-TOF Peak (Mw)	Protein (Mw)	PSI-BLAST Match	Identity	e-value	Domain (e-value)	PSORTB location (score)
7149.66Da	RpmI (7157.74Da)	50S ribosomal protein L35	100.00%	2e-37	Ribosomal_L35p (1.44e-21)	Cytoplasmic (9.26)
7171.62Da	YacG (7174.98Da)	DNA gyrase inhibitor	100.00%	1e-39	DNA gyrase inhibitor YacG (2.74e-38)	Unknown

According to GO, YacG was found to be involved in DNA-templated regulation of transcription (BP), negative regulation of DNA topoisomerase activity (BP), zinc-ion binding (MF), DNA topoisomerase type II inhibitor activity (MF), metal ion binding (MF) and cytosol (CC); whereas RpmI was found to be involved in translation (BP), structural constituent of ribosome (MF) and cytosolic large ribosomal subunit (CC).

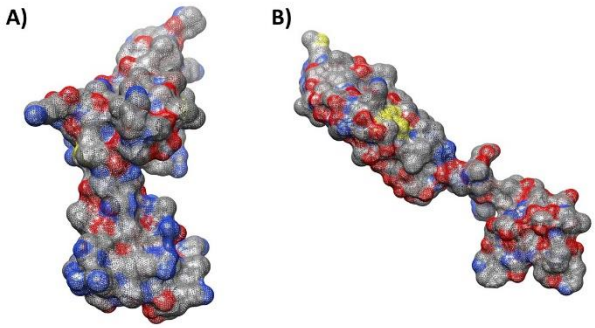


Figure 4-16. 3D models of discriminant proteins between subclinical and clinical phenotypes of persistent strains. *The models are created based on the homology structure of* ***A)*** *RpmI and* ***B)*** *YacG. The visualisation was carried out with UCSF Chimera.*

PPI network analysis of 2 discriminant proteins (RpmI and YacG) and their 17 first neighbour proteins (19 in total) are shown in Figure 4-17.

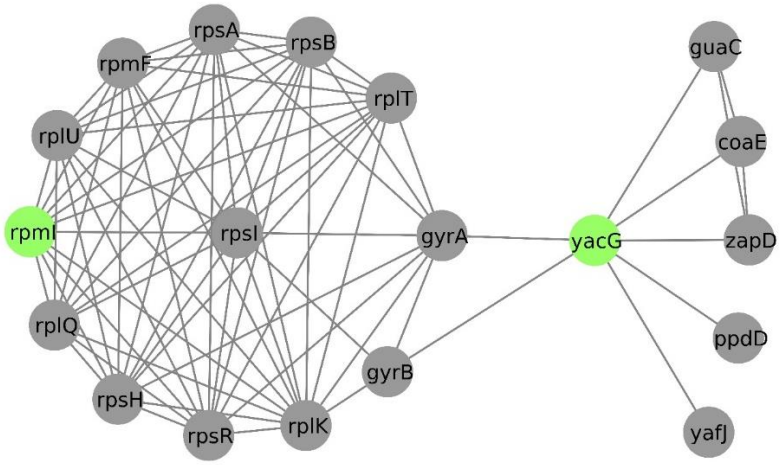


Figure 4-17. Protein-protein interaction (PPI) network related to the phenotypic profile of clinical status. *The PPI shows two discriminant proteins (RpmI and YacG) coloured in green with 17 neighbouring proteins coloured in grey. The visualisation was carried out with Cytoscape.*

4.3.4.2 Biomarker Characterisation for Persistence Profiles

Six peaks identified as providing optimal discrimination between persistent and non-persistent strains were further analysed to identify their correspondent *E. coli* proteins. When compared with Prokka annotations of analysed *E. coli* strains, six peak masses identified 30S ribosomal protein S19 (RpsS), protein YihD and four hypothetical proteins.

To further characterise the function of these proteins we did a PSI-BLAST comparative analysis (see Table 4-7), where one of the hypothetical proteins were found to belong YncJ protein family. In further analysis, hypothetical protein-3 (HP-3) was called YncJ.

Table 4-7. Top PSI-BLAST, conserved domain search and cellular location results for the six discriminant proteins between persistent and non-persistent E. coli strains.

MALDI-TOF Peak (Mw)	Protein (Mw)	PSI-BLAST Match	Identity	e-value	Domain (e-value)	PSORTB location (score)
4173.95Da	HP-1 (4171.89Da)	Hypothetical protein	100.00%	3e-21	No domain was found.	Unknown
6355.80Da	HP-2 (6359.51Da)	Hypothetical protein	100.00%	7e-29	No domain was found.	Unknown
6444.01Da	HP-3 (6453.04Da)	YncJ family protein	100.00%	2e-33	DUF2554	Unknown
9189.62Da	HP-4 (9182.58Da)	Excisionase family protein	100.00%	5e-52	DUF1233	Unknown
10256.71Da	YihD (10272.91Da)	YihD family protein	100.00%	1e-57	DUF1040	Cytoplasmic (8.96)
10298.83Da	RpsS (10299.09Da)	30S ribosomal protein S19	100.00%	8e-61	Ribosomal-S19	Cytoplasmic (9.26)

HP: Hypothetical protein, DUF: Domain of unknown function.

According to GO, YihD was found to be involved in the cytosol (CC); whereas no BP or MF terms were assigned for them. RpsS was found to be involved in translation (BP), ribosomal small subunit assembly (BP), rRNA binding (MF), structural constituent of ribosome (MF), rRNA binding (MF) and cytosolic large ribosomal subunit (CC).

GO terms were predicted by using 3D threading protein models of hypothetical proteins which resulted as follows (see Figure 4-18). For HP-1: DNA-binding (MF), DNA-binding transcription factor activity (MF), regulation of transcription (BP), regulation of DNA replication (BP) were found whereas none could be predicted for CC. For HP-2: ATP binding (MF), transporter activity (MF), toxin binding (MF), transport (BP), response to antibiotic (BP), bacteriocin immunity (BP) and plasma membrane (CC) were found. For HP-3: oxidoreductase activity (MF), transition metal ion binding (MF), establishment of localization (BP), cellular respiration (BP) and mitochondrial membrane (CC) were found. For HP-4: nucleotide binding (MF), regulation of transcription (BP) and intracellular part (CC) were found.

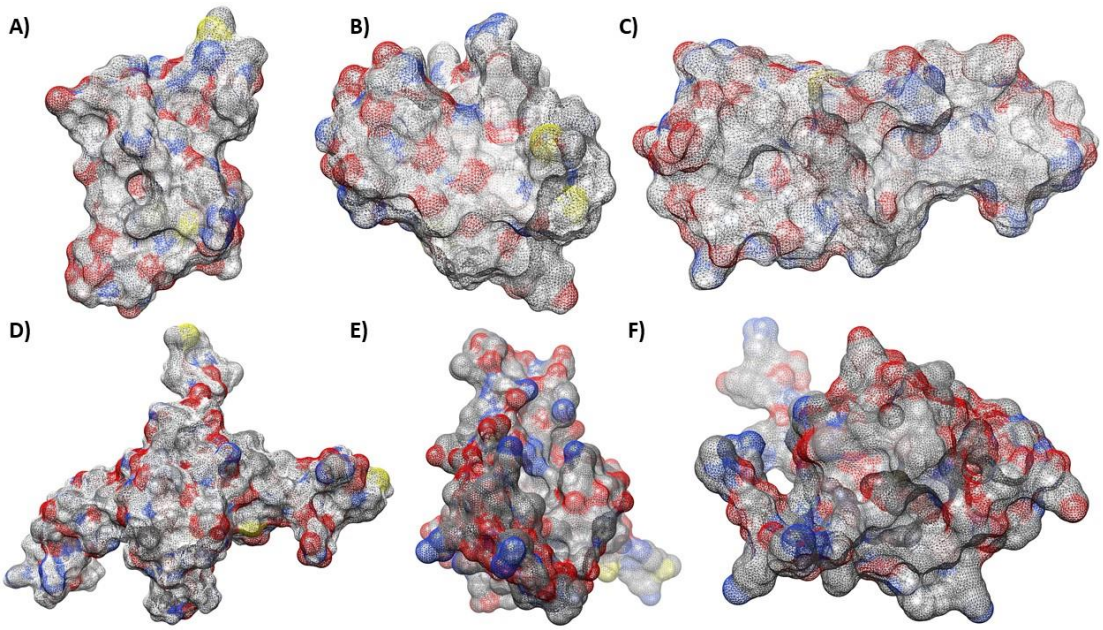


Figure 4-18. 3D models of discriminant proteins between persistent and non-persistent E. coli strains. *The models are created based on threading* ***A)*** *HP-1,* ***B)*** *HP-2,* ***C)*** *HP-3,* ***D)*** *HP-4 and homology structure of* ***E)*** *protein YihD and* ***F)*** *RpsS. The visualisation was carried out with UCSF Chimera.*

PPI network analysis could be performed for three (RpsS, YihD and YncJ) out of six proteins as no interaction of the other three hypothetical proteins could be found in the STRING database. PPI analysis of 3 discriminant proteins and their 30 first neighbour proteins (33 in total) are shown in Figure 4-19.

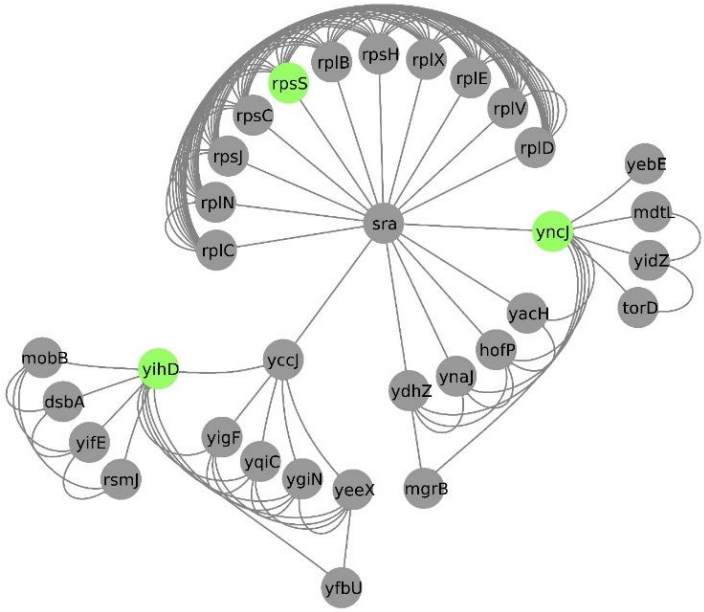


Figure 4-19. Protein-protein interaction (PPI) network related to persistence profile. *The PPI shows three discriminant proteins (RpsS, YncJ and YihD) coloured in green with 30 neighbouring proteins coloured in grey. The visualisation was carried out with Cytoscape.*

4.3.5 Functional Enrichment Analyses

Functional enrichment analyses were performed for discriminant proteins of persistency (persistence vs non-persistent) and phenotype (subclinical vs clinical) profiles and their neighbours, separately. Some of the significant annotations can be seen in Figure 4-20.

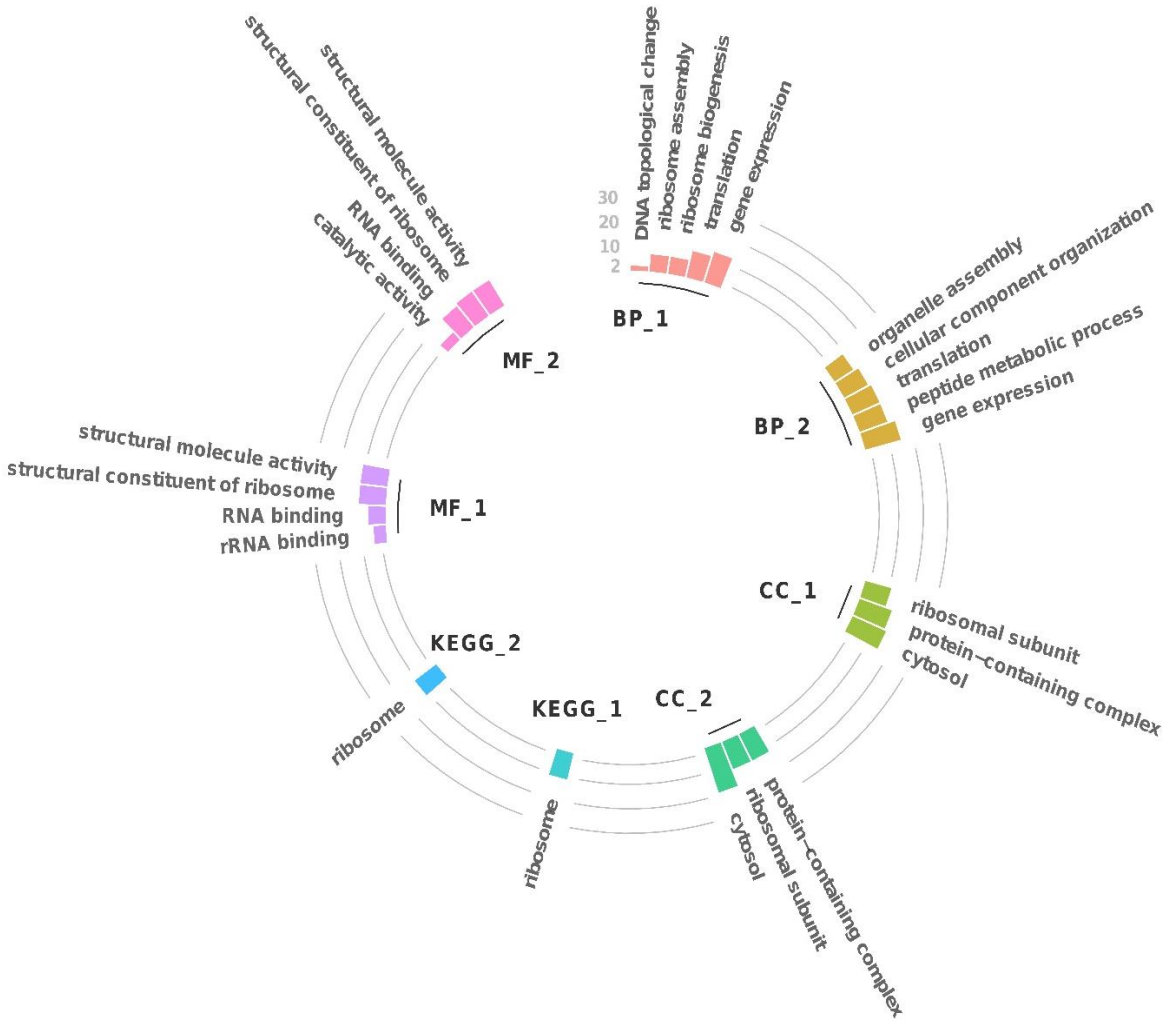


Figure 4-20. Functional enrichment analysis of phenotypic profile discriminatory network and persistency discriminatory network based on Gene Ontology and KEGG pathways. *Biological pathway (BP), molecular function (MF), cellular component (CC) and KEGG were indexed 1 for the phenotypic profile discriminatory network (subclinical vs clinical) and 2 for persistency discriminatory network (persistent vs non-persistent). This figure was generated using R package tidyverse (Wickham et al., 2019).*

4.4 DISCUSSION

This study used the recurrent mastitis samples obtained during a previous study (Bradley and Green, 2000). Recurrent clinical episodes can be a result of two different scenarios; re-infection

of the mammary gland by pathogens present in the environment or persistent infection with the same pathogen. The first objective in this work was to perform whole-genome sequencing for validation of the persistent *E. coli* samples which were collected at different time points (pre-calving and post-calving) in the same quarter. Genotypically unique strains were thought to be the result of persistent infection as there is a wide range of *E. coli* pathogens in the environment and unlikely to be re-infected with the same genotype.

Various genome comparison analyses such as FastANI, BRIG, ACT, SNPs typing etc., were performed to confirm two isolates that were collected from the same quarter in different time points were identical. The high similarity-based whole-genome analysis showed us that there were four identical cases in animal-1, animal-5, animal-8 and animal-10 which were labelled as persistent, whereas other isolates were labelled as nonpersistent and were still used to explore potential genotypic pattern between the different clinical status of mastitis pathogens.

Although the increased accessibility of the whole-genome sequencing technology, MLST is still widely used for certain epidemiological studies, as it is easy to track the ST in literature due to well-established databases (Zhou *et al.*, 2020). In this study, the MLST of 20 *E. coli* genomes were found based on Achtman and Pasteur scheme. As the Achtman scheme is the most frequently used MLST technique around the world, the findings of the Achtman scheme were compared with the literature. Amongst 20 *E. coli* isolates, the most common MLST type was ST-10, which was detected in two persistent strains (isolates 739&937 in animal-5 and isolates 766&1608 in animal-8) as well as six other strains (isolates 152, 1469, 2249, 2422, 3038 and 3109). ST-10 is a widely isolated strain from humans and farm animals (Izdebski *et al.*, 2013), and classified as zoonotic extended-spectrum beta-lactamase-producing *E. coli* (ESBL) (Lazarus *et al.*, 2015). ST-10 has been previously found in Shiga-toxin-producing *E. coli* (STEC) isolated from bovine hides and carcases in Ireland (Monaghan *et al.*, 2012; Feng *et al.*, 2017), from pigs in Australia (Kidsley *et al.*, 2018), from cattle in Tunisia (Grami *et al.*, 2014), from bovine mastitis diagnosed cows in China (Li *et al.*, 2011). Moreover, ST-10 was the most widely isolated clone from Israeli dairy cows (Blum and Leitner, 2013; Lifshitz *et al.*, 2018) and found to be more prevalent in the cows than their environment (Blum and Leitner, 2013). Persistent strain in animal-1 (isolates 2424&2474) was found to be ST-57, which was previously isolated from peritonitis syndrome outbreaks in chicken flocks (Landman *et al.*, 2014). ST-57 was detected in avian pathogenic *E. coli* (APEC) and uropathogenic *E. coli* (UPEC) strains as well (Lau *et al.*, 2008; Hussein *et al.*, 2013). Persistent strain in animal-10 (isolates 2421&2506) was found to be ST-106. ST-106 was found in enterohemorrhagic *E. coli*

(EHEC) strains isolated from calves and humans from divergent locations such as the US, Germany, Cuba and Scotland (Abu-Ali *et al.*, 2009). Isolate 3042 was found to be ST-1079, which was previously detected in ESBL, MBL (metallo-beta-lactamase) and AmpC beta-lactamase-producing isolates collected from livestock and poultry in India (Govindaraj *et al.*, 2019). Isolate 329 was found ST-297, which was previously isolated from cattle in Tanzania (Madoshi *et al.*, 2016), calves in India (Murugan *et al.*, 2019), cows and chickens in China (Chen *et al.*, 2018), verotoxin-producing *E. coli* (VTEC) strain from cheese in Kosovo (Nagy *et al.*, 2015) and from ESBL-producing isolates from wild birds in Tunisia (Ben Yahia *et al.*, 2018). Isolate 1998 was found ST-685, which was previously isolated from healthy swine in Thailand (Seenama, Thamlikitkul and Ratthawongjirakul, 2019). Isolate 2117 and 435 were found ST-4087 and ST-5880, respectively. No *E. coli* isolate with this ST types was seen in the literature. There were 9635 different STs in the pubMLST database for the Achtman scheme on the day of analysis, some of which belonged to one of the 56 clonal complexes. As the name states, ST-10 was the main type of ST-10 clonal complex which contained pathogenic or non-pathogenic strains from different sources. ST-106 was the furthest type to ST-10 with 6 loci variants. ST-106 belonged to ST-69 clonal complex which was isolated from sources including cattle around the world but mostly ESBL-producing strains (Lifshitz *et al.*, 2018; Gordon, 2013; Madoshi *et al.*, 2016). Strains of ST-69 observed in different studies were shown to be persistent with good ability to colonize and adapt different hosts (Alghoribi *et al.*, 2014; Gibreel *et al.*, 2010). ST-57 belonged to ST-350 clonal complex while other STs (ST-5880, ST-685, ST-1079, ST-297, ST-4087) did not belong to any clonal complex yet although there was only single locus variant (SLV) between ST-685 and ST-10. These findings showed that there was no specific ST type to define bovine mastitis causing *E. coli* strains as the same ST types could be isolated from various range of host and source.

It was interesting to see that highly similar but not identical isolates (isolate 3038 &3109; isolate 2378 with persistent strain in animal 10) based on whole-genome sequencing analysis (i.e FastANI) were typed as same ST according to both MLST schemes (Achtman and Pasteur scheme). One can comment that MLST is highly accurate for true negatives but with high rates of false positives. Hence, MLST should be used carefully for strain detection studies.

MLST needs to screen seven or more housekeeping genes for strain typing. As the number of loci to screen increases, the analysis time and price of the analysis increases (Overdevest *et al.*, 2012). Hence, researchers developed new typing techniques which can screen fewer loci and give the same discriminatory power with the MLST, one of them was CH typing based on two

loci: *fumC* and *fimH*. Housekeeping genes such as *fumC* evolve slowly as they are under neutral or stabilizing selection pressures (Jolley and Maiden, 2014). Instead, genes like *fimH* (type 1 adhesin) evolve quickly to adapt regarding positive selection pressure and suggested to give better performance to differentiate strains (Schwartz *et al.*, 2013). In an extensive study, almost a thousand *E. coli* isolates including commensal/pathogenic model isolates (ECOR) and newly collected isolates were CH typed (Weissman *et al.*, 2012). Comparing our results with the results of that study; the most common CH type 11-27 of our analysis were also detected in laboratory strains MG1655 and W3110, ECOR12, ECOR14 and ECOR24, which were all from Swedish human hosts (Ochman and Selander, 1984). There were three model organisms isolated from steer and two of them were found to be *fumC11*, which was the most common *fumC* type in our study. Another CH type found in one of our persistent isolates was *fumC35-fimH47*, which was also detected in the ECOR47 strain, sheep faecal from New Guinea. Other CH types found in our persistent cases *fumC11-fimH54* and *fumC31-fimH54* were not detected in any isolates of that particular study.

In the case of an outbreak, it is essential to define which pathogenic serotype the *E. coli* isolate belongs to. Serotyping has been widely used since it was first discovered and has become the gold standard (Kauffmann, 1947). Conventional serotyping relies on three immunogenic units, which are lipopolysaccharide (O antigen), capsular (K antigen) and flagellar (H antigen) (Orskov *et al.*, 1977). As the nature of this technique which screens the immunogenic structures unlike other typing methods such as PFGE, ribotyping or MLST; serotyping is the most accurate to be used (Joensen *et al.*, 2015). Serotyping result of isolates 2424&2474 (persistent strain in animal-1) was found to be O176:H32. In the literature, not much was found related to this serotype however O176 was detected in VTEC isolated from calves (Scheutz *et al.*, 2004). Serotyping result of isolates 739&937 (persistent strain in animal-5) was found to be O74:H39, which was previously observed in STEC isolated from Argentinian dairy cows (Fernández *et al.*, 2010) and beef (Constantiniu, 2002), enterohemorrhagic strains from bovine faeces (Abdulmawjood *et al.*, 2003). Notably, this serotype was seen in another persistent mastitis-causing *E. coli* isolated from a dairy farm in New York, which shared common virulence factors with ExPEC (Dogan *et al.*, 2012). Serotyping result of isolates 766&1608 (persistent in animal-8) was found O9:H17, which was previously found in ESBL-producing *E. coli* (Verschuuren *et al.*, 2020). Serotyping result of isolates 2421&2506 (persistent strain in animal-10) was found to be O17:H18, which was previously found in UPEC and STEC (Wallace-Gadsden *et al.*, 2007; Eklund, Scheutz and Siitonen, 2001). Serotyping result of isolates 3038&

3109 was found to be O135:H11, which was previously found in *E. coli* strains isolated from faeces of feedlot cattle (Diarra *et al.*, 2009). The serotype of isolate 2422 was O45:H11, which was previously found in EPEC, non-pathogenic, UPEC and attaching-effacing *E. coli* (AEEC) strains as well (Delannoy, Beutin and Fach, 2012; Malik *et al.*, 2017; Fröhlicher *et al.*, 2008; Paniagua-Contreras *et al.*, 2019). The serotype of isolate 1469 was Ont:H16, which was also isolated from a healthy dairy cow (Houser *et al.*, 2008). The serotype of isolate 2249 was O107:H54, which was also found in isolates from urban impacted coastal waters (Fernandes *et al.*, 2020). The serotype of isolate 152 was O4:H27, which was also found in isolates related to pigs with diarrhoea (Boerlin *et al.*, 2005). The serotype of isolate 1998 was found to be Ont:H34, which was also detected in EPEC strains from Australia and Brazil (Nguyen *et al.*, 2006; Gomes *et al.*, 2004); STEC strains from animals including cattle (Constantiniu, 2002); and VTEC strains from bovine skin and carcasses (Denis, Wieczorek and Osek, 2014). The serotype of isolate 2117 was found to be O127:H4, which was observed in EPEC (Jensen *et al.*, 2007), STEC (Tozzoli *et al.*, 2014), enteroaggregative hemorrhagic *E. coli* (EAHEC) (Dallman *et al.*, 2012) and commensal isolates (Ahmed, Olsen and Herrero-Fresno, 2017). The serotype of isolate 3042 was O6:H49, which was previously observed in STEC strains isolated from healthy beef and dairy cattle (Castro *et al.*, 2019). The serotype of isolate 329 was O179:H8, which was found in STEC isolated mostly from dairy cattle, beef cattle and sheep (Vimont, Delignette-Muller and Vernozy-Rozand, 2007; Beutin and Strauch, 2007; Fremaux *et al.*, 2006; Hornitzky *et al.*, 2005; Castro *et al.*, 2019); VTEC isolated from various sources such as raw milk, milk filter and meat (Scheutz *et al.*, 2004); and EPEC isolated from cheese (Júnior *et al.*, 2019). Serotyping result of isolate 435 was found to be O26:H9, which was previously detected in STEC isolated from cattle faeces (Stanford *et al.*, 2018), cattle itself and its environment (de Souza Figueiredo *et al.*, 2019). Overall, our results were consistent with the findings of Wenz *et al.*, (2006) where they failed to detect any predominant serotype in their bovine-mastitis causing *E. coli* strains.

Persistent strains in animal-1 (isolates 2424&2474), animal-5 (isolates 739&937), animal-8 (isolates 766&1608) and animal-10 (isolates 2421& were found belonging to phylogroup E, A, A and D, respectively. Amongst the remaining isolates, phylogroup A was the most common which was found in seven isolates (isolate 152, 1998, 2117, 2249, 2422, 3038 and 3109), while three isolates (isolate 329, 1469 and 3042) were found in phylogroup B1. This was consistent with the *E. coli* strains isolated from bovine mastitis cases in various geographical locations (Zude, 2014; Ghanbarpour and Oswald, 2010; Blum and Leitner, 2013; Dogan *et al.*, 2006).

There are mainly commensal, transient and acute mastitis-causing strains in phylogroup A (Dogan *et al.*, 2006; Blum *et al.*, 2015). However, persistent mastitis strains were found in this group as well (Blum *et al.*, 2015). Phylogroup B1 strains are also mainly commensal or intestinal pathogens (Gordon and Cowling, 2003). However, transient and persistent mastitis-causing strains were also observed in this group (Dogan *et al.*, 2012). It was a common belief that cow factors are the main drivers of mastitis severity rather than pathogenic factors (Burvenich *et al.*, 2003). This may be true for animal-5 and animal-8 as persistent strains in these animals were found in phylogroup A and this persistency may be a result of their weak immune system. However, persistent strains in animal-1 and animal-10 were found in phylogroup E and D, respectively, which also contained highly pathogenic *E. coli* strains such as EHEC serotype O157:H7 and enteropathogenic (EPEC) serotype O55:H7 in phylogroup E (Cooper *et al.*, 2014); extra-intestinal pathogenic (ExPEC), enteroaggregative *E. coli* (EAEC) and highly virulent mammary pathogenic strain in phylogroup D (Olson et al., 2018). Dogan *et al.*, (2012) also found persistent mastitis-causing *E. coli* strains which belonged to phylogroup D. Previously, mastitis associated *E. coli* strain was seen in phylogroup E which was also rare compared to phylogroup A and B1 (Leimbach *et al.*, 2017). Although certain phylogroups dominate the specific hosts and habitats, other factors such as changes in the environment or selection pressure may result in dominancy of less abundant clades (Smati *et al*., 2013; Scholz *et al*., 2016).

Plasmids are important elements to circulate virulence factors and AMR in the bacterial population (Guardabassi and Courvalin, 2006). In our study, IncF plasmid family was the most commonly found (IncFIA(HII), IncFIB(K), IncFIB(pECLA), IncFIB(AP001918), IncFIC(FII), IncFIC(pCoo) and IncFII(29)). This plasmid family capable of carrying transfer, multidrug resistance and virulence factors (Johnson and Nolan, 2009). IncFIB(AP001918) was previously found in *E. coli* samples isolated from retail meat (Falgenhauer *et al.*, 2016), cattle-sourced ESBL (Adator *et al.*, 2020), cattle faeces (Bumunang *et al.*, 2019), human and domestic animals (Salinas *et al.*, 2019). IncFIA was previously found in ESBL-producing *E. coli* samples from bovine mastitis cases around the world (Ali *et al.*, 2017; Freitag *et al.*, 2017; Afema *et al.*, 2018). IncFIC and IncFII plasmids were previously found in multidrug-resistant *E. coli* isolates from cattle with suspected mastitis infections in France and Germany (Brennan *et al.*, 2016). IncY (phage-like plasmid), p0111, col440II and col3M were another plasmid type detected in this analysis. IncY plasmid was another common type isolated from bovine mastitis cases around different locations in China (Ali *et al.*, 2017).

Many attempts have been made to identify virulence factors of bovine-mastitis causing *E. coli* strains. Although several studies (Fernandes *et al.*, 2011; Suojala *et al.*, 2011; Blum *et al.*, 2015) suggested that some virulence traits could be specific to bovine-mastitis causing *E. coli* strains, none of these traits has been universally agreed. In line with other studies, we could not see any pattern of virulence factors driving the force behind either bovine-mastitis causing, persistency or clinical severity. Seven virulence genes in the persistent strain of animal-1 (isolates 2424&2474), four virulence genes in the persistent strain of animal-8 (isolates 766&1608) and six virulence genes in the persistent strain of animal-10 (isolates 2421&2506) were observed whereas none was found in the persistent strain of animal-5 (isolate 739&937). There now follows a literature review about the virulence genes present in the sequenced *E. coli* genomes of this study.

AstA gene was previously found amongst 6.3% of bovine mastitis-causing *E. coli* strains (Suojala *et al.*, 2011), which was almost the same as our small scale study. It was only present in one of our persistent strains which was consistent with the study by Dogan *et al.*, (2012), but in contradistinction to findings of Leimbach *et al.*, (2015), where it was present in acute bovine mastitis strain (*E. coli* 1303) but absent in persistent mastitis-causing strain (ECC-1470). It was found to be the unique gene in bovine mastitis-causing *E. coli* isolates compared with environmental *E. coli* strains (Blum and Leitner, 2013). It was also found to be one of the most related virulence genes resulting in metritis and endometritis, the other serious dairy diseases (LeBlanc, Osawa and Dubuc, 2011).

Iss gene was the most common virulence factor found in various studies with bovine mastitis-causing *E. coli* strains, although none of the studies proved the presence of *iss* in their all isolates (Kaipainen *et al.*, 2002). However, *iss* gene was also found frequently in ExPEC, APEC and NMEC (meningitis-associated *E. coli*) strains (Johnson, Wannemuehler and Nolan, 2008; Rodriguez-Siek *et al.*, 2005). In our study, *iss* gene was found in 56.25% of the cases in both persistent and non-persistent; however, previously it was detected in only 16.7% of the 154 bovine mastitis-causing *E. coli* isolates and not in any persistent strain (Suojala *et al.*, 2011; Leimbach *et al.*, 2015).

The presence of *lpfA* gene was shown in transient and persistent mastitis-causing isolates and it was suggested to play a role in the adhesion of the pathogen (Dogan *et al.*, 2012). *LpfA* gene was previously found to be a statistically significant gene that was present in phylogenetic B and absent in group A (Blum and Leitner, 2013; Kempf *et al.*, 2016). These findings were

observed in our study as well, where the isolates with *lpfA* presence belonged to phylogenetic B1 and D; moreover absent in group A isolates. *LpfA* was shown in other pathogenic or non-pathogenic *E. coli* strains as well (Toma *et al.*, 2006; Chassaing *et al.*, 2011).

HilA-like regulator (EilA) and enteroaggregative immunoglobulin repeat protein (Air) play important roles in activating type III secretion system (T3SS) and adhesion, respectively (Sheikh *et al.*, 2006). The correspondence of these genes was shown in various studies and it was concluded that EilA activates the Air protein (Sheikh *et al.*, 2006). Both were present in only the same individuals of this study as well. Previously they were also isolated in EAEC strains and cefotaxime-resistant *E. coli* (CREC) strains from the faecal samples of suckling calves (Sheikh *et al.*, 2006; Manga *et al.*, 2019).

Type-V secretion serine protease (*espP*) was found in the persistent strain of animal-10 (isolates 2421&2506) and their closest isolate-2378. *EspP* was one of the ExPEC virulence factors (Köhler and Dobrindt, 2011; Dezfulian *et al.*, 2003), it was not detected in any bovine-mastitis causing *E. coli* strains in the study by Leimbach *et al*., (2017); but detected in 3 out of 37 bovine-mastitis causing *E. coli* strains in the study by Keane (2016).

Microcins are secreted mainly by *Enterobacteriaceae* for providing tolerance to heat, pH, protease and bactericidal activity (Rebuffat, 2012). *MchB*, *mchC* and *mchF* were found together in the persistent strain of animal-1 (isolates 2424&2474) and *mchF* alone was found in the persistent strain of animal-8 (isolates 766&1608). Three of them together were previously found in STEC and only ST-21 type of EHEC strains (Gonzalez-Escalona *et al.*, 2016; Ferdous *et al.*, 2016). *MchF* gene was found in 3 out of 37 bovine-mastitis causing *E. coli* strains (Keane, 2016).

Glutamate decarboxylase (*gad*) was shown to be related to survival in high acidic environments (Ma *et al.*, 2002; Yin *et al.*, 2012; Bergholz *et al.*, 2007) and involved in oxidative stress regulation (Bose, Venkatesh and Mande, 2017). *Gad* was another most common virulence factor which was found in 50% of the *E. coli* isolates in our study. It was previously shown present in EHEC, EPEC, enterotoxigenic (ETEC), enteroinvasive (EIEC), STEC and bovine mastitis-causing *E. coli* strains (Grant, Weagant and Feng, 2001; Richards *et al.*, 2015)

Salmochelin siderophore receptor (*IroN*) is one of the members of *E. coli* siderophore, which play a role in iron uptake so that bacteria can compete with immune cells and increase bacterial growth (Faber and Bäumler, 2014). Moreover, it was suggested to take part in an invasion mechanism (Feldmann *et al.*, 2007). *IroN* was found only in two persistent strains in this study

which were coming from animal-1 (isolates 2424&2474) and animal-8 (isolates 766&1608). Other researchers also found *iroN* present in their bovine mastitis-causing *E. coli* strains with the prevalence of 6% and 25% in transient and persistent mastitis cases, respectively (Dogan *et al.*, 2012; Kempf *et al.*, 2016). However, none of the bovine mastitis-causing *E. coli* strains did contain this gene in another study (Leimbach *et al.*, 2017). *IroN* gene was detected in ExPEC, UPEC and NMEC as well (Guzman-Hernandez *et al.*, 2016; Peigne *et al.*, 2009; Feldmann *et al.*, 2007).

Overall, no association was found between clinical severity of mastitis and virulence factor of mastitis-causing *E. coli*, which was consistent with previous studies (Suojala *et al.*, 2011; Lehtolainen *et al.*, 2003; Wenz *et al.*, 2006). It was suggested that persistent mastitis *E. coli* strains could have other bacterial factors other than virulence genes (Shpigel, Elazar and Rosenshine, 2008). However, there was no other gene which was particularly specific to any phenotypic group of *E. coli* isolates was found in our study. This may show that the weak influence of the pathogen genome to develop different outcomes of bovine mastitis. Cases like the same strains infecting the different cows but only some resulting in clinical mastitis shows the contribution of host factors. In the current study, isolate-2378 from animal-9 was highly similar to the persistent case in the animal-10 (isolates 2421&2506), although it could not persist or flare-up later in the lactation period. Cases like different strains infecting the quarters at dry period time, but only some resulting in persistent episodes through lactation strongly suggest the contribution of pathogen factors. However, it is still unknown to what extent these factors play a role in clinical severity (Rainard *et al.*, 2018). In this study, no specific pattern was shown between different phenotypes of mastitis pathogens based on their genomic profiles. It was, then, aimed to show the interaction between host and pathogen, by the constant host, based on proteome profiles of the pathogen in different clinical status.

Higher predictions were successfully performed by relatively simpler models in this current study. This is not surprising as it is a common experience that simpler models work better for smaller datasets and as the model gets complex, more data is needed. MLP was previously shown not to perform well with small datasets (Silva and Adeodato, 2011). DT and RF were previously shown to be successful small datasets but the dataset was still relatively bigger than ours (Shaikhina *et al.*, 2019). One of the main reasons for the low performance of RF classifier should be related to a limited number of features as RF needs a higher number of features. DT was also known to improve prediction performance together with an increase in training datasets (Morgan *et al.*, 2003). AdaBoost was previously shown to be more successful with larger

datasets by reducing bias and variance compared to small datasets (Zhang, Wang and Zhang, 2011). Prediction performances of QDA and LDA are also highly related to the size of data that is trained. This should explain the low success in our study with these models.

Although some classifiers like LR, LSVM and NB gave reasonable prediction performance for both analyses, we cannot talk about random testing which could disguise the learning curve and result in overfitting. It is no doubt that large datasets avoid overfitting and give better generalization; hence, this experiment should be repeated on a bigger scale. In ML studies, big datasets have been thought to be a key factor of high prediction performance (Müller and Guido, 2016). Moreover, the larger sample size was concluded to be improving prediction accuracy no matter what algorithm or feature type was employed (Cui and Gong, 2018). However, it is not always possible to collect large datasets due to high cost, time pressure, laborious methodology, scarce resource or simply biological ethics. Therefore, researchers are sometimes constrained to performing ML on limited datasets, to obtain preliminary results at least. Contrary to common belief, big data is not always the solution and there is room for improving the performance with small datasets. NCV was shown to give robust and unbiased performance compared to other validation methods such as K-fold cross-validation which still needs a big amount of data to deal with bias (Vabalas *et al.*, 2019). We also employed NCV for our study and had good performance with certain classifiers. The other issue to cope with employing ML on small datasets high dimensionality as neuroimaging, microarrays or genomic studies are subject to small sample size but produce lots of features (Arbabshirani *et al.*, 2017; Libbrecht and Noble, 2015). This was not the issue in our study as only the peaks with high intensity and statistically different ones were employed and there were two features for phenotype discrimination of persistent strains and six features for persistency detection. Hyperparameter optimisation was another key point to cope with the overfitting problem (Vabalas *et al.*, 2019), which were tuned for each classifier in this study. Another point to consider for assessing the performance of classifiers is the standard deviation of the prediction metrics. The performance of small datasets may likely have high variance if outliers are present. Confidence bounds in an experiment with around 30 sample size were expected to be 15% (Varoquaux, 2018). LR, LSVM, NB, RBF SVM, LDA and QDA had all their prediction metrics (sensitivity, specificity, AUC, accuracy, kappa) in this range while others failed for persistence profiling. For phenotype profiling, all the classifiers but NN and RF gave prediction performance with less than 15% error bar.

While our primary aim was to develop ML-powered diagnostic discriminating different phenotypes of the same genotype holding persistent *E. coli* strains, we also characterized the molecular determinants and mechanism underlying the clinical pattern to understand how bacteria transform their phenotype in the mammary gland. Our findings in differentiating subclinical and clinical phenotypes of *E. coli* showed that two MALDI-TOF peaks correspond to ribosomal protein (RpmI) and DNA gyrase inhibitor protein (YacG). YacG protein inhibits DNA gyrase activities such as supercoiling and relaxation, which could alter the physiology of the living organism. Hence, the function of YacG was thought to be as a control mechanism of the cell division and DNA replication in response to cell growth and stress signals (Sengupta and Nagaraja, 2008; Vos *et al.*, 2014). Protein expression of YacG was shown to be altered by an environmental stimulus in *E. coli* (Stenger, 2019) and other species (Yang *et al.*, 2019b). However, these studies of *yacG* were performed *in vitro* and it is not known why and when *yacG* is expressed *in vivo*. A common suggestion for its abundance may be related to stress, dormancy or persistency (Hobson and Berger, 2019). *RpmI* encodes the L35 protein which is the component of 50S subunit of ribosome and was found to be non-essential in *E. coli* (Shoji *et al.*, 2011). However, RpmI was one of the biomarkers that were detected to discriminate *Campylobacter coli* clades by using MALDI-TOF (Emele *et al.*, 2019b). This part of the study showed that phenotypic differences between the subclinical and clinical status of the same genotypes were detectable by proteomic MALDI-TOF spectral profiles. This difference seems reasonable as the selection pressure in the mammary gland would drive bacteria to adjust their phenotype as either protein abundance or modification.

Our findings in differentiating persistent and non-persistent phenotypes of *E. coli* showed that six MALDI-TOF peaks correspond to ribosomal protein (RpsS), protein YihD and four hypothetical proteins of which two of them could be matched with YncJ family protein and excisionase family protein by PSI-BLAST analysis. Protein expression of RpsS was shown to be altered by an environmental stimulus which shows RpsS involves in the stress response of *E. coli* (Stenger, 2019; Božik *et al.*, 2018). *RpsS* was shown essential for the survival of *E. coli* on LB media at 30˚C and 37˚C (Shoji *et al.*, 2011). It was one of the biomarkers which were found to differentiate *C. coli* and *C. fetus* clades by using MALDI-TOF (Emele *et al.*, 2019a; Emele *et al.*, 2019b). It was also found to be upregulated in mastitis *E. coli* strains cultivated with MAC-T cells (Zude, 2014). Protein expression of YihD was shown to be altered by environmental stimulus (Stenger, 2019) and in depletion studies, it was shown to be involved in ribosome biogenesis (Vlasblom *et al.*, 2014). Hypothetical protein-3 was found to belong to

YncJ family protein. The function of YncJ is still unknown but suggested to be related to stress response (Raivio, Leblanc and Price, 2013). *YncJ* gene was found up-regulated in fluoroquinolone-resistant *E. coli* isolates (Yamane *et al.*, 2012), and was also shown to be affected by environmental stress such as different concentrations of external copper (Yamamoto and Ishihama, 2005). Hypothetical protein-4 was found to have an excisionase family domain which was phage-encoded. This may be plausible for explaining the persistence character of some strains. The domains of other hypothetical proteins were not found; hence, no further information could be given about them.

Overall, we demonstrated that there was no genotypic pattern amongst bovine mastitis-causing *E. coli* strains to cause different phenotypic outcomes of persistency or clinical severity. It was shown that biological changes in the mammary environment force the pathogen to adapt its protein abundance by comparing MALDI profiles of different phenotypic groups. By using ML, we were able to show some of the biomarkers in a limited range, which may inspire further studies to design diagnostic tools or antimicrobial agents for bovine mastitis-causing *E. coli*.

CHAPTER 5 MASS SPECTROMETRY AND MACHINE LEARNING FOR THE ACCURATE DIAGNOSIS OF BENZYLPENICILLIN AND MULTIDRUG RESISTANCE OF *STAPHYLOCOCCUS AUREUS* IN BOVINE MASTITIS

This chapter is published in PLOS Computational Biology (https://doi.org/10.1371/journal.pcbi.1009108) with the title above by Necati Esener, Alexandre M. Guerra, Katharina Giebel, Daniel Lea, Martin J. Green, Andrew J. Bradley and Tania Dottorini. The authors' contributions were as follows: MJG and AJB provided the original data. TD, MJG and AJB conceived and designed the data analysis procedures. NE, AMG, KG, and DL carried on the data analysis. NE, AMG and TD wrote the manuscript. All authors reviewed the manuscript.

In **Chapter 5**, the main aim was to find a fast and more accurate alternative to standard susceptibility tests, to profile multidrug and benzylpenicillin resistance in *S. aureus* isolates. Data preparation of the MALDI-TOF spectra was performed by an in-house script written in MATLAB platform. Pre-processed data was analysed with ten supervised ML algorithms that were available in the sci-kit learn library in Python: LR, LSVM, RBF SVM, MLP NN, RF, DT, AdaBoost, NB, LDA and QDA. We tested the power of MALDI-TOF MS combined with ML techniques to present the antimicrobial profile of *S. aureus* associated with bovine mastitis. Here for the first time, we demonstrate that this approach can be used to develop diagnostic solutions that can discriminate with high performance between benzylpenicillin- and multidrug-resistant and susceptible bovine mastitis-causing *S. aureus* isolates.

5.1 INTRODUCTION

A new generation of 'superbugs' caused by ever-increasing AMR is a growing challenge in modern human and veterinary medicine (HM Government, 2019). The effects of neglecting AMR are extensive and have not only an economic impact, but also concern global health, environment, food sustainability and safety, and socioeconomic status. It is estimated that 700,000 deaths each year globally are caused only by AMR infections and this figure is expected to reach 10 million by 2050 (O'Neill, 2016).

S. aureus is a major opportunistic pathogen, infecting both humans and a wide variety of animals including dairy cattle, which have been recently proven to pose an important zoonotic potential, being the principal animal reservoir of novel human epidemic clones (Richardson *et al.*, 2018). Worldwide, *S. aureus* is one of the most frequently isolated pathogens of bovine mastitis, which remains a significant problem in the dairy industry by affecting productivity, profitability, animal health and welfare (Heikkilä *et al.*, 2018). The majority of bovine mastitis infections caused by *S. aureus* exhibit subclinical and chronic manifestations resulting in long-term intramammary persistence (Schukken *et al.*, 2011). *S. aureus* can reproduce swiftly upon entering the mammary gland and induce immune reactions that can lead to tissue injuries (Sutra and Poutrel, 1994). Most of the time, the immune response of the cow itself cannot successfully eliminate the *S. aureus* infection and treatment is needed (Sutra and Poutrel, 1994). Existing *S. aureus* vaccines are not considered as a preventive solution due to their yet unproven effectiveness against infections (Rainard *et al.*, 2018).

Antibiotics such as beta-lactams (e.g. benzylpenicillin), tetracyclines, cephalosporins (e.g. cefquinome), macrolides (e.g. erythromycin, tilmicosin, tylosin) and lincosamides are the most commonly used treatment for bovine mastitis (Gentilini *et al.*, 2000; Watts *et al.*, 1995). In the most recent UK-VARRS report (2020) which covers the data of 34% of all dairy farms in the UK, beta-lactams (32%) were found to be the most used antibiotic class for dairy cows. In a recent study with German dairy cows (Doehring and Sundrum, 2019), 44% of the clinical mastitis cases were found to be treated with penicillins. Similarly, 41% of the mastitis cases were found to be treated with penicillins in an extensive study with European dairy farms (De Briyne *et al.*, 2014). In the same study, at least 22.93% of the other dairy-related disorders were found to be treated with penicillins. The first examples of using benzylpenicillin for bovine mastitis treatment can be traced back to the 1940s (Aarestrup and Jensen, 1998). However, penicillin-resistant *S. aureus* strains, carrying a penicillinase/beta-lactamase emerged shortly after its first clinical usage and by the early 1950s, they became pandemic (Aarestrup and Jensen, 1998). More than 90% of current human-associated isolates (Peacock and Paterson, 2015) and varying from 84% to 92% of dairy-related isolates were observed to be penicillin-resistant (Feng *et al.*, 2016; Kalayu *et al.*, 2020). Although the UK surveillance reports between 2016 and 2019 showed that penicillin resistance in *S. aureus* was relatively low (22.25% on average) in British dairy cattle, the occurrence of penicillin resistance showed an increase through these years; 12.9%, 20.5%, 27.8% and 27.8% in 2016, 2017, 2018 and 2019, respectively. It should be also

noted that penicillin resistance of *S. aureus* associated with bovine mastitis was always the most common amongst tested antimicrobial classes (UK-VARSS, 2020; UK-VARSS, 2019).

It is not uncommon in the dairy cattle industry to give antibiotics to healthy animals to prevent the insurgence of diseases, and to sick animals often without certainty about the actual bacterial origin of the disease. Even when the bacterial origin is defined, broad-spectrum antibiotics are often used, instead of targeting the specific bacterial strain causing the illness. Underlying such prescription practices is the lack of fast, affordable and effective diagnostic solutions, which leaves the veterinarian to primarily rely on educated guesses. These practices have serious consequences, amongst which is the appearance and diffusion of multidrug antibiotic resistance profiles in the pathogen population. Preventing and controlling AMR is crucial as decreasing antimicrobial usage will lead to slower development and spread of the resistance. It is crucial that people and animals take the right medicine at the right time, in the right dosage and over the right period. Inappropriate usage of the antimicrobial drug causes resistance in infectious organisms.

S. aureus is capable of acquiring new resistance traits by the integration into its genome of exogenous genetic material via horizontal gene transfer and mutational events (Jensen and Lyon, 2009; Pantosti, Sanchini and Monaco, 2007). In *Staphylococcus* spp, the major targets underlying mechanisms of resistance are the cell envelope, the ribosome and nucleic acids (Foster, 2017). However, several studies have identified hypothetical proteins as also being associated with drug resistance specifically in *S. aureus* (Holden *et al.*, 2004). Characterising the proteins, alone or in combination, that contribute to the resistance, can potentially lead to improved diagnostic tools and therapeutics against antibiotic-resistant *S. aureus* and may hold the key to unlocking this global health problem.

In the dairy industry, several antimicrobial susceptibility profiling techniques have been used, which includes dilution (in broth, agar or milk) and agar diffusion (aka Kirby-Bauer) (Khan, Siddiqui and Park, 2019). These techniques have their specific advantages and disadvantages. Broth microdilution and agar dilution are the gold standard techniques; furthermore, agar diffusion is commonly preferred due to its low price and easy-use (Constable and Morin, 2003; de Jong *et al.*, 2018). Broth dilution is a quantitative technique which supplies MIC (minimum inhibitory concentration) values for the organism cultured on broth medium that includes antimicrobial. Broth dilution technique has two types, macrodilution and microdilution, although macrodilution technique is laborious and costly, new commercial kits (e.g. Sensititre® ARIS

2X and VITEK® 2) are available for microdilution analysis (Constable and Morin, 2003). Milk dilution is also a quantitative technique like broth dilution but supplies MIC values for the organism that are cultured on the milk instead. Again, there are commercially available test kits to perform milk dilution (e.g. MASTik®). Milk dilution technique is relatively economic, fast and more comparable to *in vivo* conditions compared to broth dilution or agar diffusion. However, the mastitis agent cannot be identified at the species-level unless additional culturing is performed (Constable and Morin, 2003). Moreover, milk dilution assay does not always give the same results of broth dilution and agar diffusion depending on the pathogen especially for *S. aureus* (Constable and Morin, 2003). Similarly, agar dilution is a quantitative technique which supplies MIC values for the organism cultured on agar that includes antimicrobial; however, this technique is relatively costly and difficult to perform (Constable and Morin, 2003). For agar diffusion (aka Kirby-Bauer) technique mastitis pathogen is incubated on agar with known antimicrobials over nightly, the inhibition zone is measured and transformed to MIC values for specific antimicrobial (Constable and Morin, 2003). However, such diagnostic tools are not affordable and quick enough to offer real-time control of invasive infections. The development of fast, affordable and effective diagnostic solutions capable of discriminating between antibiotic-resistant and susceptible *S. aureus* strains would be of huge benefit for effective disease detection and treatment.

MALDI-TOF has been an alternative way of detecting antibiotic resistance due to its low-cost and speed (Hrabák, Chudáčková and Walková, 2013). Antibiotic resistance profiles of several organisms could be determined by MALDI-TOF (Axelsson, Rehnstam-Holm and Nilson, 2019; Cordovana *et al.*, 2019; Nisa *et al.*, 2019), and, in combination with ML techniques, larger datasets could be analysed faster, more easily and economically (Tang *et al.*, 2019; Sharaha *et al.*, 2019).

The objective of this study was to find a fast and more accurate alternative to standard susceptibility tests, to profile multidrug and benzylpenicillin resistance in *S. aureus* isolates. To this end, we tested the discriminatory power given by the combination of supervised ML and MALDI-TOF, complemented by a PPI network and a protein structural analysis workflow. Here for the first time, we demonstrate that this approach can be used to develop diagnostic solutions that can discriminate with high performance between benzylpenicillin and multidrug-resistant and susceptible bovine mastitis-causing *S. aureus* isolates.

5.2 METHODS

5.2.1 Data Source

The data for the current study were obtained from the previous large-scale study as for Chapter 3 (Green *et al.*, 2007) of the UK dairy herds within the scope of the control plan of mastitis. The farm selection and sample collection were mentioned in the Methods section of Chapter 3; hence, will not be repeated here.

The samples were from 24 herds each in a different farm (24 farms) where 23 farms were in England (most of them in the south) and one farm was in Wales (Llangathen, Carmarthen). The locations of the farms and their respective number of cows are shown in Figure 5-1. 82 *S. aureus* isolates were collected from 67 animals that were diagnosed with bovine mastitis in 24 different farms, in England and Wales between March 2004 and May 2005. The animals with mastitis were either primiparous (n=9) or multiparous (n=73, median parity=4). On the day of sample collection, the days in milk of the cows varied from 1 to 569 days with a median of 160 days.

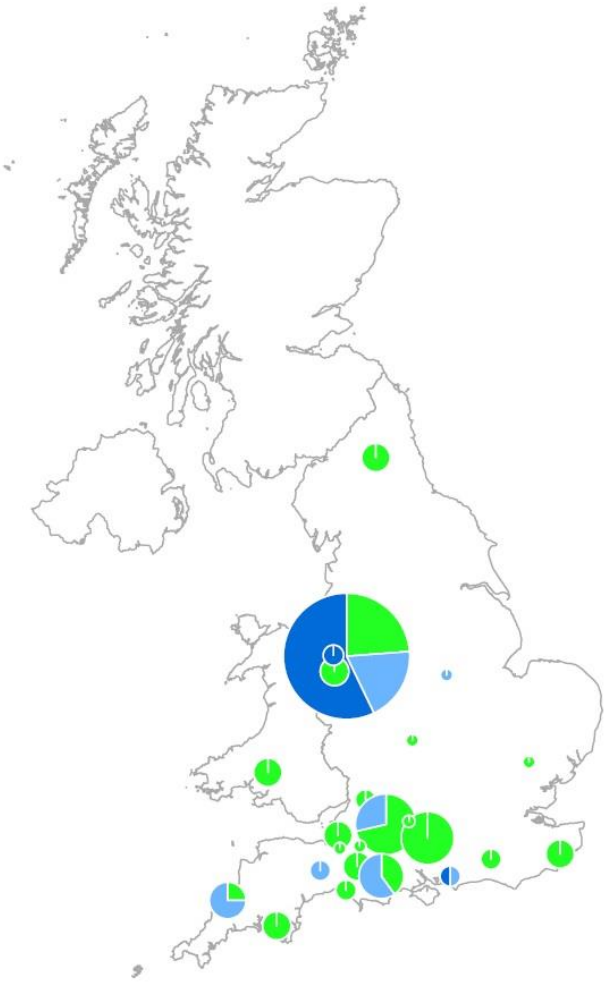


Figure 5-1. Location of the enrolled farms in the United Kingdom that provided Staphylococcus aureus isolates. *The circles represent the location of the farms and the diameters indicate the number of isolates provided by these farms. The highest number of isolates provided by a single farm was 21, while the lowest was 1. Green colour represents the susceptible S. aureus isolates, whereas dark and light blue represent multidrug-resistant and benzylpenicillin-resistant only S. aureus isolates, respectively. This figure was generated in R (R Core Team, 2019) using the sp (Pebesma and Bivand, 2005), mapdata (Deckmyn, 2018) and mapplots (Gerritsen, 2014) packages.*

5.2.2 Antimicrobial Susceptibility Testing

Bovine mastitis-causing *S. aureus* isolates were tested using a VITEK 2 AST-GP79 card per isolate by QMMS. Each card was filled with at least one positive control well with no antibiotic and multiple wells with increasing concentrations of the following antibiotics: benzylpenicillin, cefoxitin, oxacillin, cefalotin, ceftiofur, cefquinome, amikacin, gentamicin, kanamycin, neomycin, enrofloxacin, clindamycin, erythromycin, tilmicosin, tylosin, tetracycline, florfenicol and trimethoprim/sulfamethoxazole. Using the VITEK 2, the growth and viability of the isolates were measured in all wells compared to the control wells. Relative bacterial growth in each antibiotic well was calculated and compared with the positive control wells. The MIC

values were calculated by comparing the growth of the bacteria to the growth of isolates with known MICs. *S. aureus* isolates were labelled as either resistant or susceptible according to their antibiotic resistance profiles based on Clinical and Laboratory Standards Institute (CLSI) breakpoints (VET01-S3) (Watts *et al.*, 2008).

5.3 RESULTS

5.3.1 Antimicrobial Susceptibility Testing

VITEK analysis showed that the cohort consisted of 31 benzylpenicillin resistant and 51 benzylpenicillin susceptible isolates. Amongst the resistant isolates, 16 isolates were found to be only benzylpenicillin-resistant, while 15 isolates were found to be resistant to at least two more antibiotics, in addition to benzylpenicillin (multidrug-resistant). Only one isolate was resistant to two antibiotics in total. As shown in Figure 5-2 out of 15 multidrug-resistant isolates, 11 isolates were resistant to benzylpenicillin, clindamycin, erythromycin, tilmicosin and tylosin; 1 isolate was resistant to benzylpenicillin, clindamycin, tilmicosin and tylosin; 1 isolate was resistant to benzylpenicillin, tetracycline and tilmicosin; 1 isolate was resistant to benzylpenicillin and tetracycline, and 1 isolate was resistant to benzylpenicillin, cefalotin, cefoxitin and oxacillin. 51 isolates were found to be susceptible to all antibiotics used in this study which were benzylpenicillin, cefoxitin, oxacillin, cefalotin, ceftiofur, cefquinome, amikacin, gentamicin, kanamycin, neomycin, enrofloxacin, clindamycin, erythromycin, tilmicosin, tylosin, tetracycline, florfenicol and trimethoprim/sulfamethoxazole.

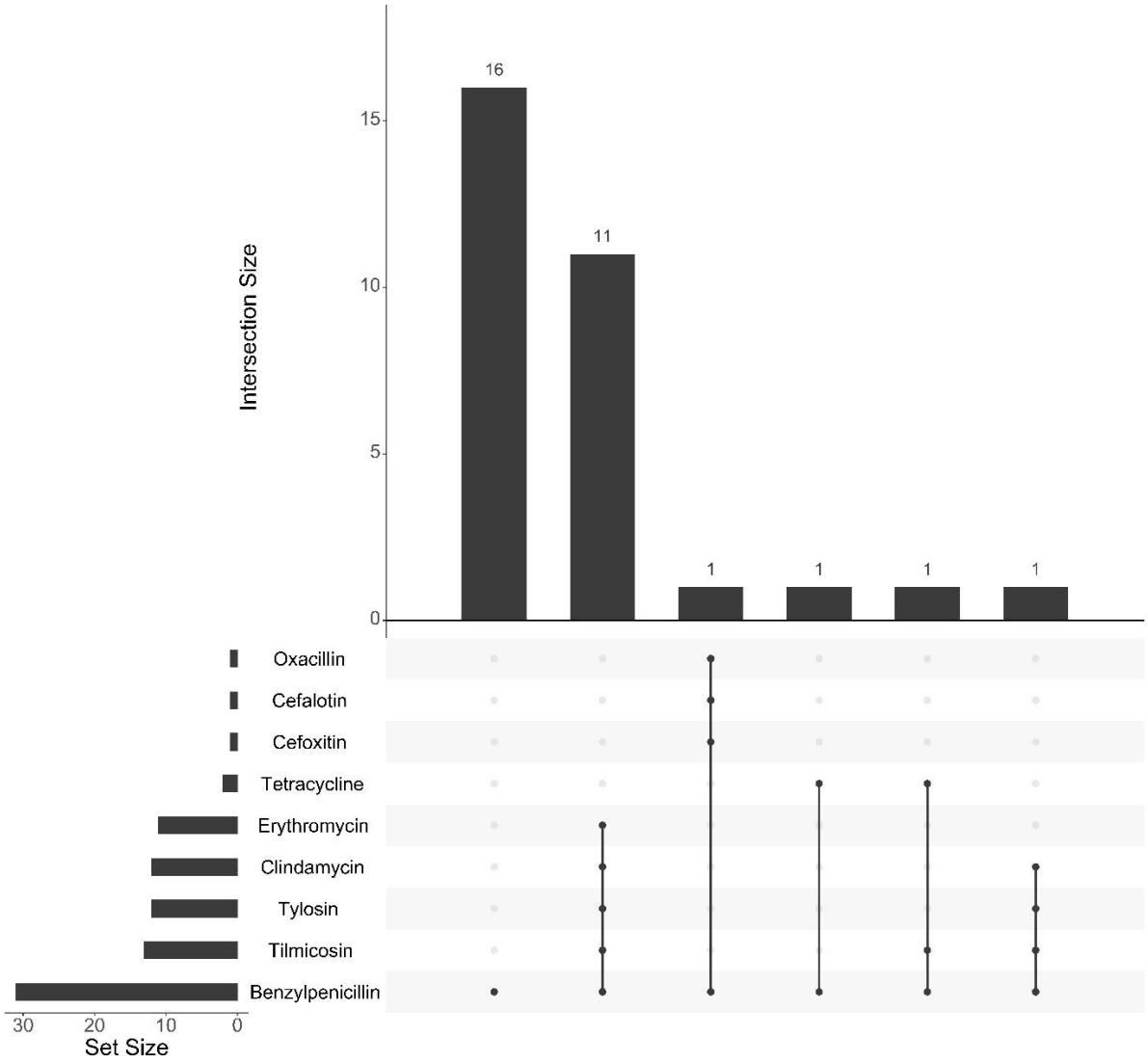


Figure 5-2. UpSet diagram summarizing the profile of antimicrobial-resistant S. aureus isolates. *The total size of resistant S. aureus isolates is shown on the left bar plot. The multi-resistant profile is visualised by the bottom plot and the occurrence is represented on the top bar plot. This figure was generated by using R package UpSetR (Conway, Lex and Gehlenborg, 2017).*

5.3.2 Generation of MALDI-TOF Peak Lists and Set-Up of the Classifiers

A total of 312 MALDI-TOF raw data spectra had been obtained from 82 *S. aureus* isolates, on average 4 replicate spectra per isolate. The peak lists, i.e. the lists of paired mass/charge (*m/z*) ratios and corresponding intensity values, were extracted from the raw spectra as specified in the Methods Section.

5.3.3 Analysis of Multidrug-Resistant vs Susceptible Isolates

We first focused on investigating the possibility to develop a classifier to verify if MALDI-TOF peak lists associated with isolates could be used to predict their multidrug phenotype.

Specifically, we considered the spectra of 15 multidrug-resistant isolates (all resistant to benzylpenicillin and at least one more antibiotic) and 51 susceptible isolates (susceptible to all antibiotics tested in this study). A total of 249 raw spectra were analysed. The pre-processing led to the identification of four different peaks found to appear in at least 30% of all number of spectra (see Table 5-1).

Table 5-1. Peak statistic report for the analysis of multidrug-resistant vs susceptible isolates.

Mass (kDa)	PTTA	PWKW	Ave1	Ave2	StdDev1	StdDev2	PA	PA1	PA2
4.807	3.78E-12	1.34E-07	7.27	19.55	5.89	3.72	66.88	35.71	98.04
6.422	0.00036	0.041891	6.92	10.30	4.54	2.00	45.31	35.71	54.90
6.891	0.02021	0.12752	31.98	43.04	23.96	14.89	80.18	64.29	96.07
9.621	6.81E-08	3.73E-07	32.39	43.00	3.28	6.23	100.00	100.00	100.00

PTTA is the *p*-value of Welch's *t*-test.
PKWK is the *p*-value of Wilcoxon test.
Ave1 is the intensity average of class 'Resistant'; **Ave2** is the intensity average of class 'Susceptible'.
StdDev1 is the intensity standard deviation of class 'Resistant'; **StdDev2** is the intensity standard deviation of class 'Susceptible'.
PA is the overall proportion of appearance; **PA1** is the proportion of appearance of class 'Resistant'; **PA2** is the proportion of appearance of class 'Susceptible'.

Due to the unbalanced nature of this specific data set (76% of samples were susceptible and only 24% were resistant), the undersampling method was employed to build robust classifiers (Lemaître, Nogueira and Aridas, 2017). At each one of the 30 runs, 15 samples were randomly chosen out of the initial 51 susceptible samples and a final balanced (50% resistant, 50% susceptible) dataset was generated. The four peaks were then used as features to build ten classifiers and to develop predictive models for the multidrug phenotype. Before the classification, features were standardised (mean centred and unit variance scaled) then resistant and susceptible isolates were labelled as positive and negative, respectively. Amongst the investigated ML approaches, LDA, LSVM and RBF SVM were found to be the top three best performance showing algorithms. Diagnostic systems trained on individual isolates coming from 24 different farms achieved up to (mean result values of test data): accuracy = 96.81±0.43%, sensitivity = 99.88±0.41%, specificity = 95.96±0.52%, and kappa = 91.83±1.37% in LDA algorithm. Detailed performance results of all classifiers on test data can be found in Figure 5-3.

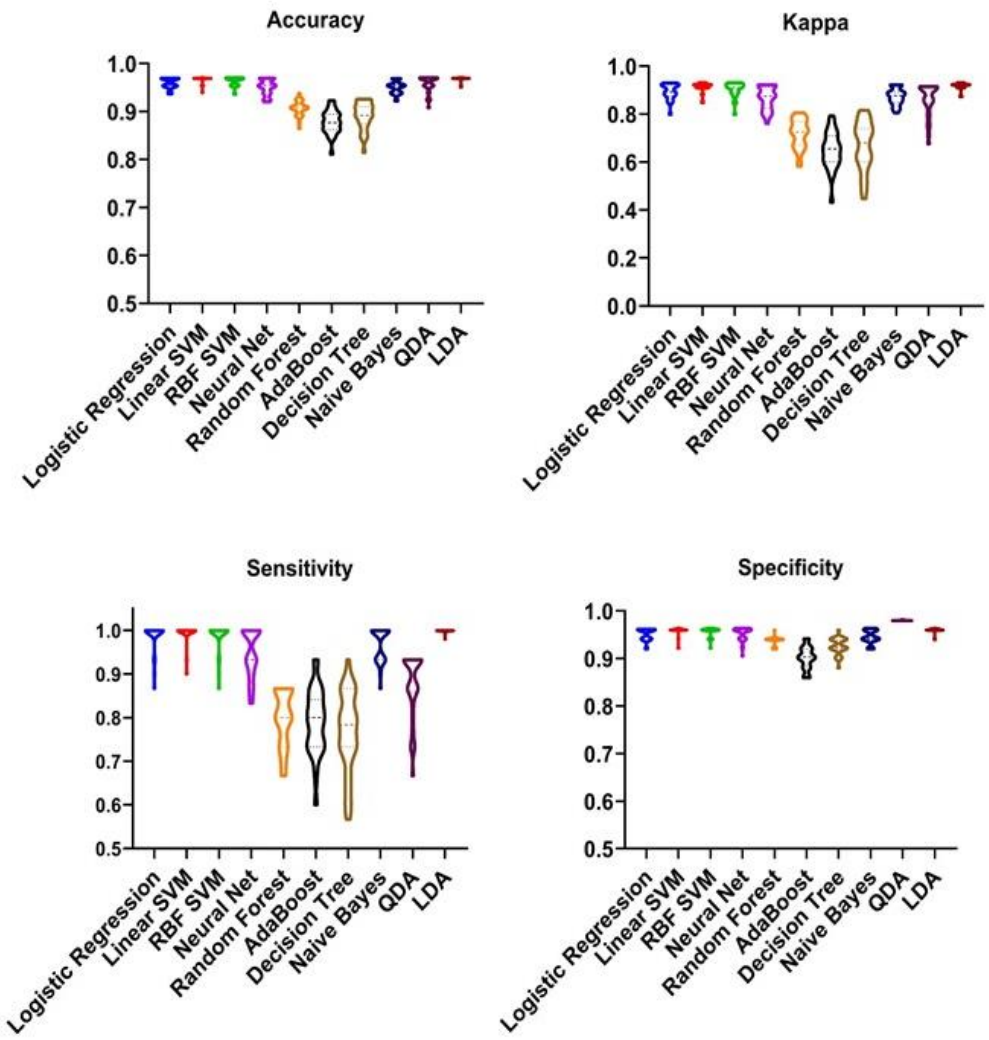


Figure 5-3. Prediction performance results of classifiers of multidrug-resistant vs susceptible S. aureus isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, random forest, AdaBoost, decision tree, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify the multidrug resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on four metrics (from left to right): accuracy, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. These graphs were generated in GraphPad Prism v8.*

5.3.4 Analysis of Benzylpenicillin-Resistant Only vs Susceptible Isolates

Next, antimicrobial susceptibility profiling for benzylpenicillin only was investigated. This was to isolate specific patterns underlying resistance to this specific antibiotic. Benzylpenicillin was chosen because it was the only antibiotic for which we had singly resistant isolates.

To this aim, the spectra of the 16 benzylpenicillin-resistant only and 51 susceptible isolates (susceptible to all antibiotics tested in this study) were first pre-processed as described in the

Methods Section. Five peaks were found in at least 30% of the overall number of spectra (Table 5-2).

Table 5-2. Peak statistic report for the analysis of benzylpenicillin-resistant only vs susceptible isolates.

Mass (kDa)	PTTA	PWKW	Ave1	Ave2	StdDev1	StdDev2	PA	PA1	PA2
4.305	0.258564	0.213998	10.20	9.34	2.60	2.64	34.33	37.50	33.33
4.807	7.02E-08	5.96E-07	12.94	19.55	4.02	3.72	92.54	75.00	98.04
6.422	0.39999	0.50342	10.81	10.30	2.44	2.00	58.21	68.75	54.90
6.891	5.69E-12	8.31E-08	10.00	43.04	8.80	14.89	76.12	56.16	96.07
9.621	1.81E-10	3.35E-08	29.84	43.00	5.54	6.23	100.00	100.00	100.00

PTTA is the *p*-value of Welch's *t*-test.
PKWK is the *p*-value of Wilcoxon test.
Ave1 is the intensity average of class 'Resistant'; **Ave2** is the intensity average of class 'Susceptible'.
StdDev1 is the intensity standard deviation of class 'Resistant'; **StdDev2** is the intensity standard deviation of class 'Susceptible'.
PA1 is the proportion of appearance of class 'Resistant'; **PA2** is the proportion of appearance of class 'Susceptible'.

Due to the unbalanced nature of this specific data set (76% of samples are susceptible and only 24% are resistant), the undersampling method was employed to build robust classifiers (Lemaître, Nogueira and Aridas, 2017). At each one of the 30 runs, 16 samples were randomly chosen out of the initial 51 susceptible samples and a final balanced (50% resistant, 50% susceptible) dataset was generated. The five peaks were then used as features to build ten classifiers and to develop predictive models for the benzylpenicillin phenotype. Before the classification, features were standardised (mean centred and unit variance scaled) then resistant and susceptible isolates were labelled as positive and negative, respectively. Amongst the investigated ML approaches RBF SVM, NN and LR were those that achieved the best performance. Diagnostic systems trained on individual isolates coming from 24 different farms achieved up to (mean result values of test data); accuracy = 97.54±1.91%, sensitivity = 99.93±0.25%, specificity = 95.04±3.83%, and kappa = 95.04±3.83% in RBF SVM algorithm. Detailed performance results of all classifiers on test data can be found in Figure 5-4.

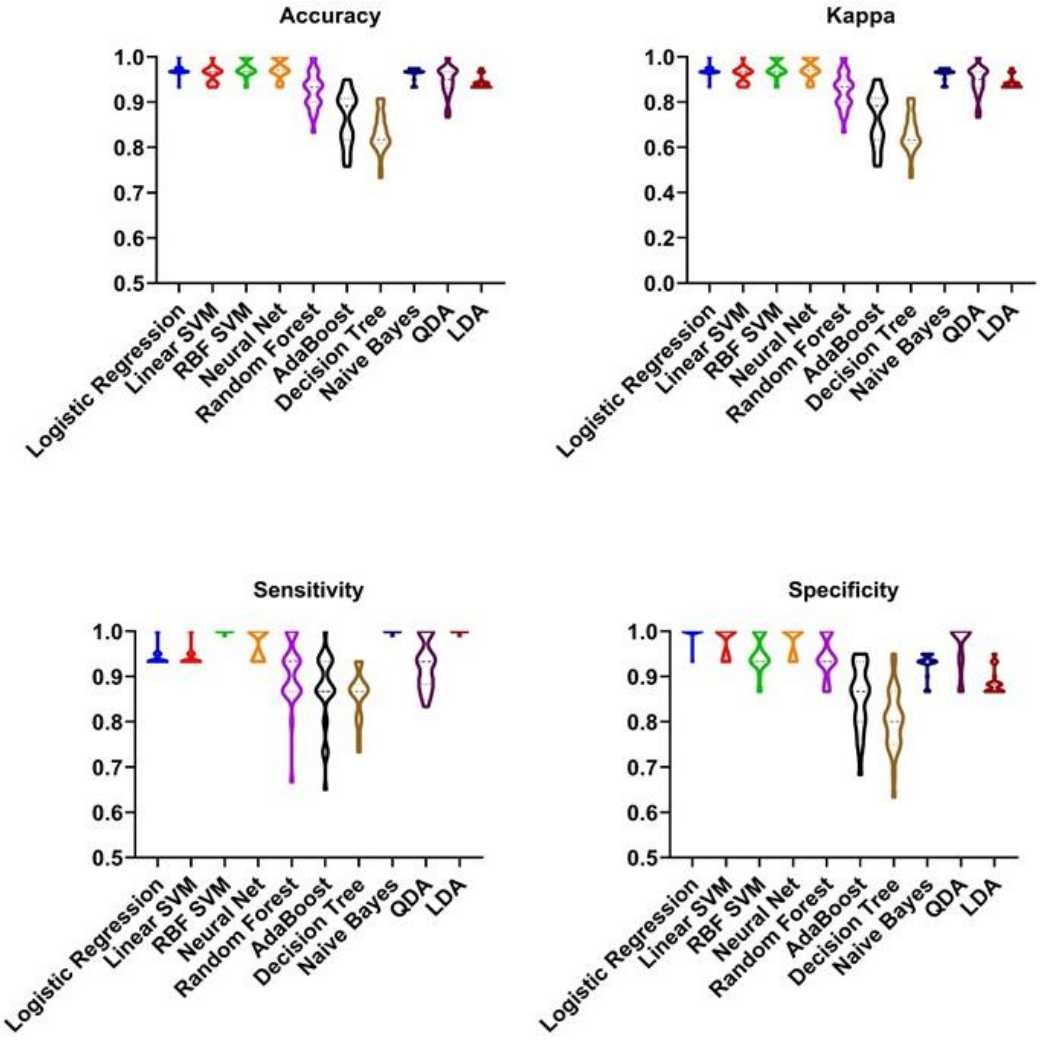


Figure 5-4. Prediction performance results of classifiers of benzylpenicillin-resistant vs susceptible S. aureus isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, random forest, AdaBoost, decision tree, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify the multidrug resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on four metrics (from left to right): accuracy, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. These graphs were generated in GraphPad Prism v8.*

Notably, four peaks (4.807kDa, 6.422kDa, 6.891kDa and 9.621kDa) were found common in the analyses of benzylpenicillin-resistant vs susceptible and multidrug-resistant vs susceptible isolates. When comparing the intensities of these four peaks in the two datasets (resistant vs. susceptible) we observed that 4.807kDa, 6.891kDa and 9.621kDa had a higher average in susceptible isolates consistently while 6.422kDa had a higher average of intensity in

benzylpenicillin-resistant only isolates class. 4.305kDa which was specific to benzylpenicillin-resistant only analysis had higher average intensity in resistant than susceptible isolates.

5.3.5 Biomarker Characterisation

The five peaks identified as providing optimal discrimination between benzylpenicillin-resistant only and susceptible isolates were further analysed to identify their correspondent *S. aureus* proteins. It should be noted that the four peaks identified as providing optimal discrimination between multidrug-resistant and susceptible were also amongst these peaks. When compared to the reference *S. aureus* Newbould 305 (ATCC 29740) proteome, the five peak masses identified the following five *S. aureus* proteins: two hypothetical proteins (molecular weights of 4801.95 and 6901.37Da), RpmJ, RpmD and DNA-binding protein HU. The molecular weights of the corresponding proteins changed slightly from those in the original spectra as a result of the search criteria outlined in the Methods. To further characterise the function of these proteins PSI-BLAST comparative analysis was computed; all discriminant proteins with 100% coverage and significant e-values are shown in Table 5-3.

Table 5-3. Top PSI-BLAST, conserved domain search and cellular location results for the five discriminant proteins.

MALDI-TOF Peak (MW)	Protein (MW)	PSI-BLAST Match	Identity	e- value	Domain (e-value)	PSORTB location (score)
4305.59Da	rpmJ (4305.36Da)	50S ribosomal protein L36	100.00%	4e-16	Ribosomal_L36 (1.2e-19)	Cytoplasmic (10.00)
4807.21Da	HP1 (4801.95Da)	Uncharacterized protein	100.00%	4e-14	No conserved domain identified	Cytoplasmic membrane (9.55)
6422.37Da	rpmD (6422.48Da)	50S ribosomal protein L30	100.00%	4e-33	Ribosomal_L30 (3.4e-21)	Cytoplasmic (9.67)
6891.17Da	HP2 (6901.37Da)	Membrane protein	100.00%	1e-07	No conserved domain identified	Cytoplasmic membrane (9.55)
9621.26Da	DNA-binding protein HBsu (9626.01Da)	HU family DNA-binding protein	100.00%	2e-56	Bacterial DNA-binding protein (6.2e-37)	Cytoplasmic (9.67)

HP: Hypothetical protein.

To better understand the functions and roles of these proteins within the drug resistance phenotype, we characterised the molecular functions (MF), cellular components (CC), and

biological processes (BP) they may carry out. Here, RpmJ and RpmD are the 50S ribosomal proteins L36 and L30, respectively. HU is a histone-like DNA-binding protein, which interacts with DNA to protect from denaturation (Mishra and Horswill, 2017). For the hypothetical proteins, we used 3D threading methods to predict the GO functions (Figure 5-5). The hypothetical protein of 4801.95Da was annotated as intracellular protein transport (BP), proteolysis (BP), homophilic cell adhesion via plasma membrane adhesion molecules (BP) and ion binding (MF). The hypothetical protein of 6901.37Da was annotated as being involved with the small molecule metabolic process (BP), antibiotic metabolic process (BP), lipid transport (BP) and ion binding (MF).

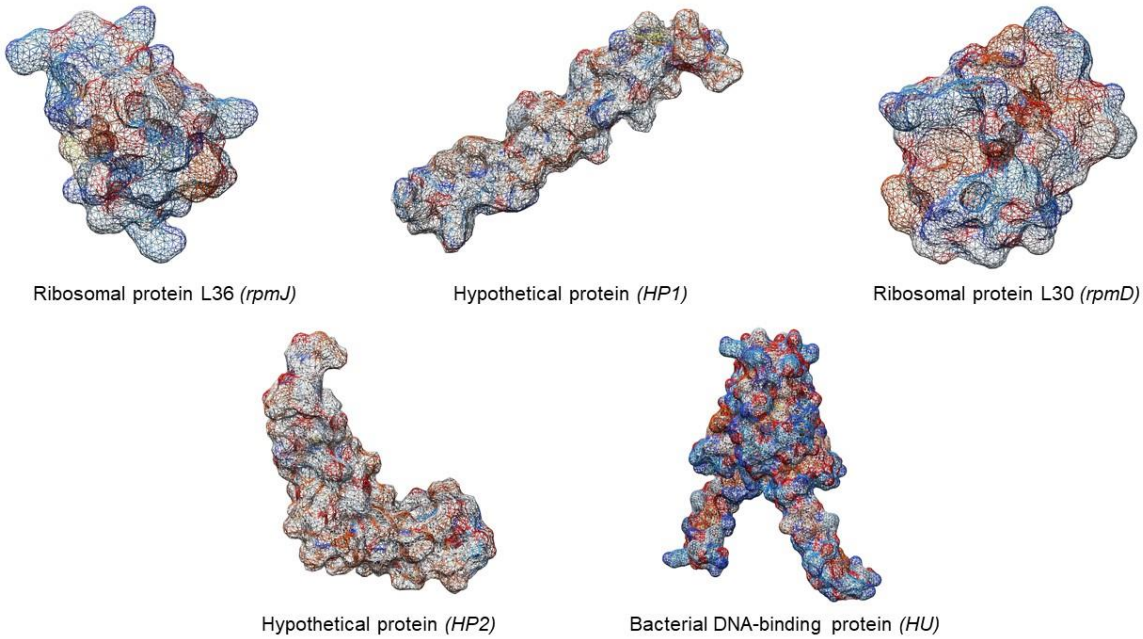


Figure 5-5. The 3D structures of the five proteins found to correspond to the significant MALDI-TOF peaks identified by the classifiers between benzylpenicillin-resistant and susceptible S. aureus isolates. *Top row from left to right: homology models of ribosomal protein L36p (RpmJ, Mw: 4305.36Da), threading model of hypothetical protein (HP1, Mw: 4801.95Da) and homology model of ribosomal protein L30p (RpmD, Mw: 6422.48Da). Bottom row from left to right: threading model of hypothetical protein (HP2, Mw: 6901.37Da) and homology model of bacterial DNA-binding protein (HU, Mw: 9626.01Da). The visualisation was carried out with UCSF Chimera.*

Next, the drug resistance interactome was investigated by building the PPI network. The benzylpenicillin PPI network, including the four significant proteins (RpmJ, RpmD, HU and HP2) and their 149 first neighbours, was generated (Figure 5-6). It should be noted that HP1 could not be found in the *S. aureus* proteome available in STRING database. Potential AMR genes

in the *S. aureus* proteome (see Methods for details) that were found to be interacting with the PPI network of interest are listed in Table 5-4.

Table 5-4. Potential antimicrobial-resistant proteins in Staphylococcus aureus proteome matched with resistant proteins in ResFinder v3.1 database. *Protein names, accession code, sequence similarity and shared domains are listed.*

Antimicrobial Class	ResFinder v3.1	*S. aureus* Proteome		Homology		
	AMR protein	Protein Name	Accession	Sequence Similarity	Shared Domain	Accession
Beta-lactam	*mecA_1_NC_002745_1*	*mecA*	AID38487.1	99.40%	*FtsI*	COG0768
Beta-lactam	*penA_1_AF515059_1*	*pbpA*	AID39620.1	27.60%	*FtsI*	COG0768
Beta-lactam	*blaGOB-2_1_AF189296*	*MBL*	AID40009.1	34.50%	*GloB*	COG0491
Beta-lactam	*blaPEDO-2_1_KP109678_1*	*blaZ*	AID41282.1	99.50%	*PenP*	COG2367
Tetracycline	*otr(C)_1_AY509111*	*ABC-2*	AID38698.1	27.20%	*CcmA*	COG1133
Tetracycline	*tet(M)_6_M21136_1*	*tetM*	AID38887.1	100.00%	*FusA*	COG0480
Tetracycline	*otr(A)_1_X53401*	*fusA*	AID39042.1	25.60%	*FusA*	COG0480
Tetracycline	*tet(42)_1_EU523697_1*	*norA*	AID39204.1	27.80%	*MFS*	CI28910
Macrolide	*erm(A)_1_X03216_1*	*ermA*	AID40123.1	100.00%	*RsmA*	COG0030

Tetracycline resistance protein (TetM) and elongation factor G (FusA) which are related to tetracycline resistance, were found to be the first neighbours of RpmJ and RpmD based on the experimental findings of the homologs in *E. coli* (Gagarinova *et al.*, 2016; Antipov *et al.*, 2017). Additional four proteins (MecA, BlaZ, PbpA and metallo-beta-lactamase (MBL)) related with beta-lactams, rRNA adenine N-6-methyltransferase (ErmA), related with macrolides resistance, and multidrug efflux pump (NorA) and ABC transporter protein (ABC-2) were found to interact with some first neighbours of the discriminant proteins in the network. Penicillin-binding protein 2 prime (MecA) was shown to share a common interactor, cell division protein (DivIB), with the discriminant protein RpmD. The interactions of MecA-DivIB (interaction score: 0.639) and DivIB-RpmD (interaction score: 0.864) are based on experimental/biological data coming from homologs in other species (Rowland *et al.*, 2010). MecA was also shown to share a common interactor, DNA polymerase I (PolA), with the discriminant protein HU. While the interaction of MecA-PolA was based on text mining (interaction score: 0.432), the PolA-HU was based on experimental/biological data (interaction score: 0.668) obtained from homologs in other species (Lopez-Causape *et al.*, 2017; Ramstein *et al.*, 2003). Text-mining represents the interaction of the proteins that are intermittently mentioned together in scientific publications (Szklarczyk *et al.*, 2018). PolA was the only protein which links (based on text

mining) HU to beta-lactamase (BlaZ) (interaction score: 0.425) (Wang *et al.*, 2014; Lopez-Causape *et al.*, 2017).

ErmA was shown to share common nodes (ribosomal proteins) with the discriminant proteins RpmD and RpmJ. ErmA was shown, based on text mining, to also interact with PolA, linked to HU as previously described, (interaction score: 0.611) (McCarthy *et al.*, 2011) and to other proteins (RpsA, MetG and GuaA), based on co-expression, gene fusion and co-occurrence (interaction scores >0.400). Gene co-expression relies on the assumption that proteins in the same network show a similar expression pattern (Jansen, Greenbaum and Gerstein, 2002; Ge *et al.*, 2001). Gene fusion technique relies on the assumption that if a fusion protein has two component homologous proteins which are not neighbours, they are highly likely to interact with each other (Enright *et al.*, 1999). Gene co-occurrence relies on the assumption that the proteins are co-occurred in the organism which is close to each other at the phylogenetic tree (Huynen *et al.*, 2000).

NorA was shown to share a common interactor, DNA topoisomerase (TopA) with the discriminant protein HU. ABC-2 was shown to share common interactors, signal recognition particle proteins FfH and FtsY with discriminant proteins RpmD and RpmJ. MBL was shown to share a common interactor, putative fatty oxidation complex protein (AID38649.1), with discriminant protein RpmJ based on co-expression, gene fusion and co-occurrence (interaction scores > 0.400).

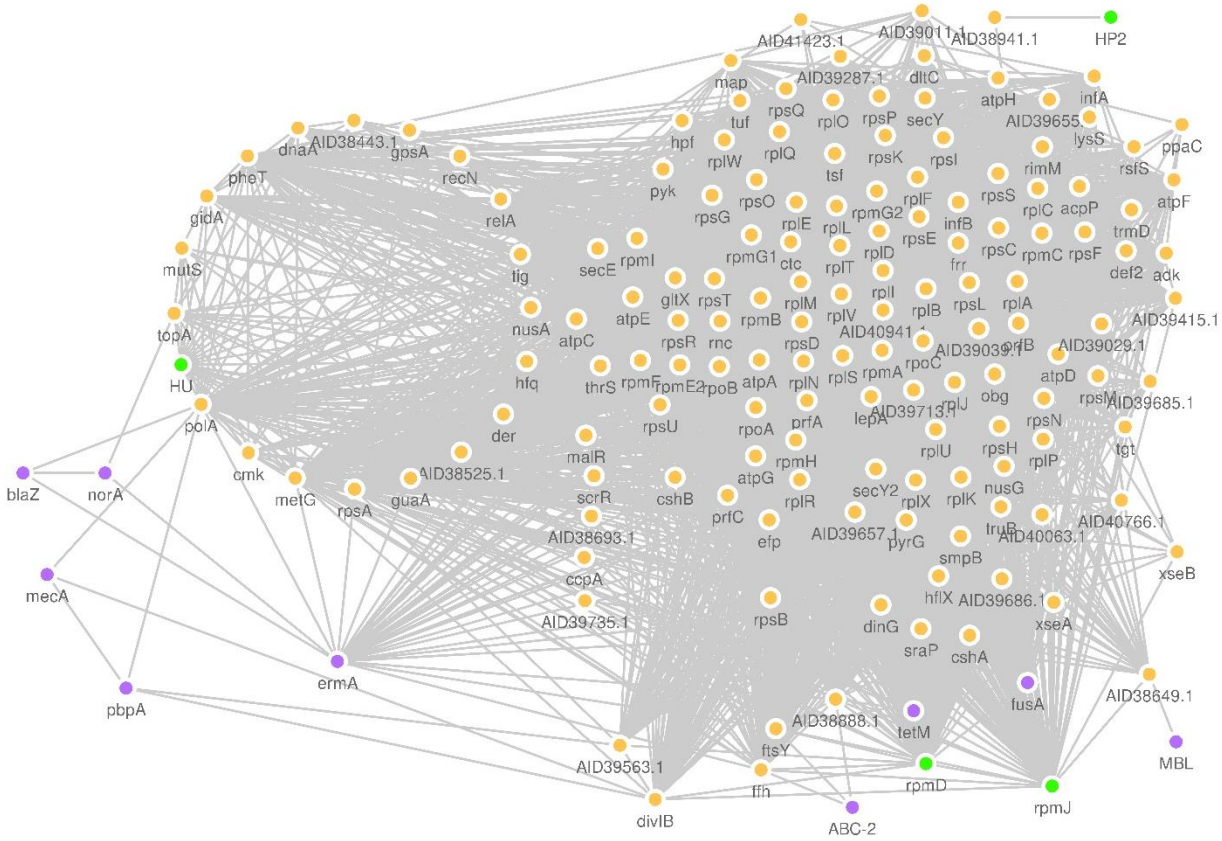


Figure 5-6. Protein-protein interaction (PPI) network of the Staphylococcus aureus proteins that are found to be discriminant between benzylpenicillin resistant and susceptible isolates. *The PPI network shows the four discriminant proteins (RpmJ, RpmD, HU, hypothetical protein 2 (HP2) and their first neighbour interactors in green and orange colours, respectively. Purple nodes (BlaZ, NorA, MecA, PbpA, ErmA, ABC-2, TetM, FusA and MBL) represent the antibiotic-resistant proteins predicted by ResFinder v3.1. Amongst antibiotic-resistant proteins TetM and FusA were found to be first neighbour interactors of discriminant proteins RpmJ and RpmD, while the others share common first neighbour interactors with the discriminant proteins as follows: BlaZ shares PolA with HU; MecA shares PolA with HU, and DivIB with RpmD and RpmJ; PbpA shares PolA with HU, and DivIB with RpmD and RpmJ; NorA shares TopA with HU; ABC-2 shares FfH and FtsY with RpmD and RpmJ; MBL shares AID38649.1 with RpmJ; and ErmA shares PolA, MetG, RpsA and GuaA with HU, and small ribosomal proteins such as RpsA with RpmD and RpmJ. The visualisation was carried out with Cytoscape.*

Notably, the PPI analysis of the benzylpenicillin-resistant proteome, 153 proteins – a total of 4 discriminant proteins and 149 first neighbour proteins – showed higher connectivity (clustering coefficient 0.728) than the randomly selected 4 proteins (this is repeated 10 times, only 4 proteins were selected randomly as this was the number of discriminant proteins in the network of interest) and their 106 first neighbour proteins (average of 10 times, same criteria with the benzylpenicillin-resistant proteome was applied for determining the first neighbour proteins) in *S. aureus* proteome network (clustering coefficient 0.068). The average number of neighbours per protein was 68.719 in the benzylpenicillin-resistant proteome network and 2.834 in the randomly built *S. aureus* proteome network. In terms of network density, the values were 0.452 (benzylpenicillin-resistant proteome network) and 0.041 (randomly built *S. aureus* proteome network) and for the network heterogeneity, the values were 0.528 benzylpenicillin-resistant proteome network) and 3.051 (randomly built *S. aureus* proteome network).

Functional enrichment analysis of benzylpenicillin network in *S. aureus* proteome was performed and the results based on GO terms and KEGG pathways can be seen in Figure 5-7.

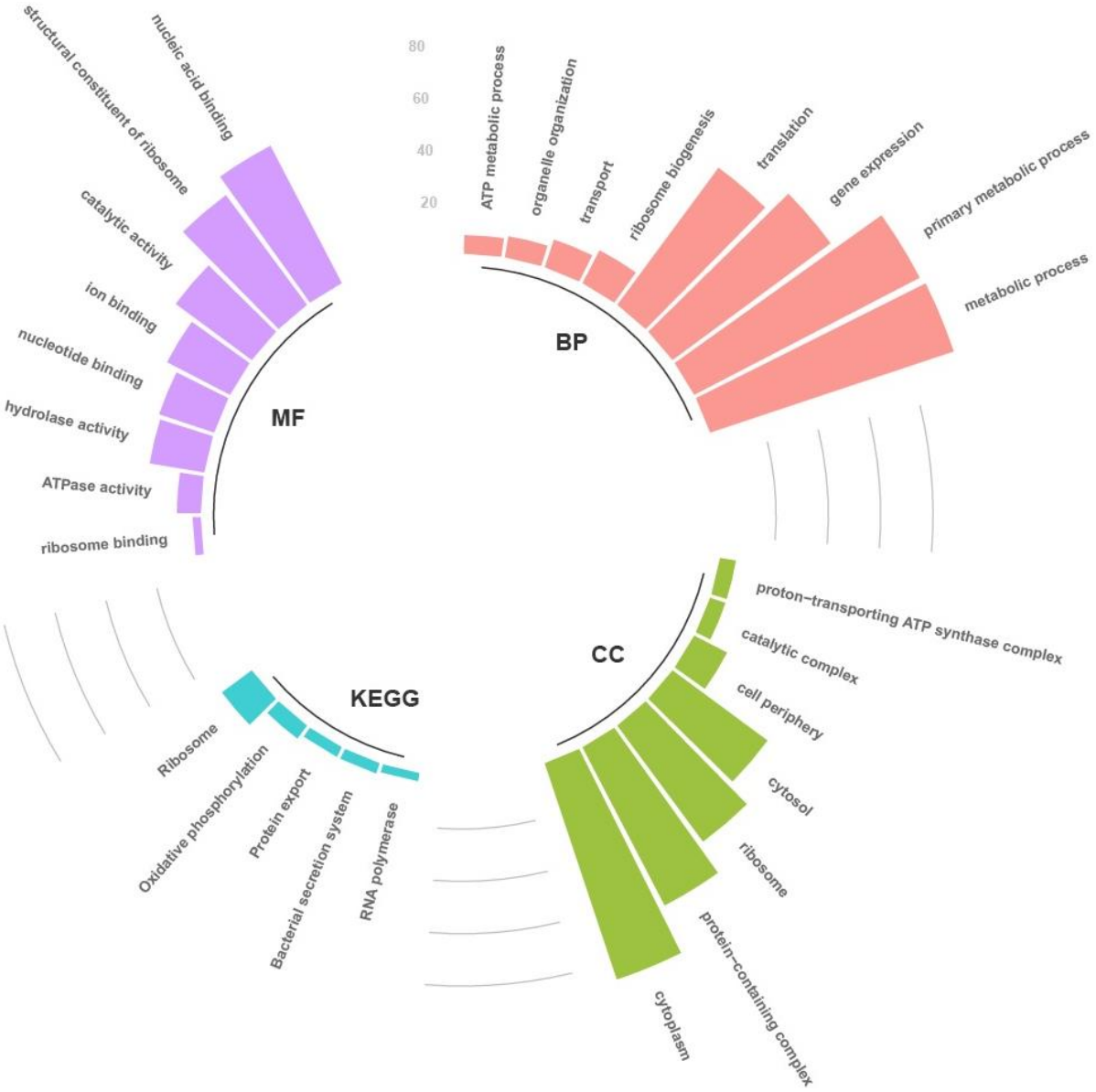


Figure 5-7. Functional enrichment analysis of the benzylpenicillin network in S. aureus based on Gene Ontology and KEGG pathways. *The network contains the 4 discriminant proteins and their 149 first neighbours. This figure was generated using R package tidyverse (Wickham et al., 2019).*

5.4 DISCUSSION

Antimicrobial-resistant *S. aureus* infections are a major concern in human and veterinary medicine. Recently, dairy cattle have been shown to be an important risk factor for zoonotic transfer (Richardson *et al.*, 2018). Fast, affordable and effective diagnostic solutions which can detect the specific *S. aureus* strains and their antimicrobial-resistant and susceptible profiles are key to support effective and targeted treatment selection.

Motivated by identifying the most effective method to discriminate (multidrug- and benzylpenicillin-) resistant and susceptible *S. aureus* strains, we approached the task in a principled way by applying optimization techniques to overcome uncertainty in data features and by using a wide repertoire of classification methods. Diagnostic systems trained on individual isolates coming from 24 different farms for analysing benzylpenicillin profiles achieved up to (mean result values of test data): accuracy: 97.54±1.91%, sensitivity: 99.93±0.25%, specificity: 95.04±3.83%, and kappa: 95.04±3.83% in RBF SVM algorithm. Again, diagnostic systems trained on individual isolates coming from 24 different farms for analysing multidrug profiles achieved up to (mean result values of test data): accuracy: 96.81±0.43%, sensitivity: 99.88±0.41%, specificity: 95.96±0.52%, and kappa: 91.83±1.37% in LDA algorithm. Although RBF SVM and LDA were found to be the best prediction performing models for profiling benzylpenicillin only and multidrug resistance, respectively; LR, LSVM, MLP NN, NB and QDA also resulted in kappa values over 85.00%. RF for multidrug profiling (kappa 85.11% for benzylpenicillin only) and AdaBoost and DT for both benzylpenicillin only and multidrug profiling gave relatively poor performance (kappa values lower than 85.00%). As being ensemble models RF, a classifier working in the principle of bagging, and AdaBoost, a classifier working in the principle of boosting, they need to use a subsample of an already small dataset; hence, might have not had enough data points to be trained well enough (Genuer, Poggi and Tuleau, 2008; Li, Wang and Sung, 2004). DT was also shown to give a limited performance with small datasets (Morgan *et al.*, 2003). It is suggested that stable classifiers should have less than 10% standard deviation for small datasets (e.g. less than 100 data points) between observations (Varoquaux, 2018). Therefore, it was also checked for each model employed in our studies. All classifiers except DT in the multidrug profiling analysis met these criteria. The other interesting finding was that all the learners except LDA had a much higher performance for discriminating benzylpenicillin-resistant only vs susceptible isolates rather than multidrug-resistant vs susceptible isolates. This may be explained by the fact that comparison based on single antimicrobial profile has less complexity.

While our primary aim was to develop ML-powered diagnostics discriminating resistant and susceptible isolates of bovine mastitis-causing *S. aureus*, we also characterized the molecular determinants and interactions underlying the identified antibiotic resistance and susceptibility patterns.

Our findings showed that the five MALDI-TOF peaks recognized as significant by the trained classifiers were found to correspond to two ribosomal proteins (RpmJ and RpmD), DNA-

binding HU protein and two hypothetical proteins. DNA-binding HU protein, RpmD and two hypothetical proteins were also found to give the best discrimination between multidrug-resistant and susceptible profiles of *S. aureus*.

The notion that components of the ribosome are important in the growth rate and antibiotic resistance of bacteria is a well-known concept (Gomez *et al.*, 2017). Among those determinants involved in intrinsic resistance, ribosomal proteins have been found to deal with the general response to stress (Olivares Pacheco *et al.*, 2013). Similarly, recent findings highlighted the existence of ribosomal mutations conferring resistance to antibiotics of several classes not targeting the ribosome (Gomez *et al.*, 2017). Specifically, it has been shown that ribosomal mutations can contribute to the evolution of multidrug-resistant profiles, by inducing ribosomal misassembly, which in turn leads to a systematic transcriptional cell alteration, ultimately impacting resistance to multiple antibiotics by interfering with different cellular pathways (Gomez *et al.*, 2017). RpmJ was shown to be up-regulated in *P. aeruginosa* when treated with ciprofloxacin and fluoroquinolone (Babin *et al.*, 2017) and similarly in *S. epidermidis* (Zhu *et al.*, 2010). Moreover, *rpmJ* was shown to confer intrinsic multidrug resistance to a varied set of antibiotics (nitrofurantoin, sulfamethoxazole, rifampicin, tetracycline, vancomycin, ampicillin, colistin, erythromycin) in *E. coli*, where deletion of this gene caused the bacteria to become more sensitive than wild type (Liu *et al.*, 2010). In comparison, fewer literature works have been published about *rpmD* and antibiotic resistance. *RpmD* was shown downregulated in *S. aureus* strains which had the antibiotic tolerance related LytSR system silenced (Sharma-Kuinkel *et al.*, 2009).

The discriminant protein DNA-binding HU protein was found essential in the bacterial survival and growth of *S. aureus* (Chaudhuri *et al.*, 2009). It was also previously found to be correlated to antibiotic resistance by being upregulated in the mutant *S. aureus* isolates with silenced serine/threonine kinase PknB, which also has a penicillin-binding domain (Donat *et al.*, 2009). Besides the proteins with known functions, we also identified two hypothetical proteins, but we were unable to find any evidence linking them to antibiotic resistance in the previous literature. Although it was not possible for us to identify the function of these hypothetical proteins, by applying PSI-BLAST and PSORTb v3.0 together with 3D threading modelling searches, the hypothetical proteins were predicted to be involved in pathways such as antibiotic metabolic process, lipid/protein transport and ion binding. In future studies, molecular biology techniques such as knock-out strategy may be performed to test whether they are actually AMR genes (Hadjadj *et al.*, 2019).

Two genes play a great role in resistant against penicillin in *S. aureus*; *blaZ* and *mecA* (Jensen and Lyon, 2009). The resistance mechanism of *blaZ* gene is the inactivation of the penicillin by improving the hydrolysis activity of the beta-lactam rings (Olsen, Christensen and Aarestrup, 2006). In a recent study, *blaZ* was detected in all of the MRSA strains isolated from bovine mastitis cases in China (Yang *et al.*, 2020). Furthermore, *blaZ* was found in high frequencies of *S. aureus* strains associated with bovine mastitis around the world (Jamali, Radmehr and Ismail, 2014; Aslantaş and Demir, 2016; Pérez *et al.*, 2020b). In this study, BlaZ was found to have common interactors with discriminant proteins in *S. aureus* PPI networks. Another beta-lactam-resistant gene was *mecA*, which encodes an alternative penicillin-binding protein with an altered antibiotic action target (Sawant, Gillespie and Oliver, 2009). In a recent study, *mecA* was detected in all of the MRSA strains isolated from bovine mastitis cases in China (Yang *et al.*, 2020). Furthermore, *mecA* was also found in *S. aureus* strains associated with bovine mastitis around the world (Haubert *et al.*, 2017; Aslantaş and Demir, 2016; Jamali, Radmehr and Ismail, 2014). In this study, MecA was found to have common interactors with discriminant proteins in *S. aureus* PPI networks.

By proving protection of ribosomal structure (Gao *et al.*, 2011), *tetM*, is one of the most prevalent tetracycline resistance genes in *S. aureus* isolated from both humans and animals (De Vries *et al.*, 2009). It was detected in all of the MRSA strains isolated from bovine mastitis cases in China (Yang *et al.*, 2020). Furthermore, *tetM* was found in several *S. aureus* strains associated with bovine mastitis around the world (Feng *et al.*, 2016; Haubert *et al.*, 2017; Jamali, Radmehr and Ismail, 2014; Aslantaş and Demir, 2016). In this study, TetM was found to be interacting directly with discriminant ribosomal proteins (both RpmJ and RpmD) in *S. aureus* PPI networks.

Erythromycin belongs to the macrolide class of antibiotics which inhibits bacterial protein synthesis by blocking peptidyl transferase (Vázquez-Laslop and Mankin, 2018). In *S. aureus*; *ermA*, *ermB* and *ermC* have been reported as macrolide-resistant genes which regulate ribosomal alterations (Jensen and Lyon, 2009; Gatermann, Koschinski and Friedrich, 2007). These genes have also been found in bovine mastitis associated *S. aureus* isolates, where *ermC* is more frequent (Pérez *et al.*, 2020b; Jamali, Radmehr and Ismail, 2014). Conversely, another study found *ermA* gene in all *S. aureus* isolates whereas *ermC* was present in none of them (Shamila-Syuhada *et al.*, 2016). Similar results were observed by a Turkish study where *ermA* was present in 70% of the bovine mastitis-causing *S. aureus* isolates (Aslantaş and Demir,

2016). In this study, ErmA was found to have common interactors with discriminant proteins in *S. aureus* PPI networks.

We were not surprised that known genes such as *blaZ* and *mecA* conferring resistance to penicillin in *S. aureus* were not amongst the MALDI-TOF peaks recognized as significant by the trained classifiers. This is because the mass range resolution of the MALDI-TOF was set to be between 2kDa and 12kDa, and the BlaZ and MecA are the proteins with molecular weights higher than 20kDa. However, our PPI cluster analysis results showed that these proteins known to confer resistance have all been found to interact with most of the proteins corresponding to the MALDI-TOF peaks and to form a highly connected benzylpenicillin proteome network.

While our approach successfully developed a diagnostic solution to identify AMR signatures, there are limitations to our method which future work may build upon. For one, the working range of 2-12kDa does not give the possibility to study the complete *S. aureus* proteome in relation to a specific phenotype. Besides, we acknowledge that our data were collected from farms only in England and Wales. However, this should not pose a restriction on our method's ability to predict resistance or susceptibility in other farms across the globe. If it is given a sufficiently diverse distribution of data to train the supervised learning algorithms, this would reduce any geographical bias that could affect predictive capability. Finally, we defined multidrug-resistant isolates as those being resistant to benzylpenicillin and at least one other antibiotic. Therefore, there is a bias towards peaks determining resistance or susceptibility to benzylpenicillin, which may explain why all 4 multidrug discriminant peaks occurred within the set of benzylpenicillin-only discriminant peaks.

Overall, we demonstrated that the combination of supervised ML and MALDI-TOF MS can be used to develop an effective computational diagnostic solution that can discriminate between benzylpenicillin/multidrug-resistant and susceptible *S. aureus* strains. Our solution could save time and money with respect to traditional susceptibility testing which is not viable for day-to-day monitoring of antibiotic resistance. Our solution could support farmers with timely, accurate and targeted treatment selection.

CHAPTER 6 DISCRIMINATION OF *ENTEROCOCCUS FAECALIS* AND *ENTEROCOCCUS FAECIUM* ISOLATES BASED ON ANTIMICROBIAL PROFILE

This study was conducted by Necati Esener, Alexandre M. Guerra, Katharina Giebel, Martin J. Green, Andrew J. Bradley and Tania Dottorini to be submitted in an open journal. The authors' contributions were as follows: MJG and AJB provided the original data. TD, MJG, AJB and NE conceived and designed the data analysis procedures. NE, AMG and KG carried on the data analysis. NE wrote the manuscript. TD and MJG contributed with comments and amendments.

In **Chapter 6**, the main aim was to test the power of MALDI-TOF MS coupled with ML, for profiling AMR in a more general perspective (several types of antimicrobials and different organisms), which was shown to work for *S. aureus* (Esener *et al.*, 2021). Data preparation of the MALDI-TOF spectra was performed by an in-house script written in MATLAB platform. Pre-processed data was analysed with ten supervised ML algorithms that were available in the sci-kit learn library in Python: LR, LSVM, RBF SVM, MLP NN, RF, DT, AdaBoost, NB, LDA and QDA. We tested the power of MALDI-TOF MS combined with ML techniques to present the antimicrobial profile of *E. faecalis* and *E. faecium*. The antimicrobial profile of *E. faecalis* was checked for benzylpenicillin, chloramphenicol, clindamycin, erythromycin, tetracycline and TMP/SMX, whereas the antimicrobial profile of *E. faecium* was checked for benzylpenicillin, cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin. Here, we showed MALDI-TOF MS coupled with ML has the potential of differentiating *E. faecalis* and *E. faecium* based on a single antimicrobial profile.

6.1 INTRODUCTION

Enterococcus is a large genus containing more than 50 species (Ludwig, Schleifer and Whitman, 2015); but *Enterococcus faecalis* and *Enterococcus faecium* are the most common species in dairy-related habitats (Gelsomino *et al.*, 2001). Both can be isolated from faecal samples of animals including mammals as they are involved in intestinal microbiota; however, once they are released into the environment, they can survive in various media such as soil, sand, water, forage, plants, vegetables and milk (Aarestrup, Butaye and Witte, 2002). Both species show viability in extreme pH, temperature, salinity and media; for instance, they can

still show an increase in numbers even after refrigeration or pasteurization of milk (Giraffa, 2003). They can also inhabit the skin of dairy animals, around the teats ending up in bulk milk tanks during milking (Petersson-Wolfe *et al.*, 2008). *Enterococcus* spp. have been suggested as environmental contaminants instead of mastitis agents (Petersson-Wolfe *et al.*, 2008; Rysanek, Zouharova and Babak, 2009); however, other studies have shown them as mastitis pathogens (Petersson-Wolfe, Wolf and Hogan, 2009; McBride *et al.*, 2007).

Dairy animals are housed during winter and recycled manure solids can be widely used as organic bedding material in the UK, EU and US (Bradley *et al.*, 2014). No matter how strict the conditions for use of recycled manure solids are set, there is always a risk of new infections, transfer of virulence and AMR originating from inefficiently prepared material (Bradley *et al.*, 2014).

Antibiotic treatment has been the most efficient way to fight bacterial diseases in animals. In a recent study with dairy cows Germany (Doehring and Sundrum, 2019), 44%, 37%, 6%, 1% and 1% of the clinical mastitis cases were found to be treated with penicillins, cephalosporins, (fluoro)quinolones, lincosamides and pyrimidines, respectively. In an extensive study with European dairy farms (De Briyne *et al.*, 2014), 41%, 33%, and 6% of the mastitis cases were found to be treated with penicillins, cephalosporins and macrolides, respectively. In the same study, at least 23% of the other dairy-related disorders were found to be treated with penicillins, 6% with cephalosporins, 11% with macrolides, 12% with tetracyclines, 9% with (fluoro)quinolones and 1% with lincosamides. In the latest UK surveillance report, antimicrobial usage by the dairy farms was found as follows: 32% penicillins/ 1st generation cephalosporins, 19% aminoglycosides, 22% tetracyclines, 10% macrolides, 12% trimethoprim/sulphonamides, 4% amphenicols, 0.4% 3rd/4th generation cephalosporins, 0.4% fluoroquinolones and 1% others (UK-VARSS, 2020).

Treatment of the diseases caused by *Enterococcus* is becoming more difficult, as the AMR profile of the species rises as a result of huge antimicrobial usage (Giraffa, Carminati and Neviani, 1997). In a study with German dairy farms, of 64 *E. faecalis* isolates, 86% were resistant to tetracycline, 17% to erythromycin and 9% to chloramphenicol (Werner *et al.*, 2012). In the same study, 19% and 14% of 37 *E. faecium* isolates were resistant to tetracycline and erythromycin, respectively. Meanwhile, none of the *E. faecium* isolates was resistant to chloramphenicol. In a study of dairy farms in Finland, of 63 *Enterococcus* isolates, both *E. faecalis* and *E. faecium*, 73%, 19% and 25.4% were resistant to tetracycline, erythromycin and multi-

antimicrobials respectively, while none of them was resistant to penicillin G (aka benzylpenicillin) (Pitkälä *et al.*, 2004). In Italy, *E. faecalis* isolates resistant to tetracycline, erythromycin, chloramphenicol and trimethoprim/sulfamethoxazole were 65.8%, 28.9%,18.4% and 2.6% of 38 *E. faecalis* strains, respectively (Cariolato, Andrighetto and Lombardi, 2008). Meanwhile, *E. faecium* isolates resistant to erythromycin, penicillin, tetracycline, chloramphenicol and trimethoprim/sulfamethoxazole were 40%, 27.8%, 20%, 2.3% and 2.3% of 43 *E. faecium* strains, respectively. In Portugal, resistance to enrofloxacin, erythromycin, chloramphenicol, tetracycline and benzylpenicillin were widely observed in dairy, clinical veterinary and human samples (de Fátima Silva Lopes *et al.*, 2005).

AMR mechanism can either be intrinsically present in the core genome of bacteria; or result from DNA mutations or acquisition of genetic mobile elements between intra and inter-species (Van Hoek *et al.*, 2011). As many antimicrobial compounds are naturally produced molecules, bacteria have evolved resistance mechanisms intrinsically; therefore, this intrinsic resistance is not a main concern in the current pandemic of AMR (Munita and Arias, 2016). However, horizontal gene transfer enables resistance to certain antimicrobials in those expected to be susceptible, and hence is a great concern (Munita and Arias, 2016). It has been increasing for decades due to selection pressure through increased antimicrobial use in human and veterinary medicine (von Wintersdorff *et al.*, 2016). Antimicrobial use for food-producing animals was shown to be greater than human use, 73% of the total antimicrobial consumption in the world; however, it is noteworthy to acknowledge that this varies a lot between regions (i.e. much lower in EU countries than low- and middle-income countries) (Van Boeckel *et al.*, 2019). Moreover, antimicrobial use for livestock animals is expected to rise by 11.5% by the year 2030 (Tiseo *et al.*, 2020). Increased AMR due to usage in large quantities is not only limited to the system of these animals, their carers and their consumers but spread wider communities via groundwaters (Chee-Sanford *et al.*, 2001; Smith *et al.*, 2013).

Enterococcus species show intrinsically low resistance to beta-lactams, lincosamides, aminoglycosides and trimethoprim/sulfamethoxazole (Garrido, Gálvez and Pulido, 2014; Hollenbeck and Rice, 2012; Arias and Murray, 2012). However, the high resistance profile amongst *Enterococcus* spp. is believed to be mainly as a result of horizontal gene transfer (Hershberger *et al.*, 2005) which has been shown *in vitro* (Eaton and Gasson, 2001) and *in vivo* studies (Lester *et al.*, 2006). Acquired resistance of *Enterococcus* by horizontal gene transfer included several antimicrobials like tetracycline, erythromycin, chloramphenicol and vancomycin (Cho *et al.*, 2020a; Cho *et al.*, 2020b; Conwell *et al.*, 2017). Tetracycline resistance of *Enterococcus* is

mainly associated with *tetM*, which is often due to transposable elements (Aarestrup *et al.*, 2000). Erythromycin resistance of *Enterococcus* is mostly associated with *ermB*, which is located in plasmid and mainly reported to be spread from *Enterococcus* to many other bacterial species (Munita and Arias, 2016; Roberts, 2008). Plasmid exchange between *Enterococcus* isolates was at high frequency even though there was no selection pressure (Cocconcelli, Cattivelli and Gazzola, 2003). Vancomycin resistance of *Enterococcus* is mainly associated with *vanA* and *vanB*, which are mainly located on mobile elements such as plasmids and transposons (Garrido, Gálvez and Pulido, 2014; Woodford, 2001). Vancomycin-resistant *S. aureus* is evolved due to plasmid from *E. faecalis* (Gardete and Tomasz, 2014)

Identification of the pathogens at strain level and their antimicrobial profile is important when it comes to choosing the right treatment (Zadoks *et al.*, 2003) as the resistance increase is correlated with the excessive usage of antibiotics on farms (Hershberger *et al.*, 2005). In veterinary medicine, antibiotic susceptibility testing is performed by phenotypic methods, such as disk diffusion, epsilometer test, automated system VITEK® 2, macrodilution and microdilution; or genotypic methods such as PCR, DNA microarray, DNA chips and whole-genome sequencing (Khan, Siddiqui and Park, 2019). However, such diagnostic tools are neither affordable nor quick enough to offer real-time control of invasive infections. MALDI-TOF has become an alternative way of detecting AMR due to its low-cost and speed (Hrabák, Chudáčková and Walková, 2013). Antimicrobial profiles of several organisms could be determined by MALDI-TOF (Axelsson, Rehnstam-Holm and Nilson, 2019; Cordovana *et al.*, 2019; Nisa *et al.*, 2019). MALDI-TOF spectra can contain more than a hundred peaks which cannot be identified by visual inspection. Therefore, ML which can learn from complex datasets and offer a solution for even non-linear classification issues is useful. MALDI-TOF coupled with ML has been popular recently so that larger datasets can be analysed fast and cheaply (Tang *et al.*, 2019; Sharaha *et al.*, 2019). Thus, vancomycin-resistant and susceptible *E. faecium* isolates could be differentiated by using ML algorithms on MALDI-TOF data (Griffin *et al.*, 2012; Wang *et al.*, 2020).

The objective of this study was to find an alternative to standard susceptibility tests, to profile benzylpenicillin, chloramphenicol, clindamycin, erythromycin, tetracycline and TMP/SMX resistance in *E. faecalis*; and benzylpenicillin, cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin resistance in *E. faecium* isolates. Doing so enables diagnosis of the antimicrobial profiles of *Enterococcus* species, monitor over time and alter strategies around disease prevention. To this end, we tested the discriminatory power of MALDI-TOF coupled with

several supervised ML algorithms: LR, NN, RF, AdaBoost, DT, NB, LDA, QDA, RBF and LSVM. The prediction performance was evaluated by sensitivity, specificity, accuracy, AUC and Cohen's kappa. Furthermore, we analysed the discriminant proteins between resistant and susceptible profiles of *Enterococcus* proteomes individually using bioinformatic tools. We showed that the peaks from MALDI-TOF spectra which have been employed by the several ML models as the most relevant for discrimination between susceptible and resistant to treatment, actually correspond to ribosomal subunits, DNA binding and bacterial two-component regulatory proteins, suggesting that these proteins may be correlated with different antimicrobial profiles. To the best of our knowledge, other than vancomycin, no studies have employed ML in the analysis of MALDI-TOF spectra for the classification of *Enterococcus* species based on their antimicrobial profile.

6.2 METHODS

6.2.1 Data Source

For this study, 111 *E. faecalis* and 88 *E. faecium* isolates were collected from milk and recycled manure solids bedding materials between 2015 and 2017. Susceptibility testing for all *Enterococcus* pathogens was conducted in the QMMS laboratory by using VITEK 2 (Ligozzi *et al.*, 2002). Briefly, antibiotic susceptibility testing cards were filled with positive control and increasing concentration of the following antibiotics: benzylpenicillin, chloramphenicol, clindamycin, erythromycin, tetracycline and trimethoprim/sulfamethoxazole for *E. faecalis* isolates; benzylpenicillin, cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin for *E. faecium* isolates. The growth of *Enterococcus* isolates in the control wells was observed and growth at an appropriate rate was confirmed. MIC was then measured by comparing the growth of *Enterococcus* pathogens with known MIC values. *E. faecalis* and *E. faecium* isolates were labelled as either resistant or susceptible according to MIC breakpoints adapted from CLSI (Watts *et al.*, 2008) and EUCAST (EUCAST, 2019).

6.3 RESULTS

6.3.1 Data Source

Antimicrobial-resistant/susceptible profiles for a total of 111 *E. faecalis* isolates are shown in Figure 6-1-A. While 38.74% (n=43) of the *E. faecalis* isolates were resistance to only one certain antibiotic, 39.64% (n=44) were resistant to more than one antibiotic. The distribution

of the multi-resistant profiles is shown in Figure 6-1-B. 21.62% (n=24) of the *E. faecalis* isolates were found to be non-resistant (either susceptible, intermediate or not known) to any antibiotics used in this study.

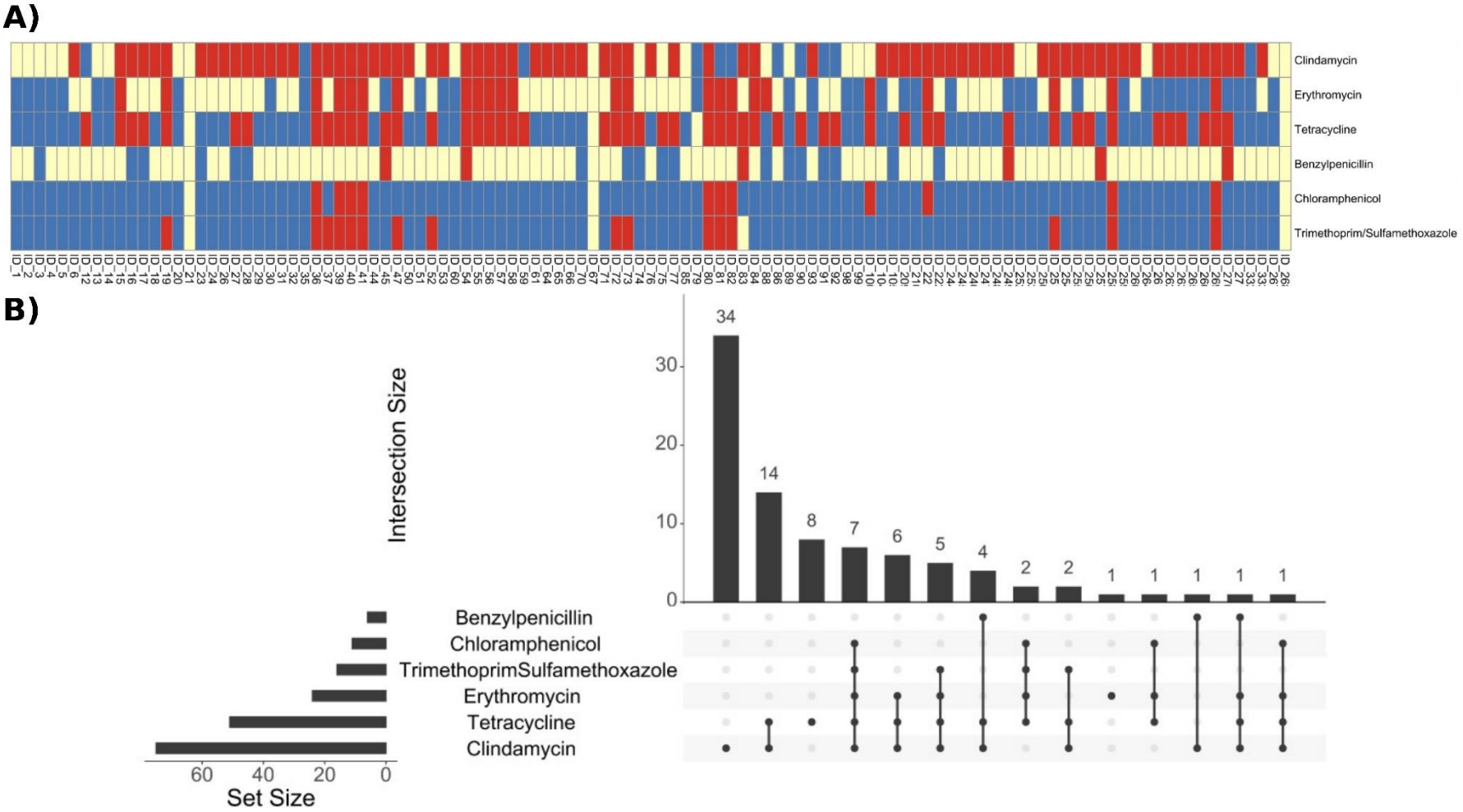


Figure 6-1. The antimicrobial-resistant/susceptible profiles of E. faecalis isolates. A) *Red-coloured wells indicate the resistance whereas blue coloured wells indicate the susceptibility for a certain antibiotic. Yellow coloured wells indicate either intermediate or terminated analysis before the results obtained.* ***B)*** *UpSet diagram summarizing the profile of resistant E. faecalis isolates. The total size of resistant E. faecalis isolates is shown on the left bar plot. The multi-resistant profile is visualised by the bottom plot and the occurrence is represented on the top bar plot.*

E. faecalis isolates, where technical count refers to the replicates of biological samples during MALDI-TOF analysis, were profiled as follows:

- Benzylpenicillin: 18 biological (70 technical) and 6 biological (26 technical) isolates as being susceptible and resistant classes, respectively.

- Chloramphenicol: 97 biological (392 technical) and 11 biological (44 technical) isolates as being susceptible and resistant classes, respectively.

- Erythromycin: 38 biological (150 technical) and 24 biological (99 technical) isolates as being susceptible and resistant classes, respectively.

- Tetracycline: 56 biological (220 technical) and 51 biological (210 technical) isolates as being susceptible and resistant classes, respectively.

- Clindamycin: 11 biological (44 technical) and 75 biological (301 technical) isolates as being susceptible and resistant classes, respectively.

- TMP/SMX: 91 biological (365 technical) and 16 biological (64 technical) isolates as being susceptible and resistant classes, respectively. The detailed information about these *E. faecalis* isolates can be seen in Table 6-1.

Table 6-1. MIC values of Enterococcus faecalis isolates against benzylpenicillin, chloramphenicol, erythromycin, tetracycline, clindamycin and TMP/SMX. *In this table, only susceptible and resistant isolates which are used in further machine learning analysis are shown, whereas intermediate isolates are not shown.*

	MIC Values	0.12		0.25		0.5		1		4		8		16		64		Susceptible Total		Resistant Total	
Antibiotic		Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding				
Benzylpenicillin	biological							17	1			5	0	1	0			18		6	
	technical							67	3			22	0	4	0			70		26	
	profile							Susceptible				Resistant		Resistant							
Chloramphenicol	biological									15	2	69	11			11	0	97		11	
	technical									54	7	287	44			44	0	392		44	
	profile									Susceptible		Susceptible				Resistant					
Erythromycin	biological			15	0	23	0					24	0					38		24	
	technical			58	0	92	0					99	0					150		99	
	profile			Susceptible		Susceptible						Resistant									
Tetracycline	biological							43	13					51	0			56		51	
	technical							168	52					210	0			220		210	
	profile							Susceptible						Resistant							
Clindamycin	biological	6	0	5	0					63	12							11		75	
	technical	23	0	21	0					253	48							44		301	
	profile	Susceptible		Susceptible						Resistant											
TMP/SMX	biological													78	13	16	0	91		16	
	technical													314	51	64	0	365		64	
	profile													Susceptible*		Resistant**					

*****MIC values of ≤20 µg/ml.

**MIC values of ≥160 µg/ml.

Antimicrobial-resistant/susceptible profiles for a total of 88 *E. faecium* isolates are shown in Figure 6-2-A. While 11.36% (n=10) of the *E. faecium* isolates were resistant to only one certain antibiotic, 86.36% (n=76) were resistant to more than one antibiotic. The distribution of the multi-resistant profiles is shown in Figure 6-2-B. 2.27% (n=2) of the *E. faecium* isolates were found to be non-resistant (either susceptible, intermediate or not known) to any antibiotics used in this study.

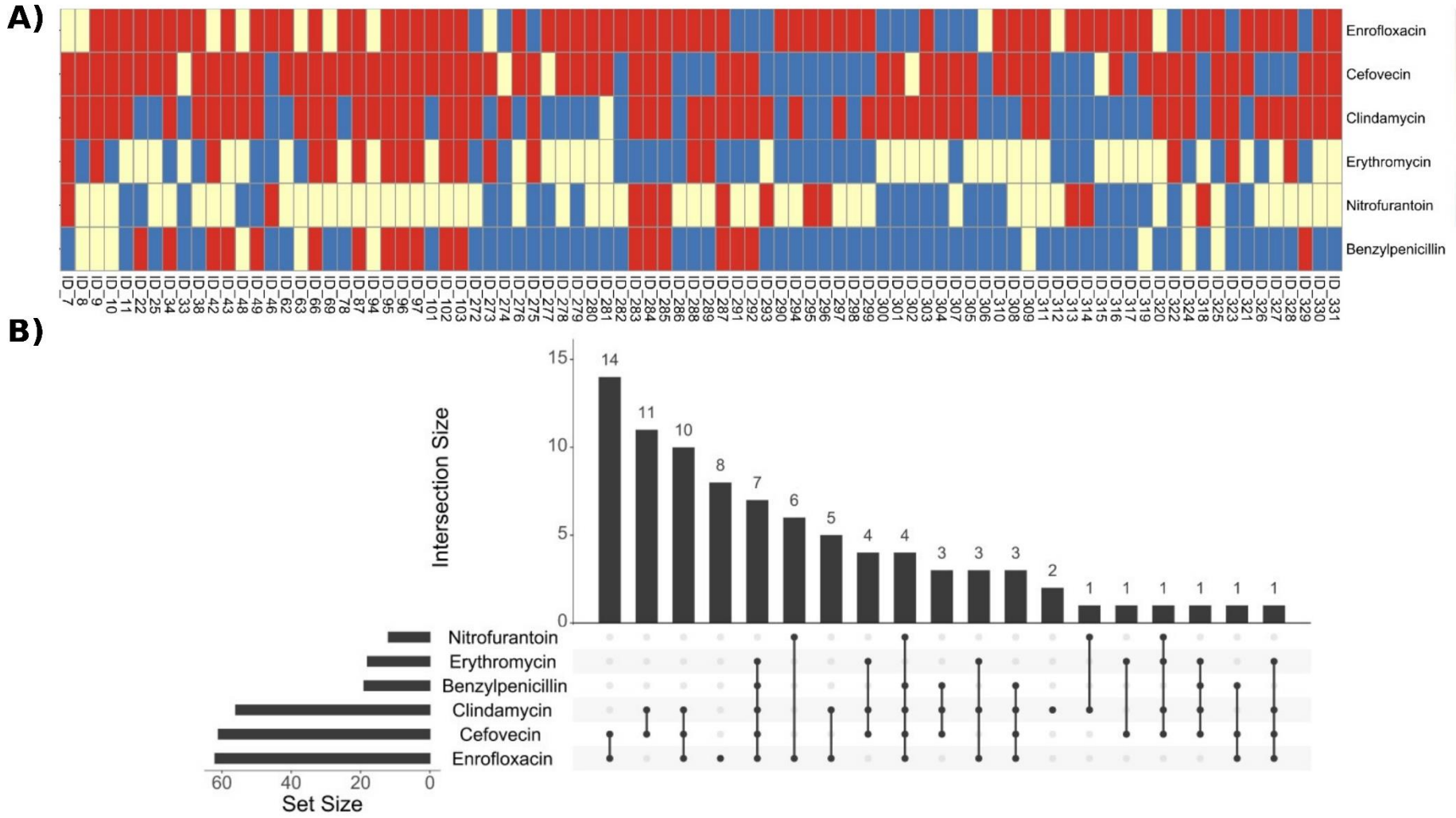


Figure 6-2. The antimicrobial-resistant/susceptible profiles of E. faecium isolates. A) *Red-coloured wells indicate the resistance whereas blue coloured wells indicate the susceptibility for a certain antibiotic. Yellow coloured wells indicate either intermediate or terminated analysis before the results obtained.* ***B)*** *UpSet diagram summarizing the profile of resistant E. faecium isolates. The total size of resistant E. faecium isolates is shown on the left bar plot. The multi-resistant profile is visualised by the bottom plot and the occurrence is represented on the top bar plot.*

E. faecium isolates, where technical count refers to the replicates of biological samples during MALDI-TOF analysis, were profiled as follows:

- Benzylpenicillin: 59 biological (224 technical) and 19 biological (74 technical) isolates as being susceptible and resistant classes, respectively.

- Cefovecin: 22 biological (82 technical) and 61 biological (237 technical) isolates as being susceptible and resistant classes, respectively.

- Enrofloxacin: 15 biological (56 technical) and 62 biological (238 technical) isolates as being susceptible and resistant classes, respectively.

- Nitrofurantoin: 25 biological (96 technical) and 12 biological isolates (46 technical replicates) isolates as being susceptible and resistant classes, respectively.

- Clindamycin: 31 biological (118 technical) and 56 biological (214 technical) isolates as being susceptible and resistant classes, respectively.

- Erythromycin: 32 biological (122 technical) and 18 biological (66 technical) isolates as being susceptible and resistant classes, respectively. The detailed information about these *E. faecium* isolates can be seen in Table 6-2.

Table 6-2. MIC values of Enterococcus faecium isolates against benzylpenicillin, cefovecin, enrofloxacin, nitrofurantoin, clindamycin and erythromycin. *In this table, only susceptible and resistant isolates which are used in further machine learning analysis are shown, whereas intermediate isolates are not shown.*

	MIC Values	0.12		0.25		0.5		1		2		4		8		16		32		128		256		Susceptible Total	Resistant Total
Antibiotic		Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding	Milk	Bedding		
Benzylpenicillin	biological	25	2	10	7	1	7	4	3					3	1	12	3							**59**	**19**
	technical	92	8	39	26	4	26	16	13					13	4	46	11							**224**	**74**
	profile	Susceptible		Susceptible		Susceptible		Susceptible						Resistant		Resistant									
Cefovecin	biological					18	1	1	0	1	1			37	24									**22**	**61**
	technical					67	4	3	0	4	4			145	92									**82**	**237**
	profile					Susceptible		Susceptible		Susceptible				Resistant											
Enrofloxacin	biological					10	5			14	4	34	10											**15**	**62**
	technical					37	19			51	15	133	39											**56**	**238**
	profile					Susceptible				Resistant		Resistant													
Nitrofurantoin	biological															6	0	12	7	8	2	2	0	**25**	**12**
	technical															23	0	47	26	30	8	8	0	**96**	**46**
	profile															Susceptible		Susceptible		Resistant		Resistant			
Clindamycin	biological	6	3	16	2	0	1	1	2			19	37											**31**	**56**
	technical	24	11	60	8	0	3	4	8			72	142											**118**	**214**
	profile	Susceptible		Susceptible		Susceptible						Resistant													
Erythromycin	biological			18	4	8	2							9	9									**32**	**18**
	technical			66	17	31	8							35	31									**122**	**66**
	profile			Susceptible		Susceptible								Resistant											

6.3.2 Generation of Peak List and Algorithm Tuning

To compare each antimicrobial profile (resistant vs susceptible), the relative intensity height filter was set to 1, where the maximum intensity value was set to 100 and others normalized accordingly, as previously suggested (Ressom *et al.*, 2007). Peak selection was performed based on Welch's *t*-test for the normally distributed data of the following analyses; *E. faecalis*: benzylpenicillin, chloramphenicol and clindamycin; *E. faecium*: benzylpenicillin and clindamycin. *E. faecalis* erythromycin analysis gave the best prediction performance results by having used all of the peaks (including peaks found to be statistically non-significant by Welch's *t*-test and Wilcoxon test, but still present in 30% of the spectra) whereas for the rest of the analyses (*E. faecalis*: tetracycline and TMP/SMX; *E. faecium*: cefovecin, enrofloxacin, erythromycin and nitrofurantoin) peak selection was performed based on Wilcoxon test, which is more robust to non-normally distributed data (Wilcoxon, 1992).

6.3.3 Analyses with *E. faecalis* Isolates

Resistant and susceptible *E. faecalis* to each antimicrobial class – benzylpenicillin, chloramphenicol, clindamycin, erythromycin, tetracycline and TMP/SMX – were labelled as positive and negative, respectively. Due to the imbalanced nature between resistant and susceptible classes of benzylpenicillin, chloramphenicol, clindamycin, erythromycin and TMP/SMX datasets, each was resampled by oversampling the minority class to build robust classifiers. The minority class was the resistant isolates for benzylpenicillin, chloramphenicol, erythromycin and TMP/SMX datasets; and susceptible isolates were clindamycin. Feature (peak) selection was performed in each analysis except the erythromycin dataset. This was done by selecting the statistically significant peaks that appear in at least 30% of all number of the spectra. The detailed information is shown in Table 6-3.

Table 6-3. The counts of the biological and technical (spectra) replicate in each class, feature selection and re-balancing techniques in analyses with E. faecalis isolates.

Analysis	Resistant isolate (spectra count)	Susceptible isolate (spectra count)	Feature Selection	Resampling Technique
Benzylpenicillin	6 (26)	18 (70)	*ptta<0.05*	Oversampling the resistant class
Chloramphenicol	11 (44)	97 (392)	*ptta<0.05*	Oversampling the resistant class
Clindamycin	75 (301)	11 (44)	*ptta<0.05*	Oversampling the susceptible class
Erythromycin	24 (99)	38 (150)	-	Oversampling the resistant class
Tetracycline	51 (210)	56 (220)	*pwkw<0.05*	-
TMP/SMX	16 (64)	91 (365)	*pwkw<0.05*	Oversampling the resistant class

Ptta is the *p*-value of Welch's *t*-test and *pwkw* is the *p*-value of Wilcoxon test.

The prediction performances after 30 observations of the most powerful algorithms in each discrimination analysis between resistant and susceptible *E. faecalis* isolates are shown and summarized in Figure 6-3 and Table 6-4, respectively.

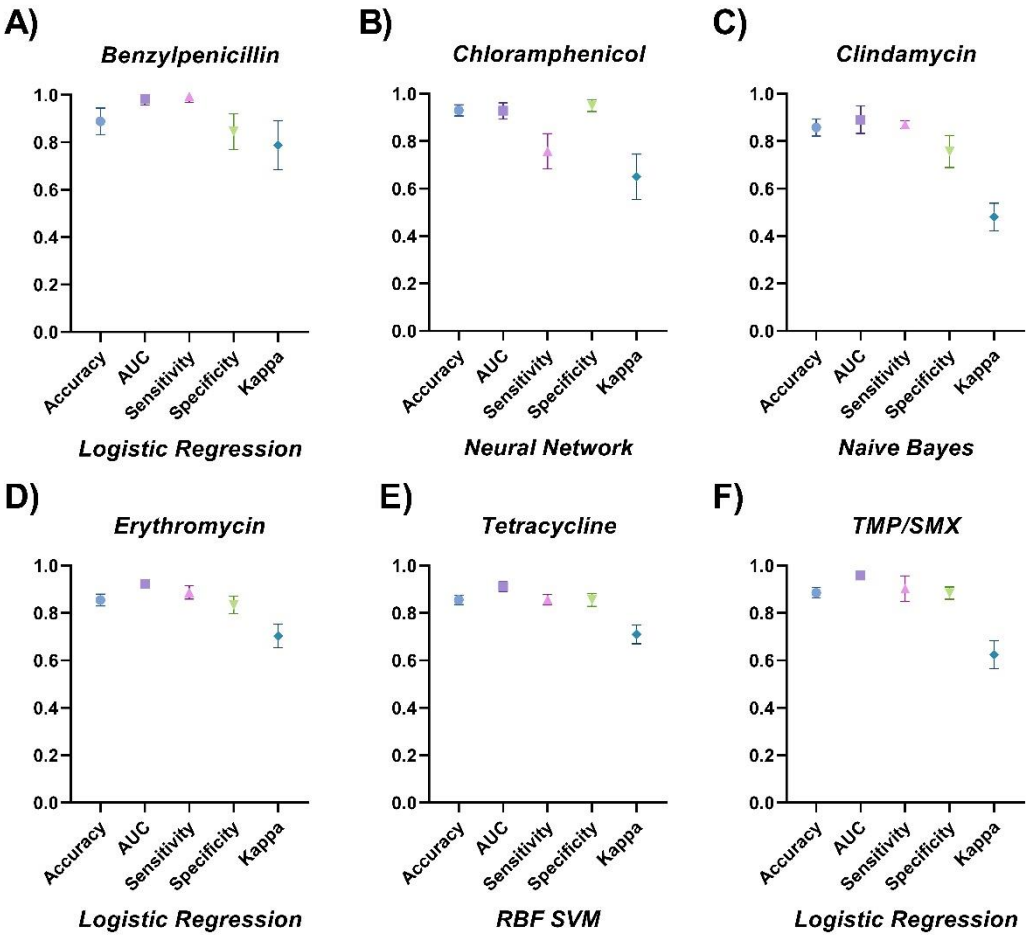


Figure 6-3. Best prediction performances (accuracy, AUC, sensitivity, specificity and Kappa) in discrimination of resistant and susceptible profiles of E. faecalis isolates. A) *Logistic regression was found to be the best algorithm to discriminate benzylpenicillin resistant and susceptible E. faecalis isolates.* ***B)*** *MLP neural network was found to be the best algorithm to discriminate chloramphenicol resistant and susceptible E. faecalis isolates.* ***C)*** *Naive Bayes was found to be the best algorithm to discriminate clindamycin resistant and susceptible E. faecalis isolates.* ***D)*** *Logistic regression was found to be the best algorithm to discriminate erythromycin resistant and susceptible E. faecalis isolates.* ***E)*** *RBF SVM was found to be the best algorithm to discriminate tetracycline resistant and susceptible E. faecalis isolates.* ***F)*** *Logistic regression was found to be the best algorithm to discriminate TMP/SMX (trimethoprim/sulfamethoxazole) resistant and susceptible E. faecalis isolates. These graphs were generated in GraphPad Prism v8.*

Table 6-4. Best prediction performers and their exact performance values for discrimination of resistant and susceptible profiles of E. faecalis isolates.

Analysis	Best Algorithm	Accuracy	AUC	Kappa	Sensitivity	Specificity
Benzylpenicillin-Resistant vs Susceptible	Logistic Regression[a]	88.84±5.59%	98.44±2.74%	78.82±10.35%	99.33±2.54%	84.55±7.59%
Chloramphenicol-Resistant vs Susceptible	MLP Neural Network[b]	93.00±2.34%	92.85±3.37%	65.04±9.52%	75.77±7.42%	95.01±2.48%
Clindamycin-Resistant vs Susceptible	Naive Bayes[c]	85.77±3.54%	89.09±5.81%	48.01±5.82%	87.03±1.71%	75.67±6.79%
Erythromycin-Resistant vs Susceptible	Logistic Regression[d]	85.47±2.51%	92.09±1.68%	70.33±4.99%	88.70±2.81%	83.39±3.64%
Tetracycline-Resistant vs Susceptible	RBF SVM[e]	85.52±1.99%	91.03±2.11%	71.02±3.96%	85.61±2.23%	85.46±2.67%
TMP/SMX-Resistant vs Susceptible	Logistic Regression[f]	88.58±2.14%	95.76±1.36%	62.36±5.93%	90.31±5.34%	88.36±2.57%

[a] See Figure 6-3-A. The prediction performance of other ML algorithms can be seen in Supplementary Figure 1.
[b] See Figure 6-3-B. The prediction performance of other ML algorithms can be seen in Supplementary Figure 2.
[c] See Figure 6-3-C. The prediction performance of other ML algorithms can be seen in Supplementary Figure 3.
[d] See Figure 6-3-D. The prediction performance of other ML algorithms can be seen in Supplementary Figure 4.
[e] See Figure 6-3-E. The prediction performance of other ML algorithms can be seen in Supplementary Figure 5.
[f] See Figure 6-3-F. The prediction performance of other ML algorithms can be seen in Supplementary Figure 6.

6.3.4 Analyses with *E. faecium* Isolates

Resistant and susceptible *E. faecium* isolates to each antimicrobial class – benzylpenicillin, cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin – were labelled as positive and negative, respectively. Due to the imbalanced nature between resistant and susceptible classes of benzylpenicillin, cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin datasets, each was resampled by oversampling the minority class to build robust classifiers. The minority class was the resistant isolates for benzylpenicillin, erythromycin and nitrofurantoin datasets; and susceptible isolates were cefovecin, clindamycin and enrofloxacin. Feature (peak) selection was performed in each analysis by selecting the statistically significant peaks that appear in at least 30% of all number of spectra. The detailed information is shown in Table 6-5.

Table 6-5. The counts of the biological and technical (spectra) replicate in each class, feature selection and re-balancing techniques in analyses with E. faecium isolates.

Analysis	Resistant isolate (spectra count)	Susceptible isolate (spectra count)	Feature Selection	Resampling Technique
Benzylpenicillin	19 (74)	59 (224)	*ptta<0.05*	Oversampling the resistant class
Cefovecin	61 (237)	22 (82)	*pwkw<0.05*	Oversampling the susceptible class
Clindamycin	56 (214)	31 (118)	*ptta<0.05*	Oversampling the susceptible class
Enrofloxacin	62 (238)	15 (56)	*pwkw<0.05*	Oversampling the susceptible class
Erythromycin	18 (66)	32 (122)	*pwkw<0.05*	Oversampling the resistant class
Nitrofurantoin	12 (46)	25 (96)	*pwkw<0.05*	Oversampling the resistant class

Ptta is the *p*-value of Welch's *t*-test and *pwkw* is the *p*-value of Wilcoxon test.

The prediction performances after 30 observations of the most powerful algorithms in each discrimination analysis between resistant and susceptible *E. faecium* isolates are shown and summarized in Figure 6-4 and Table 6-6, respectively.

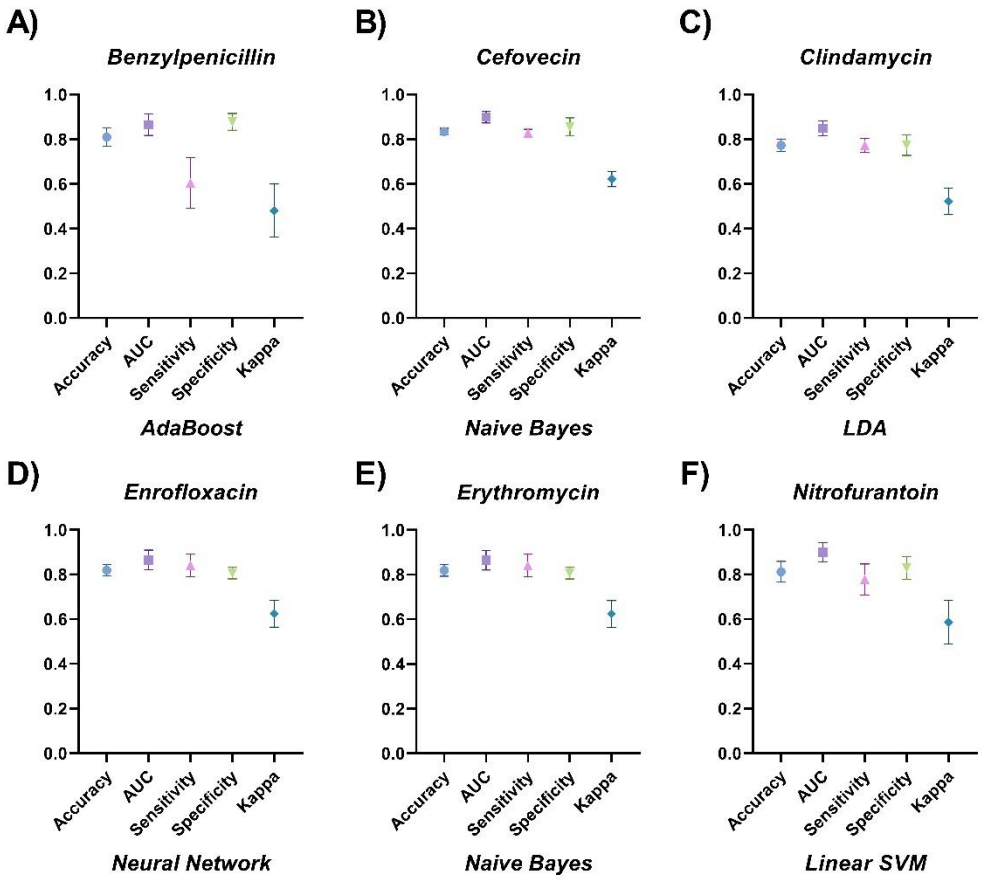


Figure 6-4. Best prediction performances in discrimination of resistant and susceptible profiles of E. faecium isolates. ***A)*** *AdaBoost was found to be the best algorithm to discriminate benzylpenicillin-resistant and susceptible E. faecium isolates.* ***B)*** *Naïve Bayes was found to be the best algorithm to discriminate cefovecin-resistant and susceptible E. faecium isolates.* ***C)*** *LDA (linear discriminant analysis) was found to be the best algorithm to discriminate clindamycin-resistant and susceptible E. faecium isolates.* ***D)*** *MLP neural network was found to be the best algorithm to discriminate enrofloxacin-resistant and susceptible E. faecium isolates.* ***E)*** *Naïve Bayes was found to be the best algorithm to discriminate erythromycin-resistant and susceptible E. faecium isolates.* ***F)*** *LSVM was found to be the best algorithm to discriminate nitrofurantoin-resistant and susceptible E. faecium isolates. These graphs were generated in GraphPad Prism v8.*

Table 6-6. Best prediction performers and their exact performance values for discrimination of resistant and susceptible profiles of E. faecium isolates.

Analysis	Best Algorithm	Accuracy	AUC	Kappa	Sensitivity	Specificity
Benzylpenicillin-Resistant vs Susceptible	AdaBoost[a]	80.92±4.12%	86.52±4.85%	48.06±11.88%	60.50±11.35%,	87.80±3.78%
Cefovecin-Resistant vs Susceptible	Naive Bayes[b]	83.47±1.41%	89.98±2.66%	62.21±3.43%	82.72±1.72%	85.57±4.04%
Clindamycin-Resistant vs Susceptible	LDA[c]	77.33±2.78%	84.87±3.24%	52.23±5.95%	77.33±3.09%	77.40±4.65%
Enrofloxacin-Resistant vs Susceptible	MLP Neural Network[d]	81.85±2.58%	86.46±4.40%	62.42±5.93%	84.10±5.02%	80.66±2.65%
Erythromycin-Resistant vs Susceptible	Naive Bayes[e]	81.85±2.58%	86.46±4.40%	62.42±5.93%	84.10±5.02%	80.66±2.65%
Nitrofurantoin-Resistant vs Susceptible	LSVM[f]	81.22±4.64%	89.90±4.37%	58.65±9.74%	77.77±6.91%	82.93±5.14%

[a] See Figure 6-4-A. The prediction performance of other ML algorithms can be seen in Supplementary Figure 7.
[b] See Figure 6-4-B. The prediction performance of other ML algorithms can be seen in Supplementary Figure 8.
[c] See Figure 6-4-C. The prediction performance of other ML algorithms can be seen in Supplementary Figure 9.
[d] See Figure 6-4-D. The prediction performance of other ML algorithms can be seen in Supplementary Figure 10.
[e] See Figure 6-4-E. The prediction performance of other ML algorithms can be seen in Supplementary Figure 11.
[f] See Figure 6-4-F. The prediction performance of other ML algorithms can be seen in Supplementary Figure 12.

6.3.5 Biomarker Characterisation

The peaks with the relative intensity of 10 and above, which is the average value in almost every analysis and provides a representative number of biological biomarkers in each analysis, were taken from the statistical report of each analysis, as characterizing all the peaks would not be convenient. The discriminant peaks found in the comparisons of resistant and susceptible profiles of each antibiotic for *E. faecalis* and *E. faecium* were cross-matched with the proteins are shown in Table 6-7 and Table 6-8, respectively.

Table 6-7. Discriminant peaks in each antimicrobial analysis of E. faecalis with the corresponding proteins and top PSI-BLAST match of these proteins with their cellular location.

MALDI-TOF Peak (Mw)	Protein (Mw)	PSI-BLAST Match	PSORTB location (score)	Discriminant Profile
4440.35Da	PTS family porter (4440.13Da)	PTS family porter	Unknown (2.5)	BENZ, CLIN
4766.09Da	HP-1 (4769.57Da)	Hypothetical protein	Unknown (2.5)	BENZ, CHLO, ERYT, TETRA
5557.37Da	LuxR family protein (5547.38Da)	Bacterial response regulator	Unknown (2.5)	BENZ, CHLO, ERYT, TETRA, TMP/SMX
6224.00Da	RpmD (6224.27Da)	50S ribosomal protein L30	Cytoplasmic (9.67)	ERYT, TETRA, TMP/SMX
6399.48Da	RpmF (6399.48Da)	50S ribosomal protein L32	Cytoplasmic (9.67)	CHLO, ERYT
6669.68Da	HP-2 (6670.51Da)	Hypothetical protein	Cytoplasmic (7.5)	ERYT
6858.79Da	RpmB (6857.07Da)	50S ribosomal protein L28	Cytoplasmic (9.67)	ERYT, TETRA, CLIN, TMP/SMX
7022.47Da	RpsZ (7022.32Da)	30S ribosomal S14	Cytoplasmic (9.97)	CHLO, ERYT, TMP/SMX
7327.72Da	RpmC (7329.55Da)	50S ribosomal L29	Cytoplasmic (9.97)	ERYT, TETRA, TMP/SMX
7574.28Da	RpmI (7569.00Da)	50S ribosomal L35	Cytoplasmic (9.67)	TETRA, CLIN, TMP/SMX
8106.05Da	InfA (8105.42Da)	Translation initiation factor 1A/IF-1	Cytoplasmic (10)	TETRA, TMP/SMX
9111.40Da	RpsR (9110.64Da)	30S ribosomal S18	Cytoplasmic (9.97)	CHLO, ERYT, TETRA, TMP/SMX
9524.71Da	HU (9524.89Da)	DNA-binding protein HBsu	Cytoplasmic (9.97)	CHLO, CLIN, ERYT, TETRA, TMP/SMX
10510.85Da	RpsO (10511.97Da)	30S ribosomal S15	Cytoplasmic (9.67)	TETRA, TMP/SMX
1111.30Da	RplX (11115.04Da)	50S ribosomal L24	Cytoplasmic (9.67)	CHLO, BENZ, ERYT, TETRA, TMP/SMX

HP: hypothetical protein, BENZ: benzylpenicillin, CHLO: chloramphenicol, CLIN: clindamycin, ERYT: erythromycin, TETRA: tetracycline, TMP/SMX: trimethoprim/ sulfamethoxazole.

Table 6-8. Discriminant peaks in each antimicrobial analysis of E. faecium with the corresponding proteins and top PSI-BLAST match of these proteins with their cellular location.

MALDI-TOF Peak (Mw)	Protein (Mw)	PSI-BLAST Match	PSORTB location (score)	Discriminant Profile
4488.85Da	HP-1 (4487.34Da)	Putative metal homeostasis protein	Unknown (2.5)	BENZ, CEFO, ERYT
5036.95Da	HP-2 (5047.09Da)	Enterocin L50 family leaderless bacteriocin	Cytoplasmic (9.55)	CEFO, ENROF, NITRO
6507.02Da	RpmF (6510.57Da)	50S ribosomal protein L32	Cytoplasmic (9.67)	BENZ, CLIN
7273.74Da	CsbD (7287.14Da)	CsbD-like family protein	Unknown (2.5)	ENROF
7324.94Da	SHK (7337.42Da)	Two-component system sensor histidine kinase	Cytoplasmic (7.5)	CEFO, ERYT, CLIN, NITRO
8977.49Da	RpsR (8980.54Da)	30S ribosomal protein S18	Cytoplasmic (9.97)	BENZ, CEFO, CLIN
9056.35Da	RpsT (9059.33Da)	30S ribosomal protein S20	Cytoplasmic (9.67)	BENZ, CLIN
9547.57Da	HU (9550.88Da)	Bacterial DNA-binding protein	Cytoplasmic (9.97)	BENZ, CLIN
10068.06Da	PTS-IIB (10079.74Da)	PTS sugar transporter subunit IIB	Cytoplasmic (7.5)	CEFO, ENROF, NITRO, ERYT
10934.36Da	RplX (10937.90Da)	50S ribosomal L24	Cytoplasmic (9.97)	BENZ

HP: hypothetical protein, BENZ: benzylpenicillin, CEFO: cefovecin, CLIN: clindamycin, ENROF: enrofloxacin, ERYT: erythromycin, NITRO: nitrofurantoin.

Analyses of *E. faecalis* isolates based on their antimicrobial profiles resulted in two common proteins (LuxR family protein and RplX) to discriminate chloramphenicol-resistant vs susceptible, tetracycline-resistant vs susceptible, benzylpenicillin-resistant vs susceptible, erythromycin-resistant vs susceptible and TMP/SMX-resistant vs susceptible profiles (see Figure 6-5-A). Bacterial DNA-binding HU protein was found common to discriminate chloramphenicol-resistant vs susceptible, tetracycline-resistant vs susceptible, clindamycin-resistant vs susceptible, erythromycin-resistant vs susceptible and TMP/SMX-resistant vs susceptible profiles. RpsR was found in the discrimination analyses of tetracycline-resistant vs susceptible, erythromycin-resistant vs susceptible, chloramphenicol-resistant vs susceptible and TMP/SMX-

resistant vs susceptible profiles. HP-1 (Mw: 4769.57Da) was found common to discriminate chloramphenicol-resistant vs susceptible, tetracycline-resistant vs susceptible, benzylpenicillin-resistant vs susceptible, and erythromycin-resistant vs susceptible. RpmB was found common to discriminate clindamycin-resistant vs susceptible, tetracycline-resistant vs susceptible, TMP/SMX-resistant vs susceptible, and erythromycin-resistant vs susceptible. Two common proteins (RpmC and RpmD) were found in the discrimination analyses of erythromycin-resistant vs susceptible, tetracycline-resistant vs susceptible and TMP/SMX-resistant vs susceptible profiles. RpsZ was found common in the discrimination of chloramphenicol-resistant vs susceptible, erythromycin-resistant vs susceptible and TMP/SMX-resistant vs susceptible profiles. RpmI was found common to discriminate tetracycline-resistant vs susceptible, clindamycin-resistant vs susceptible and TMP/SMX-resistant vs susceptible profiles. Two common proteins (InfA and RpsO) were found in the discrimination analyses of tetracycline-resistant vs susceptible and TMP/SMX-resistant vs susceptible profiles. RpmF was found common in the discrimination of chloramphenicol-resistant vs susceptible and erythromycin-resistant vs susceptible profiles. PTS family porter protein was found common to discriminate benzylpenicillin-resistant vs susceptible and clindamycin-resistant vs susceptible profiles. HP-2 (Mw: 6670.51Da) was found unique to the analysis of erythromycin-resistant vs susceptible profiles (see Figure 6-5-A).

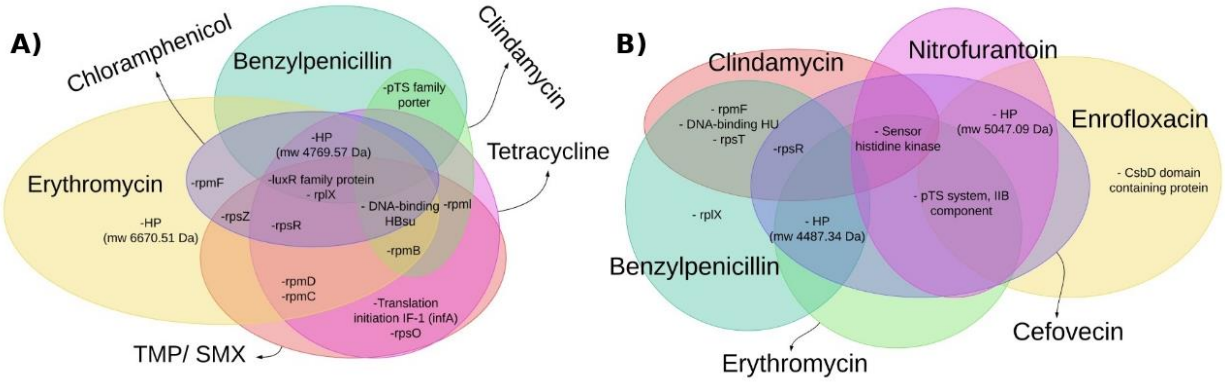


Figure 6-5. Discriminant proteins of Enterococcus faecalis and Enterococcus faecium between resistant and susceptible profiles of each antibiotic. *These figures were generated using Lucidchart.com.*

Analyses of *E. faecium* isolates based on their antimicrobial profiles resulted in that PTS system IIB component was common to discriminate cefovecin-resistant vs susceptible, nitrofurantoin-resistant vs susceptible, erythromycin-resistant vs susceptible, and enrofloxacin-resistant vs susceptible profiles (see Figure 6-5-B). Sensor histidine kinase (SHK) protein was

found common in the discrimination of erythromycin-resistant vs susceptible, nitrofurantoin-resistant vs susceptible, clindamycin-resistant vs susceptible and cefovecin-resistant vs susceptible profiles. RpsR was found common in the discrimination of clindamycin-resistant vs susceptible, benzylpenicillin-resistant vs susceptible and cefovecin-resistant vs susceptible profiles. HP-1 (Mw: 4487.34Da) was found common in the discrimination of benzylpenicillin-resistant vs susceptible, erythromycin-resistant vs susceptible, and cefovecin-resistant vs susceptible profiles. HP-2 (Mw: 5047.09Da) was fund common in the discrimination of cefovecin-resistant vs susceptible, nitrofurantoin-resistant vs susceptible, and enrofloxacin-resistant vs susceptible profiles. Three common proteins (RpmF, DNA-binding HU and RpsT) were found to discriminate clindamycin-resistant vs susceptible and benzylpenicillin-resistant vs susceptible profiles. RplX and a CsbD domain-containing protein were found unique to benzylpenicillin-resistant vs susceptible and enrofloxacin-resistant vs susceptible profiles, respectively (see Figure 6-5-B).

6.3.5.1 Functional Characterisation of Discriminant Proteins

3D protein modelling is mainly used to estimate the biological functions of the proteins as the protein structure governs the interaction of it with ligands or other molecules (Lopez *et al.*, 2007). In this section, SWISS-MODEL (homology modelling) was used to predict 3D modelling structures of known-function proteins, and I-TASSER server (threading/folding recognition) was used to predict 3D modelling structures and GO terms of less known-function proteins. 3D models of these discriminant proteins are shown in Figure 6-6 and Figure 6-7 for *E. faecalis* and *E. faecium*, respectively.

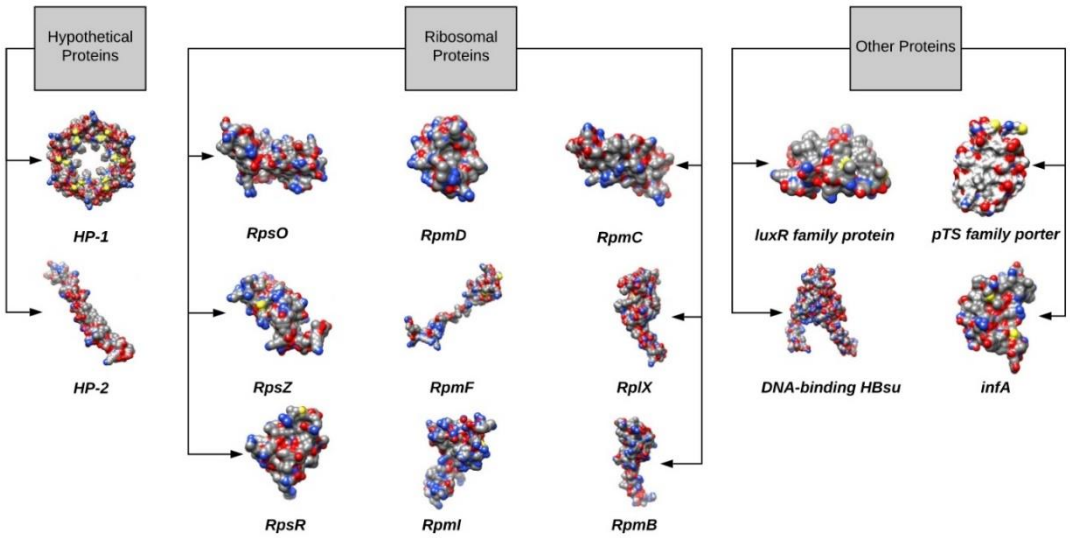


Figure 6-6. 3D structures of the discriminant proteins of Enterococcus faecalis between resistant and susceptible profiles of each antibiotic. *Hypothetical proteins: HP-1 (Mw: 4769.57Da), HP-2 (Mw: 6670.51Da). Ribosomal proteins: 30S ribosomal S15 (RpsO), 30S ribosomal S14 (RpsZ), 30S ribosomal S18 (RpsR), 50S ribosomal L30 (RpmD), 50S ribosomal L32 (RpmF), 50S ribosomal L35 (RpmI), 50S ribosomal L29 (RpmC), 50S ribosomal L24 (RplX) and 50S ribosomal L28 (RpmB). Other proteins: LuxR family protein, DNA-binding protein HBsu, PTS family porter, translation initiation factor IF-1 (InfA). The visualisation was carried out with UCSF Chimera and Lucidchart.com.*

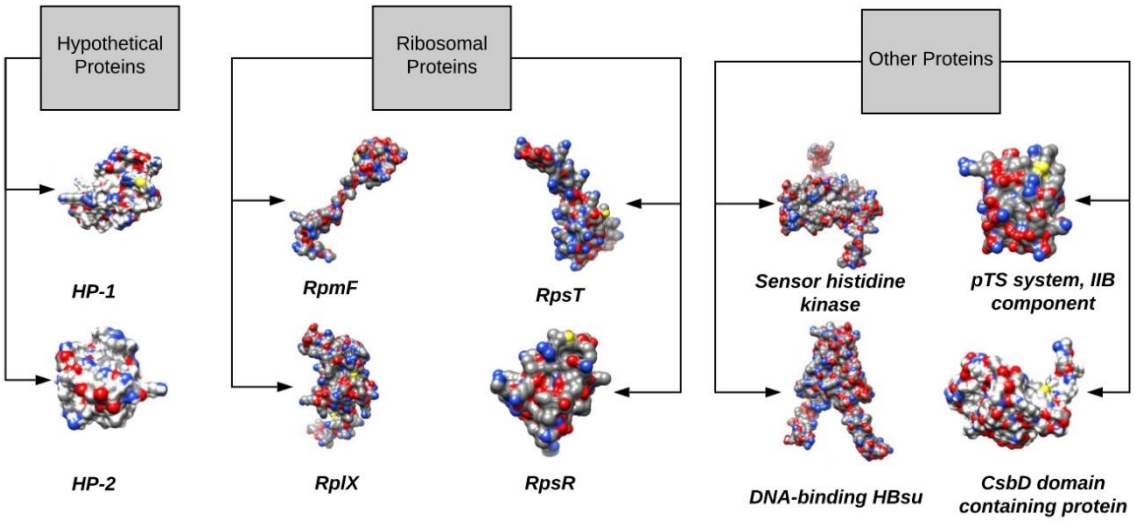


Figure 6-7. 3D structures of the discriminant proteins of Enterococcus faecium between resistant and susceptible profiles of each antibiotic. *Hypothetical proteins: HP-1 (Mw: 4487.34Da) and HP-2 (Mw: 5047.09Da). Ribosomal proteins: 50S ribosomal L32 (RpmF), 50S ribosomal protein L24 (RplX), 30S ribosomal S20 (RpsT) and 30S ribosomal protein S18 (RpsR). Other proteins: sensor histidine kinase, DNA-binding protein HBsu, pTS system IIB component and CsbD domain-containing protein. The visualisation was carried out with UCSF Chimera and Lucidchart.com.*

PTS (phosphoenolpyruvate: carbohydrate phosphotransferase system) protein was found to be discriminant between several antimicrobial-resistant and susceptible profiles of both *E. faecalis* and *E. faecium* isolates in this study. PTS family porters are involved in the transportation and phosphorylation of carbohydrates as well as signal transduction and regulatory functions (Ruiz-Cruz *et al.*, 2016). In some bacteria species, *pTS* genes were proved to have virulence effects (Hava and Camilli, 2002; Jones, Knoll and Rubens, 2000). Moreover, some classes of PTS porters were suggested to have a function in bacteria for survival in animal hosts especially in epithelial cells (Zuniga *et al.*, 2005). The function of PTS family porters and biofilm formation has been shown to be associated in gram-positive bacteria (Sutrina, McGeary and Bourne, 2007), which may be important as biofilm-forming organisms are known to show increased antimicrobial-resistant (Sobisch *et al.*, 2019).

DNA-binding HU protein was found to be discriminant between several antimicrobial-resistant and susceptible profiles of both *E. faecalis* and *E. faecium* isolates in this study. DNA-binding HU protein is considered to be more significant in gram-positive bacteria than gram negatives (Macvanin and Adhya, 2012). CsbD protein is involved in the cellular response to environmental stress conditions (Zuber, 2001). A Csbd-like domain-containing protein was previously found to be one of the discriminant proteins in several processes such as pathogenic vs non-pathogenic *E. coli* strains (Fagerquist *et al.*, 2010) and aflatoxigenic vs non-aflatoxigenic *Aspergillus flavus* strains (Pennerman *et al.*, 2019).

LuxR family protein was found discriminant for resistant and susceptible profiles of *E. faecalis* for the following antibiotics benzylpenicillin, chloramphenicol, erythromycin, tetracycline and TMP/SMX. LuxR family proteins are known to be regulators involved in the quorum-sensing system (Nasser and Reverchon, 2007). In different species, LuxR homologs were suggested to be involved in controlling oxidative stress (Hudaiberdiev *et al.*, 2015), which could be an explanation for the discriminant profile in our experiment, as bactericidal antimicrobials induce oxidative stress (Kohanski, Dwyer and Collins, 2010; Marrakchi, Liu and Andreescu, 2014).

SHK is part of two-component signal transduction systems (TCSs) in bacteria, which respond to changes in environmental conditions (Gao and Stock, 2009). TCSs are essential for cell survival and are also responsible for pathogenicity, biofilm formation and quorum sensing (Gotoh *et al.*, 2010). The TCSs of some bacteria species, including *E. faecalis*, are recommended as novel drug targets (Gotoh *et al.*, 2010; Bem *et al.*, 2015; Tiwari *et al.*, 2017). Furthermore,

AMR genes of vancomycin are regulated by TCSs in *Enterococcus* species (Mueller-Premru *et al.*, 2009; Koteva *et al.*, 2010).

RpsO was one of the discriminant peaks between resistant and susceptible *E. faecalis* isolates of tetracycline and TMP/SMX in the current study. However, it was found to be one of the four discriminant transcripts between azithromycin, a macrolide class antibiotic, resistant and susceptible isolates of *Neisseria gonorrhoeae* (Wadsworth *et al*., 2019). RpmC and RpmD were found to be discriminant protein between erythromycin, tetracycline and TMP/SMX-resistant and susceptible profiles of *E. faecalis* in the current study. The abundance of *rpmC* showed significant alterations for environmental conditions such as copper stress (Tarrant *et al.*, 2019), antimicrobial activity (Choi *et al.*, 2011; Alves *et al.*, 2019) or presence of other organisms (Luppens *et al.*, 2008). Expression of *rpmD* was observed to be downregulated in studies where antibiotic tolerance related systems were knocked out (Sharma-Kuinkel *et al.*, 2009). RpmB was one of the discriminant peaks between resistant and susceptible *E. faecalis* isolates of erythromycin, tetracycline, clindamycin and TMP/SMX in our study. However, *rpmB* was shown to be inhibited by streptomycin, an aminoglycoside antibiotic, in *M. tuberculosis* (Fan *et al.*, 2014). In our study, RpmI was found discriminant between tetracycline, clindamycin and TMP/SMX-resistant vs susceptible profiles of *E. faecalis*. *RpmI* was suggested to be essential for protease restoration (Kitten and Willis, 1996), which play an important role in AMR profile in some species (Fernández *et al.*, 2012).

RpsR was found to be discriminant between several antimicrobial profiles of *E. faecalis* and *E. faecium* such as erythromycin, tetracycline, clindamycin, benzylpenicillin etc. *RpsR* gene in *E. faecalis* was shown to be upregulated by the treatment of glyphosate -an active ingredient of widely used herbicides (Saunders and Pezeshki, 2015)- and its primary breakdown product of aminomethylphosphonic acid (Stenger, 2019). Similar results with our study about the role of *rpsR* in resistance to erythromycin (Xu *et al.*, 2010) and multidrug were found earlier (Pathania *et al.*, 2009). RplX was found to be discriminant between benzylpenicillin and susceptible profiles of *E. faecium*. Moreover, it was found to differentiate *E. faecalis* isolates based on the resistance profile of several antibiotics including benzylpenicillin. In previous studies with other species, *rplX* gene was shown to be differentially expressed in the presence of antimicrobial agent lupulone and betulinaldehyde (Wei *et al.*, 2014; Chung, Chung and Navaratnam, 2013).

Proteins mainly act as a team rather than individual to perform their biological functions at cellular and system levels (De Las Rivas and Fontanillo, 2010; Berggård, Linse and James, 2007; Eisenberg *et al.*, 2000). In this study, PPI was used to outline protein complexes and learn their biological pathways in detail. PPI analysis found that the clustering coefficient of 295 proteins - a total of 15 discriminant proteins and 270 first neighbour proteins – was 0.611, while randomly selected 15 proteins (this is repeated 10 times) and their 368 first neighbour proteins (on average) network in *E. faecalis* proteome had a clustering coefficient of 0.035. The clustering coefficient provides a scale of the interconnectivity of a network (between 0 and 1, where 1 is for all neighbours connected and 0 for no connection of neighbours) (Ravasz *et al.*, 2002). The average number of neighbours per protein for the clusters of interest and the randomly built PPI network of *E. faecalis* was 68.298 and 2.232, respectively. The average number of neighbours represents the mean connectivity value of a protein in the network (Assenov *et al.*, 2008). The clusters of interest and the randomly built PPI network had 0.232 and 0.006 for network density and 0.716 and 4.105 for network heterogeneity, respectively. Network density is the normalized version of the average number of neighbours, and proteins with no connectivity are rated 0, while proteins with lots of connections are given values closer to 1 (Assenov *et al.*, 2008). Network heterogeneity indicates the variance of connectivity and the more hub nodes, the greater value it has (Dong and Horvath, 2007).

For *E. faecium*, it was found that the clustering coefficient of 345 proteins - a total of 8 discriminant proteins and 337 first neighbour proteins – was 0.383 while randomly selected 8 proteins (this is repeated 10 times) and their 358 first neighbour proteins network in *E. faecium* proteome had a clustering coefficient of 0.012. The average number of neighbours per protein in the clusters of interest and the randomly built PPI network was 4.591 and 2.211, respectively. The clusters of interest and the randomly built PPI network had 0.013 and 0.007 for network density and 3.517 and 4.168 for network heterogeneity, respectively.

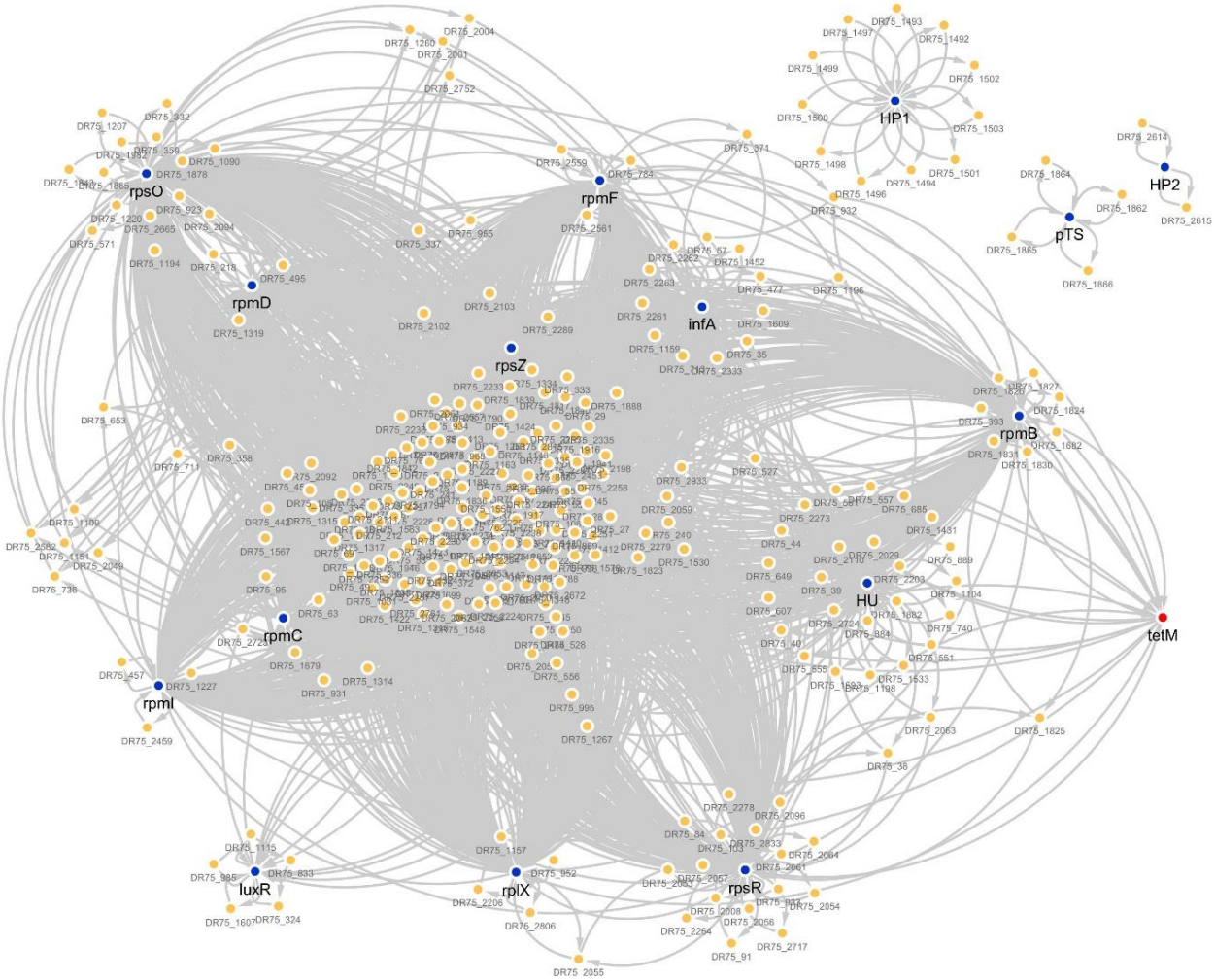


Figure 6-8. The protein-protein interaction (PPI) network showing 295 Enterococcus faecalis proteins. *The blue circles represent the 15 discriminant proteins (RpmB, RpmC, RpmD, RpmF, RpmI, RpsO, RpsZ, RpsR, RplX, HU (DNA-binding protein), pTS family porter, luxR family protein, InfA (translation initiation factor IF-1), HP1 and HP2 (hypothetical proteins)), while yellow circles are the E. faecalis proteins with which the discriminant proteins interact (first shell interacting partners). Amongst these first shell interacting partners, the red circle is TetM (tetracycline resistance protein) is related to antimicrobial resistance.*

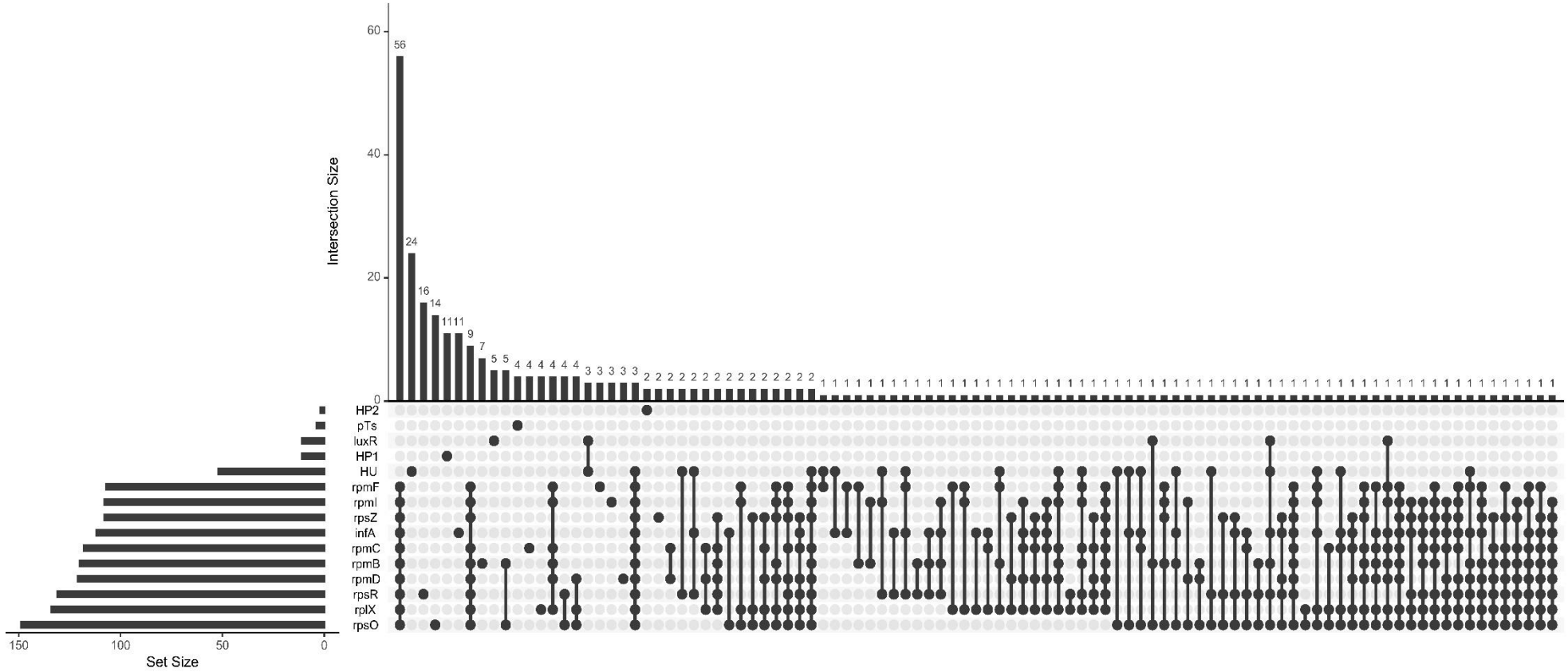


Figure 6-9. UpSet diagram summarizing the interacting sets of discriminant proteins in Enterococcus faecalis. *The total size of E. faecalis proteins with which the discriminant proteins interact (first shell interacting partners) is shown on the left bar plot. Every possible interaction is visualised by the bottom plot and the occurrence is represented on the top bar plot.*

PPI network could be created for all discriminant proteins of *E. faecalis*; however, HP-1 (Mw: 4769.57Da), HP-2 (Mw: 6670.51Da) and PTS family porter did not show cluster with other discriminant proteins (see Figure 6-8). Ribosomal proteins (RpmB, RpmC, RpmD, RpmF, RpmI, RpsO, RpsZ, RpsR and RplX), LuxR family protein, InfA (translation initiation factor) and HU (DNA-binding protein) were found to be interacting directly or via common neighbour nodes (proteins). The occurrence of the common first shell interacting partners between discriminant *E. faecalis* proteins is shown in Figure 6-9.

Moreover, TetM was found directly interacting with discriminant ribosomal proteins (RpmB, RpmC, RpmD, RpmF, RpmI, RpsO, RpsZ, RpsR and RplX) and translation initiation factor (InfA) in the PPI network (see Figure 6-8). Although the neighbour interaction of TetM with our discriminant *E. faecalis* proteins have not been experimentally validated yet, homologs of them were shown in *E. coli* K12 W3110 (Gagarinova *et al.*, 2016; Butland *et al.*, 2005; Hu *et al.*, 2009).

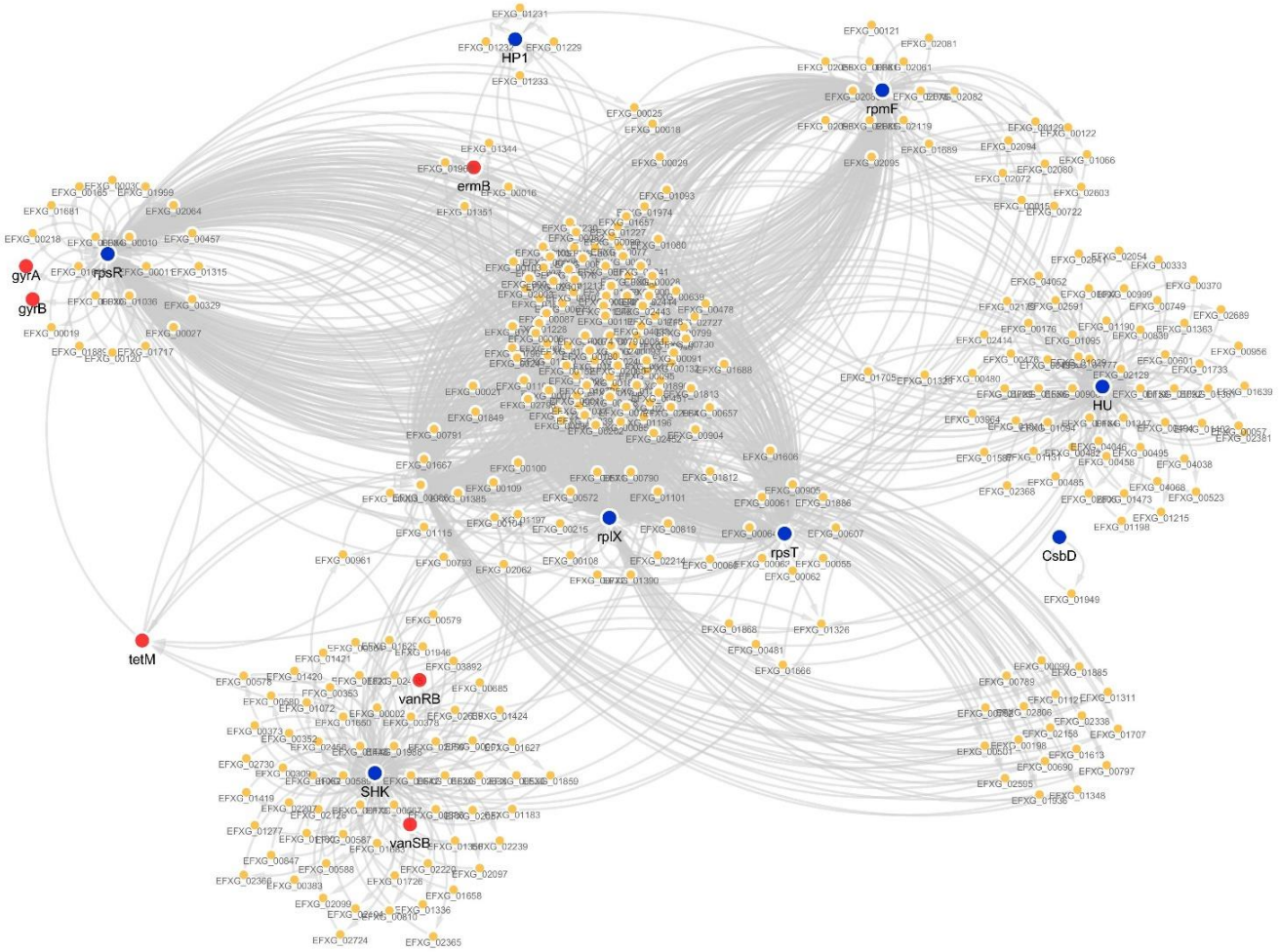


Figure 6-10*. The protein-protein interaction (PPI) network showing 345 Enterococcus faecium proteins.* *The blue circles represent the 8 discriminant proteins (RpsR, RpmF, RplX, RpsT, HU (DNA-binding protein), HP1 (hypothetical protein with Mw of 4487.34Da), sensor histidine kinase (SHK), CsbD like domain-containing protein, while yellow circles are the E. faecium proteins with which the discriminant proteins interact (first shell interacting partners). Amongst these first shell interacting partners, six proteins (red circles): TetM (tetracycline resistance protein), GyrA and GyrB (quinolone resistance proteins), ErmB (macrolide resistance protein), VanSB and VanRB (vancomycin resistance proteins) are related to antimicrobial resistance.*

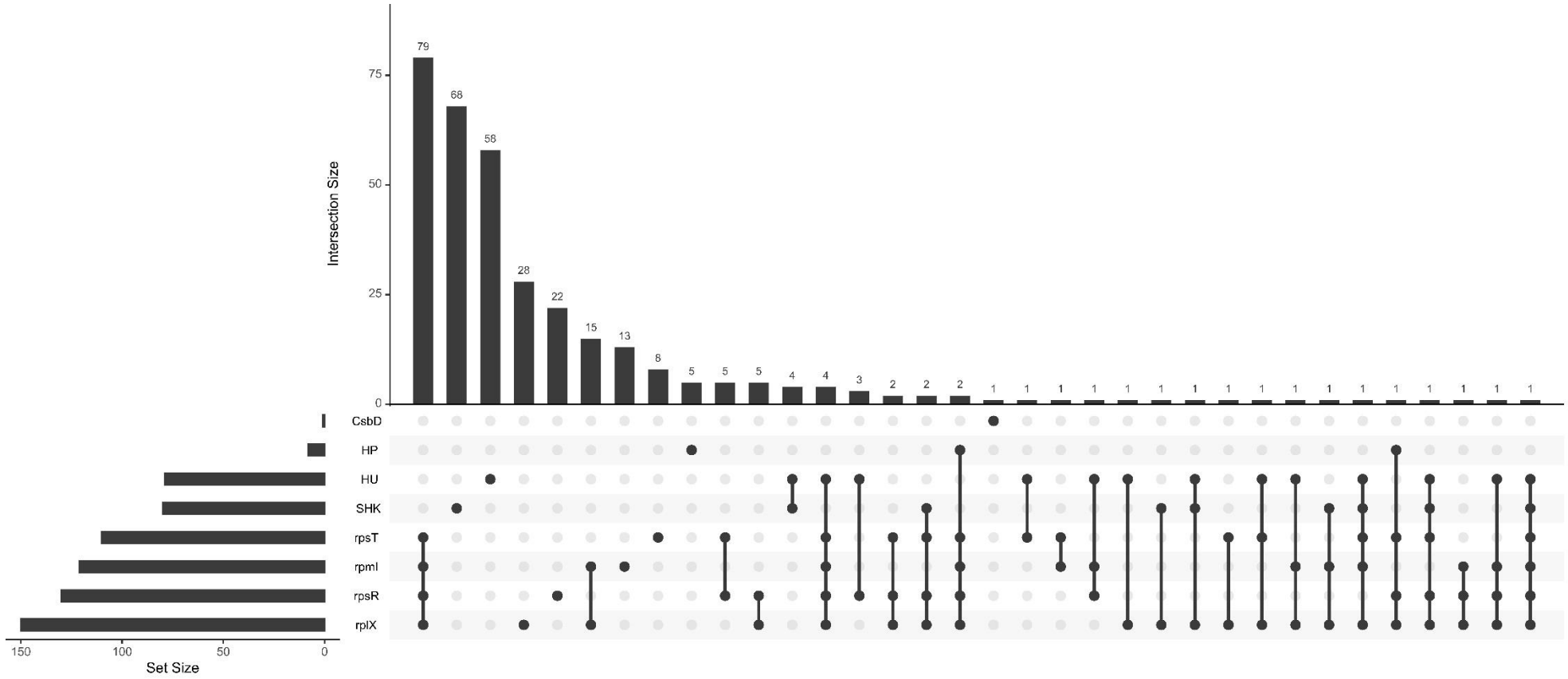


Figure 6-11. UpSet diagram summarizing the interacting sets of discriminant proteins in Enterococcus faecium. *The total size of E. faecium proteins with which the discriminant proteins interact (first shell interacting partners) is shown on the left bar plot. Every possible interaction is visualised by the bottom plot and the occurrence is represented on the top bar plot.*

PPI network could be created for eight out of ten discriminant proteins of *E. faecium* (RpsR, RpmF, RplX, RpsT, HU, HP-1 (Mw: 4487.34Da), SHK and CsbD like domain-containing protein), while HP-2 (Mw: 5047.09Da) and PTS system IIB component did not contain any interactions higher than medium confidence score (see Figure 6-10). Discriminant *E. faecium* proteins were found to be interacting directly or via common neighbour nodes (proteins) as seen in Figure 6-11.

Moreover, AMR proteins in the PPI network of discriminant proteins were found as follows: TetM (tetracycline resistance protein) that directly interacts with the discriminant ribosomal proteins; GyrA and GyrB (quinolone resistance proteins) that directly interact with RpsR; ErmB (macrolide resistance protein) which directly interacts with RpsR and RpsT; VanSB and VanRB (vancomycin resistance proteins) which directly interact with SHK (see Figure 6-10). Although the neighbour interactions of VanSB and VanRB with our discriminant *E. faecium* protein SHK have not been experimentally validated yet, homologs of these proteins in *E. coli* K12 W3110 and MG1655 were found to be interacting (Yamamoto *et al.*, 2005; Babu *et al.*, 2018). Again, the neighbour interaction of TetM with our discriminant *E. faecium* ribosomal proteins - RpsR, RpsT, RplX and RpmF - have not been experimentally validated yet but homologs of them were shown in *E. coli* K12 W3110 (Hu *et al.*, 2009; Butland *et al.*, 2005). No experimental interactions between ErmB and RpsR, ErmB and RpsT, GyrA and RpsR, and GyrB and RpsR were proved in *E. faecium* or other organisms yet. These interactions were predicted through co-expression and/or text-mining in STRING database (Szklarczyk *et al.*, 2018).

Gene category techniques like GO annotation and KEGG enables profiling over-represented features in a given set of proteins (Bauer, 2017). Functional enrichment analyses of the proteins involved in the PPI network of *E. faecalis* and *E. faecium* are shown in Figure 6-12. Proteins in both networks took part in the similar biological pathway, molecular function, cellular component and KEGG pathway with similar counts.

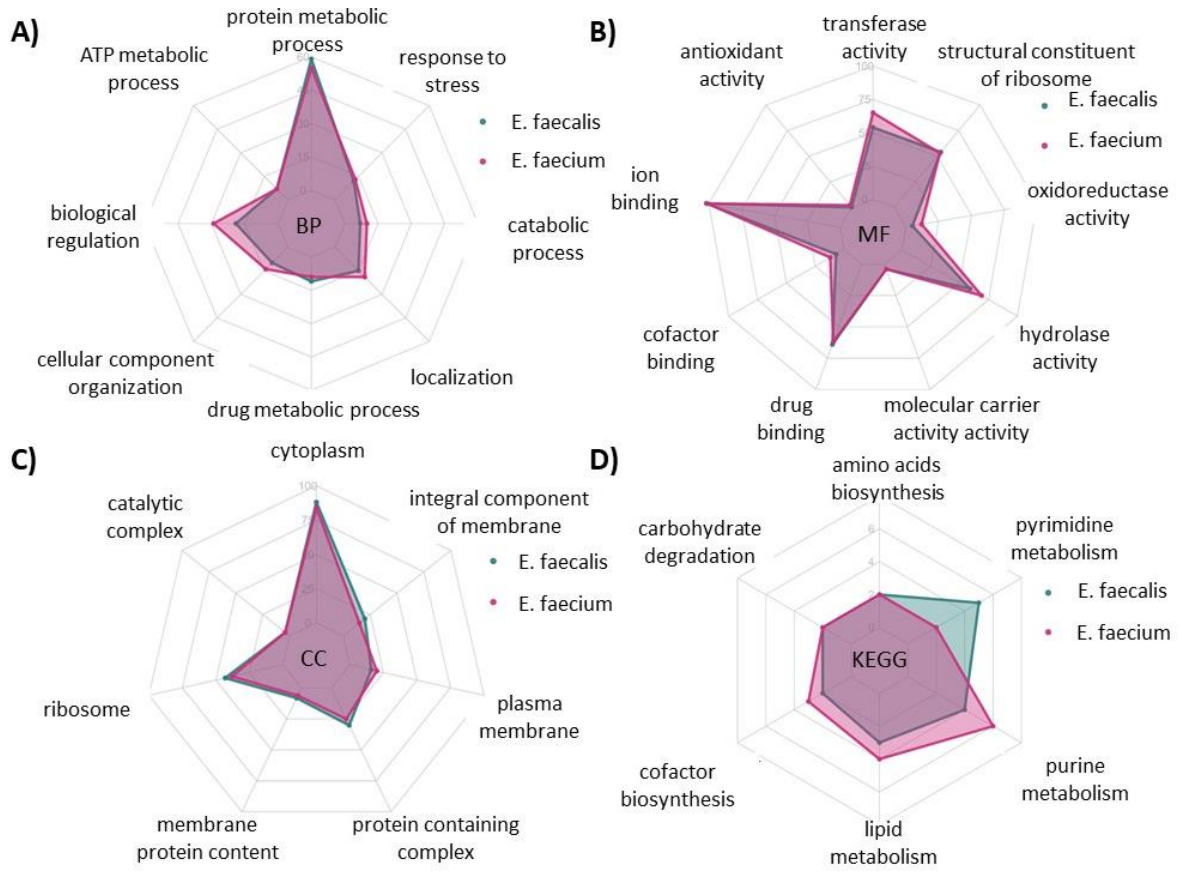


Figure 6-12. Functional enrichment analyses of the genes encoding the 295 Enterococcus faecalis proteins and 345 Enterococcus faecium proteins present in the PPIs. *In **A)** Biological pathway (BP), **B)** Molecular function (MF), **C)** Cellular component (CC) and **D)** KEGG pathway; the enriched categories and the number of the genes populating them are shown. The figure was generated using R package fmsb (Nakazawa and Nakazawa, 2019).*

Moreover, individual GO terms for not well-characterized discriminant proteins of *E. faecalis* were estimated by threading technique (see Methods section for details).

1) GO terms assigned for PTS family porter are: negative regulation of immune effector process (BP), cytokine receptor binding (MF), and extracellular region part (CC).
2) GO terms assigned for HP-1 (hypothetical protein with Mw of 4769.57Da): phosphorylation (BP), ATP binding (MF), metal ion binding (MF) and nucleoid (CC).
3) GO terms assigned for HP-2 (hypothetical protein with Mw of 6670.51Da): oxidation-reduction process (BP), oxidoreductase activity (MF), transition metal ion binding (MF), and integral component of membrane (CC).

Individual GO terms for not well-characterized discriminant proteins of *E. faecium* were also estimated by the threading technique (see Methods section for details).

1) GO terms assigned for HP-1 (hypothetical protein with Mw of 4487.34Da) are: translocation (BP), protein transporter activity (MF), zinc ion binding (MF), protein domain specific binding (MF) and cytosol (CC).
2) GO terms assigned for HP-2 (hypothetical protein with Mw of 5047.09Da) are: DNA replication (BP), ATP binding (MF) and integral component of membrane (CC).
3) GO terms assigned for CsbD domain-containing protein: cellular protein metabolic process (BP), protein binding (MF), DNA polymerase activity (MF), metal ion binding (MF) and cytoplasmic part (CC).

6.4 DISCUSSION

In this study, *E. faecalis* and *E. faecium* isolates did not show resistance to only one particular antibiotic, but rather, presented a multidrug resistance profile. However, generating multidrug-resistant vs susceptible classes was not possible as (1) there were limited *E. faecalis* or *E. faecium* isolates that showed resistance to certain antibiotics together, (2) *E. faecalis* isolates that were susceptible to all antibiotics were few in number; and (3) there was no *E. faecium* isolate that showed no resistance to all antibiotics screened in this study. Hence, the isolates were grouped as resistant and susceptible according to the sole antibiotic profile and then analysed. This is plausible because, in the field, one could not screen all antibiotics to define the antimicrobial profile of a pathogen. Instead, the several antibiotics used in their management scheme are screened to define whether a pathogen is resistant to that particular antimicrobial. This study could represent a real-life scenario that is generally seen in dairy farms.

Out of 12 different resistant vs susceptible antimicrobial profile comparison; LR and NB gave the best performance for 3 of the analyses each, of which *E. faecalis*: benzylpenicillin, erythromycin and TMP/SMX for LR and *E. faecalis* clindamycin; *E. faecium* cefovecin and erythromycin for NB. MLP NN gave the best performance for 2 analyses which were *E. faecalis* chloramphenicol and *E. faecium* erythromycin. Meanwhile, RBF SVM, LDA, Adaboost and LSVM and gave the best performance for 1 analysis each, of which were *E. faecalis*: tetracycline; *E. faecium* clindamycin, benzylpenicillin and nitrofurantoin, respectively. However, DT, RF and QDA were not found in any of the analysis as the best performance predictor. RF, as a classifier working in the principle of bagging, might not have had enough data points to be trained well enough (Genuer, Poggi and Tuleau, 2008). DT was also shown to give a limited performance with small datasets (Morgan *et al.*, 2003). QDA performance was surprising as, in general, QDA outperforms LDA as it can fit in the data better. However, in this study, LDA

outperformed QDA in ten of the twelve analyses, which was seen in other studies as well when small datasets were the case (Wu *et al.*, 1996; Decruyenaere *et al.*, 2015).

Classification of *E. faecalis* isolates based on antimicrobial profile gave satisfying results for benzylpenicillin, chloramphenicol, clindamycin, erythromycin, tetracycline and TMP/SMX, after 30 observations for each analysis. The confidence level between the observations in the experiments with around 100 samples is expected to be less than 10% (Varoquaux, 2018), where all prediction performance metrics of these analyses provided this criterion as well.

Classification of *E. faecium* isolates based on antimicrobial profile gave satisfying results for cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin, after 30 observations for each analysis. The prediction performance metrics of these analyses fell into the 10% confidence bound as well. Classification results of benzylpenicillin-resistant vs susceptible isolates of *E. faecium* were relatively lower (i.e. kappa of 48.06%) and more than 10% confidence bound between multiple observations. This may be a result of the labelling technique of the isolates as some benzylpenicillin susceptible isolates had MIC values very close to the breakpoints. In fact, when the extreme cases were compared (benzylpenicillin-resistant vs most susceptible), the results were significantly improved (data not shown). This was also observed in *S. aureus* when the degree of AMR increased, the similarity to spectra of wild-type strain was decreased (Muroi *et al.*, 2012).

The prediction performance of the learners could be improved with more peaks (features) by setting the relative intensity height filter at a lower threshold (for instance, the height filter was set to 10 in the previous chapter). Improved accuracy with the employment of more peaks was also observed in the study by Wang and colleagues (2020), where they continuously increased the number of peaks and the performance reached a plateau when the peak number was over a hundred. Application of a large number of peaks rather than a few can also offer solutions for reproducibility of MALDI-TOF spectra which has been a huge problem (Croxatto, Prod'hom and Greub, 2012). If some discriminant peaks are not produced for any reason, alternative peaks could still be used for classification (Wang *et al.*, 2020).

This study showed that differences in the antimicrobial profiles of *Enterococcus* species could be detected in MALDI-TOF spectral profiles by using ML. This seems reasonable as the antibiotic stress would drive bacteria to adjust their protein abundance and/or modification. While the primary aim was to develop ML-powered diagnostics discriminating resistant and susceptible isolates of *Enterococcus* species, we also characterized the molecular determinants and

mechanisms underlying the patterns. Although more peaks were employed to discriminate based on antimicrobial profiles, only those peaks with certain relative intensity were offered as biomarker candidates. Fifteen and twelve peaks were identified by the classifiers as the most discriminant between resistant and susceptible profiles of *E. faecalis* and *E. faecium*, respectively. These peaks were cross-matched with the proteins in the proteomes of these species, as understanding the roles of these proteins may help the discriminatory power of MALDI-TOF.

As the identified biomarkers of this study are not the product of specific AMR genes, it may be hard to explain the antimicrobial profiles of the isolates. However, other factors have also been shown to contribute to AMR or adaptation in response to selective pressure (Miller, Munita and Arias, 2014). Similar results to our study were found when the proteome of another mastitis-causing pathogen *S. xylosus* under the effect of tylosin was analysed (Liu *et al.*, 2019). The proteins which had increased expressions under tylosin included several ribosomal proteins and a translation initiation factor protein. In another study with MRSA, *rpmB* and *rplX* were amongst the differentially regulated genes under treatment of therapeutic agents; stigmasterol and luperol (Adnan, Ibrahim and Yaacob, 2017). Moreover, different peak intensities in MALDI profiles of resistant and susceptible isolates were observed, even where there was no prior antimicrobial treatment (Mather *et al.*, 2016). Having observed the similar expression pattern between the discriminate peaks and resistant genes may explain the biologic mechanism behind AMR, these biomarkers could be targeted for new drug development.

Clustering coefficient values are greater in real networks than randomly selected networks (Albert and Barabási, 2002), which can be used to explain the importance of discriminant proteins and their neighbours in the PPI networks of each organism. Our results indicated that the networks of interest both for *E. faecalis* and *E. faecium* had a remarkable clustering coefficient which was much higher than randomly built PPI networks in each proteome. This shows the modular organization where the nodes in the network of interest are internally more connected compared to the rest of the proteome (Ravasz *et al.*, 2002). Our results for the network of interest for *E. faecalis* showed that both the average number of neighbours and network density values were higher than the randomly built PPI network of *E. faecalis*. Although these results were also higher in the network of interest for *E. faecium*, they were much closer to the values of the randomly built PPI network. Heterogeneous networks are accepted to be robust against random decays but susceptible to targeted interventions (Shang, 2014). The network heterogeneity of the network of interest for *E. faecalis* was almost six times smaller than the value of the randomly built network of *E. faecalis*. However, network heterogeneity of the network of

interest for *E. faecium* was similar with the randomly built network of *E. faecium.* This means that the network of interest for *E. faecium* have some hubs which supply the connection between nodes and thus may be specifically targeted (Dong and Horvath, 2007). Overall, topological properties of networks of interest for *E. faecium* and *E. faecalis* indicate that the same approach may fail to treat both infections.

PPI analysis showed that 12 out of 15 discriminant proteins of *E. faecalis* interact with each other either directly or indirectly (by sharing the same first neighbour proteins). It is no surprise that ribosomal proteins interact with each other but DNA-binding HU protein, InfA and LuxR family protein were also in the same network. The other three discriminant proteins were less characterized (i.e. hypothetical proteins) and this may be the reason why they were not found to have interaction with the rest, above medium confidence level. Moreover, TetM was the first neighbour of the discriminant ribosomal proteins. This may suggest that the expression of vital proteins participating in discrimination of resistant and susceptible profiles of *E. faecalis* are co-ordinated and may be targeted by new drug therapies (Petta *et al.*, 2016; Carro, 2018).

In the PPI network of *E. faecium*, all the discriminant proteins except CsbD were shown to be interacting with each other directly or indirectly (sharing first neighbours). Again, findings of direct interaction between ribosomal proteins were not surprising but other discriminant nodes SHK, HU and HP were also involved in the same network. Moreover, AMR regulatory proteins VanRB and VanSB, which activate transcription of vancomycin resistance genes *vanA*, *vanH* and *vanX* (Arthur *et al*., 1997), were found to be the first neighbour of SHK, TetM to be the first neighbour of discriminant ribosomal proteins, GyrA and GyrB to be the first neighbour of RpsR, and ErmB to be the first neighbour of RpsT and RpsR. This may strongly suggest that the expression of vital proteins participating in discrimination of resistant and susceptible profiles of a species (*E. faecium* in this case) are co-ordinated.

Functional enrichment analyses of both *E. faecalis* and *E. faecium* PPI networks resulted in the same GO functions and KEGG pathways in similar counts. There were important functions such as response to stress, drug metabolic process, antioxidant activity, drug binding etc., which may explain the biological mechanism behind the potential biomarkers.

The primary target of most synthetic and natural antimicrobials is bacterial ribosomal subunits (Schlunzen *et al.*, 2001). In this study 13 out of 25 discriminant proteins were part of ribosomal units and direct interactions of antimicrobial-resistant proteins were shown in the PPI network. This may strongly suggest that the discriminatory power of several ML algorithms could

correctly select the discriminatory peaks from MALDI profiles. The proteomes analysed in this study were limited to 12kDa due to the working range of the MALDI-TOF equipment, which prevented the observation of higher mass proteins.

AMR of *Enterococcus* is highly associated with antimicrobial usage (Beukers *et al.*, 2017; Klibi *et al.*, 2015). There are various categorization systems of antimicrobials according to several organizations such as WHO (World Health Organisation), AMEG (Antimicrobial Advice *ad hoc* Expert Group) and OIE (World Organisation for Animal Health). Although category naming/numbering vary between these classification systems, the antimicrobials which are recommended to avoid and reduce veterinary use fall into the same priority levels (EMA/AMEG, 2019). Antimicrobials tested in this study fall into the following categories: cefovecin and enrofloxacin for restricted use; chloramphenicol, erythromycin and clindamycin for cautious use; nitrofurantoin, benzylpenicillin, trimethoprim/sulfamethoxazole and tetracycline for prudent use (EMA/AMEG, 2019).

Chloramphenicol inhibits protein biosynthesis by binding to the peptidyl transferase centre of the large ribosomal unit (Schlunzen *et al.*, 2001). The resistance to chloramphenicol is mostly a result of the activity of the chloramphenicol acetyl-transferase gene, *cat* (Schwarz *et al.*, 2004), which was previously found on a transmissible plasmid in *E. faecalis* RE25 (Schwarz, Perreten and Teuber, 2001). However, there are other resistance mechanisms including efflux systems, interruptions by phosphotransferases, mutations of the ribosomal proteins or permeability barrier success (Abushaheen *et al.*, 2020). The model strain we used in this study for the *E. faecalis* proteome did not have the *cat* gene; thus, we could not show the interaction of discriminant proteins with it.

Erythromycin belongs to the macrolide class of antibiotics which inhibits bacterial protein synthesis by blocking peptidyl transferase just like chloramphenicol (Vázquez-Laslop and Mankin, 2018). Although macrolide resistance occurs due to the contributions of many genes, such as macrolide and streptogramin B resistant (*msr*), macrolide efflux (*mef*) and virginiamycin factor A (*vga*), erythromycin ribosome methylase B (*ermB*) is the most common gene observed in resistant isolates (Roberts *et al.*, 1999). Amongst erythromycin resistance genes, *ermB* is the most commonly found in *Enterococcus* spp. isolated from animal (Petersen and Dalsgaard, 2003), food (Gazzola *et al.*, 2012; Rizzotti *et al.*, 2005) and environmental sources (Cho *et al.*, 2020a). In our study, we did not observe ErmB interacting with the discriminant *E. faecalis* proteins, but we did observe an interaction in the PPI network of *E. faecium*.

Tetracycline inhibits protein synthesis by binding the small subunit of bacterial ribosomes (Chukwudi, 2016). Tetracycline resistance occurs as a result of three different mechanisms: ribosomal protection mechanism coding genes (e.g. *tetM*), efflux pump functioning genes (e.g. *tetL*) and enzymatic inactivation genes (e.g. *tetX*) (Van Hoek *et al.*, 2011). Amongst tetracycline resistance genes, *tetM* is the most commonly found in *Enterococcus* spp. isolated from human (Aarestrup *et al.*, 2000), animal (Jackson *et al.*, 2010; Cauwerts *et al.*, 2007), food (Huys *et al.*, 2004; Frazzon *et al.*, 2010) and environmental sources (Cho *et al.*, 2020a; Sadowy and Luczkiewicz, 2014). It is thought to be a highly important gene for spreading tetracycline resistance amongst *Enterococcus* (Rizzotti *et al.*, 2009). In this study, TetM was found to be interacting directly with discriminant ribosomal proteins in both *E. faecalis* and *E. faecium* PPI networks.

Benzylpenicillin inhibits the synthesis of peptidoglycan (the main component of the bacterial cell wall) by targeting penicillin-binding proteins (PBPs) (Sauvage *et al.*, 2002). Penicillin is one of the most active beta-lactams to combat *Enterococcus* infections (Miller, Munita and Arias, 2014). Bacterial species in the *Enterococcus* genus contain several PBPs; in particular, 6 genes (*pbpA*, *pbpB*, *pbpF*, *pbpZ*, *pbp5* and *ponA*) are responsible for PBPs synthesis in *E. faecalis* and *E. faecium* (Duez *et al.*, 2004). Particularly, owing to *pbp5* gene, *Enterococcus* species have intrinsically low-resistance to beta-lactam antibiotics (Sifaoui *et al.*, 2001; Arbeloa *et al.*, 2004), where penicillin resistance of *E. faecium* is higher than *E. faecalis* (Garrido, Gálvez and Pulido, 2014). In our study, we could not find the direct interaction of our discriminant proteins with these genes in *E. faecalis* and *E. faecium*, but PbpA and PonA were found to be interacting with some first neighbour proteins in both *E. faecalis* and *E. faecium* PPI networks (data not shown).

Cefovecin belongs to the cephalosporin antibiotic class but is relatively new (third generation). Cefovecin is already known to inhibit the synthesis of the cell wall by targeting PBPs just like other cephalosporin family antibiotics. Intrinsic tolerance in *Enterococcus* for cephalosporin is thought to be associated with penicillin-binding proteins especially *pbp5* gene (Rice *et al.*, 2009; Arbeloa *et al.*, 2004). Bacterial two-component regulatory systems (TCS) such as histidine kinase, response regulator (CroS/R), transmembrane kinase and cognate phosphatase (IreK/P) were also shown to be related to cephalosporin resistance of *Enterococcus* (Miller, Munita and Arias, 2014; Kellogg and Kristich, 2018).

Enrofloxacin is a synthetic antibiotic that belongs to the fluoroquinolone antibiotic family. Fluoroquinolones work by targeting two bacterial enzymes, DNA gyrase and topoisomerase

IV, which play important roles during DNA replication (Hooper and Jacoby, 2016). Resistance to fluoroquinolones was found to be correlated with the mutations of four genes - DNA gyrase A subunit (*gyrA*) and B subunit (*gyrB*), and topoisomerase IV A subunit (*parC*) and B subunit (*parE*) - in (fluoro)quinolone-resistant *E. faecium* and *E. faecalis* isolates (Oyamada *et al.*, 2006). In our study, GyrA and GyrB were found to be the first neighbours of the discriminant ribosomal protein RpsR between resistant and susceptible profiles of *E. faecium* while ParC and ParE were found to be interacting with some first neighbours of discriminant proteins in the PPI network (data not shown).

Nitrofurantoin inhibits bacteria by damaging bacterial DNA, RNA and cell wall protein synthesis (Waller and Sampson, 2017). Nitrofurantoin susceptibility was shown to be correlated with the presence of nitrofuran reductases. When these reductases transform nitrofuran into electrophilic intermediates, bacterial ribosomal proteins are targeted, resulting in inhibition of protein synthesis. Deletion of two nitrofuran reductase genes (*nfsA* and *nfsB*) in *E. coli* increased the resistance profile compared to the wild type (Sandegren *et al.*, 2008). Resistance without the loss of nitrofuran reductases was also reported which was the result of plasmid-mediated *oqxAB* nitrofurantoin (Ho *et al.*, 2015). In this study, nitrofuran reductases were not found to be interacting directly with our discriminant proteins between resistant and susceptible antimicrobial profiles.

Clindamycin is a semisynthetic antibiotic acquired by chlorination of lincomycin. It belongs to the lincosamide antimicrobial family and has a broad spectrum against gram-positive and anaerobes. Clindamycin targets the large units in bacterial ribosomes and inhibits peptidyl transferase activity (Dhawan and Thadepalli, 1982). Clindamycin resistance occurs mainly as a result of the altered ribosome due to adenine demethylation in 23S rRNA which can also result in multi-drug resistance. In addition to this common resistance mechanism, lincosamide inactivation nucleotidylation genes *linA* and *linB* were found related to clindamycin resistance in *E. faecium* (Bozdogan *et al.*, 1999). *E. faecalis* was also considered to be resistant to clindamycin owing to an intrinsic gene of lincosamides and streptogramins A (*lsa*) (Dina, Malbruny and Leclercq, 2003), rRNA adenine N-6-methyltransferases *ermA* and *ermB* (Malhotra-Kumar *et al.*, 2009). In this study, ErmB was found to be interacting directly with our discriminant proteins between resistant and susceptible antimicrobial profiles.

Trimethoprim (TMP), a pyrimidine class antibiotic, and sulfamethoxazole (SMX), a sulphonamide class antibiotic, both target bacterial folic acid synthesis but at different stages; SMX prevents the transformation of dihydropteroate from para-aminobenzoic acid by inhibiting

dihydropteroate synthetase whereas TMP prevents the transformation of tetrahydrofolate from dihydrofolate by inhibiting dihydrofolate reductase. These steps happen sequentially and therefore the combination of TMP/SMX was offered to provide a synergetic effect (Minato *et al.*, 2018). Although *in vitro* susceptibility tests can show the susceptible profile of *Enterococcus* to trimethoprim-sulfamethoxazole, like shown previously (Zervos and Schaberg, 1985) or in our study, they cannot be effective against *Enterococcus* infections *in vivo* as *Enterococcus* can bypass the synergistically working mechanism of these antibiotics by absorbing the folic acid from the environment (Miller, Munita and Arias, 2014). Trimethoprim resistance can be acquired by horizontal transfer of dihydrofolate reductase (*dfr*) genes (Bergmann *et al.*, 2014). In *E. faecalis* studies, dihydrofolate reductase genes (*dfrE* and *dfrF*) were shown to be providing intrinsic resistance to TMP/SMX (Coque *et al.*, 1999). In this study, dihydrofolate reductases were not found to be interacting directly with our discriminant proteins between resistant and susceptible antimicrobial profiles. However, DfrE was found to be interacting with the first neighbours of discriminant proteins in the *E. faecium* PPI network (data not shown).

In conclusion, discrimination of antimicrobial profiles of *Enterococcus* species based on MALDI-TOF profiles combined with various ML algorithms gave satisfactory results. Some of the discriminant proteins between resistant and susceptible isolates were related to AMR, stress response and drug binding activities as an individual protein or PPI network. As future work, a wider range of proteomes could be employed to improve the prediction performance of ML algorithms.

CHAPTER 7 DISCUSSION

This research used MALDI-TOF coupled with ML to discriminate mastitis pathogens at subspecies level in dairy herds of England and Wales. The purpose of the research was to provide an alternative diagnostic tool to current techniques (see section 1.3 for details) used for mastitis pathogens identification. Esener *et al.,* (2018) was the first paper to perform MALDI-TOF MS coupled with ML algorithms to discriminate bovine mastitis pathogens. In the whole thesis, several mastitis agents were tested both gram-positive (*S. uberis, S. aureus, E. faecalis* and *E. faecium*) and gram-negative (*E. coli*).

In Chapter 3, the only commercially available software that combines MALDI-TOF MS analysis with ML was used. However, the data manipulation for the pre-and post-analysis was limited to pre-defined settings. Moreover, only three supervised machine learning classifiers were available which offered limited room for exploring the biological mechanism underlying the data. In Chapters 4, 5 and 6, in-house scripts written in MATLAB and Python were used for data preparation and machine learning analyses, respectively.

Typeability of MALDI-TOF MS coupled with ML was tested for transmission of *S. uberis* in Chapter 3, and for disease phenotype of *E. coli* in Chapter 4. Antimicrobial susceptibility testing ability of MALDI-TOF MS coupled with ML was tested for *S. aureus* in Chapter 5, and for *Enterococcus* spp. in Chapter 6.

Transmission behaviour (contagious and environmental) of *S. uberis* has been shown to vary at the subspecies level, contrary to early studies that suggested only an environmental route (Davies *et al.*, 2016). Chapter 3 of this study investigated the application of MALDI-TOF MS coupled with ML algorithms in commercial software ClinProTools to differentiate transmission route (contagious and environmental) of *S. uberis* isolates at both the herd (intra-farm) and country-level (inter-farm).

This chapter has shown that discriminatory power based on MALDI profiles is higher in intra-farm analysis, especially when GA is used, which can be potentially developed in screening solutions at the herd level. Inter-farm classification appeared to be much weaker; hence, further improvements are needed to use it in screening solutions applicable at the population level. The explanation for that may be the divergent evolution of *S. uberis* isolates between the dairy herds in the UK, as there are no significant differences between contact and management control (i.e. bedding material, antimicrobial therapy etc). Previous studies have found divergent evolution

in other organisms such as *Campylobacter* spp., *S. aureus* and *H. pylori* (Wittwer *et al.*, 2005; Sordelli *et al.*, 2000; Falush *et al.*, 2003). This may be the reason why the same discriminant peaks combination does not work equally well for every study farm.

The investigation of MALDI profiles of clinical mastitis isolates has shown that prompt diagnosis of the transmission route at the early stages of an outbreak is possible. Considering that potentially contagious transmission of *S. uberis* has been found in two-thirds of commercial herds and proved to be the dominant transmission route in a third of UK herds (Davies *et al.*, 2016), the development of such a diagnostic tool is necessary. MALDI-TOF based methods would allow clinicians to identify the most appropriate control measures rapidly during an outbreak of disease. Potentially, it could reduce the incidence of clinical disease, reduce associated production losses and the costs associated with the treatment; and improve the efficiency of labour and resource allocation on the farm.

Chapter 4 of this study investigated the genotypic and phenotypic characteristics of bovine mastitis-causing *E. coli* strains. To this end, 20 *E. coli* strains were sequenced which were isolated from 10 cows (same cow same quarter) that had different clinical outcomes: subclinical and clinical mastitis. MALDI-TOF MS coupled with several ML algorithms – LR, LSVM, RBF SVM, MLP NN, DT, RF, AdaBoost, NB, LDA and QDA– were then performed to discriminate proteomic characteristics of this mastitis agent based on clinical outcome (clinical and subclinical) and disease phenotype (persistent and non-persistent).

It was demonstrated that there was no genotypic pattern amongst *E. coli* strains to cause different phenotypic outcomes of persistency or clinical severity. Biological changes in the mammary environment (i.e. host immune response) forced the pathogen to adapt its protein abundance. This finding accords with a recent study which showed that host response is needed for mastitis pathogen *S. uberis* to infect the mammary cells and cause mastitis (Archer *et al.*, 2020). By using machine learning, some of the biomarkers in a limited range could be shown, which may inspire further studies to design diagnostic tools or antimicrobial agents for bovine mastitis-causing *E. coli*.

As a limitation of this study, the sample size was too small to draw broad conclusions. However, this resulted from stringent data selection criteria and generating more data (i.e. same cow, same quarter, different bacterial phenotype) was not possible. It was at least shown that there was a proteomic difference in the mastitis pathogens even though their genomics is almost completely identical.

Future research should be undertaken to investigate whether similar solutions based on the analysis of MALDI-TOF MS coupled with ML may be used to develop screening tools to identify early signs of mastitis or related risk factors. The analysis of MALDI-TOF peaks has demonstrated good prediction performance for binary classification of bovine mastitis-causing *E. coli* with different phenotypic characters. There is abundant room for further progress in determining the proteome of this pathogen on a larger scale (not limited to MALDI-TOF MS range) but considering biomarkers of this study.

Antimicrobial-resistant *S. aureus* infections are a major concern in human and veterinary medicine, where dairy cattle are an important risk factor for zoonotic transfer. Fast, affordable and effective diagnostic solutions able to detect the specific *S. aureus* strains and their antimicrobial-resistant and susceptibility profiles are key to effective and targeted treatment selection. Chapter 5 investigated the application of MALDI-TOF MS coupled with ML algorithms to discriminate *S. aureus* strains that are resistant or susceptible to multidrug and benzylpenicillin. The task was approached in a principled way by applying optimization techniques to overcome uncertainty in data features and by using a wide repertoire of classification methods. In general, most of the machine learning approaches tested achieved high performance and had kappa values over 85.00%.

Previous studies have shown biomarkers between MRSA and MSSA isolated from humans by using MALDI-TOF either with ML (Sogawa *et al.*, 2017; Tang *et al.*, 2019) or without ML (Du *et al.*, 2002; Edwards-Jones *et al.*, 2000). Biomarkers found in these studies did not match with our discriminant peaks. This result may be explained by the fact that there was only one MRSA strain in the current study. In a recent study, van Oosten and Klein (2020) performed MALDI-TOF coupled with ML to identify proteomic signatures of *S. aureus* when treated with different antimicrobials such as benzylpenicillin, erythromycin, gentamicin, neomycin, tetracycline, trimethoprim etc. Peaks at *m/z* 4306.7Da (4305.59Da in our study), 4812.5Da (4807.21Da in our study), 6889.1Da (6891.1Da in our study) and 9627.6Da (9621.26Da in our study) were found common to be significant features for the antimicrobial screening of *S. aureus*. Wang and colleagues (2018a) found the following peaks at *m/z* 4305Da (4305.59Da in our study), 4813Da (4807.21Da in our study), 6422Da (6423.37Da in our study), 6887Da (6891.1Da in our study) and 9625Da (9621.26Da in our study) amongst the relevant features to distinguish VSSA from VISA and hVISA isolates. Another study (Mather *et al.*, 2016) which performed MALDI-TOF coupled with ML showed that peaks at *m/z* 4815Da (4807.21Da in

our study), 6425Da (6423.37Da in our study) and 9626Da (9621.26Da in our study) were amongst the features to differentiate VISA, hVISA and VSSA.

The most important limitation lies in the fact that the working range of 2-12kDa prevents inspection of the complete *S. aureus* proteome regarding its specific antimicrobial profile. Also, it is important to acknowledge that the data were collected from only English and Welsh farms. However, this should not limit the potential of the current method to determine antimicrobial profiles in other farms across the globe, which could be easily achieved when diversely distributed data (i.e. to weaken geographical bias) is supplied to train the ML algorithms. Lastly, multidrug-resistant isolates were all resistant to benzylpenicillin. Hence, there is a bias towards peaks determining resistance or susceptibility to benzylpenicillin, which could explain why all four multidrug discriminant peaks occurred within the set of benzylpenicillin-only discriminant peaks.

Having shown the good performance of MALDI-TOF MS coupled with ML on differentiating the antimicrobial profile of *S. aureus* (Chapter 5), it was aimed to apply the same technique to other bovine mastitis pathogens such as *Enterococcus* spp. Chapter 6 investigated to provide an alternative to standard susceptibility tests which would profile benzylpenicillin, chloramphenicol, clindamycin, erythromycin, tetracycline and TMP/SMX resistance in *E. faecalis*; and benzylpenicillin, cefovecin, clindamycin, enrofloxacin, erythromycin and nitrofurantoin resistance in *E. faecium* isolates. Like in Chapter 5, this chapter approached the task in a principled way by applying optimization techniques to overcome uncertainty in data features and by using a wide repertoire of classification methods. Out of 12 different resistant vs susceptible antimicrobial profile comparisons, LR and NB gave the best performance for 3 of the current analyses each: for LR *E. faecalis* benzylpenicillin, erythromycin and TMP/SMX; for NB, *E. faecalis* clindamycin, *E. faecium* cefovecin and erythromycin. MLP NN gave the best performance for 2 analyses: *E. faecalis* chloramphenicol and *E. faecium* erythromycin. Meanwhile, RBF SVM, LDA, AdaBoost and LSVM gave the best performance for 1 analysis each: *E. faecalis* tetracycline, *E. faecium* clindamycin, benzylpenicillin and nitrofurantoin, respectively.

In a previous study with larger sample size, Wang and colleagues (2020) compared vancomycin-resistant and susceptible *E. faecium* isolates that were sourced from human patients. Peaks at *m/z* 5038Da (5036.95Da in our study), 6512Da (6507.02Da in our study), 7275Da (7273.74Da in our study), 7330Da (7324.94Da in our study), 10075Da (10068.06Da in our study) and 10941Da (10934.36Da in our study) were found common to be significant features with our study. New broad-spectrum antimicrobials can be designed targeting these

biomarkers. Another study found *m/z* 6036.59Da and 4526.45Da to be statistically significant peaks between tetracycline-resistant and susceptible *Enterococcus* spp., which were different from the current study (Sabença *et al.*, 2020). This discrepancy could be attributed to the use of MALDI profiles at the genus level isolated from food sources in the other study. Exact *m/z* matches of the biomarkers between MALDI-TOF experiments are not common since MALDI-TOF instruments, calibration settings, wearing out of laser and detector, sample preparation and spectra pre-processing vary hugely in the studies (Sauget *et al.*, 2017).

It was demonstrated that the combination of supervised ML and MALDI-TOF MS can also be used to develop an effective computational diagnostic solution that can discriminate between resistant and susceptible profiles of *E. faecalis* and *E. faecium* for more antimicrobial classes besides benzylpenicillin (as shown for *S. aureus* in Chapter 5).

To compare the results of the analyses between chapters, kappa was used as it measures the performance considering both positive and negative classes. Sensitivity and specificity alone could not be used for comparison of different analyses, as the former provides results which consider only the positive class, whereas the latter provides results which consider the only negative class. Moreover, the use of accuracy for comparison may be confusing as it may be higher for imbalanced datasets although this would not mean healthy prediction of each class – accuracy will be high as the classifier will mainly predict the datasets of the frequent class correctly (Müller and Guido, 2016). It should be noted that AUC also considers both classes, but it was not used as ClinProTools does not calculate this value.

Comparison of the findings between chapters shows that the kappa values of biotyping (from Chapter 3 and 4) were relatively lower than values of antimicrobial susceptibility analysis (in Chapter 5 and 6). There are possible explanations for these findings. A possible explanation for the low kappa value of Chapter 3 might be that the model of inter-farm analysis was validated by external data coming from different herds which did not contribute any spectra to train the model. This is plausible as the intra-farm analyses had remarkable prediction performance (kappa of 92.62%). Another possible explanation for the low performance in Chapter 3 may be due to limited parameter settings as hyperparameter tuning was restricted within the options defined by commercial software. While working with MALDI-TOF MS technology, it is important to generate more than one spectrum per sample as a low-quality spectrum may be missing certain *m/z* values due to noise (López Fernández *et al.*, 2016). Similarity comparison between technical replicates is important as differences have been observed between technical replicates due to the performance of the instrument (Oberle *et al.*, 2016). One possible

explanation for the relatively low performance of the analysis in Chapter 4 might be that CCI analysis was not performed to measure similarities between technical spectra, unlike in other chapters. Quality control of the spectra would have increased analysis success. The reason why CCI was not performed for the analysis in Chapter 4 was not to lose more spectra from the small-sized dataset. Another possible explanation may be the small size of the sample as it shrank through stringent data selection criteria. Moreover, in the analysis of clinical status, the proteomes of genotypically identical *E. coli* strains were compared. Hence, the low prediction performance is plausible as there is relatively limited room for differentiation.

Binary classification models are generally not designed for skewed data and perform better in balanced classes (Zhang, 2010; He and Garcia, 2009). In binary classification, if the datasets in one class are more frequent than in the other one, they are called imbalanced classes (Müller and Guido, 2016). This causes a bias towards the class that has more datasets (majority class) (Prati, Batista and Monard, 2009). Therefore, additional steps were needed to cope with the bias due to the nature of imbalanced datasets. As previously stated, the manipulation of the data was limited within the commercial software ClinProTools, so no additional pre-processing to balance the datasets in positive and negative classes could be performed in Chapter 3. The datasets in positive and negative classes were fairly balanced in Chapter 4. However, the datasets in positive and negative classes were imbalanced for both of the analyses in Chapters 5 and 6.

In the current work, the undersampling approach was used to balance benzylpenicillin/multi-drug-resistant and susceptible classes of *S. aureus* in Chapter 5; and the oversampling approach was used to balance single antimicrobial-resistant and susceptible classes of *E. faecalis* and *E. faecium* in Chapter 6. Oversampling and undersampling techniques provide improved prediction performance by balancing the classes. They do not have any obvious superiority over each other and can be decided based on the problem (Xia *et al.*, 2019). However, when undersampling is used, it ignores some data points in the majority class which may be important for analysis, and therefore this may reduce the prediction performance (Buda, Maki and Mazurowski, 2018). On the other hand, oversampling may be computationally expensive as it increases the training datasets for the minority class (Le *et al.*, 2019). Furthermore repeated data points of the minority class may result in overfitting (Azadbakht, Fraser and Khoshelham, 2018). In this thesis, the approaches performed to cope with imbalanced datasets were basic. There are also more advanced techniques (i.e. oversampling by ADASYN and SMOTE or cleaning undersampling using methods such as Tomek's links, edited nearest neighbours,

condensed nearest neighbours, instance hardness threshold etc.). However, these techniques generate new datasets instead of repeating the existing samples for oversampling or deleting the existing data points for undersampling. These techniques were not selected in the current study as this might have disturbed the biological mechanism behind the phenotypic classes. Further researchers, who do not focus on the biological relationship between the classes, can employ one of these approaches in their work.

The biomarker characterisation was performed through cross-matching of *m/z* and molecular weight of protein in databases such as UniProt, NCBI and PATRIC. Peak-protein cross-matching by comparing observed and theoretical mass has been widely used in proteomic studies (Borgaonkar *et al.*, 2010; Shin *et al.*, 2008; Pusztai *et al.*, 2004). This kind of approach is tentative as only molecular weights of unmodified proteins are considered in these databases, although an assortment of posttranslational modifications such as methylation, acetylation, carboxylation, and protease cleavage may develop in the cell (Meister, 2009). To cope with these issues, the theoretical N-terminal methionine cleavage was considered in the analyses of Chapters 4, 5 and 6 like previous studies (Arnold *et al.*, 1999). The molecular weight of a protein is not always the same as the addition and removal of molecules or isotopes of elements vary the mass by 1 to 22Da; moreover, phosphorylation of a protein may cause two peaks separated from each other by 80Da (Coombes, Baggerly and Morris, 2007). The peaks become broader as the *m/z* values increase; hence, peak-protein cross-match should be performed by a formula considering the individual peak mass rather than an identical value for all peak-protein cross-matching (Coombes, Baggerly and Morris, 2007). To cope with these issues, the mass range of 0.2% has been considered as the threshold for peak-protein cross-matching in the analyses of Chapters 3, 4, 5 and 6. The value of 0.2% was previously used as a threshold for peak-protein cross-matching, although other studies employed a wider range (0.5%) instead (Borgaonkar *et al.*, 2010; Shin *et al.*, 2008). Since the current study was limited to computational approaches, it was not possible to test biomarkers by wet-lab experiments such as knock-out and knock-down. Further studies regarding the role of biomarkers, especially less known ones (i.e hypothetical proteins), would be worthwhile.

It was not possible to assess bovine mastitis agents directly from the milk samples; therefore, they were first cultured and then MALDI profiled. Although previous studies aimed to analyse bovine mastitis pathogens including *E. coli*, *S. aureus*, *S. uberis* and *E. faecalis* directly from the milk, a certain amount of CFU was needed which was significantly above the clinical mastitis diagnosis threshold (Barreiro *et al.*, 2012; Barreiro *et al.*, 2017). MALDI-TOF coupled

with ML analysis takes only 1 day which is needed for bacterial growth. Sample preparation for MALDI-TOF takes around 10 min/sample and analysis itself takes 2 min/sample (Barreiro *et al.*, 2010). ML analysis depends on the algorithm, for which computation times were generally short (less than 2 min) on a machine with Intel(R) Core (TM) i7-8665U CPU @ 1.90GHz; RAM: 16Gb; Windows 10. GA (on average 2.5h) and MLP NN (on average 4h) were exceptions. These values were also reported in a similar study (Maciel-Guerra *et al.*, 2021). On the other hand, traditional biochemical tests used for typing takes between 3 to 6 days after single colony growth which needs an additional day (Barreiro *et al.*, 2010). MALDI-TOF coupled with ML analysis also provides a 50% time reduction compared to conventional antimicrobial susceptibility tests (Wang *et al.*, 2020).

MALDI-TOF MS instrument is expensive as a capital investment; however, once installed the application per identification is much more economic than conventional microbiology tests (Gaillot *et al.*, 2011). The cost of MALDI-TOF MS analysis was shown to be almost 20 times more economic than conventional phenotypic techniques per sample (Cherkaoui *et al.*, 2010). Moreover, it was shown to decrease reagent waste, unnecessary antibiotic use, labour costs and the need for secondary verification such as sequencing (Gaillot *et al.*, 2011; Tan *et al.*, 2012; Nagel *et al.*, 2014).

Since the mass range of the MALDI approach could not exceed 12kDa, it was not possible to observe differences between the heavy proteins in the proteomes of the analysed classes, which could provide other biomarkers. A potential alternative approach to increase the mass range beyond current values may be the application of high-intensity focused ultrasound-assisted proteomic analysis (Gekenidis *et al.*, 2014). The proteins with higher molecular weights could be detected by this technique. In another study, the upper limit of the detection was increased up to 75kDa for both gram-positive and negative bacteria by using unconventional chemicals for pre-treatment (Madonna *et al.*, 2000).

Notwithstanding the relatively limited proteins of the proteome, MALDI-TOF offers valuable insights into discriminating different phenotypes based on whole-cell and cell extract measurements which are most abundant with ribosomal proteins (Ryzhov and Fenselau, 2001). It is estimated that 20% of the proteins in the cell are ribosomal proteins and the conventional working mass range of MALDI-TOF consists of mainly ribosomal and a few other housekeeping proteins (Stump *et al.*, 2003; Murray, 2012). Additionally, previous studies suggested that ribosomal proteins are stable biomarkers as they were not affected in terms of peak mass under several growth conditions (Arnold and Reilly, 1999; Wunschel *et al.*, 2005a).

Additionally, the abundance of the ribosomal proteins has been shown to vary according to the phenotype of bacteria (Arnold *et al.*, 1999; Sousa *et al.*, 2020).

Interlaboratory experiments concluded that when appropriate controls of the conditions were provided, spectral reproducibility could be ensured (Valentine *et al.*, 2005; Wunschel *et al.*, 2005a; Wunschel *et al.*, 2005b). However, some biomarkers other than ribosomal proteins could be affected in terms of presence/absence by sample and matrix preparation, protein extraction techniques or experimental conditions (Wang *et al.*, 1998). As the intensity of proteins is also considered to define biomarkers, which may be influenced by minor experimental variations, it is recommended that all samples should be treated in the same laboratory; if this is not possible extreme caution should be shown to minimize bias between laboratories (Wang *et al.*, 1998).

Moreover, appropriate pre-processing steps should be performed after acquiring the raw spectra from a MALDI-TOF instrument. Previously, researchers used to think that if there was a strong biomarker between groups, it would be detected no matter how the pre-processing was performed (Borgaonkar *et al.*, 2010; Wagner, Naik and Pothen, 2003). However, pre-processing is essential to draw biological conclusions from the raw data. Even small differences between MALDI-profiles may alter the performance of the classification; therefore, all classes should be prepared in the same pre-processing pipeline (Chung *et al.*, 2019).

A typical MALDI spectrum contains hundreds of peaks (Coombes, Baggerly and Morris, 2007); however, not all of them have relevant biological information. Visual inspection of these many peaks is not possible; therefore, ML techniques are needed (Morris *et al.*, 2005; Mather *et al.*, 2016; Wang *et al.*, 2018b). In the experiments of this study, the MALDI-TOF coupled with ML approach identified a total of 46 biomarker proteins which were as follows: 8 biomarkers from the differentiation analysis of *S. uberis* based on transmission route (see Chapter 3), 6 biomarkers from the differentiation analysis of *E. coli* based on disease phenotype and 2 biomarkers from the differentiation analysis of *E. coli* based on clinical outcome (see Chapter 4), 5 biomarkers from the antimicrobial susceptibility profiling of *S. aureus* (see Chapter 5), 15 biomarkers from the antimicrobial susceptibility profiling of *E. faecalis* and 10 biomarkers from the antimicrobial susceptibility profiling of *E. faecium* (see Chapter 6). Out of the 46 biomarkers, 19 ribosomal proteins, 3 DNA-binding proteins, 2 bacteriocins, 2 phosphotransferase system proteins, 1 ATP synthase protein, 1 bacterial response regulator (LuxR family protein), 1 translation initiation factor, 1 sensor histidine kinase, 1 stress response protein (CsbD domain containing) and 1 DNA gyrase inhibitor (YacG) were found as known

functional domains; whereas 12 hypothetical and 2 DUF (domain of unknown function) proteins were found as unknown functional domains. Detailed information about the functions of these biomarkers was given in the Results sections of the relevant chapters.

These biomarkers accord with earlier observations by MALDI-TOF MS, which showed mainly ribosomal proteins, DNA-binding proteins, stress-associated proteins (i.e. CsbD family proteins) and hypothetical proteins (Ojima-Kato *et al.*, 2017; van Oosten and Klein, 2020). Cheng and colleagues (2018) compared the spectra coming from 14 genera (81 species, 403 strains) by using GA and found ribosomal and DNA-binding proteins as the ten most discriminant ones for bacterial identification. However, non-ribosomal proteins were also found to show different expression according to different situations (Cheng, Qiao and Horvatovich, 2018).

Notwithstanding the mass range limitations of the analyses, this study offered biomarkers to differentiate certain phenotypes. Biomarker characterisation should be supported by the functions and pathways they are involved in, as the interactions or differential expressions of other proteins can help to understand the biological mechanism of the disease (Swan *et al.*, 2013). The products of structural analyses, functional enrichment analyses and PPI analyses of biomarkers are not repeated here and can be found in the Results sections of each chapter. In Chapters 5 and 6 – where respectively the classifications for antimicrobial susceptibility profiles of *S. aureus* and *Enterococcus* spp. were performed - known resistance proteins were not amongst the biomarkers found by the classifiers. As previously stated this was not expected as the molecular weights of these proteins were out of the mass range of standard MALDI-TOF analysis (DeMarco and Ford, 2013). Previous antimicrobial susceptibility profiling studies by using MALDI-TOF MS and ML also did not find the biomarkers related to AMR proteins (Sousa *et al.*, 2020). However, the PPI cluster analysis of Chapter 5 showed that these proteins known to confer resistance have been found to interact with most of the biomarkers and to form a highly connected benzylpenicillin proteome network. Again, the PPI cluster analysis of Chapter 6 showed that some antimicrobial-resistant proteins and biomarkers were interacting with each other directly or via common first interactors.

MALDI-TOF MS technology has also been offered as an alternative to the current antimicrobial susceptibility test (detailed in the Introduction). In the dairy industry, phenotypic tests are still commonly used; however, breakpoints of the antimicrobials are decided based on human pathogens or animal pathogens associated with other than bovine mastitis. Determination and validation of antimicrobial susceptibility breakpoints depend on pathogen organism, disease

type, the tissue and organ where the infection occurs and the host species (Cameron *et al.*, 2016). Currently, antimicrobial susceptibility profiles of bovine mastitis pathogens are mainly decided based on the clinical breakpoints defined by CLSI. However, available breakpoints for bovine mastitis pathogens are limited to ceftiofur, penicillin/novobiocin and pirlimycin only (Ruegg *et al.*, 2015). Breakpoints in penicillin/novobiocin and pirlimycin are available for *S. aureus*, *S. agalactiae*, *S. dysgalactiae* and *S. uberis* IMIs, whereas breakpoints in ceftiofur are available for the same pathogens and *E. coli* IMIs (Cameron *et al.*, 2016). As indicated in the study by Constable and Morin (2003), to determine the breakpoints for mastitis pathogens, certain requirements are still missing. for instance, pharmacokinetics and pharmacodynamics data for treatment of clinical mastitis is missing or the recommended breakpoints are found by oral or intravenous administration route in humans; however, other routes such as intra-mammary, subcutaneous or intramuscular routes are preferred for dairy cow treatment (Léger *et al.*, 2017). Hence, clinical breakpoints for bovine mastitis pathogens are decided based on indications other than mastitis or even on species other than bovine, which could lead to mis-interpretation of the analysis results (Schwarz *et al.*, 2010). In a study of bovine mastitis pathogens in Belgian dairy farms, antimicrobial susceptibility profiling breakpoints were not available for 42% of the cases and more than 80% of cases were not veterinary specific but from human host data (Supré, Lommelen and De Meulemeester, 2014). Another important limitation of the use of mastitis-non-specific breakpoints is that comparison between studies is not possible as the analysis may use different breakpoints; for example, clinical breakpoints are not the same in French and British antimicrobial susceptibility profile surveillances of mastitis agents (de Jong *et al.*, 2018). Clinicians are greatly in need of breakpoints that refer to antimicrobial susceptibility profiles of mastitis pathogens. MALDI-TOF coupled with ML is not dependent on breakpoints and therefore can provide a solution for identifying the antimicrobial susceptibility profiles of the organisms with no need of disease-specific clinical breakpoints. In Chapters 5 and 6, it was shown that the proteome of mastitis pathogens *S. aureus*, *E. faecalis* and *E. faecium* could be discriminated based on their antimicrobial phenotype.

Overall, in this study, several ML techniques were successfully performed to classify MALDI-TOF data coming from the bovine mastitis pathogens with different phenotypes. There is abundant room for further progress in the use of MALDI-TOF coupled with ML. MALDI spectra obtained by uniformed techniques (same instrument, same data preparation etc.) can be deposit in online libraries with their metadata which enables larger datasets with more variation (e.g. geographical, time series etc.) and fairly balanced phenotypes. More complex ML techniques

such as deep learning, which demand large datasets and more computational power, can then be used for achieving even better prediction performance.

REFERENCES

Aarestrup, F. M., Agerso, Y., Gerner–Smidt, P., Madsen, M. and Jensen, L. B. (2000) 'Comparison of antimicrobial resistance phenotypes and resistance genes in Enterococcus faecalis and Enterococcus faecium from humans in the community, broilers, and pigs in Denmark', *Diagnostic Microbiology and Infectious Disease,* 37(2), pp. 127-137.

Aarestrup, F. M., Butaye, P. and Witte, W. (2002) 'Nonhuman reservoirs of enterococci', *The enterococci*: American Society of Microbiology, pp. 55-99.

Aarestrup, F. M. and Jensen, N. E. (1996) 'Genotypic and phenotypic diversity of Streptococcus dysgalactiae strains isolated from clinical and subclinical cases of bovine mastitis', *Veterinary Microbiology,* 53(3), pp. 315-323.

Aarestrup, F. M. and Jensen, N. E. (1998) 'Development of Penicillin Resistance among Staphylococcus aureus Isolated from Bovine Mastitis in Denmark and Other Countries', *Microbial Drug Resistance,* 4(3), pp. 247-256.

Aarestrup, F. M., Wegener, H. C., Jensen, N. E., Jonsson, O., Myllys, V., Thorberg, B. M., Waage, S. and Rosdahl, V. T. (1997) 'A study of phage- and ribotype patterns of Staphylococcus aureus isolated from bovine mastitis in the Nordic countries', *Acta Vet Scand,* 38(3), pp. 243-52.

Abdi, H. (2007) 'Discriminant correspondence analysis', *Encyclopedia of measurement and statistics*, pp. 270-275.

Abdulmawjood, A., Bülte, M., Cook, N., Roth, S., Schönenbrücher, H. and Hoorfar, J. (2003) 'Toward an international standard for PCR-based detection of Escherichia coli O157: Part 1. Assay development and multi-center validation', *Journal of Microbiological Methods,* 55(3), pp. 775-786.

Abebe, R., Hatiya, H., Abera, M., Megersa, B. and Asmare, K. (2016) 'Bovine mastitis: prevalence, risk factors and isolation of Staphylococcus aureus in dairy herds at Hawassa milk shed, South Ethiopia', *BMC Veterinary Research,* 12(1), pp. 270.

Abu-Ali, G. S., Lacher, D. W., Wick, L. M., Qi, W. and Whittam, T. S. (2009) 'Genomic diversity of pathogenic Escherichia coli of the EHEC 2 clonal complex', *BMC genomics,* 10(1), pp. 296.

Abushaheen, M. A., Muzaheed, Fatani, A. J., Alosaimi, M., Mansy, W., George, M., Acharya, S., Rathod, S., Divakar, D. D., Jhugroo, C., Vellappally, S., Khan, A. A., Shaik, J. and Jhugroo, P. (2020) 'Antimicrobial resistance, mechanisms and its clinical significance', *Disease-a-Month,* 66(6), pp. 100971.

Adator, H. E., Walker, M., Narvaez-Bravo, C., Zaheer, R., Goji, N., Cook, R. S., Tymensen, L., Hannon, J. S., Church, D., Booker, W. C., Amoako, K., Nadon, A. C., Read, R. and McAllister, A. T. (2020) 'Whole Genome Sequencing Differentiates Presumptive Extended Spectrum Beta-Lactamase Producing Escherichia coli along Segments of the One Health Continuum', *Microorganisms,* 8(3).

Adnan, S.-N.-A., Ibrahim, N. and Yaacob, W. A. (2017) 'Transcriptome analysis of methicillin-resistant Staphylococcus aureus in response to stigmasterol and lupeol', *Journal of Global Antimicrobial Resistance,* 8, pp. 48-54.

Afema, J. A., Ahmed, S., Besser, T. E., Jones, L. P., Sischo, W. M. and Davis, M. A. (2018) 'Molecular Epidemiology of Dairy Cattle-Associated <span class="named-content genus-species" id="named-

content-1">Escherichia coli Carrying bla_{CTX-M} Genes in Washington State', *Applied and Environmental Microbiology,* 84(6), pp. e02430-17.

Aghamohammadi, M., Haine, D., Kelton, D. F., Barkema, H. W., Hogeveen, H., Keefe, G. P. and Dufour, S. (2018) 'Herd-Level Mastitis-Associated Costs on Canadian Dairy Farms', *Frontiers in Veterinary Science,* 5, pp. 100.

Aguilar-Ayala, D. A., Cnockaert, M., André, E., Andries, K., Gonzalez-Y-Merchand, J. A., Vandamme, P., Palomino, J. C. and Martin, A. (2017) 'In vitro activity of bedaquiline against rapidly growing nontuberculous mycobacteria', *Journal of Medical Microbiology,* 66(8), pp. 1140-1143.

Ahmed, S., Olsen, J. E. and Herrero-Fresno, A. (2017) 'The genetic diversity of commensal Escherichia coli strains isolated from non-antimicrobial treated pigs varies according to age group', *PloS one,* 12(5), pp. e0178623-e0178623.

Al-Masoud, N., Xu, Y., Nicolaou, N. and Goodacre, R. (2014) 'Optimization of matrix assisted desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS) for the characterization of Bacillus and Brevibacillus species', *Analytica Chimica Acta,* 840, pp. 49-57.

Albert, R. and Barabási, A.-L. (2002) 'Statistical mechanics of complex networks', *Reviews of modern physics,* 74(1), pp. 47.

Alemayehu, T. and Hailemariam, M. (2020) 'Prevalence of vancomycin-resistant enterococcus in Africa in one health approach: a systematic review and meta-analysis', *Scientific Reports,* 10(1), pp. 20542.

Algammal, A. M., Enany, M. E., El-Tarabili, R. M., Ghobashy, M. O. I. and Helmy, Y. A. (2020) 'Prevalence, Antimicrobial Resistance Profiles, Virulence and Enterotoxin-Determinant Genes of MRSA Isolated from Subclinical Bovine Mastitis Samples in Egypt', *Pathogens,* 9(5), pp. 362.

Alghoribi, M. F., Gibreel, T. M., Dodgson, A. R., Beatson, S. A. and Upton, M. (2014) 'Galleria mellonella infection model demonstrates high lethality of ST69 and ST127 uropathogenic E. coli', *PloS one,* 9(7).

Ali, T., Rahman, S. U., Zhang, L., Shahid, M., Han, D., Gao, J., Zhang, S., Ruegg, P. L., Saddique, U. and Han, B. (2017) 'Characteristics and genetic diversity of multi-drug resistant extended-spectrum beta-lactamase (ESBL)-producing Escherichia coli isolated from bovine mastitis', *Oncotarget,* 8(52), pp. 90144-90163.

Alikhan, N.-F., Petty, N. K., Zakour, N. L. B. and Beatson, S. A. (2011) 'BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons', *BMC genomics,* 12(1), pp. 402.

Almeida, R. A., Kerro-Dego, O., Prado, M. E., Headrick, S. I., Lewis, M. J., Siebert, L. J., Pighetti, G. M. and Oliver, S. P. (2015) 'Protective effect of anti-SUAM antibodies on Streptococcus uberis mastitis', *Veterinary Research,* 46(1), pp. 133.

Almeida, R. A., Luther, D. A., Park, H.-M. and Oliver, S. P. (2006) 'Identification, isolation, and partial characterization of a novel Streptococcus uberis adhesion molecule (SUAM)', *Veterinary Microbiology,* 115(1), pp. 183-191.

Alpaydin, E. (2020) *Introduction to machine learning.* MIT press.

Alves, F. C. B., Albano, M., Andrade, B. F. M. T., Chechi, J. L., Pereira, A. F. M., Furlanetto, A., Rall, V. L. M., Fernandes, A. A. H., dos Santos, L. D. and Barbosa, L. N. (2019) 'Comparative Proteomics of

Methicillin-Resistant Staphylococcus aureus Subjected to Synergistic Effects of the Lantibiotic Nisin and Oxacillin', *Microbial Drug Resistance*.

Anderson, K. L., Lyman, R., Moury, K., Ray, D., Watson, D. W. and Correa, M. T. (2012) 'Molecular epidemiology of Staphylococcus aureus mastitis in dairy heifers', *Journal of Dairy Science,* 95(9), pp. 4921-4930.

Anderson, T. W. and Darling, D. A. (1954) 'A test of goodness of fit', *Journal of the American statistical association,* 49(268), pp. 765-769.

Andrews, S. (2014) *FastQC A Quality Control tool for High Throughput Sequence Data.*

Antipov, S. S., Tutukina, M. N., Preobrazhenskaya, E. V., Kondrashov, F. A., Patrushev, M. V., Toshchakov, S. V., Dominova, I., Shvyreva, U. S., Vrublevskaya, V. V., Morenkov, O. S., Sukharicheva, N. A., Panyukov, V. V. and Ozoline, O. N. (2017) 'The nucleoid protein Dps binds genomic DNA of Escherichia coli in a non-random manner', *PLoS One,* 12(8), pp. e0182800.

Apparao, M. D., Ruegg, P. L., Lago, A., Godden, S., Bey, R. and Leslie, K. (2009) 'Relationship between in vitro susceptibility test results and treatment outcomes for gram-positive mastitis pathogens following treatment with cephapirin sodium', *Journal of Dairy Science,* 92(6), pp. 2589-2597.

Arabnia, H. R. and Tran, Q. N. (2015) *Emerging trends in computational biology, bioinformatics, and systems biology: algorithms and software tools.* Morgan Kaufmann.

Araújo, T. F. and Ferreira, C. L. d. L. F. (2013) 'The genus Enterococcus as probiotic: safety concerns', *Brazilian Archives of Biology and Technology,* 56, pp. 457-466.

Arbabshirani, M. R., Plis, S., Sui, J. and Calhoun, V. D. (2017) 'Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls', *NeuroImage,* 145, pp. 137-165.

Arbeloa, A., Segal, H., Hugonnet, J.-E., Josseaume, N., Dubost, L., Brouard, J.-P., Gutmann, L., Mengin-Lecreulx, D. and Arthur, M. (2004) 'Role of class A penicillin-binding proteins in PBP5-mediated β-lactam resistance in Enterococcus faecalis', *Journal of bacteriology,* 186(5), pp. 1221-1228.

Archer, N., Egan, S. A., Coffey, T. J., Emes, R. D., Addis, M. F., Ward, P. N., Blanchard, A. M. and Leigh, J. A. (2020) 'A Paradox in Bacterial Pathogenesis: Activation of the Local Macrophage Inflammasome Is Required for Virulence of Streptococcus uberis', *Pathogens,* 9(12), pp. 997.

Archer, S. C., Bradley, A. J., Cooper, S., Davies, P. L. and Green, M. J. (2017) 'Prediction of Streptococcus uberis clinical mastitis risk using Matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS) in dairy herds', *Preventive veterinary medicine,* 144, pp. 1-6.

Argaw, A. (2016) 'Review on epidemiology of clinical and subclinical mastitis on dairy cows', *Food Sci. Qual. Manag,* 52, pp. 56-65.

Arias, C. A. and Murray, B. E. (2012) 'The rise of the Enterococcus: beyond vancomycin resistance', *Nat Rev Microbiol,* 10(4), pp. 266-78.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. and Wishart, D. S. (2016) 'PHASTER: a better, faster version of the PHAST phage search tool', *Nucleic Acids Res,* 44(W1), pp. W16-21.

Arneberg, R., Rajalahti, T., Flikka, K., Berven, F. S., Kroksveen, A. C., Berle, M., Myhr, K.-M., Vedeler, C. A., Ulvik, R. J. and Kvalheim, O. M. (2007) 'Pretreatment of mass spectral profiles: application to proteomic data', *Analytical chemistry,* 79(18), pp. 7014-7026.

Arnold, R. J., Karty, J. A., Ellington, A. D. and Reilly, J. P. (1999) 'Monitoring the Growth of a Bacteria Culture by MALDI-MS of Whole Cells', *Analytical Chemistry,* 71(10), pp. 1990-1996.

Arnold, R. J. and Reilly, J. P. (1998) 'Fingerprint matching of E. coli strains with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of whole cells using a modified correlation approach', *Rapid Communications in Mass Spectrometry,* 12(10), pp. 630-636.

Arnold, R. J. and Reilly, J. P. (1999) 'Observation of Escherichia coli ribosomal proteins and their posttranslational modifications by mass spectrometry', *Anal Biochem,* 269(1), pp. 105-12.

Arruda, A. G., Godden, S., Rapnicki, P., Gorden, P., Timms, L., Aly, S. S., Lehenbauer, T. W. and Champagne, J. (2013) 'Randomized noninferiority clinical trial evaluating 3 commercial dry cow mastitis preparations: I. Quarter-level outcomes', *Journal of dairy science,* 96(7), pp. 4419-4435.

Arthur, M., Depardieu, F., Gerbaud, G., Galimand, M., Leclercq, R. and Courvalin, P. (1997) 'The VanS sensor negatively controls VanR-mediated transcriptional activation of glycopeptide resistance genes of Tn1546 and related elements in the absence of induction', *Journal of Bacteriology,* 179(1), pp. 97.

Asakura, K., Azechi, T., Sasano, H., Matsui, H., Hanaki, H., Miyazaki, M., Takata, T., Sekine, M., Takaku, T., Ochiai, T., Komatsu, N., Shibayama, K., Katayama, Y. and Yahara, K. (2018) 'Rapid and easy detection of low-level resistance to vancomycin in methicillin-resistant Staphylococcus aureus by matrix-assisted laser desorption ionization time-of-flight mass spectrometry', *PloS one,* 13(3), pp. e0194212-e0194212.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. and Eppig, J. T. (2000) 'Gene Ontology: tool for the unification of biology', *Nature genetics,* 25(1), pp. 25-29.

Aslantaş, Ö. and Demir, C. (2016) 'Investigation of the antibiotic resistance and biofilm-forming ability of Staphylococcus aureus from subclinical bovine mastitis cases', *Journal of Dairy Science,* 99(11), pp. 8607-8613.

Assareh, A., Volkert, L. G. and Li, J. 'Feature selections using AdaBoost: Application in gene-gene interaction detection'. *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 4-7 Oct. 2012, 831-837.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. and Albrecht, M. (2008) 'Computing topological parameters of biological networks', *Bioinformatics,* 24(2), pp. 282-284.

Aurrecoechea, C., Barreto, A., Basenko, E. Y., Brestelli, J., Brunk, B. P., Cade, S., Crouch, K., Doherty, R., Falke, D., Fischer, S., Gajria, B., Harb, O. S., Heiges, M., Hertz-Fowler, C., Hu, S., Iodice, J., Kissinger, J. C., Lawrence, C., Li, W., Pinney, D. F., Pulman, J. A., Roos, D. S., Shanmugasundram, A., Silva-Franco, F., Steinbiss, S., Stoeckert, C. J., Jr., Spruill, D., Wang, H., Warrenfeltz, S. and Zheng, J. (2017) 'EuPathDB: the eukaryotic pathogen genomics database resource', *Nucleic acids research,* 45(D1), pp. D581-D591.

Axelsson, C., Rehnstam-Holm, A.-S. and Nilson, B. (2019) 'Rapid detection of antibiotic resistance in positive blood cultures by MALDI-TOF MS and an automated and optimized MBT-ASTRA protocol for Escherichia coli and Klebsiella pneumoniae', *Infectious Diseases*, pp. 1-9.

Ayodele, T. O. (2010) 'Types of machine learning algorithms', *New advances in machine learning,* 3, pp. 19-48.

Azadbakht, M., Fraser, C. S. and Khoshelham, K. (2018) 'Synergy of sampling techniques and ensemble classifiers for classification of urban environments using full-waveform LiDAR data', *International Journal of Applied Earth Observation and Geoinformation,* 73, pp. 277-291.

Babin, B. M., Atangcho, L., van Eldijk, M. B., Sweredoski, M. J., Moradian, A., Hess, S., Tolker-Nielsen, T., Newman, D. K. and Tirrell, D. A. (2017) 'Selective Proteomic Analysis of Antibiotic-Tolerant Cellular Subpopulations in Pseudomonas aeruginosa Biofilms', *mBio,* 8(5), pp. e01593-17.

Babu, M., Bundalovic-Torma, C., Calmettes, C., Phanse, S., Zhang, Q., Jiang, Y., Minic, Z., Kim, S., Mehla, J. and Gagarinova, A. (2018) 'Global landscape of cell envelope protein complexes in Escherichia coli', *Nature biotechnology,* 36(1), pp. 103-112.

Badie, G., Heithoff, D. M., Sinsheimer, R. L. and Mahan, M. J. (2007) 'Altered levels of Salmonella DNA adenine methylase are associated with defects in gene expression, motility, flagellar synthesis, and bile resistance in the pathogenic strain 14028 but not in the laboratory strain LT2', *Journal of bacteriology,* 189(5), pp. 1556-1564.

Bai, J., Fan, Z. C., Zhang, L. P., Xu, X. Y. and Zhang, Z. L. 'Classification of Methicillin-Resistant and Methicillin-Susceptible Staphylococcus Aureus Using an Improved Genetic Algorithm for Feature Selection Based on Mass Spectra'. 2017, 57-63.

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., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. and Pevzner, P. A. (2012) 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *J Comput Biol,* 19(5), pp. 455-77.

Banko, M. and Brill, E. 'Scaling to very very large corpora for natural language disambiguation'. 2001: Association for Computational Linguistics, 26-33.

Barkema, H. W., Schukken, Y. H., Lam, T., Beiboer, M. L., Wilmink, H., Benedictus, G. and Brand, A. (1998) 'Incidence of clinical mastitis in dairy herds grouped in three categories by bulk milk somatic cell counts', *Journal of dairy science,* 81(2), pp. 411-419.

Barkema, H. W., Schukken, Y. H. and Zadoks, R. N. (2006) 'Invited review: The role of cow, pathogen, and treatment regimen in the therapeutic success of bovine Staphylococcus aureus mastitis', *Journal of Dairy Science,* 89(6), pp. 1877-1895.

Barlow, J. W., Zadoks, R. N. and Schukken, Y. H. (2013) 'Effect of lactation therapy on Staphylococcus aureus transmission dynamics in two commercial dairy herds', *BMC Vet Res,* 9, pp. 28.

Barreiro, J. R., Braga, P. A., Ferreira, C. R., Kostrzewa, M., Maier, T., Wegemann, B., Böettcher, V., Eberlin, M. N. and dos Santos, M. V. (2012) 'Nonculture-based identification of bacteria in milk by protein fingerprinting', *Proteomics,* 12(17), pp. 2739-45.

Barreiro, J. R., Ferreira, C. R., Sanvido, G. B., Kostrzewa, M., Maier, T., Wegemann, B., Böttcher, V., Eberlin, M. N. and dos Santos, M. V. (2010) 'Identification of subclinical cow mastitis pathogens in milk by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry', *Journal of dairy science,* 93(12), pp. 5661-5667.

Barreiro, J. R., Gonçalves, J. L., Braga, P. A. C., Dibbern, A. G., Eberlin, M. N. and Veiga Dos Santos, M. (2017) 'Non-culture-based identification of mastitis-causing bacteria by MALDI-TOF mass spectrometry', *J Dairy Sci,* 100(4), pp. 2928-2934.

Baseggio, N., Mansell, P. D., Browning, J. W. and Browning, G. F. (1997) 'Strain differentiation of isolates of streptococci from bovine mastitis by pulsed-field gel electrophoresis', *Molecular and cellular probes,* 11(5), pp. 349-354.

Bauer, S. (2017) 'Gene-category analysis', *Methods Mol. Biol,* 1446, pp. 175-188.

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. and Clermont, O. (2018) 'ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping', *Microb Genom,* 4(7).

Bem, A. E., Velikova, N., Pellicer, M. T., Baarlen, P. v., Marina, A. and Wells, J. M. (2015) 'Bacterial Histidine Kinases as Novel Antibacterial Drug Targets', *ACS Chemical Biology,* 10(1), pp. 213-224.

Ben Braïek, O. and Smaoui, S. (2019) 'Enterococci: between emerging pathogens and potential probiotics', *BioMed Research International,* 2019.

Ben Yahia, H., Ben Sallem, R., Tayh, G., Klibi, N., Ben Amor, I., Gharsa, H., Boudabbous, A. and Ben Slama, K. (2018) 'Detection of CTX-M-15 harboring Escherichia coli isolated from wild birds in Tunisia', *BMC Microbiology,* 18(1), pp. 26.

Benkova, M., Soukup, O. and Marek, J. (2020) 'Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice', *Journal of applied microbiology,* 129(4), pp. 806-822.

Berggård, T., Linse, S. and James, P. (2007) 'Methods for the detection and analysis of protein–protein interactions', *Proteomics,* 7(16), pp. 2833-2842.

Bergholz, T., Tarr, C., M Christensen, L., J Betting, D. and S Whittam, T. (2007) *Recent Gene Conversions between Duplicated Glutamate Decarboxylase Genes (gadA and gadB) in Pathogenic Escherichia coli.*

Bergmann, R., van der Linden, M., Chhatwal, G. S. and Nitsche-Schmitz, D. P. (2014) 'Factors That Cause Trimethoprim Resistance in Streptococcus pyogenes', *Antimicrobial Agents and Chemotherapy,* 58(4), pp. 2281.

Berry, D. P. and Meaney, W. J. (2006) 'Interdependence and distribution of subclinical mastitis and intramammary infection among udder quarters in dairy cattle', *Preventive veterinary medicine,* 75(1-2), pp. 81-91.

Bes, M., Guerin-Faublee, V., Meugnier, H., Etienne, J. and Freney, J. (2000) 'Improvement of the identification of staphylococci isolated from bovine mammary infections using molecular methods', *Veterinary microbiology,* 71(3), pp. 287-294.

Beukers, A. G., Zaheer, R., Goji, N., Amoako, K. K., Chaves, A. V., Ward, M. P. and McAllister, T. A. (2017) 'Comparative genomics of Enterococcus spp. isolated from bovine feces', *BMC microbiology,* 17(1), pp. 1-18.

Beutin, L. and Strauch, E. (2007) 'Identification of sequence diversity in the Escherichia coli fliC genes encoding flagellar types H8 and H40 and its use in typing of Shiga toxin-producing E. coli O8, O22, O111, O174, and O179 strains', *J Clin Microbiol,* 45(2), pp. 333-9.

Blowey, R. W. and Edmondson, P. (2010) *Mastitis control in dairy herds.* Cabi.

Blum, S. E., Goldstone, R. J., Connolly, J. P. R., Répérant-Ferter, M., Germon, P., Inglis, N. F., Krifucks, O., Mathur, S., Manson, E., McLean, K., Rainard, P., Roe, A. J., Leitner, G. and Smith, D. G. E. (2018) 'Postgenomics Characterization of an Essential Genetic Determinant of Mammary Pathogenic Escherichia coli', *mBio,* 9(2), pp. e00423-18.

Blum, S. E., Heller, E. D., Sela, S., Elad, D., Edery, N. and Leitner, G. (2015) 'Genomic and Phenomic Study of Mammary Pathogenic Escherichia coli', *PloS one,* 10(9), pp. e0136387-e0136387.

Blum, S. E. and Leitner, G. (2013) 'Genotyping and virulence factors assessment of bovine mastitis Escherichia coli', *Vet Microbiol,* 163(3-4), pp. 305-12.

Boehmke, B. and Greenwell, B. M. (2019) *Hands-on machine learning with R.* CRC Press.

Boerlin, P. (1997) 'Applications of multilocus enzyme electrophoresis in medical microbiology', *Journal of microbiological methods,* 28(3), pp. 221-231.

Boerlin, P., Travis, R., Gyles, C. L., Reid-Smith, R., Heather Lim, N. J., Nicholson, V., McEwen, S. A., Friendship, R. and Archambault, M. (2005) 'Antimicrobial Resistance and Virulence Genes of Escherichia coli Isolates from Swine in Ontario', *Applied and Environmental Microbiology,* 71(11), pp. 6753.

Bogni, C., Odierno, L., Raspanti, C., Giraudo, J., Larriestra, A., Reinoso, E., Lasagno, M., Ferrari, M., Ducrós, E. and Frigerio, C. (2011) 'War against mastitis: Current concepts on controlling bovine mastitis pathogens', *Science against microbial pathogens: Communicafing current research and technological advances*, pp. 483-494.

Boireau, C., Cazeau, G., Jarrige, N., Calavas, D., Madec, J. Y., Leblond, A., Haenni, M. and Gay, É. (2018) 'Antimicrobial resistance in bacteria isolated from mastitis in dairy cattle in France, 2006-2016', *J Dairy Sci,* 101(10), pp. 9451-9462.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics,* 30(15), pp. 2114-20.

Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A. and Pevzner, P. A. (2013) 'N-terminal protein processing: a comparative proteogenomic analysis', *Molecular & cellular proteomics : MCP,* 12(1), pp. 14-28.

Borgaonkar, S. P., Hocker, H., Shin, H. and Markey, M. K. (2010) 'Comparison of normalization methods for the identification of biomarkers using MALDI-TOF and SELDI-TOF mass spectra', *OMICS A Journal of Integrative Biology,* 14(1), pp. 115-126.

Bose, T., Venkatesh, K. V. and Mande, S. S. (2017) 'Computational Analysis of Host-Pathogen Protein Interactions between Humans and Different Strains of Enterohemorrhagic Escherichia coli', *Front Cell Infect Microbiol,* 7, pp. 128.

Botrel, M.-A., Haenni, M., Morignat, E., Sulpice, P., Madec, J.-Y. and Calavas, D. (2010) 'Distribution and antimicrobial resistance of clinical and subclinical mastitis pathogens in dairy cows in Rhône-Alpes, France', *Foodborne pathogens and disease,* 7(5), pp. 479-487.

Bouchard, D., Peton, V., Almeida, S., Le Maréchal, C., Miyoshi, A., Azevedo, V., Berkova, N., Rault, L., François, P., Schrenzel, J., Even, S., Hernandez, D. and Le Loir, Y. (2012) 'Genome Sequence of Staphylococcus aureus

Newbould 305, a Strain Associated with Mild Bovine Mastitis', *Journal of Bacteriology,* 194(22), pp. 6292.

Bouchet, V., Huot, H. and Goldstein, R. (2008) 'Molecular genetic basis of ribotyping', *Clinical microbiology reviews,* 21(2), pp. 262-273.

Bozdogan, B., Berrezouga, L., Kuo, M.-S., Yurek, D. A., Farley, K. A., Stockman, B. J. and Leclercq, R. (1999) 'A New Resistance Gene, linB, Conferring Resistance to Lincosamides by Nucleotidylation in Enterococcus faecium HM1025', *Antimicrobial Agents and Chemotherapy,* 43(4), pp. 925.

Božik, M., Cejnar, P., Šašková, M., Nový, P., Maršík, P. and Klouček, P. (2018) 'Stress response of Escherichia coli to essential oil components – insights on low-molecular-weight proteins from MALDI-TOF', *Scientific Reports,* 8(1), pp. 13042.

Bradley, A., Biggs, A., Green, M. and Lam, T. J. G. M. (2012) 'Control of mastitis and enhancement of milk quality', pp. 117-168.

Bradley, A. J. (2002) 'Bovine mastitis: an evolving disease', *The Veterinary Journal,* 164(2), pp. 116-128.

Bradley, A. J., Breen, J. E., Payne, B., White, V. and Green, M. J. (2015) 'An investigation of the efficacy of a polyvalent mastitis vaccine using different vaccination regimens under field conditions in the United Kingdom', *J Dairy Sci,* 98(3), pp. 1706-20.

Bradley, A. J. and Green, M. J. (2000) 'A study of the incidence and significance of intramammary enterobacterial infections acquired during the dry period', *J Dairy Sci,* 83(9), pp. 1957-65.

Bradley, A. J. and Green, M. J. (2001a) 'Adaptation of Escherichia coli to the bovine mammary gland', *Journal of clinical microbiology,* 39(5), pp. 1845-1849.

Bradley, A. J. and Green, M. J. (2001b) 'Aetiology of clinical mastitis in six Somerset dairy herds', *Veterinary Record,* 148(22), pp. 683-686.

Bradley, A. J. and Green, M. J. (2004) 'The importance of the nonlactating period in the epidemiology of intramammary infection and strategies for prevention', *Veterinary Clinics: Food Animal Practice,* 20(3), pp. 547-568.

Bradley, A. J., Leach, K. A., Archer, S. C., Breen, J. E., Green, M. J., Ohnstad, I. and Tuer, S. (2014) 'Scoping study on the potential risks (and benefits) of using recycled manure solids as bedding for dairy cattle'.

Bradley, A. J., Leach, K. A., Breen, J. E., Green, L. A. and Green, M. J. (2007) 'Survey of the incidence and etiology of mastitis on dairy farms in England and Wales. 2007', *Vet Rec,* 160, pp. 253-258.

Bradley, A. J., Leach, K. A., Green, M. J., Gibbons, J., Ohnstad, I. C., Black, D. H., Payne, B., Prout, V. E. and Breen, J. E. (2018) 'The impact of dairy cows' bedding material and its microbial content on the quality and safety of milk–A cross sectional study of UK farms', *International journal of food microbiology,* 269, pp. 36-45.

Breiman, L. (2001) 'Random forests', *Machine learning,* 45(1), pp. 5-32.

Brennan, E., Martins, M., McCusker, M. P., Wang, J., Alves, B. M., Hurley, D., El Garch, F., Woehrlé, F., Miossec, C., McGrath, L., Srikumar, S., Wall, P. and Fanning, S. (2016) 'Multidrug-Resistant Escherichia coli in Bovine Animals, Europe', *Emerging infectious diseases,* 22(9), pp. 1650-1652.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., 3rd, Stevens, R., Vonstein, V., Wattam, A. R. and Xia, F. (2015) 'RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes', *Sci Rep,* 5, pp. 8365.

Brightling, P. B., Dyson, R. D., Hope, A. F. and Penry, J. (2009) 'A national programme for mastitis control in Australia: Countdown Downunder', *Irish Veterinary Journal,* 62(4), pp. 1-7.

Brown, C. T. and Irber, L. (2016) 'sourmash: a library for MinHash sketching of DNA', *J. Open Source Softw,* 1, pp. 27.

Bruker Daltonics 2011. ClinProTools 3.0: User Manual. Bremen: Bruker Daltonik GmbH.

Brynildsrud, O., Bohlin, J., Scheffer, L. and Eldholm, V. (2016) 'Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary', *Genome Biol,* 17(1), pp. 238.

Buda, M., Maki, A. and Mazurowski, M. A. (2018) 'A systematic study of the class imbalance problem in convolutional neural networks', *Neural Networks,* 106, pp. 249-259.

Bumunang, E. W., McAllister, T. A., Zaheer, R., Ortega Polo, R., Stanford, K., King, R., Niu, Y. D. and Ateba, C. N. (2019) 'Characterization of Non-O157 Escherichia coli from Cattle Faecal Samples in the North-West Province of South Africa', *Microorganisms,* 7(8).

Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S.-Y., Mulder, N. and Hunter, S. (2012) 'Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation', *Database,* 2012.

Burvenich, C., Van Merris, V., Mehrzad, J., Diez-Fraile, A. and Duchateau, L. (2003) 'Severity of E. coli mastitis is mainly determined by cow factors', *Veterinary research,* 34(5), pp. 521-564.

Butland, G., Peregrín-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. and Emili, A. (2005) 'Interaction network containing conserved and essential protein complexes in Escherichia coli', *Nature,* 433(7025), pp. 531-537.

Caballero, B., Trugo, L. C. and Finglas, P. M. (2003) *Encyclopedia of food sciences and nutrition.* Academic.

Cameron, M., Barkema, H. W., De Buck, J., De Vliegher, S., Chaffer, M., Lewis, J. and Keefe, G. P. (2017) 'Identification of bovine-associated coagulase-negative staphylococci by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using a direct transfer protocol', *Journal of dairy science,* 100(3), pp. 2137-2147.

Cameron, M., Saab, M., Heider, L., McClure, J., Rodriguez-Lecompte, J. C. and Sanchez, J. (2016) 'Antimicrobial susceptibility patterns of environmental Streptococci recovered from bovine milk samples in the maritime provinces of Canada', *Frontiers in veterinary science,* 3, pp. 79.

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F. and Hasman, H. (2014) 'In silico detection and typing of plasmids using PlasmidFinder

and plasmid multilocus sequence typing', *Antimicrobial agents and chemotherapy,* 58(7), pp. 3895-3903.

Carin, L. (2020) *Logistic Regression - Simple Introduction to Machine Learning*. Available at: https://www.coursera.org/learn/machine-learning-duke/lecture/8N63I/logistic-regression (Accessed: October 08 2020).

Cariolato, D., Andrighetto, C. and Lombardi, A. (2008) 'Occurrence of virulence factors and antibiotic resistances in Enterococcus faecalis and Enterococcus faecium collected from dairy and human samples in North Italy', *Food Control,* 19(9), pp. 886-892.

Carro, L. (2018) 'Protein-protein interactions in bacteria: a promising and challenging avenue towards the discovery of new antibiotics', *Beilstein journal of organic chemistry,* 14, pp. 2881-2896.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M.-A., Barrell, B. G. and Parkhill, J. (2005) 'ACT: the Artemis Comparison Tool', *Bioinformatics (Oxford, England),* 21(16), pp. 3422-3423.

Casadesús, J. and Low, D. (2006) 'Epigenetic gene regulation in the bacterial world', *Microbiology and molecular biology reviews,* 70(3), pp. 830-856.

Castro, V. S., Figueiredo, E. E. d. S., Stanford, K., McAllister, T. and Conte-Junior, C. A. (2019) 'Shiga-toxin producing Escherichia coli in Brazil: A systematic review', *Microorganisms,* 7(5), pp. 137.

Cauwerts, K., Decostere, A., De Graef, E. M., Haesebrouck, F. and Pasmans, F. (2007) 'High prevalence of tetracycline resistance in Enterococcus isolates from broilers carrying the erm(B) gene', *Avian Pathol,* 36(5), pp. 395-9.

Cawley, G. C. and Talbot, N. L. C. (2010) 'On over-fitting in model selection and subsequent selection bias in performance evaluation', *Journal of Machine Learning Research,* 11(Jul), pp. 2079-2107.

Cervinkova, D., Vlkova, H., Borodacova, I., Makovcova, J., Babak, V., Lorencova, A., Vrtkova, I., Marosevic, D. and Jaglic, Z. (2013) 'Prevalence of mastitis pathogens in milk from clinically healthy cows', *Veterinarni medicina,* 58(11), pp. 567-575.

Chang, C., Coggill, P., Bateman, A., Finn, R. D., Cymborowski, M., Otwinowski, Z., Minor, W., Volkart, L. and Joachimiak, A. (2009) 'The structure of pyogenecin immunity protein, a novel bacteriocin-like immunity protein from Streptococcus pyogenes', *BMC structural biology,* 9(1), pp. 75.

Chassaing, B., Rolhion, N., de Vallée, A., Sa'ad, Y. S., Prorok-Hamon, M., Neut, C., Campbell, B. J., Söderholm, J. D., Hugot, J.-P. and Colombel, J.-F. (2011) 'Crohn disease–associated adherent-invasive E. coli bacteria target mouse and human Peyer's patches via long polar fimbriae', *The Journal of clinical investigation,* 121(3), pp. 966-975.

Chaudhuri, R. R., Allen, A. G., Owen, P. J., Shalom, G., Stone, K., Harrison, M., Burgis, T. A., Lockyer, M., Garcia-Lara, J. and Foster, S. J. (2009) 'Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH)', *BMC genomics,* 10(1), pp. 291.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research,* 16, pp. 321-357.

Chee-Sanford, J. C., Aminov, R. I., Krapac, I. J., Garrigues-Jeanjean, N. and Mackie, R. I. (2001) 'Occurrence and diversity of tetracycline resistance genes in lagoons and groundwater underlying two swine production facilities', *Applied and environmental microbiology,* 67(4), pp. 1494-1502.

Chen, C., Hou, J., Tanner, J. J. and Cheng, J. (2020) 'Bioinformatics methods for mass spectrometry-based proteomics data analysis', *International journal of molecular sciences,* 21(8), pp. 2873.

Chen, L., Wang, L., Yassin, A. K., Zhang, J., Gong, J., Qi, K., Ganta, R. R., Zhang, Y., Yang, Y., Han, X. and Wang, C. (2018) 'Genetic characterization of extraintestinal Escherichia coli isolates from chicken, cow and swine', *AMB Express,* 8(1), pp. 117.

Cheng, D., Qiao, L. and Horvatovich, P. (2018) 'Toward Spectral Library-Free Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry Bacterial Identification', *J Proteome Res,* 17(6), pp. 2124-2130.

Cheng, W. N. and Han, S. G. (2020) 'Bovine mastitis: risk factors, therapeutic strategies, and alternative treatments', *Asian-Australasian Journal of Animal Sciences*.

Cherkaoui, A., Hibbs, J., Emonet, S., Tangomo, M., Girard, M., Francois, P. and Schrenzel, J. (2010) 'Comparison of Two Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry Methods with Conventional Phenotypic Identification for Routine Identification of Bacteria to the Species Level', *Journal of Clinical Microbiology,* 48(4), pp. 1169.

Cho, S., Barrett, J. B., Frye, J. G. and Jackson, C. R. (2020a) 'Antimicrobial Resistance Gene Detection and Plasmid Typing Among Multidrug Resistant Enterococci Isolated from Freshwater Environment', *Microorganisms,* 8(9), pp. 1338.

Cho, S., Hiott, L. M., McDonald, J. M., Barrett, J. B., McMillan, E. A., House, S. L., Adams, E. S., Frye, J. G. and Jackson, C. R. (2020b) 'Diversity and antimicrobial resistance of Enterococcus from the Upper Oconee Watershed, Georgia', *Journal of Applied Microbiology,* 128(4), pp. 1221-1233.

Choi, E.-K., Park, Y. e., Choi, B. W., Kim, K.-S., Yang, H. J., Ahn, K. S. and Jang, H.-J. (2011) 'Genome-wide gene expression analysis of Patrinia scabiosaefolia reveals an antibiotic effect', *BioChip Journal,* 5(3), pp. 246.

Chukwudi, C. U. (2016) 'rRNA Binding Sites and the Molecular Mechanism of Action of the Tetracyclines', *Antimicrobial agents and chemotherapy,* 60(8), pp. 4433-4441.

Chung, C.-R., Wang, H.-Y., Lien, F., Tseng, Y.-J., Chen, C.-H., Lee, T.-Y., Horng, J.-T. and Lu, J.-J. (2019) 'Incorporating statistical test and machine intelligence into strain typing of Staphylococcus haemolyticus based on matrix-assisted laser desorption ionization-time of flight mass spectrometry', *Frontiers in Microbiology,* 10, pp. 2120.

Chung, P. Y., Chung, L. Y. and Navaratnam, P. (2013) 'Identification, by gene expression profiling analysis, of novel gene targets in Staphylococcus aureus treated with betulinaldehyde', *Research in Microbiology,* 164(4), pp. 319-326.

Claydon, M. A., Davey, S. N., Edwards-Jones, V. and Gordon, D. B. (1996) 'The rapid identification of intact microorganisms using mass spectrometry', *Nature Biotechnology,* 14(11), pp. 1584-1586.

Clermont, O., Gordon, D. and Denamur, E. (2015) 'Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes', *Microbiology,* 161(5), pp. 980-988.

Cocconcelli, P. S., Cattivelli, D. and Gazzola, S. (2003) 'Gene transfer of vancomycin and tetracycline resistances among Enterococcus faecalis during cheese and sausage fermentations', *International Journal of Food Microbiology,* 88(2), pp. 315-323.

Coffey, T. J., Pullinger, G. D., Urwin, R., Jolley, K. A., Wilson, S. M., Maiden, M. C. and Leigh, J. A. (2006) 'First insights into the evolution of Streptococcus uberis: a multilocus sequence typing scheme that enables investigation of its population biology', *Applied and environmental microbiology,* 72(2), pp. 1420-1428.

Collado, R., Prenafeta, A., González-González, L., Pérez-Pons, J. A. and Sitjà, M. (2016) 'Probing vaccine antigens against bovine mastitis caused by Streptococcus uberis', *Vaccine,* 34(33), pp. 3848-3854.

Compton, C. W. R., Heuer, C., Parker, K. and McDougall, S. (2007) 'Risk Factors for Peripartum Mastitis in Pasture-Grazed Dairy Heifers', *Journal of Dairy Science,* 90(9), pp. 4171-4180.

Constable, P. D. and Morin, D. E. (2003) 'Treatment of clinical mastitis: Using antimicrobial susceptibility profiles for treatment decisions', *Veterinary Clinics: Food Animal Practice,* 19(1), pp. 139-155.

Constantiniu, S. (2002) 'Escherichia coli enterohemoragic–an emerged pathogen of human infections Part II. Non-o157 Escherichia coli enterohemorrhagic', *J. Prev. Med,* 10, pp. 57-73.

Contreras, G. A. and Rodríguez, J. M. (2011) 'Mastitis: comparative etiology and epidemiology', *Journal of mammary gland biology and neoplasia,* 16(4), pp. 339-356.

Conway, J. R., Lex, A. and Gehlenborg, N. (2017) 'UpSetR: an R package for the visualization of intersecting sets and their properties', *Bioinformatics*.

Conwell, M., Daniels, V., Naughton, P. J. and Dooley, J. S. G. (2017) 'Interspecies transfer of vancomycin, erythromycin and tetracycline resistance among Enterococcus species recovered from agrarian sources', *BMC Microbiology,* 17(1), pp. 19.

Cook, N. B., Pionek, D. A. and Sharp, P. (2005) 'An assessment of the benefits of Orbeseal® when used in combination with dry cow antibiotic therapy in three commercial dairy herds', *Bovine Practitioner,* 39(2), pp. 83.

Coombes, K. R., Baggerly, K. A. and Morris, J. S. (2007) 'Pre-processing mass spectrometry data', *Fundamentals of Data Mining in Genomics and Proteomics*: Springer, pp. 79-102.

Cooper, K. K., Mandrell, R. E., Louie, J. W., Korlach, J., Clark, T. A., Parker, C. T., Huynh, S., Chain, P. S., Ahmed, S. and Carter, M. Q. (2014) 'Comparative genomics of enterohemorrhagic Escherichia coli O145: H28 demonstrates a common evolutionary lineage with Escherichia coli O157: H7', *BMC genomics,* 15(1), pp. 17.

Coque, T. M., Singh, K. V., Weinstock, G. M. and Murray, B. E. (1999) 'Characterization of Dihydrofolate Reductase Genes from Trimethoprim-Susceptible and Trimethoprim-Resistant Strains ofEnterococcus faecalis', *Antimicrobial agents and chemotherapy,* 43(1), pp. 141-147.

Cordovana, M., Pranada, A. B., Ambretti, S. and Kostrzewa, M. (2019) 'MALDI-TOF bacterial subtyping to detect antibiotic resistance', *Clinical Mass Spectrometry,* 14, pp. 3-8.

Cortes, C. and Vapnik, V. (1995) 'Machine learning', *Support vector networks,* 20(3), pp. 25.

Croxatto, A., Prod'hom, G. and Greub, G. (2012) 'Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology', *FEMS microbiology reviews,* 36(2), pp. 380-407.

Cucarella, C., Tormo, M. Á., Úbeda, C., Trotonda, M. P., Monzón, M., Peris, C., Amorena, B., Lasa, Í. and Penadés, J. R. (2004) 'Role of biofilm-associated protein bap in the pathogenesis of bovine Staphylococcus aureus', *Infection and immunity,* 72(4), pp. 2177-2185.

Cui, Z. and Gong, G. (2018) 'The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features', *Neuroimage,* 178, pp. 622-637.

Cuny, C., Wieler, L. H. and Witte, W. (2015) 'Livestock-associated MRSA: the impact on humans', *Antibiotics,* 4(4), pp. 521-543.

Dallman, T., Smith, G. P., O'Brien, B., Chattaway, M. A., Finlay, D., Grant, K. A. and Jenkins, C. (2012) 'Characterization of a verocytotoxin-producing enteroaggregative Escherichia coli serogroup O111:H21 strain associated with a household outbreak in Northern Ireland', *J Clin Microbiol,* 50(12), pp. 4116-9.

Daly, M., Power, E., Björkroth, J., Sheehan, P., O'Connell, A., Colgan, M., Korkeala, H. and Fanning, S. (1999) 'Molecular Analysis of Pseudomonas aeruginosa: Epidemiological Investigation of Mastitis Outbreaks in Irish Dairy Herds', *Applied and Environmental Microbiology,* 65(6), pp. 2723.

Daniel, Z., Witold Stanisław, P., Katarzyna, W.-M. and Wilhelm, G. (2016) 'Identification of Cows Susceptible to Mastitis based on Selected Genotypes by Using Decision Trees and A Generalized Linear Model', *Acta Veterinaria,* 66(3), pp. 317-335.

Darling, A. C. E., Mau, B., Blattner, F. R. and Perna, N. T. (2004) 'Mauve: multiple alignment of conserved genomic sequence with rearrangements', *Genome research,* 14(7), pp. 1394-1403.

Davies, P. L., Leigh, J. A., Bradley, A. J., Archer, S. C., Emes, R. D. and Green, M. J. (2016) 'Molecular epidemiology of Streptococcus uberis clinical mastitis in dairy herds: strain heterogeneity and transmission', *Journal of clinical microbiology,* 54(1), pp. 68-74.

Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M. and Wattam, A. R. (2016) 'Antimicrobial resistance prediction in PATRIC and RAST', *Scientific reports,* 6, pp. 27930.

De Briyne, N., Atkinson, J., Borriello, S. P. and Pokludová, L. (2014) 'Antibiotics used most commonly to treat animals in Europe', *Veterinary Record,* 175(13), pp. 325.

de Fátima Silva Lopes, M., Ribeiro, T., Abrantes, M., Figueiredo Marques, J. J., Tenreiro, R. and Crespo, M. T. B. (2005) 'Antimicrobial resistance profiles of dairy and clinical isolates and type strains of enterococci', *International Journal of Food Microbiology,* 103(2), pp. 191-198.

de Jong, A., Garch, F. E., Simjee, S., Moyaert, H., Rose, M., Youala, M. and Siegwart, E. (2018) 'Monitoring of antimicrobial susceptibility of udder pathogens recovered from cases of clinical mastitis in dairy cows across Europe: VetPath results', *Vet Microbiol,* 213, pp. 73-81.

De Las Rivas, J. and Fontanillo, C. (2010) 'Protein–protein interactions essentials: key concepts to building and analyzing interactome networks', *PLoS Comput Biol,* 6(6), pp. e1000807.

de Souza Figueiredo, E. E., Yang, X., Zhang, P., Reuter, T. and Stanford, K. (2019) 'Comparison of heating block and water bath methods to determine heat resistance in Shiga-toxin producing Escherichia coli with and without the locus of heat resistance', *J Microbiol Methods,* 164, pp. 105679.

De Vries, L. E., Christensen, H., Skov, R. L., Aarestrup, F. M. and Agersø, Y. (2009) 'Diversity of the tetracycline resistance gene tet (M) and identification of Tn 916-and Tn 5801-like (Tn 6014) transposons in Staphylococcus aureus from humans and animals', *Journal of Antimicrobial Chemotherapy,* 64(3), pp. 490-500.

Deckmyn, M. A. (2018) 'Package ‘mapdata’'.

Decruyenaere, A., Decruyenaere, P., Peeters, P., Vermassen, F., Dhaene, T. and Couckuyt, I. (2015) 'Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods', *BMC Medical Informatics and Decision Making,* 15(1), pp. 83.

Dego, O. K., Van Dijk, J. E. and Nederbragt, H. (2002) 'Factors involved in the early pathogenesis of bovine Staphylococcus aureus mastitis with emphasis on bacterial adhesion and invasion. A review', *Veterinary Quarterly,* 24(4), pp. 181-198.

Delannoy, S., Beutin, L. and Fach, P. (2012) 'Use of clustered regularly interspaced short palindromic repeat sequence polymorphisms for specific detection of enterohemorrhagic Escherichia coli strains of serotypes O26: H11, O45: H2, O103: H2, O111: H8, O121: H19, O145: H28, and O157: H7 by real-time PCR', *Journal of clinical microbiology,* 50(12), pp. 4035-4040.

Deluyker, H. A., Van Oye, S. N. and Boucher, J. F. (2005) 'Factors affecting cure and somatic cell count after pirlimycin treatment of subclinical mastitis in lactating cows', *Journal of Dairy Science,* 88(2), pp. 604-614.

DeMarco, M. L. and Ford, B. A. (2013) 'Beyond identification: emerging and future uses for MALDI-TOF mass spectrometry in the clinical microbiology laboratory', *Clin Lab Med,* 33(3), pp. 611-28.

Denis, E., Wieczorek, K. and Osek, J. (2014) 'Molecular characterization of non-0157 verotoxigenic Escherichia coli isolated from slaughtered cattle in Poland', *Medycyna Weterynaryjna,* 70(11).

Devriese, L. A., Hommez, J., Laevens, H., Pot, B., Vandamme, P. and Haesebrouck, F. (1999) 'Identification of aesculin-hydrolyzing streptococci, lactococci, aerococci and enterococci from subclinical intramammary infections in dairy cows', *Veterinary Microbiology,* 70(1), pp. 87-94.

Dezfulian, H., Batisson, I., Fairbrother, J. M., Lau, P. C., Nassar, A., Szatmari, G. and Harel, J. (2003) 'Presence and characterization of extraintestinal pathogenic Escherichia coli virulence genes in F165-positive E. coli strains isolated from diseased calves and pigs', *J Clin Microbiol,* 41(4), pp. 1375-85.

Diarra, M. S., Giguere, K., Malouin, F., Lefebvre, B., Bach, S., Delaquis, P., Aslam, M., Ziebell, K. A. and Roy, G. (2009) 'Genotype, serotype, and antibiotic resistance of sorbitol-negative Escherichia coli isolates from feedlot cattle', *Journal of food protection,* 72(1), pp. 28-36.

Dina, J., Malbruny, B. and Leclercq, R. (2003) 'Nonsense Mutations in the lsa-Like Gene in Enterococcus faecalis Isolates Susceptible to Lincosamides and Streptogramins A', *Antimicrobial Agents and Chemotherapy,* 47(7), pp. 2307.

Ding, S. and Li, S. 'PSO Parameters Optimization Based Support Vector Machines for Hyperspectral Classification'. *2009 First International Conference on Information Science and Engineering*, 26-28 Dec. 2009, 4066-4069.

Dinges, M. M., Orwin, P. M. and Schlievert, P. M. (2000) 'Exotoxins of Staphylococcus aureus', *Clinical microbiology reviews,* 13(1), pp. 16-34.

Dingwell, R. T., Leslie, K. E., Duffield, T. F., Schukken, Y. H., DesCoteaux, L., Keefe, G. P., Kelton, D. F., Lissemore, K. D., Shewfelt, W. and Dick, P. (2003) 'Efficacy of intramammary tilmicosin and risk factors for cure of Staphylococcus aureus infection in the dry period', *Journal of dairy science,* 86(1), pp. 159-168.

Dodd, F. H. and Jackson, E. R. (1971) *The control of bovine mastitis.* National Institute for Research in Dairying.

Doehring, C. and Sundrum, A. (2019) 'The informative value of an overview on antibiotic consumption, treatment efficacy and cost of clinical mastitis at farm level', *Preventive Veterinary Medicine,* 165, pp. 63-70.

Dogan, B., Klaessig, S., Rishniw, M., Almeida, R. A., Oliver, S. P., Simpson, K. and Schukken, Y. H. (2006) 'Adherent and invasive Escherichia coli are associated with persistent bovine mastitis', *Veterinary microbiology,* 116(4), pp. 270-282.

Dogan, B., Rishniw, M., Bruant, G., Harel, J., Schukken, Y. H. and Simpson, K. W. (2012) 'Phylogroup and lpfA influence epithelial invasion by mastitis associated Escherichia coli', *Vet Microbiol,* 159(1-2), pp. 163-70.

Dogan, B., Schukken, Y. H., Santisteban, C. and Boor, K. J. (2005) 'Distribution of serotypes and antimicrobial resistance genes among Streptococcus agalactiae isolates from bovine and human hosts', *Journal of clinical microbiology,* 43(12), pp. 5899-5906.

Donat, S., Streker, K., Schirmeister, T., Rakette, S., Stehle, T., Liebeke, M., Lalk, M. and Ohlsen, K. (2009) 'Transcriptome and functional analysis of the eukaryotic-type serine/threonine kinase PknB in Staphylococcus aureus', *Journal of bacteriology,* 191(13), pp. 4056-4069.

Dong, J. and Horvath, S. (2007) 'Understanding network concepts in modules', *BMC Systems Biology,* 1(1), pp. 24.

Dongen, S. (2000) 'A cluster algorithm for graphs'.

Dorn, M., e Silva, M. B., Buriol, L. S. and Lamb, L. C. (2014) 'Three-dimensional protein structure prediction: Methods and computational strategies', *Computational biology and chemistry,* 53, pp. 251-276.

Dosogne, H., Vangroenweghe, F. and Burvenich, C. (2002) 'Potential mechanism of action of J5 vaccine in protection against severe bovine coliform mastitis', *Veterinary research,* 33(1), pp. 1-12.

Douglas, V. L., Fenwick, S. G., Pfeiffer, D. U., Williamson, N. B. and Holmes, C. W. (2000) 'Genomic typing of Streptococcus uberis isolates from cases of mastitis, in New Zealand dairy cows, using pulsed-field gel electrophoresis', *Veterinary microbiology,* 75(1), pp. 27-41.

Down, P. M., Bradley, A. J., Breen, J. E., Hudson, C. D. and Green, M. J. (2016) 'Current management practices and interventions prioritised as part of a nationwide mastitis control plan', *Veterinary record,* 178(18), pp. 449-449.

Down, P. M., Green, M. J. and Hudson, C. D. 'Rate of transmission and the cost of clinical mastitis'. 2013, 15.

Drain, P. K., Bajema, K. L., Dowdy, D., Dheda, K., Naidoo, K., Schumacher, S. G., Ma, S., Meermeier, E., Lewinsohn, D. M. and Sherman, D. R. (2018) 'Incipient and subclinical tuberculosis: a clinical

review of early stages and progression of infection', *Clinical microbiology reviews,* 31(4), pp. e00021-18.

Du, Z., Yang, R., Guo, Z., Song, Y. and Wang, J. (2002) 'Identification of Staphylococcus aureus and Determination of Its Methicillin Resistance by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry', *Analytical Chemistry,* 74(21), pp. 5487-5491.

Ducheyne, E., Charlier, J., Vercruysse, J., Rinaldi, L., Biggeri, A., Demeler, J., Brandt, C., de Waal, T., Selemetas, N. and Höglund, J. (2015) 'Modelling the spatial distribution of Fasciola hepatica in dairy cattle in Europe', *Geospatial health,* 9(2), pp. 261-270.

Duda, R. O., Hart, P. E. and Stork, D. G. (2012) *Pattern classification.* John Wiley & Sons.

Duez, C., Hallut, S., Rhazi, N., Hubert, S., Amoroso, A., Bouillenne, F., Piette, A. and Coyette, J. (2004) 'The ponA gene of Enterococcus faecalis JH2-2 codes for a low-affinity class A penicillin-binding protein', *Journal of bacteriology,* 186(13), pp. 4412-4416.

Dufour, S., Dohoo, I. R., Barkema, H. W., DesCôteaux, L., DeVries, T. J., Reyher, K. K., Roy, J. P. and Scholl, D. T. (2012) 'Manageable risk factors associated with the lactational incidence, elimination, and prevalence of Staphylococcus aureus intramammary infections in dairy cows', *Journal of Dairy Science,* 95(3), pp. 1283-1300.

Döpfer, D., Barkema, H. W., Lam, T., Schukken, Y. H. and Gaastra, W. (1999) 'Recurrent clinical mastitis caused by Escherichia coli in dairy cows', *Journal of dairy science,* 82(1), pp. 80-85.

Eaton, T. J. and Gasson, M. J. (2001) 'Molecular screening of Enterococcusvirulence determinants and potential for genetic exchange between food and medical isolates', *Appl. Environ. Microbiol.,* 67(4), pp. 1628-1635.

Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E. and Petrovski, K. R. (2019) 'Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep learning and gradient-boosted trees outperform other models', *Computers in biology and medicine,* 114, pp. 103456.

Ebrahimie, E., Ebrahimi, F., Ebrahimi, M., Tomlinson, S. and Petrovski, K. R. (2018) 'Hierarchical pattern recognition in milking parameters predicts mastitis prevalence', *Computers and Electronics in Agriculture,* 147, pp. 6-11.

Edwards-Jones, V., Claydon, M. A., Evason, D. J., Walker, J., Fox, A. J. and Gordon, D. B. (2000) 'Rapid discrimination between methicillin-sensitive and methicillin-resistant Staphylococcus aureus by intact cell mass spectrometry', *Journal of medical microbiology,* 49(3), pp. 295-300.

Eijsink, V. G., Axelsson, L., Diep, D. B., Håvarstein, L. S., Holo, H. and Nes, I. F. (2002) 'Production of class II bacteriocins by lactic acid bacteria; an example of biological warfare and communication', *Antonie Van Leeuwenhoek,* 81(1-4), pp. 639-654.

Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. (2000) 'Protein function in the post-genomic era', *Nature,* 405(6788), pp. 823-826.

Eklund, M., Scheutz, F. and Siitonen, A. (2001) 'Clinical isolates of non-O157 Shiga toxin-producing Escherichia coli: serotypes, virulence characteristics, and molecular profiles of strains of the same serotype', *Journal of clinical microbiology,* 39(8), pp. 2829-2834.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A. and Smart, A. (2018) 'The Pfam protein families database in 2019', *Nucleic acids research,* 47(D1), pp. D427-D432.

Elhadidy, M. and Elsayyad, A. (2013) 'Uncommitted role of enterococcal surface protein, Esp, and origin of isolates on biofilm production by Enterococcus faecalis isolated from bovine mastitis', *Journal of Microbiology, Immunology and Infection,* 46(2), pp. 80-84.

Elhadidy, M. and Zahran, E. (2014) 'Biofilm mediates E nterococcus faecalis adhesion, invasion and survival into bovine mammary epithelial cells', *Letters in applied microbiology,* 58(3), pp. 248-254.

EMA/AMEG (2019) *Categorisation of antibiotics in the European Union*. Available at: https://www.ema.europa.eu/en/documents/report/categorisation-antibiotics-european-union-answer-request-european-commission-updating-scientific_en.pdf.

Emele, M. F., Karg, M., Hotzel, H., Graaf-van Bloois, L., Groß, U., Bader, O. and Zautner, A. E. (2019a) 'Differentiation of Campylobacter fetus subspecies by proteotyping', *European Journal of Microbiology and Immunology,* 9(2), pp. 62-71.

Emele, M. F., Možina, S. S., Lugert, R., Bohne, W., Masanta, W. O., Riedel, T., Groß, U., Bader, O. and Zautner, A. E. (2019b) 'Proteotyping as alternate typing method to differentiate Campylobacter coli clades', *Scientific reports,* 9(1), pp. 1-11.

Emerson, D., Agulto, L., Liu, H. and Liu, L. (2008) 'Identifying and Characterizing Bacteria in an Era of Genomics and Proteomics', *BioScience,* 58(10), pp. 925-936.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. (1999) 'Protein interaction maps for complete genomes based on gene fusion events', *Nature,* 402(6757), pp. 86-90.

Enright, M. C., Day, N. P. J., Davies, C. E., Peacock, S. J. and Spratt, B. G. (2000) 'Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones ofStaphylococcus aureus', *Journal of clinical microbiology,* 38(3), pp. 1008-1015.

Erbas, G., Parin, U., Turkyilmaz, S., Ucan, N., Ozturk, M. and Kaya, O. (2016) 'Distribution of antibiotic resistance genes in Enterococcus spp. isolated from mastitis bovine milk', *Acta Veterinaria,* 66(3), pp. 336-346.

Erskine, R. J., Wagner, S. and DeGraves, F. J. (2003) 'Mastitis therapy and pharmacology', *The Veterinary clinics of North America. Food animal practice,* 19(1), pp. 109-38.

Esener, N., Green, M. J., Emes, R. D., Jowett, B., Davies, P. L., Bradley, A. J. and Dottorini, T. (2018) 'Discrimination of contagious and environmental strains of Streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning', *Scientific Reports,* 8(1), pp. 17517.

Esener, N., Guerra, A. M., Giebel, K., Lea, D., Green, M. J., Bradley, A. J. and Dottorini, T. (2021) 'Mass spectrometry and machine learning for the accurate diagnosis of benzylpenicillin and multidrug resistance of Staphylococcus aureus in bovine mastitis', *PLOS Computational Biology,* 17(6), pp. e1009108.

EUCAST (2019) *The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters, version 9.0, 2019*. Available at: http://www.eucast.org/clinical_breakpoints/ (Accessed: 2019.

Faber, F. and Bäumler, A. J. (2014) 'The impact of intestinal inflammation on the nutritional environment of the gut microbiota', *Immunology letters,* 162(2), pp. 48-53.

Fagerquist, C. K., Garbus, B. R., Miller, W. G., Williams, K. E., Yee, E., Bates, A. H., Boyle, S., Harden, L. A., Cooley, M. B. and Mandrell, R. E. (2010) 'Rapid Identification of Protein Biomarkers of Escherichia coli O157:H7 by Matrix-Assisted Laser Desorption Ionization-Time-of-Flight−Time-of-Flight Mass Spectrometry and Top-Down Proteomics', *Analytical Chemistry,* 82(7), pp. 2717-2725.

Fairbrother, J.-H., Dufour, S., Fairbrother, J. M., Francoz, D., Nadeau, É. and Messier, S. (2015) 'Characterization of persistent and transient Escherichia coli isolates recovered from clinical mastitis episodes in dairy cows', *Veterinary microbiology,* 176(1-2), pp. 126-133.

Falgenhauer, L., Waezsada, S.-E., Gwozdzinski, K., Ghosh, H., Doijad, S., Bunk, B., Spröer, C., Imirzalioglu, C., Seifert, H., Irrgang, A., Fischer, J., Guerra, B., Käsbohrer, A., Overmann, J., Goesmann, A. and Chakraborty, T. (2016) 'Chromosomal Locations of mcr-1 and bla CTX-M-15 in Fluoroquinolone-Resistant Escherichia coli ST410', *Emerging infectious diseases,* 22(9), pp. 1689-1691.

Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., Kidd, M., Blaser, M. J., Graham, D. Y., Vacher, S. and Perez-Perez, G. I. (2003) 'Traces of human migrations in Helicobacter pylori populations', *science,* 299(5612), pp. 1582-1585.

Fan, X., Tang, X., Yan, J. and Xie, J. (2014) 'Identification of idiosyncratic Mycobacterium tuberculosis ribosomal protein subunits with implications in extraribosomal function, persistence, and drug resistance based on transcriptome data', *Journal of Biomolecular Structure and Dynamics,* 32(10), pp. 1546-1551.

Fang, R., Cui, J., Cui, T., Guo, H., Ono, H. K., Park, C.-H., Okamura, M., Nakane, A. and Hu, D.-L. (2019) 'Staphylococcal enterotoxin C is an important virulence factor for mastitis', *Toxins,* 11(3), pp. 141.

Feldmann, F., Sorsa, L. J., Hildinger, K. and Schubert, S. (2007) 'The salmochelin siderophore receptor IroN contributes to invasion of urothelial cells by extraintestinal pathogenic Escherichia coli in vitro', *Infect Immun,* 75(6), pp. 3183-7.

Feng, P., Delannoy, S., Lacher, D. W., Bosilevac, J. M. and Fach, P. (2017) 'Characterization and Virulence Potential of Serogroup O113 Shiga Toxin–Producing Escherichia coli Strains Isolated from Beef and Cattle in the United States', *Journal of Food Protection,* 80(3), pp. 383-391.

Feng, Y., Qi, W., Wang, X.-r., Ling, W., Li, X.-p., Luo, J.-y., Zhang, S.-d. and Li, H.-s. (2016) 'Genetic characterization of antimicrobial resistance in Staphylococcus aureus isolated from bovine mastitis cases in Northwest China', *Journal of integrative agriculture,* 15(12), pp. 2842-2847.

Fenlon, C., O'Grady, L., Dunnion, J., Shalloo, L., Butler, S. T. and Doherty, M. L. (2016) 'A comparison of machine learning techniques for predicting insemination outcome in Irish dairy cows'.

Fenlon, C., O'Grady, L., Mee, J. F., Butler, S. T., Doherty, M. L. and Dunnion, J. (2017) 'A comparison of 4 predictive models of calving assistance and difficulty in dairy heifers and cows', *Journal of Dairy Science,* 100(12), pp. 9746-9758.

Ferdous, M., Friedrich, A. W., Grundmann, H., de Boer, R. F., Croughs, P. D., Islam, M. A., Kluytmans-van den Bergh, M. F. Q., Kooistra-Smid, A. M. D. and Rossen, J. W. A. (2016) 'Molecular characterization and phylogeny of Shiga toxin–producing Escherichia coli isolates obtained from two

Dutch regions using whole genome sequencing', *Clinical Microbiology and Infection,* 22(7), pp. 642-e1.

Fernandes, J. B., Zanardo, L. G., Galvao, N. N., Carvalho, I. A., Nero, L. A. and Moreira, M. A. (2011) 'Escherichia coli from clinical mastitis: serotypes and virulence factors', *J Vet Diagn Invest,* 23(6), pp. 1146-52.

Fernandes, M. R., Sellera, F. P., Moura, Q., Esposito, F., Sabino, C. P. and Lincopan, N. (2020) 'Identification and genomic features of halotolerant extended-spectrum-β-lactamase (CTX-M)-producing Escherichia coli in urban-impacted coastal waters, Southeast Brazil', *Marine Pollution Bulletin,* 150, pp. 110689.

Fernández, D., Irino, K., Sanz, M. E., Padola, N. L. and Parma, A. E. (2010) 'Characterization of Shiga toxin-producing Escherichia coli isolated from dairy cows in Argentina', *Letters in Applied Microbiology,* 51(4), pp. 377-382.

Fernández, L., Breidenstein, E. B. M., Song, D. and Hancock, R. E. W. (2012) 'Role of intracellular proteases in the antibiotic resistance, motility, and biofilm formation of Pseudomonas aeruginosa', *Antimicrobial agents and chemotherapy,* 56(2), pp. 1128-1132.

Field, T. R., Ward, P. N., Pedersen, L. H. and Leigh, J. A. (2003) 'The hyaluronic acid capsule of Streptococcus uberis is not required for the development of infection and clinical mastitis', *Infect Immun,* 71(1), pp. 132-9.

Fillingame, R. H., Oldenburg, M. and Fraga, D. (1991) 'Mutation of alanine 24 to serine in subunit c of the Escherichia coli F1F0-ATP synthase reduces reactivity of aspartyl 61 with dicyclohexylcarbodiimide', *Journal of Biological Chemistry,* 266(31), pp. 20934-20939.

Finch, J. M., Hill, A. W., Field, T. R. and Leigh, J. A. (1994) 'Local vaccination with killed Streptococcus uberis protects the bovine mammary gland against experimental intramammary challenge with the homologous strain', *Infection and Immunity,* 62(9), pp. 3599.

Finch, J. M., Winter, A., Walton, A. W. and Leigh, J. A. (1997) 'Further studies on the efficacy of a live vaccine against mastitis caused by Streptococcus uberis', *Vaccine,* 15(10), pp. 1138-1143.

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S. and Stoeckert Jr, C. J. (2011) 'Using OrthoMCL to Assign Proteins to OrthoMCL-DB Groups or to Cluster Proteomes Into New Ortholog Groups', *Current Protocols in Bioinformatics,* 35(1), pp. 6.12.1-6.12.19.

Fisher, R. A. (1936) 'The use of multiple measurements in taxonomic problems', *Annals of eugenics,* 7(2), pp. 179-188.

Fitzgerald, J., Meaney, W., Hartigan, P., Smyth, C. J. and Kapur, V. (1997) 'Fine structure molecular epidemiology analysis of Staphylococcus aureus recovered from cows', *Epidemiology and infection,* 119, pp. 261-9.

Forsyth, R., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., Wall, D., Wang, L., Brown-Driver, V. and Froelich, J. M. (2002) 'A genome-wide strategy for the identification of essential genes in Staphylococcus aureus', *Molecular microbiology,* 43(6), pp. 1387-1400.

Foster, T. J. (2017) 'Antibiotic resistance in Staphylococcus aureus. Current status and future prospects', *FEMS Microbiology Reviews,* 41(3), pp. 430-449.

Fournier, C., Kuhnert, P., Frey, J., Miserez, R., Kirchhofer, M., Kaufmann, T., Steiner, A. and Graber, H. U. (2008) 'Bovine Staphylococcus aureus: Association of virulence genes, genotypes and clinical outcome', *Research in Veterinary Science,* 85(3), pp. 439-448.

Fox, L. K. and Gay, J. M. (1993) 'Contagious mastitis', *Veterinary Clinics of North America: Food Animal Practice,* 9(3), pp. 475-487.

Francisco, A. P., Bugalho, M., Ramirez, M. and Carriço, J. A. (2009) 'Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach', *BMC bioinformatics,* 10(1), pp. 152.

Franz, C. M., Huch, M., Abriouel, H., Holzapfel, W. and Gálvez, A. (2011) 'Enterococci as probiotics and their implications in food safety', *Int J Food Microbiol,* 151(2), pp. 125-40.

Frazzon, A. P. G., Gama, B. A., Hermes, V., Bierhals, C. G., Pereira, R. I., Guedes, A. G., d'Azevedo, P. A. and Frazzon, J. (2010) 'Prevalence of antimicrobial resistance and molecular characterization of tetracycline resistance mediated by tet(M) and tet(L) genes in Enterococcus spp. isolated from food in Southern Brazil', *World Journal of Microbiology and Biotechnology,* 26(2), pp. 365-370.

Freitag, C., Michael, G. B., Kadlec, K., Hassel, M. and Schwarz, S. (2017) 'Detection of plasmid-borne extended-spectrum β-lactamase (ESBL) genes in Escherichia coli isolates from bovine mastitis', *Veterinary Microbiology,* 200, pp. 151-156.

Fremaux, B., Raynaud, S., Beutin, L. and Rozand, C. V. (2006) 'Dissemination and persistence of Shiga toxin-producing Escherichia coli (STEC) strains on French dairy farms', *Veterinary microbiology,* 117(2-4), pp. 180-191.

Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C. and Meinnel, T. (2006) 'The proteomics of N-terminal methionine cleavage', *Mol Cell Proteomics,* 5(12), pp. 2336-49.

Fröhlicher, E., Krause, G., Zweifel, C., Beutin, L. and Stephan, R. (2008) 'Characterization of attaching and effacing Escherichia coli (AEEC) isolated from pigs and sheep', *BMC microbiology,* 8(1), pp. 144.

Fursova, K., Sorokin, A., Sokolov, S., Dzhelyadin, T., Shulcheva, I., Shchannikova, M., Nikanova, D., Artem'eva, O., Zinovieva, N. and Brovko, F. (2020) 'Virulence Factors and Phylogeny of Staphylococcus aureus Associated With Bovine Mastitis in Russia Based on Genome Sequences', *Frontiers in Veterinary Science,* 7, pp. 135.

Gagarinova, A., Stewart, G., Samanfar, B., Phanse, S., White, C. A., Aoki, H., Deineko, V., Beloglazova, N., Yakunin, A. F., Golshani, A., Brown, E. D., Babu, M. and Emili, A. (2016) 'Systematic Genetic Screens Reveal the Dynamic Global Functional Organization of the Bacterial Translation Machinery', *Cell Rep,* 17(3), pp. 904-916.

Gagnon, M., Hamelin, L., Fréchette, A., Dufour, S. and Roy, D. (2020) 'Effect of recycled manure solids as bedding on bulk tank milk and implications for cheese microbiological quality', *Journal of Dairy Science,* 103(1), pp. 128-140.

Gaillot, O., Blondiaux, N., Loïez, C., Wallet, F., Lemaître, N., Herwegh, S. and Courcol, R. J. (2011) 'Cost-effectiveness of switch to matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine bacterial identification', *Journal of clinical microbiology,* 49(12), pp. 4412-4412.

Gao, J., Ferreri, M., Liu, X. Q., Chen, L. B., Su, J. L. and Han, B. (2011) 'Development of multiplex polymerase chain reaction assay for rapid detection of Staphylococcus aureus and selected antibiotic

resistance genes in bovine mastitic milk samples', *Journal of Veterinary Diagnostic Investigation,* 23(5), pp. 894-901.

Gao, R. and Stock, A. M. (2009) 'Biological insights from structures of two-component proteins', *Annu Rev Microbiol,* 63, pp. 133-54.

Gao, X., Fan, C., Zhang, Z., Li, S., Xu, C., Zhao, Y., Han, L., Zhang, D. and Liu, M. (2019) 'Enterococcal isolates from bovine subclinical and clinical mastitis: Antimicrobial resistance and integron-gene cassette distribution', *Microbial Pathogenesis,* 129, pp. 82-87.

García-Solache, M. and Rice, L. B. (2016) 'Genome Sequence of the Multiantibiotic-Resistant Enterococcus faecium Strain C68 and Insights on the pLRM23 Colonization Plasmid', *Genome Announc,* 4(3).

García-Álvarez, L., Holden, M. T. G., Lindsay, H., Webb, C. R., Brown, D. F. J., Curran, M. D., Walpole, E., Brooks, K., Pickard, D. J. and Teale, C. (2011) 'Meticillin-resistant Staphylococcus aureus with a novel mecA homologue in human and bovine populations in the UK and Denmark: a descriptive study', *The Lancet infectious diseases,* 11(8), pp. 595-603.

Gardete, S. and Tomasz, A. (2014) 'Mechanisms of vancomycin resistance in Staphylococcus aureus', *The Journal of Clinical Investigation,* 124(7), pp. 2836-2840.

Gardner, S. N. and Hall, B. G. (2013) 'When Whole-Genome Alignments Just Won't Work: kSNP v2 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial Genomes', *PLOS ONE,* 8(12), pp. e81760.

Garrido, A. M., Gálvez, A. and Pulido, R. P. (2014) 'Antimicrobial resistance in enterococci', *Journal of Infectious Diseases and Therapy*.

Gatermann, S. G., Koschinski, T. and Friedrich, S. (2007) 'Distribution and expression of macrolide resistance genes in coagulase-negative staphylococci', *Clinical microbiology and infection,* 13(8), pp. 777-781.

Gaudet, P., Škunca, N., Hu, J. C. and Dessimoz, C. (2017) 'Primer on the gene ontology', *The Gene Ontology Handbook*: Humana Press, New York, NY, pp. 25-37.

Gazzola, S., Fontana, C., Bassi, D. and Cocconcelli, P. S. (2012) 'Assessment of tetracycline and erythromycin resistance transfer during sausage fermentation by culture-dependent and -independent methods', *Food Microbiology,* 30(2), pp. 348-354.

Ge, H., Liu, Z., Church, G. M. and Vidal, M. (2001) 'Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae', *Nature genetics,* 29(4), pp. 482-486.

Gekenidis, M.-T., Studer, P., Wüthrich, S., Brunisholz, R. and Drissner, D. (2014) 'Beyond the matrix-assisted laser desorption ionization (MALDI) biotyping workflow: in search of microorganism-specific tryptic peptides enabling discrimination of subspecies', *Applied and environmental microbiology,* 80(14), pp. 4234-4241.

Gelsomino, R., Vancanneyt, M., Condon, S., Swings, J. and Cogan, T. M. (2001) 'Enterococcal diversity in the environment of an Irish Cheddar-type cheesemaking factory', *International journal of food microbiology,* 71(2-3), pp. 177-188.

Gentilini, E., Denamiel, G., Llorente, P., Godaly, S., Rebuelto, M. and DeGregorio, O. (2000) 'Antimicrobial susceptibility of Staphylococcus aureus isolated from bovine mastitis in Argentina', *Journal of dairy science,* 83(6), pp. 1224-1227.

Genuer, R., Poggi, J.-M. and Tuleau, C. (2008) 'Random Forests: some methodological insights', *arXiv preprint arXiv:0811.3619*.

Georrge, J. J. and Umrania, V. V. (2012) 'Subtractive Genomics Approach to Identify Putative Drug Targets and Identification of Drug-like Molecules for Beta Subunit of DNA Polymerase III in Streptococcus Species', *Applied Biochemistry and Biotechnology,* 167(5), pp. 1377-1395.

Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytiä-Trees, E., Ribot, E. M. and Swaminathan, B. (2006) 'PulseNet USA: A Five-Year Update', *Foodborne Pathogens and Disease,* 3(1), pp. 9-19.

Gerritsen, H. (2014) 'mapplots: Data visualisation on maps', *R package version,* 1(1).

Ghanbarpour, R. and Oswald, E. (2010) 'Phylogenetic distribution of virulence genes in Escherichia coli isolated from bovine mastitis in Iran', *Research in veterinary science,* 88(1), pp. 6-10.

Gibreel, T. M., Dodgson, A. R., Cheesbrough, J., Fox, A. J., Bolton, F. J. and Upton, M. (2010) 'Population structure, virulence potential and antibiotic susceptibility of uropathogenic Escherichia coli from Northwest England', *Journal of Antimicrobial Chemotherapy,* 67(2), pp. 346-356.

Gil, R., Silva, F. J., Peretó, J. and Moya, A. (2004) 'Determination of the core of a minimal bacterial gene set', *Microbiology and Molecular Biology Reviews,* 68(3), pp. 518-537.

Gilbert, F. B., Fromageau, A., Gélineau, L. and Poutrel, B. (2006a) 'Differentiation of bovine Staphylococcus aureus isolates by use of polymorphic tandem repeat typing', *Veterinary microbiology,* 117(2-4), pp. 297-303.

Gilbert, F. B., Fromageau, A., Lamoureux, J. and Poutrel, B. (2006b) 'Evaluation of tandem repeats for MLVA typing of Streptococcus uberis isolated from bovine mastitis', *BMC Veterinary Research,* 2(1), pp. 33.

Gilchrist, T. L., Smith, D. G. E., Fitzpatrick, J. L., Zadoks, R. N. and Fontaine, M. C. (2013) 'Comparative molecular analysis of ovine and bovine Streptococcus uberis isolates', *Journal of dairy science,* 96(2), pp. 962-970.

Giraffa, G. (2003) 'Functionality of enterococci in dairy products', *International Journal of Food Microbiology,* 88(2), pp. 215-222.

Giraffa, G., Carminati, D. and Neviani, E. (1997) 'Enterococci isolated from dairy products: a review of risks and potential technological use', *Journal of Food Protection,* 60(6), pp. 732-738.

Goldberg, D. E. and Holland, J. H. (1988) 'Genetic algorithms and machine learning', *Machine learning,* 3(2), pp. 95-99.

Gomes, T. A. T., Irino, K., Girão, D. M., Girão, V. B. C., Guth, B. E. C., Vaz, T. M. I., Moreira, F. C., Chinarelli, S. H. and Vieira, M. A. M. (2004) 'Emerging enteropathogenic Escherichia coli strains?', *Emerging infectious diseases,* 10(10), pp. 1851-1855.

Gomez, J. E., Kaufmann-Malaga, B. B., Wivagg, C. N., Kim, P. B., Silvis, M. R., Renedo, N., Ioerger, T. R., Ahmad, R., Livny, J., Fishbein, S., Sacchettini, J. C., Carr, S. A. and Hung, D. T. (2017) 'Ribosomal

mutations promote the evolution of antibiotic resistance in a multidrug environment', *eLife,* 6, pp. e20420.

Gonano, M. and Winter, P. 'Phenotypic and molecular identification of Streptococcus species isolated from milk of intramammary infected dairy cows in Austria'. 2008, 191-198.

Gonzalez, R. N., Cullor, J. S., Jasper, D. E., Farver, T. B., Bushnell, R. B. and Oliver, M. N. (1989) 'Prevention of clinical coliform mastitis in dairy cows by a mutant Escherichia coli vaccine', *Canadian Journal of Veterinary Research,* 53(3), pp. 301.

Gonzalez-Escalona, N., Toro, M., Rump, L. V., Cao, G., Nagaraja, T. G. and Meng, J. (2016) 'Virulence Gene Profiles and Clonal Relationships of Escherichia coli O26:H11 Isolates from Feedlot Cattle as Determined by Whole-Genome Sequencing', *Applied and environmental microbiology,* 82(13), pp. 3900-3912.

Gonçalves, J. L., Tomazi, T., Barreiro, J. R., Braga, P. A. d. C., Ferreira, C. R., Araújo Junior, J. P., Eberlin, M. N. and dos Santos, M. V. (2014) 'Identification of Corynebacterium spp. isolated from bovine intramammary infections by matrix-assisted laser desorption ionization-time of flight mass spectrometry', *Veterinary Microbiology,* 173(1-2), pp. 147-151.

Gordon, D. M. (2013) 'Chapter 1 - The ecology of Escherichia coli', in Donnenberg, M.S. (ed.) *Escherichia coli (Second Edition)*. Boston: Academic Press, pp. 3-20.

Gordon, D. M. and Cowling, A. (2003) 'The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects', *Microbiology,* 149(12), pp. 3575-3586.

Gotoh, Y., Eguchi, Y., Watanabe, T., Okamoto, S., Doi, A. and Utsumi, R. (2010) 'Two-component signal transduction as potential drug targets in pathogenic bacteria', *Curr Opin Microbiol,* 13(2), pp. 232-9.

Govindaraj, G., Hiremath, J., Reddy, G. B. M., Siju, S. J., Yogishardhaya, R. and Prajapati, A. (2019) 'ICAR NIVEDI Annual Report 2018-19'.

Grami, R., Dahmen, S., Mansour, W., Mehri, W., Haenni, M., Aouni, M. and Madec, J.-Y. (2014) 'blaCTX-M-15-Carrying F2:A-:B- Plasmid in Escherichia coli from Cattle Milk in Tunisia', *Microbial Drug Resistance,* 20(4), pp. 344-349.

Grant, M. A., Weagant, S. D. and Feng, P. (2001) 'Glutamate decarboxylase genes as a prescreening marker for detection of pathogenic Escherichia coli groups', *Applied and environmental microbiology,* 67(7), pp. 3110-3114.

Grant, R. G. and Finch, J. M. (1997) 'Phagocytosis of Streptococcus uberis by bovine mammary gland macrophages', *Research in veterinary science,* 62(1), pp. 74-78.

Green, M., Breen, J., Hudson, C., Black, H., Cross, K. and Bradley, A. (2012) 'The DairyCo mastitis control plan: A nationwide scheme for mastitis control in Great Britain', *Cattle Practice,* 20, pp. 65-71.

Green, M., Huxley, J., Madouasse, A., Browne, W., Medley, G., Bradley, A., Biggs, A., Breen, J., Burnell, M. and Hayton, A. (2008) 'Making good decisions on dry cow management to improve udder health–Synthesising evidence in a Bayesian framework', *Cattle practice: journal of the British Cattle Veterinary Association,* 16, pp. 200.

Green, M. J., Green, L. E., Schukken, Y. H., Bradley, A. J., Peeler, E. J., Barkema, H. W., De Haas, Y., Collis, V. J. and Medley, G. F. (2004) 'Somatic cell count distributions during lactation predict clinical mastitis', *Journal of dairy science,* 87(5), pp. 1256-1264.

Green, M. J., Leach, K. A., Breen, J. E., Green, L. E. and Bradley, A. J. (2007) 'National intervention study of mastitis control in dairy herds in England and Wales', *The Veterinary Record,* 160(9), pp. 287-293.

Griffin, P. M., Price, G. R., Schooneveldt, J. M., Schlebusch, S., Tilse, M. H., Urbanski, T., Hamilton, B. and Venter, D. (2012) 'Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify vancomycin-resistant enterococci and investigate the epidemiology of an outbreak', *Journal of clinical microbiology,* 50(9), pp. 2918-2931.

Grzesiak, W., Zaborski, D., Sablik, P., Żukiewicz, A., Dybus, A. and Szatkowska, I. (2010) 'Detection of cows with insemination problems using selected classification models', *Computers and electronics in agriculture,* 74(2), pp. 265-273.

Gröhn, Y. T., González, R. N., Wilson, D. J., Hertl, J. A., Bennett, G., Schulte, H. and Schukken, Y. H. (2005) 'Effect of pathogen-specific clinical mastitis on herd life in two New York State dairy herds', *Preventive Veterinary Medicine,* 71(1), pp. 105-125.

Gröhn, Y. T., Wilson, D. J., González, R. N., Hertl, J. A., Schulte, H., Bennett, G. and Schukken, Y. H. (2004) 'Effect of Pathogen-Specific Clinical Mastitis on Milk Yield in Dairy Cows', *Journal of Dairy Science,* 87(10), pp. 3358-3374.

Guardabassi, L. and Courvalin, P. (2006) 'Modes of antimicrobial action and mechanisms of bacterial resistance', *Antimicrobial resistance in bacteria of animal origin*: American Society of Microbiology, pp. 1-18.

Guex, N., Peitsch, M. C. and Schwede, T. (2009) 'Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective', *ELECTROPHORESIS,* 30(S1), pp. S162-S173.

Gurjar, A., Gioia, G., Schukken, Y., Welcome, F., Zadoks, R. and Moroni, P. (2012) 'Molecular diagnostics applied to mastitis problems on dairy farms', *Veterinary Clinics: Food Animal Practice,* 28(3), pp. 565-576.

Gurjar, A. A., Klaessig, S., Salmon, S. A., Yancey Jr, R. J. and Schukken, Y. H. (2013) 'Evaluation of an alternative dosing regimen of a J-5 mastitis vaccine against intramammary Escherichia coli challenge in nonlactating late-gestation dairy cows', *Journal of dairy science,* 96(8), pp. 5053-5063.

Guzman-Hernandez, R., Contreras-Rodriguez, A., Hernandez-Velez, R., Perez-Martinez, I., Lopez-Merino, A., Zaidi, M. B. and Estrada-Garcia, T. (2016) 'Mexican unpasteurised fresh cheeses are contaminated with Salmonella spp., non-O157 Shiga toxin producing Escherichia coli and potential uropathogenic E. coli strains: A public health risk', *International Journal of Food Microbiology,* 237, pp. 10-16.

Géron, A. (2019) *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media.

Günther, J., Petzl, W., Bauer, I., Ponsuksili, S., Zerbe, H., Schuberth, H.-J., Brunner, R. M. and Seyfert, H.-M. (2017) 'Differentiating Staphylococcus aureus from Escherichia coli mastitis: S. aureus triggers

unbalanced immune-dampening and host cell invasion immediately after udder infection', *Scientific Reports,* 7(1), pp. 4811.

Gürler, H., Findik, A., Gültiken, N., Ay Serhan, S., Çiftçi, A., Koldaş, E., Arslan, S. and Findik, M. (2015) 'Investigation On The Etiology Of Subclinical Mastitis In Jersey And Hybrid Jersey Dairy Cows', *Acta Veterinaria,* 65(3), pp. 358-370.

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M. and Harris, S. R. (2017) 'Phandango: an interactive viewer for bacterial population genomics', *Bioinformatics,* 34(2), pp. 292-293.

Hadjadj, L., Baron, S. A., Diene, S. M. and Rolain, J.-M. (2019) 'How to discover new antibiotic resistance genes?', *Expert Review of Molecular Diagnostics,* 19(4), pp. 349-362.

Hammer, B., Strickert, M. and Villmann, T. (2005) 'Supervised neural gas with general similarity measure', *Neural Processing Letters,* 21(1), pp. 21-44.

Hamzah, A. M. and Kadim, H. K. (2018) 'Isolation and identification of Enterococcus faecalis from cow milk samples and vaginal swab from human', *Entomol Zool Sci,* 6, pp. 218-222.

Hanna, R., Aleksandra, L.-P., Maria, K. and Marcin, W. (2019) 'Occurrence of enterococci in mastitic cow's milk and their antimicrobial resistance', *Journal of Veterinary Research,* 63(1), pp. 93-97.

Hart, P. (1968) 'The condensed nearest neighbor rule (Corresp.)', *IEEE transactions on information theory,* 14(3), pp. 515-516.

Hastie, T., Rosset, S., Zhu, J. and Zou, H. (2009) 'Multi-class adaboost', *Statistics and its Interface,* 2(3), pp. 349-360.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Haubert, L., Kroning, I. S., Iglesias, M. A. and da Silva, W. P. (2017) 'First report of the Staphylococcus aureus isolate from subclinical bovine mastitis in the South of Brazil harboring resistance gene dfrG and transposon family Tn916-1545', *Microbial Pathogenesis,* 113, pp. 242-247.

Hava, D. L. and Camilli, A. (2002) 'Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors', *Mol Microbiol,* 45(5), pp. 1389-406.

Haveri, M., Roslöf, A., Rantala, L. and Pyörälä, S. (2007) 'Virulence genes of bovine Staphylococcus aureus from persistent and nonpersistent intramammary infections with different clinical characteristics', *Journal of Applied Microbiology,* 103(4), pp. 993-1000.

Haveri, M., Taponen, S., Vuopio-Varkila, J., Salmenlinna, S. and Pyörälä, S. (2005) 'Bacterial genotype affects the manifestation and persistence of bovine Staphylococcus aureus intramammary infection', *Journal of clinical microbiology,* 43(2), pp. 959-961.

He, H., Bai, Y., Garcia, E. A. and Li, S. (2008) 'ADASYN: Adaptive synthetic sampling approach for imbalanced learning'. 2008: IEEE, 1322-1328.

He, H. and Garcia, E. A. (2009) 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering,* 21(9), pp. 1263-1284.

Heald, C. W., Kim, T., Sischo, W. M., Cooper, J. B. and Wolfgang, D. R. (2000) 'A Computerized Mastitis Decision Aid Using Farm-Based Records: An Artificial Neural Network Approach', *Journal of Dairy Science,* 83(4), pp. 711-720.

Heikkilä, A. M., Liski, E., Pyörälä, S. and Taponen, S. (2018) 'Pathogen-specific production losses in bovine mastitis', *Journal of dairy science,* 101(10), pp. 9493-9504.

Hempstalk, K., McParland, S. and Berry, D. P. (2015) 'Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows', *Journal of Dairy Science,* 98(8), pp. 5262-5273.

Hershberger, E., Oprea, S. F., Donabedian, S. M., Perri, M., Bozigar, P., Bartlett, P. and Zervos, M. J. (2005) 'Epidemiology of antimicrobial resistance in enterococci of animal origin', *Journal of Antimicrobial Chemotherapy,* 55(1), pp. 127-130.

Hillenkamp, F., Karas, M., Beavis, R. C. and Chait, B. T. (1991) 'Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers', *Anal Chem,* 63(24), pp. 1193a-1203a.

HM Government (2019) *Tackling antimicrobial resistance 2019–2024*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784894/UK_AMR_5_year_national_action_plan.pdf (Accessed: December, 2020.

Ho, P.-L., Ng, K.-Y., Lo, W.-U., Law, P. Y., Lai, E. L.-Y., Wang, Y. and Chow, K.-H. (2015) 'Plasmid-Mediated OqxAB Is an Important Mechanism for Nitrofurantoin Resistance in Escherichia coli', *Antimicrobial agents and chemotherapy,* 60(1), pp. 537-543.

Hobson, M. J. and Berger, J. M. (2019) 'Caught in the Open: A Domain Insertion of M. tuberculosis Gyrase Suppresses ATPase Dimerization', *Structure,* 27(4), pp. 561-563.

Hoe, F. G. H. and Ruegg, P. L. (2005) 'Relationship between antimicrobial susceptibility of clinical mastitis pathogens and treatment outcome in cows', *Journal of the American Veterinary Medical Association,* 227(9), pp. 1461-1468.

Hoekstra, J., Rutten, V., Sommeling, L., van Werven, T., Spaninks, M., Duim, B., Benedictus, L. and Koop, G. (2018) 'High Production of LukMF' in Staphylococcus aureus Field Strains Is Associated with Clinical Bovine Mastitis', *Toxins,* 10(5), pp. 200.

Hogan, J. S., Smith, K. L., Hoblet, K. H., Schoenberger, P. S., Todhunter, D. A., Hueston, W. D., Pritchard, D. E., Bowman, G. L., Heider, L. E. and Brockett, B. L. (1989) 'Field survey of clinical mastitis in low somatic cell count herds', *Journal of Dairy Science,* 72(6), pp. 1547-1556.

Holden, M. T., Feil, E. J., Lindsay, J. A., Peacock, S. J., Day, N. P., Enright, M. C., Foster, T. J., Moore, C. E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S. D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., Doggett, J., Dowd, L., Feltwell, T., Hance, Z., Harris, B., Hauser, H., Holroyd, S., Jagels, K., James, K. D., Lennard, N., Line, A., Mayes, R., Moule, S., Mungall, K., Ormond, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Sanders, M., Sharp, S., Simmonds, M., Stevens, K., Whitehead, S., Barrell, B. G., Spratt, B. G. and Parkhill, J. (2004) 'Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance', *Proc Natl Acad Sci U S A,* 101(26), pp. 9786-91.

Holland, J. H. (1975) 'Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence', *Ann Arbor, MI: University of Michigan Press*, pp. 439-444.

Holland, R. D., Wilkes, J. G., Rafii, F., Sutherland, J. B., Persons, C. C., Voorhees, K. J. and Lay, J. O. (1996) 'Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry', *Rapid Communications in Mass Spectrometry,* 10(10), pp. 1227-1232.

Hollenbeck, B. L. and Rice, L. B. (2012) 'Intrinsic and acquired resistance mechanisms in enterococcus', *Virulence,* 3(5), pp. 421-569.

Holliday, G. L., Davidson, R., Akiva, E. and Babbitt, P. C. (2017) 'Evaluating functional annotations of enzymes using the gene ontology', *The Gene Ontology Handbook*: Humana Press, New York, NY, pp. 111-132.

Homan, W. L., Tribe, D., Poznanski, S., Li, M., Hogg, G., Spalburg, E., Van Embden, J. D. and Willems, R. J. (2002) 'Multilocus sequence typing scheme for Enterococcus faecium', *J Clin Microbiol,* 40(6), pp. 1963-71.

Hooper, D. C. and Jacoby, G. A. (2016) 'Topoisomerase Inhibitors: Fluoroquinolone Mechanisms of Action and Resistance', *Cold Spring Harbor perspectives in medicine,* 6(9), pp. a025320.

Hornitzky, M. A., Mercieca, K., Bettelheim, K. A. and Djordjevic, S. P. (2005) 'Bovine Feces from Animals with Gastrointestinal Infections Are a Source of Serologically Diverse Atypical Enteropathogenic Escherichia coli and Shiga Toxin-Producing E. coli Strains That Commonly Possess Intimin', *Applied and Environmental Microbiology,* 71(7), pp. 3405.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013) *Applied logistic regression.* John Wiley & Sons.

Hossain, M., Egan, S. A., Coffey, T., Ward, P. N., Wilson, R., Leigh, J. A. and Emes, R. D. (2015) 'Virulence related sequences; insights provided by comparative genomics of Streptococcus uberis of differing virulence', *BMC Genomics,* 16(1), pp. 334.

Houser, B. A., Donaldson, S. C., Padte, R., Sawant, A. A., DebRoy, C. and Jayarao, B. M. (2008) 'Assessment of Phenotypic and Genotypic Diversity of Escherichia coli Shed by Healthy Lactating Dairy Cattle', *Foodborne Pathogens and Disease,* 5(1), pp. 41-51.

Hrabák, J., Chudáčková, E. and Walková, R. (2013) 'Matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry for detection of antibiotic resistance mechanisms: from research to routine diagnosis', *Clinical Microbiology Reviews,* 26(1), pp. 103-114.

Hu, P., Janga, S. C., Babu, M., Diaz-Mejia, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., Chandran, S., Christopoulos, C., Nazarians-Armavil, A., Nasseri, N. K., Musso, G., Ali, M., Nazemof, N., Eroukova, V., Golshani, A., Paccanaro, A., Greenblatt, J. F., Moreno-Hagelsieb, G. and Emili, A. (2009) 'Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins', *PLoS Biol,* 7(4), pp. e96.

Huang, T.-S., Lee, S. S.-J., Lee, C.-C. and Chang, F.-C. (2020) 'Detection of carbapenem-resistant Klebsiella pneumoniae on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach', *PLOS ONE,* 15(2), pp. e0228459.

Hudaiberdiev, S., Choudhary, K. S., Vera Alvarez, R., Gelencsér, Z., Ligeti, B., Lamba, D. and Pongor, S. (2015) 'Census of solo LuxR genes in prokaryotic genomes', *Frontiers in Cellular and Infection Microbiology,* 5, pp. 20.

Huijps, K., Lam, T. J. G. M. and Hogeveen, H. (2008) 'Costs of mastitis: facts and perception', *Journal of Dairy Research,* 75(1), pp. 113-120.

Hussein, A. H. M., Ghanem, I. A. I., Eid, A. A. M., Ali, M. A., Sherwood, J. S., Li, G., Nolan, L. K. and Logue, C. M. (2013) 'Molecular and Phenotypic Characterization of Escherichia coli Isolated from Broiler Chicken Flocks in Egypt', *Avian Diseases,* 57(3), pp. 602-611.

Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000) 'Predicting protein function by genomic context: quantitative evaluation and qualitative inferences', *Genome research,* 10(8), pp. 1204-1210.

Huys, G., D'Haene, K., Collard, J.-M. and Swings, J. (2004) 'Prevalence and molecular characterization of tetracycline resistance in Enterococcus isolates from food', *Appl. Environ. Microbiol.,* 70(3), pp. 1555-1562.

Hyde, R. M., Down, P. M., Bradley, A. J., Breen, J. E., Hudson, C., Leach, K. A. and Green, M. J. (2020) 'Automated prediction of mastitis infection patterns in dairy herds using machine learning', *Scientific reports,* 10(1), pp. 1-8.

Ibtisam, El-Zubeir, E. M., Kutzer, P. and El-Owni, O. A. O. (2010) 'Frequencies and antibiotic susceptibility patterns of bacteria causing mastitis among cows and their environment in Khartoum State', *Research Journal of Microbiology,* 5(5), pp. 381-389.

Idriss, S. E., Foltys, V., Tančin, V., Kirchnerová, K., Tančinová, D. and Zaujec, K. (2014) 'Mastitis pathogens and their resistance against antimicrobial agents in dairy cows in Nitra, Slovakia', *Slovak Journal of Animal Science,* 47(1), pp. 33-38.

Izdebski, R., Baraniak, A., Fiett, J., Adler, A., Kazma, M., Salomon, J., Lawrence, C., Rossini, A., Salvia, A., Vidal Samso, J., Fierro, J., Paul, M., Lerman, Y., Malhotra-Kumar, S., Lammens, C., Goossens, H., Hryniewicz, W., Brun-Buisson, C., Carmeli, Y. and Gniadkowski, M. (2013) 'Clonal structure, extended-spectrum beta-lactamases, and acquired AmpC-type cephalosporinases of Escherichia coli populations colonizing patients in rehabilitation centers in four countries', *Antimicrob Agents Chemother,* 57(1), pp. 309-16.

Jackson, C. R., Fedorka-Cray, P. J., Davis, J. A., Barrett, J. B., Brousse, J. H., Gustafson, J. and Kucher, M. (2010) 'Mechanisms of antimicrobial resistance and genetic relatedness among enterococci isolated from dogs and cats in the United States', *Journal of Applied Microbiology,* 108(6), pp. 2171-2179.

Jain, C., Dilthey, A., Koren, S., Aluru, S. and Phillippy, A. M. (2017) 'A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases', *bioRxiv*, pp. 103812.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. and Aluru, S. (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature Communications,* 9(1), pp. 5114.

Jamali, H., Radmehr, B. and Ismail, S. (2014) 'Short communication: Prevalence and antibiotic resistance of Staphylococcus aureus isolated from bovine clinical mastitis', *Journal of Dairy Science,* 97(4), pp. 2226-2230.

Jansen, R., Greenbaum, D. and Gerstein, M. (2002) 'Relating whole-genome expression data with protein-protein interactions', *Genome research,* 12(1), pp. 37-46.

Jayarao, B. M., Gillespie, B. E., Lewis, M. J., Dowlen, H. H. and Oliver, S. P. (1999) 'Epidemiology of Streptococcus uberis intramammary infections in a dairy herd', *Zentralbl Veterinarmed B,* 46(7), pp. 433-42.

Jenkins, C., Rentenaar, R. J., Landraud, L. and Brisse, S. (2017) '180 - Enterobacteriaceae', in Cohen, J., Powderly, W.G. and Opal, S.M. (eds.) *Infectious Diseases (Fourth Edition)*: Elsevier, pp. 1565-1578.e2.

Jensen, C., Ethelberg, S., Olesen, B., Schiellerup, P., Olsen, K. E. P., Scheutz, F., Nielsen, E. M., Neimann, J., Høgh, B. and Gerner-Smidt, P. (2007) 'Attaching and effacing Escherichia coli isolates from Danish children: clinical significance and microbiological characteristics', *Clinical microbiology and infection,* 13(9), pp. 863-872.

Jensen, S. O. and Lyon, B. R. (2009) 'Genetics of antimicrobial resistance in Staphylococcus aureus', *Future Microbiol,* 4(5), pp. 565-82.

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M. and Aarestrup, F. M. (2014) 'Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli', *J Clin Microbiol,* 52(5), pp. 1501-10.

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M. and Scheutz, F. (2015) 'Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data', *J Clin Microbiol,* 53(8), pp. 2410-26.

Johnsen, L., Fimland, G. and Nissen-Meyer, J. (2005) 'The C-terminal domain of pediocin-like antimicrobial peptides (class IIa bacteriocins) is involved in specific recognition of the C-terminal part of cognate immunity proteins and in determining the antimicrobial spectrum', *Journal of Biological Chemistry,* 280(10), pp. 9243-9250.

Johnson, T. J. and Nolan, L. K. (2009) 'Pathogenomics of the virulence plasmids of Escherichia coli', *Microbiol Mol Biol Rev,* 73(4), pp. 750-74.

Johnson, T. J., Wannemuehler, Y. M. and Nolan, L. K. (2008) 'Evolution of the iss gene in Escherichia coli', *Appl. Environ. Microbiol.,* 74(8), pp. 2360-2369.

Jolley, K. A. and Maiden, M. C. J. (2010) 'BIGSdb: Scalable analysis of bacterial genome variation at the population level', *BMC Bioinformatics,* 11(1), pp. 595.

Jolley, K. A. and Maiden, M. C. J. (2014) 'Using MLST to study bacterial variation: prospects in the genomic era', *Future microbiology,* 9(5), pp. 623-630.

Jones, A. L., Knoll, K. M. and Rubens, C. E. (2000) 'Identification of Streptococcus agalactiae virulence genes in the neonatal rat sepsis model using signature-tagged mutagenesis', *Mol Microbiol,* 37(6), pp. 1444-55.

Jones, D. T., Taylort, W. R. and Thornton, J. M. (1992) 'A new approach to protein fold recognition', *Nature,* 358(6381), pp. 86-89.

Jones, N., Bohnsack, J. F., Takahashi, S., Oliver, K. A., Chan, M.-S., Kunst, F., Glaser, P., Rusniok, C., Crook, D. W. M., Harding, R. M., Bisharat, N. and Spratt, B. G. (2003) 'Multilocus sequence typing system for group B streptococcus', *Journal of clinical microbiology,* 41(6), pp. 2530-2536.

Juhász-Kaszanyitzky, E., Jánosi, S., Somogyi, P., Dán, A., van der Graaf-van Bloois, L., van Duijkeren, E. and Wagenaar, J. A. (2007) 'MRSA transmission between cows and humans', *Emerging infectious diseases,* 13(4), pp. 630-632.

Jørgensen, H. J., Nordstoga, A. B., Sviland, S., Zadoks, R. N., Sølverød, L., Kvitle, B. and Mørk, T. (2016) 'Streptococcus agalactiae in the environment of bovine dairy herds–rewriting the textbooks?', *Veterinary microbiology,* 184, pp. 64-72.

Júnior, J. C. R., Silva, F. F., Lima, J. B. A., Ossugui, E. H., Junior, P. I. T., Campos, A., Navarro, A., Tamanini, R., Ribeiro, J. and Alfieri, A. A. (2019) 'Molecular characterization and antimicrobial resistance of pathogenic Escherichia coli isolated from raw milk and Minas Frescal cheeses in Brazil', *Journal of dairy science,* 102(12), pp. 10850-10854.

Kaake, R. M., Wang, X. and Huang, L. (2010) 'Profiling of protein interaction networks of protein complexes using affinity purification and quantitative mass spectrometry', *Mol Cell Proteomics,* 9(8), pp. 1650-65.

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. and Lund, O. (2014) 'Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms', *PLOS ONE,* 9(8), pp. e104984.

Kaipainen, T., Pohjanvirta, T., Shpigel, N. Y., Shwimmer, A., Pyorala, S. and Pelkonen, S. (2002) 'Virulence factors of Escherichia coli isolated from bovine clinical mastitis', *Vet Microbiol,* 85(1), pp. 37-46.

Kalayu, A. A., Woldetsadik, D. A., Woldeamanuel, Y., Wang, S.-H., Gebreyes, W. A. and Teferi, T. (2020) 'Burden and antimicrobial resistance of S. aureus in dairy farms in Mekelle, Northern Ethiopia', *BMC Veterinary Research,* 16(1), pp. 20.

Kallow, W., Erhard, M., Lima, N., Santos, I. M., Serra, R., Venâncio, A., Friedl, T., Müller, J., de Hoog, G. S. and Verkley, G. J. M. (2006) 'Microbial strain characterisation by MALDI-TOF MS-possibilities and limits'.

Kamphuis, C., Mollenhorst, H., Heesterbeek, J. A. P. and Hogeveen, H. (2010) 'Detection of clinical mastitis with sensor data from automatic milking systems is improved by using decision-tree induction', *Journal of Dairy Science,* 93(8), pp. 3616-3627.

Kampstra, P. (2008) 'Beanplot: A boxplot alternative for visual comparison of distributions', *Journal of statistical software,* 28(1), pp. 1-9.

Kanehisa, M. and Goto, S. (2000) 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic acids research,* 28(1), pp. 27-30.

Karas, M., Bachmann, D., Bahr, U. and Hillenkamp, F. (1987) 'Matrix-assisted ultraviolet laser desorption of non-volatile compounds', *International journal of mass spectrometry and ion processes,* 78, pp. 53-68.

Karas, M. and Hillenkamp, F. (1988) 'Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons', *Analytical Chemistry,* 60(20), pp. 2299-2301.

Karumanchi, S. A., August, P. and Podymow, T. (2010) 'Renal complications in normal pregnancy', *Comprehensive clinical nephrology*: Elsevier, pp. 504-515.

Kateete, D. P., Kabugo, U., Baluku, H., Nyakarahuka, L., Kyobe, S., Okee, M., Najjuka, C. F. and Joloba, M. L. (2013) 'Prevalence and Antimicrobial Susceptibility Patterns of Bacteria from Milkmen and Cows with Clinical Mastitis in and around Kampala, Uganda', *PLOS ONE,* 8(5), pp. e63413.

Katholm, J. and Rattenborg, E. (2009) 'Surveillance of the B streptococcal infection in Danish dairy herds', *Dansk Veterinærtidsskrift,* 92(19), pp. 24-31.

Katz, L. S., Griswold, T., Williams-Newkirk, A. J., Wagner, D., Petkau, A., Sieffert, C., Van Domselaar, G., Deng, X. and Carleton, H. A. (2017) 'A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens', *Frontiers in Microbiology,* 8, pp. 375.

Kauffmann, F. (1947) 'The serology of the coli group', *Journal of Immunology,* 57(1), pp. 71-100.

Keane, O. M. (2016) 'Genetic diversity, the virulence gene profile and antimicrobial resistance of clinical mastitis-associated Escherichia coli', *Research in Microbiology,* 167(8), pp. 678-684.

Keefe, G. (2012) 'Update on control of Staphylococcus aureus and Streptococcus agalactiae for management of mastitis', *Veterinary Clinics: Food Animal Practice,* 28(2), pp. 203-216.

Kellogg, S. L. and Kristich, C. J. (2018) 'Convergence of PASTA Kinase and Two-Component Signaling in Response to Cell Wall Stress in Enterococcus faecalis', *Journal of bacteriology,* 200(12), pp. e00086-18.

Kempf, F., Slugocki, C., Blum, S. E., Leitner, G. and Germon, P. (2016) 'Genomic comparative study of bovine mastitis Escherichia coli', *PLoS One,* 11(1), pp. e0147954.

Khan, Z. A., Siddiqui, M. F. and Park, S. (2019) 'Current and Emerging Methods of Antibiotic Susceptibility Testing', *Diagnostics (Basel, Switzerland),* 9(2), pp. 49.

Kidsley, A. K., Abraham, S., Bell, J. M., O'Dea, M., Laird, T. J., Jordan, D., Mitchell, P., McDevitt, C. A. and Trott, D. J. (2018) 'Antimicrobial Susceptibility of Escherichia coli and Salmonella spp. Isolates From Healthy Pigs in Australia: Results of a Pilot National Survey', *Frontiers in Microbiology,* 9, pp. 1207.

Kitten, T. and Willis, D. K. (1996) 'Suppression of a sensor kinase-dependent phenotype in Pseudomonas syringae by ribosomal proteins L35 and L20', *Journal of Bacteriology,* 178(6), pp. 1548.

Kjos, M., Salehian, Z., Nes, I. F. and Diep, D. B. (2010) 'An extracellular loop of the mannose phosphotransferase system component IIC is responsible for specific targeting by class IIa bacteriocins', *Journal of bacteriology,* 192(22), pp. 5906-5913.

Klaas, I. C. and Zadoks, R. N. (2018) 'An update on environmental mastitis: Challenging perceptions', *Transboundary and emerging diseases,* 65, pp. 166-185.

Klibi, N., Aouini, R., Borgo, F., Said, L. B., Ferrario, C., Dziri, R., Boudabous, A., Torres, C. and Slama, K. B. (2015) 'Antibiotic resistance and virulence of faecal enterococci isolated from food-producing animals in Tunisia', *Annals of Microbiology,* 65(2), pp. 695-702.

Klimienė, I., Ružauskas, M., Špakauskas, V., Mockeliūnas, R., Pereckienė, A. and Butrimaitė-Ambrozevičienė, Č. (2011) 'Prevalence of gram positive bacteria in cow mastitis and their susceptibility to beta-lactam antibiotics', *Veterinarija ir Zootechnika,* 56(78).

Kohanski, M. A., Dwyer, D. J. and Collins, J. J. (2010) 'How antibiotics kill bacteria: from targets to networks', *Nature reviews. Microbiology,* 8(6), pp. 423-435.

Kolde, R. and Kolde, M. R. (2015) 'Package 'pheatmap'', *R Package,* 1(7).

Komano, T., Utsumi, R. and Kawamukai, M. (1991) 'Functional analysis of the fic gene involved in regulation of cell division', *Research in Microbiology,* 142(2), pp. 269-277.

Koteva, K., Hong, H.-J., Wang, X. D., Nazi, I., Hughes, D., Naldrett, M. J., Buttner, M. J. and Wright, G. D. (2010) 'A vancomycin photoprobe identifies the histidine kinase VanSsc as a vancomycin receptor', *Nature chemical biology,* 6(5), pp. 327.

Koul, A., Dendouga, N., Vergauwen, K., Molenberghs, B., Vranckx, L., Willebrords, R., Ristic, Z., Lill, H., Dorange, I. and Guillemont, J. (2007) 'Diarylquinolines target subunit c of mycobacterial ATP synthase', *Nature chemical biology,* 3(6), pp. 323.

Kreausukon, K., Fetsch, A., Kraushaar, B., Alt, K., Müller, K., Krömker, V., Zessin, K. H., Käsbohrer, A. and Tenhagen, B. A. (2012) 'Prevalence, antimicrobial resistance, and molecular characterization of methicillin-resistant Staphylococcus aureus from bulk tank milk of dairy herds', *Journal of Dairy Science,* 95(8), pp. 4382-4388.

Krishnamurthy, T. and Ross, P. L. (1996) 'Rapid identification of bacteria by direct matrix-assisted laser desorption/ionization mass spectrometric analysis of whole cells', *Rapid communications in mass spectrometry,* 10(15), pp. 1992-1996.

Krukowski, H., Henryka, L., Zastempowska, E. W. A., Smulski, S. and Bis-Wencel, H. (2020) 'Etiological agents of bovine mastitis in Poland', *Medycyna Weterynaryjna,* 76, pp. 6339-2020.

Krömker, V., Reinecke, F., Paduch, J.-H. and Grabowski, N. (2014) 'Bovine Streptococcus uberis intramammary infections and mastitis'.

Kubat, M. (2017) *An introduction to machine learning.* Springer.

Kumari, T., Bhakat, C. and Choudhary, R. K. (2018) 'A review on subclinical mastitis in dairy cattle', *Int. J. Pure Appl. Biosci,* 6(2), pp. 1291-1299.

Kurjogi, M. M. and Kaliwal, B. B. (2014) 'Epidemiology of bovine mastitis in cows of Dharwad district', *International scholarly research notices,* 2014.

Kuroishi, T., Komine, K.-I., Kai, K., Itagaki, M., Kobayashi, J., Ohta, M., Kamata, S.-I. and Kumagai, K. (2003) 'Concentrations and specific antibodies to staphylococcal enterotoxin-C and toxic shock syndrome toxin-1 in bovine mammary gland secretions, and inflammatory response to the intramammary inoculation of these toxins', *Journal of veterinary medical science,* 65(8), pp. 899-906.

Kuyucuoğlu, Y. (2011) 'Antibiotic resistance of enterococci isolated from bovine subclinical mastitis', *Eurasian Journal of Veterinary Sciences,* 27(4), pp. 231-234.

Köhler, C.-D. and Dobrindt, U. (2011) 'What defines extraintestinal pathogenic Escherichia coli?', *International Journal of Medical Microbiology,* 301(8), pp. 642-647.

Kühn, I., Iversen, A., Finn, M., Greko, C., Burman, L. G., Blanch, A. R., Vilanova, X., Manero, A., Taylor, H. and Caplin, J. (2005) 'Occurrence and relatedness of vancomycin-resistant enterococci in animals, humans, and the environment in different European regions', *Applied and environmental microbiology,* 71(9), pp. 5383-5390.

Lakew, M., Tolosa, T. and Tigre, W. (2009) 'Prevalence and major bacterial causes of bovine mastitis in Asella, South Eastern Ethiopia', *Tropical animal health and production,* 41(7), pp. 1525.

Lam, T., Van Den Borne, B. H. P., Jansen, J., Huijps, K., Van Veersen, J. C. L., Van Schaik, G. and Hogeveen, H. (2013) 'Improving bovine udder health: A national mastitis control program in the Netherlands', *Journal of dairy science,* 96(2), pp. 1301-1311.

Landin, H., Mörk, M. J., Larsson, M. and Waller, K. P. (2015) 'Vaccination against Staphylococcus aureus mastitis in two Swedish dairy herds', *Acta Veterinaria Scandinavica,* 57(1), pp. 81.

Landman, W. J. M., Buter, G. J., Dijkman, R. and van Eck, J. H. H. (2014) 'Molecular typing of avian pathogenic Escherichia coli colonies originating from outbreaks of E. coli peritonitis syndrome in chicken flocks', *Avian Pathology,* 43(4), pp. 345-356.

Lasch, P., Fleige, C., Stämmler, M., Layer, F., Nübel, U., Witte, W. and Werner, G. (2014) 'Insufficient discriminatory power of MALDI-TOF mass spectrometry for typing of Enterococcus faecium and Staphylococcus aureus isolates', *Journal of Microbiological Methods,* 100, pp. 58-69.

Lau, S. H., Reddy, S., Cheesbrough, J., Bolton, F. J., Willshaw, G., Cheasty, T., Fox, A. J. and Upton, M. (2008) 'Major uropathogenic Escherichia coli strain isolated in the northwest of England identified by multilocus sequence typing', *Journal of clinical microbiology,* 46(3), pp. 1076-1080.

Laverde Gomez, J. A., van Schaik, W., Freitas, A. R., Coque, T. M., Weaver, K. E., Francia, M. V., Witte, W. and Werner, G. (2011) 'A multiresistance megaplasmid pLG1 bearing a hylEfm genomic island in hospital Enterococcus faecium isolates', *International Journal of Medical Microbiology,* 301(2), pp. 165-175.

Lazarus, B., Paterson, D. L., Mollinger, J. L. and Rogers, B. A. (2015) 'Do human extraintestinal Escherichia coli infections resistant to expanded-spectrum cephalosporins originate from food-producing animals? A systematic review', *Clin Infect Dis,* 60(3), pp. 439-52.

Le Maréchal, C., Seyffert, N., Jardin, J., Hernandez, D., Jan, G., Rault, L., Azevedo, V., François, P., Schrenzel, J., van de Guchte, M., Even, S., Berkova, N., Thiéry, R., Fitzgerald, J. R., Vautor, E. and Le Loir, Y. (2011) 'Molecular Basis of Virulence in Staphylococcus aureus Mastitis', *PLOS ONE,* 6(11), pp. e27354.

Le, T., Vo, M. T., Vo, B., Lee, M. Y. and Baik, S. W. (2019) 'A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction', *Complexity,* 2019.

LeBlanc, S. J., Osawa, T. and Dubuc, J. (2011) 'Reproductive tract defense and disease in postpartum dairy cows', *Theriogenology,* 76(9), pp. 1610-1618.

Lehtolainen, T., Pohjanvirta, T., Pyörälä, S. and Pelkonen, S. (2003) 'Association between virulence factors and clinical course of Escherichia coli mastitis', *Acta Veterinaria Scandinavica,* 44(4), pp. 203.

Leigh, J. A., Egan, S. A., Ward, P. N., Field, T. R. and Coffey, T. J. (2010) 'Sortase anchored proteins of Streptococcus uberis play major roles in the pathogenesis of bovine mastitis in dairy cattle', *Veterinary research,* 41(5), pp. 63.

Leigh, J. A. and Field, T. R. (1991) 'Killing of Streptococcus uberis by bovine neutrophils following growth in chemically defined media', *Veterinary research communications,* 15(1), pp. 1-6.

Leigh, J. A., Field, T. R. and Williams, M. R. (1990) 'Two strains of Streptococcus uberis, of differing ability to cause clinical mastitis, differ in their ability to resist some host defence factors', *Research in veterinary science,* 49(1), pp. 85-87.

Leigh, J. A. and Lincoln, R. A. (1997) 'Streptococcus uberis acquires plasmin activity following growth in the presence of bovine plasminogen through the action of its specific plasminogen activator', *FEMS Microbiology Letters,* 154(1), pp. 123-129.

Leimbach, A., Poehlein, A., Vollmers, J., Görlich, D., Daniel, R. and Dobrindt, U. (2017) 'No evidence for a bovine mastitis Escherichia coli pathotype', *BMC Genomics,* 18(1), pp. 359.

Leimbach, A., Poehlein, A., Witten, A., Scheutz, F., Schukken, Y., Daniel, R. and Dobrindt, U. (2015) 'Complete genome sequences of Escherichia coli strains 1303 and ECC-1470 isolated from bovine mastitis', *Genome Announc.,* 3(2), pp. e00182-15.

Leitner, G., Koren, O., Jacoby, S., Merin, U. and Silanikove, N. (2012) 'Options for handling chronic subclinical mastitis during lactation in modern dairy farms', *Israel journal of veterinary medicine,* 67(3), pp. 162-169.

Lemaître, G., Nogueira, F. and Aridas, C. K. (2017) 'Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning', *The Journal of Machine Learning Research,* 18(1), pp. 559-563.

Lester, C. H., Frimodt-Møller, N., Sørensen, T. L., Monnet, D. L. and Hammerum, A. M. (2006) 'In vivo transfer of the vanA resistance gene from an Enterococcus faecium isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers', *Antimicrobial agents and chemotherapy,* 50(2), pp. 596-599.

Letunic, I. and Bork, P. (2016) 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic Acids Res,* 44(W1), pp. W242-5.

Letunic, I. and Bork, P. (2018) '20 years of the SMART protein domain annotation resource', *Nucleic Acids Research,* 46(D1), pp. D493-D496.

Li, D.-X., Zhang, S.-M., Hu, G.-Z., Wang, Y., Liu, H.-B., Wu, C.-M., Shang, Y.-H., Chen, Y.-X. and Du, X.-D. (2011) 'Tn3-associated rmtB together with qnrS1, aac(6')-Ib-cr and blaCTX-M-15 are co-located on an F49:A-:B- plasmid in an Escherichia coli ST10 strain in China', *Journal of Antimicrobial Chemotherapy,* 67(1), pp. 236-238.

Li, X., Tang, J., Zhang, Q., Gao, B., Yang, J. J., Song, S., Wu, W., Zhang, W., Yao, P., Deng, N., Deng, L., Xie, Y., Qian, H. and Wu, H. (2020) 'Power-efficient neural network with artificial dendrites', *Nature Nanotechnology*.

Li, X., Wang, L. and Sung, E. (2004) 'Improving adaboost for classification on small training sample sets with active learning', *Korea*.

Libbrecht, M. W. and Noble, W. S. (2015) 'Machine learning applications in genetics and genomics', *Nature Reviews Genetics,* 16(6), pp. 321-332.

Lifshitz, Z., Sturlesi, N. a., Parizade, M., Blum, S. E., Gordon, M., Taran, D. and Adler, A. (2018) 'Distinctiveness and Similarities Between Extended-Spectrum β-Lactamase-Producing Escherichia coli Isolated from Cattle and the Community in Israel', *Microbial Drug Resistance,* 24(6), pp. 868-875.

Ligozzi, M., Bernini, C., Bonora, M. G., de Fatima, M., Zuliani, J. and Fontana, R. (2002) 'Evaluation of the VITEK 2 System for Identification and Antimicrobial Susceptibility Testing of Medically Relevant Gram-Positive Cocci', *Journal of Clinical Microbiology,* 40(5), pp. 1681.

Lim, S.-k., Joo, Y.-s., Moon, J.-s., Lee, A.-r., Nam, H.-m., Wee, S.-h. and Koh, H.-b. (2004) 'Molecular Typing of Enterotoxigenic Staphylococcus aureus Isolated from Bovine Mastitis in Korea', *Journal of Veterinary Medical Science,* 66(5), pp. 581-584.

Liu, A., Tran, L., Becket, E., Lee, K., Chinn, L., Park, E., Tran, K. and Miller, J. H. (2010) 'Antibiotic sensitivity profiles determined with an Escherichia coli gene knockout collection: generating an antibiotic bar code', *Antimicrobial agents and chemotherapy,* 54(4), pp. 1393-1403.

Liu, Q., Sung, A. H., Qiao, M., Chen, Z., Yang, J. Y., Yang, M. Q., Huang, X. and Deng, Y. (2009) 'Comparison of feature selection and classification for MALDI-MS data', *BMC genomics,* 10(S1), pp. S3.

Liu, X., Wang, J., Chen, M., Che, R., Ding, W., Yu, F., Zhou, Y., Cui, W., Xiaoxu, X., God'spower, B.-O. and Li, Y. (2019) 'Comparative proteomic analysis reveals drug resistance of Staphylococcus xylosus ATCC700404 under tylosin stress', *BMC Veterinary Research,* 15(1), pp. 224.

Lopez, G., Rojas, A., Tress, M. and Valencia, A. (2007) 'Assessment of predictions submitted for the CASP7 function prediction category', *Proteins: Structure, Function, and Bioinformatics,* 69(S8), pp. 165-174.

Lopez-Benavides, M. G., Williamson, J. H., Pullinger, G. D., Lacy-Hulbert, S. J., Cursons, R. T. and Leigh, J. A. (2007) 'Field observations on the variation of Streptococcus uberis populations in a pasture-based dairy farm', *Journal of dairy science,* 90(12), pp. 5558-5566.

Lopez-Causape, C., Sommer, L. M., Cabot, G., Rubio, R., Ocampo-Sosa, A. A., Johansen, H. K., Figuerola, J., Canton, R., Kidd, T. J., Molin, S. and Oliver, A. (2017) 'Evolution of the Pseudomonas aeruginosa mutational resistome in an international Cystic Fibrosis clone', *Sci Rep,* 7(1), pp. 5555.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J. and Marchler-Bauer, A. (2020) 'CDD/SPARCLE: the conserved domain database in 2020', *Nucleic acids research,* 48(D1), pp. D265-D268.

Ludwig, W., Schleifer, K. H. and Whitman, W. B. (2015) 'Enterococcaceae fam. nov', *Bergey's Manual of Systematics of Archaea and Bacteria*, pp. 1-2.

Luppens, S. B. I., Kara, D., Bandounas, L., Jonker, M. J., Wittink, F. R. A., Bruning, O., Breit, T. M., Ten Cate, J. M. and Crielaard, W. (2008) 'Effect of Veillonella parvula on the antimicrobial resistance and gene expression of Streptococcus mutans grown in a dual-species biofilm', *Oral Microbiology and Immunology,* 23(3), pp. 183-189.

Léger, D. F., Newby, N. C., Reid-Smith, R., Anderson, N., Pearl, D. L., Lissemore, K. D. and Kelton, D. F. (2017) 'Estimated antimicrobial dispensing frequency and preferences for lactating cow therapy by Ontario dairy veterinarians', *The Canadian veterinary journal = La revue veterinaire canadienne,* 58(1), pp. 26-34.

López Fernández, H., Reboiro-Jato, M., Pérez Rodríguez, J. A., Fdez-Riverola, F. and Glez-Peña, D. (2016) 'Implementing effective machine learning-based workflows for the analysis of mass spectrometry data', *Journal of Integrated OMICS,* 6(1), pp. 23-27.

Ma, J., Cocchiaro, J. and Lee, J. C. (2004) 'Evaluation of Serotypes of Staphylococcus aureus Strains Used in the Production of a Bovine Mastitis Bacterin', *Journal of Dairy Science,* 87(1), pp. 178-182.

Ma, Z., Richard, H., Tucker, D. L., Conway, T. and Foster, J. W. (2002) 'Collaborative regulation of Escherichia coli glutamate-dependent acid resistance by two AraC-like regulators, GadX and GadW (YhiW)', *Journal of bacteriology,* 184(24), pp. 7001-7012.

Maciel-Guerra, A., Esener, N., Giebel, K., Lea, D., Green, M. J., Bradley, A. J. and Dottorini, T. (2021) 'Prediction of Streptococcus uberis clinical mastitis treatment success in dairy herds by means of mass spectrometry and machine-learning', *Scientific Reports,* 11(1), pp. 7736.

Macvanin, M. and Adhya, S. (2012) 'Architectural organization in E. coli nucleoid', *Biochim Biophys Acta,* 1819(7), pp. 830-5.

Madonna, A. J., Basile, F., Ferrer, I., Meetani, M. A., Rees, J. C. and Voorhees, K. J. (2000) 'On-probe sample pretreatment for the detection of proteins above 15 KDa from whole cell bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry', *Rapid Communications in Mass Spectrometry,* 14(23), pp. 2220-2229.

Madoshi, B. P., Kudirkiene, E., Mtambo, M. M. A., Muhairwa, A. P., Lupindu, A. M. and Olsen, J. E. (2016) 'Characterisation of commensal Escherichia coli isolated from apparently healthy cattle and their attendants in Tanzania', *PLoS One,* 11(12).

Maeda, Y., Sugiyama, Y., Kogiso, A., Lim, T.-K., Harada, M., Yoshino, T., Matsunaga, T. and Tanaka, T. (2018) 'Colony Fingerprint-Based Discrimination of Staphylococcus species with Machine Learning Approaches', *Sensors,* 18(9).

Mahmmod, Y. S., Nonnemann, B., Svennesen, L., Pedersen, K. and Klaas, I. C. (2018) 'Typeability of MALDI-TOF assay for identification of non-aureus staphylococci associated with bovine intramammary infections and teat apex colonization', *Journal of Dairy Science,* 101(10), pp. 9430-9438.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M. and Spratt, B. G. (1998) 'Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms', *Proceedings of the National Academy of Sciences of the United States of America,* 95(6), pp. 3140-3145.

Makarova, K. S., Ponomarev, V. A. and Koonin, E. V. (2001) 'Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins', *Genome Biology,* 2(9), pp. research0033.1.

Malhotra-Kumar, S., Mazzariol, A., Van Heirstraeten, L., Lammens, C., de Rijk, P., Cornaglia, G. and Goossens, H. (2009) 'Unusual resistance patterns in macrolide-resistant Streptococcus pyogenes harbouring erm(A)', *J Antimicrob Chemother,* 63(1), pp. 42-6.

Malik, A., Nagy, B., Kugler, R. and Szmolka, A. (2017) 'Pathogenic potential and virulence genotypes of intestinal and faecal isolates of porcine post-weaning enteropathogenic Escherichia coli', *Research in Veterinary Science,* 115, pp. 102-108.

Manga, I., Hasman, H., Smidkova, J., Medvecky, M., Dolejska, M. and Cizek, A. (2019) 'Fecal Carriage and Whole-Genome Sequencing-Assisted Characterization of CMY-2 Beta-Lactamase-Producing Escherichia coli in Calves at Czech Dairy Cow Farm', *Foodborne Pathog Dis,* 16(1), pp. 42-53.

Marrakchi, M., Liu, X. and Andreescu, S. (2014) 'Oxidative stress and antibiotic resistance in bacterial pathogens: state of the art, methodologies, and future trends', *Advancements of Mass Spectrometry in Biomedical Research*: Springer, pp. 483-498.

Martinez, G., Harel, J., Higgins, R., Lacouture, S., Daignault, D. and Gottschalk, M. (2000) 'Characterization of Streptococcus agalactiae isolates of bovine and human origin by randomly amplified polymorphic DNA analysis', *Journal of clinical microbiology,* 38(1), pp. 71-78.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. and Šali, A. (2000) 'Comparative protein structure modeling of genes and genomes', *Annual review of biophysics and biomolecular structure,* 29(1), pp. 291-325.

Martínez, A. M. and Kak, A. C. (2001) 'Pca versus lda', *IEEE transactions on pattern analysis and machine intelligence,* 23(2), pp. 228-233.

Masalha, M., Borovok, I., Schreiber, R., Aharonowitz, Y. and Cohen, G. (2001) 'Analysis of transcription of the Staphylococcus aureus aerobic class Ib and anaerobic class III ribonucleotide reductase genes in response to oxygen', *Journal of bacteriology,* 183(24), pp. 7260-7272.

Maslow, J. N., Maury Ellis, M. and Arbeit, R. D. (1993) 'Molecular Epidemiology: Application of Contemporary Techniques to the Typing of Microorganisms', *Clinical Infectious Diseases,* 17(2), pp. 153-162.

Mather, C. A., Werth, B. J., Sivagnanam, S., SenGupta, D. J. and Butler-Wu, S. M. (2016) 'Rapid Detection of Vancomycin-Intermediate Staphylococcus aureus by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry', *Journal of clinical microbiology,* 54(4), pp. 883-890.

Matsumoto-Nakano, M. and Kuramitsu, H. K. (2006) 'Role of Bacteriocin Immunity Proteins in the Antimicrobial Sensitivity of Streptococcus mutans', *Journal of Bacteriology,* 188(23), pp. 8095-8102.

McBride, S. M., Fischetti, V. A., LeBlanc, D. J., Moellering Jr, R. C. and Gilmore, M. S. (2007) 'Genetic diversity among Enterococcus faecalis', *PloS one,* 2(7), pp. e582.

McCarthy, A. J., Witney, A. A., Gould, K. A., Moodley, A., Guardabassi, L., Voss, A., Denis, O., Broens, E. M., Hinds, J. and Lindsay, J. A. (2011) 'The distribution of mobile genetic elements (MGEs) in MRSA CC398 is associated with both host and country', *Genome Biol Evol,* 3, pp. 1164-74.

McDonald, J. S. and Anderson, A. J. (1981) 'Experimental intramammary infection of the dairy cow with Escherichia coli during the nonlactating period', *American journal of veterinary research,* 42(2), pp. 229-231.

McDougall, S. (2003) 'Intramammary treatment of clinical mastitis of dairy cows with a combination of lincomycin and neomycin, or penicillin and dihydrostreptomycin', *New Zealand Veterinary Journal,* 51(3), pp. 111-116.

McLachlan, G. (2004) *Discriminant analysis and statistical pattern recognition.* John Wiley & Sons.

McMillan, D. J., Bessen, D. E., Pinho, M., Ford, C., Hall, G. S., Melo-Cristino, J. and Ramirez, M. (2010) 'Population genetics of Streptococcus dysgalactiae subspecies equisimilis reveals widely dispersed clones and extensive recombination', *PLoS One,* 5(7), pp. e11741.

Meister, A. (2009) *Advances in enzymology and related areas of molecular biology.* John Wiley & Sons.

Mellmann, A., Bimet, F., Bizet, C., Borovskaya, A. D., Drake, R. R., Eigner, U., Fahr, A. M., He, Y., Ilina, E. N., Kostrzewa, M., Maier, T., Mancinelli, L., Moussaoui, W., Prévost, G., Putignani, L., Seachord, C. L., Tang, Y. W. and Harmsen, D. (2009) 'High Interlaboratory Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry-Based Species Identification of Nonfermenting Bacteria', *Journal of Clinical Microbiology,* 47(11), pp. 3732.

Mellmann, A., Cloud, J., Maier, T., Keckevoet, U., Ramminger, I., Iwen, P., Dunn, J., Hall, G., Wilson, D. and Lasala, P. (2008) 'Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry in comparison to 16S rRNA gene sequencing for species identification of nonfermenting bacteria', *Journal of clinical microbiology,* 46(6), pp. 1946-1954.

Messenger, A. M., Barnes, A. N. and Gray, G. C. (2014) 'Reverse Zoonotic Disease Transmission (Zooanthroponosis): A Systematic Review of Seldom-Documented Human Biological Threats to Animals', *PLOS ONE,* 9(2), pp. e89055.

Middleton, J. R., Ma, J., Rinehart, C. L., Taylor, V. N., Luby, C. D. and Steevens, B. J. (2006) 'Efficacy of different Lysigin (TM) formulations in the prevention of Staphylococcus aureus intramammary infection in dairy heifers', *The Journal of dairy research,* 73(1), pp. 10.

Miekley, B., Traulsen, I. and Krieter, J. (2013) 'Mastitis detection in dairy cows: the application of support vector machines', *The Journal of Agricultural Science,* 151(6), pp. 889-897.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. and Gurevich, A. (2018) 'Versatile genome assembly evaluation with QUAST-LG', *Bioinformatics,* 34(13), pp. i142-i150.

Miller, W. R., Munita, J. M. and Arias, C. A. (2014) 'Mechanisms of antibiotic resistance in enterococci', *Expert review of anti-infective therapy,* 12(10), pp. 1221-1236.

Minato, Y., Dawadi, S., Kordus, S. L., Sivanandam, A., Aldrich, C. C. and Baughn, A. D. (2018) 'Mutual potentiation drives synergy between trimethoprim and sulfamethoxazole', *Nature Communications,* 9(1), pp. 1003.

Minogue, T. D., Daligault, H. E., Davenport, K. W., Broomall, S. M., Bruce, D. C., Chain, P. S., Coyne, S. R., Chertkov, O., Freitas, T., Gibbons, H. S., Jaissle, J., Koroleva, G. I., Ladner, J. T., Palacios, G. F., Rosenzweig, C. N., Xu, Y. and Johnson, S. L. (2014) 'Complete Genome Assembly of Enterococcus faecalis 29212, a Laboratory Reference Strain', *Genome Announc,* 2(5).

Mishra, S. and Horswill, A. R. (2017) 'Heparin Mimics Extracellular DNA in Binding to Cell Surface-Localized Proteins and Promoting Staphylococcus aureus Biofilm Formation', *mSphere,* 2(3), pp. e00135-17.

Misra, N., Wines, T. F., Knopp, C. L., Hermann, R., Bond, L., Mitchell, B., McGuire, M. A. and Tinker, J. K. (2018) 'Immunogenicity of a Staphylococcus aureus-cholera toxin A(2)/B vaccine for bovine mastitis', *Vaccine,* 36(24), pp. 3513-3521.

Mitchell, T. M. (1997) 'Machine learning. 1997', *Burr Ridge, IL: McGraw Hill,* 45(37), pp. 870-877.

Monaghan, Á., Byrne, B., Fanning, S., Sweeney, T., McDowell, D. and Bolton, D. J. (2012) 'Serotypes and virulotypes of non-O157 shiga-toxin producing Escherichia coli (STEC) on bovine hides and carcasses', *Food Microbiology,* 32(2), pp. 223-229.

Monistero, V., Barberio, A., Biscarini, F., Cremonesi, P., Castiglioni, B., Graber, H. U., Bottini, E., Ceballos-Marquez, A., Kroemker, V., Petzer, I. M., Pollera, C., Santisteban, C., Veiga Dos Santos, M.,

Bronzo, V., Piccinini, R., Re, G., Cocchi, M. and Moroni, P. (2020) 'Different distribution of antimicrobial resistance genes and virulence profiles of Staphylococcus aureus strains isolated from clinical mastitis in six countries', *Journal of Dairy Science,* 103(4), pp. 3431-3446.

Montironi, I. D., Moliva, M. V., Campra, N. A., Raviolo, J. M., Bagnis, G., Cariddi, L. N. and Reinoso, E. B. (2020) 'Characterization of an Enterococcus faecium strain in a murine mastitis model', *Journal of Applied Microbiology,* 128(5), pp. 1289-1300.

Moody, J. and Darken, C. J. (1989) 'Fast learning in networks of locally-tuned processing units', *Neural computation,* 1(2), pp. 281-294.

Morgan, J., Daugherty, R., Hilchie, A. and Carey, B. (2003) 'Sample size and modeling accuracy of decision tree based data mining tools', *Acad Inf Manag Sci J,* 6, pp. 71-99.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A. and Kobayashi, R. (2005) 'Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum', *Bioinformatics,* 21(9), pp. 1764-1775.

Mueller, J. P. and Massaron, L. (2016) *Machine learning for dummies.* John Wiley & Sons.

Mueller-Premru, M., Zidar, N., Cvitković Špik, V., Krope, A. and Kikelj, D. (2009) 'Benzoxazine Series of Histidine Kinase Inhibitors as Potential Antimicrobial Agents with Activity against Enterococci', *Chemotherapy,* 55(6), pp. 414-417.

Mullarky, I. K., Su, C., Frieze, N., Park, Y. H. and Sordillo, L. M. (2001) 'Staphylococcus aureus agr genotypes with enterotoxin production capabilities can resist neutrophil bactericidal activity', *Infection and immunity,* 69(1), pp. 45-51.

Munita, J. M. and Arias, C. A. (2016) 'Mechanisms of Antibiotic Resistance', *Microbiology spectrum,* 4(2), pp. 10.1128/microbiolspec.VMBF-0016-2015.

Munoz, M. A. and Zadoks, R. N. (2007) 'Patterns of fecal shedding of Klebsiella by dairy cows', *Journal of dairy science,* 90(3), pp. 1220-1224.

Muroi, M., Shima, K., Igarashi, M., Nakagawa, Y. and Tanamoto, K.-i. (2012) 'Application of matrix-assisted laser desorption ionization-time of flight mass spectrometry for discrimination of laboratory-derived antibiotic-resistant bacteria', *Biological and Pharmaceutical Bulletin,* 35(10), pp. 1841-1845.

Murray, P. R. (2010) 'Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: usefulness for taxonomy and epidemiology', *Clinical Microbiology and Infection,* 16(11), pp. 1626-1630.

Murray, P. R. (2012) 'What is new in clinical microbiology—microbial identification by MALDI-TOF mass spectrometry: a paper from the 2011 William Beaumont Hospital Symposium on molecular pathology', *The journal of molecular diagnostics,* 14(5), pp. 419-423.

Murugan, M. S., Sinha, D. K., Vinodh Kumar, O. R., Yadav, A. K., Pruthvishree, B. S., Vadhana, P., Nirupama, K. R., Bhardwaj, M. and Singh, B. R. (2019) 'Epidemiology of carbapenem-resistant Escherichia coli and first report of blaVIM carbapenemases gene in calves from India', *Epidemiology and Infection,* 147, pp. e159.

Myllys, V., Ridell, J., Björkroth, J., Biese, I. and Pyörälä, S. (1997) 'Persistence in bovine mastitis of Staphylococcus aureus clones as assessed by random amplified polymorphic DNA analysis, ribotyping and biotyping', *Veterinary microbiology,* 57(2-3), pp. 245-251.

Müller, A. C. and Guido, S. (2016) *Introduction to machine learning with Python: a guide for data scientists.* " O'Reilly Media, Inc.".

Nagel, J. L., Huang, A. M., Kunapuli, A., Gandhi, T. N., Washer, L. L., Lassiter, J., Patel, T. and Newton, D. W. (2014) 'Impact of antimicrobial stewardship intervention on coagulase-negative Staphylococcus blood cultures in conjunction with rapid diagnostic testing', *Journal of clinical microbiology,* 52(8), pp. 2849-2854.

Nagy, B., Szmolka, A., Smole Možina, S., Kovač, J., Strauss, A., Schlager, S., Beutlich, J., Appel, B., Lušicky, M., Aprikian, P., Pászti, J., Tóth, I., Kugler, R. and Wagner, M. (2015) 'Virulence and antimicrobial resistance determinants of verotoxigenic Escherichia coli (VTEC) and of multidrug-resistant E. coli from foods of animal origin illegally imported to the EU by flight passengers', *International Journal of Food Microbiology,* 209, pp. 52-59.

Nakano, S., Matsumura, Y., Ito, Y., Fujisawa, T., Chang, B., Suga, S., Kato, K., Yunoki, T., Hotta, G. and Noguchi, T. (2015) 'Development and evaluation of MALDI-TOF MS-based serotyping for Streptococcus pneumoniae', *European Journal of Clinical Microbiology & Infectious Diseases,* 34(11), pp. 2191-2198.

Nakazawa, M. and Nakazawa, M. M. (2019) 'Package 'fmsb'', *See https://cran. r-project. org/web/packages/fmsb/fmsb. pdf*.

Nanamiya, H. and Kawamura, F. (2010) 'Towards an Elucidation of the Roles of the Ribosome during Different Growth Phases in <I>Bacillus subtilis</I>', *Bioscience, Biotechnology, and Biochemistry,* 74(3), pp. 451-461.

Nascimento, A., Ko, A. I., Martins, E. A. L., Monteiro-Vitorello, C. B., Ho, P. L., Haake, D. A., Verjovski-Almeida, S., Hartskeerl, R. A., Marques, M. V. and Oliveira, M. C. (2004) 'Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis', *Journal of bacteriology,* 186(7), pp. 2164-2172.

Nascimento, J. d. S., Fagundes, P. C., Brito, M. A. V. d. P., Santos, K. R. N. d. and Bastos, M. d. C. d. F. (2005) 'Production of bacteriocins by coagulase-negative staphylococci involved in bovine mastitis', *Veterinary Microbiology,* 106(1), pp. 61-71.

Nasser, W. and Reverchon, S. (2007) 'New insights into the regulatory mechanisms of the LuxR family of quorum sensing regulators', *Analytical and Bioanalytical Chemistry,* 387(2), pp. 381-390.

Natori, Y., Nanamiya, H., Akanuma, G., Kosono, S., Kudo, T., Ochi, K. and Kawamura, F. (2007) 'A fail-safe system for the ribosome under zinc-limiting conditions in Bacillus subtilis', *Molecular Microbiology,* 63(1), pp. 294-307.

Nemeth, J., Muckle, C. A. and Gyles, C. L. (1994) 'In vitro comparison of bovine mastitis and fecal Escherichia coli isolates', *Veterinary Microbiology,* 40(3), pp. 231-238.

Nguyen, R. N., Taylor, L. S., Tauschek, M. and Robins-Browne, R. M. (2006) 'Atypical enteropathogenic Escherichia coli infection and prolonged diarrhea in children', *Emerging infectious diseases,* 12(4), pp. 597-603.

Nickerson, S. C., Owens, W. E., Tomita, G. M. and Widel, P. W. (1999) 'Vaccinating dairy heifers with a Staphylococcus aureus bacterin reduces mastitis at calving', *Large Animal Practice*.

Nisa, S., Bercker, C., Midwinter, A. C., Bruce, I., Graham, C. F., Venter, P., Bell, A., French, N. P., Benschop, J., Bailey, K. M. and Wilkinson, D. A. (2019) 'Combining MALDI-TOF and genomics in the study of methicillin resistant and multidrug resistant Staphylococcus pseudintermedius in New Zealand', *Scientific Reports,* 9(1), pp. 1271.

Nüesch-Inderbinen, M., Käppeli, N., Morach, M., Eicher, C., Corti, S. and Stephan, R. (2019) 'Molecular types, virulence profiles and antimicrobial resistance of Escherichia coli causing bovine mastitis', *Veterinary Record Open,* 6(1), pp. e000369.

O'Rourke, D. (2009) 'Nutrition and udder health in dairy cows: a review', *Irish Veterinary Journal,* 62(4), pp. S15.

Oberle, M., Wohlwend, N., Jonas, D., Maurer, F. P., Jost, G., Tschudin-Sutter, S., Vranckx, K. and Egli, A. (2016) 'The Technical and Biological Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Based Typing: Employment of Bioinformatics in a Multicenter Study', *PLOS ONE,* 11(10), pp. e0164260.

Ochman, H. and Selander, R. K. (1984) 'Standard reference strains of Escherichia coli from natural populations', *Journal of bacteriology,* 157(2), pp. 690-693.

Ojima-Kato, T., Yamamoto, N., Nagai, S., Shima, K., Akiyama, Y., Ota, J. and Tamura, H. (2017) 'Application of proteotyping Strain Solution™ ver. 2 software and theoretically calculated mass database in MALDI-TOF MS typing of Salmonella serotype', *Appl Microbiol Biotechnol,* 101(23-24), pp. 8557-8569.

Olde Riekerink, R. G. M., Barkema, H. W. and Stryhn, H. (2007) 'The Effect of Season on Somatic Cell Count and the Incidence of Clinical Mastitis', *Journal of Dairy Science,* 90(4), pp. 1704-1715.

Olivares Pacheco, J., Bernardini, A., Garcia-Leon, G., Corona, F., Sanchez, M. B. and Martinez, J. (2013) 'The intrinsic resistome of bacterial pathogens', *Frontiers in Microbiology,* 4, pp. 103.

Oliveira, L., Hulland, C. and Ruegg, P. L. (2013) 'Characterization of clinical mastitis occurring in cows on 50 large dairy herds in Wisconsin', *Journal of dairy science,* 96(12), pp. 7538-7549.

Oliver, S. P., Gillespie, B. E., Headrick, S. J., Lewis, M. J. and Dowlen, H. H. (2005) 'Prevalence, risk factors, and strategies for controlling mastitis in heifers during the periparturient period', *Int. J. Appl. Res. Vet. Med,* 3, pp. 150-162.

Oliver, S. P., Gillespie, B. E. and Jayarao, B. M. (1998) 'Detection of new and persistent Streptococcus uberis and Streptococcus dysgalactiae intramammary infections by polymerase chain reaction-based DNA fingerprinting', *FEMS microbiology letters,* 160(1), pp. 69-73.

Oliver, S. P., Pighetti, G. M. and Almeida, R. A. (2011) 'Mastitis Pathogens | Environmental Pathogens', in Fuquay, J.W. (ed.) *Encyclopedia of Dairy Sciences (Second Edition)*. San Diego: Academic Press, pp. 415-421.

Olsen, J. E., Christensen, H. and Aarestrup, F. M. (2006) 'Diversity and evolution of blaZ from Staphylococcus aureus and coagulase-negative staphylococci', *Journal of Antimicrobial Chemotherapy,* 57(3), pp. 450-460.

Olson, M. A., Siebach, T. W., Griffitts, J. S., Wilson, E. and Erickson, D. L. (2018) 'Genome-Wide Identification of Fitness Factors in Mastitis-Associated Escherichia coli', *Applied and Environmental Microbiology,* 84(2), pp. e02190-17.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. and Phillippy, A. M. (2016) 'Mash: fast genome and metagenome distance estimation using MinHash', *Genome biology,* 17(1), pp. 132.

Orskov, I., Orskov, F., Jann, B. and Jann, K. (1977) 'Serology, chemistry, and genetics of O and K antigens of Escherichia coli', *Bacteriological reviews,* 41(3), pp. 667.

Otto, M. (2013) 'Staphylococcal infections: mechanisms of biofilm maturation and detachment as critical determinants of pathogenicity', *Annual review of medicine,* 64, pp. 175-188.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F. and Stevens, R. (2014) 'The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)', *Nucleic Acids Res,* 42(Database issue), pp. D206-14.

Overdevest, I., Heck, M., Van Der Zwaluw, K., Willemsen, I., Van De Ven, J., Verhulst, C. and Kluytmans, J. (2012) 'Comparison of SpectraCell RA Typing and Multilocus Sequence Typing for Extended-Spectrum-β-Lactamase-Producing Escherichia coli', *Journal of clinical microbiology,* 50(12), pp. 3999-4001.

Oyamada, Y., Ito, H., Fujimoto, K., Asada, R., Niga, T., Okamoto, R., Inoue, M. and Yamagishi, J. (2006) 'Combination of known and unknown mechanisms confers high-level resistance to fluoroquinolones in Enterococcus faecium', *J Med Microbiol,* 55(Pt 6), pp. 729-36.

O'Neill, J. (2016) *The Review on Antimicrobial Resistance: Final report and recommendations*. Available at: https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf (Accessed: December, 2020.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A. and Parkhill, J. (2015) 'Roary: rapid large-scale prokaryote pan genome analysis', *Bioinformatics,* 31(22), pp. 3691-3693.

Panchal, G., Ganatra, A., Kosta, Y. P. and Panchal, D. (2011) 'Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers', *International Journal of Computer Theory and Engineering,* 3(2), pp. 332-337.

Paniagua-Contreras, G. L., Monroy-Pérez, E., Díaz-Velásquez, C. E., Uribe-García, A., Labastida, A., Peñaloza-Figueroa, F., Domínguez-Trejo, P., García, L. R., Vaca-Paniagua, F. and Vaca, S. (2019) 'Whole-genome sequence analysis of multidrug-resistant uropathogenic strains of Escherichia coli from Mexico', *Infection and drug resistance,* 12, pp. 2363-2377.

Panina, E. M., Mironov, A. A. and Gelfand, M. S. (2003) 'Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins', *Proceedings of the National Academy of Sciences,* 100(17), pp. 9912-9917.

Pantoja, J. C. F., Hulland, C. and Ruegg, P. L. (2009) 'Somatic cell count status across the dry period as a risk factor for the development of clinical mastitis in the subsequent lactation', *Journal of dairy science,* 92(1), pp. 139-148.

Pantosti, A., Sanchini, A. and Monaco, M. (2007) 'Mechanisms of antibiotic resistance in Staphylococcus aureus'.

Parker, A. M., Shukla, A., House, J. K., Hazelton, M. S., Bosward, K. L., Kokotovic, B. and Sheehy, P. A. (2016) 'Genetic characterization of Australian Mycoplasma bovis isolates through whole genome sequencing analysis', *Veterinary Microbiology,* 196, pp. 118-125.

Pathania, R., Zlitni, S., Barker, C., Das, R., Gerritsma, D. A., Lebert, J., Awuah, E., Melacini, G., Capretta, F. A. and Brown, E. D. (2009) 'Chemical genomics in Escherichia coli identifies an inhibitor of bacterial lipoprotein targeting', *Nature Chemical Biology,* 5, pp. 849.

Paulin-Curlee, G. G., Singer, R. S., Sreevatsan, S., Isaacson, R., Reneau, J., Foster, D. and Bey, R. (2007) 'Genetic diversity of mastitis-associated Klebsiella pneumoniae in dairy cows', *Journal of dairy science,* 90(8), pp. 3681-3689.

Peacock, S. J. and Paterson, G. K. (2015) 'Mechanisms of Methicillin Resistance in Staphylococcus aureus', *Annu Rev Biochem,* 84, pp. 577-601.

Pebesma, E. and Bivand, R. S. (2005) 'S classes and methods for spatial data: the sp package', *R news,* 5(2), pp. 9-13.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) 'Scikit-learn: Machine learning in Python', *Journal of machine learning research,* 12(Oct), pp. 2825-2830.

Peigne, C., Bidet, P., Mahjoub-Messai, F., Plainvert, C., Barbe, V., Medigue, C., Frapy, E., Nassif, X., Denamur, E., Bingen, E. and Bonacorsi, S. (2009) 'The plasmid of Escherichia coli strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic E. coli plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model', *Infect Immun,* 77(6), pp. 2272-84.

Pennerman, K. K., Yin, G., Bennett, W. J. and Hua, T. S.-S. (2019) 'Aspergillus flavus NRRL 35739, a Poor Biocontrol Agent, May Have Increased Relative Expression of Stress Response Genes', *Journal of Fungi,* 5(2).

Pereira, U. P., Oliveira, D. G. S., Mesquita, L. R., Costa, G. M. and Pereira, L. J. (2011) 'Efficacy of Staphylococcus aureus vaccines for bovine mastitis: A systematic review', *Veterinary Microbiology,* 148(2), pp. 117-124.

Pereyre, S., Tardy, F., Renaudin, H., Cauvin, E., Machado, L. D. P. N., Tricot, A., Benoit, F., Treilles, M. and Bebear, C. (2013) 'Identification and subtyping of clinically relevant human and ruminant mycoplasmas by use of matrix-assisted laser desorption ionization–time of flight mass spectrometry', *Journal of Clinical Microbiology,* 51(10), pp. 3314-3323.

Perrig, M. S., Ambroggio, M. B., Buzzola, F. R., Marcipar, I. S., Calvinho, L. F., Veaute, C. M. and Barbagelata, M. S. (2015) 'Genotyping and study of the pauA and sua genes of Streptococcus uberis isolates from bovine mastitis', *Revista Argentina de Microbiología,* 47(4), pp. 282-294.

Perrig, M. S., Veaute, C., Renna, M. S., Pujato, N., Calvinho, L., Marcipar, I. and Barbagelata, M. S. (2017) 'Assessment of the potential utility of different regions of Streptococcus uberis adhesion molecule (SUAM) for mastitis subunit vaccine development', *Microbial Pathogenesis,* 105, pp. 273-279.

Persson, Y., Nyman, A.-K. J. and Grönlund-Andersson, U. (2011) 'Etiology and antimicrobial susceptibility of udder pathogens from cases of subclinical mastitis in dairy cows in Sweden', *Acta Veterinaria Scandinavica,* 53(1), pp. 36.

Pesesky, M. W., Hussain, T., Wallace, M., Patel, S., Andleeb, S., Burnham, C.-A. D. and Dantas, G. (2016) 'Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in Gram-negative Bacilli from whole genome sequence data', *Frontiers in microbiology,* 7, pp. 1887.

Petersen, A. and Dalsgaard, A. (2003) 'Species composition and antimicrobial resistance genes of Enterococcus spp., isolated from integrated and traditional fish farms in Thailand', *Environmental Microbiology,* 5(5), pp. 395-402.

Petersson-Wolfe, C. S., Adams, S., Wolf, S. L. and Hogan, J. S. (2008) 'Genomic typing of enterococci isolated from bovine mammary glands and environmental sources', *Journal of dairy science,* 91(2), pp. 615-619.

Petersson-Wolfe, C. S., Tholen, A. R., Currin, J. and Leslie, K. E. (2013) 'Practical methods for mastitis control', *WCDS Advances in Dairy Technology,* 25, pp. 341-358.

Petersson-Wolfe, C. S., Wolf, S. L. and Hogan, J. S. (2007) 'In Vitro Growth of Enterococci of Bovine Origin in Bovine Mammary Secretions from Various Stages of Lactation1', *Journal of Dairy Science,* 90(9), pp. 4226-4231.

Petersson-Wolfe, C. S., Wolf, S. L. and Hogan, J. S. (2009) 'Experimental challenge of bovine mammary glands with Enterococcus faecium during early and late lactation', *Journal of dairy science,* 92(7), pp. 3158-3164.

Petkau, A., Mabon, P., Sieffert, C., Knox, N. C., Cabral, J., Iskander, M., Iskander, M., Weedmark, K., Zaheer, R. and Katz, L. S. (2017) 'SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology', *Microbial genomics,* 3(6).

Petrella, S., Cambau, E., Chauffour, A., Andries, K., Jarlier, V. and Sougakoff, W. (2006) 'Genetic basis for natural and acquired resistance to the diarylquinoline R207910 in mycobacteria', *Antimicrobial agents and chemotherapy,* 50(8), pp. 2853-2856.

Petrovski, K. R., Buneski, G. and Trajcev, M. (2006) 'A review of the factors affecting the costs of bovine mastitis', *Journal of the South African Veterinary Association,* 77(2), pp. 52-60.

Petrovski, K. R., Heuer, C., Parkinson, T. J. and Williamson, N. B. (2009) 'The incidence and aetiology of clinical bovine mastitis on 14 farms in Northland, New Zealand', *New Zealand veterinary journal,* 57(2), pp. 109-115.

Petrovski, K. R., Laven, R. A. and Lopez-Villalobos, N. (2011) 'A descriptive analysis of the antimicrobial susceptibility of mastitis-causing bacteria isolated from samples submitted to commercial diagnostic laboratories in New Zealand (2003–2006)', *New Zealand Veterinary Journal,* 59(2), pp. 59-66.

Petta, I., Lievens, S., Libert, C., Tavernier, J. and De Bosscher, K. (2016) 'Modulation of Protein–Protein Interactions for the Development of Novel Therapeutics', *Molecular Therapy,* 24(4), pp. 707-718.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. (2004) 'UCSF Chimera--a visualization system for exploratory research and analysis', *J Comput Chem,* 25(13), pp. 1605-12.

Phuektes, P., Mansell, P. D., Dyson, R. S., Hooper, N. D., Dick, J. S. and Browning, G. F. (2001) 'Molecular epidemiology of Streptococcus uberis isolates from dairy cows with mastitis', *Journal of clinical microbiology,* 39(4), pp. 1460-1466.

Piccinini, R., Borromeo, V. and Zecconi, A. (2010) 'Relationship between S. aureus gene pattern and dairy herd mastitis prevalence', *Veterinary Microbiology,* 145(1), pp. 100-105.

Pichette-Jolette, S., Millette, G., Demontier, E., Bran-Barrera, D., Cyrenne, M., Ster, C., Haine, D., Keefe, G., Malouin, F. and Roy, J. P. (2019) 'Partial prediction of the duration and the clinical status of Staphylococcus aureus bovine intramammary infections based on the phenotypic and genotypic analysis of isolates', *Veterinary microbiology,* 228, pp. 188-195.

Piepers, S., Peeters, K., Opsomer, G., Barkema, H. W., Frankena, K. and De Vliegher, S. (2011) 'Pathogen group specific risk factors at herd, heifer and quarter levels for intramammary infections in early lactating dairy heifers', *Preventive Veterinary Medicine,* 99(2), pp. 91-101.

Piessens, V., Van Coillie, E., Verbist, B., Supré, K., Braem, G., Van Nuffel, A., De Vuyst, L., Heyndrickx, M. and De Vliegher, S. (2011) 'Distribution of coagulase-negative Staphylococcus species from milk and environment of dairy cows differs between herds', *Journal of Dairy Science,* 94(6), pp. 2933-2944.

Pitkälä, A., Haveri, M., Pyörälä, S., Myllys, V. and Honkanen-Buzalski, T. (2004) 'Bovine mastitis in Finland 2001—prevalence, distribution of bacteria, and antimicrobial resistance', *Journal of dairy science,* 87(8), pp. 2433-2441.

Poelarends, J. J., Hogeveen, H., Sampimon, O. C. and Sol, J. (2001) 'Monitoring subclinical mastitis in Dutch dairy herds'.

Prado, M. E., Almeida, R. A., Ozen, C., Luther, D. A., Lewis, M. J., Headrick, S. J. and Oliver, S. P. (2011) 'Vaccination of dairy cows with recombinant Streptococcus uberis adhesion molecule induces antibodies that reduce adherence to and internalization of S. uberis into bovine mammary epithelial cells', *Veterinary Immunology and Immunopathology,* 141(3), pp. 201-208.

Prati, R. C., Batista, G. E. and Monard, M. C. 'Data mining with imbalanced class distributions: concepts and methods'. 2009, 359-376.

Price, L. B., Stegger, M., Hasman, H., Aziz, M., Larsen, J., Andersen, P. S., Pearson, T., Waters, A. E., Foster, J. T., Schupp, J., Gillece, J., Driebe, E., Liu, C. M., Springer, B., Zdovc, I., Battisti, A., Franco, A., Żmudzki, J., Schwarz, S., Butaye, P., Jouy, E., Pomba, C., Porrero, M. C., Ruimy, R., Smith, T. C., Robinson, D. A., Weese, J. S., Arriola, C. S., Yu, F., Laurent, F., Keim, P., Skov, R. and Aarestrup, F. M. (2012) 'Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock', *mBio,* 3(1), pp. e00305-11.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2009) 'FastTree: computing large minimum evolution trees with profiles instead of a distance matrix', *Mol Biol Evol,* 26(7), pp. 1641-50.

Pusztai, L., Gregory, B. W., Baggerly, K. A., Peng, B., Koomen, J., Kuerer, H. M., Esteva, F. J., Symmans, W. F., Wagner, P. and Hortobagyi, G. N. (2004) 'Pharmacoproteomic analysis of prechemotherapy

and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma', *Cancer: Interdisciplinary International Journal of the American Cancer Society,* 100(9), pp. 1814-1822.

Pyörälä, S. and Taponen, S. (2009) 'Coagulase-negative staphylococci—Emerging mastitis pathogens', *Veterinary Microbiology,* 134(1), pp. 3-8.

Pérez, V. K. C., Costa, G. M. d., Guimarães, A. S., Heinemann, M. B., Lage, A. P. and Dorneles, E. M. S. (2020a) 'Relationship between virulence factors and antimicrobial resistance in Staphylococcus aureus from bovine mastitis', *Journal of Global Antimicrobial Resistance,* 22, pp. 792-802.

Pérez, V. K. C., Custódio, D. A. C., Silva, E. M. M., de Oliveira, J., Guimarães, A. S., Brito, M. A. V. P., Souza-Filho, A. F., Heinemann, M. B., Lage, A. P. and Dorneles, E. M. S. (2020b) 'Virulence factors and antimicrobial resistance in Staphylococcus aureus isolated from bovine mastitis in Brazil', *Brazilian Journal of Microbiology*.

R Core Team (2019) 'R: A language and environment for statistical computing', *R Foundation for Statistical Computing*.

Rainard, P., Foucras, G., Fitzgerald, J. R., Watts, J. L., Koop, G. and Middleton, J. R. (2018) 'Knowledge gaps and research priorities in Staphylococcus aureus mastitis control', *Transboundary and Emerging Diseases,* 65(S1), pp. 149-165.

Raivio, T. L., Leblanc, S. K. D. and Price, N. L. (2013) 'The Escherichia coli Cpx envelope stress response regulates genes of diverse function that impact antibiotic resistance and membrane integrity', *Journal of bacteriology,* 195(12), pp. 2755-2767.

Raizada, R. and Lee, Y. (2013) 'Smoothness without Smoothing: Why Gaussian Naive Bayes Is Not Naive for Multi-Subject Searchlight Studies', *PloS one,* 8, pp. e69566.

Ralhan, R., Desouza, L. V., Matta, A., Tripathi, S. C., Ghanny, S., Datta Gupta, S., Bahadur, S. and Siu, K. W. (2008) 'Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry', *Mol Cell Proteomics,* 7(6), pp. 1162-73.

Ramstein, J., Hervouet, N., Coste, F., Zelwer, C., Oberto, J. and Castaing, B. (2003) 'Evidence of a thermal unfolding dimeric intermediate for the Escherichia coli histone-like HU proteins: thermodynamics and structure', *J Mol Biol,* 331(1), pp. 101-21.

Randall, L. P., Lemma, F., Koylass, M., Rogers, J., Ayling, R. D., Worth, D., Klita, M., Steventon, A., Line, K., Wragg, P., Muchowski, J., Kostrzewa, M. and Whatmore, A. M. (2015) 'Evaluation of MALDI-ToF as a method for the identification of bacteria in the veterinary diagnostic laboratory', *Res Vet Sci,* 101, pp. 42-9.

Rao Reddy Neelapu, N. and Pavani, T. (2013) 'Identification of novel drug targets in HpB38, HpP12, HpG27, Hpshi470, HpSJM180 strains of Helicobacter pylori: an in silico approach for therapeutic intervention', *Current drug targets,* 14(5), pp. 601-611.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002) 'Hierarchical organization of modularity in metabolic networks', *Science,* 297(5586), pp. 1551-5.

Rebuffat, S. 2012. Microcins in action: amazing defence strategies of Enterobacteria. Portland Press Limited.

Reinoso, E., Bettera, S., Frigerio, C., DiRenzo, M., Calzolari, A. and Bogni, C. (2004) 'RAPD-PCR analysis of Staphylococcus aureus strains isolated from bovine and human hosts', *Microbiological research,* 159(3), pp. 245-255.

Reksen, O., Solverod, L., Branscum, A. J. and Osteras, O. (2006) 'Relationships between milk culture results and treatment for clinical mastitis or culling in Norwegian dairy cattle', *J Dairy Sci,* 89(8), pp. 2928-37.

Reller, L. B., Weinstein, M., Jorgensen, J. H. and Ferraro, M. J. (2009) 'Antimicrobial Susceptibility Testing: A Review of General Principles and Contemporary Practices', *Clinical Infectious Diseases,* 49(11), pp. 1749-1755.

Ressom, H. W., Varghese, R. S., Drake, S. K., Hortin, G. L., Abdel-Hamid, M., Loffredo, C. A. and Goldman, R. (2007) 'Peak selection from MALDI-TOF mass spectra using ant colony optimization', *Bioinformatics,* 23(5), pp. 619-626.

Reyher, K. K., Dufour, S., Barkema, H. W., Des Côteaux, L., Devries, T. J., Dohoo, I. R., Keefe, G. P., Roy, J. P. and Scholl, D. T. (2011) 'The National Cohort of Dairy Farms—A data collection platform for mastitis research in Canada', *Journal of dairy science,* 94(3), pp. 1616-1626.

Reyher, K. K., Haine, D., Dohoo, I. R. and Revie, C. W. (2012) 'Examining the effect of intramammary infections with minor mastitis pathogens on the acquisition of new intramammary infections with major mastitis pathogens—A systematic review and meta-analysis', *Journal of Dairy Science,* 95(11), pp. 6483-6502.

Rhee, S. Y., Wood, V., Dolinski, K. and Draghici, S. (2008) 'Use and misuse of the gene ontology annotations', *Nature Reviews Genetics,* 9(7), pp. 509-515.

Riaboff, L., Poggi, S., Madouasse, A., Couvreur, S., Aubin, S., Bédère, N., Goumand, E., Chauvin, A. and Plantier, G. (2020) 'Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data', 169.

Rice, L. B., Carias, L. L., Rudin, S., Hutton, R., Marshall, S., Hassan, M., Josseaume, N., Dubost, L., Marie, A. and Arthur, M. (2009) 'Role of class A penicillin-binding proteins in the expression of beta-lactam resistance in Enterococcus faecium', *J Bacteriol,* 191(11), pp. 3649-56.

Richards, V. P., Lefébure, T., Pavinski Bitar, P. D., Dogan, B., Simpson, K. W., Schukken, Y. H. and Stanhope, M. J. (2015) 'Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic Escherichia coli', *PloS one,* 10(3), pp. e0119799-e0119799.

Richardson, E. J., Bacigalupe, R., Harrison, E. M., Weinert, L. A., Lycett, S., Vrieling, M., Robb, K., Hoskisson, P. A., Holden, M. T. G., Feil, E. J., Paterson, G. K., Tong, S. Y. C., Shittu, A., van Wamel, W., Aanensen, D. M., Parkhill, J., Peacock, S. J., Corander, J., Holmes, M. and Fitzgerald, J. R. (2018) 'Gene exchange drives the ecological success of a multi-host bacterial pathogen', *Nature Ecology & Evolution,* 2(9), pp. 1468-1478.

Riekerink, R. G. M. O., Barkema, H. W., Kelton, D. F. and Scholl, D. T. (2008) 'Incidence rate of clinical mastitis on Canadian dairy farms', *Journal of dairy science,* 91(4), pp. 1366-1377.

Riekerink, R. G. M. O., Barkema, H. W., Scholl, D. T., Poole, D. E. and Kelton, D. F. (2010) 'Management practices associated with the bulk-milk prevalence of Staphylococcus aureus in Canadian dairy farms', *Preventive Veterinary Medicine,* 97(1), pp. 20-28.

Rishishwar, L., Petit, R. A., Kraft, C. S. and Jordan, I. K. (2014) 'Genome sequence-based discriminator for vancomycin-intermediate Staphylococcus aureus', *Journal of bacteriology,* 196(5), pp. 940-948.

Rizzardi, K., Wahab, T. and Jernberg, C. (2013) 'Rapid subtyping of Yersinia enterocolitica by matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) for diagnostics and surveillance', *Journal of clinical microbiology,* 51(12), pp. 4200-4203.

Rizzotti, L., La Gioia, F., Dellaglio, F. and Torriani, S. (2009) 'Molecular diversity and transferability of the tetracycline resistance gene tet (M), carried on Tn916-1545 family transposons, in enterococci from a total food chain', *Antonie Van Leeuwenhoek,* 96(1), pp. 43-52.

Rizzotti, L., Simeoni, D., Cocconcelli, P., Gazzola, S., Dellaglio, F. and Torriani, S. (2005) 'Contribution of Enterococci to the Spread of Antibiotic Resistance in the Production Chain of Swine Meat Commodities', *Journal of Food Protection,* 68(5), pp. 955-965.

Roberts, M. C. (2008) 'Update on macrolide–lincosamide–streptogramin, ketolide, and oxazolidinone resistance genes', *FEMS Microbiology Letters,* 282(2), pp. 147-159.

Roberts, M. C., Sutcliffe, J., Courvalin, P., Jensen, L. B., Rood, J. and Seppala, H. (1999) 'Nomenclature for macrolide and macrolide-lincosamide-streptogramin B resistance determinants', *Antimicrobial agents and chemotherapy,* 43(12), pp. 2823-2830.

Rocha, L. S., Silva, D. M., Silva, M. P., Vidigal, P. M. P., Silva, J. C. F., Guerra, S. T., Ribeiro, M. G., Mendes, T. A. d. O. and Ribon, A. d. O. B. (2019) 'Comparative genomics of Staphylococcus aureus associated with subclinical and clinical bovine mastitis', *PLOS ONE,* 14(8), pp. e0220804.

Rodrigues, C., Passet, V., Rakotondrasoa, A. and Brisse, S. (2018) 'Identification of Klebsiella pneumoniae, Klebsiella quasipneumoniae, Klebsiella variicola and related phylogroups by MALDI-TOF mass spectrometry', *Frontiers in microbiology,* 9, pp. 3000.

Rodrigues, N. M. B., Bronzato, G. F., Santiago, G. S., Botelho, L. A. B., Moreira, B. M., Coelho, I. d. S., Souza, M. M. S. d. and Coelho, S. d. M. d. O. (2017) 'The Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry (MALDI-TOF MS) identification versus biochemical tests: a study with enterobacteria from a dairy cattle environment', *Brazilian Journal of Microbiology,* 48(1), pp. 132-138.

Rodriguez-Siek, K. E., Giddings, C. W., Doetkott, C., Johnson, T. J. and Nolan, L. K. (2005) 'Characterizing the APEC pathotype', *Veterinary research,* 36(2), pp. 241-256.

Roer, L., Johannesen, T. B., Hansen, F., Stegger, M., Tchesnokova, V., Sokurenko, E., Garibay, N., Allesøe, R., Thomsen, M. C. F., Lund, O., Hasman, H. and Hammerum, A. M. (2018) 'CHTyper, a Web Tool for Subtyping of Extraintestinal Pathogenic Escherichia coli Based on the fumC and fimH Alleles', *Journal of Clinical Microbiology,* 56(4), pp. e00063-18.

Ronco, T., Klaas, I. C., Stegger, M., Svennesen, L., Astrup, L. B., Farre, M. and Pedersen, K. (2018) 'Genomic investigation of Staphylococcus aureus isolates from bulk tank milk and dairy cows with clinical mastitis', *Veterinary Microbiology,* 215, pp. 35-42.

Rossitto, P. V., Ruiz, L., Kikuchi, Y., Glenn, K., Luiz, K., Watts, J. L. and Cullor, J. S. (2002) 'Antibiotic susceptibility patterns for environmental streptococci isolated from bovine mastitis in central California dairies', *Journal of dairy science,* 85(1), pp. 132-138.

Rowland, S. L., Wadsworth, K. D., Robson, S. A., Robichon, C., Beckwith, J. and King, G. F. (2010) 'Evidence from artificial septal targeting and site-directed mutagenesis that residues in the extracytoplasmic beta domain of DivIB mediate its interaction with the divisomal transpeptidase PBP 2B', *J Bacteriol,* 192(23), pp. 6116-25.

Roy, C. R. and Cherfils, J. (2015) 'Structure and function of Fic proteins', *Nature Reviews Microbiology,* 13, pp. 631.

Ruegg, P. L. 'The application of evidence based veterinary medicine to mastitis therapy'. 2010, 78-93.

Ruegg, P. L. (2017) 'A 100-Year Review: Mastitis detection, management, and prevention', *Journal of Dairy Science,* 100(12), pp. 10381-10397.

Ruegg, P. L., Oliveira, L., Jin, W. and Okwumabua, O. (2015) 'Phenotypic antimicrobial susceptibility and occurrence of selected resistance genes in gram-positive mastitis pathogens isolated from Wisconsin dairy cows', *J Dairy Sci,* 98(7), pp. 4521-34.

Ruiz-Cruz, S., Espinosa, M., Goldmann, O. and Bravo, A. (2016) 'Global Regulation of Gene Expression by the MafR Protein of Enterococcus faecalis', *Frontiers in Microbiology,* 6, pp. 1521.

Ruiz-Garbajosa, P., Bonten, M. J., Robinson, D. A., Top, J., Nallapareddy, S. R., Torres, C., Coque, T. M., Cantón, R., Baquero, F., Murray, B. E., del Campo, R. and Willems, R. J. (2006) 'Multilocus sequence typing scheme for Enterococcus faecalis reveals hospital-adapted genetic complexes in a background of high rates of recombination', *J Clin Microbiol,* 44(6), pp. 2220-8.

Russell, S. J. and Norvig, P. (2010) *Artificial Intelligence: A Modern Approach.* Third Edition edn.: Prentice Hall.

Ryman, V. E., Nickerson, S. C., Hurley, D. J., Berghaus, R. D. and Kautz, F. M. (2013) 'Influence of horn flies (Haematobia irritans) on teat skin condition, intramammary infection, and serum anti-S. aureus antibody titres in holstein heifers', *Research in Veterinary Science,* 95(2), pp. 343-346.

Rysanek, D., Zouharova, M. and Babak, V. (2009) 'Monitoring major mastitis pathogens at the population level based on examination of bulk tank milk samples', *Journal of Dairy Research,* 76(1), pp. 117-123.

Ryzhov, V. and Fenselau, C. (2001) 'Characterization of the protein subset desorbed by MALDI from whole bacterial cells', *Analytical chemistry,* 73(4), pp. 746-750.

Różańska, H., Lewtak-Piłat, A., Kubajka, M. and Weiner, M. (2019) 'Occurrence of Enterococci in Mastitic Cow's Milk and their Antimicrobial Resistance', *Journal of veterinary research,* 63(1), pp. 93-97.

Sabença, C., Sousa, T. d., Oliveira, S., Viala, D., Theron, L., Chambon, C., Hébraud, M., Beyrouthy, R., Bonnet, R. and Caniça, M. (2020) 'Next-generation sequencing and MALDI mass spectrometry in the study of multiresistant processed meat vancomycin-resistant enterococci (VRE)', *Biology,* 9(5), pp. 89.

Sadowy, E. and Luczkiewicz, A. (2014) 'Drug-resistant and hospital-associated Enterococcus faecium from wastewater, riverine estuary and anthropogenically impacted marine catchment basin', *BMC Microbiology,* 14(1), pp. 66.

Sakwinska, O., Giddey, M., Moreillon, M., Morisset, D., Waldvogel, A. and Moreillon, P. (2011) 'Staphylococcus aureus Host Range and Human-Bovine Host Shift', *Applied and Environmental Microbiology,* 77(17), pp. 5908.

Salinas, L., Cárdenas, P., Johnson, T. J., Vasco, K., Graham, J. and Trueba, G. (2019) 'Diverse Commensal Escherichia coli Clones and Plasmids Disseminate Antimicrobial Resistance Genes in Domestic Animals and Children in a Semirural Community in Ecuador', *mSphere,* 4(3), pp. e00316-19.

Sambongi, Y., Iko, Y., Tanabe, M., Omote, H., Iwamoto-Kihara, A., Ueda, I., Yanagida, T., Wada, Y. and Futai, M. (1999) 'Mechanical rotation of the c subunit oligomer in ATP synthase (F0F1): direct observation', *Science,* 286(5445), pp. 1722-1724.

Sandegren, L., Lindqvist, A., Kahlmeter, G. and Andersson, D. I. (2008) 'Nitrofurantoin resistance mechanism and fitness cost in Escherichia coli', *Journal of Antimicrobial Chemotherapy,* 62(3), pp. 495-503.

Sangaiah, A. K. (2019) *Deep Learning and Parallel Computing Environment for Bioengineering Systems.* Academic Press.

Sauget, M., Valot, B., Bertrand, X. and Hocquet, D. (2017) 'Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria?', *Trends in Microbiology,* 25(6), pp. 447-455.

Saunders, L. and Pezeshki, R. (2015) 'Glyphosate in runoff waters and in the root-zone: a review', *Toxics,* 3(4), pp. 462-480.

Sauvage, E., Kerff, F., Fonzé, E., Herman, R., Schoot, B., Marquette, J. P., Taburet, Y., Prevost, D., Dumas, J., Leonard, G., Stefanic, P., Coyette, J. and Charlier, P. (2002) 'The 2.4-Å crystal structure of the penicillin-resistant penicillin-binding protein PBP5fm from Enterococcus faecium in complex with benzylpenicillin', *Cellular and Molecular Life Sciences CMLS,* 59(7), pp. 1223-1232.

Savage, E., Chothe, S., Lintner, V., Pierre, T., Matthews, T., Kariyawasam, S., Miller, D., Tewari, D. and Jayarao, B. (2017) 'Evaluation of three bacterial identification systems for species identification of bacteria isolated from bovine mastitis and bulk tank milk samples', *Foodborne pathogens and disease,* 14(3), pp. 177-187.

Sawant, A. A., Gillespie, B. E. and Oliver, S. P. (2009) 'Antimicrobial susceptibility of coagulase-negative Staphylococcus species isolated from bovine milk', *Veterinary microbiology,* 134(1-2), pp. 73-81.

Scali, F., Camussone, C., Calvinho, L. F., Cipolla, M. and Zecconi, A. (2015) 'Which are important targets in development of S. aureus mastitis vaccine?', *Res Vet Sci,* 100, pp. 88-99.

Schabauer, L., Wenning, M., Huber, I. and Ehling-Schulz, M. (2014) 'Novel physico-chemical diagnostic tools for high throughput identification of bovine mastitis associated gram-positive, catalase-negative cocci', *BMC Veterinary Research,* 10(1), pp. 156.

Schaellibaum, D. M. (1999) 'Mastitis pathogens isolated in Switzerland 1987-1996', *International Dairy Federation*.

Schapire, R. E. and Freund, Y. (2013) 'Boosting: Foundations and algorithms', *Kybernetes*.

Schaumann, R., Knoop, N., Genzel, G. H., Losensky, K., Rosenkranz, C., Stîngu, C. S., Schellenberger, W., Rodloff, A. C. and Eschrich, K. (2012) 'A step towards the discrimination of beta-lactamase-

producing clinical isolates of Enterobacteriaceae and Pseudomonas aeruginosa by MALDI-TOF mass spectrometry', *Medical science monitor : international medical journal of experimental and clinical research,* 18(9), pp. MT71-MT77.

Scheutz, F., Cheasty, T., Woodward, D. and Smith, H. R. (2004) 'Designation of O174 and O175 to temporary O groups OX3 and OX7, and six new E. coli O groups that include Verocytotoxin-producing E. coli (VTEC): O176, O177, O178, O179, O180 and O181', *APMIS,* 112(9), pp. 569-84.

Schlotter, K., Ehricht, R., Hotzel, H., Monecke, S., Pfeffer, M. and Donat, K. (2012) 'Leukocidin genes lukF-P83 and lukM are associated with taphylococcus aureus clonal complexes 151, 479 and 133 isolated from bovine udder infections in Thuringia, Germany', *Veterinary research,* 43(1), pp. 42.

Schlunzen, F., Zarivach, R., Harms, J., Bashan, A., Tocilj, A., Albrecht, R., Yonath, A. and Franceschi, F. (2001) 'Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria', *Nature,* 413(6858), pp. 814-21.

Schmidtchen, A., Frick, I. M., Andersson, E., Tapper, H. and Björck, L. (2002) 'Proteinases of common pathogenic bacteria degrade and inactivate the antibacterial peptide LL-37', *Molecular microbiology,* 46(1), pp. 157-168.

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L. and Segata, N. (2016) 'Strain-level microbial epidemiology and population genomics from shotgun metagenomics', *Nat Methods,* 13(5), pp. 435-8.

Schubert, S. and Kostrzewa, M. (2017) 'MALDI-TOF MS in the microbiology laboratory: current trends', *Curr Issues Mol Biol,* 23, pp. 17-20.

Schukken, Y., Chuff, M., Moroni, P., Gurjar, A., Santisteban, C., Welcome, F. and Zadoks, R. (2012) 'The “other” gram-negative bacteria in mastitis: Klebsiella, serratia, and more', *Veterinary Clinics: Food Animal Practice,* 28(2), pp. 239-256.

Schukken, Y. H., Günther, J., Fitzpatrick, J., Fontaine, M. C., Goetze, L., Holst, O., Leigh, J., Petzl, W., Schuberth, H. J., Sipka, A., Smith, D. G. E., Quesnell, R., Watts, J., Yancey, R., Zerbe, H., Gurjar, A., Zadoks, R. N. and Seyfert, H. M. (2011) 'Host-response patterns of intramammary infections in dairy cows', *Veterinary Immunology and Immunopathology,* 144(3), pp. 270-289.

Schwartz, D. J., Kalas, V., Pinkner, J. S., Chen, S. L., Spaulding, C. N., Dodson, K. W. and Hultgren, S. J. (2013) 'Positively selected FimH residues enhance virulence during urinary tract infection by altering FimH conformation', *Proceedings of the National Academy of Sciences,* 110(39), pp. 15530-15537.

Schwarz, F. V., Perreten, V. and Teuber, M. (2001) 'Sequence of the 50-kb Conjugative Multiresistance Plasmid pRE25 from Enterococcus faecalis RE25', *Plasmid,* 46(3), pp. 170-187.

Schwarz, S., Kehrenberg, C., Doublet, B. and Cloeckaert, A. (2004) 'Molecular basis of bacterial resistance to chloramphenicol and florfenicol', *FEMS microbiology reviews,* 28(5), pp. 519-542.

Schwarz, S., Silley, P., Simjee, S., Woodford, N., van Duijkeren, E., Johnson, A. P. and Gaastra, W. (2010) 'Assessing the antimicrobial susceptibility of bacteria obtained from animals', *Vet Microbiol: Vol. 1-2*. Netherlands, pp. 1-4.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., Koonin, E. V. and Altschul, S. F. (2001) 'Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements', *Nucleic acids research,* 29(14), pp. 2994-3005.

Schürch, A. C., Arredondo-Alonso, S., Willems, R. J. L. and Goering, R. V. (2018) 'Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches', *Clinical Microbiology and Infection,* 24(4), pp. 350-354.

Seemann, T. (2014) 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics,* 30(14), pp. 2068-9.

Seemann, T. 2015. Snippy: fast bacterial variant calling from NGS reads.

Seenama, C., Thamlikitkul, V. and Ratthawongjirakul, P. (2019) 'Multilocus sequence typing and blaESBL characterization of extended-spectrum beta-lactamase-producing Escherichia coli isolated from healthy humans and swine in Northern Thailand', *Infection and Drug Resistance,* 12, pp. 2201.

Seib, K. L., Jen, F. E. C., Tan, A., Scott, A. L., Kumar, R., Power, P. M., Chen, L.-T., Wu, H.-J., Wang, A. H. J. and Hill, D. M. C. (2015) 'Specificity of the ModA11, ModA12 and ModD1 epigenetic regulator N6-adenine DNA methyltransferases of Neisseria meningitidis', *Nucleic acids research,* 43(8), pp. 4150-4162.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N. and Whittam, T. S. (1986) 'Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics', *Appl Environ Microbiol,* 51(5), pp. 873-84.

Sengupta, S. and Nagaraja, V. (2008) 'YacG from Escherichia coli is a specific endogenous inhibitor of DNA gyrase', *Nucleic Acids Research,* 36(13), pp. 4310-4316.

Serra, J. (1983) *Image analysis and mathematical morphology.* Academic Press, Inc.

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P. and Weigel, K. (2014) 'Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms', *Journal of dairy science,* 97(2), pp. 731-742.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R. and Khovanova, N. (2019) 'Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation', *Biomedical Signal Processing and Control,* 52, pp. 456-462.

Shamila-Syuhada, A. K., Rusul, G., Wan-Nadiah, W. A. and Chuah, L.-O. (2016) 'Prevalence and Antibiotics Resistance of Staphylococcus aureus Isolates Isolated from Raw Milk Obtained from Small-Scale Dairy Farms in Penang, Malaysia', *Pakistan Veterinary Journal,* 36(1).

Shang, Y. (2014) 'Unveiling robustness and heterogeneity through percolation triggered by random-link breakdown', *Physical Review E,* 90(3), pp. 032820.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome research,* 13(11), pp. 2498-2504.

Sharaha, U., Rodriguez-Diaz, E., Sagi, O., Riesenberg, K., Lapidot, I., Segal, Y., Bigio, I. J., Huleihel, M. and Salman, A. (2019) 'Detection of Extended-Spectrum β-Lactamase-Producing Escherichia coli Using Infrared Microscopy and Machine-Learning Algorithms', *Analytical chemistry,* 91(3), pp. 2525-2530.

Sharma-Kuinkel, B. K., Mann, E. E., Ahn, J.-S., Kuechenmeister, L. J., Dunman, P. M. and Bayles, K. W. (2009) 'The Staphylococcus aureus LytSR two-component regulatory system affects biofilm formation', *Journal of bacteriology,* 191(15), pp. 4767-4775.

Sharma-Kuinkel, B. K., Rude, T. H. and Fowler, V. G., Jr. (2016) 'Pulse Field Gel Electrophoresis', *Methods in molecular biology (Clifton, N.J.),* 1373, pp. 117-130.

Sharp, P. M. (1994) 'Identification of genes encoding ribosomal protein L33 from Bacillus licheniformis, Thermus thermophilus and Thermotoga maritima', *Gene,* 139(1), pp. 135-136.

Sheikh, J., Dudley, E. G., Sui, B., Tamboura, B., Suleman, A. and Nataro, J. P. (2006) 'EilA, a HilA-like regulator in enteroaggregative Escherichia coli', *Molecular Microbiology,* 61(2), pp. 338-350.

Shin, H., Sheu, B., Joseph, M. and Markey, M. K. (2008) 'Guilt-by-association feature selection: identifying biomarkers from proteomic profiles', *Journal of biomedical informatics,* 41(1), pp. 124-136.

Shoji, S., Dambacher, C. M., Shajani, Z., Williamson, J. R. and Schultz, P. G. (2011) 'Systematic Chromosomal Deletion of Bacterial Ribosomal Protein Genes', *Journal of Molecular Biology,* 413(4), pp. 751-761.

Shpigel, N. Y., Elazar, S. and Rosenshine, I. (2008) 'Mammary pathogenic Escherichia coli', *Curr Opin Microbiol,* 11(1), pp. 60-5.

Shum, L. W. C., McConnel, C. S., Gunn, A. A. and House, J. K. (2009) 'Environmental mastitis in intensive high-producing dairy herds in New South Wales', *Australian Veterinary Journal,* 87(12), pp. 469-475.

Sifaoui, F., Arthur, M., Rice, L. and Gutmann, L. (2001) 'Role of penicillin-binding protein 5 in expression of ampicillin resistance and peptidoglycan structure in Enterococcus faecium', *Antimicrob Agents Chemother,* 45(9), pp. 2594-7.

Silva, I. B. V. d. and Adeodato, P. J. L. 'PCA and Gaussian noise in MLP neural network training improve generalization in problems with small and unbalanced data sets'. *The 2011 International Joint Conference on Neural Networks*, 31 July-5 Aug. 2011, 2664-2669.

Singh, S., Goswami, P., Singh, R. and Heller, K. J. (2009) 'Application of molecular identification tools for Lactobacillus, with a focus on discrimination between closely related species: A review', *LWT - Food Science and Technology,* 42(2), pp. 448-457.

Smati, M., Clermont, O., Le Gal, F., Schichmanoff, O., Jauréguy, F., Eddi, A., Denamur, E. and Picard, B. (2013) 'Real-time PCR for quantitative analysis of human commensal Escherichia coli populations reveals a high frequency of subdominant phylogroups', *Appl. Environ. Microbiol.,* 79(16), pp. 5005-5012.

Smith, K. L. and Hogan, J. S. (1993) 'Environmental mastitis', *Vet Clin North Am Food Anim Pract,* 9(3), pp. 489-98.

Smith, M. R., Martinez, T. and Giraud-Carrier, C. (2014) 'An instance level analysis of data complexity', *Machine learning,* 95(2), pp. 225-256.

Smith, T. C., Gebreyes, W. A., Abley, M. J., Harper, A. L., Forshey, B. M., Male, M. J., Martin, H. W., Molla, B. Z., Sreevatsan, S., Thakur, S., Thiruvengadam, M. and Davies, P. R. (2013) 'Methicillin-

Resistant Staphylococcus aureus in Pigs and Farm Workers on Conventional and Antibiotic-Free Swine Farms in the USA', *PLOS ONE,* 8(5), pp. e63704.

Sobisch, L.-Y., Rogowski, K. M., Fuchs, J., Schmieder, W., Vaishampayan, A., Oles, P., Novikova, N. and Grohmann, E. (2019) 'Biofilm Forming Antibiotic Resistant Gram-Positive Pathogens Isolated From Surfaces on the International Space Station', *Frontiers in microbiology,* 10, pp. 543-543.

Sogawa, K., Watanabe, M., Ishige, T., Segawa, S., Miyabe, A., Murata, S., Saito, T., Sanda, A., Furuhata, K. and Nomura, F. (2017) 'Rapid Discrimination between Methicillin-Sensitive and Methicillin-Resistant Staphylococcus aureus Using MALDI-TOF Mass Spectrometry', *Biocontrol Science,* 22(3), pp. 163-169.

Sol, J., Sampimon, O. C., Barkema, H. W. and Schukken, Y. H. (2000) 'Factors associated with cure after therapy of clinical mastitis caused by Staphylococcus aureus', *Journal of dairy science,* 83(2), pp. 278-284.

Soni, J., Ansari, U., Sharma, D. and Soni, S. (2011) 'Predictive data mining for medical diagnosis: An overview of heart disease prediction', *International Journal of Computer Applications,* 17(8), pp. 43-48.

Sordelli, D. O., Buzzola, F. R., Gomez, M. I., Steele-Moore, L., Berg, D., Gentilini, E., Catalano, M., Reitz, A. J., Tollersrud, T., Denamiel, G., Jeric, P. and Lee, J. C. (2000) 'Capsule Expression by Bovine Isolates ofStaphylococcus aureus from Argentina: Genetic and Epidemiologic Analyses', *Journal of Clinical Microbiology,* 38(2), pp. 846.

Sousa, T., Viala, D., Théron, L., Chambon, C., Hébraud, M., Poeta, P. and Igrejas, G. (2020) 'Putative Protein Biomarkers of Escherichia coli Antibiotic Multiresistance Identified by MALDI Mass Spectrometry', *Biology (Basel),* 9(3).

Spohr, M., Rau, J., Friedrich, A., Klittich, G., Fetsch, A., Guerra, B., Hammerl, J. A. and Tenhagen, B. A. (2011) 'Methicillin-Resistant Staphylococcus aureus (MRSA) in Three Dairy Herds in Southwest Germany', *Zoonoses and Public Health,* 58(4), pp. 252-261.

Stanford, K., Reuter, T., Hallewell, J., Tostes, R., Alexander, T. W. and McAllister, T. A. (2018) 'Variability in Characterizing Escherichia coli from Cattle Feces: A Cautionary Tale', *Microorganisms,* 6(3).

Steeneveld, W., van der Gaag, L. C., Barkema, H. W. and Hogeveen, H. (2009) 'Providing probability distributions for the causal pathogen of clinical mastitis using naive Bayesian networks', *Journal of Dairy Science,* 92(6), pp. 2598-2609.

Stenger, K. S. (2019) 'Identification of the response pathways of Escherichia coli and Enterococcus faecalis to glyphosate and it's major breakdown product Aminomethyl phosphonic acid (AMPA)'.

Stump, M. J., Jones, J. J., Fleming, R. C., Lay, J. O. and Wilkins, C. L. (2003) 'Use of double-depleted 13C and 15N culture media for analysis of whole cell bacteria by MALDI time-of-flight and Fourier transform mass spectrometry', *Journal of the American Society for Mass Spectrometry,* 14(11), pp. 1306-1314.

Sukhnanand, S., Dogan, B., Ayodele, M. O., Zadoks, R. N., Craver, M. P., Dumas, N. B., Schukken, Y. H., Boor, K. J. and Wiedmann, M. (2005) 'Molecular subtyping and characterization of bovine and human Streptococcus agalactiae isolates', *J Clin Microbiol,* 43(3), pp. 1177-86.

Sun, Z., Samarasinghe, S. and Jago, J. (2010) 'Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks', *Journal of Dairy Research,* 77(2), pp. 168-175.

Suojala, L., Kaartinen, L. and Pyörälä, S. (2013) 'Treatment for bovine Escherichia coli mastitis – an evidence-based approach', *Journal of Veterinary Pharmacology and Therapeutics,* 36(6), pp. 521-531.

Suojala, L., Pohjanvirta, T., Simojoki, H., Myllyniemi, A. L., Pitkala, A., Pelkonen, S. and Pyorala, S. (2011) 'Phylogeny, virulence factors and antimicrobial susceptibility of Escherichia coli isolated in clinical bovine mastitis', *Vet Microbiol,* 147(3-4), pp. 383-8.

Supré, K., Lommelen, K. and De Meulemeester, L. (2014) 'Antimicrobial susceptibility and distribution of inhibition zone diameters of bovine mastitis pathogens in Flanders, Belgium', *Veterinary Microbiology,* 171(3), pp. 374-381.

Surveillance Intelligence Unit (2020) *Veterinary Investigation Diagnosis Analysis (VIDA) Annual Report 2019- Diagnoses by Year 2012-2019.* Available at: https://public.tableau.com/profile/siu.apha#!/vizhome/VIDAAnnualReport2019/VIDAAnnualReport2019 (Accessed: 2020, October 12.).

Sutra, L. and Poutrel, B. (1994) 'Virulence factors involved in the pathogenesis of bovine intramammary infections due to Staphylococcus aureus', *Journal of Medical Microbiology,* 40(2), pp. 79-89.

Sutrina, S. L., McGeary, T. and Bourne, C. A. (2007) 'The Phosphoenolpyruvate:Sugar Phosphotransferase System and Biofilms in Gram-Positive Bacteria', *Journal of Molecular Microbiology and Biotechnology,* 12(3-4), pp. 269-272.

Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. and Bacardit, J. (2013) 'Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology', *OMICS: A Journal of Integrative Biology,* 17(12), pp. 595-610.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H. and Bork, P. (2018) 'STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets', *Nucleic acids research,* 47(D1), pp. D607-D613.

Tan, K. E., Ellis, B. C., Lee, R., Stamper, P. D., Zhang, S. X. and Carroll, K. C. (2012) 'Prospective evaluation of a matrix-assisted laser desorption ionization–time of flight mass spectrometry system in a hospital clinical microbiology laboratory for identification of bacteria and yeasts: a bench-by-bench study for assessing the impact on time to identification and cost-effectiveness', *Journal of clinical microbiology,* 50(10), pp. 3301-3308.

Tang, W., Ranganathan, N., Shahrezaei, V. and Larrouy-Maumus, G. (2019) 'MALDI-TOF mass spectrometry on intact bacteria combined with a refined analysis framework allows accurate classification of MSSA and MRSA', *PloS one,* 14(6).

Taponen, S., Björkroth, J. and Pyörälä, S. (2008) 'Coagulase-negative staphylococci isolated from bovine extramammary sites and intramammary infections in a single dairy herd', *The Journal of Dairy Research,* 75(4), pp. 422.

Taponen, S., Simojoki, H., Haveri, M., Larsen, H. D. and Pyörälä, S. (2006) 'Clinical characteristics and persistence of bovine mastitis caused by different species of coagulase-negative staphylococci identified with API or AFLP', *Veterinary Microbiology,* 115(1), pp. 199-207.

Tarrant, E., P Riboldi, G., McIlvin, M. R., Stevenson, J., Barwinska-Sendra, A., Stewart, L. J., Saito, M. A. and Waldron, K. J. (2019) 'Copper stress in Staphylococcus aureus leads to adaptive changes in central carbon metabolism', *Metallomics : integrated biometal science,* 11(1), pp. 183-200.

Tassi, R., McNeilly, T. N., Fitzpatrick, J. L., Fontaine, M. C., Reddick, D., Ramage, C., Lutton, M., Schukken, Y. H. and Zadoks, R. N. (2013) 'Strain-specific pathogenicity of putative host-adapted and nonadapted strains of Streptococcus uberis in dairy cattle', *Journal of dairy science,* 96(8), pp. 5129-5145.

Tassi, R., McNeilly, T. N., Sipka, A. and Zadoks, R. N. (2015) 'Correlation of hypothetical virulence traits of two Streptococcus uberis strains with the clinical manifestation of bovine mastitis', *Veterinary Research,* 46(1), pp. 123.

te Witt, R., van Belkum, A., MacKay, W. G., Wallace, P. S. and van Leeuwen, W. B. (2010) 'External quality assessment of the molecular diagnostics and genotyping of meticillin-resistant Staphylococcus aureus', *European journal of clinical microbiology & infectious diseases,* 29(3), pp. 295-300.

Teixeira, L. M., Carvalho, M. G., Espinola, M. M., Steigerwalt, A. G., Douglas, M. P., Brenner, D. J. and Facklam, R. R. (2001) 'Enterococcus porcinus sp. nov. and Enterococcus ratti sp. nov., associated with enteric disorders in animals', *Int J Syst Evol Microbiol,* 51(Pt 5), pp. 1737-1743.

Tenaillon, O., Skurnik, D., Picard, B. and Denamur, E. (2010) 'The population genetics of commensal Escherichia coli', *Nature Reviews Microbiology,* 8(3), pp. 207-217.

Tendolkar, P. M., Baghdayan, A. S. and Shankar, N. (2003) 'Pathogenic enterococci: new developments in the 21st century', *Cellular and Molecular Life Sciences CMLS,* 60(12), pp. 2622-2636.

Tenhagen, B.-A., Hansen, I., Reinecke, A. and Heuwieser, W. (2009) 'Prevalence of pathogens in milk samples of dairy cows with clinical mastitis and in heifers at first parturition', *Journal of dairy research,* 76(2), pp. 179-187.

Tenhagen, B. A., Köster, G., Wallmann, J. and Heuwieser, W. (2006) 'Prevalence of mastitis pathogens and their resistance against antimicrobial agents in dairy cows in Brandenburg, Germany', *Journal of Dairy Science,* 89(7), pp. 2542-2551.

Thomas, P. D. (2017) 'The gene ontology and the meaning of biological function', *The Gene Ontology Handbook*: Humana Press, New York, NY, pp. 15-24.

Thongkam, J., Xu, G., Zhang, Y. and Huang, F. 'Breast cancer survivability via AdaBoost algorithms'. 2008, 55-64.

Tikofsky, L. L. and Zadoks, R. N. (2005) 'Cross-Infection Between Cats and Cows: Origin and Control of Streptococcus canis Mastitis in a Dairy Herd', *Journal of Dairy Science,* 88(8), pp. 2707-2713.

Tiseo, K., Huber, L., Gilbert, M., Robinson, T. P. and Van Boeckel, T. P. (2020) 'Global Trends in Antimicrobial Use in Food Animals from 2017 to 2030', *Antibiotics,* 9(12).

Tiwari, S., Jamal, S. B., Hassan, S. S., Carvalho, P. V. S. D., Almeida, S., Barh, D., Ghosh, P., Silva, A., Castro, T. L. P. and Azevedo, V. (2017) 'Two-Component Signal Transduction Systems of Pathogenic Bacteria As Targets for Antimicrobial Therapy: An Overview', *Frontiers in Microbiology,* 8, pp. 1878.

Todhunter, D. A., Smith, K. L. and Hogan, J. S. (1995) 'Environmental Streptococcal Intramammary Infections of the Bovine Mammary Gland1', *Journal of dairy science,* 78(11), pp. 2366-2374.

Toma, C., Higa, N., Iyoda, S., Rivas, M. and Iwanaga, M. (2006) 'The long polar fimbriae genes identified in Shiga toxin-producing Escherichia coli are present in other diarrheagenic E. coli and in the standard E. coli collection of reference (ECOR) strains', *Research in microbiology,* 157(2), pp. 153-161.

Tomazi, T., Gonçalves, J. L., Barreiro, J. R., de Campos Braga, P. A., e Silva, L. F. P., Eberlin, M. N. and Dos Santos, M. V. (2014) 'Identification of coagulase-negative staphylococci from bovine intramammary infection by matrix-assisted laser desorption ionization–time of flight mass spectrometry', *Journal of clinical microbiology,* 52(5), pp. 1658-1663.

Tomek, I. (1976) 'Two modifications of CNN'.

Tomita, T., Meehan, B., Wongkattiya, N., Malmo, J., Pullinger, G., Leigh, J. and Deighton, M. (2008) 'Identification of Streptococcus uberis Multilocus Sequence Types Highly Associated with Mastitis', *Applied and Environmental Microbiology,* 74(1), pp. 114.

Tozzoli, R., Grande, L., Michelacci, V., Ranieri, P., Maugliani, A., Caprioli, A. and Morabito, S. (2014) 'Shiga toxin-converting phages and the emergence of new pathogenic Escherichia coli: a world in motion', *Frontiers in cellular and infection microbiology,* 4, pp. 80.

Tyers, M. and Mann, M. (2003) 'From genomics to proteomics', *Nature,* 422(6928), pp. 193-197.

UK-VARSS (2019) *Veterinary Antibiotic Resistance and Sales Surveillance Report (UK-VARSS 2018)*, New Haw, Addlestone: Veterinary Medicines Directorate. Available at: www.gov.uk/government/collections/veterinary-antimicrobial- resistance-and-sales-surveillance.

UK-VARSS (2020) *Veterinary Antibiotic Resistance and Sales Surveillance Report (UK-VARSS 2019)*, New Haw, Addlestone: Veterinary Medicines Directorate. Available at: www.gov.uk/government/collections/veterinary-antimicrobial- resistance-and-sales-surveillance.

UniProt, C. (2018) 'UniProt: a worldwide hub of protein knowledge', *Nucleic acids research,* 47(D1), pp. D506-D515.

Unnerstad, H. E., Lindberg, A., Waller, K. P., Ekman, T., Artursson, K., Nilsson-Öst, M. and Bengtsson, B. (2009) 'Microbial aetiology of acute clinical mastitis and agent-specific risk factors', *Veterinary microbiology,* 137(1-2), pp. 90-97.

Urwin, R. and Maiden, M. C. J. (2003) 'Multi-locus sequence typing: a tool for global epidemiology', *Trends in microbiology,* 11(10), pp. 479-487.

Vabalas, A., Gowen, E., Poliakoff, E. and Casson, A. J. (2019) 'Machine learning algorithm validation with a limited sample size', *PloS one,* 14(11).

Vakkamäki, J., Taponen, S., Heikkilä, A.-M. and Pyörälä, S. (2017) 'Bacteriological etiology and treatment of mastitis in Finnish dairy herds', *Acta Veterinaria Scandinavica,* 59(1), pp. 33.

Valentine, N., Wunschel, S., Wunschel, D., Petersen, C. and Wahl, K. (2005) 'Effect of culture conditions on microorganism identification by matrix-assisted laser desorption ionization mass spectrometry', *Applied and environmental microbiology,* 71(1), pp. 58-64.

van Belkum, A., Chatellier, S., Girard, V., Pincus, D., Deol, P. and Dunne Jr, W. M. (2015) 'Progress in proteomics for clinical microbiology: MALDI-TOF MS for microbial species identification and more', *Expert review of proteomics,* 12(6), pp. 595-605.

van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., Fussing, V., Green, J., Feil, E., Gerner-Smidt, P., Brisse, S. and Struelens, M. (2007) 'Guidelines for the validation and application of typing methods for use in bacterial epidemiology', *Clin Microbiol Infect,* 13 Suppl 3, pp. 1-46.

Van Boeckel, T. P., Pires, J., Silvester, R., Zhao, C., Song, J., Criscuolo, N. G., Gilbert, M., Bonhoeffer, S. and Laxminarayan, R. (2019) 'Global trends in antimicrobial resistance in animals in low- and middle-income countries', *Science,* 365(6459), pp. eaaw1944.

van den Bogaard, A. E. and Stobberingh, E. E. (2000) 'Epidemiology of resistance to antibiotics: Links between animals and humans', *International Journal of Antimicrobial Agents,* 14(4), pp. 327-335.

van den Borne, B. H. P., Halasa, T., van Schaik, G., Hogeveen, H. and Nielen, M. (2010) 'Bioeconomic modeling of lactational antimicrobial treatment of new bovine subclinical intramammary infections caused by contagious pathogens', *Journal of Dairy Science,* 93(9), pp. 4034-4044.

Van Hoek, A. H. A. M., Mevius, D., Guerra, B., Mullany, P., Roberts, A. P. and Aarts, H. J. M. (2011) 'Acquired antibiotic resistance genes: an overview', *Frontiers in microbiology,* 2, pp. 203.

van Oosten, L. N. and Klein, C. D. (2020) 'Machine Learning in Mass Spectrometry: A MALDI-TOF MS Approach to Phenotypic Antibacterial Screening', *Journal of Medicinal Chemistry*.

Vanderhaeghen, W., Cerpentier, T., Adriaensen, C., Vicca, J., Hermans, K. and Butaye, P. (2010) 'Methicillin-resistant Staphylococcus aureus (MRSA) ST398 associated with clinical and subclinical mastitis in Belgian cows', *Veterinary Microbiology,* 144(1), pp. 166-171.

VanderPlas, J. (2016) *Python data science handbook: Essential tools for working with data.* " O'Reilly Media, Inc.".

Vapnik, V. N. (1995) 'The nature of statistical learning', *Theory*.

Varoquaux, G. (2018) 'Cross-validation failure: small sample sizes lead to large error bars', *Neuroimage,* 180, pp. 68-77.

Veh, K. A., Klein, R. C., Ster, C., Keefe, G., Lacasse, P., Scholl, D., Roy, J. P., Haine, D., Dufour, S., Talbot, B. G., Ribon, A. O. B. and Malouin, F. (2015) 'Genotypic and phenotypic characterization of Staphylococcus aureus causing persistent and nonpersistent subclinical bovine intramammary infections during lactation or the dry period', *Journal of Dairy Science,* 98(1), pp. 155-168.

Verbeke, J., Piepers, S., Supré, K. and De Vliegher, S. (2014) 'Pathogen-specific incidence rate of clinical mastitis in Flemish dairy herds, severity, and association with herd hygiene', *Journal of Dairy Science,* 97(11), pp. 6926-6934.

Verschuuren, T. D., Bruijning-Verhagen, P. C. J., Bosch, T., Schürch, A. C., Willems, R. J. L., Bonten, M. J. M. and Kluytmans, J. (2020) 'Extended-spectrum beta-lactamase (ESBL)-producing and non-ESBL-producing Escherichia coli isolates causing bacteremia in the Netherlands (2014-2016) differ in clonal

distribution, antimicrobial resistance gene and virulence gene content', *PloS one,* 15(1), pp. e0227604-e0227604.

Vimont, A., Delignette-Muller, M. L. and Vernozy-Rozand, C. (2007) 'Supplementation of enrichment broths by novobiocin for detecting Shiga toxin-producing Escherichia coli from food: a controversial use', *Letters in Applied Microbiology,* 44(3), pp. 326-331.

Vlasblom, J., Zuberi, K., Rodriguez, H., Arnold, R., Gagarinova, A., Deineko, V., Kumar, A., Leung, E., Rizzolo, K., Samanfar, B., Chang, L., Phanse, S., Golshani, A., Greenblatt, J. F., Houry, W. A., Emili, A., Morris, Q., Bader, G. and Babu, M. (2014) 'Novel function discovery with GeneMANIA: a new integrated resource for gene function prediction in Escherichia coli', *Bioinformatics,* 31(3), pp. 306-310.

von Wintersdorff, C. J. H., Penders, J., van Niekerk, J. M., Mills, N. D., Majumder, S., van Alphen, L. B., Savelkoul, P. H. M. and Wolffs, P. F. G. (2016) 'Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer', *Frontiers in Microbiology,* 7, pp. 173.

Vos, S. M., Lyubimov, A. Y., Hershey, D. M., Schoeffler, A. J., Sengupta, S., Nagaraja, V. and Berger, J. M. (2014) 'Direct control of type IIA topoisomerase activity by a chromosomally encoded regulatory protein', *Genes & development,* 28(13), pp. 1485-1497.

Vrioni, G., Tsiamis, C., Oikonomidis, G., Theodoridou, K., Kapsimali, V. and Tsakris, A. (2018) 'MALDI-TOF mass spectrometry technology for detecting biomarkers of antimicrobial resistance: current achievements and future perspectives', *Annals of translational medicine,* 6(12), pp. 240-240.

Vázquez-Laslop, N. and Mankin, A. S. (2018) 'How Macrolide Antibiotics Work', *Trends in Biochemical Sciences,* 43(9), pp. 668-684.

Wadsworth, C. B., Sater, M. R. A., Bhattacharyya, R. P. and Grad, Y. H. (2019) 'Impact of Species Diversity on the Design of RNA-Based Diagnostics for Antibiotic Resistance in Neisseria gonorrhoeae', *Antimicrobial Agents and Chemotherapy,* 63(8), pp. e00549-19.

Wagner, M., Naik, D. and Pothen, A. (2003) 'Protocols for disease classification from mass spectrometry data', *PROTEOMICS,* 3(9), pp. 1692-1698.

Walker, J. M. (2005) *The proteomics protocols handbook.* Springer.

Wallace-Gadsden, F., Johnson, J. R., Wain, J. and Okeke, I. N. (2007) 'Enteroaggregative Escherichia coli related to uropathogenic clonal group A', *Emerging infectious diseases,* 13(5), pp. 757-760.

Waller, D. G. and Sampson, T. (2017) *Medical pharmacology and therapeutics E-Book.* Elsevier Health Sciences.

Wang, E. and Samarasinghe, S. (2005) 'On-line detection of mastitis in dairy herds using artificial neural networks'.

Wang, H.-Y., Chen, C.-H., Lee, T.-Y., Horng, J.-T., Liu, T.-P., Tseng, Y.-J. and Lu, J.-J. (2018a) 'Rapid Detection of Heterogeneous Vancomycin-Intermediate Staphylococcus aureus Based on Matrix-Assisted Laser Desorption Ionization Time-of-Flight: Using a Machine Learning Approach and Unbiased Validation', *Frontiers in Microbiology,* 9, pp. 2393.

Wang, H.-Y., Lee, T.-Y., Tseng, Y.-J., Liu, T.-P., Huang, K.-Y., Chang, Y.-T., Chen, C.-H. and Lu, J.-J. (2018b) 'A new scheme for strain typing of methicillin-resistant Staphylococcus aureus on the basis

of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using machine learning approach', *PloS one,* 13(3), pp. e0194289-e0194289.

Wang, H.-Y., Lu, K.-P., Chung, C.-R., Tseng, Y.-J., Lee, T.-Y., Chang, T.-H., Wu, M.-H., Lin, T.-W., Liu, T.-P. and Lu, J.-J. (2020) 'Rapidly predicting vancomycin resistance of Enterococcus faecium through MALDI-TOF MS spectra obtained in real-world clinical microbiology laboratory', *bioRxiv*.

Wang, J.-Y., Sarker, A. H., Cooper, P. K. and Volkert, M. R. (2004) 'The single-strand DNA binding activity of human PC4 prevents mutagenesis and killing by oxidative DNA damage', *Molecular and cellular biology,* 24(13), pp. 6084-6093.

Wang, W.-L., Liu, J., Huo, Y.-B. and Ling, J.-Q. (2013) 'Bacteriocin immunity proteins play a role in quorum-sensing system regulated antimicrobial sensitivity of Streptococcus mutans UA159', *Archives of Oral Biology,* 58(4), pp. 384-390.

Wang, Z., Russon, L., Li, L., Roser, D. C. and Long, S. R. (1998) 'Investigation of spectral reproducibility in direct analysis of bacteria proteins by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry', *Rapid Communications in Mass Spectrometry,* 12(8), pp. 456-464.

Wang, Z., Zhou, H., Wang, H., Chen, H., Leung, K. K., Tsui, S. and Ip, M. (2014) 'Comparative genomics of methicillin-resistant Staphylococcus aureus ST239: distinct geographical variants in Beijing and Hong Kong', *BMC Genomics,* 15, pp. 529.

Ward, P. N., Field, T. R., Rapier, C. D. and Leigh, J. A. (2003) 'The activation of bovine plasminogen by PauA is not required for virulence of Streptococcus uberis', *Infection and immunity,* 71(12), pp. 7193-7196.

Ward, P. N., Holden, M. T. G., Leigh, J. A., Lennard, N., Bignell, A., Barron, A., Clark, L., Quail, M. A., Woodward, J., Barrell, B. G., Egan, S. A., Field, T. R., Maskell, D., Kehoe, M., Dowson, C. G., Chanter, N., Whatmore, A. M., Bentley, S. D. and Parkhill, J. (2009) 'Evidence for niche adaptation in the genome of the bovine pathogen Streptococcus uberis', *BMC Genomics,* 10(1), pp. 54.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T. (2018) 'SWISS-MODEL: homology modelling of protein structures and complexes', *Nucleic Acids Res,* 46(W1), pp. W296-w303.

Waters, C. M., Antiporta, M. H., Murray, B. E. and Dunny, G. M. (2003) 'Role of the Enterococcus faecalis GelE protease in determination of cellular chain length, supernatant pheromone levels, and degradation of fibrin and misfolded surface proteins', *Journal of bacteriology,* 185(12), pp. 3613-3623.

Watts, J. L. (1988) 'Etiological agents of bovine mastitis', *Veterinary microbiology,* 16(1), pp. 41-66.

Watts, J. L., Lowery, D. E., Teel, J. F. and Rossbach, S. (2000) 'Identification of Corynebacterium bovis and other Coryneforms Isolated from Bovine Mammary Glands', *Journal of Dairy Science,* 83(10), pp. 2373-2379.

Watts, J. L., Salmon, S. A., Yancey, R. J., Jr., Nickerson, S. C., Weaver, L. J., Holmberg, C., Pankey, J. W. and Fox, L. K. (1995) 'Antimicrobial Susceptibility of Microorganisms Isolated from the Mammary Glands of Dairy Heifers', *Journal of Dairy Science,* 78(7), pp. 1637-1648.

Watts, J. L., Shryock, T. R., Apley, M., Brown, S. D., Gray, J. T., Heine, H., Hunter, R. P., Mevius, D. J., Paich, M. G. and Silley, P. (2008) 'Performance standards for antimicrobial disk and dilution susceptibility tests for bacteria isolated from animals; approved standard—third edition'.

Wei, J., Liang, J., Shi, Q., Yuan, P., Meng, R., Tang, X., Yu, L. and Guo, N. (2014) 'Genome-wide transcription analyses in Mycobacterium tuberculosis treated with lupulone', *Brazilian Journal of Microbiology,* 45, pp. 333-342.

Weissman, S. J., Johnson, J. R., Tchesnokova, V., Billig, M., Dykhuizen, D., Riddell, K., Rogers, P., Qin, X., Butler-Wu, S., Cookson, B. T., Fang, F. C., Scholes, D., Chattopadhyay, S. and Sokurenko, E. (2012) 'High-resolution two-locus clonal typing of extraintestinal pathogenic Escherichia coli', *Appl Environ Microbiol,* 78(5), pp. 1353-60.

Welker, M. (2011) 'Proteomics for routine identification of microorganisms', *Proteomics,* 11(15), pp. 3143-3153.

Wenz, J. R., Barrington, G. M., Garry, F. B., Ellis, R. P. and Magnuson, R. J. (2006) 'Escherichia coli Isolates' Serotypes, Genotypes, and Virulence Genes and Clinical Coliform Mastitis Severity', *Journal of Dairy Science,* 89(9), pp. 3408-3412.

Wenz, J. R., Barrington, G. M., Garry, F. B., McSweeney, K. D., Dinsmore, R. P., Goodell, G. and Callan, R. J. (2001) 'Bacteremia associated with naturally occurring acute coliform mastitis in dairy cows', *Journal of the American Veterinary Medical Association,* 219(7), pp. 976-981.

Werner, G., Fleige, C., Feßler, A. T., Timke, M., Kostrzewa, M., Zischka, M., Peters, T., Kaspar, H. and Schwarz, S. (2012) 'Improved identification including MALDI-TOF mass spectrometry analysis of group D streptococci from bovine mastitis and subsequent molecular characterization of corresponding Enterococcus faecalis and Enterococcus faecium isolates', *Veterinary Microbiology,* 160(1), pp. 162-169.

Wickham, H. (2011) 'ggplot2', *Wiley Interdisciplinary Reviews: Computational Statistics,* 3(2), pp. 180-185.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L. and Hester, J. (2019) 'Welcome to the Tidyverse', *Journal of open source software,* 4(43), pp. 1686.

Wilcoxon, F. (1992) 'Individual comparisons by ranking methods', *Breakthroughs in statistics*: Springer, pp. 196-202.

Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A. and Tingey, S. V. (1990) 'DNA polymorphisms amplified by arbitrary primers are useful as genetic markers', *Nucleic acids research,* 18(22), pp. 6531-6535.

Wilson, D. J., Grohn, Y. T., Bennett, G. J., González, R. N., Schukken, Y. H. and Spatz, J. (2007a) 'Comparison of J5 vaccinates and controls for incidence, etiologic agent, clinical severity, and survival in the herd following naturally occurring cases of clinical mastitis', *Journal of dairy science,* 90(9), pp. 4282-4288.

Wilson, D. J., Grohn, Y. T., Bennett, G. J., González, R. N., Schukken, Y. H. and Spatz, J. (2008) 'Milk production change following clinical mastitis and reproductive performance compared among J5 vaccinated and control dairy cattle', *J Dairy Sci,* 91(10), pp. 3869-79.

Wilson, D. J., Mallard, B. A., Burton, J. L., Schukken, Y. H. and Grohn, Y. T. (2009) 'Association of Escherichia coli J5-Specific Serum Antibody Responses with Clinical Mastitis Outcome for J5 Vaccinate and Control Dairy Cattle', *Clinical and Vaccine Immunology,* 16(2), pp. 209.

Wilson, D. J., Mallard, B. A., Burton, J. L., Schukken, Y. H. and Gröhn, Y. T. (2007b) 'Milk and serum J5-specific antibody responses, milk production change, and clinical effects following intramammary Escherichia coli challenge for J5 vaccinate and control cows', *Clin Vaccine Immunol,* 14(6), pp. 693-9.

Wilson, D. L. (1972) 'Asymptotic properties of nearest neighbor rules using edited data', *IEEE Transactions on Systems, Man, and Cybernetics,* (3), pp. 408-421.

Wisconsin Veterinary Diagnostic Laboratory (2020) *Interpretation of Bulk Tank Milk Results*. Available at: https://www.wvdl.wisc.edu/wp-content/uploads/2017/09/Bulk-Tank-guidelines.pdf (Accessed: 28 August 2020).

Wittwer, M., Keller, J., Wassenaar, T. M., Stephan, R., Howald, D., Regula, G. and Bissig-Choisat, B. (2005) 'Genetic diversity and antibiotic resistance patterns in a Campylobacter population isolated from poultry farms in Switzerland', *Applied and Environmental Microbiology,* 71(6), pp. 2840-2847.

Wolf, C., Kusch, H., Monecke, S., Albrecht, D., Holtfreter, S., von Eiff, C., Petzl, W., Rainard, P., Bröker, B. M. and Engelmann, S. (2011) 'Genomic and proteomic characterization of Staphylococcus aureus mastitis isolates of bovine origin', *PROTEOMICS,* 11(12), pp. 2491-2502.

Woodford, N. (2001) 'Epidemiology of the genetic elements responsible for acquired glycopeptide resistance in enterococci', *Microb Drug Resist,* 7(3), pp. 229-36.

Worby, C. A., Mattoo, S., Kruger, R. P., Corbeil, L. B., Koller, A., Mendez, J. C., Zekarias, B., Lazar, C. and Dixon, J. E. (2009) 'The Fic Domain: Regulation of Cell Signaling by Adenylylation', *Molecular Cell,* 34(1), pp. 93-103.

Wu, S., Skolnick, J. and Zhang, Y. (2007) 'Ab initio modeling of small proteins by iterative TASSER simulations', *BMC biology,* 5(1), pp. 17.

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S. and Erni, F. (1996) 'Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data', *Analytica Chimica Acta,* 329(3), pp. 257-265.

Wu, X., Hou, S., Zhang, Q., Ma, Y., Zhang, Y., Kan, W. and Zhao, X. (2016) 'Prevalence of virulence and resistance to antibiotics in pathogenic enterococci isolated from mastitic cows', *Journal of Veterinary Medical Science,* 78(11), pp. 1663-1668.

Wunschel, D. S., Hill, E. A., McLean, J. S., Jarman, K., Gorby, Y. A., Valentine, N. and Wahl, K. (2005a) 'Effects of varied pH, growth rate and temperature using controlled fermentation and batch culture on matrix assisted laser desorption/ionization whole cell protein fingerprints', *Journal of microbiological methods,* 62(3), pp. 259-271.

Wunschel, S. C., Jarman, K. H., Petersen, C. E., Valentine, N. B., Wahl, K. L., Schauki, D., Jackman, J., Nelson, C. P. and White, E. (2005b) 'Bacterial analysis by MALDI-TOF mass spectrometry: An inter-laboratory comparison', *Journal of the American Society for Mass Spectrometry,* 16(4), pp. 456-462.

Xia, W., Ma, C., Liu, J., Liu, S., Chen, F., Yang, Z. and Duan, J. (2019) 'High-Resolution Remote Sensing Imagery Classification of Imbalanced Data Using Multistage Sampling Method and Deep Neural Networks', *Remote Sensing,* 11, pp. 2523.

Xiang, Z. (2006) 'Advances in homology protein structure modeling', *Current Protein and Peptide Science,* 7(3), pp. 217-227.

Xiao, D., Zhao, F., Zhang, H., Meng, F. and Zhang, J. (2014) 'Novel strategy for typing Mycoplasma pneumoniae isolates by use of matrix-assisted laser desorption ionization–time of flight mass spectrometry coupled with ClinProTools', *Journal of clinical microbiology,* 52(8), pp. 3038-3043.

Xu, H. H., Trawick, J. D., Haselbeck, R. J., Forsyth, R. A., Yamamoto, R. T., Archer, R., Patterson, J., Allen, M., Froelich, J. M. and Taylor, I. (2010) 'Staphylococcus aureus TargetArray: comprehensive differential essential gene expression as a mechanistic tool to profile antibacterials', *Antimicrobial agents and chemotherapy,* 54(9), pp. 3659-3670.

Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R. and Ishihama, A. (2005) 'Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli', *Journal of Biological Chemistry,* 280(2), pp. 1448-1456.

Yamamoto, K. and Ishihama, A. (2005) 'Transcriptional response of Escherichia coli to external copper', *Molecular Microbiology,* 56(1), pp. 215-227.

Yamane, T., Enokida, H., Hayami, H., Kawahara, M. and Nakagawa, M. (2012) 'Genome-wide transcriptome analysis of fluoroquinolone resistance in clinical isolates of Escherichia coli', *International Journal of Urology,* 19(4), pp. 360-368.

Yan, J., Koc, M. and Lee, J. (2004) 'A prognostic algorithm for machine performance assessment and its application', *Production Planning & Control,* 15(8), pp. 796-801.

Yang, F., Zhang, S., Shang, X., Li, H., Zhang, H., Cui, D., Wang, X., Wang, L., Yan, Z. and Sun, Y. (2020) 'Short communication: Detection and molecular characterization of methicillin-resistant Staphylococcus aureus isolated from subclinical bovine mastitis cases in China', *Journal of Dairy Science,* 103(1), pp. 840-845.

Yang, F., Zhang, S., Shang, X., Wang, X., Yan, Z., Li, H. and Li, J. (2019a) 'Short communication: Antimicrobial resistance and virulence genes of Enterococcus faecalis isolated from subclinical bovine mastitis cases in China', *Journal of Dairy Science,* 102(1), pp. 140-144.

Yang, J. and Zhang, Y. (2015) 'I-TASSER server: new development for protein structure and function predictions', *Nucleic acids research,* 43(W1), pp. W174-W181.

Yang, S.-K., Yusoff, K., Ajat, M., Thomas, W., Abushelaibi, A., Akseer, R., Lim, S.-H. E. and Lai, K.-S. (2019b) 'Disruption of KPC-producing Klebsiella pneumoniae membrane via induction of oxidative stress by cinnamon bark (Cinnamomum verum J. Presl) essential oil', *PloS one,* 14(4).

Yasui, Y., Pepe, M., Thompson, M. L., Adam, B. L., Wright, G. L., Jr., Qu, Y., Potter, J. D., Winget, M., Thornquist, M. and Feng, Z. (2003) 'A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection', *Biostatistics,* 4(3), pp. 449-63.

Yin, X., Feng, Y., Lu, Y., Chambers, J. R., Gong, J. and Gyles, C. L. (2012) 'Adherence and associated virulence gene expression in acid-treated Escherichia coli O157 : H7 in vitro and in ligated pig intestine', *Microbiology,* 158(Pt 4), pp. 1084-93.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J. and Brinkman, F. S. (2010) 'PSORTb 3.0: improved protein subcellular localization prediction with

refined localization subcategories and predictive capabilities for all prokaryotes', *Bioinformatics,* 26(13), pp. 1608-15.

Yu, Z. N., Wang, J., Ho, H., Wang, Y. T., Huang, S. N. and Han, R. W. (2020) 'Prevalence and antimicrobial-resistance phenotypes and genotypes of Escherichia coli isolated from raw milk samples from mastitis cases in four regions of China', *Journal of Global Antimicrobial Resistance,* 22, pp. 94-101.

Zadoks, R. and Fitzpatrick, J. (2009) 'Changing trends in mastitis', *Ir Vet J,* 62 Suppl 4, pp. S59-70.

Zadoks, R. N., Allore, H. G., Barkema, H. W., Sampimon, O. C., Wellenberg, G. J., Gröhn, Y. T. and Schukken, Y. H. (2001) 'Cow-and quarter-level risk factors for Streptococcus uberis and Staphylococcus aureus mastitis', *Journal of Dairy Science,* 84(12), pp. 2649-2663.

Zadoks, R. N., Allore, H. G., Hagenaars, T. J., Barkema, H. W. and Schukken, Y. H. (2002) 'A mathematical model of Staphylococcus aureus control in dairy herds', *Epidemiology & Infection,* 129(2), pp. 397-416.

Zadoks, R. N., Gillespie, B. E., Barkema, H. W., Sampimon, O. C., Oliver, S. P. and Schukken, Y. H. (2003) 'Clinical, epidemiological and molecular characteristics of Streptococcus uberis infections in dairy herds', *Epidemiology and Infection,* 130(2), pp. 335-349.

Zadoks, R. N., Middleton, J. R., McDougall, S., Katholm, J. and Schukken, Y. H. (2011) 'Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans', *J Mammary Gland Biol Neoplasia,* 16(4), pp. 357-72.

Zadoks, R. N. and Schukken, Y. H. (2006) 'Use of molecular epidemiology in veterinary practice', *The Veterinary clinics of North America. Food animal practice,* 22(1), pp. 229-261.

Zadoks, R. N., Schukken, Y. H. and Wiedmann, M. (2005) 'Multilocus sequence typing of Streptococcus uberis provides sensitive and epidemiologically relevant subtype information and reveals positive selection in the virulence gene pauA', *Journal of clinical microbiology,* 43(5), pp. 2407-2417.

Zadoks, R. N., Tikofsky, L. L. and Boor, K. J. (2005) 'Ribotyping of Streptococcus uberis from a dairy's environment, bovine feces and milk', *Veterinary microbiology,* 109(3-4), pp. 257-265.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M. and Larsen, M. V. (2012) 'Identification of acquired antimicrobial resistance genes', *J Antimicrob Chemother,* 67(11), pp. 2640-4.

Zervos, M. J. and Schaberg, D. R. (1985) 'Reversal of the in vitro susceptibility of enterococci to trimethoprim-sulfamethoxazole by folinic acid', *Antimicrob Agents Chemother,* 28(3), pp. 446-8.

Zhang, C.-X., Wang, G.-W. and Zhang, J.-S. (2011) 'An empirical bias–variance analysis of DECORATE ensemble method at different training sample sizes', *Journal of Applied Statistics - J APPL STAT,* 39, pp. 1-22.

Zhang, H. (2005) 'Exploring conditions for the optimality of naive Bayes', *International Journal of Pattern Recognition and Artificial Intelligence,* 19(02), pp. 183-198.

Zhang, Y. (2010) *New advances in machine learning.* BoD–Books on Demand.

Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., Achtman, M., Brown, D., Chattaway, M., Dallman, T., Delahay, R. and Kornschober, C. (2020) 'The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity', *Genome Research,* 30(1), pp. 138-152.

Zhu, T., Lou, Q., Wu, Y., Hu, J., Yu, F. and Qu, D. (2010) 'Impact of the Staphylococcus epidermidis LytSR two-component regulatory system on murein hydrolase activity, pyruvate utilization and global transcriptional profile', *BMC microbiology,* 10, pp. 287-287.

Zhu, Y.-Z., Li, Q.-T., Wang, L., Zhong, Y., Ding, G.-H., Li, G., Jia, P.-L., Shi, T.-L. and Guo, X.-K. (2008) 'Gene expression profiling-based in silico approach to identify potential vaccine candidates and drug targets against B. pertussis and B. parapertussis', *OMICS A Journal of Integrative Biology,* 12(3), pp. 161-169.

Zou, J., Hong, G., Guo, X., Zhang, L., Yao, C., Wang, J. and Guo, Z. (2011) 'Reproducible Cancer Biomarker Discovery in SELDI-TOF MS Using Different Pre-Processing Algorithms', *PLOS ONE,* 6(10), pp. e26294.

Zuber, P. (2001) 'A peptide profile of the Bacillus subtilis genome', *Peptides,* 22(10), pp. 1555-1577.

Zude, I. (2014) 'Characterization of virulence-associated traits of Escherichia coli bovine mastitis isolates'.

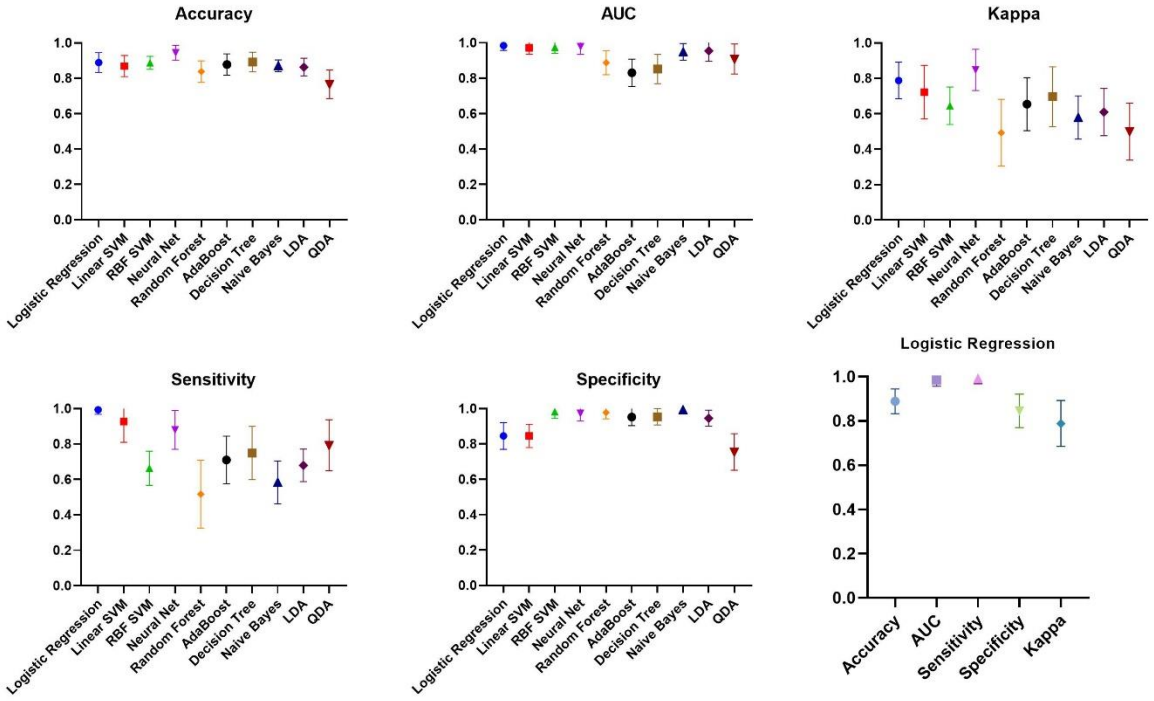
Zuniga, M., Comas, I., Linaje, R., Monedero, V., Yebra, M. J., Esteban, C. D., Deutscher, J., Perez-Martinez, G. and Gonzalez-Candelas, F. (2005) 'Horizontal gene transfer in the molecular evolution of mannose PTS transporters', *Mol Biol Evol,* 22(8), pp. 1673-85.

Østerås, O., Edge, V. L. and Martin, S. W. (1999) 'Determinants of success or failure in the elimination of major mastitis pathogens in selective dry cow therapy', *Journal of Dairy Science,* 82(6), pp. 1221-1231.

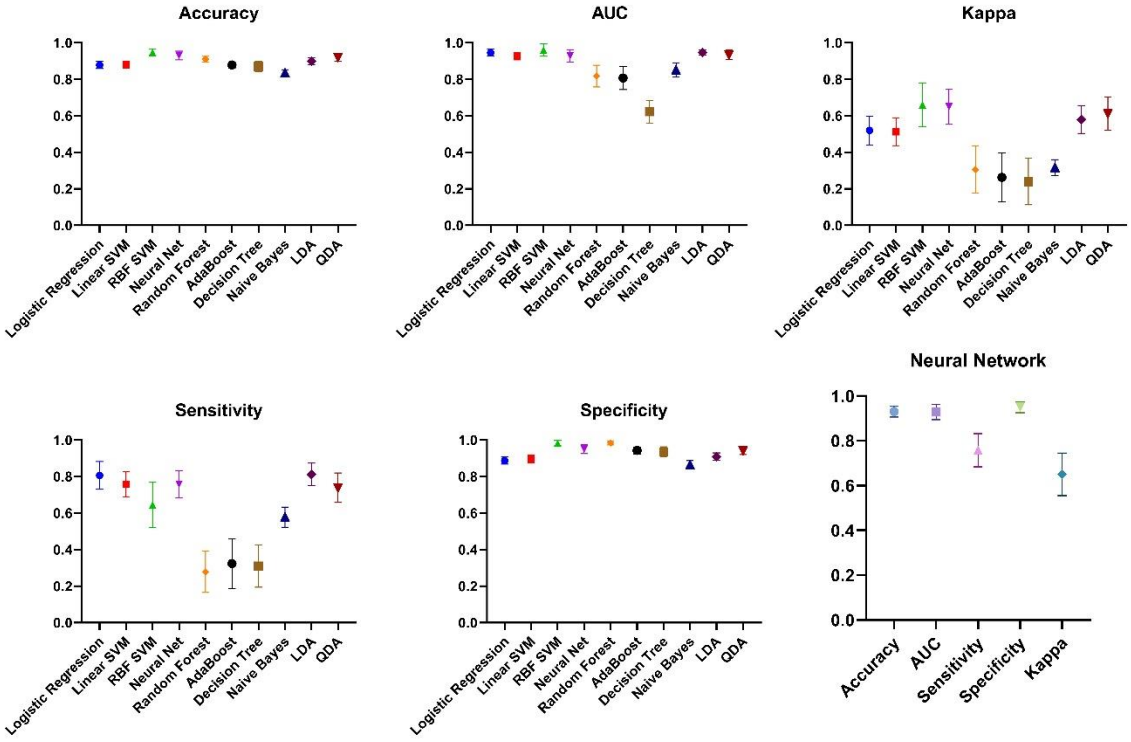
Østerås, O. and Sølverød, L. (2009) 'Norwegian mastitis control programme', *Irish Veterinary Journal,* 62(4), pp. 1-8.

Østerås, O., Sølverød, L. and Reksen, O. (2006) 'Milk culture results in a large Norwegian survey—effects of season, parity, days in milk, resistance, and clustering', *Journal of Dairy Science,* 89(3), pp. 1010-1023.

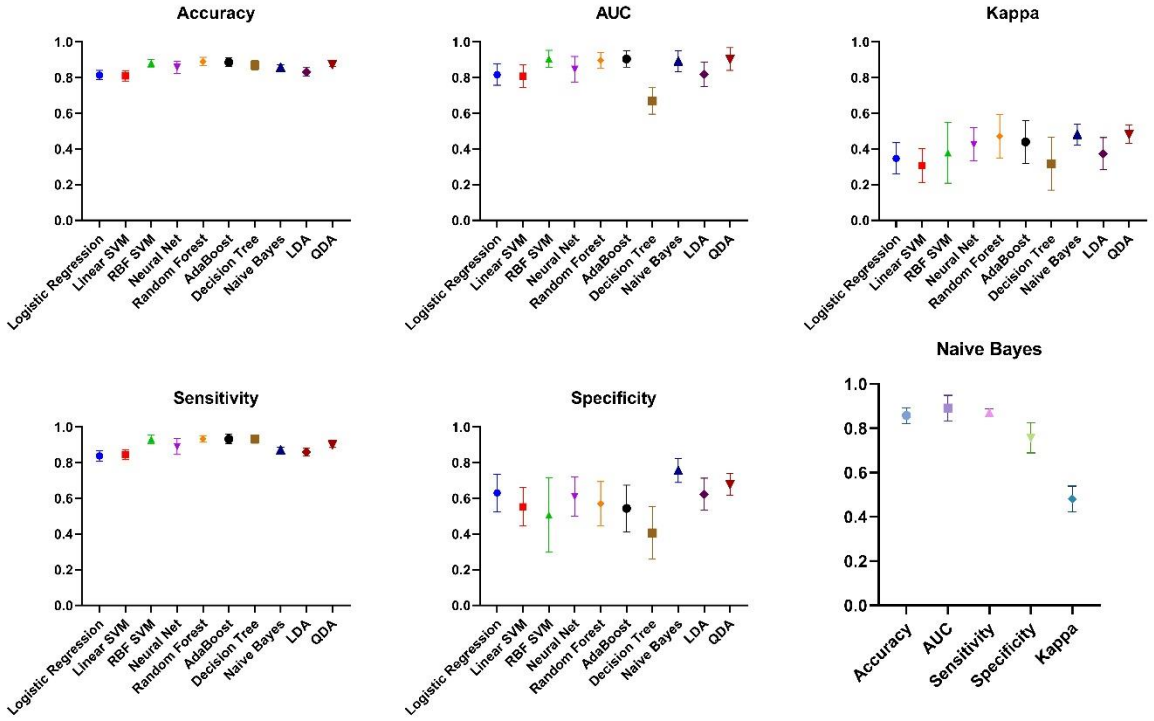
SUPPLEMENTARY FILES



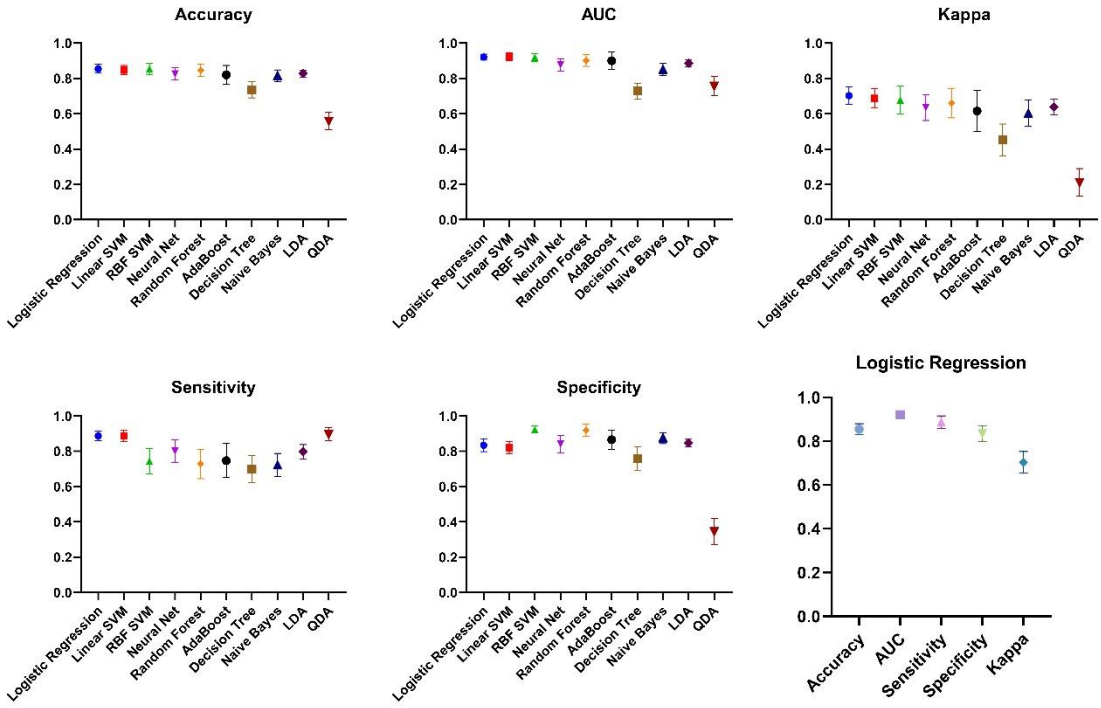
Supplementary Figure 1. Prediction performance results of algorithms to discriminate benzylpenicillin-resistant and sensitive E. faecalis isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify benzylpenicillin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Logistic regression was found to give the best performance.*



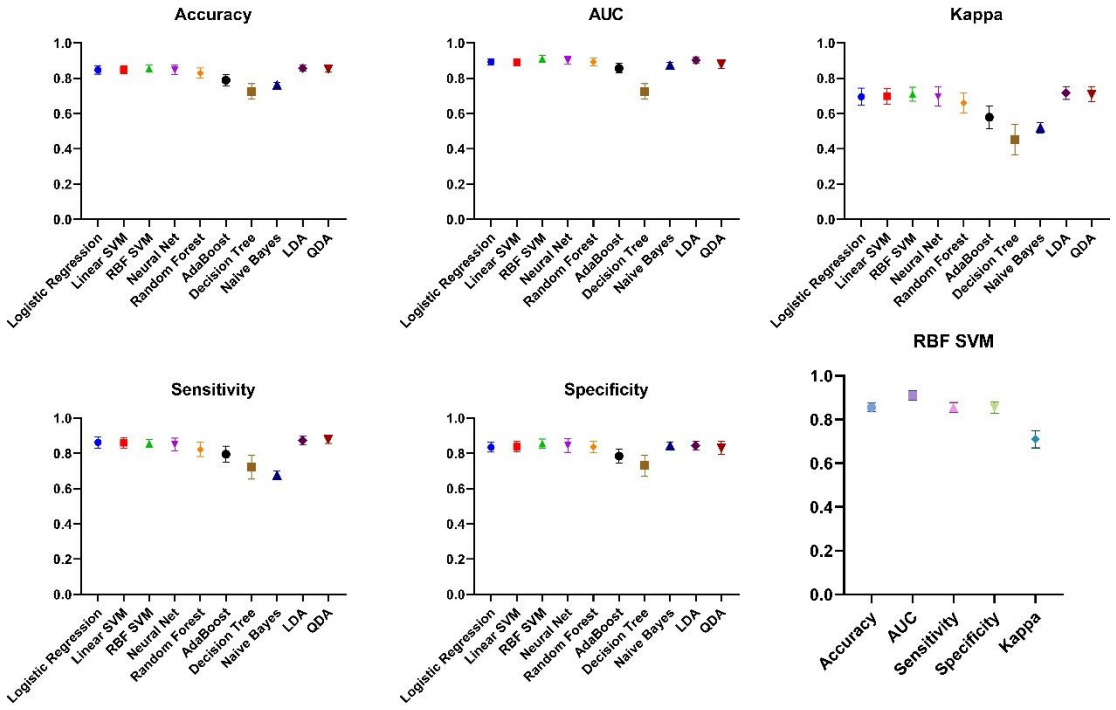
Supplementary Figure 2. Prediction performance of the several algorithms to discriminate chloramphenicol-resistant and sensitive E. faecalis isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify chloramphenicol resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. MLP neural network was found to give the best performance.*



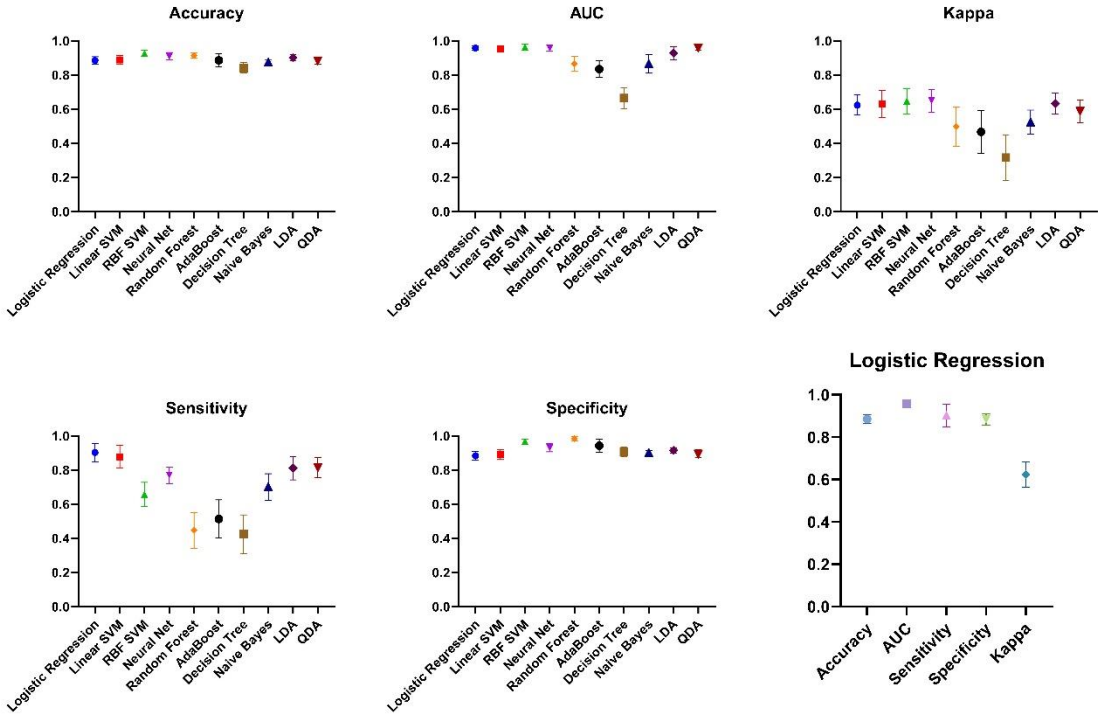
Supplementary Figure 3. Prediction performance of the several algorithms to discriminate clindamycin-resistant and sensitive E. faecalis isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify clindamycin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Naïve Bayes was found to give the best performance.*



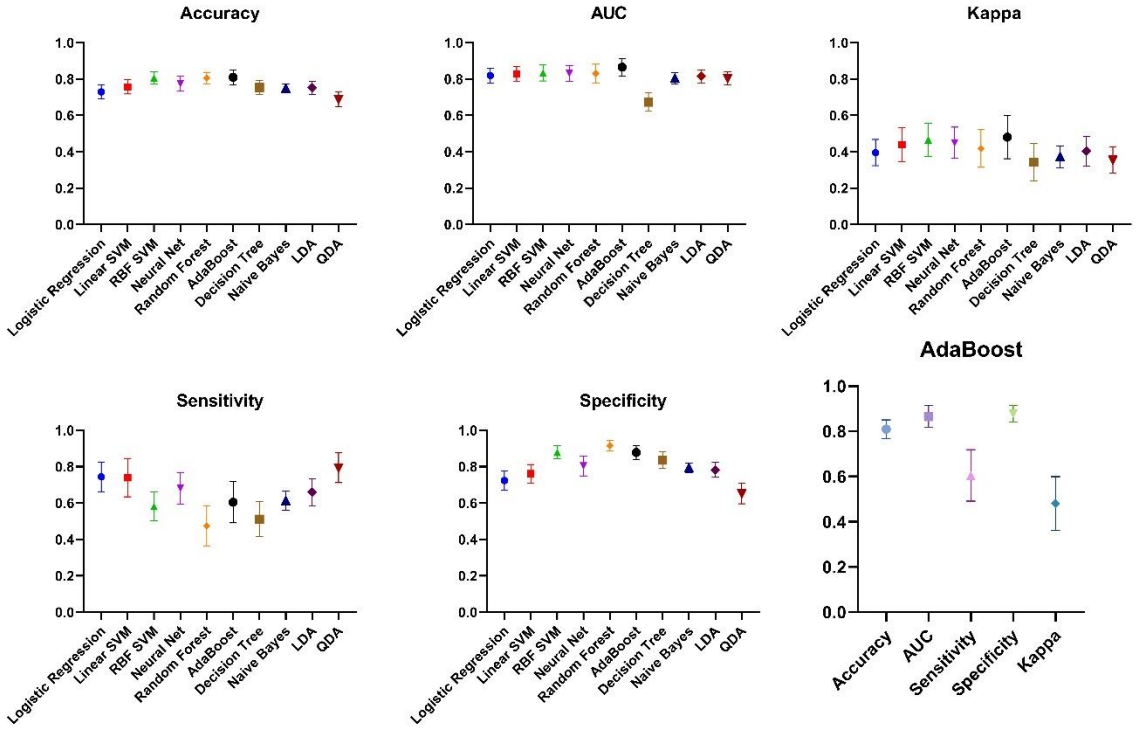
Supplementary Figure 4. Prediction performance of the several algorithms to discriminate erythromycin-resistant and sensitive E. faecalis isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify erythromycin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Logistic regression was found to give the best performance.*



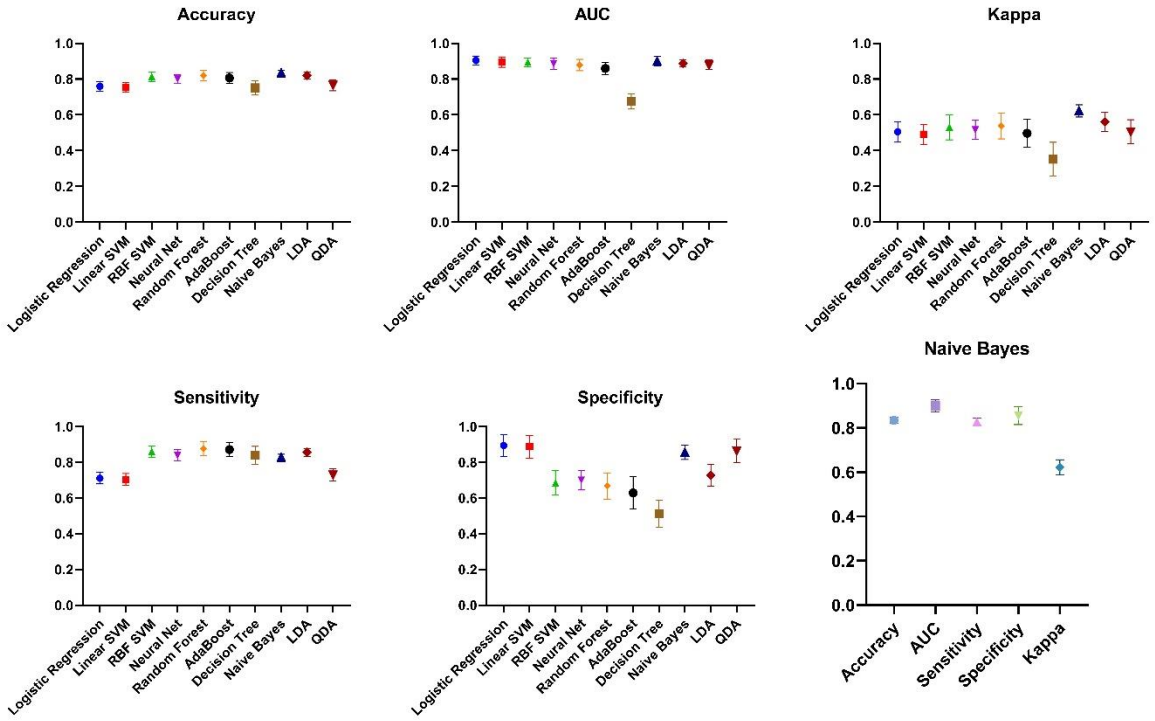
Supplementary Figure 5. Prediction performance of the several algorithms to discriminate tetracycline-resistant and sensitive E. faecalis isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify tetracycline resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. RBF SVM was found to give the best performance.*



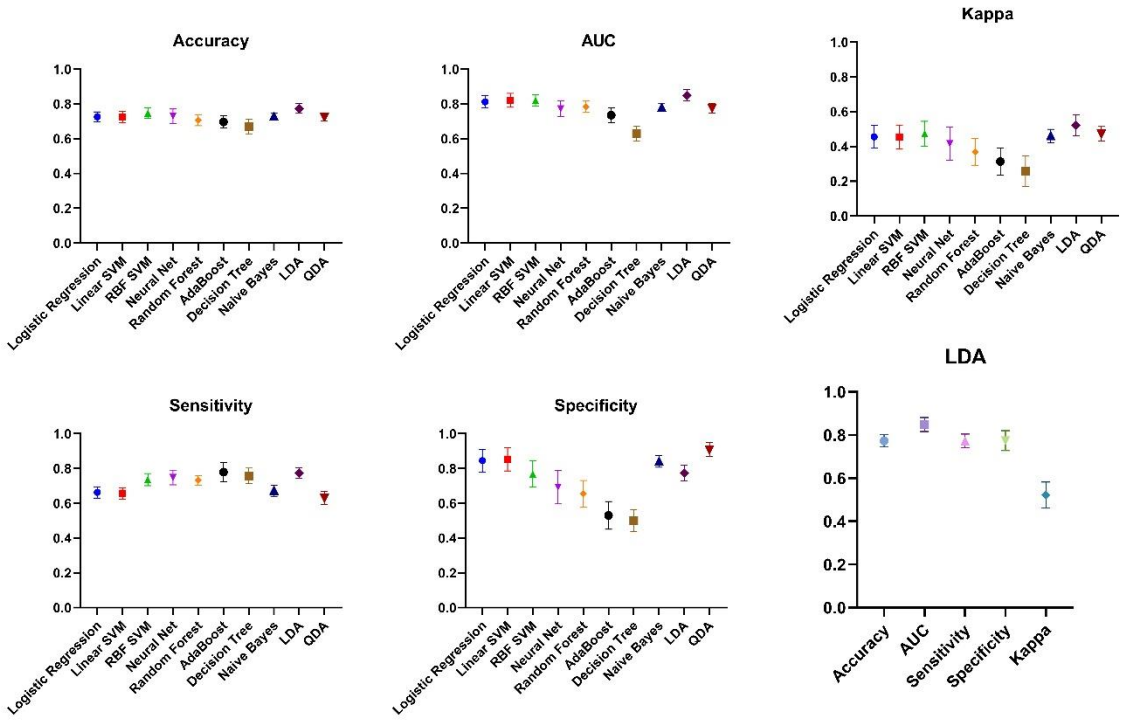
Supplementary Figure 6. Prediction performance of the several algorithms to discriminate TMP/SMX-resistant and sensitive E. faecalis isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify TMP/SMX resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Logistic regression was found to give the best performance.*



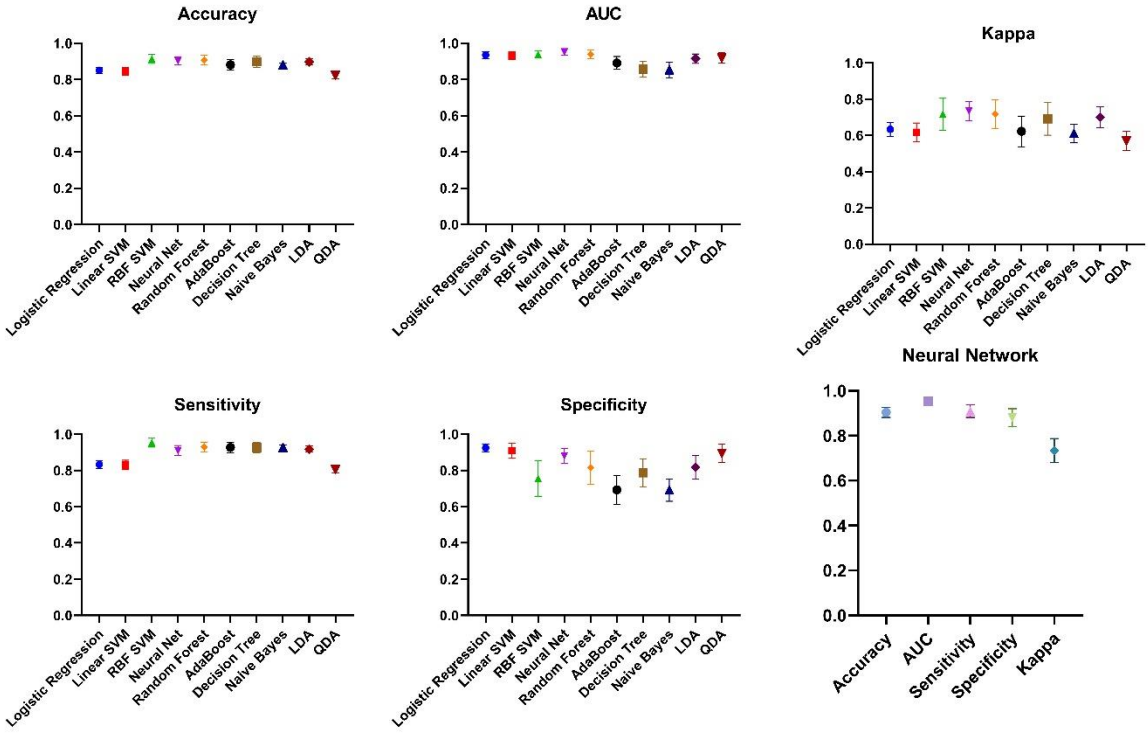
Supplementary Figure 7. Prediction performance of the several algorithms to discriminate benzylpenicillin-resistant and sensitive E. faecium isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify benzylpenicillin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. AdaBoost was found to give the best performance.*



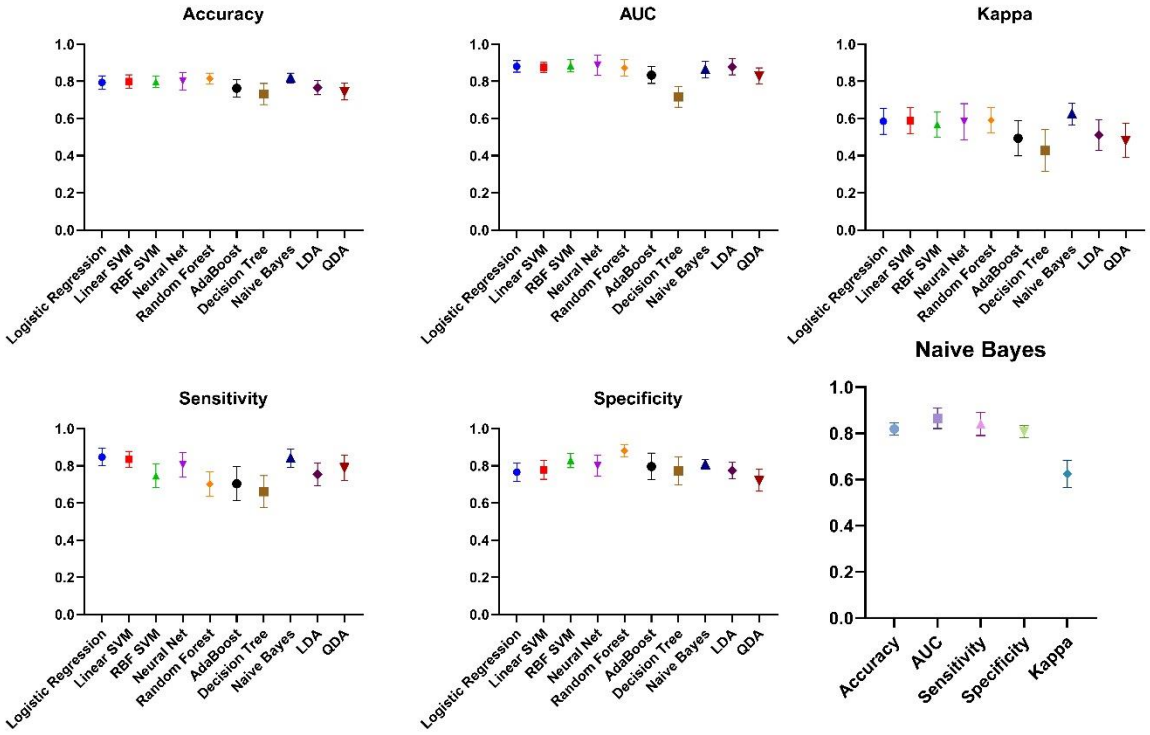
Supplementary Figure 8. Prediction performance of the several algorithms to discriminate cefovecin-resistant and sensitive E. faecium isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify cefovecin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Naïve Bayes was found to give the best performance.*



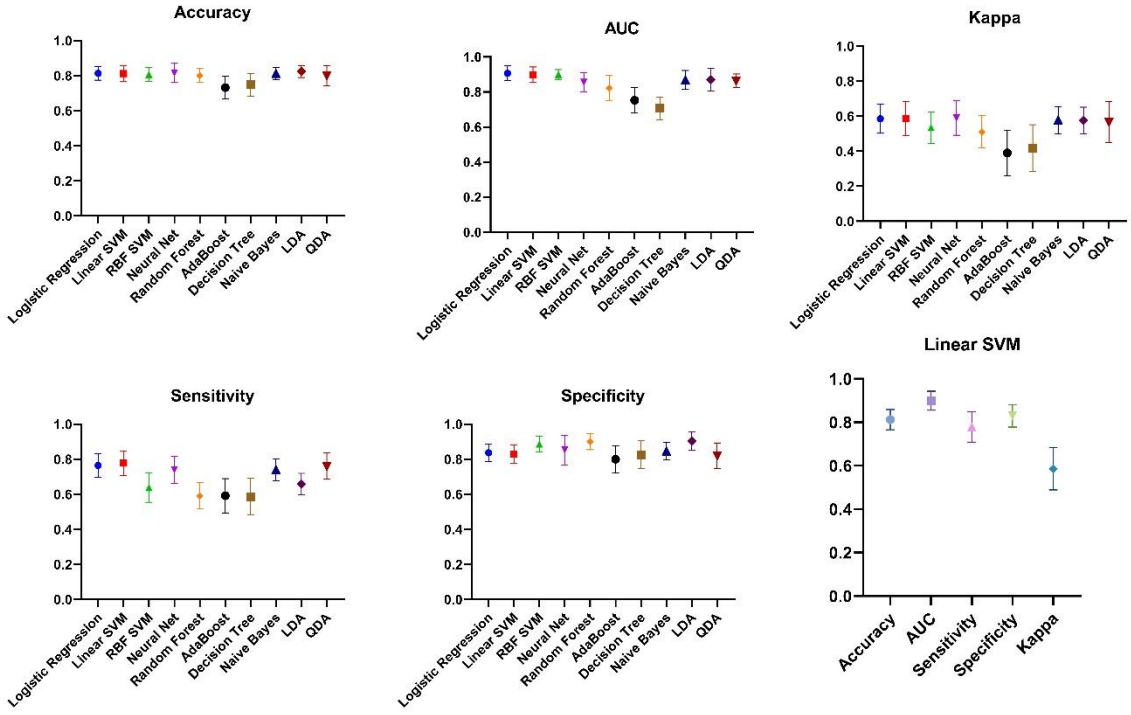
Supplementary Figure 9. Prediction performance of the several algorithms to discriminate clindamycin-resistant and sensitive E. faecium isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify clindamycin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. LDA was found to give the best performance.*



Supplementary Figure 10. Prediction performance of the several algorithms to discriminate enrofloxacin-resistant and sensitive E. faecium isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify enrofloxacin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. MLP neural network was found to give the best performance.*



Supplementary Figure 11. Prediction performance of the several algorithms to discriminate erythromycin-resistant and sensitive E. faecium isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify erythromycin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Naïve Bayes was found to give the best performance.*



Supplementary Figure 12. Prediction performance of the several algorithms to discriminate nitrofurantoin-resistant and sensitive E. faecium isolates. *Ten different algorithms (logistic regression, linear SVM, RBF SVM, MLP neural network, decision tree, random forest, AdaBoost, naïve Bayes, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA)) that were used to classify nitrofurantoin resistance profiles are shown on the X-axis. The prediction performance of these algorithms was measured based on five metrics (from left to right): accuracy, AUC, kappa, sensitivity and specificity. The scores for each metric (Y-axis) are between 0 and 1. Linear SVM was found to give the best performance.*