Streptococcus equi subsp. *zooepidemicus* in New Zealand Horses

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Biological Sciences

at

The University of Waikato

by

Olivia Patty

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Abstract

Streptococcus equi subsp. *zooepidemicus* (*S. zooepidemicus*) is an opportunistic commensal bacterium of the horse mucosa causing disease in susceptible populations. This research sets out to enhance the understanding of *S. zooepidemicus* infections within the New Zealand horse population through strain-associated genomic and phenotypic characteristics. Retrospectively, *S. zooepidemicus* is the second most frequently isolated bacterium in New Zealand horses, being mainly associated with respiratory and urogenital tract infections. However, it also has the ability to colonise and or infect a diverse range of other anatomical regions of New Zealand horses. *S. zooepidemicus* isolates were shown to be genetically diverse via strain typing using either the single polymorphic gene, szp or standard multilocus sequence typing (MLST). Both typing schemes were shown to lack the discriminatory power needed to completely separate the isolates which was resolved by employing core genome MLST. Importantly, none of these typing methods could link a particular strain to disease status or anatomical region. However, when disease occurred the number of colonising strains tended to reduce to just the infective strain, indicating bacterial competition and strain fitness advantages. Factors available to *S. zooepidemicus* to colonise and persist in the horse include biofilm formation and cellular invasion with the majority of strains tested *in vitro*, able to form biofilms and all tested strains capable of invasion.

Bioinformatic analysis showed that the pangenome of *S. zooepidemicus* is open, which demonstrates the ability of species to accessorise its genome enabling the species to diversify and exploit different hosts and environment. An average of 45 new genes are discovered with the addition of each new genome. Unfortunately, no single feature of the accessory genome analysed was able to differentiate the New Zealand horse *S. zooepidemicus* strains as commensals or pathogens or differentiate strains from respiratory disease or uterine infection. As a result, this study defines *S. zooepidemicus* as an opportunistic pathogen that can reside in the healthy equine population and infect susceptible hosts and are not anatomically, geographically or host bound.

Acknowledgements

Firstly, thank you to the New Zealand Equine Research Foundation for supporting and funding this project and the opportunities you made available to me. Thank you for all the other research opportunities you facilitate for the health of New Zealand equines. I am further grateful to all the New Zealand equine veterinarians and to the two Waikato pathology laboratories, New Zealand Veterinary Pathology (NZVP) and SVS Laboratories, for providing the specimens making this project possible. Additional thanks to Dr Isobel Gibson (NZVP) for guidance with and provision of the equine sample records used in the retrospective study. Thank you to Bill Fleury (Department of Conservation) and the staff at the Tūrangi Hautu factory of Tongariro Prison (Department of Corrections) for arranging collection of the Kaimanawa horse samples and to Dr Erica Gee (Massey University) and Dr Murray Brightwell (Cambridge Equine Hospital) for their guidance in taking those samples. In addition, I would like to thank Dr Lee Morris (EquiBreed) and Dr Alexandra Leander (Waikato Equine Veterinary Centre) for their support in arranging samples and answering all my questions regarding horses and NZ veterinarians. Also, thank you to Alex for her friendship and support while doing her research at the University of Waikato and beyond.

A huge thanks goes to Dr Andrew Waller (formerly from Animal Health Trust, UK and now from Intervacc AB, Sweden). Thank you for your support in fielding all my questions, providing guidance, sequencing the majority of my *S. zooepidemicus* genomes, introducing me to Shovill and running the genomes through Pathogenwatch. I am grateful to have met you at the International Equine Infectious Diseases Conferences and the Getting to Grips with Strangles Workshops of the Havemeyer Foundation and also grateful for the kindness you have shown me.

Dr Ray Cursons!!!! This would not have happened without you! You enabled, guided, encouraged and provided all the opportunities. It has been quite the journey filled with many adventures! Thank you for sticking with me even through your “soft” retirement. I do not

know how to express my gratitude enough. You are one-of-a-kind and I am truly grateful for you!

To the best office mate I could have ever asked for, Dr Sari Karppinen! Thank you for your friendship, advise, wisdom and comic relief. And let us not forget your dear Jimmy's contribution to this work and Jouni's macro he wrote for my Excel data. Miss you!!!

To my managers, work colleagues and fellow students, thank you for your continued encouragement and support throughout this undertaking. Special thanks goes to Dr Joanna Hicks, Emma Voss and Sharna McCleary for taking the time to read my thesis and provide valuable feedback. This thesis is better because of you! Dr Ian McDonald, thank you for stepping up and helping me through the final important processes. I needed help and you were there without hesitation! Thank you!!

Bruce, as with my MSc, you never wavered in your support and confidence in me. I would not have been able to achieve this without you backing me! You are my everything! To Azra and Marshall, thank you for sharing the computer with me ☺ and for putting up with all the time I spent focused on this thesis. Azra, thank you for always checking up on me and asking how it was going. Marshall, your computer skills really saved me! Thank you for spending so much time trying to figure out what it was I wanted the computer to do and writing me scripts so it would do it! My family, I love you three!!

All my schooling, from elementary through to today I attribute to my parents, Sue and Roy. Thank you for your support and belief in my abilities from the very start, I'll never forget my roots. This PhD thesis is dedicated to you both! I wonder how many hamburgers Grandpa owes me now?!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SI (Systeme Internationale d'Unités) abbreviations are used throughout this thesis. Further abbreviations are listed below.

AHT	Animal Health Trust
AIC	akaike information likelihood-based criteria
aLRT	approximate likelihood-ratio statistical test
BHI	brain heart infusion
BLAST	basic local alignment search tool
bp	base pair
Cas	CRISPR-associated
CC	clonal complex
cfu	colony-forming units
cg	core genome
CI	confidence intervals
CM	complete media
CM-	complete media minus treatment antibiotics
CM+	complete media plus treatment antibiotics
COG	clusters of orthologous groups
CRISPR	clustered regular interspaced short palindromic repeat
DLV	double locus variants
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
EPS	extracellular polymeric substances
g	times gravity
GAS	Group A *Streptococcus*
GBS	Group B *Streptococcus*
GCS	Group C *Streptococcus*
GGS	Group G *Streptococcus*
GTR + G + I	general time reversible plus gamma and invariable sites models
GWAS	genome wide association studies
HGT	horizontal gene transfer
ICE	integrative and conjugative elements
indel	small insertions and deletions
iTOL	interactive tree of life
kDa	kilodalton
MLST	multilocus sequence typing
MOI	multiplicity of infection
NCBI	National Center for Biotechnology Information
NNI	nearest neighbour interchange
NZVP	New Zealand Veterinary Pathology

OD	optical density
OD_c	optical density biofilm cut-off value
PAI	pathogenicity island
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PFGE	pulsed-field gel electrophoresis
RPM	revolutions per minute
RNA	ribonucleic acid
SD	standard deviation
SDS	sodium dodecyl sulphate
SH	Shimodaira-Hasegawa
SLV	single locus variants
SMS	smart model selection
SNP	single nucleotide polymorphism
ST	strain type
SVS	SVS Laboratories
TA	toxin-antitoxin
TE	tris ethylenediaminetetraacetic acid
THB	Todd Hewitt broth
v/*v*	volume per volume

Definitions

Terms used throughout this thesis are defined as follows:

Commensal	bacteria benefiting from living in a host without causing harm or benefits to that host. Part of the host's normal flora.
Opportunistic pathogen	normally a commensal, potentially causing disease in times of stress when host's innate resistance is weakened.
Colonised	the presence of bacteria without eliciting an immune response. No clinical manifestations.
Infection	proliferating bacteria inducing an immune response with resultant clinical manifestations.
Pathogenicity	the state of being capable of causing disease.
Fitness factors	genes contributing to the survival of a bacterium in a given niche (an increase in fitness).
Virulence factors	types of fitness factors that are specifically attributed to the pathogenicity of a bacterium. Increases virulence of the bacteria and is usually associated with an increase in fitness.

Chapter 1

S. equi subsp. *zooepidemicus*: Background

1.1 Pyogenic Streptococci

The genus *Streptococcus* contains approximately 72 species of catalase-negative, Gram-positive cocci with a wide range of epidemiological and ecological characteristics (Richards *et al.*, 2014). Based on 16S rRNA sequencing, they are classed into Pyogenic, Bovis, Mutans, Mitis-Sanguinis, Anginosus, Salivarius or Other groups (Gobbetti & Calasso, 2014; Janda, 2014). Members of the pyogenic group have been further classified into Lancefield groups based on the presence and type of surface antigens and are associated with pus-inducing (pyogenic) infections in humans and animals (Sitkiewicz & Hryniewicz, 2010; Gobbetti & Calasso, 2014) (Sitkiewicz & Hryniewicz, 2010). The major groups of β-haemolytic, pyogenic streptococci are Lancefield Groups A, B, C, and G (**Table 1.1**). Group A *Streptococcus* (GAS), *Streptococcus pyogenes*, is an important human pathogen resulting in over half a million yearly global deaths (Sitkiewicz & Hryniewicz, 2010). *S. pyogenes* can cause infections of mucosal surfaces and skin such as pharyngitis/tonsillitis, impetigo and cellulitis with some infections causing post-infectious sequelae such as rheumatic fever and glomerulonephritis (Sitkiewicz & Hryniewicz, 2010). It is also the cause of severe invasive infections including streptococcal toxic shock syndrome, necrotising fasciitis, pneumonia and bacteraemia (Sitkiewicz & Hryniewicz, 2010). Group B *Streptococcus* (GBS), *Streptococcus agalactiae*, is a common major bovine pathogen causing mastitis as well as a human pathogen causing serious life-threatening disease in adults with underlying medical conditions and in neonates and young infants (Sitkiewicz & Hryniewicz, 2010). Group C and G *Streptococcus* (GCS and

GGS) are regarded as opportunistic animal pathogens which are emerging as important human pathogens causing invasive and often fatal disease (Sitkiewicz & Hryniewicz, 2010). The GCS contains *Streptococcus dysgalactiae* subsp. *dysgalactiae*, *Streptococcus dysgalactiae* subsp. *equisimilis*, *Streptococcus equi* subsp. *equi*, S*. equi* subsp. *zooepidemicus*, *S. equi* subsp. *ruminatorum, Streptococcus phocae* and *Streptococcus marimammalium* (**Table 1.1**). *S. equi* subsp. *zooepidemicus* is the species of focus for this thesis and for ease of reading, the three subspecies of *S. equi* will be referred to by their subspecies name: *S. equi* subsp. *equi* as *S. equi*, *S. equi* subsp. *zooepidemicus* as *S. zooepidemicus*, and *S. equi* subsp. *ruminatorum* as *S. ruminatorum*.

Table 1.1: The *Streptococcus* pyogenic group

Species	Lancefield Group*	HEM**	Host	Reference
S. pyogenes	A	β	human	(Sitkiewicz & Hryniewicz, 2010; Janda, 2014)
S. canis	G	β	dog, human, other animals	(Whatmore *et al.*, 2001; Janda, 2014)
S. dysgalactiae subsp. *equisimilis*	A, C, G, L	β	pig, cattle, horse, human	(Sitkiewicz & Hryniewicz, 2010; Janda, 2014)
S. dysgalactiae subsp. *dysgalactiae*	C, L	β	pig, cattle, human	(Sitkiewicz & Hryniewicz, 2010; Janda, 2014)
S. equi subsp. *equi*	C	β	horse, occasionally human and dog	(Sitkiewicz & Hryniewicz, 2010; Janda, 2014)
S. equi subsp. *zooepidemicus*	C, G	β	several animal species, human	(Sitkiewicz & Hryniewicz, 2010; Janda, 2014)
S. equi subsp. *ruminatorum*	C	β	sheep, goat, hyena, zebra, primate, human	(Speck *et al.*, 2008; Meyer *et al.*, 2011; Janda, 2014; Guo *et al.*, 2020)
S. ictaluri	-	γ	fish	(Collins *et al.*, 2000; Janda, 2014)
S. pseudoporcinus	B	β	human	(Janda, 2014)
S. porcinus	E, P, U, V, -	β	pig, human	(Janda, 2014)
S. uberis	E, P, B, G, -	α, γ	cattle	(Groschup *et al.*, 1991; Janda, 2014)
S. parauberis	P, U, G, E, -	α, γ	cattle	(Groschup *et al.*, 1991; Collins *et al.*, 2000; Gobbetti & Calasso, 2014; Janda, 2014)
S. iniae	-	α, β	fish, human	(Gobbetti & Calasso, 2014; Janda, 2014)
S. urinalis	-	γ	human	(Collins *et al.*, 2000; Janda, 2014)
S. agalactiae	B	α, γ, β	cattle	(Sitkiewicz & Hryniewicz, 2010; Gobbetti & Calasso, 2014; Janda, 2014)
S. castoresus	A	β	beaver	(Janda, 2014)
S. didelphis	-	β	opossum	(Gobbetti & Calasso, 2014; Janda, 2014)
S. halichoeri	B	γ	seal, human, badger, dog, fox, mink , finnraccon	(Janda, 2014; Aaltonen *et al.*, 2020)
S. hongkongensis	G	γ	fish, human	(Janda, 2014)
S. phocae	C, F, -	β	seal, porpoise	(Collins *et al.*, 2000; Janda, 2014)
S. fryi	M	β	dog	(Tomida *et al.*, 2011; Gobbetti & Calasso, 2014)
S. marimammalium	C	β	seal	(Lawson *et al.*, 2005; Tomida *et al.*, 2011; Gobbetti & Calasso, 2014)

*Letter designations of Lancefield grouping based on composition of cell wall antigens where a single dash indicates an un-groupable species. **Haemolysis (HEM): alpha (α) = incomplete haemolysis with the oxidising of haemoglobin resulting in a greenish zone around colonies, beta (β) = complete lysis of red blood cells resulting in a clearing zone around colonies, and gamma (γ) = lack of visible haemolysis.

1.2 Evolution of *Streptococcus equi* subspecies

Early evolution of *Streptococcus* was characterized by genome expansion through gene gain followed by a period of streamlining through gene loss, diversifying the major groups within this genus (Richards *et al.*, 2014). The majority of the present day species then evolved through a period of genome expansion (gene gain) with horizontal gene transfer (HGT) suspected to have played a large role in *Streptococcus* evolution (Richards *et al.*, 2014). The pyogenic group showed a high gene turnover (gene gain/loss) through a pattern of rapid adaptive radiation by HGT, with little correlation between the evolutionary relationships seen by the wide range of hosts and environments inhabited by the species of this group (Richards *et al.*, 2014). Genes involved in the uptake of sugar (phosphotransferase system) and the metabolism of carbohydrates have been fundamental in the evolution of this pyogenic group (Richards *et al.*, 2014). Further, genes involved in pathogenesis were also important within the group evolution with indications that they were acquired by HGT after the group branched (Richards *et al.*, 2014).

Mobile genetic elements have been instrumental in shaping the genomes of *S. zooepidemicus* through chromosomal inversions, recombination and HGT resulting in a diverse species (Beres *et al.*, 2008; Holden *et al.*, 2009; Ma *et al.*, 2013). There is strong evidence of genetic exchange between *S. equi* and *S. zooepidemicus* and with other species of the pyogenic group, in particular *S. pyogenes* through the identification of extensive shared gene content as well as with *S. agalactiae* and *S. dysgalactiae* (Beres *et al.*, 2008; Holden *et al.*, 2009; Waller & Robinson, 2013). Beres *et al*. (2008) postulates that *S. zooepidemicus* is naturally-competent with the presence of many homologues of competence genes in the *S. zooepidemicus* MGCS10565 genome, a strain responsible for an epidemic of nephritis in Brazil caused by the consumption of contaminated cheese (Beres *et al.*, 2008). They identified 22 of 23 streptococcal essential genes for natural transformation in this strain. Natural competence

would assist in the uptake of foreign DNA and evolution/adaptation of the species. However, there is also evidence that *S. zooepidemicus* maintains a stable genome by resisting the up-take of foreign DNA, balancing the need of both gene gain and genome stability. The lack of or relatively few prophages in *S. zooepidemicus* genomes (Beres *et al.*, 2008; Waller & Robinson, 2013) is suggestive of the subspecies resistance to the uptake and incorporation of foreign DNA through phage-mediated HGT, preventing stable prophage integration (Beres *et al.*, 2008). This lack of prophages is a trend seen in other naturally transformable streptococci and is postulated to be an effect of the presence of the competence system (Beres *et al.*, 2008). Clustered regular interspaced short palindromic repeat (CRISPR) elements and CRISPR-associated (Cas) genes, CRISPR-Cas systems, are also present in *S. zooepidemicus* genomes further providing resistance to the incorporation of prophages (Beres *et al.*, 2008; Holden *et al.*, 2009; Waller & Robinson, 2013). CRISPRs will be a part of the balancing act between gene gain, allowing for adaptation, and genome stability. Similarity of the CRISPR-Cas spacer sequences provides evidence that the pyogenic group share a common phage pool which would facilitate the exchange of DNA between the *Streptococcus* species through phage mediated HGT (Waller & Robinson, 2013).

The ability of the highly-related subspecies to exploit distinct niches gives evidence to the role of adaptive evolution in diversifying this species, for example S. *equi* is restricted to horses, *S. ruminatorum* is found in mastitic sheep and goats, and *S. zooepidemicus* infects a wide range of hosts (Richards *et al.*, 2014). The three subspecies share greater than 98 % DNA homology (Fernández *et al.*, 2004; Timoney, 2004). *S. zooepidemicus* is thought to be the archetype of *S. equi* based on multilocus enzyme electrophoresis (Jorm *et al.*, 1994) and the composition of 16S-23S RNA gene intergenic spacer sequences (Chanter *et al.*, 1997). Comparative genomics suggests that the acquisition of a novel system involving the uptake of iron was a key speciation event in the evolution of *S. equi* from *S. zooepidemicus* becoming host restricted to the horse (Holden *et al.*, 2009).

1.3 *Streptococcus equi* subspecies

Strangles is a well-known and highly contagious upper respiratory tract infection of equines caused by the bacterium *S. equi*. As its name suggests, it often involves abscessation of the head and neck lymph nodes which become large and painful affecting breathing and swallowing (Taylor & Wilson, 2006). The first description of the disease goes back to 1251 (Timoney, 1993b; Sweeney *et al.*, 2005) and has been referenced to in Europe since the 13th century (Taylor & Wilson, 2006). It is now a worldwide problem and one of the most diagnosed infectious diseases in horses, posing major welfare and economic impacts. In 2008 in the United Kingdom there were over 700 outbreaks of strangles, with some outbreaks involving over 200 horses and having significant associated costs (Waller *et al.*, 2011). *S. equi* can become established in the guttural pouch of the horse via the formation of chondroids, inspissated pus harboring bacteria, resulting in persistent asymptomatic infections which significantly influences the survival and spread of this disease on a global scale (Waller *et al.*, 2011).

Polymerase chain reaction (PCR) amplification and sequencing of the variable N-terminal region of the antiphagocytic M-protein (Sem) is used to identify the different strains of *S. equi* causing an infection (Anzai *et al.*, 2005; Kelly *et al.*, 2006; Ijaz *et al.*, 2011), through the use of the *S. zooepidemicus* multilocus sequence typing (MLST) database (http://pubmLst.org/szooepidemicus/). This allows easy comparison of strains found throughout the world (Jolley & Maiden, 2010). Using this method, two new *S. equi* Sem strains were found, *sem*-100 and *sem*-99, in the New Zealand horse population (Patty & Cursons, 2014). Strain *sem*-100 appeared more prevalent, causing nine out of the thirteen outbreaks in the study and appearing on both the North and South Island of New Zealand. Strain *sem*-99 was only found on the North Island in a total of four outbreaks. However, due to the low numbers of outbreaks, the dominance of *sem*-100 is yet to be confirmed. Continued surveillance is required to detect new strains emerging, which may have biosecurity implications. In comparison, there has been very little identification of *S. zooepidemicus* strains present within New Zealand and their associations with infection.

Conversely to *S. equi*, *S. zooepidemicus* is considered an opportunistic pathogen that causes disease only when its host is under situations of stress, such as viral infection, heat stress, or tissue injury (Timoney, 2004). In horses, *S. zooepidemicus* has been associated with respiratory disease, foal pneumonia, endometritis, and abortion (Mir *et al.*, 2013a) and has been isolated from equine joints, lymph nodes, nasal cavities, lungs, and reproductive tracts (Timoney, 2004; Casagrande Proietti *et al.*, 2011). In addition, it can be a serious zoonotic pathogen for people in contact with infected horses (Timoney, 2004; Pelkonen *et al.*, 2013) and is also the cause of acute infection in a number of other animals including but not limited to: dogs (Chalker *et al.*, 2012; Velineni *et al.*, 2014c), chickens (Roy *et al.*, 2013), goats (Mir *et al.*, 2013b), cats (Blum *et al.*, 2010), pigs (Ma *et al.*, 2013), and seals (Akineden *et al.*, 2005).

Like *S. zooepidemicus*, *S. ruminatorum* also has a varied host range. It was first described in 2004 as a novel subspecies isolated from sheep and goats with clinical and subclinical mastitis (Fernández *et al.*, 2004). It has since been isolated from hyenas and zebras as the likely causative agent of an outbreak of strangles-like symptoms in hyenas (Speck *et al.*, 2008). Further, two cases of human disease have been reported (Marchandin *et al.*, 2007; Meyer *et al.*, 2011). Most recently, it was discovered in the blood of a deceased rhesus marcaque (Guo *et al.*, 2020). As a novel subspecies, there is still a lot to learn about *S. ruminatorum*.

1.4 *Streptococcus zooepidemicus* and Equine Disease

Beta (β)-haemolytic streptococci are frequently associated with equine diseases. For instance, *S. zooepidemicus* was the most frequently isolated streptococci (72 % of the cases) in equine disease cases where β-haemolytic streptococci were isolated in the United States (Erol *et al.*, 2012). The *S. zooepidemicus* isolates were mostly recovered from foetal tissue, placentas and the genital system, as well as lymph nodes, the respiratory tract, joints, and other organs (Erol *et al.*, 2012). *S. equisimilis* was isolated in 21.3 % of the cases and at 5.8 %, *S. equi* was the least isolated, recovered from lymph node abscesses, the upper and lower respiratory tract, and guttural pouches as well as in two abortion cases. In 0.9 % of the cases the β-haemolytic streptococci remained unidentified, mainly isolated from foetuses and the genital tract (Erol

et al., 2012). Although frequently isolated, a positive *S. zooepidemicus* result creates uncertainty around it being the causative pathogen or just a commensal, due to its opportunistic lifestyle.

1.4.1 Respiratory Infections and *S. zooepidemicus*

Respiratory disease is an important cause of economic loss in the horse breeding industry and can have a major impact on racehorse performance and training. Respiratory disease is generally of viral or bacterial aetiology, however, the role of bacteria in respiratory disease is harder to assess and is less understood than viral infections (Christley *et al.*, 2001). In addition, it is thought that various other factors are involved in the disease. One study was unable to link viral, bacterial, or mycoplasma infection to 58 % of respiratory cases (Christley *et al.*, 2001). They theorised that either the bacterial infection had already resolved or the case was due to non-infectious airway disorders such as allergic irritation or non-allergic irritation following infectious or environmental challenge. Although associations can be made, the mere presence of bacteria isolated from respiratory infections does not confirm it as the primary causative agent, as many bacteria are part of the normal microbial flora.

From previous studies on bacteria associated with respiratory tract infections, *S. zooepidemicus* was shown to be one of the most commonly isolated bacterial species from horses with upper respiratory disease (Laus *et al.*, 2007; Jannatabadi *et al.*, 2008; Ryu *et al.*, 2011; Mir *et al.*, 2013a) and lower respiratory disease (Christley *et al.*, 2001; Newton *et al.*, 2003; Wood *et al.*, 2005b, 2005a). However, the association of *S. zooepidemicus* with respiratory infections appears to decrease with age (Barquero *et al.*, 2010), which may be indicative of age-acquired immunity. Post-mortem analysis of racehorses with fatal pneumonia resulted in the isolation of *S. zooepidemicus* in 72 % of the cases and in particular, was the main bacterial agent isolated from areas of lung abscess and necrosis (Carvallo *et al.*, 2017).

1.4.2 Uterine Infections and *S. zooepidemicus*

Bacterial infections occur in 25 % to 60 % of barren mares usually presenting in three forms: (1) venereal infections caused by the spread of a virulent pathogen between mares via live

cover by a stallion, (2) chronic infections due to persistent opportunistic bacteria in mares and (3) post-breeding infections (persistent breeding-induced endometritis) caused by genital commensals of the mare after natural mating or artificial insemination (Causey, 2006). Although there is a lack of information on bacterial and uterine interactions, there is compelling evidence that the uterine environment and its ability to clear bacteria play a role in whether a mare is more susceptible or resistant to bacterial infections (Causey, 2006).

Organisms associated with uterine infections include commensals (*S. zooepidemicus* and *E. coli*), pathogenic bacteria (*Klebsiella pneumoniae* types 1, 2 and 5, *Pseudomonas aeruginosa* and *Taylorella equigenitalis*) and anaerobic bacteria (*Bacteroides fragilis*) (Causey, 2006). Regardless of inflammation, isolation of bacteria from the uterus is strongly associated with decreased pregnancy rates (Riddle *et al.*, 2007). β-haemolytic streptococci and *E. coli* are the most frequently isolated organisms from the uterus of mares with fertility problems (Albihn *et al.*, 2003; Riddle *et al.*, 2007; Frontoso *et al.*, 2008) and are also the most common organisms isolated from mares regardless of fertility problems (Nielsen, 2005; Overbeck *et al.*, 2011; Urošević *et al.*, 2011; Walter *et al.*, 2012). β-haemolytic streptococci have been shown to have a higher association with clinical endometritis than with repeat breeding, however the opposite was true for *E. coli* (Albihn *et al.*, 2003). In addition, uterine cultures of β-haemolytic streptococci were associated with a higher number of positive cytological results than *E. coli*, shown by the presence of polymorphonuclear neutrophils indicating inflammation (Riddle *et al.*, 2007; Overbeck *et al.*, 2011; Walter *et al.*, 2012). *S. zooepidemicus* makes up the majority of the isolated β-haemolytic streptococci from the uterus of mares with fertility problems and endometritis (68 % to 92 %), with *S. equi* and *S. dysgalactiae* subsp. *equisimilis* following at two to three percent and six to 13 % respectively (Albihn *et al.*, 2003; Casagrande Proietti *et al.*, 2011).

A study from Denmark has highlighted the ability of *S. zooepidemicus* to establish dormant infection in the endometrium of sub-clinically infected mares (Petersen *et al.*, 2015). Barren mares, without clinical symptoms of endometritis or reproductive abnormalities that were culture negative for *S. zooepidemicus,* were treated with a novel bacterial growth medium (bActivate). This medium significantly induced active growth of *S. zooepidemicus* within these barren mares providing evidence of a dormant subclinical infection. Further, two different

S. zooepidemicus strains isolated from clinical endometritis could form persister cells, tolerant to 32-times the minimal inhibitory concentrations of penicillin, which may explain how the establishment of dormant undiagnosed uterine infections could occur (Petersen *et al.*, 2015).

1.4.3 Treatment of *S. zooepidemicus* Infections

To understand the current treatments for a *S. zooepidemicus* infection in New Zealand equines, a survey was conducted with veterinarians (n = 19) in October 2018 as part of this PhD project. Although there is no standard treatment for bacterial infections as they are tailored to the individual horse and stage of the disease process, the survey indicated that penicillin for respiratory infections (**Figure 1.1** A) and ceftiofur or gentamicin for uterine infections (**Figure 1.1** B) were the primary choices for treatment of *S. zooepidemicus*. One noted that antibiotics were not routinely used for treatment. When asked about usage of non-antibiotic treatments for *S. zooepidemicus*, respiratory infection treatments included antimicrobials (e.g. iodine), nonsteroidal anti-inflammatory drugs (e.g. phenylbutazone and flunixin), and mucolytics (e.g. DMSO). Non-antibiotic treatments for uterine *S. zooepidemicus* infections included saline flushing, oxytocin, kerosene and no treatment.

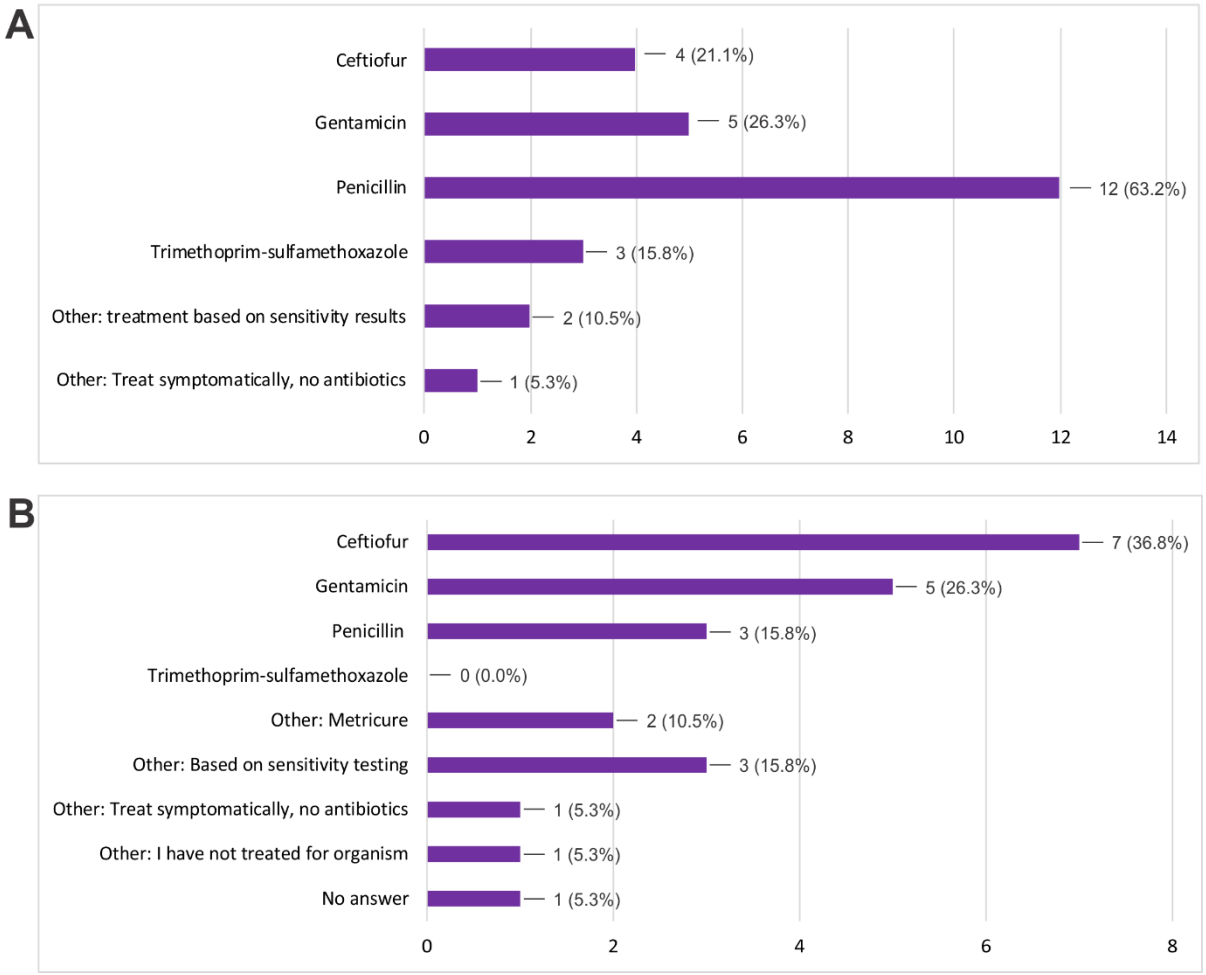


Figure 1.1: Survey of 19 New Zealand equine veterinarians on antibiotic treatments for *S. zooepidemicus* respiratory (A) and uterine (B) infections.

Streptococcus species isolated from respiratory infections (or colonisation) of young New Zealand horses, of which 40.3 % were identified as *S. zooepidemicus*, were equally susceptible to penicillin and ceftiofur *in vitro* (97.1 % and 97.0 % susceptibility respectively) with gentamicin following at 91.6 % (Toombs-Ruane *et al.*, 2015). Nearly 50 % of these isolates were resistant to trimethoprim-sulfonamide. Penicillin was therefore recommended as a first-line antimicrobial in New Zealand horses with suspected respiratory infection pending bacterial culture and susceptibility results (Toombs-Ruane *et al.*, 2015) which is reflected in the survey results. Although the antibiotic susceptibility of New Zealand reproductive *S. zooepidemicus* isolates remains to be investigated, 13 United States isolates from mares with postpartum metritis were 100 % susceptible to ceftiofur, penicillin, gentamicin and

trimethoprim-sulfonamides *in vitro* (Ferrer & Palomares, 2018). However, in an *ex vivo* model of the equine postpartum uterus, ceftiofur lacked bactericidal activity at levels representative of systemic administration against *S. zooepidemicus* whereas, a mixed formulation of penicillin and gentamicin was bactericidal at levels representative of local administration (Von Dollen *et al.*, 2019). This is indicative that the fluid uterine environment will affect antibiotic activity and the preference of ceftiofur in New Zealand should be accessed for clinical efficacy.

Through the survey, it was noted that equine veterinarians were dubious that *S. zooepidemicus* could be the primary pathogen in equine disease. Studies of *S. zooepidemicus,* in particular strain typing and relationship to equine disease, is lacking in New Zealand to validate this viewpoint.

1.5 *Streptococcus zooepidemicus* Strain Typing

Various molecular methods have been used to delineate strains of *S. zooepidemicus* in equines. An early PCR typing method targeted the 16S-23S rRNA gene intergenic region and involved four separate reactions (Chanter et al. 1997). Later, a PCR based method was developed towards the antigenic M-like Szp hypervariable regions involving five different reactions (Walker & Timoney, 1998). The Szp typing scheme was later simplified by developing a PCR-restriction length polymorphism (RFLP) approach using the DdeI restriction enzyme (Anzai *et al.*, 2002) as well as simply sequencing the *szp* gene (Anzai *et al.*, 2000; Ijaz *et al.*, 2011). In addition, pulsed-field gel electrophoresis (PFGE) has also been used, digesting genomic DNA with the SmaI restriction enzyme (Luque *et al.*, 2006). In 2008, a MLST scheme was developed (Webb *et al.*, 2008), which resulted in the establishment of a database (http://pubmLst.org/*szooepidemicus*/), allowing the direct comparison of strains between different labs (Jolley & Maiden, 2010). The latest attempts to identify different strains has involved the sequencing of the fibrinogen- and plasminogen-binding *szm* gene, which has been shown to vary between strains (Velineni & Timoney, 2013b), and is a partial homolog to the *sem* gene which is already used in the typing of *S. equi*.

To date, various typing studies have shown conflicting evidence of either opportunistic sources of *S. zooepidemicus* infections or infectious external sources of infection. There is evidence of *S. zooepidemicus* strains being able to both colonise and persist in the upper respiratory tract and are representative of the pneumonia-associated strains suggestive of an opportunistic source of infection (Anzai *et al.*, 2000; Anzai *et al.*, 2002). Conversely, there is also evidence showing clonal succession over time with some respiratory strains increasing in prevalence and others decreasing, suggestive of an infectious external source of infection (Newton *et al.*, 2008; Barquero *et al.*, 2010). However, high spatial strain variation between properties in both uterine (Luque *et al.*, 2006) and respiratory (Barquero *et al.*, 2010) infections suggest an opportunistic disease. Regardless of the internal opportunistic or external source of the strain, it will be through the acquisition and regulation of genes, particularly of virulence genes together with the host immunity that determines the success of a strain to proliferate and cause infection.

As a result of the development of the *S. zooepidemicus* MLST database, 130 allelic profiles of disease-associated isolates from horses, pigs, humans, and dogs over several countries were defined showing the genetic diversity of this species (Webb *et al.*, 2008) and this diversity is continuing to expand. The allelic profiles now stand at 413 strain types (STs) (http://pubmLst.org/szooepidemicus/, assessed 15/8/2020). Various trends have been noted through the clustering of STs such as strains from acute fatal hemorrhagic pneumonia in dogs related with some of the strain types found in various equine isolates (Webb *et al.*, 2008). It was also unexpectedly noted that strain types clustering with the strangles abscess-forming species *S. equi*, were significantly associated with cases of uterine infection or abortion in horses rather than isolates from non-strangles lymph node abscesses (Webb *et al.*, 2008).

1.5.1 *S. zooepidemicus* Strain Pathogenicity in Equines

Recent investigations have highlighted the pathogenic potential of particular *S. zooepidemicus* strains or a subgroup of related strains. One study identified a single MLST sequence type (ST-24) as the primary aetiological agent responsible for an outbreak of upper respiratory disease in Icelandic horses in 2009 (Lindahl *et al.*, 2013). Through Szp amino acid sequences and MLST analysis, the clone causing the disease was different to isolates from

healthy horses on the same property. This strain was also isolated from one of the affected horses eight months post-resolution suggesting the infectious strain could persist in horses after recovery or the horse had been re-infected by the same strain (Lindahl *et al.*, 2013). Similarly, yet on a larger scale, an epidemic of respiratory disease affecting almost the entire population of Icelandic horses in April 2010, was caused by a single strain of *S. zooepidemicus* (ST-209) in the absence of viral infection (Björnsdóttir *et al.*, 2012). This strain was again isolated from healthy Icelandic horses three years after the outbreak indicating the ability of pathogenic strains to colonise and persist in healthy horses after infection, becoming endemic in the population (Björnsdóttir *et al.*, 2017).

A second study characterised the Szm protein of a mucoid *S. zooepidemicus* strain from a clonal outbreak of equine respiratory disease in New Caledonia (Velineni & Timoney, 2013b). The outbreak persisted for 10 months and involved weanling and adult horses on at least 13 different properties. Interestingly, this outbreak occurred after the importation of six thoroughbred weanlings from New Zealand suggesting the emergence of a novel strain into a naïve susceptible population (Velineni *et al.*, 2014b). The isolates from this outbreak were mucoid with identical Szp, Szm amino acid sequences and MLST typing (ST-307). Comparison of ST-307 strain with those on the MLST database placed it into a clonal complex with 83 % of isolates recovered from the lower respiratory tract (Waller, 2014).

A final study suggests that a subgroup of genetically-related clones, rather than a single clone, causes uterine disease (Rasmussen *et al.*, 2013). These authors compared isolates between mares with and without clinical endometritis as well as isolates within mares, between colonies from the same sample and between different regions of the reproductive tract. PFGE was first used to group the isolates and then chosen isolates, representing the genetic diversity, were submitted to MLST analysis resulting in 21 different STs. The uterine isolates clustered significantly together forming a subgroup of endometritis associated with *S. zooepidemicus* strains and are genetically different from isolates of the caudal reproductive tract (Rasmussen *et al.*, 2013).

Overall, strain typing of *S. zooepidemicus* has shown huge genetic diversity within this species with some STs being more pathogenic than others. Changes in ST prevalence indicated

stronger virulence of some strains over others and/or acquired immunity of the host species. A pathogenic strain can become persistent/latent in its host post-resolution, and it has been speculated that within a horse the resident population of *S. zooepidemicus* includes pathogenic strains that reflect the history of infection, rather than just a population of harmless commensals (Waller, 2014). There is evidence of strains forming clusters or groups that are associated with disease and evidence that sources of infection can be either from endogenous or exogenous strains. Infection can also arise from novel strains forming through horizontal gene transfer between endogenous strains within its host (Velineni *et al.*, 2014a). The question that now remains is what factors influence the different fitness of particular strains in different disease and environment situations.

1.5.2 *S. zooepidemicus* Fitness Factors

The success of commensals and pathogens depends on fitness factors such as adhesion, biofilm production and/or resistance to phagocytosis allowing for increased survival and/or virulence. Through *S. zooepidemicus* genetic comparisons, the diversity within the species has been further defined and has given indication to potential fitness factors, which are important in the abilities of different strains to colonise different hosts and tissues. Such fitness factors also have implications for control and treatment of *S. zooepidemicus* infections. Fitness factors vary between strains of *S. zooepidemicus*, with some of these variances discussed below.

Phospholipase A2 (SlaA), a streptococcal virulence toxin, was encoded in some *S. zooepidemicus* strain types (44 of 140) compared to a second putative phospholipase A2 toxin, SlaB, found in all *S. equi* and *S. zooepidemicus* strains tested (Holden *et al.*, 2009). Paillot *et al*. (Paillot *et al.*, 2010a) identified three novel superantigen-encoding genes in *S. zooepidemicus*; virulence factors in the immunopathogenesis of streptococci. Forty-nine percent of a diverse group of *S. zooepidemicus* strain types contained at least one of the three novel superantigens. The presence of such novel superantigens was significantly associated with the isolation of *S. zooepidemicus* from equine cases of non-strangles lymph node abscessation. The absence of superantigens, however, does not necessarily imply a reduced pathogenicity. The highly virulent MLST ST-24 strain, responsible for the Icelandic clonal

horse respiratory outbreak, did not contain any superantigens (Lindahl *et al.*, 2013). In comparing the genomes of *S. equi* and a virulent human and a healthy horse strain of *S. zooepidemicus* (MGCS10565 and H70 respectively), with a virulent pig strain (ATCC35246), a group of novel pig pathogenic genes grouped together tightly in four regions termed pathogenicity islands (PAIs); three containing a toxin-antitoxin system and one containing a restriction modification system (Ma *et al.*, 2013). Deletion of part of one of the PAIs, which included genes belonging to a toxin-antitoxin system, increased host survival in experimentally-infected mice indicating its role in bacterial virulence (Ma *et al.*, 2013). Furthermore the virulence-associated protein E, VapE, along with proteins involved in adherence, were unique to the ATCC35246 pig strain (Ma *et al.*, 2013).

Pili play an important role in bacterial adhesion to their host and studies have shown differences in the pili loci between strains (Beres *et al.*, 2008; Holden *et al.*, 2009; Ma *et al.*, 2013) as well as within a pilus locus among a diverse group of *S. zooepidemicus* strain types (Holden *et al.*, 2009). This is suggestive of the pili's role in the ability to infect different hosts and tissues. Four loci have been identified, *FimI*, *FimII*, *FimIII* and *FimIV*, with only *FimI* found in all *S. zooepidemicus* and *S. equi* strains analysed (n=254) (Steward *et al.*, 2012). Over-expression of the *FimI* genes significantly increased bacterial adhesion suggestive of its importance in attachment and colonisation within a host (Steward *et al.*, 2012).

Biofilm production is another fitness factor involved in bacteria colonisation and persistence in hosts. Yi *et al*. (2013) identified *S. zooepidemicus* genes up-regulated *in vivo* during experimental infection of pigs with isolate ATCC35246 using the selective capture of transcribed sequences (SCOTS) method (Yi *et al.*, 2013). One of the up-regulated genes produces a fibronectin-binding protein, Fnz, a cell adhesion protein in host-pathogen interactions. The ATCC35246 strain was consistently able to form biofilms with a *fnz* knockout mutant strain having a significantly reduced biofilm formation (Ma *et al.*, 2013). The *fnz* mutant strain also demonstrated a significant reduction in its ability to adhere to Hep-2 mammalian cells and reduced virulence in experimentally-infected mice. Further, other virulence genes associated with cell adhesion were significantly down-regulated in the *fnz* mutant strain including *szp* (M-like protein), *zag* (IgG-binding protein) and *gapdh* (glyceraldehyde-3-phosphate dehydrogenase) (Ma *et al.*, 2013). An additional study

determined that most proteins expressed during biofilm formation were associated with metabolism, adhesion and stress conditions (Yi *et al.*, 2014). Collectively this suggests biofilm formation is an important virulence factor in the pig *S. zooepidemicus* strain.

Capsule production also plays a role in the virulence of bacteria. *S. zooepidemicus* typically produces small dry colonies at 37 °C but the level of capsule production (hyaluronic acid) increases at 23 °C (Velineni & Timoney, 2014). Although hyaluronic acid is antiphagocytic, its increased production at ambient temperatures may also reflect an adaptation to life on the skin and mucosal surfaces by preventing dehydration of the bacterium (Velineni & Timoney, 2014). The encapsulation of *S. zooepidemicus* varies with stably mucoid strains less common yet more likely to be the source of an epizootic outbreak (Salasia *et al.*, 2004; Lindahl *et al.*, 2013; Velineni *et al.*, 2014b). Mice vaccinated with recombinant NC78 proteins challenged with *S. zooepidemicus* NC78, the respiratory mucoid strain ST-307 originating from New Zealand, revealed proteins that are important in the establishment of *S. zooepidemicus* infection (Velineni & Timoney, 2013a). Three proteins in particular, provided high levels of resistance in mice; Szm, Szp and HylC (hyaluronidase). Both *szm* and *szp* are highly variable and have been previously discussed in their use in strain typing (section 1.5). Of note, it has also been found that the hyaluronidase, *hyal*, is associated with virulence of the mucoid ATCC35246 pig stain, and is speculated to facilitate bacterial spread in the host tissue (Zhou *et al.*, 2013). DnaK, part of the molecular chaperone complex, was identified in both mucoid strains, NC78 and ATCC35246 (Marcellin *et al.*, 2009; Velineni *et al.*, 2014b) but absent in the non-mucoid isolates suggestive of its involvement in the virulence of the mucoid strain (Velineni *et al.*, 2014b).

The fitness of individual bacterial strains is also affected by host-microbe interactions. The actual effectiveness of the bacteria's repertoire of fitness factors will be influenced by the host environment. Host-pathogen interactions are dynamic as indicated by the previously mentioned changes in ST occurrences over time and novel strains causing outbreaks in naïve horses. These are reflections of the adaptive changes in the immunity of those particular host sub-populations.

1.6 Summary

Respiratory and uterine disease play an important role in the equine industry with large welfare and economic implications. β-haemolytic streptococci are often involved in these diseases with *S. equi*, a well-known clonal pathogen of the upper respiratory tract, causing the equine disease strangles, whereas *S. zooepidemicus* is associated with both respiratory and uterine diseases. Unlike *S. equi*, *S. zooepidemicus* is highly genetically diverse and generally regarded as a common commensal of the horse mucosa, opportunistically causing secondary infections when bacterial-host interactions are compromised. Current research has revealed some strains to be more virulent than others in certain horse populations causing primary infection, as well as some strains associated with particular diseases. Acquisition and expression of different genes will influence the different strains ability to infect different hosts and tissues.

1.7 Aim

The overarching aim of this thesis is to enhance our understanding of the epidemiology of *S. zooepidemicus* in respiratory and uterine diseases within the New Zealand horse population. Investigation of both genomic and phenotypic characteristics of strains associated with disease will greatly increase our understanding of the different strains and their pathogenic potential. Advancing our knowledge of the *S. zooepidemicus* strains present in New Zealand will help determine the bacterium's importance and risk factors associated with particular horse diseases, ultimately contributing to their successful monitoring and management.

In order to achieve this aim, four objectives were set out:

Objective 1: Determine the relative prevalence of *S. zooepidemicus* in the New Zealand horse population compared with other equine bacterial pathogens.

Objective 2: Use strain typing to distinguish between strains and determine any correlation between strain and disease status.

Objective 3: Investigate phenotypic variation of strain types through biofilm and mammalian cell invasion assays.

Objective 4: Characterise the genomic variation of *S. zooepidemicus* strains using genome sequencing to identify genetic determinants of pathogenicity.

Chapter 2

Prevalence and Strains of *Streptococcus zooepidemicus* in the New Zealand Horse Population

2.1 Introduction

Streptococcus zooepidemicus can be isolated from both healthy and infected horses and is considered an opportunistic commensal, only causing disease when the host is under situations of stress (Timoney, 2004). It is particularly associated with horse respiratory and uterine infections, consistent with isolation from lymph nodes, nasal cavities, lungs and reproductive tracts (Timoney, 2004; Casagrande Proietti *et al.*, 2011). There has been little epidemiological investigation of this bacterium in New Zealand horses, yet S*. zooepidemicus* was the most frequently isolated bacterium in respiratory samples sent to commercial diagnostic laboratories, from New Zealand horses aged four weeks to three years, between 2004-2014 (Toombs-Ruane *et al.*, 2015). Conversely, isolation of *S. zooepidemicus* from the respiratory tract of healthy New Zealand horses is low at 6.5 % (Acke *et al.*, 2015). An association was found between a history of respiratory disease over a six month period in New Zealand dogs and horses with the isolation of *S. zooepidemicus*, although small numbers made the significance questionable (Acke *et al.*, 2015).

Of the various molecular methods used to delineate strains of *S. zooepidemicus*, the *szp* virulence gene (Walker & Timoney, 1998) and the MLST scheme (Webb *et al.*, 2008) are the most accepted. Strain typing has allowed researchers to investigate the epidemiology of individual strains involved with disease and have shown that rather than being a clonal

population, *S. zooepidemicus* isolates represent a wide genetic diversity of strain types (Webb *et al.*, 2008). Initially, the role of *S. zooepidemicus* as a primary pathogen in equines was debatable but recently this view is starting to be accepted with evidence that specific strains are the primary source of infection in respiratory (Lindahl *et al.*, 2013) and uterine disease (Rasmussen *et al.*, 2013). One study in particular reported a 10-month long respiratory outbreak involving weanling and adult horses on various different properties in New Caledonia (Velineni *et al.*, 2014c). Interestingly, this outbreak was associated with the importation of six thoroughbred weanlings from New Zealand suggesting the emergence of a novel "immigrant" strain into an immunological naïve horse population resulting in disease (Velineni *et al.*, 2014c).

2.1.1 Aim

Currently there is little knowledge on the genetic diversity of *S. zooepidemicus* strains isolated from New Zealand horses. To comprehend the role of *S. zooepidemicus* in horse respiratory and uterine disease in New Zealand, prevalence data obtained from submitted clinical samples to diagnostic laboratories, together with an analysis of the genetic diversity of strains within New Zealand equids, were investigated.

2.2 Methods

2.2.1 Prevalence of *S. zooepidemicus* Retrospective Study

Equine sample records from New Zealand Veterinary Pathology (NZVP, Hamilton, New Zealand) between 5 January 2009 and 1 September 2014 were retrospectively assessed. All horse samples regardless of age, sex or breed and from all anatomical sites were included for analysis. Two samples submitted from the same horse at the same time were considered together. Excluded from the analysis were samples assumed to be from "clinically well" horses described as "screening prior to export" and samples taken from stable/yard or mixed samples (multiple horses). All isolates from mixed infections were included. There was no knowledge of pre-treatment with antibiotics nor was there consideration for coinfection with viruses, fungi or parasites. Samples were classed as respiratory, urogenital, musculoskeletal

and soft tissue, digestive, and unknown. Respiratory included all samples likely to reflect the respiratory system based on specimen descriptions and clinical details. They included respiratory associated abscess/lymph nodes and guttural pouch samples. Urogenital samples included uterus, umbilical/placenta/foetal, vaginal, urine/urethral, penis/semen, and mastitis/milk samples. Musculoskeletal and soft tissue were grouped together as an underlying wound cause could not be ruled out. These included joint, tendon, bone, hoof, skin/hair, wound, ear/eye, abdominal fluid/seroma, mouth ulcers, and vaccine injection/post-surgical infection sites. The digestive class included faecal and stomach/intestine contents. Unknown included samples where not enough anatomical detail was available to designate a class including abscesses with no information on sample location. Data is described by counts and percentages with 95 % confidence intervals (CI). Pearson's chi-squared test was used to determine p-values for isolation sites.

2.2.2 University of Waikato *S. zooepidemicus* Isolate Collection

Cultured *S. zooepidemicus* isolates came from a collection established at the University of Waikato and held at minus 80 °C storage. These isolates included, wherever possible, up to five randomly chosen *S. zooepidemicus* colonies per horse sample. The isolates were classed by sample group as respiratory disease, respiratory abscess, respiratory apparently healthy, and uterus. The majority of isolates were cultured on sheep blood agar plates from 51 veterinarian submitted horse samples during 2010-2017. These samples included nasal swabs, nasopharyngeal swabs, abscess swabs, a tracheal wash, a guttural pouch wash, uterus swabs, and uterine fluid. On four occasions, veterinarians submitted multiple individual horse samples from a herd or a single property with a suspected respiratory disease outbreak. A fifth "partial herd" was sampled which included isolates from 15 feral Kaimanawa horse nasopharyngeal swabs, collected post-mortem from the 2014 Department of Conservation muster for culling purposes. This particular group of isolates made up the "respiratory apparently healthy" group as respiratory disease was not observed at the time within this feral group. A further 42 single uterus isolates were received from two Hamilton veterinary diagnostic laboratories, NZVP and SVS, in 2016. Overall, the samples equated to 286 *S. zooepidemicus* isolates from 108 individual horses originating from Auckland, Waikato,

Southland, Manawatū-Whanganui, and Masterton regions of New Zealand. Details of all isolates can be found in **Table S2. 1**.

2.2.3 *Szp* and MLST Typing of the University of Waikato *S. zooepidemicus* Isolates

DNA was extracted using Presto™ Mini gDNA Bacteria Kit (Geneaid, GBB300) following manufacturer's instructions with the addition of 4 mg/mL lysozyme (Sigma, L-7001) and 3000 U/mL mutanolysin (Sigma, M9901-5KU) to the Gram+ Buffer. Each isolate was typed based on the nucleotide sequence of the *szp* virulence gene (Walker & Timoney, 1998; Velineni *et al.*, 2014c). The *szp* PCR consisted of 0.5 U HOT FIREPol® DNA polymerase (Solis Biodyne, 01-02-01000), 1 x Buffer B1, 2 mM $MgCl_2$, 0.2 mM dNTPs (Genscript, C01582-10), 0.5 µM of IGSzPF and IGSzPR primers (Ijaz *et al.*, 2011) and 50 ng of the extracted DNA template. PCR parameters were as follows: an initial activation/denaturation at 95 °C for 15 min, followed by 30 cycles of denaturation at 95 °C for 15 s, annealing at 60 °C for 30 s, and extending at 72 °C for 2 min, ending with a final extension at 72 °C for 5 min. The resulting amplified products, averaging around 1220 bp, were treated to an enzyme clean-up by adding 1 U of rAPid alkaline phosphatase (Roche, 4898133001) and 10 U exonuclease 1 (Fermentas, EN0581) for 30 min at 37 °C and then inactivated at 85 °C for 15 min. Both forward and reverse amplified DNA strands were sequenced using the same PCR primers. Geneious R11.1.4 (https://www.geneious.com/) was used to align (using ClustalW) and analyse the sequences. *Szp* types were assigned to each isolate based on single nucleotide polymorphisms (SNPs). The NCBI Basic Local Alignment Search Tool (BLAST) (https://blast.ncbi.nlm.nih.gov/) was used to compare the *szp* strain type sequences with those in the nucleotide collection database to give a global comparison. A subset of this *szp* typing pool, consisting of a range of *szp* types and isolates from different anatomical isolation sources, were additionally used for MLST typing using the previously described method (Webb *et al.*, 2008). This included amplifying and sequencing the internal fragments of seven housekeeping genes: carbamate kinase (*arc*C), ribonucleoside-diphosphate reductase (*nrdE*), prolyl-tRNA synthetase (*proS*), signal peptidase I (*spi*), thymidylate kinase (*tdk*), triosephosphate isomerase (*tpi*), and acetyl-CoA acetyltransferase (*yqiL*). The MLST sequencing data from this subset was then submitted to the *S. zooepidemicus* MLST database (https://pubmLst.org/*szooepidemicus*/) sited at the University of Oxford (Jolley *et al.*, 2018), to be assigned sequence types (STs).

2.2.4 Phylogenetic Analysis of *S. zooepidemicus* MLST Sequence Types (STs)

Allelic profiles were used to infer genetic relationship among strains with PHYLOViZ 2.0 software (Nascimento *et al.*, 2016) using the geoBURST algorithm (Francisco *et al.*, 2009) creating a full minimum spanning tree. The algorithm was applied to the MLST numeric profiles to implement a model of clonal expansion and diversification based on the number of differences between the profiles. This draws links between STs that correspond to the most probable pattern of descent (Francisco *et al.*, 2009). A clonal complex (CC) was defined as a group of STs that are linked together with single and double locus variants (SLV and DLV) within their profiles.

Maximum likelihood trees, implemented by PhyML 3.0, were created to infer evolutionary relationships (Guindon *et al.*, 2010). For tree construction a General Time Reversible model (GTR) along with both gamma distribution (G) and proportion of invariable sites models (I), GTR + G + I, were used as determined by the Smart Model Selection (SMS) in PhyML with Akaike Information likelihood-based criteria (AIC) (Guindon *et al.*, 2010; Lefort *et al.*, 2017). A BioNJ starting tree with Nearest Neighbor Interchange (NNI) starting tree topology improvement was implemented within PhyML 3.0. The approximate Likelihood-Ratio statistical Test (aLRT) with Shimodaira-Hasegawa (SH)-like option was selected for branch support estimation (Anisimova & Gascuel, 2006; Guindon *et al.*, 2010). The aLRT branch support is considered significant when it is larger than 0.90-0.99 and with the SH-like option, branch supports greater than 0.9 tend to be similar to bootstrap proportions greater than 0.75 (Guindon *et al.*, 2010). Trees were constructed with ClustalW-aligned nucleotide sequences. Resultant trees were then visualised and annotated using an online tool, interactive Tree Of Life (iTOL) (Letunic & Bork, 2019).

2.3 Results

2.3.1 Retrospective Study of *S. zooepidemicus* Prevalence

With the data received from NZVP diagnostic lab, there was no knowledge of pre-treatment with antibiotics nor were other infectious agents (viruses, fungi, parasites) or biological

factors considered, therefore a causative link between disease and isolation could not be made in the retrospective analysis. Over a period of five and a half years, 1608 clinical horse samples were submitted to NZVP for culture, resulting in 3816 bacterial isolates. *S. zooepidemicus* isolates were highly-prevalent being the second most frequently isolated bacterium within the samples at 27.8 % (95 % CI: 25.6 % to 30.0 %) with alpha-haemolytic streptococci being the most prevalent at 32.2 % (95 % CI: 30.1 to 34.6%) (**Table 2.1**). No significant difference occurred in the isolation frequency of *S. zooepidemicus* between respiratory and urogenital sites (p = 0.195) (**Figure 2.1**). However, *S. zooepidemicus* had a significantly greater probability to be isolated from respiratory (p = 0.0002) and urogenital (p = 0.002) sites than from the musculoskeletal and soft tissue sites (**Figure 2.1**). Although not as prevalent, *S. zooepidemicus* was also found in a range of musculoskeletal and soft tissues sites including abdominal fluid, abscesses from hoof and eye sites, eye infections, post-surgical and vaccine sites, joint fluid, skin disease, and wounds.

Table 2.1: Proportion of bacterial isolates of horse samples submitted for culture to NZVP (2009-2015).

Species Isolated	Gram Strain	Total Culture Positive n	Proportion Positive per Total Samples %	(95 % CI)
Alpha-haemolytic Streptococci	+	518	32.2	(29.9-34.5)
Bacillus species	+	280	17.4	(15.6-19.3)
Enterococcus species	+	198	12.3	(10.7-13.9)
Non-haemolytic Streptococci	+	82	5.1	(4.0-6.2)
Rhodococcus equi	+	11	0.7	(0.3-1.1)
Staphylococcus aureus	+	362	22.5	(20.5-24.6)
Staphylococcus species (Coagulase-negative)	+	284	17.7	(15.8-19.5)
Streptococcus dysgalactiae subsp. *equisimilus*	+	219	13.6	(11.9-15.3)
Streptococcus equi subsp. *equi*	+	61	3.8	(2.9-4.7)
Streptococcus equi subsp. *zooepidemicus*	+	447	27.8	(25.6-30.0)
Other (Gram-positive) *	+	158	9.8	(8.4-11.3)
Enterobacter species	-	272	16.9	(15.1-18.7)
Escherichia coli	-	341	21.2	(19.2-23.2)
Klebsiella species	-	134	8.3	(7.0-9.7)
Pseudomonas species	-	123	7.6	(6.4-8.9)
Salmonella species	-	1	0.1	(-0.1-0.2)
Other (Gram-negative) **	-	318	19.8	(17.9-21.7)
Unknown Gram-neg or -pos		7	0.4	(0.1-0.8)
Total isolates		3816		
Total samples		1608		

*Other Gram-positive include: *Actinomyces* species, *Arcanobacterium* species, *Bergeyella zoohelcum,* β-haemolytic Streptococci, *Brevibacterium linens, Corynebacterium* species, *Dermatophilus congolensis, Hafnia alvei, Lactobacillus* species, *Micrococcus* species and *Staphylococcus intermedius/pseudintermedius.* **Other Gram-negative include: *Acinetobacter* species, *Actinobacillus* species, *Aeromonas* species, *Alcaligenes* species, *Citrobacter* species, *Flavobacterium* species, *Moraxella* species, *Morganella morganii, Pasteurella* species, *Proteus* species, *Providencia* species, *Serratia* species and *Stenotrophomonas maltophilia.*

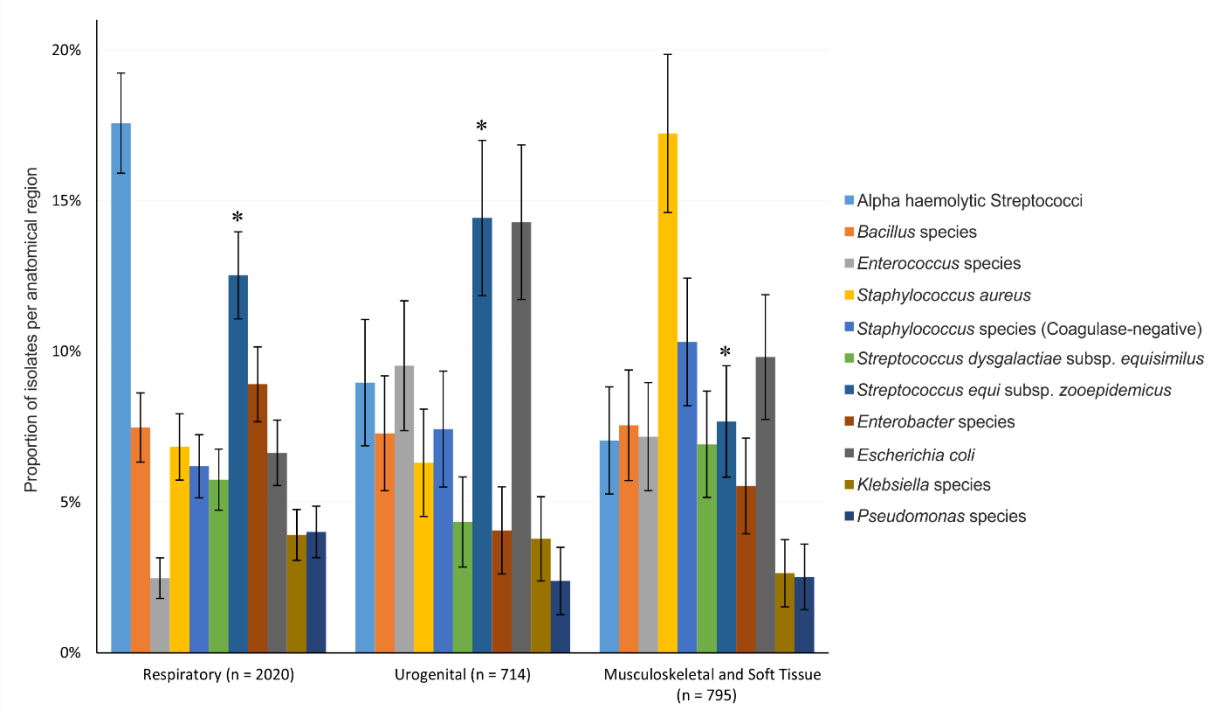


Figure 2.1: Prevalence graph of the proportion of total bacterial isolates with respect to three anatomical regions: respiratory, urogenital, musculoskeletal and soft tissue (top 11 shown). Error bars represent 95 % CI. **S. zooepidemicus* has a significantly greater probability to be isolated from respiratory ($p = 0.0002$) and urogenital ($p = 0.002$) sites than from the musculoskeletal and soft tissue sites. No significant difference of *S. zooepidemicus* isolation probability between respiratory and urogenital sites ($p = 0.195$). Faecal and samples with unknown origin were excluded from this analysis.

2.3.2 A Priori Sampling Information for the University of Waikato *S. zooepidemicus* Isolate Collection

The University of Waikato's collection of *S. zooepidemicus* isolates was established from veterinarian submitted samples, diagnostic laboratory cultured isolates and post-mortem collected Kaimanawa horse nasopharyngeal swabs. This sampling method would have led to certain biases that should be noted and considered when analysing the data. The veterinarian collected and submitted samples along with the diagnostic laboratory isolates caused a location bias with the majority of samples coming from the Waikato region. This Waikato bias however is reflective of the location of the highest population of New Zealand thoroughbreds. Some samples did originate from other regions including Auckland, Masterton, Southland,

and the feral Kaimanawa population in the Manawatū-Whanganui region. There was also no control over how the samples were collected, but the methods did reflect standard New Zealand veterinary practice. Multiple isolates were not collected from all samples and in particular, single isolates, rather than samples, were provided from the two diagnostic laboratories. A high number of isolates with one strain type may only reflect multiple horses from one property involved in an outbreak rather than a reflection of the genetic diversity of *S. zooepidemicus* in that particular herd or of the entire sample population.

2.3.3 Genetic Diversity of the University of Waikato *S. zooepidemicus* Isolate Collection

2.3.3.1 *Szp* typing – Samples with Multiple Isolates

All 286 isolates of the University of Waikato *S. zooepidemicus* collection, from the 108 horses, were *szp* typed resulting in 48 *szp* strains. These sequences were uploaded to the National Center for Biotechnology Information (NCBI) database with assigned accession numbers listed in **Table S2. 2**. Multiple isolates (up to five isolated colonies were picked wherever possible) were obtained and analysed from 52 of the 108 horses (**Table S2. 1**). Thirty-eight of the horses with multiple isolates were classed as clinical from the respiratory disease (n = 28), respiratory abscess (n = 1) or uterus (n = 9) sample groups. The nine uterine samples included both clinical and subclinical samples. The other 14 samples with multiple isolates were classed as respiratory apparently healthy from the Kaimanawa feral herd. The number of *szp* strain types isolated from a single clinical sample ranged from one to two, as compared with one to four strains from the respiratory healthy samples. A strain was considered to be clonally expanded if it was the only strain type identified from multiple isolates of a single sample. Importantly, there was a highly significant difference between clinical samples from suspected diseased horses when compared with samples from apparently healthy horses ($p < 0.0005$). Seventy-six percent (29/38) of the clinical respiratory and the uterus samples had an expansion of a single strain compared to those samples from horses considered to be healthy. Conversely, the apparently healthy horses were associated with the isolation of multiple strains in 86 % (12/14) of the samples. The *szp* typing results of these 52 samples with multiples isolates are further detailed below.

In the nine uterine samples analysed, multiple isolates were obtained per sample and this resulted in 44 isolates and six *szp* types (**Table 2.2**). Importantly, unlike the respiratory samples, only one *szp* type was isolated per sample regardless of clinical status (normal, pregnancy loss, failed pregnancy, vaginal discharge, endometritis, old infertile mare, short cycle, or not declared). *Szp*-8 was clonally expanded in two healthy, subclinical, individual horse uterine samples and was the only strain type present. Similarly, *szp*-11, -34 and -21 were also clonally expanded with no other strains present in horse samples that were either clinical or had no health status declared. There was one exception, sample WK15114 which had two *szp* types present and had an undeclared health status. One isolate from this horse uterus typed as *szp*-39 and the other four isolates were *szp*-40 (**Table 2.2**).

Table 2.2: *Szp* types of uterine samples with multiple isolates

Sample ID	Date	Clinical Details	# of isolates	# of *szp* types	*szp* type					
					8	11	21	34	39	40
WK15092	27/08/2015	normal	5	1	5					
WK15094	27/08/2015	pregnancy loss	4	1		4				
WK15096	27/08/2015	normal	5	1	5					
WK15111	8/09/2015	vaginal discharge, failed pregnancy last year	5	1				5		
WK15114	1/10/2015	not declared	5	2					1	4
WK15117	8/10/2015	not declared	5	1				5		
WK15123	8/11/2015	old infertile mare	5	1			5			
WK15125	9/11/2015	preg loss/ short cycle	5	1			5			
WK16004	15/01/2016	endometritis	5	1			5			

The apparently healthy, Kaimanawa feral herd sample population, represented 15 horses with a total of 56 *S. zooepidemicus* isolates which resulted in ten *szp* strain types (**Table 2.3**). As mentioned above, where more than one isolate was obtained per sample, the isolates were of multiple *szp* strain types. Of these 10 *szp* types, *szp*-1 and -15 each colonised a single horse; *szp*-19 and -44 each colonised two horses; *szp*-21, -29 and -30 each three horses; *szp*-25 colonised five horses; *szp*-46 colonised six horses and *szp*-42 colonised seven of the healthy horses (**Table 2.3**). These results reflect the issue of multiple *S. zooepidemicus* strains in nasopharyngeal carriage within apparently healthy horses and demonstrate the genetic

diversity of *S. zooepidemicus* strain carriage within an individual animal and within a herd (**Table 2.3**).

Table 2.3: *Szp* types of the apparently healthy, Kaimanawa samples

Sample ID	Date	Region	# of isolates	# of *szp* types	*szp* type									
					1	15	19	21	25	29	30	42	44	46
KMW001	30/05/2014	Manawatū-Whanganui	1	1					1					
KMW002	30/05/2014	Manawatū-Whanganui	4	2					1			3		
KMW003	30/05/2014	Manawatū-Whanganui	4	3				1	1			2		
KMW004	30/05/2014	Manawatū-Whanganui	5	4			1					1	2	1
KMW005	30/05/2014	Manawatū-Whanganui	2	1								2		
KMW006	30/05/2014	Manawatū-Whanganui	5	3					1		1			3
KMW007	30/05/2014	Manawatū-Whanganui	5	1				5						
KMW008	30/05/2014	Manawatū-Whanganui	3	3		1					1			1
KMW009	30/05/2014	Manawatū-Whanganui	3	2				1		2				
KMW010	30/05/2014	Manawatū-Whanganui	5	3						2		2		1
KMW011	30/05/2014	Manawatū-Whanganui	3	2									2	1
KMW012	30/05/2014	Manawatū-Whanganui	4	2					2	2				
KMW013	30/05/2014	Manawatū-Whanganui	2	2	1							1		
KMW014	30/05/2014	Manawatū-Whanganui	5	2							1			4
KMW015	30/05/2014	Manawatū-Whanganui	5	2			3					2		

This genetic diversity and carriage within herds was also observed in the four properties with respiratory outbreaks involving multiple horses. Although most clinical horses exhibited clonal expansion of a single *szp* strain (**Table S2. 1**), genetic diversity of *S. zooepidemicus* causing infection was found across animals within the herd rather than within an individual animal. One Waikato farm outbreak, referenced as W1, had 15 yearlings affected with mucopurulent nasal discharge. These horses were initially sampled with nasal swabs because they were in close contact with a confirmed strangles horse. All 15 yearlings were PCR-positive for *S. zooepidemicus* with cultured isolates obtained from ten of the 15 horses with a total of 42 isolates and three *szp* types identified (**Table 2.4**). *Szp*-22 was isolated from seven out of ten horses of which two horses also had *szp*-12 isolated. *Szp*-12 was further

isolated from two more horses and *szp*-28 was isolated from the tenth horse. Clonal expansion was seen with all three of these *szp* types (**Table 2.4**).

Table 2.4: *Szp* types of Waikato respiratory outbreak W1

ID	Date	Region	# of isolates	# of szp types	*szp* type		
					12	22	28
WK14020	19/11/2014	Waikato	1	1		1	
WK14027	21/11/2014	Waikato	5	1		5	
WK14029	21/11/2014	Waikato	2	1			2
WK14030	21/11/2014	Waikato	4	2	3	1	
WK14035	21/11/2014	Waikato	5	1		5	
WK14036	21/11/2014	Waikato	5	1	5		
WK14037	21/11/2014	Waikato	5	1		5	
WK14038	21/11/2014	Waikato	5	1	5		
WK14039	21/11/2014	Waikato	5	1		5	
WK14040	21/11/2014	Waikato	5	2	1	4	

One of the *szp*-22 isolates from the W1 respiratory outbreak was from a submandibular abscess (WK14020) (**Table 2.4**). Note that the abscess sample was also PCR-positive for *S. equi* however only *S. zooepidemicus* was able to be culture isolated. None of the other samples from this herd were PCR-positive for *S. equi*. This yearling had been vaccinated for *S. equi* as a weanling which may explain the detection of *S. equi* by PCR, as the live vaccine PINNACLE® I.N. (Zoetis) is used in New Zealand, however this was not investigated. *Szp*-22 was also isolated from another respiratory associated abscess (WK15021) on a different property in the same region three months later with no detection of *S. equi* (**Table S2. 1**). Interestingly, *szp*-22 is an SNP variant from *szp*-23 (nt position 352, G→T) that results in a point-nonsense mutation in the hypervariable region of the gene. This premature stop would presumably result in a truncated protein upstream from the LPSTGE sortase-dependent surface-anchor protein sequence (**Figure 2.2**).

Figure 2.2: *Szp*-22 nucleotide and resulting amino acid sequence created in Geneious 11.1.4 (https://www.geneious.com/). Red arrow indicates position of stop codon due to single nucleotide polymorphism (nt position 352, G→T). N-terminal, hypervariable region, PEPK repeat domain and the cell wall anchor sequence are indicated by grey horizontal arrow under the sequence.

A second respiratory outbreak on a different Waikato property, referenced as W2, involved five horses and consisted of high fever, cough and mild adventitious lung sounds (**Table 2.5**). A total of 25 isolates were obtained from this outbreak with five isolates per horse, resulting in three *szp* types. *Szp*-5 was isolated from three out of five horses, two of which also had *szp*-28 present and *szp*-23 was isolated from two out of five horses. Clonal expansion was seen with *szp*-5 and *szp*-23 (**Table 2.5**).

Table 2.5: *Szp* types of Waikato respiratory outbreak W2

Sample ID	Date	Region	# of isolates	# of szp types	*szp* type		
					5	23	28
WK15034	11/03/2015	Waikato	5	1		5	
WK15036	13/03/2015	Waikato	5	1	5		
WK15037	13/03/2015	Waikato	5	2	2		3
WK15038	13/03/2015	Waikato	5	1		5	
WK15039	13/03/2015	Waikato	5	2	2		3

The third and fourth respiratory herd outbreaks were on Auckland properties referenced as A1 and A2. Outbreak A1 involved two six-month-old foals (**Table 2.6**). Sample AK15061 had a purulent nasal and ocular discharge with an abscess at the base of its left ear. An abscess sample was provided from this horse with the resultant five isolates of *S. zooepidemicus* colonies all typing as *szp*-24. The second horse, AK15060, had purulent nasal discharge and was in contact with the abscessed foal. The nasal swab from this horse resulted in five isolates and two *szp* types, *szp*-24 and -28 (**Table 2.6**).

Table 2.6: *Szp* types of Auckland respiratory outbreak A1

Sample ID	Date	Region	# of isolates	# of szp types	*szp* type	
					24	28
AK15060	15/04/2015	Auckland	5	2	2	3
AK15061	15/04/2015	Auckland	5	1	5	

The fourth respiratory outbreak, A2 was on a different Auckland property and involved seven foals with nasal discharge, increased chest sounds and some coughing symptoms (**Table 2.7**). Five different *szp* strains resulted from the 31 isolates from the seven horse nasal samples (*szp*-2, -18, -21, -27, 36). *Szp*-2, -18, -21 and -36 were involved in clonal expansion (**Table 2.7**).

Table 2.7: *Szp* types of Auckland respiratory outbreak A2

Sample ID	Date	Region	# of isolates	# of szp types	*szp* type				
					2	18	21	27	36
AK17075	5/04/2017	Auckland	5	1					5
AK17076	5/04/2017	Auckland	5	1		5			
AK17077	5/04/2017	Auckland	2	1			2		
AK17078	5/04/2017	Auckland	5	1	5				
AK17079	5/04/2017	Auckland	5	1		5			
AK17080	5/04/2017	Auckland	4	2				3	1
AK17081	5/04/2017	Auckland	5	1	5				

The final six samples, of the 52 with multiple isolates obtained, were from horses with respiratory disease yet were not linked to a respiratory disease outbreak involving multiple horses (**Table 2.8**). Clonal expansion was seen in four of the six samples with *szp*-28, -36, -5 and -41 involved. WK14049 was from a foal with runny eyes and nasal discharge. The mother mare was said to have foaled away and when returned to the property 24 days later both mother and foal had the same clinical signs. *S. zooepidemicus* was not isolated from the mother mare and no details or samples were obtained of potential other horses on the property with respiratory disease. The rest of the samples were from horses with a cough or nasal discharge or both (**Table 2.8**).

Table 2.8: *Szp* types of respiratory disease samples with multiple isolates and not linked to an outbreak property

Sample ID	Date	Region	# of isolates	# of szp types	*szp* type							
					2	3	28	34	36	37	41	48
WG15089	12/08/2015	Masterton	5	1						5		
WK14049	8/12/2014	Waikato	2	1					2			
WK15143	1/12/2015	Waikato	5	2		4		1				
WK15144	2/12/2015	Waikato	5	1			5					
WK16002	15/01/2016	Waikato	5	1							5	
WK16140	16/09/2016	Waikato	2	2	1							1

2.3.3.2 *Szp* typing – All Samples

As mentioned above, all 286 isolates from the 108 horses were *szp* typed resulting in 48 *szp* strains. No single strain was solely associated with either respiratory or uterine infections, rather multiple different *szp* types were found in each group (**Figure 2.3**). Some of the *szp* strains were unique to sample groups whilst others were found across the sample groups. Of the 48 *szp* strains identified, ten were represented in both the uterus and the respiratory groups. The uterus group had 28 *szp* types, 18 of which were unique to the group. The respiratory disease group had 21 *szp* types, ten of which were unique. Eight of the ten respiratory apparently healthy *szp* types were unique to this group with the two remainder represented in the uterus group. The three respiratory abscess *szp* strains were all

represented in the respiratory disease group with one also found in the uterus group (*szp*-24). (**Figure 2.3**).

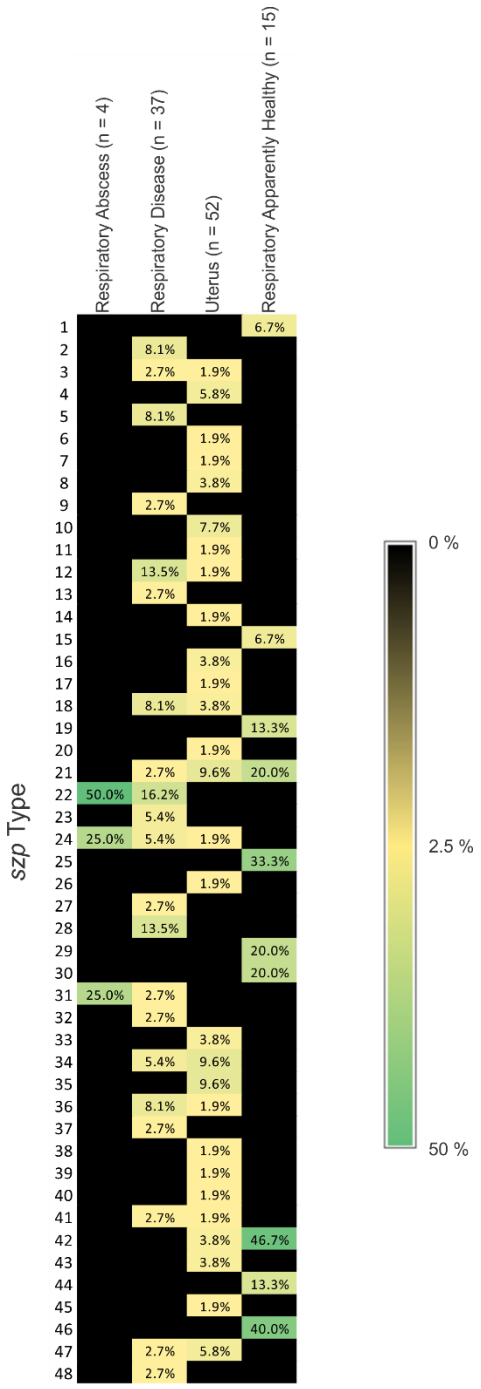


Figure 2.3: Heat map of the proportion of *szp* strains found in horses (n) for each sample group.

The NCBI alignment search tool BLAST was used to compare the 48 *szp* nucleotide sequences with those in the nucleotide collection database to give a global comparison (**Table 2.9**). This showed cross over between sample types (respiratory disease, respiratory apparently healthy, respiratory abscess, incision abscess, uterine associated). There was even some crossover between host species with canine and human *szp* sequences matching New Zealand horse isolate *szp* types. For instance, *Szp*-2 respiratory disease isolates had 100 % sequence identity with H70 isolated from a nasal swab of a healthy horse in the United Kingdom in 2000 (Holden *et al.*, 2009). Uterus *szp*-6 matched a horse respiratory United States strain. Uterus isolates of *szp*-35 matched a United States horse abortion isolate (Walker & Runyan, 2003) and an isolate from horse tonsils. The uterus isolate *szp*-38 was identical to an isolate from a horse secondary respiratory infection (Velineni *et al.*, 2014a), an isolate from a human blood sample and was also identical to the *szpSe* gene of a *S. equi* isolate. Uterus *szp*-43 had 100 % sequence identity with a horse incision abscess (Walker & Runyan, 2003) and *szp*-48 isolated from a horse nasal discharge sample matched that of a 1987 isolate from human blood with a fatal case of glomerulonephritis. Two *szp* groups with representatives in the nucleotide database remained within their original New Zealand sample type grouping. The uterus isolates *szp*-4 were a complete match to an abortion isolate (NCBI accession: AF519479) (Walker & Runyan, 2003) and *szp*-28 respiratory disease isolates were identical to a pleuropneumonia United States case (Walker & Runyan, 2003) (**Table 2.9**).

.

Table 2.9: New Zealand *szp* strain types compared to isolates in the NCBI database

NZ *szp* type	NZ sample group	NCBI 100 % *szp* sequence identity match				
		Accession	**Host**	**Source**	**Country**	**Reference**
2	respiratory disease	FM204884	horse	healthy (sample ID H70)	UK	Holden et al., 2009
4	uterus	AF519479	horse	abortion		Walker & Runyan, 2003
6	uterus	KF859910	horse	respiratory	US	
12	respiratory abscess respiratory disease uterus	AF519482	horse	cellulitis	US	Walker & Runyan, 2003
18	respiratory disease uterus	AF519484	horse	pleuropneumonia	US	Walker & Runyan, 2003
21	respiratory disease respiratory healthy uterus	AF519477 KF032042	horse horse	pleuropneumonia secondary S. zooepidemicus bronchopneumonia infection after experimental infection with equine herpes virus	US US	Walker & Runyan, 2003 Velineni *et al.*, 2014a
24	respiratory abscess respiratory disease	KF768754	dog	pneumonia	US	
28	respiratory disease	AF519470	horse	pleuropneumonia	US	Walker & Runyan, 2003
35	uterus	AF519481 KF803406	horse horse	abortion tonsils	US	Walker & Runyan, 2003
36	respiratory disease uterus	KF214275	dog	pneumonia	US	
38	uterus	LR134273 LS483380 KF032041	NA human horse	blood, host status not available blood, host status not avalible secondary S. zooepidemicus bronchopneumonia infection after experimental infection with equine herpes virus	UK UK	 Velineni *et al.*, 2014a
43	uterus	AF519480	horse	Incision abscess		Walker & Runyan, 2003
48	respiratory disease	LS483328	Human	blood, fatal glomerulonephritis		

2.3.3.3 MLST Typing

In contrast to *szp* typing, MLST typing uses amplification and sequencing of seven housekeeping genes providing more detailed typing. Therefore, MLST typing was used to further characterise 64 isolates. It is important to note these isolates were chosen as a subset of the 286 isolates and hence may lead to bias within the results and may not reflect the entire population. The 64 isolates were chosen to ensure the majority of the single polymorphic *szp* gene and all sample groups (respiratory disease, uterus, respiratory apparently healthy and respiratory abscess) were represented. From the 64 isolates 49 MLST strains (STs) were identified. Due to variation of one or more loci compared to the profiles in the database, 18 MLSTs (ST-193, 314, 330, 332, 336, 348, 360, 361, 362, 380, 381, 382, 383, 384, 385, 386, 387, 388) and six alleles (*nrdE* allele 28 and 60, *spi* allele 58 and 65, and *yqiL* allele 67 and 68) were identified as unique to these New Zealand isolates. One uterus isolate had a deletion of the

yqiL gene, WK15111_3 (**Table S2. 1**). WK15111_3 is a MLST SLV to a horse tracheal isolate from Finland (MLST ID 7723/09). As with the *szp* typing, some isolates from each sample group were found in the same STs (**Figure 2.4**). Respiratory disease and uterus isolates were combined in ST-193, ST-175, ST-147, and ST-380. Respiratory disease and respiratory abscess isolates were combined in ST-388 and ST-71. ST-2 included isolates from respiratory abscesses, respiratory disease and uterine samples.

The *szp* and MLST typing methods demonstrated a similar ability in differentiating strains albeit using a smaller sampling dataset for MLST. *Szp* typing was able to further differentiate three MLST STs into multiple strains (**Figure 2.4**). ST-24, ST-348 and ST-383 were split into two different *szp* types, all from the respiratory apparently healthy group. On the other hand, MLST typing was able to further differentiate five *szp* strains into multiple STs and into their different sample groups (**Figure 2.4**). *Szp*-3, *szp*-36, *szp*-42, *szp*-47 were each split into two different ST types and *szp*-21 isolates divided into three ST types. As such, the combination of both typing methods results in better strain discrimination (**Figure 2.4**).

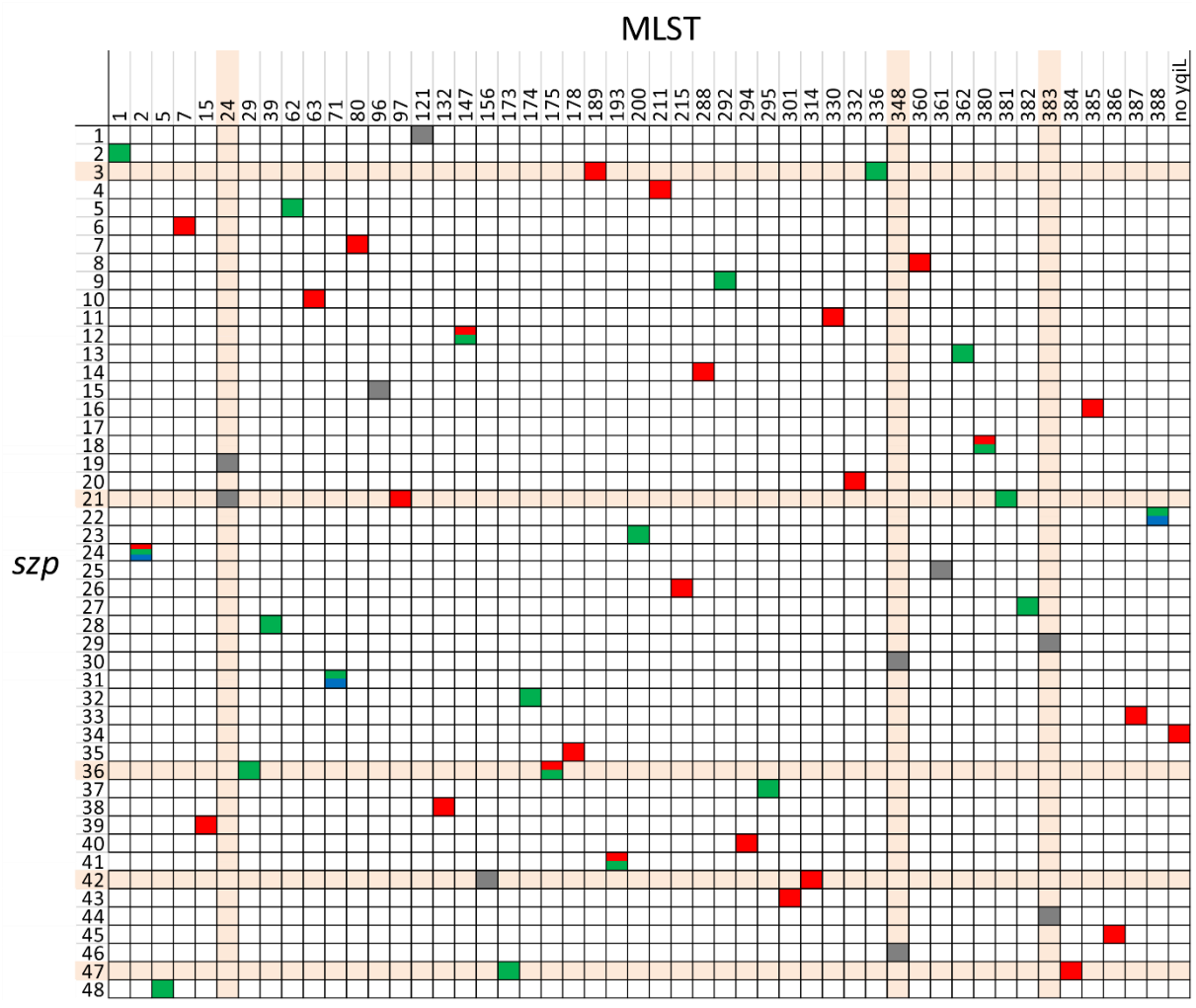


Figure 2.4: *Szp* and MLST typing comparison of New Zealand strains. Coloured by sample group: Red boxes = uterine isolates, green boxes = respiratory disease isolates, blue boxes = respiratory abscess and grey boxes = respiratory none apparent. The vertical highlighting shows MLST STs that were divided into multiple *szp* types and the horizontal highlighting shows *szp* types that were split into multiple MLST STs.

2.3.4 Population structure

The population structure of the MLST strains was investigated using the geoBurst algorithm in PHYLOVIZ software. STs with isolate data and all seven MLST genes were included in the analysis (**Table S2. 3**). Clonal complexes (CCs) were assigned to groups of single and double locus variants (SLV and DLV). The CCs were named after the predicted founder of the group from the global population analysis, and in the case of multiple founders, it was named from the ST with the most isolates (**Figure 2.5** and **Figure 2.6**). The population structure of only the

New Zealand strains in the database showed no formation of CCs that represented one of the isolation source groups (**Figure 2.5**). CC-96 and CC-144 have isolates from New Zealand respiratory infection and respiratory apparently healthy with horse and dog isolates represented (**Figure 2.5** and **Figure 2.6**). A mixture of New Zealand respiratory infection and abscess isolates made up CC-200 and CC-71 (**Figure 2.5**). Respiratory abscess isolates also group together with uterine isolates in CC-173 (**Figure 2.5**). CC-65 was a DLV between isolates from the uterus and no apparent disease samples (**Figure 2.5**).

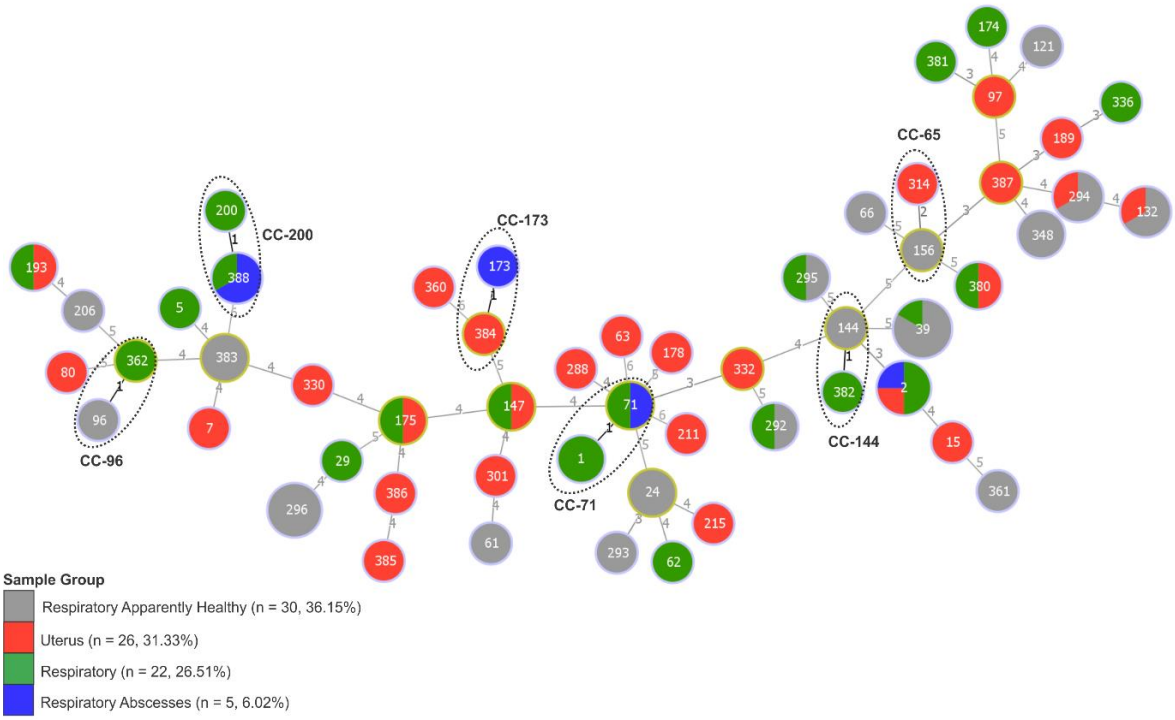


Figure 2.5: Minimum spanning tree determined by full goeBURST (global optimal eBurst; http://www.phyloviz.net) analysis of New Zealand *S. zooepidemicus* MLST allelic profiles. Circular nodes correspond to individual STs. Branch numbers indicate the number of allelic variations between STs. The branch lengths are not to scale. Clonal complexes (CCs) were defined as groups of STs with single and double locus variants (circled) and numbered based on the predicted founder in the global minimum spanning tree (**Figure 2.6**). Circle size is dependent on the number of isolates within the ST.

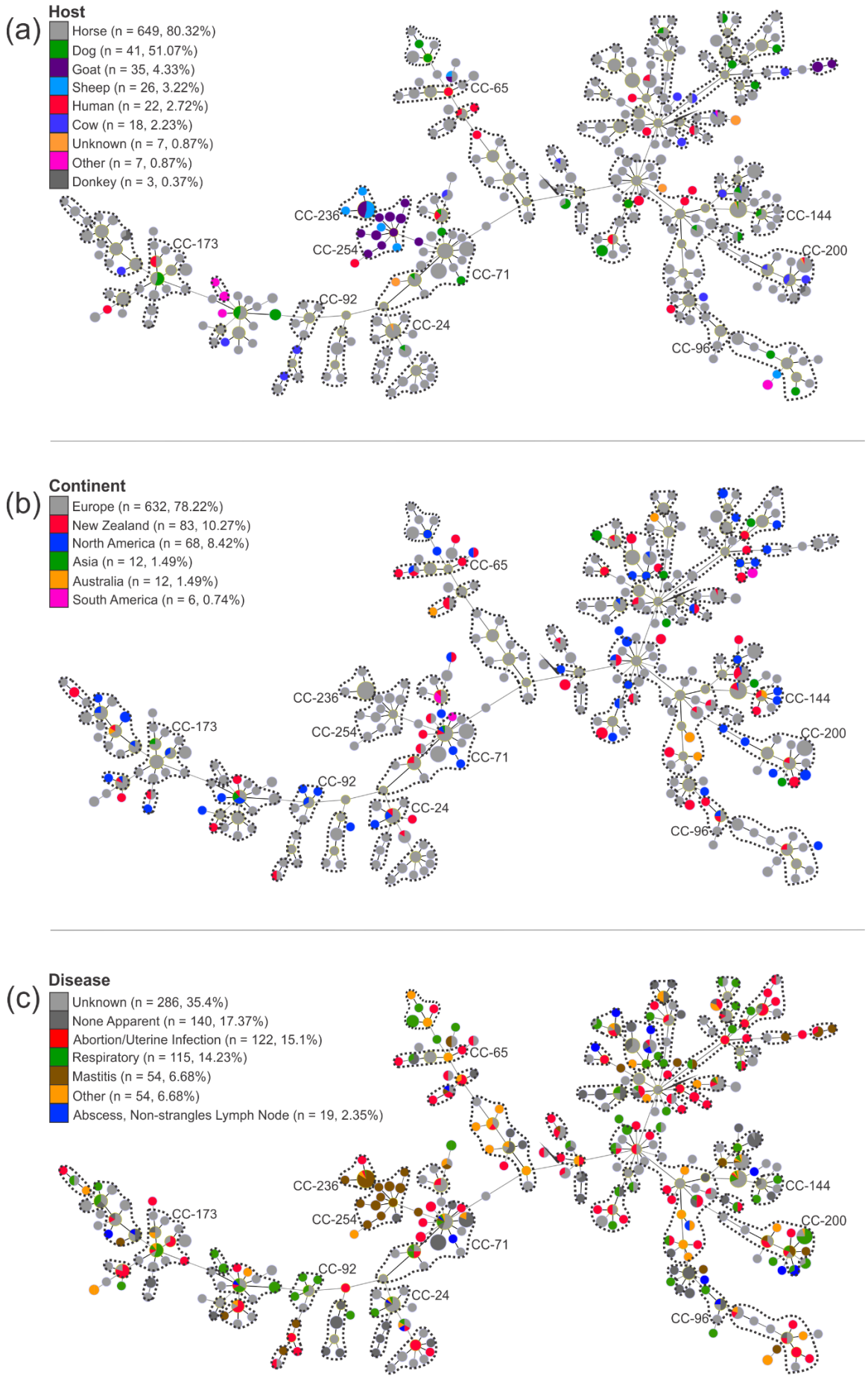

(a)
Host
Horse (n = 649, 80.32%)
Dog (n = 41, 51.07%)
Goat (n = 35, 4.33%)
Sheep (n = 26, 3.22%)
Human (n = 22, 2.72%)
Cow (n = 18, 2.23%)
Unknown (n = 7, 0.87%)
Other (n = 7, 0.87%)
Donkey (n = 3, 0.37%)
CC-65
CC-236
CC-254
CC-71
CC-144
CC-200
CC-173
CC-92
CC-24
CC-96
(b)
Continent
Europe (n = 632, 78.22%)
New Zealand (n = 83, 10.27%)
North America (n = 68, 8.42%)
Asia (n = 12, 1.49%)
Australia (n = 12, 1.49%)
South America (n = 6, 0.74%)
CC-65
CC-236
CC-254
CC-71
CC-144
CC-200
CC-173
CC-92
CC-24
CC-96
(c)
Disease
Unknown (n = 286, 35.4%)
None Apparent (n = 140, 17.37%)
Abortion/Uterine Infection (n = 122, 15.1%)
Respiratory (n = 115, 14.23%)
Mastitis (n = 54, 6.68%)
Other (n = 54, 6.68%)
Abscess, Non-strangles Lymph Node (n = 19, 2.35%)
CC-65
CC-236
CC-254
CC-71
CC-144
CC-200
CC-173
CC-92
CC-24
CC-96

Figure 2.6: Global minimum spanning tree determined by full goeBURST (global optimal eBurst; http://www.phyloviz.net) analysis of *S. zooepidemicus* MLST allelic profiles. Circular nodes correspond to individual STs. STs are colored with respect to (a) host species, (b) continent with Oceania separated into Australia and New Zealand and (c) disease status. Circle size is dependent on the number of isolates within the ST. The branch lengths are not to scale. Clonal complexes (CCs) were defined as groups of STs with single and double locus variants (circled).

When the New Zealand strains are compared globally, random scattering throughout the entire population of STs was observed (**Figure 2.6**) with no specific grouping of the New Zealand strains. This scattering was also apparent when looking at the global STs and resultant CCs in association with host species or with disease (**Figure 2.6**). The majority of the CCs could not be considered important due to bias in isolate submissions with 80 % of submissions from horse hosts (**Figure 2.6a**), 78 % from the continent of Europe (**Figure 2.6b**) and 35 % with unknown disease status (**Figure 2.6c**). However, the previously reported CC-236 and CC-254 did stand out as unique (**Figure 2.6**). They include sheep and goat mastitis isolates from Spain which acquired a novel sugar utilisation system (Steward *et al.*, 2017). CC-92 contains the ST-307 strain which matches the strain introduced via horses imported from New Zealand that caused an outbreak of respiratory disease in the New Caledonia horses (**Figure 2.6**) (Velineni *et al.*, 2014c). None of the strains in the current study were found within this clonal complex. Two of the apparently healthy respiratory isolates from the current study were ST-24 of CC-24 (**Figure 2.5**and **Figure 2.6**), the same type as the upper respiratory disease outbreak in Sweden horses (Lindahl *et al.*, 2013).

MLST ST-388 contains only the isolates with the *szp*-22 truncated protein and is within CC-200 (**Figure 2.6**). ST-388 is a SLV to ST-200 which consists of horse isolates from abortion/uterine infection, respiratory infection and unknown disease association, and cow mastitis isolates. The entire CC-200 contains the following: horse isolates from Europe, New Zealand, and North America associated with unknown disease status, abortion uterine infection, respiratory disease, non-strangles lymph node abscess and other abscess; cow isolates from Europe with mastitis and a human isolate from Europe with septicaemia (**Figure 2.6**).

To analyse this data in respect to genetic distances a maximum likelihood tree was constructed (**Figure 2.7**) based on the global MLST population. The New Zealand isolates

maintain the random scattering throughout the global population with no clustering of disease status or host species with the exception of previously defined sheep and goat mastitis complex, CC-326 and 254 (**Figure 2.7**).

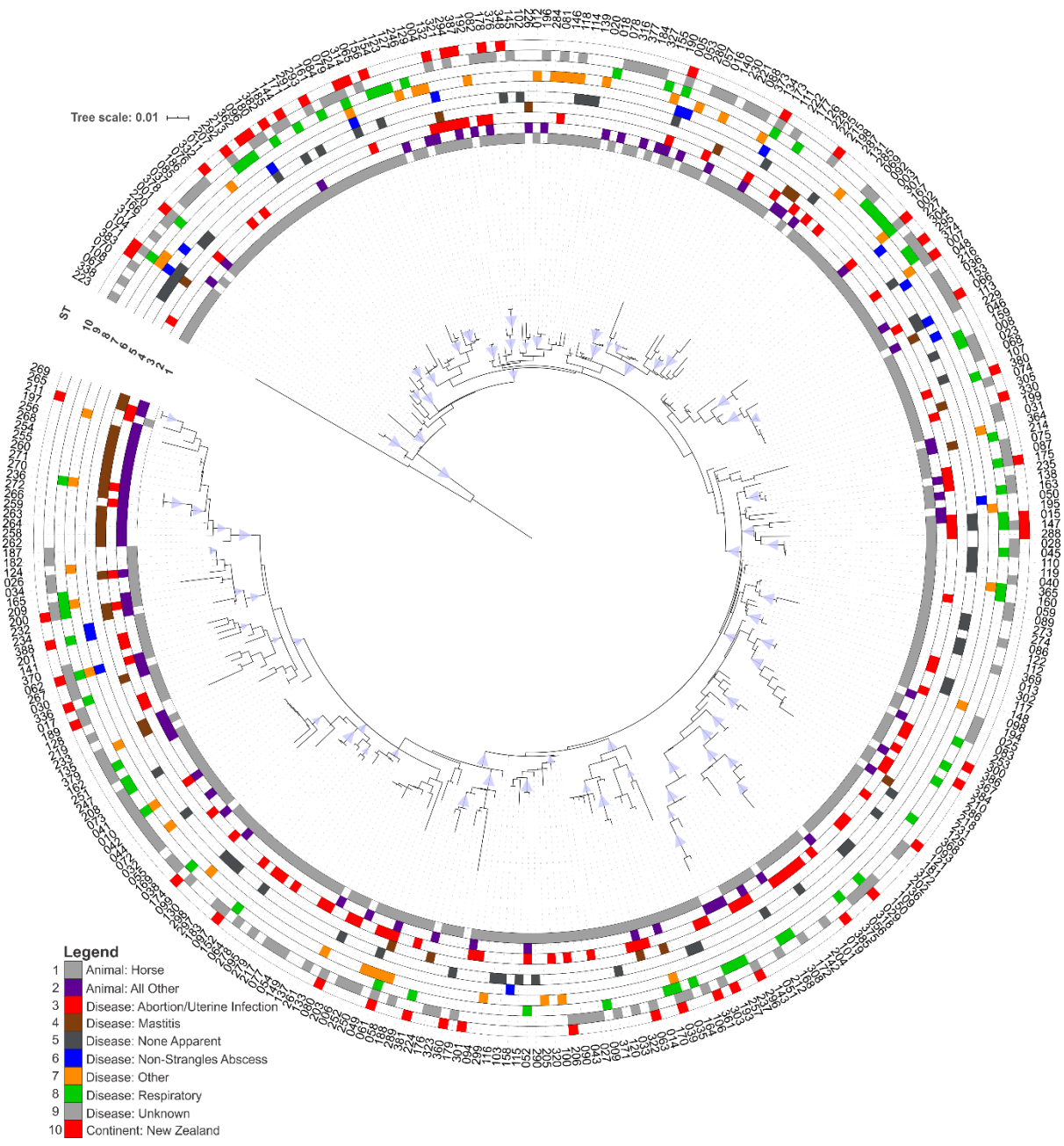
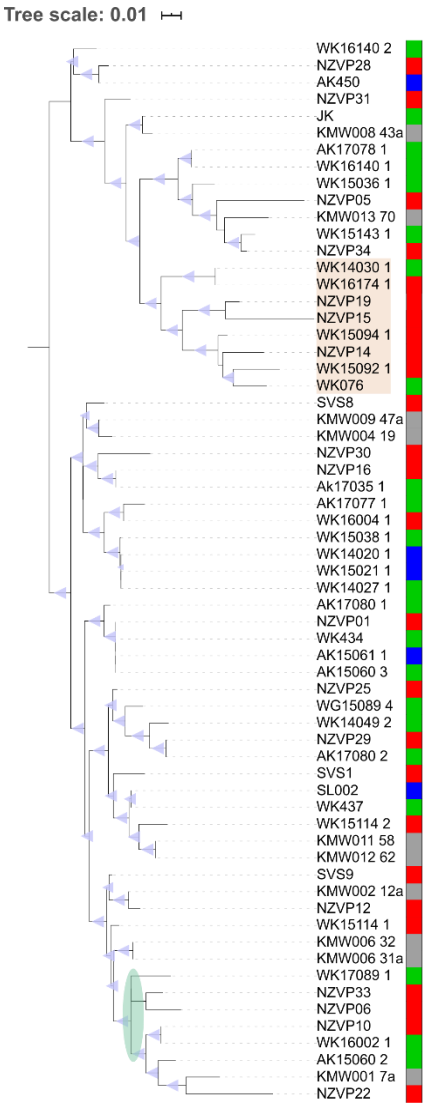
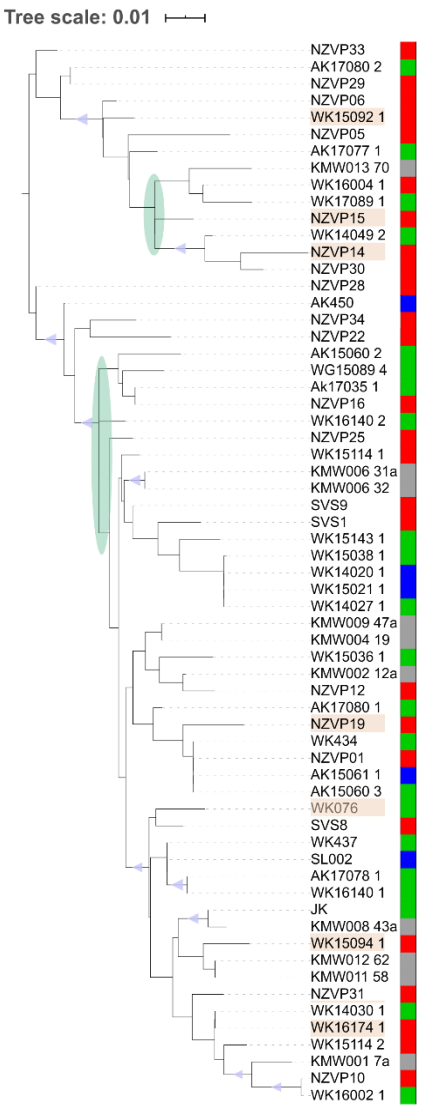


Figure 2.7: Maximum likelihood tree obtained with the concatenated nucleotide sequences of all global *S. zooepidemicus* MLST STs with complete profiles using PhyML (Guindon et al., 2010) and annotated in iTOL (Letunic & Bork, 2019). Midpoint rooted with branch aLRT SH-like support values in the range of 0.99 - 1 indicated by pale blue triangles. Rings are defined in the legend and are based on animal host, disease association and those from New Zealand. MLST STs are identified on the outer ring.

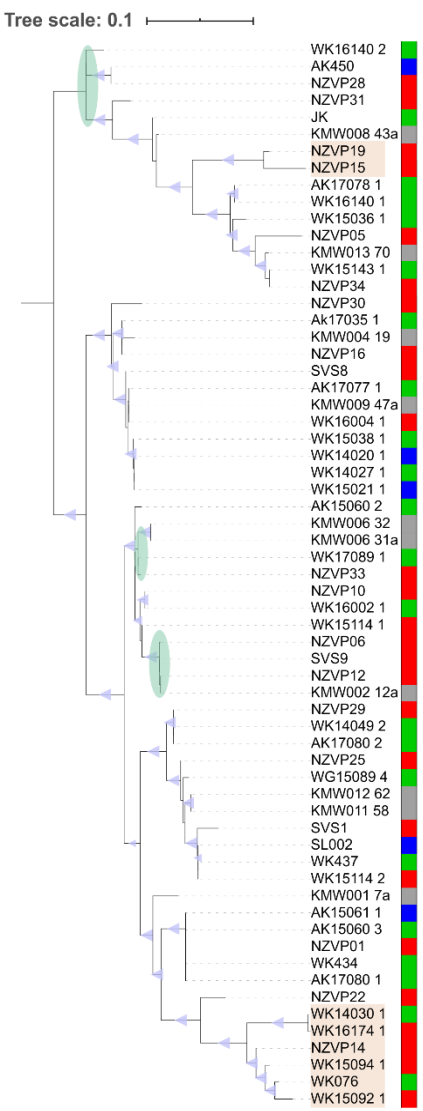
Maximum likelihood trees were then constructed to directly compare the typing methods of the New Zealand strains (**Figure 2.8**). Isolates included in this dataset had both *szp* and MLST typing data in order to create three different phylogenetic trees based on each method plus a tree combining the two methods. All three trees showed clustering of isolates from across the four groups of respiratory disease, respiratory abscess, respiratory apparently healthy and uterus. There are notable differences with the isolate clustering inferred by the different typing phylogenies. For instance, the combined MLST-*szp* tree showed a clustering of six uterus isolates with two respiratory disease isolates that was not maintained in either the MLST or *szp* tree. Both the combined MLST-*szp* and the *szp* trees had stronger branch support compared to the MLST tree. Increasing the number of genes used to create the three typing trees from one (*szp*) to seven (MLST) to the combined MLST-*szp* of eight genes resulted in fewer polytomies; unresolved branches where the internal node has more than two direct descendants (**Figure 2.8**).



MLST-*szp*



MLST



szp

Figure 2.8: Maximum likelihood trees of *S. zooepidemicus* with both *szp* and MLST data using PhyML (Guindon et al., 2010) and annotated in iTOL (Letunic & Bork, 2019). The MLST-*szp* tree was created with the concatenated MLST nucleotide sequences with the *szp* sequence. The MLST and *szp* trees are shown for comparison. Trees are midpoint rooted with aLRT SH-like branch support values in the range of 0.99 - 1 indicated by pale blue triangles. The unresolved polytomies are shaded in green ovals. A set of example isolates are highlighted in orange to show the change of clustering across the trees.

2.4 Discussion

2.4.1 Retrospective Study

The retrospective equine sample data from NZVP reveals *S. zooepidemicus* as a highly-prevalent bacterium in the New Zealand horse population, being the second most frequently isolated bacterium particularly associated with respiratory and urogenital samples. This data is highly suggestive of *S. zooepidemicus*' role as a causative agent of both respiratory and uterine infections that is echoed in other studies (Lavoie *et al.*, 1991; Clark *et al.*, 2008; Panchaud *et al.*, 2010; van Spijk *et al.*, 2016). Another New Zealand retrospective study reported *S. zooepidemicus* as the most frequently isolated bacterium in respiratory samples submitted to professional laboratories for diagnostic purposes (Toombs-Ruane *et al.*, 2015) which further supports the findings in this study. Globally, *S. zooepidemicus* is one of the most commonly isolated bacterial species from horses with upper respiratory disease (Laus *et al.*, 2007; Jannatabadi *et al.*, 2008; Ryu *et al.*, 2011; Mir *et al.*, 2013a) and lower respiratory disease (Christley *et al.*, 2001; Newton *et al.*, 2003; Wood *et al.*, 2005b, 2005a). It is also associated with uterine infection (Causey, 2006) making up the majority of the isolated β-haemolytic streptococci from the uterus of mares with fertility problems and endometritis (68 % to 92 %) (Albihn *et al.*, 2003; Casagrande Proietti *et al.*, 2011). Furthermore, the NZVP retrospective data also shows that *S. zooepidemicus* is isolated, although not as frequently, from a variety of other anatomical sites showing a diversity in its ability to infect different anatomical regions which is reflected in other studies (Erol *et al.*, 2012; van Spijk *et al.*, 2016).

2.4.2 *Streptococcus zooepidemicus* strain typing

Analysis of the University of Waikato isolate collection by *szp* and MLST typing methods were able to show the high genetic diversity of *S. zooepidemicus* within the New Zealand horse population (48 szp strains, 49 MLST STs). Both typing methods identified strains that were not restricted anatomically or by host species, and moreover, strains that were associated with both healthy and disease status. This crossover of strain types further highlights the diverse nature of these strains in thriving in different niches. The expansion of a single *szp* strain in both respiratory disease situations and uterus samples is consistent with other studies (Anzai *et al.*, 2000; Newton *et al.*, 2008; Rasmussen *et al.*, 2013).

The presence of multiple *szp* strains within healthy horses together with multiple *szp* strains clonally expanded in a respiratory outbreak, as observed with the four outbreaks reported here, is indicative of an opportunistic infection model. This observation is consistent with multiple Szp protein phenotypes isolated from the tonsils of healthy horses but single Szp phenotypes dominant in pneumonia cases of foals and donkeys (Anzai *et al.*, 2000). This is consistent with the current understanding of horse respiratory-associated *S. zooepidemicus* infection, although, a single pathogenic strain introduced into a naïve population can also be the causal agent of a respiratory outbreak (Björnsdóttir *et al.*, 2012; Lindahl *et al.*, 2013; Velineni *et al.*, 2014c). Once herd immunity is acquired to that strain it can then still persist in a healthy horse population, for example ST-24 was recovered post resolution of the respiratory disease outbreak in Sweden (Lindahl *et al.*, 2013). This suggests that the multiple strains of *S. zooepidemicus* found in a healthy individual will contain potentially pathogenic strains reflecting a history of infection (Waller, 2014). The pathogenicity of any individual strain is also influenced by host immune factors and physiological stresses together with virulence attributes of the bacterium. It is widely accepted that the host-microbe relationship is dynamic and includes many examples of asymptomatic carriage of potential pathogens (group A *S. pyogenes* in humans (Othman *et al.*, 2019), group B *Streptococcus* (Armistead *et al.*, 2019), and *S. pneumoniae (Pinto et al., 2019)*) in various host populations. Carriage/colonisation always precedes infection and in most situations involves the clonal expansion of a single pathogenic strain as observed in this research. Survival and carriage of a particular strain of *S. zooepidemicus* within a host can involve intracellular invasion and

persistence in uterine cells that protect the bacterium against the host immune responses (Skive *et al.*, 2017).

Szp and MLST typing methods were chosen in the study for the ability to compare New Zealand strains globally. Both methods were useful in tracking an outbreak but lacked the discriminative power to differentiate between colonising and pathogenic strains of *S. zooepidemicus*. The combination of typing methods improved strain differentiation. MLST uses seven housekeeping genes essential for basic cellular functions of a microbe and are also part of the core genome. In contrast, the single polymorphic *szp* gene is considered a virulence factor. It produces a cell surface-anchored protein which mediates the binding and attachment of *S. zooepidemicus* to host cells (Fan *et al.*, 2008). Szp is involved in the reduction of complement C3 deposition on the bacterial surface, masking it from the complement-mediated host defence, indicating Szp involvement in antiphagocytosis mechanisms and perhaps maintenance of the bacterium within a host (Ma *et al.*, 2012). Szp, however, is not the complete answer for successful colonisation and carriage of *S. zooepidemicus* as shown with *szp*-22. As *szp*-22 resulted in a premature stop codon before the surface-anchor sequence, a truncated gene product would be produced and theoretically would no longer play a role in phagocytosis resistance, ultimately, leading to a decrease in colonisation and virulence. Strain *szp*-22, isolated from nasal swabs and abscesses, was nevertheless able to successfully infect the horse. Its clonal expansion in these horses gives evidence to its increased fitness over other colonising strains. Thus, the production of Szp may not be advantageous in certain pathogenic niches such as the horse respiratory tract. The loss of this gene function may reflect genetic redundancy meaning it does not confer a selective advantage for this niche adaptation. Of note, none of the uterus isolates resulted in *szp* type 22.

2.4.3 Conclusion

This chapter demonstrates that *S. zooepidemicus* is highly-prevalent and genetically diverse in the New Zealand horse population with single strains clonally expanded during infection and multiple strains colonising healthy horses. However, there is a further need for *S. zooepidemicus* infections to be examined using more sensitive and discriminatory typing

methods to endeavour to distinguish between commensal and pathogenic isolates. Although MLST and *szp* typing are useful in outbreak and epidemiological investigations, they result in different phylogenies and fail to represent the real phylogeny with evidence of unresolved branches, a known issue when using only a few genes to infer phylogeny (Tsang *et al.*, 2017). With the advances in sequencing technology and the continually decreasing price, whole genome sequencing and analysis of core, accessory and pan genomes will become the new gold standard for microbial typing. The presence of accessory genes with selective niche fitness advantages could further help elucidate those genes resulting in disease-based strains of *S. zooepidemicus*. The genome of *S. zooepidemicus* is highly flexible with mobile genetic elements having played a major role in shaping the genome through chromosomal inversions, recombination, and HGT, resulting in a genetically diverse species (Beres *et al.*, 2008; Holden *et al.*, 2009; Ma *et al.*, 2013). There is strong evidence of genetic exchange between *S. equi*, *S. zooepidemicus* and other pathogens (Beres *et al.*, 2008; Holden *et al.*, 2009; Waller & Robinson, 2013). Genes involved in the uptake of sugar (phosphotransferase system) and the metabolism of carbohydrates have been fundamental in the evolution of *Streptococcus* species (Richards *et al.*, 2014). As such, the acquisition of a novel arbutin sugar utilisation system has been proposed as a fitness advantage in *S. zooepidemicus* strains infecting ruminants, in whose diet this sugar is found (Steward *et al.*, 2017). The acquisition of fitness factors (genes/systems) to utilize different energy sources or successfully evade the host immune system, will result in a competitive advantage over other strains with improved colonisation and persistence increasing survival of the strain. As such, one would expect to see phenotypic variation between strains with different abilities in host evasion mechanisms such as biofilm formation and intracellular invasion.

Chapter 3

Biofilm Formation and Intracellular Invasion Capabilities of New Zealand *Streptococcus zooepidemicus* Isolates *In Vitro*

3.1 Introduction

In order to resist the host's mucosal clearance, a bacterium must be able to adhere to the host's cells. Adherence is the first step for the establishment of bacterial colonisation and possible infection. Once adhered, the bacterium must then evade the host's immune defences and competition with the host's microbiota. Biofilm production and/or internalisation are two mechanisms bacteria can use to survive and persist in this hostile environment. For a commensal bacterium, biofilm formation and/or invasion will allow it to persist asymptomatically and become an opportunistic pathogen when favourable conditions arise, such as in immunocompromised individuals, animals under stress, or immunologically naïve populations. The ability to colonise and persist has been noted in some strains of *S. zooepidemicus* associated with equine respiratory disease (Anzai *et al.*, 2000; Anzai *et al.*, 2002; Björnsdóttir *et al.*, 2012; Lindahl *et al.*, 2013). Additionally, evidence has highlighted the ability of *S. zooepidemicus* to establish dormant uterine infections in mares by forming antibiotic resistant persister cells (Petersen *et al.*, 2015).

Streptococcal species recognised as biofilm producers include *S. dysgalactiae* subspecies *equisimilis* (Genteluci *et al.*, 2015; Ma *et al.*, 2017b), *S. pyogenes* (Roberts *et al.*, 2012),

S. agalactiae (Parker *et al.*, 2016) and *S. pneumoniae* (Sanchez *et al.*, 2011). Streptococcal biofilms are important in asymptomatic carriage and persistence of the bacterium (Roberts *et al.*, 2012) and play a role in antibiotic resistance and the severity and progression of disease (Young *et al.*, 2016). The promotion of carriage through biofilm formation will enhance transmission to immunological naïve populations. The closely-related species, *S. equi*, is known for long-term persistence and asymptomatic carriage promoting transmission of the disease known as strangles to susceptible horses (Waller et al., 2011). Streptococci are also recognised as intracellular pathogens with the ability to invade, survive and persist in host cells, enabling them to escape antibiotic treatment and the host immune system (Rohde & Cleary, 2016). Along with biofilm formation; *S. suis* (Benga *et al*., 2004), *S. pyogenes* (Podbielski et al., 2003), *S. agalactiae* (Tenenbaum *et al.*, 2005), *S. dysgalactiae* subsp. *dysgalactiae* (Alves-Barroco *et al.*, 2019), and *S. pneumoniae* (Novick *et al.*, 2017) also have the ability to invade host cells.

Evidence from several studies suggests that certain bacterial proteins involved in biofilm production and adherence to host cells are directly related to the virulence of *S. zooepidemicus*. Firstly, the *S. zooepidemicus* immunoreactive GroEL chaperone protein is involved in biofilm formation (Yi *et al.*, 2016). Immunisation with the GroEL recombinant protein provides protection against challenge with *S. zooepidemicus* in a mouse model of infection (Yi *et al.*, 2016). Secondly, the fibronectin-binding protein, FNZ, is important in both cell adhesion and biofilm formation (Ma *et al.*, 2013). A genomic deletion of FNZ had a significant reduction in its virulence in experimentally-infected mice (Ma *et al.*, 2013). Lastly, bacterial C5a peptidase plays a role in the adherence and invasion of *S. zooepidemicus* into host cells (Wei *et al.*, 2013). Like GroEL, C5a immunisation provides immunoprotection against *S. zooepidemicus* challenge (Wei *et al.*, 2013). These combined studies indicate that the processes of biofilm formation and adherence/invasion of host cells are involved in the pathogenesis of *S. zooepidemicus* as interference with genes involved in these processes affect the virulence of *S. zooepidemicus*.

The invasion of host cells is an important persistence strategy used by *S. zooepidemicus.* Evidence indicates that an intracellular phase may be one way *S. zooepidemicus* survives in the host and causes recurrent or persistent infections (Skive *et al.*, 2017). Endometritis strains

of *S. zooepidemicus* were shown to adhere, invade and survive within host cells. Invasion was associated with triggered cytoskeletal rearrangements (i.e., membrane ruffling), invagination of the host cell membrane, and engulfment of the adherent streptococcal chain (Skive *et al.*, 2017). Moreover, the *szp* typing results from Chapter 2 demonstrate that the isolation of a single strain-type of *S. zooepidemicus* from uterine samples regardless of health status may be evidence of dormant subclinical infections in the uterus of apparently healthy horses.

3.1.1 Aim

The aim of this chapter was to determine differences in biofilm and cell invasion capabilities of New Zealand *S. zooepidemicus* strains identified in the previous chapter (Chapter 2). Differences in phenotypic association were investigated from different disease status and anatomical sites, grouped as: respiratory disease, respiratory healthy, respiratory abscess, or uterus. In addition, the relationship between biofilm formation and cell invasion was investigated.

3.2 Methods

3.2.1 Biofilm Formation

3.2.1.1 Strains of *S. equi* subspecies Used to Assess Biofilm Formation

Fifty New Zealand *S. zooepidemicus* strains were characterised for biofilm formation, ensuring representation of a diverse range of *szp* types, MLST types and anatomical sites (respiratory healthy, respiratory disease, respiratory abscess, and uterus isolates) (**Table 3.1**). Strains from the closely-related *S. equi* bacterium were used for comparative purposes. These included, wild type New Zealand isolates ED99 (*sem*-99), G100 (*sem*-100) and K100 (*sem*-100) along with the PINNACLE® I.N. vaccine (Zoetis) (*sem*-2) (Patty & Cursons, 2014). Therefore, a total of 54 strains were used for the biofilm formation assays. The *S. zooepidemicus* and *S. equi* strains were grown overnight for 18 h at 37 °C with 5 % CO_2 on blood culture plates (Fort Richard, 1100). Each plate was imaged with a standard phone camera and the diameters of ten single colonies were measured using ImageJ software (v.1.51g).

Table 3.1: *S. zooepidemicus* and *S. equi* strains used in biofilm and adhesion/invasion assays

Isolate	Group	Species	*szp*	*sem*	MLST	Biofilm assay	Invasion assay
AK450	Respiratory abscess	*S. zooepidemicus*	47		173	y	Y
SL002	respiratory abscess	*S. zooepidemicus*	31		71	y	Y
WK14020_1	Respiratory abscess	*S. zooepidemicus*	22a		388	y	Y
AK15060_4	Respiratory disease	*S. zooepidemicus*	24a			y	Y
AK17076_1	Respiratory disease	*S. zooepidemicus*	18			y	
AK17078_1	Respiratory disease	*S. zooepidemicus*	2		1	y	
AK17080_1	Respiratory disease	*S. zooepidemicus*	27		382	y	
JK	Respiratory disease	*S. zooepidemicus*	13		362	y	Y
WG15089_4	Respiratory disease	*S. zooepidemicus*	37		295	y	Y
WK076	Respiratory disease	*S. zooepidemicus*	9		292	y	
WK14035_1	Respiratory disease	*S. zooepidemicus*	22b			y	
WK14036_1	Respiratory disease	*S. zooepidemicus*	12			y	
WK14049_2	Respiratory disease	*S. zooepidemicus*	36		29	y	
WK15036_1	Respiratory disease	*S. zooepidemicus*	5		62	y	
WK15037_5	Respiratory disease	*S. zooepidemicus*	28			y	
WK15038_1	Respiratory disease	*S. zooepidemicus*	23		200	y	
WK16002_1	Respiratory disease	*S. zooepidemicus*	41		193	y	
WK16140_2	Respiratory disease	*S. zooepidemicus*	48		5	y	
WK17089_1	Respiratory disease	*S. zooepidemicus*	32		174	y	
KMW001_7a	Respiratory healthy	*S. zooepidemicus*	25		361	y	Y
KMW002_12a	Respiratory healthy	*S. zooepidemicus*	42		156	y	Y
KMW003_15a	Respiratory healthy	*S. zooepidemicus*	21a			y	
KMW004_20a	Respiratory healthy	*S. zooepidemicus*	44			y	
KMW006_31a	Respiratory healthy	*S. zooepidemicus*	30		348	y	
KMW006_32	Respiratory healthy	*S. zooepidemicus*	46		348	y	Y
KMW008_43a	Respiratory healthy	*S. zooepidemicus*	15		96	y	
KMW009_50	Respiratory healthy	*S. zooepidemicus*	29			y	
KMW013_70	Respiratory healthy	*S. zooepidemicus*	1		121	y	
KMW015_77	Respiratory healthy	*S. zooepidemicus*	19			y	
NZVP05	Uterus	*S. zooepidemicus*	4		211	y	
NZVP06	Uterus	*S. zooepidemicus*	43		301	y	
NZVP1	Uterus	*S. zooepidemicus*	24b		2	y	
NZVP14	Uterus	*S. zooepidemicus*	10		63	y	
NZVP15	Uterus	*S. zooepidemicus*	7		80	y	
NZVP19	Uterus	*S. zooepidemicus*	6		7	y	
NZVP22	Uterus	*S. zooepidemicus*	26		215	y	
NZVP25	Uterus	*S. zooepidemicus*	38		132	y	
NZVP30	Uterus	*S. zooepidemicus*	16		385	y	
NZVP31	Uterus	*S. zooepidemicus*	14		288	y	
NZVP33	Uterus	*S. zooepidemicus*	45		386	y	
NZVP34	Uterus	*S. zooepidemicus*	3		189	y	
SVS1	Uterus	*S. zooepidemicus*	35		178	y	
SVS8	Uterus	*S. zooepidemicus*	20		332	y	
SVS9	Uterus	*S. zooepidemicus*	33		387	y	
WK15092_1	Uterus	*S. zooepidemicus*	8		360	y	
WK15094_1	Uterus	*S. zooepidemicus*	11		330	y	Y
WK15111_3	Uterus	*S. zooepidemicus*	34		no yqiL	y	Y
WK15114_1	Uterus	*S. zooepidemicus*	40		294	y	
WK15114_2	Uterus	*S. zooepidemicus*	39		15	y	Y
WK15125_1	Uterus	*S. zooepidemicus*	21b			y	
PINNACLE® I.N.	Vaccine 139192C	*S. equi*		2		y	Y
EDED090	Strangles	*S. equi*		99		y	Y
G22401	Strangles	*S. equi*		100(a)		y	Y
KAY07	Strangles	*S. equi*		100(b)		y	

3.2.1.2 Biofilm Formation Kinetics

Biofilm formation kinetics of each strain, measured over a 72 h time frame, were tested using a microtiter method adapted from Kwasny and Opperman (Kwasny & Opperman, 2010). Three millilitres of Todd Hewitt Broth (THB) (DIFCO, 249240) was inoculated with a single colony of each strain of *S. zooepidemicus* or *S. equi* and grown overnight for 18 h statically at 37 °C with 5 % CO_2. The overnight culture was diluted 1:100 in THB and 200 µL transferred into a well of a 96 well plate (surface-treated, flat bottom, JET BIOFIL TCP-010-096). Four replicates, using different single colonies, were set up for each strain. A separate 96 well plate was set up for each time point to be collected. Diluted cultures were grown in the 96 well plates for 2, 12, 24, 36, 48, and 72 h at 37 °C with 5 % CO_2. Negative controls on each plate included sterile THB media with no bacterial inoculate. At the indicated time points, culture media was aspirated and each well washed three times with 300 µL of sterile double distilled water to remove all non-adhered bacteria. This was done in a plate washer (Thermo Scientific Wellwash® Versa) with the following wash parameters: plate format, 3 mm aspirate height, low speed, default dispense height start:end of 6:14.4 mm, and used a final aspirate in normal mode. The biofilm was heat fixed by placing it in a 60 °C incubator for 1 h. The fixed biofilm was then stained with 200 µL of 0.05 % crystal violet (in deionised water) for 10 min followed by four washes with 400 µL of double distilled water with the plate washer using the same parameters as before. The crystal violet stain was eluted from the biofilm by addition of 200 µL of 35 % (*v/v*) acetic acid (Supelco®, 100063) for 10 min and transferred to a clean 96 well plate. The optical density (OD) at an absorbance of 590 nm of the eluted crystal violet was measured using a microplate reader (Thermo Scientific Multiskan GO).

Biofilm production was determined using an adapted method recommended for staphylococci (Stepanović *et al.*, 2007). The optical density cut-off value (OD_c) for biofilm production was set as three standard deviations above the average negative OD. Any value greater than the set OD_c indicates a positive biofilm production which were then classified into three levels: weak = $OD_c < OD \leq 2 \times OD_c$, moderate = $2 \times OD_c < OD \leq 4 \times OD_c$ and, strong = $OD > 4 \times OD_c$.

3.2.2 Intracellular Invasion

3.2.2.1 Host Cells and *S. equi* Subspecies Strains Used to Assess Intracellular Invasion

Equine dermal fibroblast (NBL6) (ATCC® CCL57™) (passage of P19-P21) and A548 human alveolar epithelial (ATCC® CCL185™) (passage of P30-P34) cell lines were used as host cells for the invasion assays. The host cells were revived from -80 °C freezer stocks into a 75 mL tissue culture flask in 15 mL complete media (CM). The CM (RPMI 1640 medium, GlutaMAX supplement, HEPES, 72400-047 Life Technologies; 0.5 µg/mL Amphotericin B, Gibco, 15290018 Life Technologies) contained 10 % heat-inactivated foetal bovine serum (F8067 Sigma) and was supplemented with 1 % PenStrep (Penicillin-Streptomycin, 10,000 U/mL, 15140122 Life Technologies). Once revived and confluency reached, cells were then seeded at 1-2x10^5 cells/mL into 25 mL tissue culture flasks (vent cap, surface-treated, JET BIOFIL TCF-012-025) and grown in 2 mL of CM to approximately 80 % confluency for the start of the invasion assay. All incubations took place at 37 °C in a 5 % CO_2 incubator.

Twelve *S. zooepidemicus* isolates were selected for the invasion assay (**Table 3.1**) representing a range of both strain types and biofilm forming capabilities and trajectories. The isolates included three respiratory healthy, three uterine, three respiratory abscess and three respiratory disease isolates. A further three strains of *S. equi* were used for a comparison (EDED090, *sem-99*; G22401, *sem*-100; and the PINNACLE® I.N. vaccine, *sem*-2) (**Table 3.1**). For the infection inoculum, a single colony of the isolate was grown overnight in Bacto™ Brain Heart Infusion (BHI) (BD, 237500) broth then diluted 1:50 in CM- (no antibiotics). As with the host cells, all incubations took place statically at 37 °C in a 5 % CO_2 incubator.

3.2.2.2 Invasion Assay Protocol

Optimisation of these assays are described in **File S3. 1**.

At approximately 80 % confluence, host cells were washed two times with 2 mL phosphate-buffered saline (PBS; 137 mM NaCl (Merck, 1.06404), 2.7 mM KCl (AJAX-Finechem (Univar), 383-500G), 10 mM Na_2HPO_4 (AJAX-Finechem (Univar), 621-500G), 1.8 mM KH_2PO_4 (BDH (AnalaR), 102034B)) to remove residual antibiotics. For each strain tested, 2 mL of the diluted infection inoculum was added to a flask with the 80 % confluent washed host cells.

The infection inoculum was also used to determine inoculum colony-forming units (CFU) as later described (**Table 3.2**). A further 2 mL of CM- was added to another flask with 80 % confluent washed host cells for a negative control of no bacteria. After two hours of bacteria to cell contact, the cells were washed twice with 2 mL of PBS to remove non-adhered bacteria. In order to kill any adhered non-internalised bacteria, the infected cells were then treated with 2 mL CM+. CM+ contained an antibiotic mixture of 200 µg/mL gentamicin (≥ 590 µg/mg potency, G3632 sigma, note a stock solution of 50 mg/mL was made up adjusted for the potency) and 100 µg/mL penicillin (~1650 U/mg, PENNA sigma). Two millilitres of CM- was added to the no bacteria control flask. After 1 h of treatment, cells were then washed twice with 2 mL PBS to remove any remaining antibiotics and FBS. One millilitre of the last wash from the CM+ treatment was collected for CFU counts as a control to ensure no extracellular bacterial survival (**Table 3.2**). Host cells and any internalised bacteria were dissociated from the flask with the addition of 500 µL trypsin-EDTA (0.25 %, phenol red, 25200056 Life technologies). At this stage, 20 µL of the dissociated cells from the negative control (no bacteria) flask were removed and put into a microfuge tube with 20 µL of CM- to neutralize the trypsin. Twenty microliters of the dissociated cells and CM- aliquot were added to 20 µL of 0.4 % trypan blue (Sigma-Aldrich, T0887) (in PBS) to count live cells per flask using a haemocytometer. To the remainder of dissociated cells, deionised water (750 µL) was added to each flask to lyse the host cells. The host cell lysate was collected and placed in a 1.5 mL microcentrifuge tube and centrifuged at 16,000 *g* for 2 min. The supernatant was removed and the pellet resuspended in 500 µL PBS. Cellular resuspensions were used to determine CFU of internalised bacteria (**Table 3.2**). The experiment was repeated four times for each strain and the results were expressed by cfu/mL of invasive bacterium, mathematically adjusted for a multiplicity of infection (MOI) of 5 in order to compare across strain and cell types.

To determine CFU, an adapted version of the single plate-serial dilution spotting technique was used (Thomas *et al.*, 2015). Briefly, 100 µL serial dilutions of 10^{-1} to 10^{-6} were prepared in PBS. For technical replicates, 5 µL of each dilution was pipetted in triplicate onto a BHI agar plate. Plates were then incubated overnight at 37 °C. Dilutions produced either zero colonies, countable colonies or colonies that were too numerous to count. The lowest dilution with a countable range of colonies was used to determine the average CFU. The CFU for each of the

following were determined with the noted dilutions spot plated: the infection inoculate (dilutions of 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-6}), the no bacteria control (undiluted), the CM+ treatment for internalised bacteria (undiluted and dilutions of 10^{-1}, 10^{-2}, and 10^{-4}) and the wash control for the CM+ treatment (undiluted) (**Table 3.2**).

Table 3.2: Dilutions used for single plate-serial dilution spotting for colony counts of the adhesion/invasion assay.

Step	Spot Dilutions	Purpose
Infection inoculate	10^{-2}, 10^{-3}, 10^{-4}, 10^{-6}	determine MOI
no bacteria control	undiluted	sterility control
CM+ treatment	undiluted, 10^{-1}, 10^{-2}, 10^{-4}	internalised bacteria
wash control of CM+	undiluted	ensure no survival of extracellular bacteria

3.2.2.3 Infection Assay Calculations

The following calculations were used to determine the cfu/mL of invaded bacteria.

Infection inoculum (cfu/flask):

$$cfu/flask = \left(\frac{average\ colonies}{dilution\ factor * volume\ plated}\right) * volume\ of\ infection\ innoculum$$

Where: average colonies is the average number of colonies in three replicated spot dilutions from the infection inoculum
Where: dilution factor equals the smallest dilution with countable colonies used for count
Where: volume plated was 0.005 mL
Where: infection inoculum was 2 mL

Total mammalian cells (cells/flask):

$$cells/flask = cell\ density * volume\ of\ disassociation\ solution$$

Where: volume of disassociation solution = 0.500 mL of trypsin
Where: cell density was calculated as:

$$cell\ density = \frac{(cells/square * dilution\ factor)}{square\ volume} = cells/mL$$

Where: cells/square is the average number of cells in 10 squares of haemocytometer
Where: dilution factor was 4 and the volume of the haemocytometer square is 0.0001 mL

Inoculum MOI of 5 adjustment factor:

$$MOI\ of\ 5 = 5 * \left(\frac{infection\ inoculum}{total\ mammalian\ cells}\right)$$

Where: Infection inoculum (cfu/flask) divided by the total mammalian cells (cells/flask) is the actual bacterial CFU per mammalian cell
Note: MOI of 5 was chosen as it was determined to be the average across the experiments

Invaded bacteria (cfu/mL):

$$\frac{cfu}{mL} = \left(\frac{average\ colonies}{dilution\ factor * volume\ plated}\right) * inoculum\ MOI\ of\ 5\ adjustment\ factor$$

Where: average colonies equals the average number of colonies in three replicated spot dilutions from the CM+ treatment (invaded bacteria)
Where: dilution factor was the smallest dilution with countable colonies used for count
Where: volume plated was 0.005 mL

3.2.3 Statistical analysis

The Kruskal-Wallis test was used for pairwise multiple comparisons with the Bonferroni Correction post hoc test for the colony diameter data. This non-parametric test was chosen due to some of the group data being non-normal distribution based on the Shapiro-Wilk test of normality. For collected biofilm data, the association between groups (respiratory abscess, respiratory disease, respiratory healthy, and uterus) with biofilm levels and kinetics was assessed by the Fisher's exact test for RxC contingency tables by implementing the Fisher-Freeman-Halton exact test of independence. For the invasion data, dependent variables were checked for normal distribution using the Shapiro-Wilk test of normality with original data and log10 transformed data. Further, Levene's test of equality of error variances was used to check the homogeneity of the data. As such, the tests suggested that the data was non-normally distributed so non-parametric tests were chosen for analysis. The Kruskal-Wallis test was used for pairwise multiple comparisons with the Bonferroni Correction post hoc test and the Mann-Whitney U test for comparison of two independent samples. A significance level of 0.05 was used for all tests. All statistical analyses were carried

out with SPSS (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.).

3.3 Results

3.3.1 *Streptococcus equi* Subspecies Colony Morphology

The diameter of ten single colonies was measured for each isolate. The JK *szp*-13 isolate of the *S. zooepidemicus* strains is phenotypically different from the other isolates with a highly mucoid colony morphology (**Figure 3.1**). It is significantly larger than the other *S. zooepidemicus* strains ($p < 0.001$) and was more similar in size to the *S. equi* wild type strains (**Figure 3.2**). The *S. equi* wild type strains produced colonies that were significantly larger than the non-JK *S. zooepidemicus* strains ($p < 0.001$). Conversely, the *S. equi* vaccine isolate was more similar to the *S. zooepidemicus* strains and significantly different from the JK and *S. equi* wild type strains ($p < 0.001$). No difference in colony size was observed between each isolation group of *S. zooepidemicus* (respiratory abscess, respiratory disease, respiratory healthy and uterus) (**Figure 3.2**).

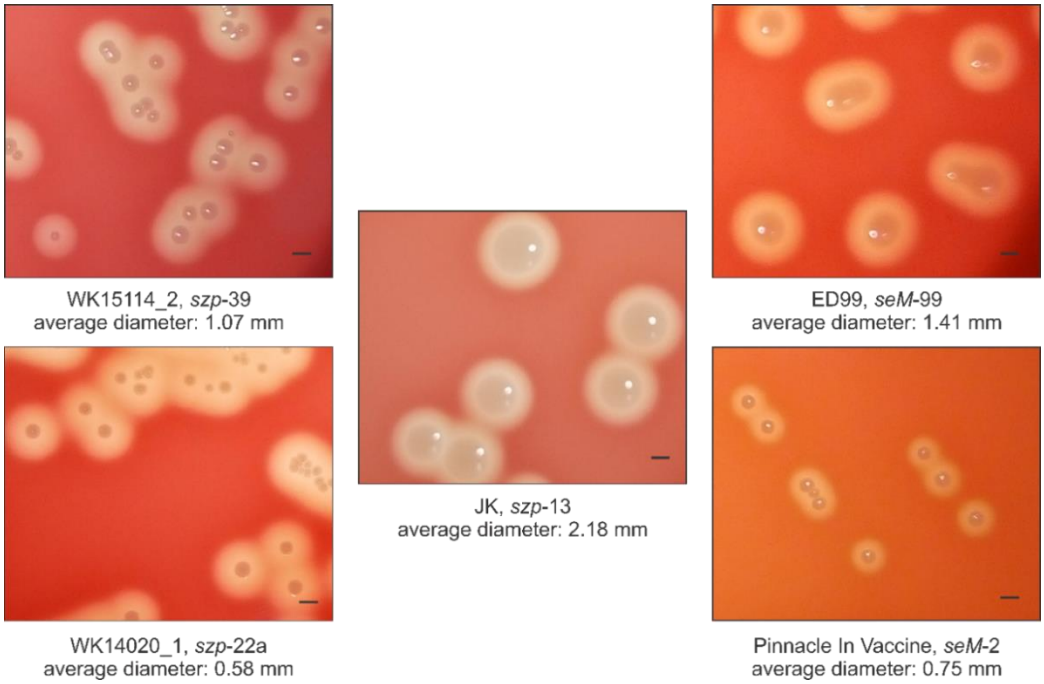


Figure 3.1: Representation of *S. zooepidemicus* and *S. equi* colony size morphology from overnight blood agar cultures grown at 37 °C with 5 % CO_2. Scale bar = 1 mm

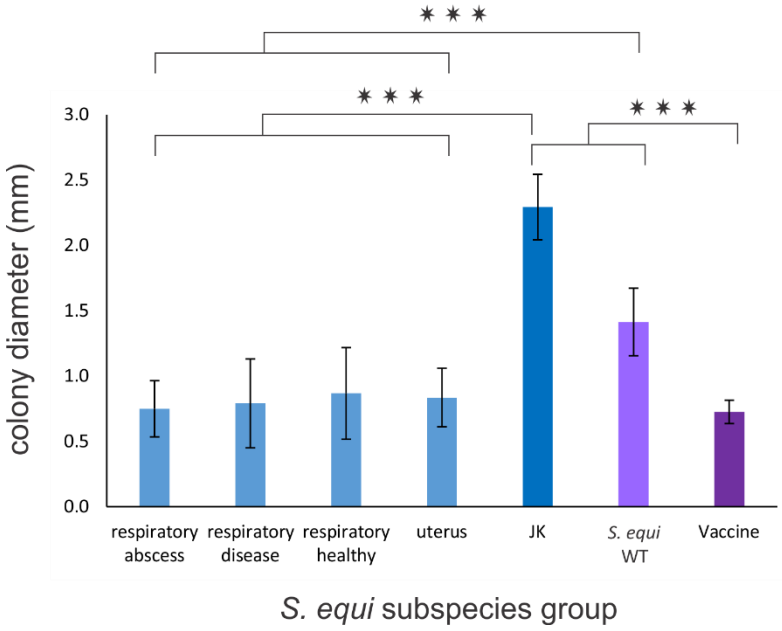


Figure 3.2: Colony diameter of the *S. equi* subspecies. Data is represented as the median colony diameter of the strains within the group, with ten colony measurements made per strain. Number of strains per group are as follow: three respiratory abscess, 15 respiratory disease, ten respiratory healthy, one JK, 21 uterus, three *S. equi* wild type (WT) and one vaccine. Error bars are ± the standard deviation. Blue bars are *S. equi* subsp. *zooepidemicus* groups and purple bars are *S. equi* subsp. *equi* groups. ***p < 0.001.

3.3.2 Biofilm Formation of the *S. equi* Subspecies

The OD_c was determined to be 0.12 at A_{590}, an OD at or below this value was considered negative for biofilm production. As such a biofilm was considered weak with an OD > 0.12 and ≤ 0.24, moderate with an OD > 0.24 and ≤ 0.48 and strong with an OD > 4.8. With the determination of these values all but three of the 50 strains of *S. zooepidemicus* tested were able to form a biofilm during some stage of the 72 h trials (**Table 3.3**). The maximum biofilm production of each strain was distributed in one of the four levels of biofilm production (**Table 3.3**). Of the 16 strains tested in the respiratory disease group, six formed strong biofilms (*szp*-9, 12, 22b, 23, 24a, 27; 37.5 %), four formed moderate biofilms (*szp*-2, 28, 36, and 37; 25.0 %), four formed weak biofilms (*szp*-5, 13, 32, and 41; 25.0%) and a further two were unable to form a biofilm (*szp*-18 and 48; 12.5 %) (**Table 3.3**). All three strains of the respiratory abscess group were able to form biofilms. *Szp*-31 (33.3 %) had produced a strong biofilm, and both *szp*-22a and *szp*-47 (66.7 %) produced a moderate biofilm (**Table 3.3**). In the respiratory healthy group of 10 strains, four (*szp*-1, 30, 44, and 46, 40.0 %) formed strong biofilms, four formed biofilms to a moderate level (*szp*-19, 21a, 25, and 29; 40.0%), and two were weak biofilm producers (*szp*-42 and 46; 20.0 %) (**Table 3.3**). Within the 21 uterus strains, five were strong biofilm producers (*szp*-4, 14, 24b, 34, and 39; 23.8 %) and 11 moderate producers (*szp*-6, 7, 8, 10, 11, 16, 21b, 35, 38, 40 and 45; 52.4 %), four were weak (*szp*-3, 20, 33, and 43; 19.0 %) and there was one unable to produce a biofilm (*szp*-26, 4.8 %) (**Table 3.3**). In comparison, *S. equi* isolates were all able to form biofilms all with their highest biofilm levels formed at 72 h. The three wild type *S. equi* isolates (*sem*-99 and *seM*-100a and b) were only able to produce a weak biofilm (**Table 3.3**). Interestingly, the assumedly less virulent and capsule deficient vaccine isolate, PINNACLE® I.N. (*sem*-2) was able to form a strong biofilm (**Table 3.3**). There was no association between the group the *S. zooepidemicus* strain belonged to (respiratory abscess, respiratory disease, respiratory healthy, and uterus) and the level of biofilm produced ($p = 0.823$). Further there was no association between the species of streptococci (*S. zooepidemicus* vs *S. equi*) and the level of biofilm produced ($p = 0.072$) (**Table 3.4**).

Table 3.3: Level of biofilm production of *S. zooepidemicus* and *S. equi* across 72 h. Coloured from pale to dark purple for increasing biofilm production based on the ODc cut-off value; grey no biofilm production (OD ≤ 0.12), pale purple for weak production (0.12 > OD ≥ 0.24), light purple for moderate production (0.24 > ODc ≥ 0.48), and dark purple for strong production (OD > 0.48). Values represent the mean of four replicates. The highest level of biofilm reach during the 72 h is indicated as one of the four levels of biofilm production.

			Average OD 590 at time point							
Strain	**szp**	**Group**	2 h	12 h	24 h	36 h	48 h	72 h	**Highest level**	
AK450	47	respiratory abscess	0.099	0.203	0.166	0.277	0.255	0.248	++	moderate
SL002 (MK)	31	respiratory abscess	0.120	0.568	0.620	0.673	0.665	0.553	+++	strong
WK14020_1	22a	respiratory abscess	0.097	0.264	0.474	0.462	0.489	0.483	++	moderate
AK15060_4	24a	respiratory disease	0.137	0.570	0.641	0.587	0.609	0.586	+++	strong
AK17076_1	18	respiratory disease	0.099	0.096	0.086	0.098	0.095	0.094	-	none
AK17078_1	2	respiratory disease	0.103	0.332	0.297	0.294	0.296	0.286	++	moderate
AK17080_1	27	respiratory disease	0.099	0.443	0.537	0.476	0.469	0.405	+++	strong
JK	13	respiratory disease	0.165	0.090	0.091	0.088	0.090	0.089	+	weak/none
WG15089_4	37	respiratory disease	0.102	0.114	0.230	0.262	0.248	0.258	++	moderate
WK076 (WD)	9	respiratory disease	0.106	0.338	0.596	0.549	0.499	0.483	+++	strong
WK14035_1	22b	respiratory disease	0.110	0.360	0.673	0.575	0.682	0.588	+++	strong
WK14036_1	12	respiratory disease	0.135	0.556	0.578	0.610	0.761	0.595	+++	strong
WK14049_2	36	respiratory disease	0.108	0.146	0.292	0.322	0.362	0.399	++	moderate
WK15036_1	5	respiratory disease	0.103	0.131	0.195	0.193	0.215	0.223	+	weak
WK15037_5	28	respiratory disease	0.106	0.276	0.322	0.317	0.290	0.313	++	moderate
WK15038_1	23	respiratory disease	0.111	0.534	0.728	0.856	0.734	0.791	+++	strong
WK16002_1	41	respiratory disease	0.097	0.167	0.168	0.154	0.171	0.172	+	weak
WK16140_2	48	respiratory disease	0.100	0.108	0.101	0.101	0.111	0.097	-	none
WK17089_1	32	respiratory disease	0.101	0.108	0.103	0.111	0.127	0.097	+	weak/none
KMW001_7a	25	respiratory healthy	0.145	0.218	0.253	0.251	0.258	0.248	++	moderate
KMW002_12a	42	respiratory healthy	0.131	0.091	0.084	0.083	0.085	0.089	+	weak/none
KMW003_15a	21a	respiratory healthy	0.124	0.113	0.283	0.223	0.202	0.212	++	moderate
KMW004_20a	44	respiratory healthy	0.162	0.445	0.495	0.363	0.390	0.334	+++	strong
KMW006_31a	30	respiratory healthy	0.148	0.491	0.493	0.461	0.436	0.448	+++	strong
KMW006_32	46	respiratory healthy	0.116	0.514	0.545	0.519	0.441	0.407	+++	strong
KMW008_43a	15	respiratory healthy	0.132	0.097	0.101	0.089	0.104	0.100	+	weak/none
KMW009_50	29	respiratory healthy	0.136	0.442	0.426	0.389	0.355	0.338	++	moderate
KMW013_70	1	respiratory healthy	0.129	0.747	0.610	0.504	0.381	0.408	+++	strong
KMW015_77	19	respiratory healthy	0.100	0.177	0.288	0.275	0.245	0.233	++	moderate
NZVP05	4	uterus	0.102	1.216	0.992	0.987	0.858	0.789	+++	strong
NZVP06	43	uterus	0.097	0.164	0.145	0.167	0.160	0.144	+	weak
NZVP1	24b	uterus	0.124	0.626	0.637	0.596	0.629	0.584	+++	strong
NZVP14	10	uterus	0.099	0.317	0.258	0.297	0.238	0.284	++	moderate
NZVP15	7	uterus	0.097	0.168	0.400	0.414	0.334	0.310	++	moderate
NZVP19	6	uterus	0.094	0.223	0.244	0.378	0.320	0.476	++	moderate
NZVP22	26	uterus	0.099	0.112	0.099	0.105	0.105	0.111	-	none
NZVP25	38	uterus	0.100	0.310	0.244	0.223	0.264	0.224	++	moderate
NZVP30	16	uterus	0.098	0.162	0.190	0.192	0.303	0.227	++	moderate
NZVP31	14	uterus	0.100	0.333	0.546	0.393	0.371	0.306	+++	strong
NZVP33	45	uterus	0.098	0.137	0.271	0.281	0.330	0.281	++	moderate
NZVP34	3	uterus	0.096	0.133	0.234	0.189	0.181	0.151	+	weak
SVS1	35	uterus	0.100	0.167	0.373	0.212	0.271	0.188	++	moderate
SVS8	20	uterus	0.099	0.097	0.193	0.161	0.233	0.174	+	weak
SVS9	33	uterus	0.096	0.193	0.165	0.159	0.176	0.148	+	weak
WK15092_1	8	uterus	0.105	0.209	0.258	0.198	0.193	0.165	++	moderate
WK15094_1	11	uterus	0.106	0.144	0.311	0.279	0.289	0.252	++	moderate
WK15111_3	34	uterus	0.110	0.599	0.569	0.482	0.395	0.366	+++	strong
WK15114_1	40	uterus	0.109	0.178	0.328	0.318	0.350	0.361	++	moderate
WK15114_2	39	uterus	0.105	0.357	0.501	0.515	0.385	0.439	+++	strong
WK15125_1	21b	uterus	0.106	0.324	0.440	0.469	0.459	0.454	++	moderate
ED99/OP19	*sem*-99	*S. equi*	0.103	0.096	0.129	0.181	0.194	0.237	+	weak
G100	*sem*-100a	*S. equi*	0.103	0.093	0.114	0.135	0.157	0.184	+	weak
K100	*sem*-100b	*S. equi*	0.106	0.100	0.114	0.118	0.128	0.140	+	weak
Vaccine(PINNACLE)	*sem*-2	*S. equi*	0.099	0.150	0.319	0.357	0.436	0.603	+++	strong

Table 3.4: Maximum biofilm production of *S. zooepidemicus* and *S. equi* isolates separated within isolation source groups being the *S. zooepidemicus* groups of respiratory abscess, respiratory disease, respiratory healthy and uterus and all the *S. equi* isolates, inclusive of the vaccine isolate, in a single group.

Group	Negative	Weak	Moderate	Strong	
	n (%)				
respiratory abscess (n=3)	0 (0.0)	0(0.0)	2 (66.7)	1 (33.3)	
respiratory disease (n=16)	2 (12.5)	4(25.0)	4 (25.0)	6 (37.5)	
respiratory healthy (n=10)	0 (0.0)	2 (20.0)	4 (40.0)	4 (40.0)	
uterus (n=21)	1 (4.8)	4 (19.0)	11 (52.4)	5 (23.8)	*
S. equi (n=4)	0 (0.0)	3 (75.0)	0 (0.0)	1 (25.0)	
Total (n=54)	3 (5.6)	19 (24.1)	21 (38.9)	17 (31.5)	

*p = 0.072.

The OD trajectories of the strains over time showed four distinct biofilm kinetics (**Figure 3.3**). The first kinetic group consisted of strains reaching one of their highest levels of biofilm productions at 12 h (**Figure 3.3 A**). From then they were able to sustain a biofilm for up to 72 h. Just over half of the *S. zooepidemicus* strains, 26 of 50 (52.0 %), were in this biofilm kinetic category with representatives from each group respiratory disease, respiratory abscess, respiratory healthy and uterus. The next kinetic group consisted of strains establishing one of their highest biofilm productions by 24 h and maintaining the biofilm up to 72 hours (**Figure 3.3 B**). Thirty-four percent (n = 17) of the strains fell into this category also having representatives from each group. The third biofilm kinetic group had biofilms starting after 24 h with an increase in biofilm production towards 72 h (**Figure 3.3 C)**. None of the *S. zooepidemicus* but all three wild type *S. equi* strains trialled fell into this kinetic group. The vaccine strain showed increasing biofilm production towards 72 h as well but started biofilm production earlier, before 12 h. The fourth biofilm kinetic group was the inability to produce or maintain a biofilm (**Figure 3.3 D**). Seven (14.0 %) of the *S. zooepidemicus* strains were in this kinetic group. These included strains that formed a very weak biofilm at only one of the time points and were unable to maintain the biofilm (*szp*-13, 15, 32 and 42). There was no association between the group which the *S. zooepidemicus* isolates belonged to and the biofilm kinetic category (p = 0.604) (**Table 3.5**). However, there was a significant difference between *S. equi* and *S. zooepidemicus* (p < 0.001); all the *S. equi* isolates followed a different kinetic biofilm trajectory than the *S. zooepidemicus* isolates (**Table 3.5**).

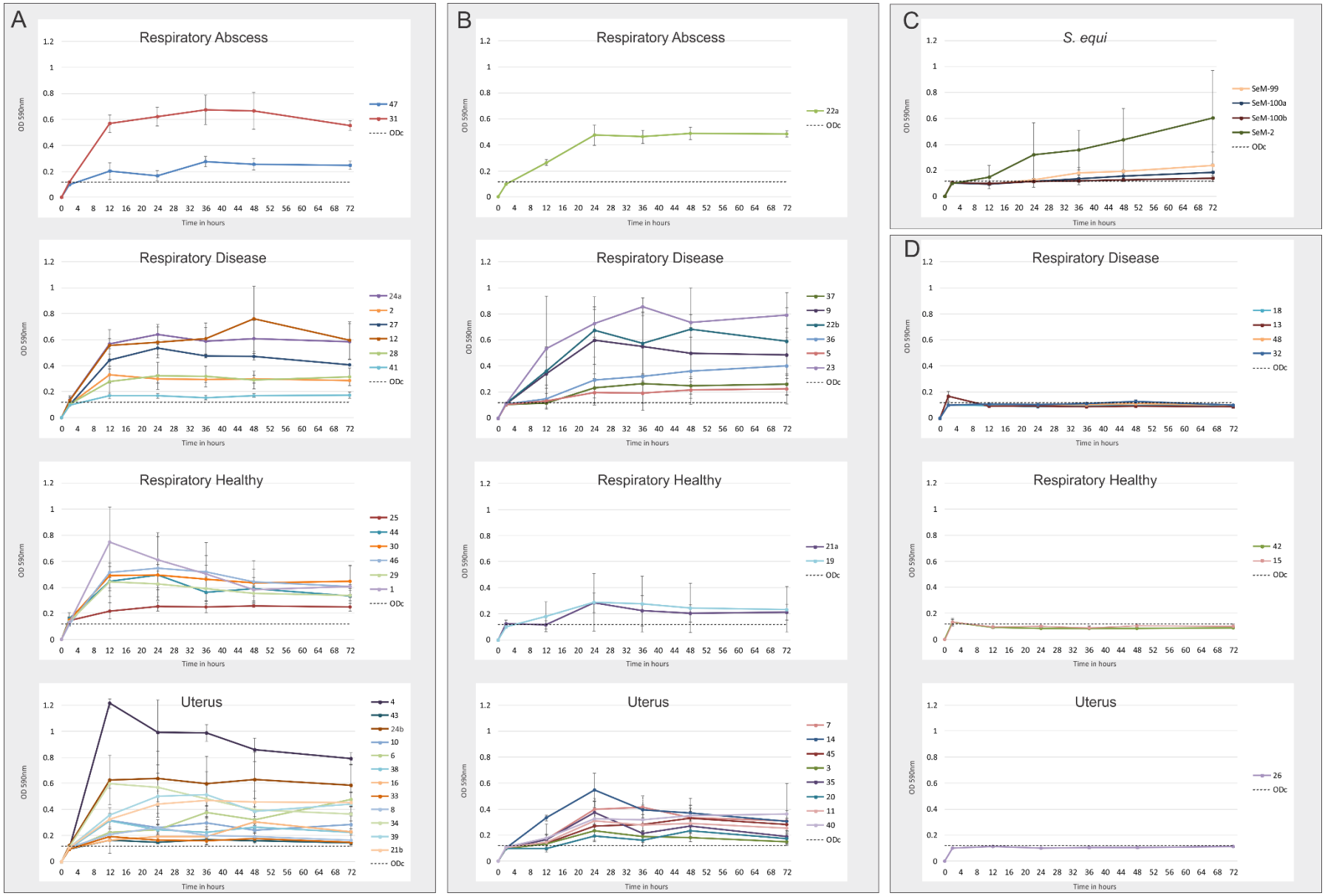

A
Respiratory Abscess
47
31
ODc
Respiratory Disease
24a
2
27
12
28
41
ODc
Respiratory Healthy
25
44
30
46
29
1
ODc
Uterus
4
43
24b
10
6
38
16
33
8
34
39
21b
ODc
B
Respiratory Abscess
22a
ODc
Respiratory Disease
37
9
22b
36
5
23
ODc
Respiratory Healthy
21a
19
ODc
Uterus
7
14
45
3
35
20
11
40
ODc
C
S. equi
SeM-99
SeM-100a
SeM-100b
SeM-2
ODc
D
Respiratory Disease
18
13
48
32
ODc
Respiratory Healthy
42
15
ODc
Uterus
26
ODc
OD 590nm
Time in hours

Figure 3.3: Biofilm kinetics of *szp* strain types. The 12 h (a) and 24 h (b) kinetic categories indicate strains with maximum production reached at that time then maintaining the biofilm afterwards. The 72 h (c) kinetic category indicates strains with production of biofilm increasing towards 72 h time point. (d) indicates no biofilm production or a weak biofilm produced at one time point only and not sustained. Horizontal dotted line represents the optical density cut-off value for biofilm production. Each *szp* strain type is denoted by its *szp* type number and corresponding coloured line. Error bars are the standard deviation of the mean of four replicates.

Table 3.5: Biofilm kinetics category of the *S. zooepidemicus* and *S. equi* isolates based on isolation source group. Negative indicates no biofilm production or a weak biofilm production at one time point only. The 12 and 24 h categories indicate biofilm maximum production reached at that time then maintain the biofilm. The 72 h categories indicates production of biofilm increasing towards 72 h time point. ***p < 0.001.

	kinetic hour				
	negative	**12 h**	**24 h**	**>72 h**	
Group		**n (%)**			
respiratory abscess (n=3)	0 (0.0)	2 (66.7)	1 (33.3)	0 (0.0)	
respiratory disease (n=16)	4 (25.0)	6 (37.5)	6 (37.5)	0 (0.0)	
respiratory healthy (n=10)	2 (20.0)	6 (60.0)	2 (20.0)	0 (0.0)	
uterus (n=21)	1 (4.8)	12 (57.1)	8 (38.1)	0 (0.0)	***
S. equi (n=4)	0 (0.0)	0 (0.0)	0 (0.0)	4 (100)	

3.3.3 Intracellular Invasion of *S. equi* Subspecies

Preliminary experiments conducted to optimise methodology for intracellular invasion assays all showed evidence of *S. zooepidemicus* and *S. equi* invading both human A549 epithelial and equine dermal fibroblast host cell lines (**File S3. 1**). Comparison of the two bacterial species revealed that *S. zooepidemicus* was able to internalise with greater efficiency than *S. equi* in the equine dermal fibroblasts ($p < 0.001$) and the A549 human lung epithelial cells ($p < 0.001$) (**Figure 3.4 A**). The cell lines used did not influence either bacterial species' ability to internalise ($p = 1$) (**Figure 3.4 A**). Due to their unusual colony phenotype both the data from the JK strain (*szp*-13, highly mucoid) and the *S. equi* vaccine strain (PINNACLE® I.N.) were removed from the rest of the wild type strains. With their data removal, the difference between the two streptococci species was maintained, still resulting in greater intracellular infection of *S. zooepidemicus* than the wildtype *S. equi* ($p < 0.001$) and the *S. equi* vaccine

strain (p = 0.001) (**Figure 3.4 B**). The mucoid *S. zooepidemicus* strain JK (*szp*-13) behaves like the wildtype *S. equi* (p = 1) with a significantly reduced internalisation (p < 0.001) when compared to the other *S. zooepidemicus* strains (**Figure 3.4**). *S. zooepidemicus* strain invasion of both host cells (range 4,758 - 2,231,625 cfu/mL, median 186,874 cfu/mL) exhibited a 469 % fold increase (log change 2.67) from the mucoid JK *szp*-13 strain (range 49 - 1,617 cfu/mL, median 398 cfu/mL) (p ≤ 0.001) (**Figure 3.4 B**). Whereas, *szp*-13 total invasion had a 38 % fold increase (log change 1.58) from the wildtype *S. equi* invasion capabilities (range 0 - 652 cfu/mL, median 11 cfu/mL) (p = 1) (**Figure 3.4 B**).

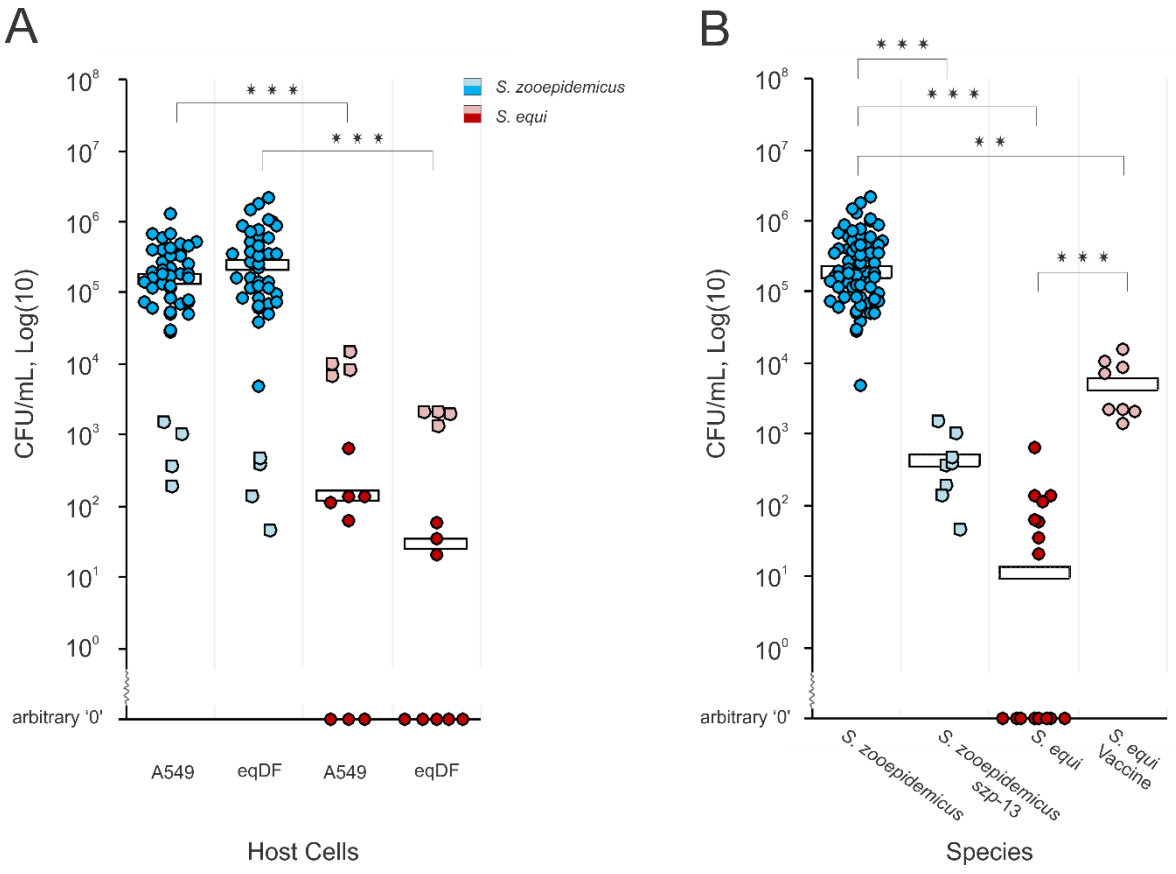


Figure 3.4: Invasion of *S. zooepidemicus* and *S. equi* in A549 and equine DF cells. A) Shows all *S. zooepidemicus* versus all *S. equi* isolates in either cell line. B) Shows comparison regardless of cell line with the unusual *S. equi* and *S. zooepidemicus* separated out (PINNACLE® I.N. *seM*-2 and JK *szp*-13 respectively). White bars represents the medium of the data points. **p = 0.001, ***p < 0.001.

Separating the *S. zooepidemicus* data based on the four disease status and isolation source groups showed no variation in invasion capabilities. Respiratory disease (minus the JK, *szp*-13, mucoid isolate), respiratory abscess, uterus and respiratory healthy groups all had the same invasion ability during infection with the human cells ($p \geq 0.846$) and with the equine cells ($p = 1$) (**Figure 3.5**). Further, there was no difference between the host cells. Each group invaded the human cells with the same efficiency as the equine host cells ($p \geq 0.089$) (**Figure 3.5**) suggesting both similar microbial adhesion proteins and host cell reception in the two infection models. The mucoid *szp*-13 strain had significantly reduced invasion capabilities ($p \leq 0.004$) than all the other groups including the respiratory disease group from which it belongs (**Figure 3.5**).

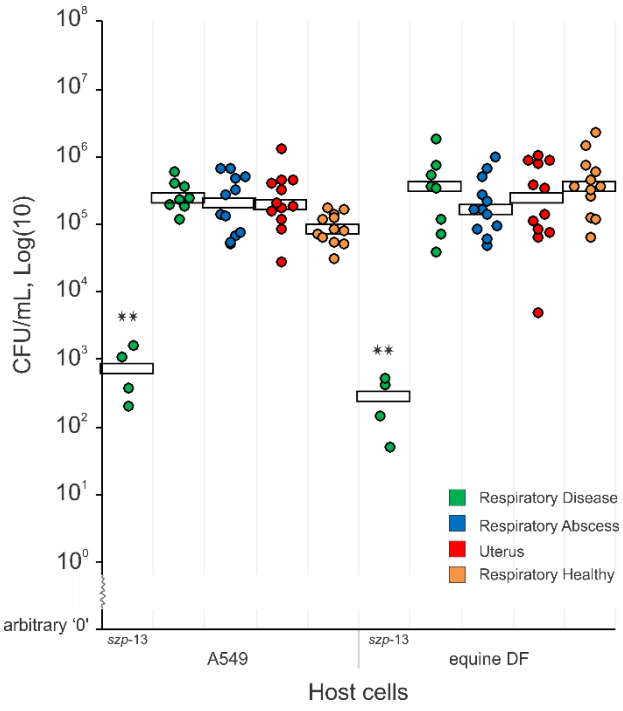


Figure 3.5: Invasion of *S. zooepidemicus* separated by isolation source group (respiratory disease, respiratory abscess, uterus and respiratory healthy). The mucoid JK isolate (szp-13) is separated out. White bars represent the medium of the data points, **p = 0.001.

Further separation of the groups into individual *szp* strains showed no variation in their abilities to internalise into the equine cells ($p \geq 0.125$) (**Figure 3.6 B**) with the exception of

szp-13 (p = 0.029). The mucoid szp-13 strain also had a significant reduction in internalising into the human A549 cells (p = 0.029). Additionally, the respiratory healthy isolate, *szp*-42, invaded with less efficiency into the human cells than a respiratory abscess isolate (szp-22a, p = 0.036) (**Figure 3.6 A**). There was no difference in the internalisation between any other strains in the human cells (p ≥ 0.190) (**Figure 3.6 A)**.

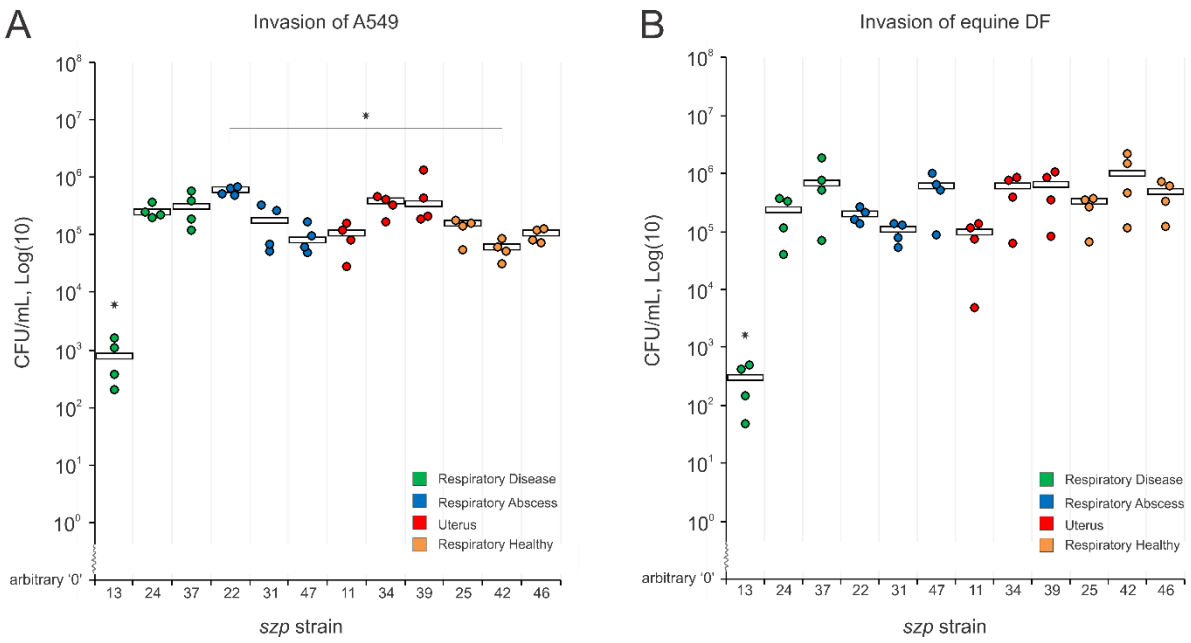


Figure 3.6: *S. zooepidemicus* strains invasion into (a) A549 and (b) equine dermal fibroblast mammalian cells. White bars represent the medium of the data points, *p < 0.05

Together, the two wild type *S. equi* strains had reduced ability to internalise than the vaccine strain in both cell lines (p = 0.029) (**Figure 3.7**). The wildtype strains invasion of both host cells (range 0 - 652 cfu/mL, median 11 cfu/mL) were a 82 % fold decrease (log change 2.65) from the vaccine strain (range 1,388 - 15,520 cfu/mL, median 4739 cfu/mL) (p ≤ 0.001) (**Figure 3.4 B**). EDED090 (*seM*-99) was able to internalise with greater efficiency than G100 (*seM*-100) into A549 human cells (p = 0.029) (**Figure 3.7**). However, there was no difference between the adhesion invasion capabilities between the wild type strains in the equine host cells (p = 0.486) (**Figure 3.7**).

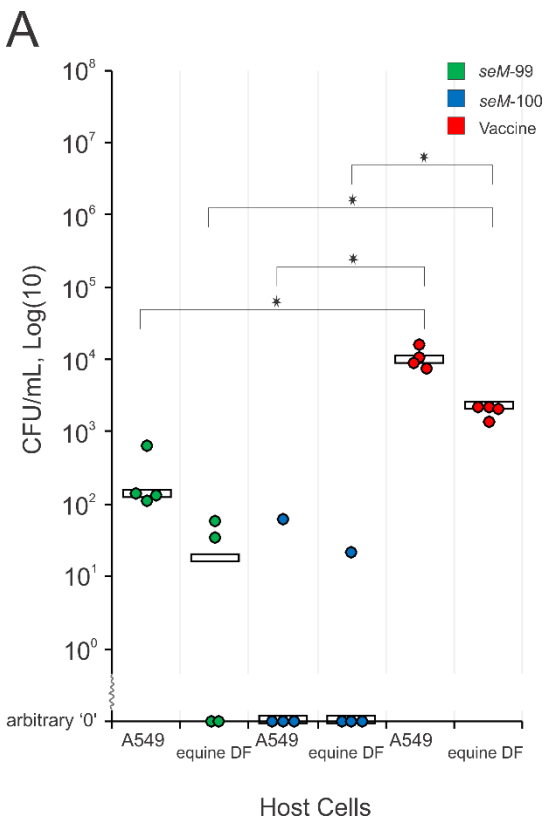


Figure 3.7: *S. equi* invasion of mammalian cell lines. White bars represent the medium of the data points, *p ≤ 0.05

3.4 Discussion

3.4.1 Biofilms of *S. equi* subspecies

Biofilms are sessile microbial bacteria communities adhered to a surface and encapsulated in a matrix of extracellular polymeric substances (EPS) making their removal by either the host immune response or antimicrobial treatment difficult. Biofilm formation is a dynamic and multifaceted process following environmental queues (Vyas *et al.*, 2019). Initial attachment of bacteria to a surface is followed by a transition from a planktonic phenotype to a sessile one with the bacteria becoming attached, producing the EPS matrix (Vyas *et al.*, 2019). As the biofilm matures its matrix enhances to assist in defence, nutrient uptake, and facilitate reproduction (Vyas *et al.*, 2019). Dispersal of the biofilm is the final stage and is responsible

for the dissemination of bacteria to new sites (Vyas *et al.*, 2019). In streptococci, biofilms play an important survival strategy that can result in asymptomatic carriage (Roberts *et al.*, 2012) and play a role in the severity and progression of disease (Young *et al.*, 2016). Abilities of biofilm formation can vary between strains of streptococcal species resulting in either disease or long-term colonisation (Sanchez *et al.*, 2011; Parker *et al.*, 2016). This suggests that there may be different biofilm forming capabilities between the New Zealand isolates of *S. zooepidemicus*, in particular between isolates collected from healthy horses versus disease. However, for the *S. zooepidemicus* isolates tested, there was no association between the anatomical group an isolate belonged to with either the level of biofilm production or the biofilm trajectory. Unexpectedly, some isolates from the uterus, respiratory disease and respiratory healthy groups were unable to form and/or maintain a biofilm. Strains associated with disease had the same probability as the isolates from healthy animals of forming strong biofilms or none-to-weak biofilms. The uterus isolates showed the same variability as the respiratory isolates.

Variances in biofilm kinetics between *S. zooepidemicus* isolates as observed in this study are known to occur (Etchecopaz *et al.*, 2016). The observation of the New Zealand *S. zooepidemicus* strains reaching their maximum biofilm production at 12 or 24 h could be a reflection of adaptation to the growth medium, different generation times of the strains and or different starting inoculum sizes. Generation times were not considered and inoculums were taken from a dilute overnight culture without standardizing the number of initial bacterial units. Nonetheless, the ability to form a biofilm was considered at any stage during the 72 h with the ability to maintain that biofilm rather than at a single time point. Further, the trajectories still suggest that *S. zooepidemicus* biofilms reach a maximum level sometime between 12 and 24 h after inoculation, at which point the biofilm goes into a maintenance stage and is likely a reflection of the stationary phase of growth. This trajectory is different from the *S. equi* strains which increased biofilm production towards 72 h with a maintenance stage not reached. However, due to the bias in numbers representing the two subspecies, further experiments would be needed to validate this finding.

The late onset and weak biofilm production for the wild type *S. equi* may reflect strains more capable of invasive disease rather than colonisation such as seen in other streptococci. Invasive strains from GBS (Group B Streptococci) are significantly more likely to form weak biofilms with strong biofilms formed by asymptomatic colonising strains (Parker *et al.*, 2016). Biofilm production in *S. pneumoniae* isolates also relates to colonisation rather than contributing to invasive disease (Sanchez *et al.*, 2011). Although known as a globally feared pathogen causing the highly infectious disease strangles, *S. equi's* ability to asymptomatically colonise has led to its global success as a pathogen. The late onset and weak biofilm production could be a reflection of experimental design not mimicking an *in vivo* setting or individual bacterial strain ability. Most *S. equi* infections are acute with approximately 10 % to 40 % leading to persistent carriage in the guttural pouch due to lack of complete clearance of the bacterium (Boyle *et al.*, 2018). The *S. equi* biofilms never reached maturation like the *S. zooepidemicus* isolates with their biofilm trajectory continuing to rise at the end of the experimental time frame. This could be a reflection of the growth medium used and different adaptation times between the strains. Future studies could look at different growth media and their ability to induce biofilm formation such as the addition of horse serum (Bustos *et al.*, 2016). As biofilm production on plastic does not mimic real life, it will be missing host factors involved in the triggering and regulation of microbial biofilm formation. Further, future work should also consider other environmental variables, which may lead to tissue tropism in biofilm production such as pH, nutrient availability and gas concentrations. More work is needed in this area to determine if *S. zooepidemicus* produces biofilms during interaction with the host and what factors govern this. Additionally, work is needed to determine whether biofilm production has any influence on clinical outcome. The biofilm results presented here showed *in vitro* phenotypic variances between strains and were used to ensure the different bacterial phenotypes were selected for use in the invasion assay.

3.4.2 Invasion Assay of *S. equi* Subspecies

The increased ability of *S. zooepidemicus* to invade mammalian cells when compared to *S. equi* may be a reflection of *S. zooepidemicus* being a colonising pathogen causing chronic infections and *S. equi* being an acute pathogen. Interestingly, the large and highly mucoid *S. zooepidemicus* isolate JK (*szp*-13) behaved more like the mucoid wild type *S. equi* isolates

than any of the *S. zooepidemicus* isolates. This isolate was from the nares of a horse with nasal discharge and a cough.

A highly mucoid colony phenotype, as seen in the *S. zooepidemicus szp*-13 and *S. equi* wild type isolates, is due to capsule production common to Gram-positive streptococci and may affect their ability to adhere and sequentially infect host cells. High levels of capsule production have been shown to decrease the level of adherence and invasion of *S. zooepidemicus* to host cells, with capsule production becoming attenuated during internalisation (Xu *et al.*, 2016). This maintenance of high level capsule production is suspected to be caused by a decrease in hyaluronidase activity which in turn would negatively impact invasion of streptococci (Holden *et al.*, 2009). This is hypothesised to be the primary reason why *S. equi* rarely progresses past the pulmonary lymphatic system. When it does, it causes severe complications due to metastatic spread of the infection becoming an invasive disease known as bastard strangles (Boyle *et al.*, 2018).

All *S. zooepidemicus* strains tested were able to invade both the human and equine cells trialled. There was no difference in invasion capabilities between respiratory disease, respiratory abscess, respiratory healthy and uterus isolation groups. In the absence of the unusual JK isolate, there was little to no variance of individual strain capabilities in the invasion of these host cells. The ability of all strains to potentially invade the host cells provides *S. zooepidemicus* with a fitness advantage at their disposal for avoiding host immune responses and thus clearance of disease. Becoming intracellular, *S. zooepidemicus* will find a haven, protected from the host's immune defences and from antimicrobial treatments. This in turn will allow it to opportunistically cause disease and transmit to others from potential asymptomatic carriage.

3.4.3 Relationship Between Biofilm and Invasion Abilities

The ability of all *S. zooepidemicus* strains to infect host cells was in contrast to the biofilm results (section 3.3.2) that showed high variability in biofilm production across these same strains. Isolates tested in the invasion assay consisted of those ranging from no biofilm production (including those with very weak biofilms that they were unable to maintain) to weak, moderate, and strong biofilm producers. This suggests that these *in vitro* assays are

not linked, and thus biofilm production on plastic is not a good indicator for *S. zooepidemicus* host cell invasion capabilities. In *S. equi*, however, there is a potential link between biofilm formation and ability to invade host cells. The wild type *S. equi* strains formed only weak biofilms compared to the strong biofilm of the PINNACLE® I.N. vaccine strain. This correlated to the decreased invasion of the wild type strains compared to that of the vaccine strain. This decreased ability in both forming biofilms and invasion might in turn be an indication of *S. equi* as an extracellular pathogen and *S. zooepidemicus* having a better ability to adopt an intracellular lifestyle.

The PINNACLE® I.N. vaccine, available for use in several countries including the USA, Canada and New Zealand, is based on a live attenuated, non-encapsulated *sem*-2 strain of *S. equi* (*Se*CR32) derived through chemical mutagenesis via N-methyl-N'-nitro-N-nitrosoguanidine treatment (Timoney, 1993a; Patty & Cursons, 2014). Concerns around its safe use have arisen with a minority of previously healthy yearling horses developing strangles after receiving the vaccine (Patty & Cursons, 2014; Cursons *et al.*, 2015). It forms typically dry non-mucoid colonies compared to wildtype *S. equi*, thus being phenotypically similar to *S. zooepidemicus* in colony appearance (**Figure 3.1**), although it has been known to revert to the encapsulated form (Walker & Timoney, 2002; Patty & Cursons, 2014). The lack of encapsulation could account for its ability to form biofilms and invade the mammalian cells, similar to *S. zooepidemicus* and conversely to the JK highly mucoid isolate. Although, off-target effects of the chemical mutagenesis could affect these results.

3.4.4 SZP Truncation and *S. zooepidemicus* Biofilm and Invasion Abilities

As discussed in the previous chapter (Chapter 2), the *szp*-22 strain with truncation of the SZP cell surface-anchored protein, which mediates binding and attachment to host cells, had the potential to affect biofilm formation and invasion of host cells. However, strain *szp*-22 from both respiratory disease and abscess isolates (a and b respectively) were able to form moderate to strong biofilms. The *szp*-22a abscess isolate was tested in the invasion assay and was able to invade the equine dermal fibroblasts to the same level as the other strains. In addition, *szp*-22a had increased internalisation within the A549 cells than a *szp*-42 from a healthy isolate. There was no negative influence in the truncation of the SZP protein on its

invasion capabilities. This gives evidence to the complexity and multifactorial aspects involved in the biofilm formation and invasion capabilities of *S. zooepidemicus* other than a single gene. Other genes, host microbiota and the host immune response will influence a strain's ability to adhere and sequentially invade or form a biofilm. For example, there are other gene products shown to be important in *S. zooepidemicus* biofilm formation such as the fibronectin-binding protein (Fnz) (Ma *et al.*, 2013). Fnz may also act as a regulator for other *S. zooepidemicus* genes associated with cell adhesion including the *szp*, a IgG-binding protein (*zag*) and a glyceraldehyde-3-phosphate dehydrogenase (*gapdh*) (Ma *et al.*, 2013). The microbiome of the horse may also play a role in the ability of *S. zooepidemicus* to adhere. For example, *S. zooepidemicus* in the presence of *Candida glabrata*, was able to form a mixed biofilm earlier and to higher levels than the individual biofilms (Etchecopaz *et al.*, 2018). The recognition-receptors of the horse (host) to the *S. zooepidemicus* surface molecules will further play a role in successful bacterial attachment and avoidance of immune attack.

3.4.5 Conclusion

An important limitation of this *in vitro* study is that it does not investigate the virulence potential of biofilm and invasion capabilities *in vivo*. These experiments do not reflect clinical infection and the isolates may act differently during normal colonisation and invasion of a host. Nevertheless, molecular techniques such as MLST and *szp* typing were not able to differentiate between colonising and infectious strains or uterus and respiratory strains. It was thus hoped that phenotypic characteristics might be more useful for discrimination. Future experiments need to confirm the relationship between these capabilities and virulence potential. It may be that the ability to form biofilms and/or invade mammalian cells *in vitro* correctly represents *S. zooepidemicus* as an opportunistic pathogen, using these techniques to colonise and hide from the horse immune surveillance system and cause disease opportunistically when conditions are ideal. Further, being a coloniser increases the opportunity of asymptomatic carriage by multiple strains. Future studies could look at treatments to disrupt or reduce colonisation and the effectiveness of current treatments used in New Zealand, as indicated in Chapter 1, on *S. zooepidemicus* biofilms and invaded mammalian cells. Although these phenotypic *in vitro* experiments could not differentiate between anatomical and disease/healthy groups, genomes may hold the key to clarify

whether there is any difference between healthy and disease isolates in pathogenic potential. Considering the accessory genome and the possession of different fitness factors will be the way forward in disease associated, strain differentiation of *S. zooepidemicus*.

Chapter 4

Genomic Diversity of New Zealand *Streptococcus zooepidemicus* Isolates

4.1 Introduction

The evolution of the β-haemolytic, pyogenic streptococci is characterised by high gene turnover (gene gain/loss), rife with HGT (Richards *et al.*, 2014). This includes *S. zooepidemicus*, where mobile genetic elements have played a major role in shaping their genomes through chromosomal inversions, recombination and HGT resulting in a diverse species (Beres *et al.*, 2008; Holden *et al.*, 2009; Ma *et al.*, 2013). There is strong evidence of genetic exchange between *S. equi* and *S. zooepidemicus* with others species of the pyrogenic group, in particular *S. pyogenes* as well as *S. agalactiae* and *S. dysgalactiae* (Beres *et al.*, 2008; Holden *et al.*, 2009; Waller & Robinson, 2013). Genes involved in pathogenesis and nutrient uptake/metabolism have been fundamental in their evolution (Richards *et al.*, 2014). This process of accessorizing their genomes has in turn allowed them successful adaptation to a wide range of hosts and environments (Jackson *et al.*, 2011; Richards *et al.*, 2014). The success of commensals and pathogens depends on repertoires of fitness factors such as adhesion, biofilm production and/or resistance to phagocytosis.

The acquisition of accessory genes have defined strains of *S. zooepidemicus*. Genome comparisons and other studies on virulence genes have indicated potential factors important to different strains abilities to colonise different hosts. For example a virulence-associated protein E (vapE), along with proteins involved in adherence, were unique to the ATCC35246 pig strain of *S. zooepidemicus* (Ma *et al.*, 2013). The acquisition of genes making up a novel arbutin sugar utilisation system (PTS) has been proposed as a fitness advantage in

S. zooepidemicus strains infecting ruminants for which this sugar is found in their diet (Steward *et al.*, 2017).

Accessory genes are often found in regions known as genomic islands, derived from HGT, and typically carry genes associated with novel functions advantageous to the bacterial isolate. There are a number of different types of genomic islands; pathogenicity islands (PAIs), metabolic islands, antibiotic resistance islands, symbiotic islands and also integrative and conjugative elements (ICE) (Jackson *et al.*, 2011; da Silva Filho *et al.*, 2018). It is thought that the key speciation event of *S. equi* evolving from *S. zooepidemicus* was the acquisition of an ICE carrying a novel iron uptake system (Holden *et al.*, 2009). In comparing the genomes of *S. equi* and a virulent human and a healthy horse strain of *S. zooepidemicus* (MGCS10565 and H70 respectively) with a virulent pig strain (ATCC35246), four novel pig PAIs were found encoding genes for three toxin-antitoxin systems and one restriction modification system (Ma *et al.*, 2013). Deletion of part of one of the PAIs, which included genes belonging to the toxin-antitoxin system, increased host survival in experimentally-infected mice indicating the PAIs role in bacterial virulence.

4.1.1 Aim

Gene gain and loss may have implications for *S. zooepidemicus* strains resulting in either a commensal or pathogenic life style. The aim of this chapter is to explore the genomes of *S. zooepidemicus* isolates from New Zealand horses, looking into the core and accessory genomes and gene presence/absence to tease out potential associations with disease status or isolation source. It further uses publicly available genomes both to expand the genomic analysis and to place the New Zealand strains within a global context.

4.2 Methods

4.2.1 *S. zooepidemicus* Samples

Fifty-one New Zealand horse *S. zooepidemicus* isolates, encompassing a range of *szp* and MLST types from the different isolation source groups, were chosen for genome sequencing

(**Table S4. 1**). A further 48 global representatives of *S. zooepidemicus* with available genome sequencing data were accessed from NCBI and used for a global comparison (**Table S4. 2**).

4.2.2 Genomic DNA Isolation and Sequencing

Due to opportunities for conducting genome sequencing arising at different stages of the research, the genomic DNA isolation and sequencing of the New Zealand isolates were performed in three different ways, outlined below.

Method 1. Three isolates were sequenced as previously described (Cursons *et al.*, 2015). DNA extraction was performed in-house and sequenced by collaborators at Animal Health Trust (AHT), United Kingdom.

A single colony of each isolate was inoculated in 10 mL of Bacto™ BHI (BD, 237500) and incubated overnight statically at 37 °C. Cells were pelleted by centrifugation at 16,000 *g* for 1 min and re-suspended in 100 µL of 10 mM Tris (AppliChem, A2264) and 1 mM EDTA (ethylenediaminetetraacetic acid, Scharlau, AC0965) (TE, pH 8.0) containing 10 µg of lysozyme (Sigma, L-7001). After incubation at 37 °C for 1 h, 350 µL of Lysis Solution A (0.1 M Tris, 0.05 M EDTA, 1.0 % SDS (sodium dodecyl sulphate, Sigma-Aldrich, 822050), and 20mM NaCl (Merck, 1.06404)) and 30 µL of proteinase K (Roche, 3115879) (60 µg/µL) was added. The solution was then incubated at 65 °C for 1 h. 350 µL of 5 M LiCl (AJAX-Finechem (Unilab), 292-500G) and 750 µL of chloroform were added and the biphasic solution was mixed on a rotating wheel for 30 min at room temperature. The solution was centrifuged at 16,000 *g* for 10 min, the top aqueous phase collected, transferred to a new tube and the DNA precipitated by adding an equal volume of isopropanol. DNA was recovered by centrifugation at 16,000 *g* for 20 min, washed with 1 mL of 70 % ethanol and resuspended in 100 µL TE, pH 8.0.

The purified DNA was sent to AHT for sequencing. DNA libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina), dual labelled using compatible indexing primer pairs from the Nextera XT index kit (Illumina), purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified using the KAPA library quantification kit for Illumina sequencing platforms (KAPABiosystems) as per the relevant manufacturer's instructions. The

libraries were then sequenced on an Illumina MiSeq genome sequencer, which generated 250 bp paired end reads.

Method 2. The second method was applied to 12 isolates. The DNA extraction was performed in-house using the following method and then genome sequenced by Beijing Genomics Institute (BGI) Tech Solutions (Hong Kong) Co., Limited.

A single colony of each isolate was inoculated into 10 mL of BHI and grown overnight statically at 37 °C. Cells were pelleted by centrifugation at 2300 *g* for 20 min. Culture media was decanted and the remaining liquid used to re-suspend the bacterial pellet prior to transferring to a 2 mL microfuge tube. The bacteria were pelleted again by centrifugation at 16,000 *g* for 1 min to remove the remaining media. The bacterial pellet was re-suspended in 200 µL of TE, pH 8.0 with 0.5 mg/mL of lysozyme (Sigma, L-7001) and 150 units/mL of mutanolysin (3000 units/mL, Sigma, M9901-5KU). The suspension was incubated for 30 min at 37 °C. 600 µL of Lysis Solution B (5 M guanidinium thiocyanate (BDH, 443243E), 0.5 % N-lauroylsarcosine sodium salt (Sigma, L-5125), 26.5M tri-sodium citrate (BDH, 102424L), 0.7 % 2-mercaptoethanol (Sigma, M-6250), pH 7.0) was added and gently mixed for 10 min on a rotating wheel at room temperature. DNA was then precipitated by the addition of an equal volume of isopropanol and recovered by centrifugation at 16,000 *g* for 15 min. The DNA pellet was washed with 1 mL 70 % ethanol and re-suspended in 100 µL TE, pH 8.0 with the addition of 0.5 µg/µL RNaseA (Roche, 10109142001) and incubated at 37 °C for 30 min. The DNA was then taken through a second purification process as follows: 570 µL Lysis Solution A (from method 1), 100 µL 5 M NaCl and 80 µL 10 % CTAB (Hexadecyltrimethylammonium bromide, Sigma-Aldrich, H5882 (in 0.7 M NaCl)) was added with a further incubation of 10 min at 65 °C. An equal volume of chloroform was then added and mixed to form an emulsion then left at room temperature for 10 min mixing gently by rotation. The solution was centrifuged at 16,000 *g* for 10 min, the aqueous phase was collected and the DNA precipitated by adding an equal volume of isopropanol. DNA was recovered by centrifugation at 16,000 *g* for 20 min, washed with 1 mL of 70 % ethanol and resuspended in 100 µL TE, pH 8.0.

In order to remove any low quality, degraded DNA, 40 µL of the purified DNA was run on a 0.7 % low melting point agarose (SeaPlaque, 50102) gel stained with RedSafe Nucleic Acid

Staining Solution (iNtRON, 21141). The high molecular weight bands were cut out and the agarose melted at 65 °C for 5-10 min. A 2.5 x volume of TE was then added to the melted agarose and DNA mixture and further incubated at 65 °C for 2 min then cooled to room temperature for 5 min. The DNA was extracted with the addition of an equal volume of phenol (pH 8) (Sigma-Aldrich, P4557), mixed to form an emulsion. The emulsion was centrifuged for 10 min at 16,000 *g*. The top aqueous layer was collected and the extraction repeated with an equal volume of chloroform added. The DNA from the top aqueous phase was precipitated by addition of 1/10 volume of 3 M sodium acetate, pH 5.2 (BDH (AnalaR), 10236) and two times the volume of ethanol, then incubated at -20 °C for 1 h. The DNA was recovered by centrifugation at 16,000 *g* for 20 min, washed with 1 mL of 70 % ethanol and resuspended in TE, pH 8.0.

The purified DNA was sent to BGI for genome sequencing. DNA libraries were prepared and sequenced on an Illumina HiSeq 2000 sequencing platform, generating 151 bp paired end reads. As indicated by the provider, this consisted of shearing the DNA into smaller fragments with a desired size by Covaris S/E210 or Bioruptor. Then the overhangs, resulting from fragmentation, are converted into blunt ends by using T4 DNA polymerase, Klenow Fragment and T4 Polynucleotide Kinase. After adding an 'A' base to the 3' end of the blunt phosphorylated DNA fragments, adapters were ligated to the ends of the DNA fragments. The desired fragments were purified through gel-electrophoresis, then selectively enriched and amplified by PCR. The index tags were introduced into the adapter at the PCR stage as appropriate and a library quality test was completed. At last, the qualified BS library was used for sequencing.

Method 3. The third method was used for the remaining 36 isolates. The DNA was extracted and genome sequenced by collaborators from AHT.

An Ames swab (Fort Richard, 108C USE) was taken from a blood culture plate (Fort Richard, 1100) of each isolate. The live cultures were sent to AHT for processing using previously described methods (Mitchell *et al.*, 2021). Briefly, DNA was extracted from single colony derived cultures using the GenElute (Sigma) kit and the Gram +ve lysis solution containing 250 units/mL mutanolysin and 2 x 10^6 units/mL lysozyme. DNA libraries were prepared using the

Nextera XT DNA sample preparation kit (Illumina), pooled and quantified using the KAPA library quantification kit for Illumina sequencing platforms as per the relevant manufacturer's instructions. The libraries were then sequenced on an Illumina MiSeq genome sequencer.

4.2.3 Genome Assembly

FastQ Illumina paired end read files were *de novo* assembled using the Shovill pipeline version 1.0.9 (https://github.com/tseemann/shovill) at default settings. This pipeline first estimates read length (seqtk v1.2-r95-dirty, https://github.com/lh3/seqtk), genome size (mash v2.2, (Ondov *et al.*, 2016)), and sequencing depth. If sequencing depth is greater than 100 x it then takes a subsample of the data using seqtk to represent a sequencing depth of 100 x. Next, it implements a kmer-based, conservative read error correction using Lighter (v1.1.2) (Song *et al.*, 2014) with the maximum number of corrections within a 20 bp window set to 1. It then applies read stitching/overlap to increase the length of reads prior to assembly using FLASH (v1.2.11) (Magoč & Salzberg, 2011) with the following set parameters; minimum overlap of 20 bp, maximum overlap equalling the maximum read length, and the maximum mismatch ratio at the default of 0.25. The assembly in Shovill was run through SPAdes (v3.14.0) (Nurk *et al.*, 2013) using a modified kmer range consisting of five different kmers. This range of five kmers is determined by the read length with a minimum kmer of 31 and a maximum set at below 75 % of the average read length and no greater than 127 reads. Default Shovill has repeat resolution enabled (Prjibelski *et al.*, 2014; Vasilinetc *et al.*, 2015) and both mismatch corrector and coverage cut-off disabled for the SPAdes assembler. Shovill inputs into SPAdes the output of FLASH, which includes the set of paired end fastq read files that were not merged and the single-read fastq file of merged extended reads. Shovill then corrects minor assembly errors using BWA (v0.7.12-r1039) (Li, 2013), samclip (v0.2, https://github.com/tseemann/samclip), samtools (v1.10, https://github.com/samtools), and Pilon (v1.23) (Walker *et al.*, 2014). Pilon fixes individual base errors (SNPs) and small insertions and deletions (indels) based on coverage, mapping and base quality. Variants will only be called if coverage is at least 25 % of the mean coverage, with a minimum alignment mapping quality of 60 and a minimum base quality of three. Finally, low coverage (coverage < 2 x) and homopolymer contigs are removed prior to Shovill producing the final assembly.

4.2.4 Phylogenomic Analysis

4.2.4.1 Core Genome MLST Phylogenetic Tree

Final genome assemblies were provided to Andrew Waller of AHT who was able to upload them into the Pathogenwatch scheme for *S. equi* (https://cgps.gitbook.io/pathogenwatch/), inclusive of subspecies, to create a core genome (cg) MLST phylogenetic tree as previously described (Mitchell *et al.*, 2021). This scheme creates a phylogeny based on alleles in the core reference genome consisting of 1286 core genes from the *S. equi* reference isolate Se4047. Mobile genetic elements, insertion sequences, sortase-processed proteins and repeats are excluded from the core genome. Alleles were assigned based on variations relative to the reference core genome. The total number of variant sites between each pair of assemblies (pairwise score matrix) (excluding indels) were used to construct the dendrogram using the Analysis of Phylogenetics and Evolution (ape) package (Paradis *et al.*, 2004) neighbour-joining implementation. The resulting tree was then midpoint rooted using the phangorm package (Schliep, 2011). The phylogenetic tree was then visualised and annotated using iTOL (Letunic & Bork, 2019).

4.2.4.2 Genome Annotation

Assemblies were annotated using Prokka (v1.14.6) at default settings and passing in the "--proteins" flag to provide the reference H70 (NC_012470) genome genbank file, ensuring consistent gene naming (Seemann, 2014). Prokka annotation follows a series of steps, first ARAGORN (v1.2.38) was applied to find and annotate tRNA and tmRNA genes based on the bacterial, archaeal and plant plastid genetic code (Laslett & Canback, 2004). MinCED (v0.2.0, https://github.com/ctSkennerton/minced) was then employed to find CRISPRs. Coding sequences were then predicted with prodigal (v2.6.3) also based on the bacterial/archaeal genetic code and ensuring partial genes at edges of sequence or genes that run into gaps were not included (Hyatt *et al.*, 2010). Using BLAST+ blastp (v2.9.0+, (Camacho *et al.*, 2009)), Prokka then annotated the coding sequences by first searching its created protein database (BLAST+ makeblastdb, v2.9.0+, (Camacho *et al.*, 2009)) made from the supplied reference genome (--proteins option). This was followed with searches through Prokka's three core databases in order: ISfinder (https://isfinder.biotoul.fr/ (Siguier *et al.*, 2006)), NCBI Bacterial Antimicrobial Resistance Reference Gene Database (https://www.ncbi.nlm.nih.gov/bioproje

ct/313047), and Swiss-Prot of the UniProtKB (https://www.uniprot.org/uniprot/). Prokka annotation then finished off by searching its curated hidden Markov model profile (HMM) libraries using hmmscan of HMMER3 (v3.1b2, (Eddy, 2011)). To determine functional annotations, the eggNOG-mapper tool (v2.0.1-4-g2466c1b) was used based on eggNOG orthology data (Huerta-Cepas *et al.*, 2017; Huerta-Cepas *et al.*, 2019) with the use of DIAMOND (Buchfink *et al.*, 2015) to perform sequence searches.

4.2.4.3 Pangenome Analysis

Pangenome analysis was carried out using Roary (v3.8.2) (Page *et al.*, 2015; Sitto & Battistuzzi, 2019). The general feature format (GFF) assembly files from Prokka were input for the Roary pipeline. Roary was run with the following parameters: threshold of isolates required to define a core gene set at 99 % and the minimum percentage identity for sequence comparisons performed by BlastP set at 95 %.

The pangenome profile was analysed using the PanGP tool (Zhao *et al.*, 2014) using the distance guide (DG) algorithm drive. The Roary 0/1 matrix output file was used as the input file after removing column and row headers. Parameters were set as follows; the sampling size was set to sample 500 different combinations based on the genome diversity of the isolates with the amplification coefficient set to 10 and then repeated 10 times. PanGP fits a continuous curve based on Heaps' law (Tettelin *et al.*, 2005; Rasko *et al.*, 2008; Tettelin *et al.*, 2008) with the following equations:

Pangenome

$$y = A_{pan}x^{B_{pan}} + C_{pan}$$

Core genome

$$y = A_{core}e^{B_{core}x} + C_{core}$$

Number of unique genes

$$y = A_{new}x^{B_{new}}$$

Where *y* denotes the size of the pangenome, core genome or number of new genes respectively, *x* denotes genome number and A, B, and C are fitting parameters. When the exponential value $B_{pan} < 0$ the pangenome is considered closed with the curve approaching a

constant and when $0 < B_{pan} > 1$ the genome is considered open with the curve increasing and unbound as new genomes are added.

4.2.5 Genome Browser

Geneious R11.1.4 (https://www.geneious.com/) was used to visualize and examine genomes. Putative fitness factors were investigated within Geneious using the blast and align (ClustalW) tools available.

4.2.6 Bacteriocin Mining

BAGEL4 webserver (http://bagel4.molgenrug.nl/) was used for *in silico* screening of the 51 New Zealand *S. zooepidemicus* genomes for the presence of putative bacteriocins (van Heel *et al.*, 2018). The DNA fasta files of each assembled genome were uploaded to the web interface of BAGEL4 for analysis. BAGEL4 identifies gene blocks in the genome that contain putative bacteriocins and other ribosomally synthesized and post-translationally modified short peptides (RiPPs) along with their corresponding genes. These gene blocks encode the toxin and associated proteins for modification, immunity, regulation and transport (van Heel *et al.*, 2018). The output was used to determine the prevalence and diversity of bacteriocin gene clusters within the New Zealand isolates. All bacteriocin predictions by BAGEL4 were considered and identified as one of three broad classes: class I post translationally modified peptides, class II unmodified peptides and class III large peptides > 10 kDa.

4.2.7 Pan-GWAS

The Scoary (v1.6.16) pipeline was used to perform pangenome wide association studies (pan-GWAS) (Brynildsrud *et al.*, 2016). Scoary analyses the gene clusters of the pangenome, scoring associations to predefined traits while accounting for population structure. The naïve Fisher's exact test with a Benjiamini-Hochberg false discovery rate adjusted p-value less than 0.05 was set to identify a gene of interest. Gene presence was then considered to have a significant association for the trait analysed with the following criteria: odds ratio > 1, sensitivity > 80 % and specificity > 80 % (Arredondo-Alonso *et al.*, 2019).

4.2.8 Statistics

Unless otherwise specified, the Kruskal-Wallis test was used for pairwise multiple comparisons with the Bonferroni Correction post hoc test and the Mann-Whitney U test for comparison of two independent samples. Gene presence associations were assessed using the Fisher's exact test for 2 x 2 contingency tables and implementing the Fisher-Freeman-Halton exact test of independence for R x C contingency tables. A significance level of 0.05 was used for all tests. All statistical analyses were carried out with SPSS (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.).

4.3 Results

4.3.1 Genome Sequencing and Assembly

As multiple DNA preparations and sequencing methods were used (section 4.2.2), the resulting genome sequencing and assemblies were compared. The main difference between the genome sequencing methods used was method 1 produced longer reads but with reduced sequencing depth (**Table 4.1**). A sequencing depth of 35 x to 50 x is considered sufficient for good coverage of small genomes (Desai *et al.*, 2013). Method 1 fell within this range and methods 2 and 3 were well above with sequencing depths greater than 100 x (**Table 4.1**). All 51 isolates sequenced produced high quality reads regardless of the DNA extraction and sequencing method with greater than 90 % of the reads resulting in a phred quality score of at least 30 which equates to a base call accuracy of 99.9 % (1/1000 error rate) (**Table 4.1**). Reads were used for *de novo* assembly, producing genomes with the expected length and GC content for the subspecies based on reference H70, NC_012470 (2.15 Mb and 41.5 % respectively). Ultimately, the data produced by method 3 resulted in the best genome assemblies based on largest sequencing depth, smallest number of contigs and largest N50 (a statistical measure of contig length with higher numbers indicating lower levels of fragmentation) (Dominguez Del Angel *et al.*, 2018) (**Table 4.1**). However, all 51 assemblies were considered to have sufficient quality for further genome analysis.

Table 4.1: Genome sequencing and assembly statistics of the 51 New Zealand horse *S. zooepidemicus* isolates. Method column refers to the sequencing method used; method 1 being an Illumina MiSeq platform with 250 bp reads through AHT, method 2 an Illumina HiSeq 2000 platform with 150 bp reads by BGI, and method 3 an Illumina MiSeq platform with 151 bp reads through AHT. Brackets indicate minimum and maximum range. Phred quality score of at least 30 equates to a base call accuracy of 99.9 %. N50 is a statistical measure of contig length.

Read Statistics

method	Isolates (n)	total bp	read length	sequencing depth	% reads ≥ 30 phred score
# 1	3	112,079,594 (74,191,900-133,277,638)	230 (35-251)	55x (37-65x)	94 % (89-98 %)
# 2	12	247,911,725 (215,421,000-252,000,000)	150 (150-150)	122x (104-130x)	93 % (88-97 %)
# 3	36	323,756,776 (265,322,402-401,258,642)	151 (151-151)	158x (132-205x)	95 % (92-98 %)

Assembly Statistics

method	Isolates (n)	genome length (bp)	N50	# contigs	% GC content
# 1	3	2,133,463 (2,064,213-2,186,154)	36,176 (32,649-41,545)	152 (148-157)	41.6 % (41.6-41.7 %)
# 2	12	2,054,182 (1,955,373-2,133,522)	78,081 (46,733-1,022,050)	86 (72-122)	41.6 % (41.4-41.7 %)
# 3	36	2,088,066 (1,968,388-2,208,175)	108,828 (56,409-271,074)	65 (41-97)	41.5 % (41.2-41.8 %)

4.3.2 cgMLST

The 1286 core genes of the Pathogenwatch scheme had on average 98.3 % (SD 0.62, min 96.8 %, max 99.7 %) representation within each New Zealand isolate. An average of 31.4 % (SD 1.74, min 27.7 %, max 34.4 %) of each New Zealand genome was accessory to this core. The cgMLST phylogenetic reconstruction showed no obvious clustering of the groups respiratory disease, respiratory abscess, respiratory healthy or uterus (**Figure 4.1**). There was high genetic diversity between isolates with the pairwise core genome single nucleotide polymorphisms (cgSNPs), on average, consisting of 26,540.8 SNPs (SD 4132.7, min 5, max 39,758). The MLST typing (Chapter 2) was consistent with the cgMLST, with ST types clustering together. Yet, cgMLST provided a more accurate phylogeny resolving all branches with better resolution to differentiate the isolates of the same ST (**Figure 4.1**). Isolates AK17078_1 and WK16140_3, both ST-1 differed by 19 cgSNPs. Isolates from ST-2, NZVP01 and WK434 differed by only five cgSNPs. KMW006_31a and KMW006_32 differed by 20

cgSNPs, both ST-348. Isolates from ST-383, KMW011_58 and KMW012_62 had a difference of 294 cgSNPs. The isolates lacking a copy of the yqiL gene, WK15143_3 and WK15111_3 (13 cgSNPs), with their remaining six alleles matching ST-132 clustered with NZVP25 of this same ST type (1968 cgSNPs and 1977 cgSNPs respectively).

The 51 isolates sequenced were obtained from single samples from 48 different horses, three of which had two isolates sequenced each (**Table S4. 1**). These three horse samples were from the respiratory disease, respiratory healthy and uterus groups (two isolates sequenced from each). Sample WK16140 was isolated from a horse with nasal discharge, isolates two and three were sequenced as each had different *szp* and MLST types. Sample WK15114 was the only uterus sample where multiple isolates were obtained that had two different strain types rather than clonal expansion of a single strain (Chapter 2). These strains had differing *szp* and MLST types (isolates one and two). The healthy sample, KMW006, had isolates 31a and 32 sequenced as both had different *szp* types yet had the same MLST type (ST-348) (section 2.3.4, **Figure 2.8**). The KMW006_31a and KMW006_32 isolates from this healthy sample were clustered together in the cgMLST analysis with only 20 SNP differences (**Figure 4.1**). These two isolates are therefore closely-related and potentially could have evolved from the other via mutation instead of being acquired from a different source. This scenario would fit with carriage of *S. zooepidemicus* in healthy individuals and random mutations acquired over time. The other pairs of isolates from the respiratory disease and uterus samples were not found clustered together but rather on distant branches (**Figure 4.1**). The WK15114 uterus isolates have 26,205 pairwise cgSNPs between them and WK16140 respiratory disease isolates have 27,207.

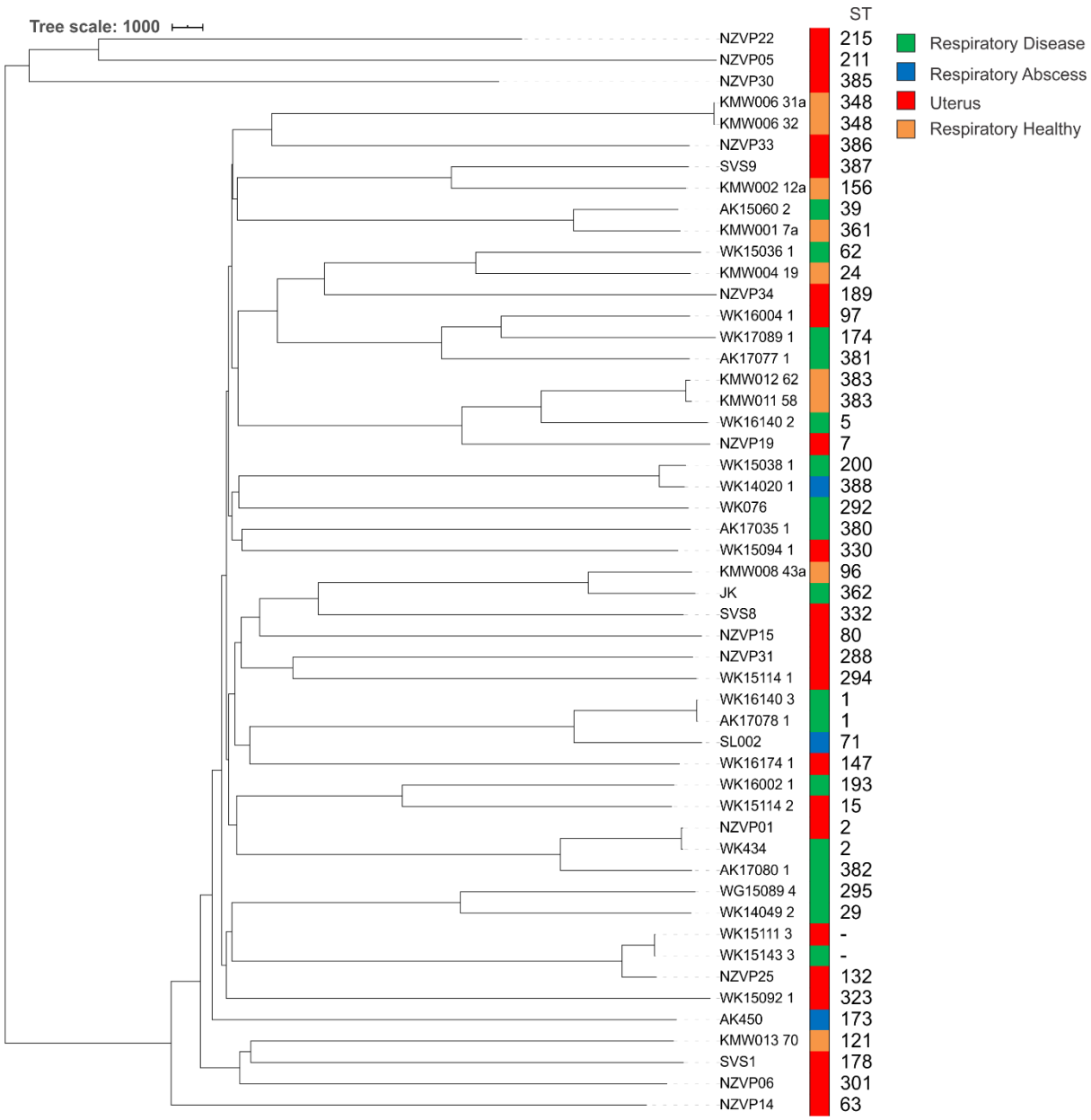


Figure 4.1: New Zealand horse *S. zooepidemicus* isolates core genome Multilocus Sequence Type (cgMLST) phylogenetic tree of 1286 reference core genes. The dendrogram was constructed on cgMLST scores using the Analysis of Phylogenetics and Evolution (ape) package (Paradis *et al.*, 2004) with neighbour-joining implementation. Tree is midpoint rooted using the phangorm package (Schliep, 2011). Annotation was done within interactive Tree of Life (iTOL) (Letunic & Bork, 2019). Scale bar represents core genome single nucleotide polymorphisms (cgSNPs).

The cgMSLT tree was recreated with the addition of 48 global representative of *S. zooepidemicus* isolates, which had available genome sequencing data in NCBI, giving a total of 99 isolates (**Figure 4.2**). The average pairwise cgSNPs was 26,145.7 SNPs (SD 4839.8, min

0, max 39,758). New Zealand isolates were scattered throughout clustering with UK, USA and Denmark isolates and consisted of horse, dog, human, pig, and cat hosts (**Figure 4.2**). This scattering of New Zealand isolates in the global cgMLST is consistent with the global MLST typing (section 2.3.4, **Figure 2.7**); again providing greater resolution between isolates with the same seven gene MLST profiles (STs) (**Figure 4.2** and **Table 4.2**). Examples of this increased resolution are outlined in **Table 4.2** and discussed below.

The ISU38408 USA horse isolate is missing the yqiL allele and clustered with both the New Zealand ST-132 and New Zealand yqiL deleted isolates with a range pairwise cgSNPs of 1064 - 2701 between them. The arcC allele was unable to be defined for SzAM60, a USA horse placenta isolate, but otherwise matches ST-301. It clusters with the New Zealand ST-301, NZP06 uterine isolate, with 530 pairwise cgSNPs. The greater capacity of resolving strain differentiation is further highlighted with ST-72 NCTC11606 (human isolate) mapping closer to ST-405 ISU3775 (goat isolate) with 137 cgSNPs than with ST-72 MGCS10565 (human isolate) with 215 cgSNPs. The ST-5 cluster contained a New Zealand respiratory disease horse isolate and an isolate from a fatal human case of *S. zooepidemicus* kidney infection with 248 pairwise cgSNPs. As identified elsewhere (Chen *et al.*, 2020), pig strains from the USA and China clustered together (ST-194) and clustered closely to guinea pig isolates related to human cases (ST-407) (**Figure 4.2**). Another pig isolate (ST-340), however, was distantly related (**Figure 4.2**).

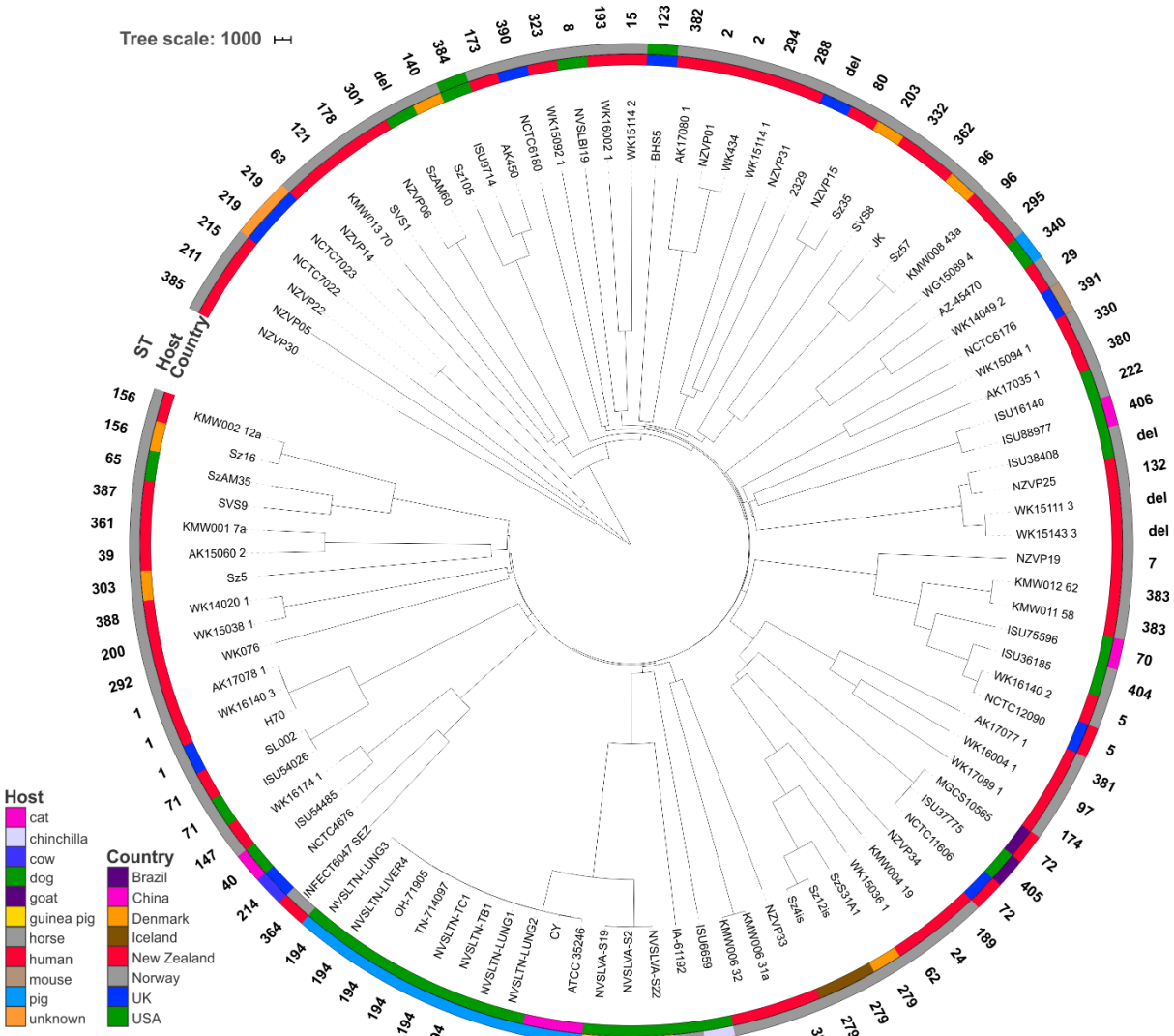


Figure 4.2: Global isolates core genome Multilocus Sequence Type (cgMLST) phylogenetic tree of 1286 reference core genes. The dendrogram was constructed on cgMLST scores using the Analysis of Phylogenetics and Evolution (ape) package *(Paradis et al., 2004)* with neighbour-joining implementation. Tree is midpoint rooted using the phangorm package *(Schliep, 2011)*. Annotation was done within the interactive Tree of Life (iTOL) *(Letunic & Bork, 2019)*. Scale bar represents core geonome single nucleotide polymorphisms (cgSNPs).

Table 4.2: Pairwise core genome single nucleotide polymorphisms (cgSNP) between *S. zooepidemicus* isolates with the same Multilocus (MLST) profile.

MLST	Pairwise cgSNP (range)	Isolate	Country	Animal	Source	Disease
ST-1	19-44	AK17078_1	New Zealand	horse	nasal	respiratory disease
		WK16140_3	New Zealand	horse	nasal	respiratory disease
		H70	UK	horse	nasal	healthy
ST-2	5	WK434	New Zealand	horse	nasal	respiratory disease
		NZVP01	New Zealand	horse	uterus	
ST-5	248	WK16140_2	New Zealand	horse	nasal	respiratory disease
		NCTC12090	UK	human	blood	fatal-glomerulonephritis
ST-71	32	SL002	New Zealand	horse	abscess	respiratory abscess
		ISU54026	USA	horse	tissue	
ST-72/ ST-405	132-215	NCTC11606	UK	human	throat	nephritis
		MGCS10565	Brazil	human	throat	nephritis
		ISU37775	USA	goat	tissue	
ST-96	401	KMW008_43a	New Zealand	horse	nasopharyngeal	apparently healthy
		Sz57	Denmark	horse	vagina	
ST-132/ yqiL deletion	1064-2701	NZVP25	New Zealand	horse	uterus	
		ISU38408	USA	horse	tissue	
		WK15111_3	New Zealand	horse	uterus	
		WK15143_3	New Zealand	Horse	Nasopharyngeal	respiratory disease
ST-156	873	KMW002_12a	New Zealand	horse	nasopharyngeal	apparently healthy
		Sz16	Denmark	horse	vagina	
ST-194	0-57	NVSLTN-TC1	USA	pig	tissue	
		CY	China	pig	milk	
		TN-714097	USA	pig	tissue	
		NVSLTN-LUNG3	USA	pig	tissue	
		NVSLTN-TB1	USA	pig	tissue	
		NVSLTN-LUNG2	USA	pig	tissue	
		ATCC 35246	China	pig		fatal
		NVSLTN-LIVER4	USA	pig	tissue	
		NVSLTN-LUNG1	USA	pig	tissue	
		OH-71905	USA	pig	tissue	
ST-219	5	NCTC7022	UK			
		NCTC7023	UK			
ST-279	104-3431	Sz4is	Iceland	horse	uterus	
		Sz12is	Iceland	horse	uterus	dormant mare
		SzS31A1	Denmark	horse	reproductive	endometritis
ST-301/ arcC undefined	530	NZVP06	New Zealand		uterus	
		SzAM60	USA	horse	Placenta	
ST-348	20	KMW006_31a	New Zealand	horse	nasopharyngeal	apparently healthy
		KMW006_32	New Zealand	horse	nasopharyngeal	apparently healthy
ST-383	294	KMW012_62	New Zealand	horse	nasopharyngeal	apparently healthy
		KMW011_58	New Zealand	horse	nasopharyngeal	apparently healthy
ST-407	0-1	NVSLVA-S19	USA	guinea pig	tissue	Associated with human infection
		NVSLVA-S2	USA	guinea pig	tissue	
		NVSLVA-S22	USA	guinea pig	tissue	

4.3.3 Pangenome

To further investigate the genome diversity of *S. zooepidemicus* infecting New Zealand horses, a pangenome was determined based on 95 % amino acid identity between the 51 New Zealand isolates. A total of 6272 genes were identified in the New Zealand pangenome (**Figure 4.3**). The core genome contained 1386 gene clusters, made up of 1280 genes present in all 51 isolates (99 %) plus 110 soft core genes defined as those found in 95 % and less than 99 % (genes common in 49-50 isolates). Although different from cgMLST, which used a pre-defined core reference genome of 1286 genes, the number of genes defining the core were similar. The accessory genome was divided into the shell with 853 genes (≤ 15 % and < 95 %, 8-48 isolates) and the cloud with 4,029 genes (< 15 %, 1-7 isolates).

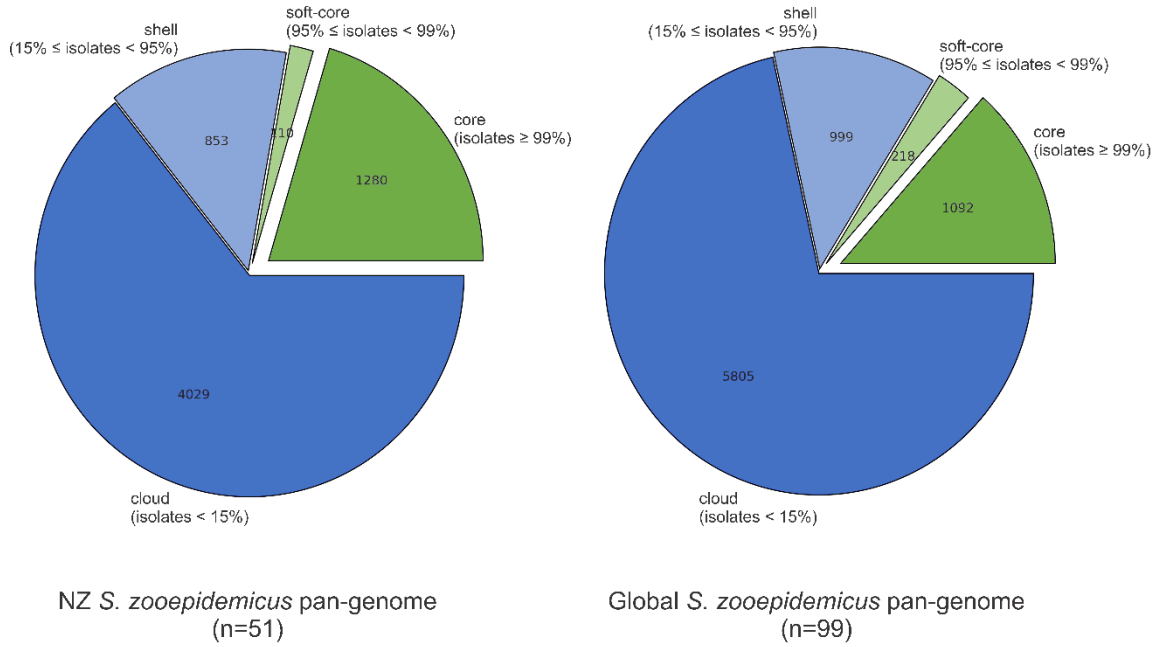


Figure 4.3: New Zealand and global *S. zooepidemicus* pangenomes graphical representation. Green pie sections represent the core genome (genes in ≥ 99 % of the isolates) and the soft-core (genes in ≥ 95 % and < 99 % of the isolates). Blue pie sections represent the accessory genome divided into the shell (genes in ≥ 15 % and < 95 % of the isolates) and the cloud (genes in < 15 % of the isolates). Analysis was computed using the Roary pipeline (v3.8.2) with a minimum percentage identity of 95 % for amino acid sequence comparisons (Page *et al.*, 2015; Sitto & Battistuzzi, 2019).

The pangenome of these 51 isolates is open (fitting curve exponent B > 0, B = 0.42) with the addition of each genome resulting in the discovery of new genes (**Figure 4.4**). The pangenome pool increases by approximately 45 new genes with each new genome added (solving for y in the fitting equation) (**Figure 4.5**), which is consistent with the pangenome created including the global *S. zooepidemicus* representatives (**Figure 4.3**). The pangenome increased by 1842 gene clusters with the inclusion of 48 global isolates, an addition of approximately 38 new genes with each isolate. The global core genome (≥ 95 % of isolates) contained 1310 genes, similar in size to the New Zealand only core genome (**Figure 4.3**). Of the global core genes, 1092 were in 99 % of the global isolates (**Figure 4.3**).

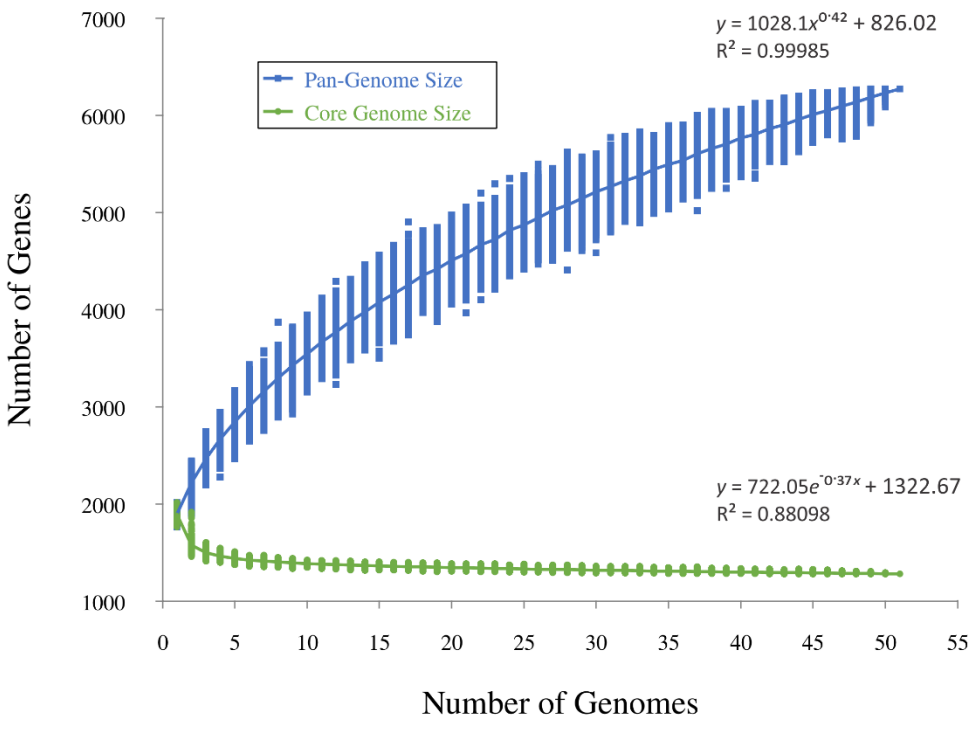


Figure 4.4: Core and pangenome plots of New Zealand 51 *S. zooepidemicus* isolates using the PanGP tool (Zhao *et al.*, 2014) with the distance guide algorithm drive. The dot plots represent the range of a sampling size set to sample 500 different combinations based on the genome diversity of the isolates with the amplification coefficient set to 10 and then repeated 10 times. The continuous curve was fitted by PanGP based on Heaps' law (Tettelin *et al.*, 2005; Rasko *et al.*, 2008; Tettelin *et al.*, 2008).

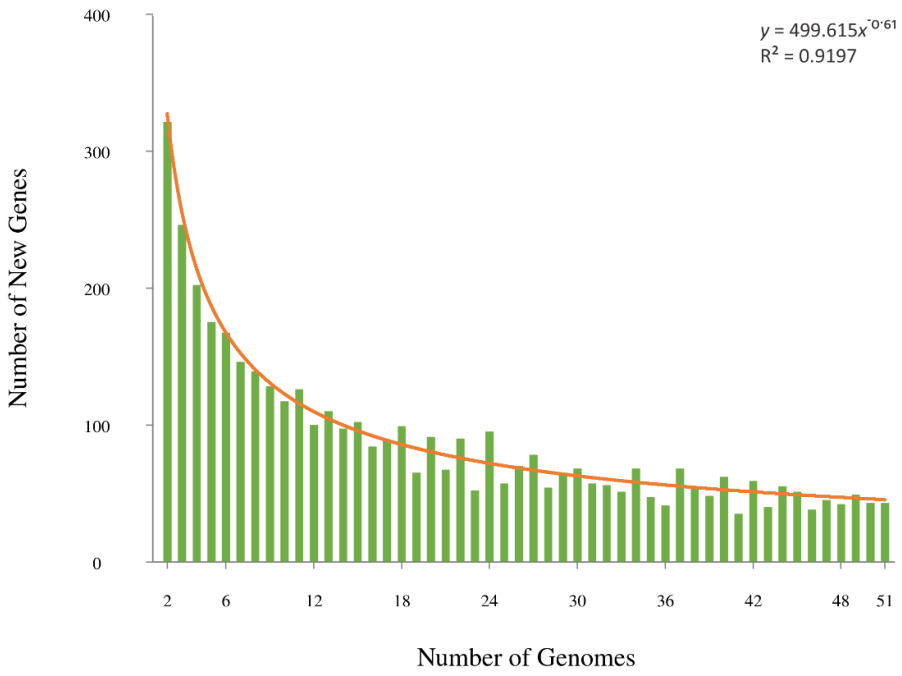


Figure 4.5: New genes plot based on the New Zealand 51 *S. zooepidemicus* isolates using the PanGP tool (Zhao *et al.*, 2014). Green bars represent the number of new isolate-specific genes added to the pangenome pool as each isolate is added. The continuous curve was fitted by PanGP with the average number of new genes converging at 45.

Each New Zealand isolate genome contained an average of 1896 genes with 508 genes making up their accessory genome, 43 of which are unique to the isolate (**Table 4.3**). Again, similar to the 31.4 % of genes in the accessory genome determined from the cgMLST (min 27.7 %, max 34.4 %), 26.8 % (min 19.7 %, max 33.2 %) of the New Zealand isolate genomes were accessory. Genome size remained consistent across isolates with no significant difference between the New Zealand isolation source groups and the number of genes in the isolates core genomes ($p = 0.172$), accessory genomes ($p = 0.730$), and their number of unique genes ($p = 0.439$) or total genes ($p = 0.681$) (**Table 4.4**). Although statistical analysis could not be performed on the global genomes with respect to host species due to some only having one representative isolate, a similar pattern can be seen with the ranges of the host representative isolates core genomes, accessory genomes, unique genes, and total genes (**Table 4.4**).

Table 4.3: Genome breakdown based on isolation source groups of the New Zealand *S. zooepidemicus* genomes. Numbers expressed as medians ± standard deviation with the minimum-maximum range in brackets. Core genes consist of genes found in at least 95 % of the genomes (core and soft genome). The accessory genes are found in less than 95 % of the genomes (shell and cloud genome).

New Zealand *S. zooepidemicus* genomes	**Total (n=51)**	**Respiratory Disease (n=17)**	**Respiratory Abscess (n=3)**	**Respiratory Healthy (n=9)**	**Uterine (n=22)**
Core genes (core+soft)	1389 ± 4.7 (1368-1390)	1390 ± 2.4 (1382-1390)	1389 ± 5.9 (1377-1390)	1389 ± 1.2 (1386-1390)	1388 ± 6.0 (1368-1390)
Accessory genes (shell+cloud)	506 ± 65.4 (373-629)	486 ± 66.1 (378-626)	574 ± 83.0 (406-589)	520 ± 45.0 (448-600)	495 ± 67.5 (373-629)
Total genes	1884 ± 65.1 (1762-2016)	1875 ± 65.9 (1768-2016)	1964 ± 80.1 (1795-1966)	1910 ± 44.8 (1837-1988)	1880 ± 67.3 (1762-2013)
Unique genes /isolate	29 ± 43.7 (0-189)	20 ± 50.0 (1-189)	50 ± 19.8 (14-60)	19 ± 28.9 (0-82)	36 ± 44.2 (2-164)

Table 4.4: Genome breakdown based on isolate host species of the global *S. zooepidemicus* genomes. Numbers expressed as medians ± standard deviation with the minimum-maximum range in brackets. Core genes consist of genes found in at least 95 % of the genomes (core and soft genome). The accessory genes are found in less than 95 % of the genomes (shell and cloud genome).

Global genomes	**Total (n=99)**	**Cat (n=3)**	**Chin-chilla (n=1)**	**Cow (n=1)**	**Dog (n=2)**	**Goat (n=1)**	**Guinea Pig (n=3)**	**Horse (n=70)**	**Human (n=4)**	**Mouse (n=1)**	**Pig (n=11)**	**Un-known (n=2)**
Core genes (core+soft)	1326 ± 5.5 (1296-1329)	1318 ± 6.5 (1310-1326)	1324	1325	1327 ± 1.5 (1325-1328)	1323	1326 ± 0.5 (1326-1327)	1326 ± 5.7 (1296-1329)	1324 ± 3.3 (1320-1329)	1327	1329 ± 1.1 (1326-1329)	1315 ± 0.0 (1315-1315)
Accessory genes (shell+cloud)	578 ± 73 (432-763)	570 ± 108.5 (500-757)	581	566	556 ± 72.5 (483-628)	432	674 ± 3.7 (671-680)	561 ± 71.4 (434-763)	565 ± 45.2 (524-622)	549	614 ± 35.9 (526-663)	693 ± 9.5 (683-702)
Total gene	1904 ± 71.7 (1755-2087)	1888 ± 102.2 (1826-2067)	1905	1891	1882 ± 71.0 (1811-1953)	1755	2001 ± 3.7 (1997-2006)	1887 ± 70.4 (1760-2087)	1892 ± 46.7 (1844-1945)	1876	1943 ± 36.6 (1852-1992)	2008 ± 9.5 (1998-2017)
Unique genes /isolate	12 ± 44.7 (0-252)	25 ± 108.2 (20-252)	71	9	65 ± 16.5 (48-81)	5	1 ± 1.7 (0-4)	21 ± 42.9 (1-231)	10 ± 5.2 (1-15)	11	2 ± 6.8 (0-25)	7 ± 6.0 (1-13)

4.3.4 Gene Distribution of *S. zooepidemicus* Pangenome

A gene presence/absence matrix created in Roary visualized the distribution of the pangenome amongst isolates. The accessory gene distribution mirrors the cgMLST phylogeny, with similar banding patterns seen in low cgSNPs isolate clusters and otherwise random banding patterns across all other isolates representing the high diversity in the phylogeny (**Figure 4.6** and **Figure 4.7**). The similar banding pattern is most obvious in the pig cluster (**Figure 4.7**).

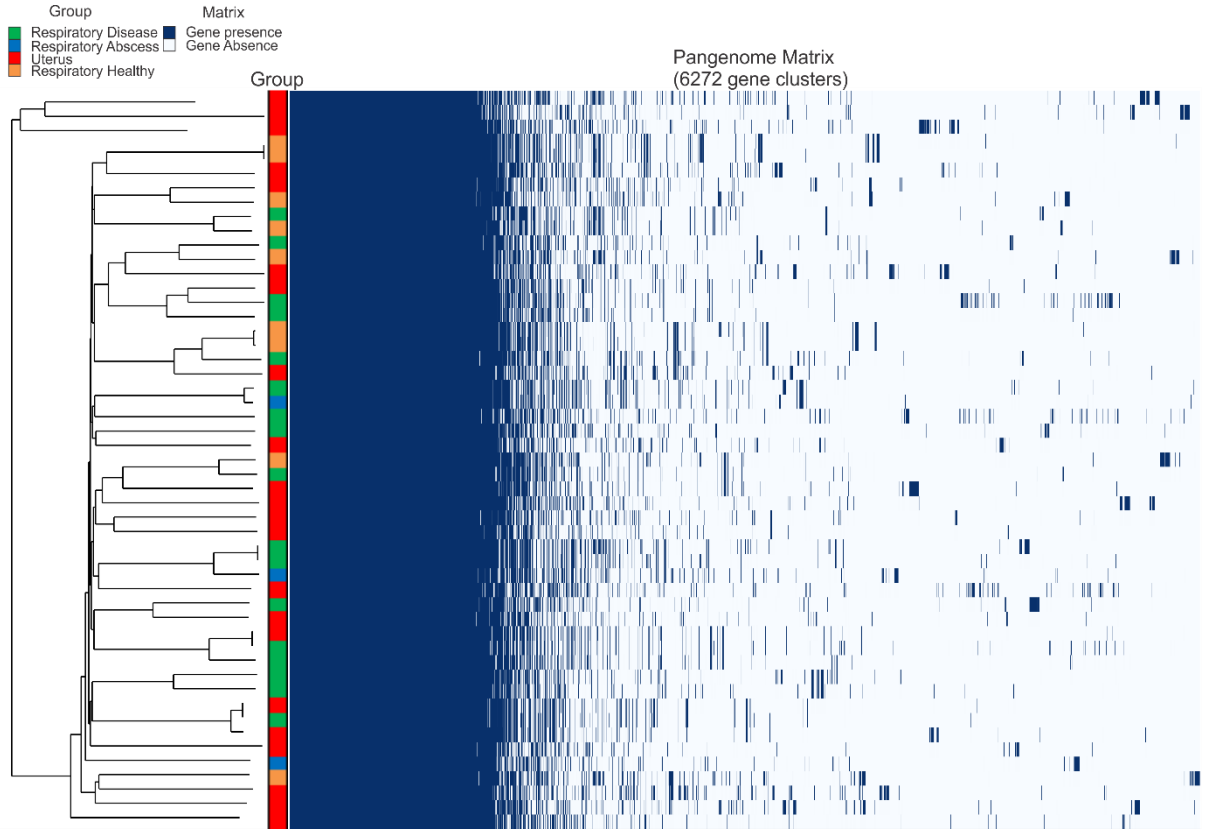


Figure 4.6: New Zealand pangenome matrix and cgMLST. New Zealand horse *S. zooepidemicus* (n=51) pangenome matrix plotted against the core genome multilocus sequence typing (cgMLST) dendrogram. Matrix was created in the Roary pipeline (v3.8.2) (Page *et al.*, 2015; Sitto & Battistuzzi, 2019), visualizing the presence/absence of the 6272 gene clusters of the New Zealand horse *S. zooepidemicus* pangenome.

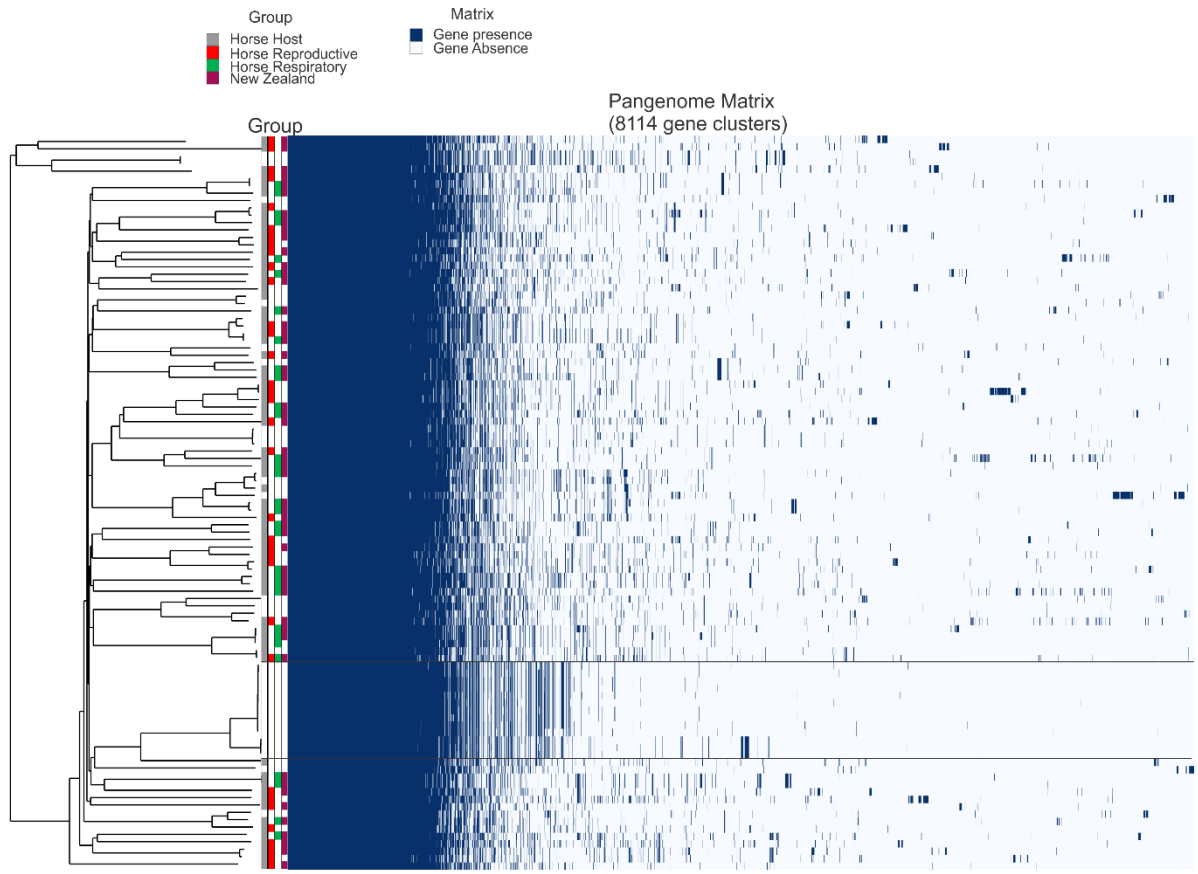


Figure 4.7: Global pangenome matrix and cgMLST. Global *S. zooepidemicus* (n=99) pangenome matrix plotted against the core genome multilocus sequence typing (cgMLST) dendrogam. Matrix was created in the Roary pipeline (v3.8.2) *(Page et al., 2015; Sitto & Battistuzzi, 2019)*, visualizing the presence/absence of the 8114 gene clusters of the global *S. zooepidemicus* pangenome. Black horizontal lines outline the pig and guinea pig/human cluster matrix.

4.3.4.1 Distribution of Fitness Factors in New Zealand Isolates

The presence/absence of potential fitness determinants in the genomes of the New Zealand isolates was investigated. Thirty-five putative fitness factors were selected from the literature and further investigated, these included factors involved in host immune evasion (antiphagocytic, non-specific T-cell activation, and cytotoxicity) and in adhesion/invasion (including biofilm formation and tissue invasion) (**Table 4.5**).

Table 4.5: Presence of *S. zooepidemicus* putative fitness factors in New Zealand isolates

Gene(s)	NZ isolates (n=51)	Description	Potential Action	Sequence Source	Reference
5Nuc	51	degradation of neutrophil extracellular traps	antiphagocytic	ATCC35246-SESEC_RS05720 (SeseC_01430)	(Ma *et al.*, 2017a)
bifA	1	Fic domain-containing protein	disrupt blood brain barrier/tissue spread	ATCC35246-SESEC_RS05335 (SeseC_01334)	(Ma *et al.*, 2019)
dnaK	51	chaperone protein (heat shock protein 70)	induction of capsule (antiphagocytic)	H70-SZO_RS08225 (SZO_15780)	Velineni et al., 2014
Enuc	51	degradation of neutrophil extracellular traps	antiphagocytic	ATCC35246-SESEC_RS04165 (SeseC_01049)	(Ma *et al.*, 2017a)
fbp	51	Fibronectin/fibrinogen-binding protein	adhesion/invasion	ATCC35246-SESEC_RS06360 (SeseC_01590)	(Kittang *et al.*, 2017)
fim I	51	pili expression loci	adhesion/invasion	MGCS10565-SEZ_RS04035 (Sez_0813), SEZ_RS04030, SEZ_RS04025, SEZ_RS04020 (Sez_0810)	Beres et al., 2008
fim II	20	pili expression loci	adhesion/invasion	MGCS10565-SEZ_RS08895 (Sez_1823) - SEZ_RS08880 (Sez_1820)	Beres et al., 2008
fim III	20	pili expression loci	adhesion/invasion	MGCS10565-SEZ_RS08925 (Sez_1830) - SEZ_RS08905 (Sez_1826)	Beres et al., 2008
fim IV	28	pili expression loci	adhesion/invasion	H70-SZO_RS09505 (SZO_18270) - SZO_RS09540 (SZO_18340)	(Holden *et al.*, 2009)
groEL	51	chaperonin GroEL	biofilm formation/adhesion	H70-SZO_RS00815 (SZO_01280)	Li Yi et al., 2016
htpZ	51	histidine trad protein (HTP)	anitphagocytic	ATCC35246-SESEC_RS03280 (SeseC_00825)	(Xu *et al.*, 2019)
hylZ	51	Beta-N-acetylglucosaminidase/ Hyaluronidase	invasion/biofilm formation	ATCC35246-SESEC_RS07890 (SeseC_02032)	(Kittang *et al.*, 2017)
ideZ	47	immunoglobulin G (IgG) endopeptidases	anitphagocytic	ATCC35246-SESEC_RS04610 (SeseC_01159)	(Lannergård & Guss, 2006)
sagB	51	Streptolysin S biosynthesis protein	cytotoxicity, activation inflammatory response, and antiphagocytic	ATCC35246-SESEC_RS02415 (SeseC_00590)	(Kittang *et al.*, 2017)
sagC	51	Streptolysin S biosynthesis protein	cytotoxicity, activation inflammatory response, and antiphagocytic	ATCC35246-SESEC_RS02420 (SeseC_00591)	(Kittang *et al.*, 2017)
sagD	51	Streptolysin S biosynthesis protein	cytotoxicity, activation inflammatory response, and antiphagocytic	ATCC35246-SESEC_RS02425 (SeseC_00592)	(Kittang *et al.*, 2017)
sagG	51	Streptolysin S biosynthesis protein	cytotoxicity, activation inflammatory response, and antiphagocytic	ATCC35246-SESEC_RS02440 (SeseC_00595)	(Kittang *et al.*, 2017)

Table 4.5 continued …….

Gene(s)	NZ isolates (n=51)	Description	Potential Action	Sequence Source	Reference
sagH	51	Streptolysin S biosynthesis protein	cytotoxicity, activation inflammatory response, and antiphagocytic	ATCC35246-SESEC_RS02445 (SeseC_00596)	(Kittang *et al.*, 2017)
sagI	51	Streptolysin S biosynthesis protein	cytotoxicity, activation inflammatory response, and antiphagocytic	ATCC35246-SESEC_RS02450 (SeseC_00597)	(Kittang *et al.*, 2017)
scpC/cepA	48	Serine endopeptidase, lactocepin, interleukin-8 protease-like protein	anitphagocytic	ATCC35246-SESEC_RS02495 (SeseC_00614)	(Kittang *et al.*, 2017)
scpZ	51	C5a peptidase	invasion	MGCS10565-SEZ_RS00005 (Sez_0001)	Wei et al., 2013
sdzB	51	Streptodornase type B	nuclease (antiphagocytic)	ATCC35246-SESEC_RS05625 (SeseC_01403)	(Kittang *et al.*, 2017)
sdzD	23	Streptodornase type D	nuclease (antiphagocytic)	ATCC35246-SESEC_RS03220 (SeseC_00808)	(Kittang *et al.*, 2017)
sfs	36	fibronectin-binding protein	adhesion/invasion	ATCC35246-SESEC_RS01955 (SeseC_00464)	Ma et al., 2013
skc_1	51	Streptokinase; plasminogen	plasminogen activating protein	MGCS10565-SEZ_RS08665 (Sez_1775)	(Kittang *et al.*, 2017)
slaA	10	Phospholipase A2	adhesion/invasion	Se4047-SEQ_RS04145 (SEQ_0849)	(Holden *et al.*, 2009)
slaB	51	putative Phospholipase A2	adhesion/invasion	H70-SZO_RS09705 (SZO_18670)	(Holden *et al.*, 2009)
spaZ	2	Protective antigen-like protein, fibrinogen-and Ig-binding protein	antiphagocytic	ATCC35246-SESEC_RS00915 (SeseC_00180)	(Kittang *et al.*, 2017)
spuZ	51	secreted enzyme	anitphagocytic	ATCC35246-SESEC_RS06875 (SeseC_01740)	(Xu *et al.*, 2019)
szeF	18	superantigen-encoding genes	non-specific T-cell activation	BHS5_01670, KC906582	Paillot et al., 2010
szeL	1	superantigen-exotoxin L	non-specific T-cell activation	AJ851369	(Alber *et al.*, 2005)
szeM	1	superantigen-exotoxin M	non-specific T-cell activation	AJ851826	Alber et al., 2005
szeN	13	superantigen-encoding genes	non-specific T-cell activation	BHS5_04190, KC906584	Paillot et al., 2010
szeP	12	superantigen-encoding genes	non-specific T-cell activation	BHS5_04180, KC906583	Paillot et al., 2010
szM	3	M-like protein	Antiphagocytic, plasminogen binding	H70-SZO_RS01125 (SZO_01900)	Velineni & Timoney, 2013a
vapE	1	virulence-associated protein E	Increased virulence, unknown function	ATCC35246-SESEC_RS05300 (SeseC_01325)	Ma et al., 2013

Seventeen of the putative fitness factors were absent in some isolates and hence part of the accessory genome of *S. zooepidemicus* (**Table 4.5** and **Figure 4.8**). The pattern of presence or absence of these factors did not follow the cgMLST phylogeny indicative of potential HGT with scattered phylogenetic distributions. Although, this could also be caused by gene loss or rapid sequence divergence (Eisen, 2000). One exception was *szM* that was identified in the cluster with two respiratory disease isolates (WK16140_3 and AK17078_1) and one respiratory abscess isolate (SL002), hence appearing to be lineage-specific. This type of restricted phylogenetic distribution could also be explained by horizontal gene transfer with the gene not being inherited from ancestral strains (Ochman *et al.*, 2000).

The three horse samples with two isolates sequenced from each (discussed earlier) were analysed for the presence/absence of the putative fitness factors. As expected, being distantly related in the cgMLST phylogeny, the WK16140 respiratory isolates contained different putative fitness factors (**Figure 4.8**). Whereas, the WK15114 uterus isolates were more similar with the only difference in putative fitness factors was the presence of the *szeF* superantigen. The cgMLST clustered isolates from the healthy KMW006 sample were identical, containing the same putative fitness factors (**Figure 4.8**).

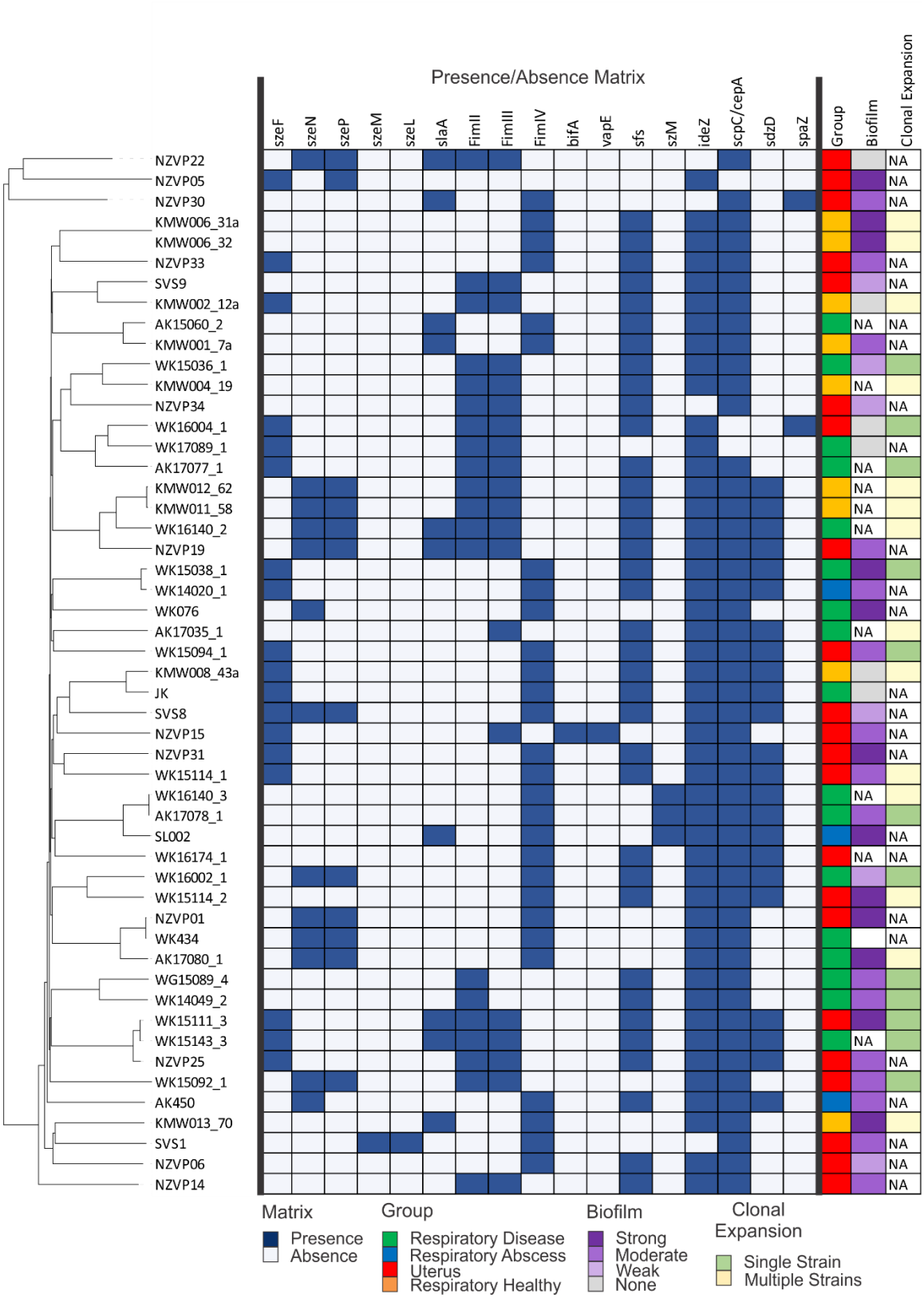


Figure 4.8: Seventeen putative fitness factors with variant occurrences among the 51 New Zealand *S. zooepidemicus* isolates, ordered against the cgMLST phylogeny. NA indicates where isolates were not tested for the particular trait so data is not available.

Each of these seventeen factors were tested for significant prevalence differences within different groupings of the New Zealand isolates. Analyses were performed on four different groupings: 1) isolate source groups being respiratory disease, respiratory healthy, respiratory abscess, and uterus; 2) anatomical source group either respiratory or uterine isolates regardless of health status; 3) clonal expansion group where the isolate was one of multiple isolated from a single sample all with the same single ST (clonally expanded) versus multiple STs isolated from a single sample as identified in the second chapter; and 4) grouped based on biofilm forming capabilities as identified in the third chapter. Isolates that were not tested for biofilm formation or clonal expansion were excluded from these analyses.

The antiphagocytic *szM* gene, found in a cgMLST cluster of three isolates as mentioned above, was found to be associated with the abscess group ($p = 0.036$). It was identified in 33.3 % (n = 1/3) of the New Zealand respiratory abscess isolates, in 11.8 % (n = 2/17) of respiratory disease isolates, but in none of the respiratory healthy (n = 0/9) or uterus (n = 0/22) isolates (**Figure 4.8**). The presence of superantigens (streptococcal pyogenic exotoxins) were previously associated with non-strangles lymph node abscessation (Paillot *et al.*, 2010b). Although, with only three sequenced abscess isolates in the study, there was no significant difference in the prevalence of superantigens across the different isolation groups ($p = 0.704$).

Superantigens, however, were associated with the inability of an isolate to form a biofilm. All of the biofilm negative isolates with genomes sequenced (n = 6/6), inclusive of those unable to maintain a biofilm, were positively associated for the presence of at least one of the superantigens compared to 55.9 % of biofilm positive isolates (n = 19/34) ($p = 0.046$). In particular, the presence of the *szeF* gene was associated with 83.3 % of biofilm negative isolates (n = 5/6) compared to 32.4 % of biofilm positive isolates (n = 11/34) ($p = 0.029$) (**Figure 4.8**). Truncated forms of *szeF* existed in the NZ isolates, KMW002_12a had a single nucleotide deletion at nucleotide position 139 as well as a 10 bp insertion (between nucleotide position 321-322) resulting in a direct repeat of AAGCGCTATA. Six other isolates (NZVP15, NZVP33, WK15094_1, WK16004_1 and WK17089_1) had the same insertion sequence, resulting in a nonsense mutation due to a frameshift. Taking into account the truncation and the likely resultant non-functionality of *szeF*, there was no longer any association with its presence in the biofilm negative isolates (n = 2/6) over the biofilm positive isolates (n = 8/34), a 33.3 %

and 23.5 % presence respectively (p = 0.474). Previous studies found SzeF less potent at stimulating equine peripheral blood mononuclear cells than other superantigens however, it was speculated that its role may be in increasing the ability of *S. zooepidemicus* to infect an extended host range (Paillot *et al.*, 2010b). No other associations were identified with the selected putative fitness factors analysed in this study (**Table 4.5**).

4.3.4.2 Presence of Toxin-Antitoxin Systems

The presence of toxin-antitoxin (TA) systems in the *S. zooepidemicus* New Zealand genomes was investigated based on their proposed roles as stress response modules and in the formation of persister cells (Fraikin *et al.*, 2020), thereby creating a bet-hedging fitness advantage during environmental stress and antibiotic tolerance. TA systems are known to be abundant in bacterial genomes and have been identified on genomic islands in *S. zooepidemicus* ATCC35246 from a pig (Ma *et al.*, 2013). Six potential TA systems of various types were identified in the Prokka annotations of the 51 *S. zooepidemicus* New Zealand genomes and verified by searching the amino acid sequence against the EMBL-EBI InterPro database of protein families (https://www.ebi.ac.uk/interpro/) or by running a BLAST search against the UniProt Protein Knowledgebase (https://www.uniprot.org/blast/) (**Table 4.6**). Five were type II TA systems, a Pez TA system and four individual ParE toxins (**Table 4.6** and **Figure 4.9**). The ParE toxins each had an associated hypothetical protein identifying as potential candidate antitoxins (**Table 4.6**). The sixth identified TA system was a type IV, the AbiEii/AbiE_4 TA (**Table 4.6** and **Figure 4.9**). The presence and absence of these TA systems varied between isolates with an average of two per genome (SD = 1.1, min = 0, max = 5). However, being only a preliminary search, the *S. zooepidemicus* genomes will contain many more candidate TA systems. For instance, the 13 *S. zooepidemicus* genomes present in TASmania, a database of bacterial genomes with annotated candidate TA systems (Akarsu *et al.*, 2019), contained an average of 46 (SD ± 6.4) candidate TA operons each.

The distribution of the six annotated putative TA systems in the New Zealand genomes were not phylogenetically restricted (**Figure 4.9**). However, the presence of the third ParE toxin was significantly higher in the respiratory healthy group than with any other group (p = 0.025). It was found in 55.5 % of the respiratory healthy isolates (n = 5/9), in none of the respiratory abscess isolates (n = 0/3), in 11.8 % of the respiratory disease (n = 2/17) and in 9.1 % of the

uterus isolates (n = 2/22) (**Figure 4.9**). No other significant associations were found with the other five TA systems. Further, there was no association between the level of biofilm formed by an isolate and the TA system present. The epsilon/zeta TA found in the ATCC35246 pig isolate (Ma *et al.*, 2013) was identified in nine of the 51 New Zealand isolates (**Figure 4.9**). These nine New Zealand isolates shared between 88.7 % to 99.4 % nucleotide identity with the pig zeta toxin (SeseC_01875) and 95.4 % to 100 % nucleotide identity for the pig epsilon antitoxin (SeseC_01876).

The three horse samples having two isolates sequenced each, as discussed earlier were analysed for the presence of TA systems. The WK16140 respiratory isolate two (WK16140_2) contains one of the annotated TA systems whereas the WK16140_3 contained five of the six TA systems (**Figure 4.9**). The two WK15114 uterus isolates contained one and two TA systems respectively. Again, as expected, the cgMLST related isolates from the healthy KMW006 sample were identical, containing the same two TA systems (**Figure 4.9**).

Table 4.6: Putative toxin-antitoxin (TA) systems annotated in the 51 New Zealand *S. zooepidemicus* genomes.

	Putative TA system gene annotation within New Zealand *S. zooepidemicus* genomes	TA type	*S. zooepidemicus* reference sequence ID	*S. zooepidemicus* reference strain	Toxin	Antitoxin	*InterPro class	*Pfam class	*UniProtKB reference (aa % identity)	# NZ isolates
	antitoxin PezA	II	SESEC_RS07345/SeseC_01876	Sz35246		helix-turn-helix	IPR001387	Clan: HTH		9
	toxin PezT		SESEC_RS07340/SeseC_01875		Zeta		IPR010488	PF06414		9
1	type II toxin-antitoxin system RelE/ParE family toxin	II	SZO_RS09880/SZO_19050	H70	ParE		IPR007712	PF05016		38
	hypothetical protein		SZO_RS09885/SZO_19060			RelB	IPR007337	PF04221		38
2	type II toxin-antitoxin system RelE/ParE family toxin	II	SZO_RS06175/SZO_11810	H70	ParE		IPR007712	PF05016		32
	hypothetical protein		SZO_RS06180/SZO_11820			RelB	undefined	undefined	A0A380KCN4 *Streptococcus hyointestinalis* (97.3 %)	32
3	type II toxin-antitoxin system RelE/ParE family toxin	II	NCTC4675_ 00902	NCTC4675	ParE		IPR007712	PF05016		9
	hypothetical protein		NCTC4675_00901			antitoxin component, ribbon-helix-helix domain protein	undefined	undefined	U2T0U2 *Leptotrichia sp. oral taxon 215 str. W9775* (74.6 %)	9
4	type II toxin-antitoxin system RelE/ParE family toxin	II	SE071780_ 00165	ATCC 39506	ParE		IPR007712	PF05016		5
	hypothetical protein		SE071780_00166			CopG family transcriptional regulator	undefined	undefined	A0A3L9DMJ5 Streptococcus sp. 28462 (42.9 %)	5
	nucleotidyl transferase AbiEii/AbiGii toxin family protein	IV	SZO_RS09285/SZO_17820	H70	AbiEii		IPR014942	PF08843		12
	abortive infection protein		SZO_RS09290/SZO_17830			AbiEi_4	IPR025159	PF13338		12

*The amino acid (aa) sequence of the identified *S. zooepidemicus* reference was used as the input sequence into the EMBL-EBI InterPro database of protein families (InterPro and Pfam class) and the the UniProt Protein Knowledgebase (UniProtKB).

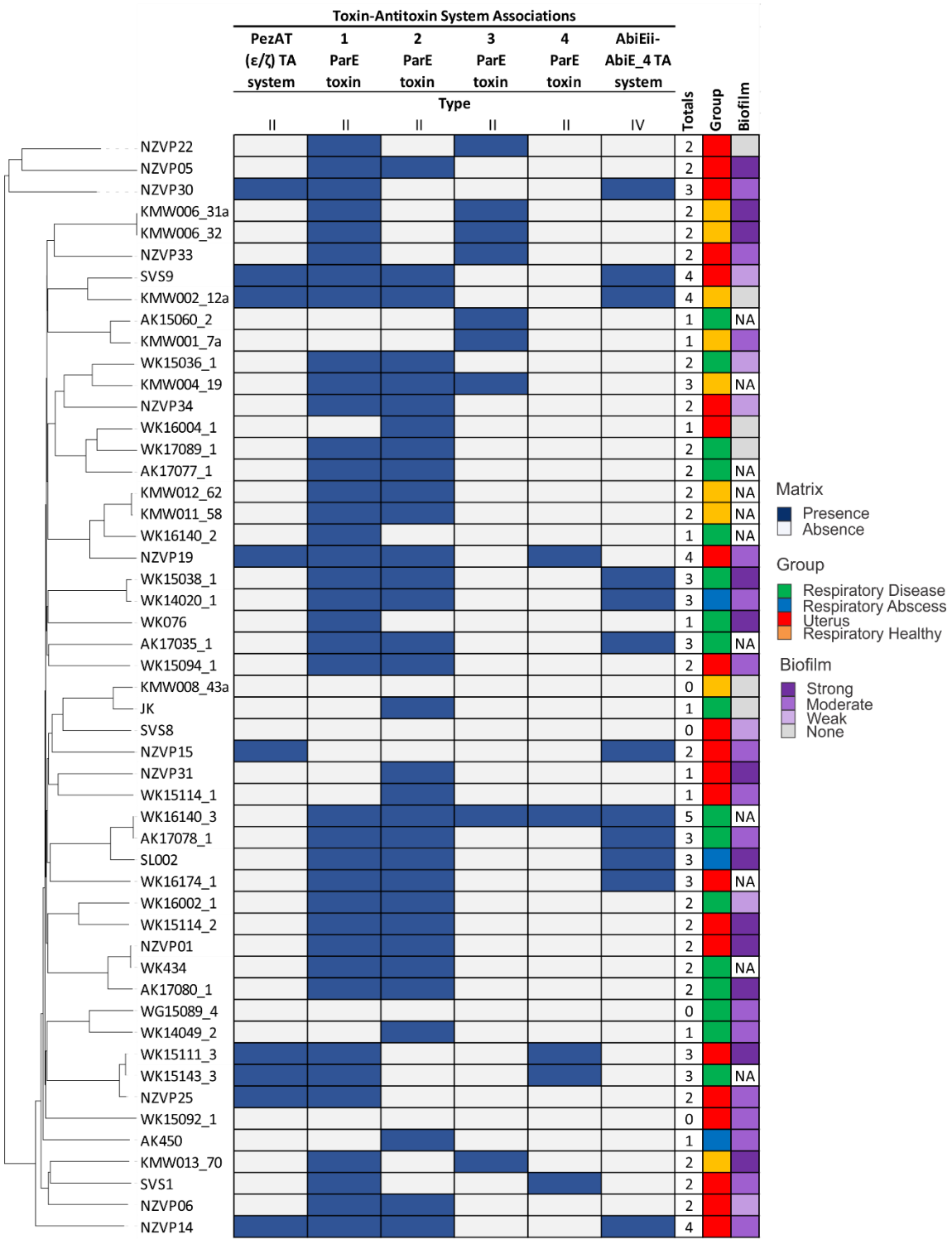


Figure 4.9: Toxin-antitoxin (TA) system associations annotated in the 51 genomes of New Zealand *S. zooepidemicus* isolates ordered against the cgMLST phylogeny. NA indicates where isolates were not tested for the particular trait so data is not available.

4.3.4.3 Hyaluronidase in the Highly Mucoid JK Isolate

Hyaluronidase is a putative virulence factor of Gram-positive bacteria. This secreted enzyme degrades hyaluronic acid, breaking down host connective tissue, facilitating spread of the bacteria and its toxins (Hynes & Walton, 2000; Holden *et al.*, 2009). The *hylC* hyaluronidase gene was present in all 51 New Zealand *S. zooepidemicus* isolates (reference sequence H70 - SZO_06680), and was further investigated as a potential culprit of isolate JK's highly mucoid appearance and impaired invasion and biofilm capabilities. A disruption in this gene has accounted for highly mucoid colonies of another *S. zooepidemicus* isolate, an ST-57 isolate with an IS element inserted 905 bp from the start codon (Holden *et al.*, 2009). JK was found to have an SNP (C → T), at position 229 nt (codon position 77) from the start codon, causing a nonsense mutation and a truncated HylC protein (**Figure 4.10**). *S. equi* has high levels of hyaluronate capsule giving the colonies their highly mucoid appearance. This is also suspected to be caused by a mutation in the *hylC* gene (SEQ_1479) of *S. equi* which contains a four base pair deletion causing a frameshift and a downstream stop codon (Holden *et al.*, 2009) (**Figure 4.10**). *S. equi* has acquired a second hyaluronate lyase encoded on a prophage giving it lower activity and a reduced substrate range (Holden *et al.*, 2009). A second hyaluronate lyase was not identified in the JK isolate.

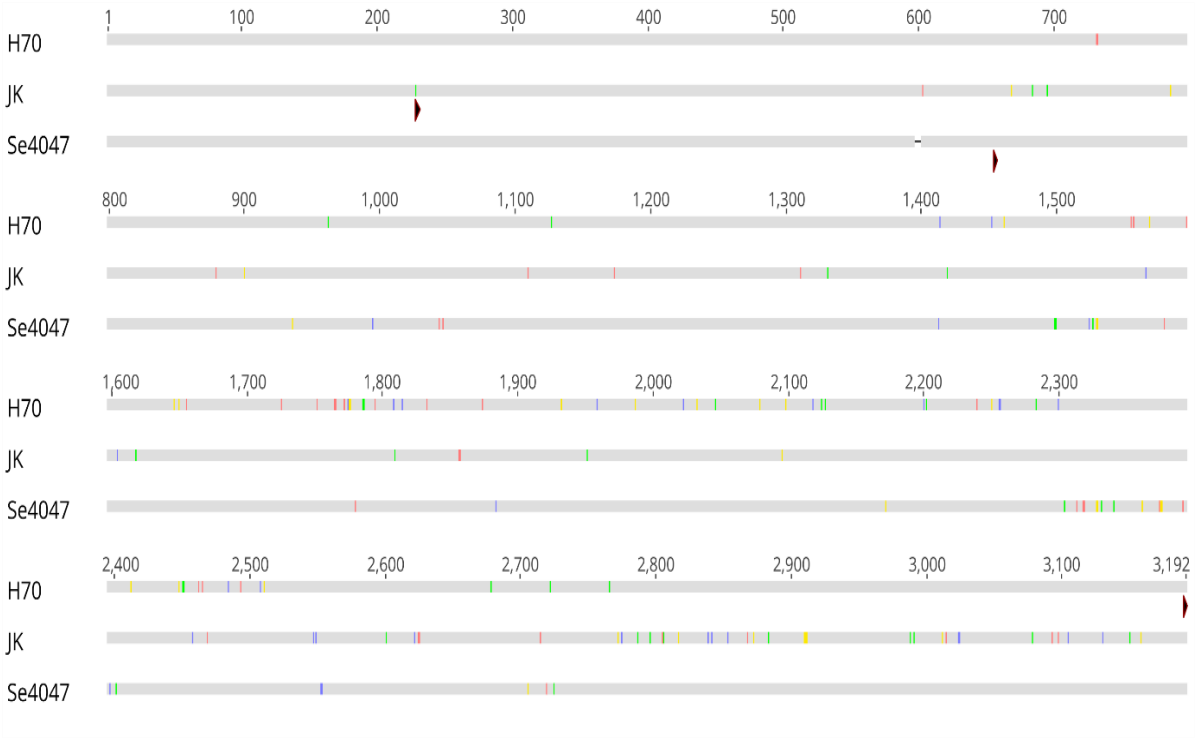


Figure 4.10: Hyaluronidase, *hylC*. A nucleotide alignment of two *S. zooepidemicus* isolates against an *S. equi* isolate, done in Geneious. Grey lines are the coding sequence with coloured bands identifying single nucleotide polymorphisms (SNPs). Position one is the A of the ATG start codon and red arrows indicate stop codons. H70 is a horse reference sequence with an uninterrupted HylC. JK is the highly mucoid isolate with impaired invasion and biofilm abilities. Se4047 is the *S. equi* reference with a four bp deletion (black dash) causing a frameshift mutation resulting in an early stop codon.

4.3.4.4 Bacteriocins in *S. zooepidemicus* Genomes

Bacteriocins were analysed due to their predicted role in giving a competitive fitness advantage over other streptococcal strains through their antimicrobial activity. New Zealand *S. zooepidemicus* isolates were found to contain several putative bacteriocins with a median of seven per isolate (SD ± 1.5). One uterus isolate (WK16174_1) however contained 15 putative bacteriocins (**Table 4.7**). There were approximately three class I and four class II bacteriocins in each isolate **(Table 4.7).** Two uterus isolates (NZVP06 and SVS1) and one respiratory healthy isolate (KMW0013_70) were the only ones to contain a class III bacteriocin. There was no significant difference in the occurrence of bacteriocins between the four different isolation groups.

Table 4.7: Presence of putative bacteriocins in New Zealand *S. zooepidemicus* horse isolates.

		New Zealand *S. zooepidemicus* isolates				
Bacteriocins		**All (n = 51)**	**Respiratory Disease (n = 17)**	**Respiratory Abscess (n = 3)**	**Respiratory Healthy (n = 9)**	**Uterus (n = 22)**
Total	median	7	7.5	8	6	7
	(min-max)	(5-15)	(5-9)	(7-8)	(6-7)	(5-15)
	SD	1.51	1.07	0.47	0.42	2.17
Class I	median	3	3.5	3	2	3
	(min-max)	(1-6)	(1-3)	(1-3)	(1-2)	(1-6)
	SD	1.20	0.85	0.82	0.47	1.65
Class II	median	4	4	5	4	4
	(min-max)	(3-6)	(3-5)	(4-5)	(3-5)	(3-6)
	SD	0.75	0.40	0.47	0.79	0.92
Class III	median	0	0	0	0	0
	(min-max)	(0-1)	(0-0)	(0-0)	(0-1)	(0-1)
	SD	0.24	0	0	0.31	0.32

The BAGEL4-identified putative bacteriocins had amino acid sequence similarities to 31 different bacteriocins within its database showing different patterns of presence/absence among the 51 New Zealand isolates (**Figure 4.11**). Detailed information on these 31 BAGEL4 bacteriocins are found in **Table S4. 3**. Only one bacteriocin, streptolysin, was identified in all isolates. All New Zealand *S. zooepidemicus* isolates with the exception of the uterus isolate NZVP22 also had a protein identified as similar to one of two alleles of the BlpU bacteriocin (bacteriocin like peptide U) in the database (**Figure 4.11**). One of two thermophilin A alleles was also found in the majority of isolates with the exception of three uterus isolates; NZVP30, NZVP05, WK15114_2 (**Figure 4.11**). As with the putative fitness factors, the pattern of presence and absence of the bacteriocins in various strains were indicative of HGT. Most of the bacteriocins had scattered phylogenetic distributions and did not follow the cgMLST phylogeny. However, there were two exceptions, which showed restricted phylogenetic distribution rather than a scattered distribution. Zoocin A was found in the cluster with two uterus isolates (SVS1 and NZVP06) and one respiratory healthy isolate (KMW013_70) (**Figure 4.11**). Penocin A was found in the cluster with two respiratory disease isolates (WK16140_3 and AK17078_1) and one respiratory abscess isolate (SL002) (**Figure 4.11**). The 31 putative bacteriocins within the 51 New Zealand genomes were identified based on amino acid sequence similarity with known bacteriocins from diverse organisms. Not only did they share sequence identity with bacteriocins from different *Streptococcus* species (*salivarius*, *pyogenes*, *thermophiles*, *uberis* and *macedonicus)* but also from *Bacillus*, *Enterococcus*,

Pediococcus and *Staphylococcus* species (**Table S4. 3**) which supports HGT with potential interspecies exchange of bacteriocin gene clusters. However, this is only indicative and GC content signatures, sequence identity (BLAST), and synteny analysis would need to be carried out, a future study pathway for *S. zooepidemicus* research.

The three horse samples having two isolates sequenced each, as discussed earlier, were analysed for the presence/absence of bacteriocins. Like the putative fitness factors, the pairs of isolates from both the WK16140 respiratory disease sample and the WK15114 uterus isolate had a different pattern of bacteriocin presence/absence (**Figure 4.11**). Again, the cgMLST related isolates from the healthy KMW006 sample were identical, containing the same putative bacteriocins (**Figure 4.11**).

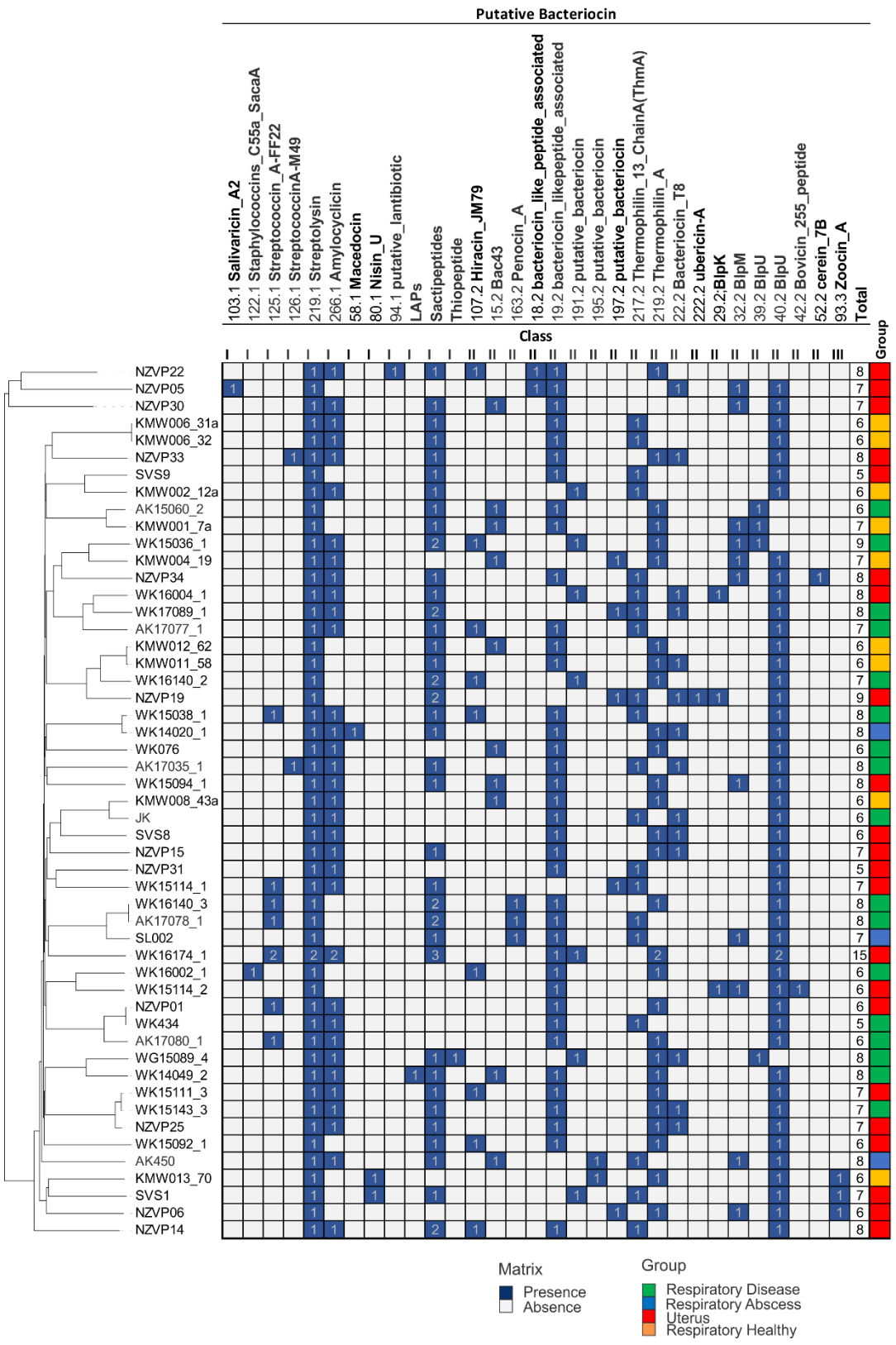


Figure 4.11: Thirty-one putative bacteriocins among the 51 New Zealand *S. zooepidemicus* isolates as identified by BAGEL4 (http://bagel4.molgenrug.nl/), ordered against the cgMLST phylogeny. Numbers within the blue presence blocks indicate the number of that particular putative bacteriocin found in the isolate. **Table S4. 3** provides details on these 31 bacteriocins.

4.3.4.5 Pan-Genome Wide Association Study of *S. zooepidemicus*

To extend the search in finding genes associated with particular traits, Scoary software was used to do a pan-GWAS. This was based on the Roary pangenome output which created gene clusters established on protein identities of 95 % sequence similarity. Pan-GWAS was performed on the same groupings as the above putative fitness factor distribution analysis of isolation, anatomical, clonal expansion and biofilm groups. No significant associations were found with the presence of any genes within these groupings based on Benjiamini-Hochberg corrected p-values of the naïve-p Fisher's exact tests ($p > 0.05$) with sensitivity and specificity set at 80 % (**Dataset S4. 1**).

For comparative purposes, the Pan-GWAS was then performed on the global representative pangenome. To validate the criteria set for association, Scoary was used to find if it would identify genes previously shown to be associated with the pig and guinea pig/human cluster. Specifically, putative fitness genes within previously identified genomic islands (Ma *et al.*, 2013) shown to be associated with these strains (Chen *et al.*, 2020). Scoary identified 182 genes associated with the pig and guinea pig/human cluster (**Dataset S4. 2)**. Of these 182 genes, all four genomic islands had genes represented with a total of 10, out of the 13 genes assessed, having a significant association with this cluster (**Table 4.8**). These results suggest that the criteria set was sufficient to be confident that any significant associations would be found, if they existed, and would have high sensitivity and specificity for that trait. However, proceeding with the global pan-GWAS analysis, no significant associations of genes were identified as specific to horses. This included gene presence comparisons within isolates from horses, whether globally or specifically from New Zealand versus all other hosts or countries and from horse respiratory isolates versus horse reproductive isolates (**Dataset S4. 2**).

Table 4.8: Scoary Pan-GWAS results of putative fitness genes associated with the pig and guinea pig/human cluster found in Genomic Islands (GI). Reference sequences from (Ma *et al.*, 2013), p-value is Benjiamini-Hochberg corrected. Green highlights genes that did not make the significance cut-off assigned: naïve-p Fisher's exact tests, Benjiamini-Hochberg corrected p-value; P > 0.05; sensitivity and specificity ≥ 80 %; and odds ratio > 1.

GI	Reference Sequence	Description	Sensitivity	Specificity	p-value
GI-1	SeseC_00898	Type II toxin-antitoxin system, antitoxin Phd	92.3	100	2.77E-12
	SeseC_00899	type II toxin-antitoxin system Doc family toxin	92.3	100	2.77E-12
GI-2	SeseC_01325	virulence-associated protein E	100	96.5	9.63E-12
	SeseC_01334	Fic/DOC domain protein, virulence protein RhuM family	100	96.5	9.63E-12
GI-3	SeseC_01908	Type IV secretory pathway, VirB4 component	100	73.3	9.64E-06
	SeseC_01912	membrane protein	100	74.4	6.55E-06
	SeseC_01914	Type IV secretory pathway, VirD4 component	100	94.2	8.46E-11
	SeseC_01916	Type IV secretory pathway,VirD4 component	100	84.9	6.18E-08
	SeseC_01875	Type II TA system zeta toxin (PezT)	100	88.4	8.54E-09
	SeseC_01876	Type II TA system epsilon antitoxin (PezA)	100	79.1	1.03E-06
GI-4	SeseC_02360	restriction endonuclease	100	96.5	9.63E-12
	SeseC_02361	restriction endonuclease control protein	100	96.5	9.63E-12
	SeseC_02362	methyltransferase (yhdJ)	100	96.5	9.63E-12

4.3.5 Functional Classification of the New Zealand *S. zooepidemicus* Pangenome

To determine if there were any differences in the repertoire of gene types present in the different groups of New Zealand isolates, clusters of orthologous groups (COGs) of gene functional categories in the core and accessory genomes of each isolate were identified (**Figure 4.12**). Of the genes identified by eggNOG, 20 % and 26 % of the core and accessory genomes (respectively) had unknown functions. The core genome of the isolates was enriched for metabolism genes when compared with the accessory genome (p > 0.001). In particular, functional categories Q (secondary metabolites biosynthesis, transport and catabolism), P (inorganic ion transport and metabolism), I (Lipid transport and metabolism), F (Nucleotide transport and metabolism), E (Amino acid transport and metabolism) and C (Energy production and conversion) were assigned to core genes by at least 2-fold when compared to the accessory genes. In contrast, the accessory genome was enriched for genes with cellular processes and signalling function when compared to the core genome (p > 0.001). In particular, functional categories V (Defence mechanisms), U (Intracellular trafficking, secretion, and vesicular transport), and N (cell motility) were assigned to the accessory genes

with a 2-fold increase than the core genes. The accessory genome also contained genes functionally linked to extracellular structures (COG W) with none identified in the core genome. These consisted of hypothetical proteins, putative collagen-binding surface-anchored proteins, putative fibronectin-binding proteins, and putative membrane proteins. However, a discrepancy was found with a core gene, a putative fibronectin/fibrinogen-binding protein (*S. zooepidemiucs* reference H70 isolate locus tag SZO_07320) found in all 51 isolates, was classed by eggNOG as COG K functional group for transcription with a gene description of RNA-binding protein homologous to an Eukaryotic snRNP.

There was no significant difference in the genes assigned to the information, storage and processing functional category between core and accessory genomes ($p = 0.192$). However, groups within this functional category varied greatly. The core genome was enriched 2-fold for genes functionally related to translation, ribosomal structure and biogenesis (J). The accessory genome was enriched 2-fold for genes functionally related to replication, recombination and repair (COG L). In the accessory genome, two uterus isolates (WK15094_1 and NZVP30) contained a hypothetical protein that was given a functional COG class of chromatin structure and dynamics (B) with an eggNOG description of positive regulation of histone acetylation. This gene shared 99 % nucleotide identity over 100 % of its length with a DNA-binding protein of a UK *S. zooepidemicus* strain, isolated from a mouse with septicaemia (locus lag NCTC6176_01814).

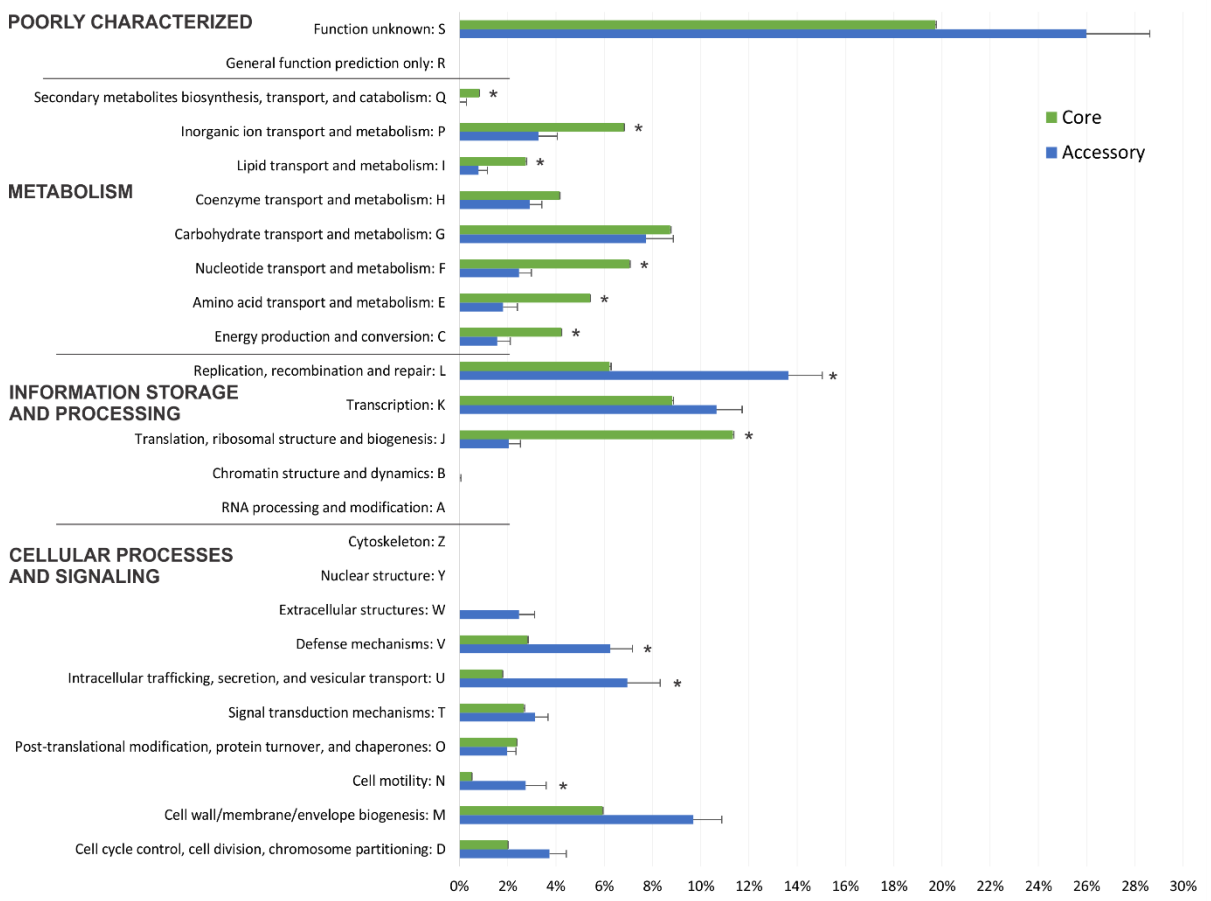


Figure 4.12: Core and accessory genome functional annotations of New Zealand *S. zooepidemicus*. Annotations were added using the eggNOG-mapper tool (v2.0.1-4-g2466c1b) *(Huerta-Cepas et al., 2017; Huerta-Cepas et al., 2019)*. Bars represent the median percentage of total genes annotated by eggNOG (n=51 isolates) for each Clusters of Orthologous Groups (COG) functional categories. Standard deviation is shown by error bars. * indicates variance between the core and accessory genome by at least a 2-fold increase.

Further, no significant differences were found in the accessory genes COG functional annotations between the groups of respiratory disease, respiratory abscess, respiratory healthy, and uterine (p-value ranging from 0.169 - 1) (**Figure 4.13**). In other words, no enrichment of a functional group could account for isolates from different anatomical isolation source or disease status.

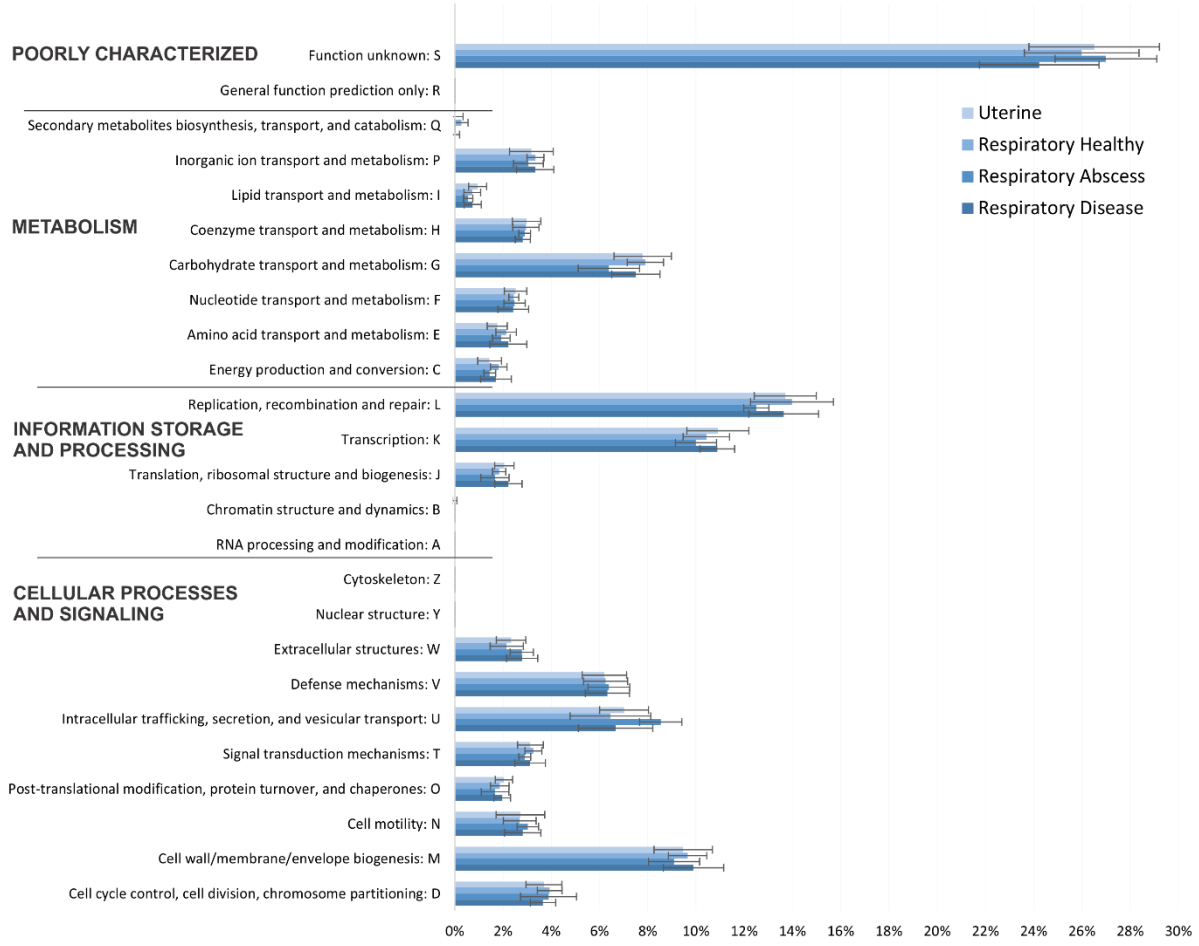


Figure 4.13: Accessory genome functional annotations breakdown based on isolate isolation source grouping of the New Zealand *S. zooepidemicus*. Annotations were done using the eggNOG-mapper tool (v2.0.1-4-g2466c1b) *(Huerta-Cepas et al., 2017; Huerta-Cepas et al., 2019)*. Bars represent the median percentage of total genes annotated by eggNOG for each Clusters of Orthologous Groups (COG) functional categories separated into isolation source groups: Respiratory Diseases (n=17), Respiratory Abscess (n=3), Respiratory Healthy (n=9), and Uterine (n=22). Standard deviation is shown by error bars.

4.4 Discussion

4.4.1 *Streptococcus zooepidemicus* Core genome Typing

Analysis of the core *S. zooepidemicus* genes through cgMLST resulted in no differentiation between New Zealand isolates involved in respiratory or uterine infection or disease status. Global analysis was able to separate the pig isolates associated with high mortality along with

the guinea pig/human isolates from all other strains as previously identified (Chen *et al.*, 2020). Yet, global analysis provided no further groupings/associations of horse isolates regardless of being of New Zealand origin or isolated from horse respiratory or reproductive tracts. This indicates that cgMLST results in strains being defined as not globally restricted nor host or anatomically restricted. The cgMLST scheme in Pathogenwatch coped with the variation seen in the *S. zooepidemicus* isolates better than the standard MLST. Population structure was able to be resolved with greater capacity due to the 1286 core genes versus seven genes, enabling the differentiation of isolates within the same ST and not being confounded by missing alleles. This increased resolution will facilitate epidemiological and outbreak surveillance studies. For instance, whole-genome sequencing and sequential cgSNP analysis was able to identify a *S. zooepidemicus* strain as the causative agent of an Icelandic horse respiratory epidemic differentiating it from those strains that were endemic (Björnsdóttir *et al.*, 2017). CgMLST also emphasised the genetic diversity within the *S. zooepidemicus* isolates, in particular, the pairwise cgSNP averaged 26,145.7 for the 99 global isolates to 26,540.8 for the 51 New Zealand isolates. These results were 288 times the genetic variation seen within *S. equi*, where the average number of pairwise cgSNPs between six Pathogenwatch clusters of 663 isolates of *S. equi* was only 90.9 cgSNPs (Mitchell *et al.*, 2021). The wide distribution of New Zealand isolates on the tree is reflective of the overall diversity of the species with no evidence of geographical restrictions. Further, these large numbers of point mutations, although showing the high diversity, may not be important in pathogenesis or even commensalism. Rather, the acquisition of novel genes may drive the evolution of pathogen or commensal traits and of respiratory or reproductive sources.

Of importance to note, the New Zealand horse isolate, WK16140_2 isolated from nasal discharge, clustered with NCTC12090 (accession: LS483328), a 1987 isolate from human blood with a fatal case of glomerulonephritis, with 248 pairwise cgSNP between them. This association was made in a previous chapter (Chapter 2) where they both had identical *szp* and ST sequences (*szp*-48 and ST-5). *S. zooepidemicus* can cause severe disease in humans, as shown with the NCTC12090 isolate, and transmission from horses to humans has been identified. For instance, although being a rare cause of infection in humans, 66 % of *S. equi* (inclusive of subspecies) meningitis infections in humans had suspected horse transmission (van Samkar *et al.*, 2016). Germany reported a case of *S. zooepidemicus* meningitis with

endophthalmitis complications, leaving the patient with severely impaired vision (Madžar *et al.*, 2015). Zoonotic transmission in this case was suspected to be from a horse with an upper respiratory tract infection whom the patient was taking care of (Madžar *et al.*, 2015). In Finland, molecular typing identified horses as the transmission source of three unrelated cases of *S. zooepidemicus* infection (Pelkonen *et al.*, 2013). The infections were severe, requiring prolonged treatment and rehabilitation with a range of clinical manifestations; sepsis, meningitis, endocarditis, purulent arthritis and psoas abscess in connection with an aortic wall infection (Pelkonen *et al.*, 2013). Norway reported a suspected zoonotic transmission to human from healthy ponies resulting in severe and very rare necrotizing myositis (Kittang *et al.*, 2017). Although we are unaware of any similar New Zealand cases, New Zealand horse owners and professionals should recognize *S. zooepidemicus* as a potential zoonosis that could lead to severe and even life threatening infections in humans.

4.4.2 The Pangenome and Accessory Putative Fitness Factors

The pangenome of *S. zooepidemicus* is open which speaks to the ability of this species to acquire new genes, diversify and exploit different hosts and environments. An average of 45 new genes were discovered with the addition of each new genome, indicating the difficulty in capturing the entire genetic diversity of this species. As such, bacterial species with open pangenomes will never be fully described (Medini *et al.*, 2005). Further, an open pangenome is synonymous with sympatric bacteria living in diverse environments with many partners from which to acquire DNA (Rouli *et al.*, 2015). As shown in Chapter 2, *S. zooepidemicus* has a wide range of hosts and can be isolated from diverse anatomical regions giving it more opportunity to encounter other strains and other microbiota increasing chances of DNA uptake.

Reductive genome evolution has been associated with pathogenicity, where pathogenic bacteria often have smaller genomes with fewer genes than less- or non-pathogenic bacteria (Weinert & Welch, 2017). Streptococcal species also exhibit this trend, such as in *S. suis* which is known as a zoonosis found in pigs. In the pig host, it is either an asymptomatic coloniser of the upper respiratory tract, or a pathogen causing respiratory tract infections and serious invasive disease within this host (Weinert *et al.*, 2015). Isolates associated with disease

contain less genes than non-clinical isolates, with putative virulence genes over-represented (Weinert *et al.*, 2015). New Zealand equine *S. zooepidemicus* isolates showed no such tendency. No differences were identified in the number of genes from isolates of healthy sources and those of respiratory disease/abscess or uterine sources.

Two putative fitness factor associations with phenotypes were identified within this thesis. *SzM* was linked to the isolation source of respiratory abscesses and superantigens were linked to biofilm negative isolates. However, the antiphagocytic *szM* gene association with the abscess group was not considered significant due to the low number of isolates tested. The finding of superantigens associated with the inability of form biofilms was unexpected. Superantigens are immune modulatory exotoxins, generating non-specific T-cell activation resulting in fever, shock and death (Proft & Fraser, 2003). Recently, the role of superantigens in *S. pyogenes* biofilm establishment has been tested resulting in nine out of 11 superantigens having biofilm inhibitory effects (Babbar *et al.*, 2019). In particular, superantigen SPeA was able to attenuate biofilm forming capabilities of all 13 isolates of *S. pyogenes* trialled (Babbar *et al.*, 2019). Biofilm formation of other *Streptococcus* species (*S. gordonii*, *S. oralis*, and *S. mutans*) were also inhibited by SPeA (Babbar *et al.*, 2019). SPeA had an effect on mature biofilms as well, with increased numbers of planktonic bacteria in the supernatant post SPeA treatment, which lead to the speculation of SpeA involvement in bacterial dissemination (Babbar *et al.*, 2019). Additional studies are needed to determine the role of superantigens in biofilm processes of *S. zooepidemicus*. This will also be relevant for *S. equi* which is known to contain superantigens (Paillot *et al.*, 2010b) and as of yet have no known effect on biofilm formation or dissemination.

As earlier described in Chapter 3, *S. zooepidemicus* isolates were able to form biofilms to varying degrees and were able to invade mammalian cells. Both biofilm formation and intracellular lifestyles have been separately associated with bacterial persister cells in many species (Fisher *et al.*, 2017). Bacterial persister cells are a subpopulation of bacteria within a genetically homogenous population that are in a dormant or slow-growing state with increased tolerances to antibiotic and other environmental stressors (Wen *et al.*, 2014; Fisher *et al.*, 2017). Antibiotic resistant persister cells have been described in *S. zooepidemicus* uterine isolates (Petersen *et al.*, 2015) which have links to intercellular invasion and

potentially biofilm formation (Skive *et al.*, 2017). Bacterial TA systems interfere with vital cellular processes, having three major functions; post-segregational killing, abortive infection, and persister cell formation (Harms *et al.*, 2018). Both post-segregational killing and abortive infection are suicide mechanisms through toxin activation (Harms *et al.*, 2018). Post-segregational killing functions to stabilise mobile elements, maintaining and allowing the spread of potential virulence and fitness genes within the population and abortive infection functions to protect bacterial populations from bacteriophage infection (Harms *et al.*, 2018). Whereas bacterial persistence, controlled by TA systems, enables the bacteria to enter a state of dormancy through controlled activation of toxins (Harms *et al.*, 2018). However, the role of TA systems in biofilm formation and persister cells is now highly contested and lacks supporting experimental evidence (Fraikin *et al.*, 2020). As such, the expression and biological function of the *S. zooepidemicus* candidate TA systems remains to be experimentally determined and their role in biofilm and persister cell formation cannot be assumed.

Nevertheless, the *S. zooepidemicus* isolates within the TASmania database (Akarsu *et al.*, 2019), contain on average 46 putative TA operons which is consistent with their multi-host, multi-system, opportunistic lifestyle. High numbers of TA systems are found in bacteria from hostile or dynamic environments and with high levels of HGT (Harms *et al.*, 2018). This study identified five putative type II TA systems, PezAT and four ParE toxins, and one putative type IV TA system, AbiEii/AbiE_4, in the 51 *S. zooepidemicus* New Zealand genomes. Type II TA systems typically produce both toxin and antitoxin proteins that form a toxin-antitoxin complex inactivating the toxin (Wen *et al.*, 2014). The toxins in type II TA systems inhibit essential cellular functions such as replication or protein synthesis with zeta toxin targeting UDP-N-acetylglucosamine involved in cell wall synthesis and ParE toxin targeting DNA gyrase involved in DNA replication (Wen *et al.*, 2014). The PezTA system found in 9 of the 51 New Zealand isolates was previously identified in the ATC35246 pig isolate within a PAI (Ma *et al.*, 2013). A mutant strain of the pig isolate was created with a partial deletion of this PAI, including the PezTA genes, and resulted in a less virulent strain in a mouse model (Ma *et al.*, 2013). Although the function of the PezTA system was not identified, the increased pathogenicity with this PAI could also indicate the importance in maintaining this virulence factor within the bacteria, a potential role of this TA system. Type IV TA systems toxin and antitoxin proteins do not form a complex as in the type II TA system, rather the antitoxin

interferes by directly binding with the toxin targets (Wen *et al.*, 2014). AbiEii toxin, of the bacterial abortive infection (Abi) innate immunity system, functions as a GTP-specific nucleotidyltransferase with reversible bacteriostasis activity playing a role in phage resistance and mobile genetic element maintenance (Dy *et al.*, 2014).

The truncation of HylC due to the SNP mutation in *S. zooepidemicus* isolate JK could explain its high levels of encapsulation due to the loss of hyaluronidase production. Similarly, the maintenance of high levels of encapsulation in both a *S. equi* (Se4047) and a *S. zooepidemicus* (ST-57) isolate, each with a disrupted *hylC* gene, was due to the loss of hyaluronidase production (Holden *et al.*, 2009). Se4047 and ST-57 both lost their mucoid appearance when grown in the presence of hyaluronidase (Holden *et al.*, 2009). Further, the reduced activity of this hyaluronate lysate may have decreased the ability of both the *S. zooepidemicus* JK and the *S. equi* isolates tested in the previous chapter to invade the mammalian cells (Chapter 3). Loss or reduced function of HylC may explain why *S. equi* infection rarely progresses beyond the lymphatic system as degradation of hyaluronic acid by hyaluronidase facilitates invasion of bacteria and their toxins (Holden *et al.*, 2009). Although the evidence suggests the HylC mutation as the cause of isolate JK's high levels of hyaluronate capsule and impaired invasion/biofilm abilities, further experiments are required to confirm this.

Bacteria also possess a range of "weapons" for inter- and intra-species conflicts providing fitness advantages by killing off other bacteria thereby reducing the competition for resources within the host's microbial community (García-Bayona & Comstock, 2018). Bacteriocins are one such weapon. They are toxins produced by bacteria that have antibiotic action on a narrow killing spectrum of similar or closely-related species (García-Bayona & Comstock, 2018). The findings in Chapter 2 of the clonal expansion of particular *S. zooepidemicus* strains (one strain present) in respiratory disease and uterus samples versus multiple strains found in the respiratory healthy individuals, was theorized to be caused by increased fitness of one strain over the others during infection. Bacteriocins may play a role in this perceived increase of fitness with the clonally expanded strains out-competing the others through the use of these antibiotics and thereby becoming the dominant strain. All New Zealand *S. zooepidemicus* isolates, that were genome sequenced, had on average seven putative bacteriocins present and they varied in the different types they possessed. Similarly, using a

different bacteriocin detection tool (BOA, Bacteriocin Operon and gene block Associator), *S. zooepidemicus* strain MGC10565 had six predicted bacteriocins (Morton *et al.*, 2015). Other streptococci also have similar findings, for instance *S. pyogenes* prevalence of bacteriocins varies from two to seven in each of the seven strains analysed (Nes *et al.*, 2007). It has been noted that a high number of predicted bacteriocins are harboured in species that inhabit different ecological niches from animal host bacteria (*S. zooepidemicus*) to soil-dwelling bacteria (*Streptomyces griseus*) and plant host bacteria (*Lesifsonia xyli*) (Morton *et al.*, 2015). This is suggestive of different functions and targets of bacteriocins dependant on their particular environment (Morton *et al.*, 2015). The finding of isolates, from the same host, containing different bacteriocins means they could potentially target each other in intra-species conflicts. However, identification of putative bacteriocins does not confirm expression of antimicrobial activity, which needs to be experimentally verified. This would be an interesting branch of research for *S. zooepidemicus* and fundamental in understanding the intra-species competition dynamics reflected in the changes in strain prevalence over time. However, it possesses several technical challenges including defining susceptible target bacteria, identifying how it is regulated and identifying any mutations that may have inactivated the functionality of the bacteriocin operon (Nes *et al.*, 2007).

Although the Pan-GWAS was able to detect genes associated with the pig and guinea pig/human cluster isolates, it resulted in no associations for the horse isolates. The pangenome was searched for genes associated with isolation source either within New Zealand or globally, genes associated with healthy or disease status, biofilm forming capabilities or clonal expansion within the New Zealand isolates. This could suggest that differences in the ability of a strain to cause disease are multifactorial and may be a combination of the host stress response and ensuing immune response. Consistent with the cgMLST findings, no evidence was found that the accessory genome can restrict a strain from becoming pathogenic or result in host or anatomical barriers.

Pathogen virulence factors can be found in 21 of the 25 different COG functional categories based on 5296 virulence factors from 126 pathogens inclusive of Gram-positive, Gram-negative, viruses, parasitic and fungal pathogens (Sayers *et al.*, 2019). They are absent from COG categories Z (cytoskeleton), W (extracellular structures), Y (nuclear structures) and

A (RNA processing and modification) (Sayers *et al.*, 2019). Pathogenic bacteria tend to have different representations of COG functional categories compared to bacteria with lower pathogenic potential when comparing genomes of pairs of closely-related species, a human pandemic bacteria and its closest non-pathogenic or non-epidemic related species (Georgiades & Raoult, 2011; Merhej *et al.*, 2013). In particular, highly pathogenic bacteria were enriched for the COG categories of replication, recombination and repair (Merhej *et al.*, 2013) and had a reduction in functional categories mostly related to metabolic activity, the production of energy and cell motility, and transcription (Georgiades & Raoult, 2011; Merhej *et al.*, 2013). No such trend was identified between the *S. zooepidemicus* New Zealand isolates from healthy versus disease versus anatomical regions, however differences in the distribution of COG categories were found between the core and accessory genomes. As expected, the core genome was enriched for metabolism genes and genes involved in translation, ribosomal structure and biogenesis. Such genes would be typical of being essential for basic cell function. The accessory genome was enriched for genes involved in replication, recombination and repair, defence mechanisms, intracellular trafficking and secretion, and cell motility. Of the genes identified by eggNOG, the accessory genomes had 26 % of its genes classified as unknown function. It was assumed that this proportion would far exceed the core genes, however they were very similar with the core genome having 20 % of the genes lacking functional classification. Hence, we are still unaware of all the biological and physiological function of many potentially important genes indispensable to *S. zooepidemicus*.

4.4.3 Conclusion

The genomic analysis of *S. zooepidemicus* genomes indicated a potential role of superantigens on the attenuation of biofilms, but was unable to identify genomic features associated with horse disease status or horse anatomical isolation site. Core and accessory genomes, cgSNPs and the presence or absence of genes, could not tease out any associations. This confirms the results of the second chapter where individual strain types (*szp*-type and MLST) were equally associated with disease and colonisation or equally associated with uterine and respiratory infection. Which is suggestive that the difference is more multifactorial and probably includes both the host immune response and physiology such as stress. Further,

this analysis did not consider gene-regulatory regions. These non-coding regions can be subject to horizontal transfer and recombination resulting in regulatory switching with phenotypic consequences (Oren *et al.*, 2014; McNally *et al.*, 2016). For instance, a single promoter inversion in *Photorhabdus luminescens* causes the switch from a commensal to a pathogen (Somvanshi *et al.*, 2012). Promotor inversion has been described in *S. zooepidemicus* as a result of the PinR serine recombinase (Steward *et al.*, 2015). This controls the transcription of a sortase-processed protein and is theorised to provide diversifying options in response to changing environments (Steward *et al.*, 2015). Analysis using a combination of core and accessory genomes with gene regulatory regions would provide enhanced resolution and a different perspective of bacterial evolution and phylogenomics (McNally *et al.*, 2016).

Further, as the regulation of genes will have fitness advantages in different environments, other methods beyond pan-GWAS might offer better resolution to gene and virulence/fitness associations. Studies should focus on identifying genes essential to *S. zooepidemicus* survival and those that are expressed during infection in different host environments. For instance, 45 *S. zooepidemicus* genes were found to be preferentially expressed during *in vivo* pig infections using the selective capture of transcribed sequences (SCOTS) technique (Yi *et al.*, 2013). SCOTS uses total RNA from infected tissue/cells and genes preferentially expressed are enriched through a series of subtractive hybridization and PCR amplification with tagged primers. The enriched cDNA is then cloned and screened by bot hybridization (Yi *et al.*, 2013). Another method is Transposon-insertion sequencing (TIS), a genome-wide transposon mutagenesis combined with next-generation sequencing that can link genotype to phenotype by estimating the fitness contribution of each gene (Cain *et al.*, 2020). TIS has been used to discover pathogen virulence factors of several streptococci; *S pyogenes*, *S. mutants* and *S. equi* (Cain *et al.*, 2020). In particular, researchers used the TIS variation method, transposon-directed insertion site sequencing (TraDIS), to identify fitness genes required for *S. equi* survival in a model representative of an equine immune response (in horse blood or hydrogen peroxide) (Charbonneau *et al.*, 2020). Fourteen genes were identified to contribute to the fitness of *S. equi* in response to the horse immune model (Charbonneau *et al.*, 2020). These techniques could be used to determine essential genes for increased fitness in different disease status and hosts of *S. zooepidemicus*.

Chapter 5

Conclusions

The overall aim of this study was to enhance the understanding of *S. zooepidemicus* in respiratory and uterine disease within the New Zealand horse population through strain associated genomic and phenotypic characteristics. This was combined with the goal of understanding the pathogenic potential of different strains and their associated factors for particular horse diseases with the intention of providing data that could contribute to their successful monitoring and management. However, rather than identifying the pathogenic potentials of different strains, this study highlighted that all strains of this bacterium have the potential to cause disease.

The main findings have retained a New Zealand focus, although they are echoed globally. *S. zooepidemicus* is highly-prevalent in New Zealand horses and highly diverse. It is the second most frequently isolated bacterium reported by New Zealand laboratories, being particularly associated with respiratory and urogenital samples. This diversity is further exemplified in the different anatomical isolation sources, including musculoskeletal and soft tissue sources, and further through their variable genomic content. Forty-eight *szp* types and 49 MLST types were identified from the 286 and 64 isolates analysed (respectively). CgMLST reflects *szp* and standard MLST typing methods, although it provides increased resolution to further differentiate strains with the same MLST, which will facilitate epidemiological contact and outbreak surveillance studies. In addition, cgMLST analysis was not confounded by missing alleles like standard MLST. Continual advances with increased speed and affordability is making whole genome sequencing the new gold standard for microbial typing. With the addition of platforms such as Pathogenwatch, this typing will easily become standardized and make data more readily available for studies and global comparisons.

Importantly, this work demonstrates that individual *S. zooepidemicus* strains can infect multiple anatomical sites of the New Zealand horse, such as both the uterus and the respiratory tracts. This held true regardless of health status with typed strains present in the uterus also found in the respiratory tract of healthy and diseased horses. *S. zooepidemicus* isolation can be in conjunction with clinical manifestations or asymptomatic colonisation, which could be sub-clinical as in the uterine of the mare. Further, a New Zealand horse isolate was highly similar to an isolate that caused a fatal human kidney disease with only 248 pairwise cgSNP between them and identical *szp* and MLST types. New Zealand horse owners and professionals should recognize the potential zoonosis that could lead to life threatening infections in humans and other animals.

Notably, when a disease occurred in the horses, the number of colonising strains tended to reduce to just the infective strain, which emphasizes the usefulness of investigating multiple bacterial colonies from a single sample. Seventy-six percent of the clinical respiratory samples and uterus samples (clinical and subclinical) had an expansion of a single strain compared to the apparently healthy status, which was associated with the isolation of multiple strains in 86 % of the healthy samples. Bacteriocins are hypothesized to be a potential method used by *S. zooepidemicus* to out-compete and kill off other strains in the host microbiota, becoming the dominant infective strain. All New Zealand *S. zooepidemicus* isolates had several putative bacteriocins identified with bacteriocin-types varying between them.

Additional survival strategies such as biofilm formation and cellular invasion are also available to *S. zooepidemicus*. Only seven of the 50 strains were unable to form and maintain a biofilm within the conditions tested and all of the 12 strains tested for cellular invasion were able to invade both human and equine cell lines. Having both biofilm and cellular invasion capabilities will enable *S. zooepidemicus* to colonise and persist by providing protection from the horse immune defences and from antibiotic treatments. TA systems in *S. zooepidemicus* may be involved in the persistence of this bacterium in healthy respiratory tracts and subclinical uterine samples through the formation of persister cells and may play a role in both biofilm and invasion processes. However, with the lack of experimental evidence, this role for *S. zooepidemicus* TA systems needs to be tested.

S. zooepidemicus has an open genome, which would allow it to diversify and exploit different niches although it maintains a relatively stable genome size with an average of 1896 genes per New Zealand isolate. No highly significant gene associations were found for any of the *S. zooepidemicus* New Zealand isolate groups of respiratory abscess, respiratory disease, respiratory healthy and uterus. Putative fitness factors were randomly dispersed throughout these different groups. However, the immune modulatory superantigens were associated with the inability of *S. zooepidemicus* to form biofilms. This is a potentially new function of *S. zooepidemicus* superantigens that warrants further investigation.

The techniques used in this study were unable to differentiate strains from healthy and disease and from respiratory and uterine sites. Although the very nature of this discovery predicts the potential of all strains to be pathogenic and that pathogenicity is not a stable characteristic. Further, it is suggestive that disease and colonisation by *S. zooepidemicus* is a complicated multifactorial process and does not simply rely on the gene repertoire of particular strains. The host immune system and condition, as well as potentially other environmental factors including the host microbial community will influence the pathogenic potential. This will be further confounded with differential gene expression in the colonisation state versus the active disease state. Subsequently, future work should focus around using a combination of genomes with gene regulatory regions and the expression of the genes they control to provide enhanced resolution of pathogenic potential.

This thesis provides the first in-depth analysis of *S. zooepidemicus* strains in New Zealand's equines, providing a foundation for further studies of this important pathogen. It defines *S. zooepidemicus* as an opportunistic pathogen that can reside in the healthy and infect the naïve; that are not anatomically, geographically or host bound. Accordingly, one cannot assume *S. zooepidemicus* is harmless and the zoonosis potential of New Zealand strains needs to be acknowledged.

References

Aaltonen, K., Kant, R., Eklund, M., Raunio-Saarnisto, M., Paulin, L., Vapalahti, O., Grönthal, T., Rantala, M., & Sironen, T. (2020). *Streptococcus halichoeri*: Comparative Genomics of an Emerging Pathogen. *International Journal of Genomics, 2020*, 8708305.

Acke, E., Midwinter, A. C., Lawrence, K., Gordon, S. J. G., Moore, S., Rasiah, I., Steward, K., French, N., & Waller, A. (2015). Prevalence of *Streptococcus dysgalactiae* subsp. *equisimilis* and *S. equi* subsp. *zooepidemicus* in a sample of healthy dogs, cats and horses. *New Zealand Veterinary Journal, 63*(5), 265-271.

Akarsu, H., Bordes, P., Mansour, M., Bigot, D.-J., Genevaux, P., & Falquet, L. (2019). TASmania: A bacterial Toxin-Antitoxin Systems database. *PLOS Computational Biology, 15*(4), e1006946.

Akineden, Ö., Hassan, A., Alber, J., El-Sayed, A., Estoepangestie, A., Lämmler, C., Weiss, R., & Siebert, U. (2005). Phenotypic and genotypic properties of *Streptococcus equi* subsp. *zooepidemicus* isolated from harbor seals (*Phoca vitulina*) from the German North Sea during the phocine distemper outbreak in 2002. *Veterinary Microbiology, 110*(1), 147-152.

Alber, J., El-Sayed, A., Estoepangestie, S., Lämmler, C., & Zschöck, M. (2005). Dissemination of the superantigen encoding genes *seeL*, *seeM*, *szeL* and *szeM* in *Streptococcus equi* subsp. *equi* and *Streptococcus equi* subsp. *zooepidemicus*. *Veterinary Microbiology, 109*(1), 135-141.

Albihn, A., Baverud, V., & Magnusson, U. (2003). Uterine microbiology and antimicrobial susceptibility in isolated bacteria from mares with fertility problems. *Acta Veterinaria Scandinavica, 44*(3-4), 121-130.

Alves-Barroco, C., Roma-Rodrigues, C., Raposo, L. R., Brás, C., Diniz, M., Caço, J., Costa, P. M., Santos-Sanches, I., & Fernandes, A. R. (2019). *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolated from milk of the bovine udder as emerging pathogens: In vitro and in vivo infection of human cells and zebrafish as biological models. *MicrobiologyOpen, 8*(1), e00623-e00623.

Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology, 55*(4), 539-52.

Anzai, T., Kuwamoto, Y., Wada, R., Sugita, S., Kakuda, T., Takai, S., Higuchi, T., & Timoney, J. F. (2005). Variation in the N-terminal region of an M-like protein of *Streptococcus equi* and evaluation of its potential as a tool in epidemiologic studies. *American Journal of Veterinary Research, 66*(12), 2167-71.

Anzai, T., Timoney, J. E., Kuwamoto, Y., Wada, R., Oikawa, M., & Higuchi, T. (2002). Polymerase chain reaction-restriction fragment length polymorphism analysis of the SzP gene of *Streptococcus*

zooepidemicus isolated from the respiratory tract of horses. *American Journal of Veterinary Research, 63*(9), 1298-301.

Anzai, T., Walker, J. A., Blair, M. B., Chambers, T. M., & Timoney, J. F. (2000). Comparison of the phenotypes of *Streptococcus zooepidemicus* isolated from tonsils of healthy horses and specimens obtained from foals and donkeys with pneumonia. *American Journal of Veterinary Research, 61*(2), 162-6.

Armistead, B., Oler, E., Adams Waldorf, K., & Rajagopal, L. (2019). The Double Life of Group B Streptococcus: Asymptomatic Colonizer and Potent Pathogen. *Journal of Molecular Biology, 431*(16), 2914-2931.

Arredondo-Alonso, S., Top, J., Schürch, A. C., McNally, A., Puranen, S., Pesonen, M., Pensar, J., Marttinen, P., Braat, J. C., Rogers, M. R. C., van Schaik, W., Kaski, S., Corander, J., & Willems, R. J. L. (2019). Genomes of a major nosocomial pathogen *Enterococcus faecium* are shaped by adaptive evolution of the chromosome and plasmidome. *bioRxiv*, 530725.

Babbar, A., Barrantes, I., Pieper, D. H., & Itzek, A. (2019). Superantigen SpeA attenuates the biofilm forming capacity of *Streptococcus pyogenes*. *Journal of Microbiology, 57*(7), 626-636.

Barquero, N., Chanter, N., Laxton, R., Wood, J. L., & Richard Newton, J. (2010). Molecular epidemiology of *Streptococcus zooepidemicus* isolated from the respiratory tracts of Thoroughbred racehorses in training. *The Veterinary Journal, 183*(3), 348-351.

Benga, L., Goethe, R., Rohde, M., & Valentin-Weigand, P. (2004). Non-encapsulated strains reveal novel insights in invasion and survival of *Streptococcus suis* in epithelial cells. *Cellular Microbiology, 6*(9), 867-881.

Beres, S. B., Sesso, R., Pinto, S. W., Hoe, N. P., Porcella, S. F., Deleo, F. R., & Musser, J. M. (2008). Genome sequence of a Lancefield group C *Streptococcus zooepidemicus* strain causing epidemic nephritis: new information about an old disease. *PLOS One, 3*(8), e3026.

Björnsdóttir, S., Harris, S. R., Svansson, V., Gunnarsson, E., Sigurðardóttir, Ó. G., Gammeljord, K., Steward, K. F., Newton, J. R., Robinson, C., Charbonneau, A. R. L., Parkhill, J., Holden, M. T. G., & Waller, A. S. (2017). Genomic dissection of an Icelandic epidemic of respiratory disease in horses and associated zoonotic cases. *mBio, 8*(4), e00826-17.

Björnsdóttir, S., Holden, M. T. G., Svansson, V., Harris, S. R., Webb, K., Robinson, C., Steward, K. F., Wright, N., Paillot, R., Newton, J. R., Gunnarsson, E., & Waller, A. S. (2012). *Streptococcus zooepidemicus*: more than just an opportunist? *Journal of Equine Veterinary Science, 32*(10, Supplement), S8.

Blum, S., Elad, D., Zukin, N., Lysnyansky, I., Weisblith, L., Perl, S., Netanel, O., & David, D. (2010). Outbreak of *Streptococcus equi* subsp. *zooepidemicus* infections in cats. *Veterinary Microbiology, 144*(1), 236-239.

Boyle, A. G., Timoney, J. F., Newton, J. R., Hines, M. T., Waller, A. S., & Buchanan, B. R. (2018). *Streptococcus equi* infections in horses: guidelines for treatment, control, and prevention of

strangles—revised consensus statement. *Journal of Veterinary Internal Medicine, 32*(2), 633-647.

Brynildsrud, O., Bohlin, J., Scheffer, L., & Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology, 17*(1), 238.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*(1), 59-60.

Bustos, C., Lanza, N., Marfil, M., Etchecopaz, A., Muñoz, A., Moras, E., & Guida, N. (2016). Expression of capsule and biofilm formation by *Streptococcus equi* subsp. *equi*. *Journal of Equine Veterinary Science, 39*, S94.

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. *Nature Reviews Genetics, 21*, 526-540.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421.

Carvallo, F. R., Uzal, F. A., Diab, S. S., Hill, A. E., & Arthur, R. M. (2017). Retrospective study of fatal pneumonia in racehorses. *Journal of Veterinary Diagnostic Investigation, 29*(4), 450-456.

Casagrande Proietti, P., Bietta, A., Coppola, G., Felicetti, M., Cook, R. F., Coletti, M., Marenzoni, M. L., & Passamonti, F. (2011). Isolation and characterization of β-haemolytic-Streptococci from endometritis in mares. *Veterinary Microbiology, 152*(1), 126-130.

Causey, R. C. (2006). Making sense of equine uterine infections: the many faces of physical clearance. *Veterinary Journal, 172*(3), 405-21.

Chalker, V. J., Waller, A., Webb, K., Spearing, E., Crosse, P., Brownlie, J., & Erles, K. (2012). Genetic diversity of *Streptococcus equi* subsp. *zooepidemicus* and doxycycline resistance in kennelled dogs. *Journal of Clinical Microbiology, 50*(6), 2134-2136.

Chanter, N., Collin, N., Holmes, N., Binns, M., & Mumford, J. (1997). Characterization of the Lancefield group C *streptococcus* 16S-23S RNA gene intergenic spacer and its potential for identification and sub-specific typing. *Epidemiology and Infection, 118*(02), 125-135.

Charbonneau, A. R. L., Taylor, E., Mitchell, C. J., Robinson, C., Cain, A. K., Leigh, J. A., Maskell, D. J., & Waller, A. S. (2020). Identification of genes required for the fitness of *Streptococcus equi* subsp. *equi* in whole equine blood and hydrogen peroxide. *Microbial Genomics, 6*(4).

Chen, X., Resende-De-Macedo, N., Sitthicharoenchai, P., Sahin, O., Burrough, E., Clavijo, M., Derscheid, R., Schwartz, K., Lantz, K., Robbe-Austerman, S., Main, R., & Li, G. (2020). Genetic characterization of *Streptococcus equi* subspecies *zooepidemicus* associated with high swine mortality in United States. *bioRxiv*, 2019.12.12.874644.

Christley, R. M., Hodgson, D. R., Rose, R. J., Wood, J. L., Reids, S. W., Whitear, K. G., & Hodgson, J. L. (2001). A case-control study of respiratory disease in Thoroughbred racehorses in Sydney, Australia. *Equine Veterinary Journal, 33*(3), 256-64.

Clark, C., Greenwood, S., Boison, J. O., Chirino-Trejo, M., & Dowling, P. M. (2008). Bacterial isolates from equine infections in western Canada (1998-2003). *The Canadian Veterinary Journal = La Revue Veterinaire Canadienne, 49*(2), 153-160.

Collins, M. D., Hutson, R. A., Falsen, E., Nikolaitchouk, N., LaClaire, L., & Facklam, R. R. (2000). An unusual Streptococcus from human urine, *Streptococcus urinalis* sp. *nov*. *International Journal of Systematic and Evolutionary Microbiology, 50*(3), 1173-1178.

Cursons, R., Patty, O., Steward, K. F., & Waller, A. S. (2015). Strangles in horses can be caused by vaccination with Pinnacle I. N. *Vaccine, 33*(30), 3440-3443.

da Silva Filho, A. C., Raittz, R. T., Guizelini, D., De Pierri, C. R., Augusto, D. W., dos Santos-Weiss, I. C. R., & Marchaukoski, J. N. (2018). Comparative analysis of genomic island prediction tools. *Frontiers in Genetics, 9*(619).

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., & Jere, A. (2013). Identification of optimum sequencing depth especially for *de novo* genome assembly of small genomes using next generation sequencing data. *PLOS ONE, 8*(4), e60204-e60204.

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., & Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research, 7*, ELIXIR-148.

Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. C., & Fineran, P. C. (2014). A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Research, 42*(7), 4590-4605.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLOS Computational Biology, 7*(10).

Eisen, J. A. (2000). Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development, 10*(6), 606-611.

Erol, E., Locke, S. J., Donahoe, J. K., Mackin, M. A., & Carter, C. N. (2012). Beta-hemolytic *Streptococcus* spp. from horses a retrospective study (2000–2010). *Journal of Veterinary Diagnostic Investigation, 24*(1), 142-147.

Etchecopaz, A., Munoz, A. J., Bustos, C. P., Mesplet, M., Guida, N., Iovannitti, C., & Retamar, G. (2018). Biofilm behavior of *Streptococcus equi* subsp. *zooepidemicus* and their interaction with *Candida Glabrata* isoloted from healthy horses. *International Journal of Infectious Diseases, 73*, 385.

Etchecopaz, A. N., Muñoz, A. J., Bustos, C. P., Martinez, G., Perez, A., Moras, E. V., Iovannitti, C. A., & Guida, N. (2016). Kinetic of biofilm formation by *Streptococcus equi* subsp *zooepidemicus*. *Journal of Equine Veterinary Science, 39*, S96-S97.

Fan, H., Wang, Y., Tang, F., & Lu, C. (2008). Determination of the mimic epitope of the M-like protein adhesin in swine *Streptococcus equi* subsp. *zooepidemicus*. *BMC Microbiology, 8*, 170.

Fernández, E., Blume, V., Garrido, P., Collins, M. D., Mateos, A., Domínguez, L., & Fernández-Garayzábal, J. F. (2004). *Streptococcus equi* subsp. *ruminatorum* subsp. *nov.*, isolated from mastitis in small ruminants. *International Journal of Systematic and Evolutionary Microbiology, 54*(6), 2291-2296.

Ferrer, M. S., & Palomares, R. (2018). Aerobic uterine isolates and antimicrobial susceptibility in mares with post-partum metritis. *Equine Veterinary Journal, 50*(2), 202-207.

Fisher, R. A., Gollan, B., & Helaine, S. (2017). Persistent bacterial infections and persister cells. *Nature Reviews Microbiology, 15*(8), 453.

Fraikin, N., Goormaghtigh, F., & Van Melderen, L. (2020). Type II toxin-antitoxin systems: evolution and revolutions. *Journal of Bacteriology, 202*(7), e00763-19.

Francisco, A. P., Bugalho, M., Ramirez, M., & Carrico, J. A. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics, 10*, 152.

Frontoso, R., De Carlo, E., Pasolini, M., van der Meulen, K., Pagnini, U., Iovane, G., & De Martino, L. (2008). Retrospective study of bacterial isolates and their antimicrobial susceptibilities in equine uteri during fertility problems. *Research in Veterinary Science, 84*(1), 1-6.

García-Bayona, L., & Comstock, L. E. (2018). Bacterial antagonism in host-associated microbial communities. *Science, 361*(6408).

Genteluci, G. L., Silva, L. G., Souza, M. C., Glatthardt, T., de Mattos, M. C., Ejzemberg, R., Alviano, C. S., Figueiredo, A. M., & Ferreira-Carvalho, B. T. (2015). Assessment and characterization of biofilm formation among human isolates of *Streptococcus dysgalactiae* subsp. *equisimilis*. *International Journal of Medical Microbiology, 305*(8), 937-47.

Georgiades, K., & Raoult, D. (2011). Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more toxin-antitoxin modules. *PLOS ONE, 6*(3), e17962.

Gobbetti, M., & Calasso, M. (2014). *Streptococcus* | Introduction. In C. A. Batt & M. L. Tortorello (Eds.), *Encyclopedia of Food Microbiology (Second Edition)* (pp. 535-553). Oxford: Academic Press.

Groschup, M. H., Hahn, G., & Timoney, J. F. (1991). Antigenic and genetic homogeneity of *Streptococcus uberis* strains from the bovine udder. *Epidemiology and infection, 107*(2), 297-310.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology, 59*(3), 307-21.

Guo, G., Wei, D., Zhang, Y., Wang, K., Liu, R., Wu, Z., Kong, L., Cheng, L., & Zhang, W. (2020). M-like protein SrM is not crucial to the virulence of a novel isolate of *Streptococcus equi* subsp. *ruminatorum* from *Macaca mulatta*. *Research in Veterinary Science, 132*, 221-228.

Harms, A., Brodersen, D. E., Mitarai, N., & Gerdes, K. (2018). Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. *Molecular Cell, 70*(5), 768-784.

Holden, M. T. G., Heather, Z., Paillot, R., Steward, K. F., Webb, K., Ainslie, F., Jourdan, T., Bason, N. C., Holroyd, N. E., Mungall, K., Quail, M. A., Sanders, M., Simmonds, M., Willey, D., Brooks, K., Aanensen, D. M., Spratt, B. G., Jolley, K. A., Maiden, M. C. J., Kehoe, M., Chanter, N., Bentley, S. D., Robinson, C., Maskell, D. J., Parkhill, J., & Waller, A. S. (2009). Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLOS Pathogens, 5*(3), e1000346.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution, 34*(8), 2115-2122.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research, 47*(D1), D309-d314.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*(1), 119.

Hynes, W. L., & Walton, S. L. (2000). Hyaluronidases of Gram-positive bacteria. *FEMS Microbiology Letters, 183*(2), 201-207.

Ijaz, M., Velineni, S., & Timoney, J. F. (2011). Selective pressure for allelic diversity in SeM of *Streptococcus equi* does not affect immunoreactive proteins SzPSe or Se18.9. *Infection, Genetics and Evolution, 11*(5), 1159-1163.

Jackson, R. W., Vinatzer, B., Arnold, D. L., Dorus, S., & Murillo, J. (2011). The influence of the accessory genome on bacterial pathogen evolution. *Mobile Genetic Elements, 1*(1), 55-65.

Janda, W. M. (2014). The Genus *Streptococcus* – Part I: Emerging Pathogens in the "Pyogenic Cocci" and the "*Streptococcus bovis*" Groups. *Clinical Microbiology Newsletter, 36*(20), 157-166.

Jannatabadi, A. A., Mohammadi, G. R., Rad, M., & Maleki, M. (2008). Molecular identification of *Streptococcus equi* subsp. *equi* and *Streptococcus equi* subsp. *zooepidemicus* in nasal swabs samples from horses suffering respiratory infections in Iran. *Pakistan Journal of Biological Sciences, 11*(3), 468-71.

Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Research, 3*, 124.

Jolley, K. A., & Maiden, M. C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics, 11*(595), 1471-2105.

Jorm, L., Love, D., Bailey, G., McKay, G., & Briscoe, D. (1994). Genetic structure of populations of β-haemolytic Lancefield group C streptococci from horses and their association with disease. *Research in Veterinary Science, 57*(3), 292-299.

Kelly, C., Bugg, M., Robinson, C., Mitchell, Z., Davis-Poynter, N., Newton, J. R., Jolley, K. A., Maiden, M. C. J., & Waller, A. S. (2006). Sequence variation of the SeM Gene of *Streptococcus equi* allows discrimination of the source of strangles outbreaks. *Journal of Clinical Microbiology, 44*(2), 480-486.

Kittang, B. R., Pettersen, V. K., Oppegaard, O., Skutlaberg, D. H., Dale, H., Wiker, H. G., & Skrede, S. (2017). Zoonotic necrotizing myositis caused by *Streptococcus equi* subsp. *zooepidemicus* in a farmer. *BMC Infectious Diseases, 17*(1), 147.

Kwasny, S. M., & Opperman, T. J. (2010). Static biofilm cultures of Gram-positive pathogens grown in a microtiter format used for anti-biofilm drug discovery. *Current Protocols in Pharmacology, Chapter 13*, Unit 13A.8.

Lannergård, J., & Guss, B. (2006). IdeE, an IgG-endopeptidase of *Streptococcus equi* ssp. *equi*. *FEMS Microbiology Letters, 262*(2), 230-235.

Laslett, D., & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research, 32*(1), 11-6.

Laus, F., Preziuso, S., Spaterna, A., Beribe, F., Tesei, B., & Cuteri, V. (2007). Clinical and epidemiological investigation of chronic upper respiratory diseases caused by beta-haemolytic Streptococci in horses. *Comparative Immunology, Microbiology and Infectious Diseases, 30*(4), 247-60.

Lavoie, J. P., Couture, L., Higgins, R., & Laverty, S. (1991). Aerobic bacterial isolates in horses in a university hospital, 1986-1988. *The Canadian Veterinary Journal = La Revue Veterinaire Canadienne, 32*(5), 292-294.

Lawson, P. A., Foster, G., Falsen, E., & Collins, M. D. (2005). *Streptococcus marimammalium* sp. *nov.*, isolated from seals. *International Journal of Systematic and Evolutionary Microbiology, 55*(1), 271-274.

Lefort, V., Longueville, J.-E., & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution, 34*(9), 2422-2424.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research, 47*(W1), W256-W259.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Lindahl, S. B., Aspán, A., Båverud, V., Paillot, R., Pringle, J., Rash, N. L., Söderlund, R., & Waller, A. S. (2013). Outbreak of upper respiratory disease in horses caused by *Streptococcus equi* subsp. *zooepidemicus* ST-24. *Veterinary Microbiology, 166*(1–2), 281-285.

Luque, I., Fernández‐Garayzábal, J., Blume, V., Maldonado, A., Astorga, R., & Tarradas, C. (2006). Molecular typing and anti-microbial susceptibility of clinical Isolates of *Streptococcus equi* ssp. *zooepidemicus* from equine bacterial endometritis. *Journal of Veterinary Medicine, Series B, 53*(9), 451-454.

Ma, F., Guo, X., & Fan, H. (2017a). Extracellular nucleases of *Streptococcus equi* subsp. *zooepidemicus* degrade neutrophil extracellular traps and impair macrophage activity of the host. *Applied and Environmental Microbiology, 83*(2), e02468-16.

Ma, J. S., Chen, S. Y., & Lo, H. H. (2017b). Biofilm formation of beta-hemolytic group G *Streptococcus dysgalactiae* subspecies *equisimilis* isolates and its association with emm polymorphism. *Acta Pathologica, Microbiologica, et Immunologica Scandinavica, 125*(11), 1027-1032.

Ma, Z., Geng, J., Yi, L., Xu, B., Jia, R., Li, Y., Meng, Q., Fan, H., & Hu, S. (2013). Insight into the specific virulence related genes and toxin-antitoxin virulent pathogenicity islands in swine streptococcosis pathogen *Streptococcus equi* ssp. *zooepidemicus* strain ATCC35246. *BMC Genomics, 14*(1), 377.

Ma, Z., Peng, J., Yu, D., Park, J. S., Lin, H., Xu, B., Lu, C., Fan, H., & Waldor, M. K. (2019). A streptococcal Fic domain-containing protein disrupts blood-brain barrier integrity by activating moesin in endothelial cells. *PLOS Pathogens, 15*(5), e1007737.

Ma, Z., Zhang, H., Zheng, J., Li, Y., Yi, L., Fan, H., & Lu, C. (2012). Interaction between M-like protein and macrophage thioredoxin facilitates antiphagocytosis for *Streptococcus equi* ssp. *zooepidemicus*. *PLoS One, 7*(2), e32099.

Madžar, D., Hagge, M., Möller, S., Regensburger, M., Lee, D.-H., Schwab, S., & Jantsch, J. (2015). Endogenous endophthalmitis complicating *Streptococcus equi* subspecies *zooepidemicus* meningitis: a case report. *BMC Research Notes, 8*(1), 184.

Magoč, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics, 27*(21), 2957-2963.

Marcellin, E., Gruber, C. W., Archer, C., Craik, D. J., & Nielsen, L. K. (2009). Proteome analysis of the hyaluronic acid-producing bacterium, *Streptococcus zooepidemicus*. *Proteome Science, 7*(1), article no. 13.

Marchandin, H., Jumas-Bilak, E., Boumzebra, A., Vidal, D., Jonquet, O., & Corne, P. (2007). Fatal *Streptococcus equi* subsp. *ruminatorum* infection in a man. *Emerging Infectious Diseases, 13*(12), 1964-1966.

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., Vehkala, M., Välimäki, N., Prentice, M. B., Ashour, A., Avram, O., Pupko, T., Dobrindt, U., Literak, I., Guenther, S., Schaufler, K., Wieler, L. H., Zhiyong, Z., Sheppard, S. K., McInerney, J. O., & Corander, J. (2016). Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLOS Genetics, 12*(9), e1006280.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development, 15*(6), 589-594.

Merhej, V., Georgiades, K., & Raoult, D. (2013). Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. *Briefings in Functional Genomics, 12*.

Meyer, A., Messer, L., De Briel, D., & Moreau, P. (2011). Second reported case of human infection with *Streptococcus equi* subsp. *ruminatorum*. *Joint Bone Spine, 78*(3), 303-305.

Mir, I., Kumar, B., Taku, A., Wani, N., Faridi, F., Dar, S., Gazal, S., Badroo, G., Zargar, A., & Iqbal, A. (2013a). The study of aerobic bacterial flora of the upper respiratory tract of equines from Jammu and Kashmir region of India. *Veterinary World, 6*(9), 623-627.

Mir, I. A., Taku, A., Iqbal, A., Ahmad, M., & Wani, N. (2013b). Clinical mastitis by *Streptococcus equi* Subsp. *zooepidemicus* in goats of Jammu. *The Indian Journal of Small Ruminants, 19*(1), 110-111.

Mitchell, C., Steward, K. F., Charbonneau, A. R. L., Walsh, S., Wilson, H., Timoney, J. F., Wernery, U., Joseph, M., Craig, D., van Maanen, K., Hoogkamer-van Gennep, A., Leon, A., Witkowski, L., Rzewuska, M., Stefańska, I., Żychska, M., van Loon, G., Cursons, R., Patty, O., Acke, E., Gilkerson, J. R., El-Hage, C., Allen, J., Bannai, H., Kinoshita, Y., Niwa, H., Becú, T., Pringle, J., Guss, B., Böse, R., Abbott, Y., Katz, L., Leggett, B., Buckley, T. C., Blum, S. E., Cruz López, F., Fernández Ros, A., Marotti Campi, M. C., Preziuso, S., Robinson, C., Newton, J. R., Schofield, E., Brooke, B., Boursnell, M., de Brauwere, N., Kirton, R., Barton, C. K., Abudahab, K., Taylor, B., Yeats, C. A., Goater, R., Aanensen, D. M., Harris, S. R., Parkhill, J., Holden, M. T. G., & Waller, A. S. (2021). Globetrotting strangles: the unbridled national and international transmission of *Streptococcus equi* between horses. *Microbial Genomics*.

Morton, J. T., Freed, S. D., Lee, S. W., & Friedberg, I. (2015). A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. *BMC Bioinformatics, 16*(1), 381.

Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carriço, J. A., & Vaz, C. (2016). PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics, 33*(1), 128-129.

Nes, I. F., Diep, D. B., & Holo, H. (2007). Bacteriocin Diversity in *Streptococcus* and *Enterococcus*. *Journal of Bacteriology, 189*(4), 1189.

Newton, J. R., Laxton, R., Wood, J. L., & Chanter, N. (2008). Molecular epidemiology of *Streptococcus zooepidemicus* infection in naturally occurring equine respiratory disease. *Veterinary Journal, 175*(3), 338-45.

Newton, J. R., Wood, J. L., & Chanter, N. (2003). A case control study of factors and infections associated with clinically apparent respiratory disease in UK Thoroughbred racehorses. *Preventive Veterinary Medicine, 60*(1), 107-32.

Nielsen, J. M. (2005). Endometritis in the mare: a diagnostic study comparing cultures from swab and biopsy. *Theriogenology, 64*(3), 510-518.

Novick, S., Shagan, M., Blau, K., Lifshitz, S., Givon-Lavi, N., Grossman, N., Bodner, L., Dagan, R., & Mizrachi Nebenzahl, Y. (2017). Adhesion and invasion of *Streptococcus pneumoniae* to primary and secondary respiratory epithelial cells. *Molecular Medicine Reports, 15*(1), 65-74.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., & Sirotkin, Y. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. In *Annual International Conference on Research in Computational Molecular Biology* (pp. 158-170): Springer.

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature, 405*(6784), 299-304.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology, 17*(1).

Oren, Y., Smith, M. B., Johns, N. I., Kaplan Zeevi, M., Biran, D., Ron, E. Z., Corander, J., Wang, H. H., Alm, E. J., & Pupko, T. (2014). Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences, 111*(45), 16112.

Othman, A. M., Assayaghi, R. M., Al-Shami, H. Z., & Saif-Ali, R. (2019). Asymptomatic carriage of *Streptococcus pyogenes* among school children in Sana'a city, Yemen. *BMC Research Notes, 12*(1), 339.

Overbeck, W., Witte, T., & Heuwieser, W. (2011). Comparison of three diagnostic methods to identify subclinical endometritis in mares. *Theriogenology, 75*(7), 1311-1318.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics, 31*(22), 3691-3693.

Paillot, R., Darby, A. C., Robinson, C., Wright, N. L., Steward, K. F., Anderson, E., Webb, K., Holden, M. T. G., Efstratiou, A., Broughton, K., Jolley, K. A., Priestnall, S. L., Marotti Campi, M. C., Hughes, M. A., Radford, A., Erles, K., & Waller, A. S. (2010a). Identification of three novel superantigen-encoding genes in *Streptococcus equi* subsp. *zooepidemicus*, *szeF*, *szeN*, and *szeP*. *Infection and Immunity, 78*(11), 4817.

Paillot, R., Robinson, C., Steward, K., Wright, N., Jourdan, T., Butcher, N., Heather, Z., & Waller, A. S. (2010b). Contribution of each of four superantigens to *Streptococcus equi*-induced mitogenicity, gamma interferon synthesis, and immunity. *Infection and Immunity, 78*(4), 1728.

Panchaud, Y., Gerber, V., Rossano, A., & Perreten, V. (2010). Bacterial infections in horses: a retrospective study at the University Equine Clinic of Bern. *Schweizer Archiv fur Tierheilkunde, 152*(4), 176-82.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics, 20*(2), 289-290.

Parker, R. E., Laut, C., Gaddy, J. A., Zadoks, R. N., Davies, H. D., & Manning, S. D. (2016). Association between genotypic diversity and biofilm production in group B *Streptococcus*. *BMC Microbiology, 16*(1), 86.

Patty, O. A., & Cursons, R. T. M. (2014). The molecular identification of *Streptococcus equi* subsp. *equi* strains isolated within New Zealand. *New Zealand Veterinary Journal, 62*(2), 63-67.

Pelkonen, S., Lindahl, S. B., Suomala, P., Karhukorpi, J., Vuorinen, S., Koivula, I., Väisänen, T., Pentikäinen, J., Autio, T., & Tuuminen, T. (2013). Transmission of *Streptococcus equi* subspecies *zooepidemicus* infection from horses to humans. *Emerging infectious diseases, 19*(7), 1041-1048.

Petersen, M. R., Skive, B., Christoffersen, M., Lu, K., Nielsen, J. M., Troedsson, M. H., & Bojesen, A. M. (2015). Activation of persistent *Streptococcus equi* subspecies *zooepidemicus* in mares with subclinical endometritis. *Veterinary Microbiology, 179*(1-2), 119-125.

Pinto, T. C. A., Neves, F. P. G., Souza, A. R. V., Oliveira, L. M. A., Costa, N. S., Castro, L. F. S., Mendonça-Souza, C. R. d. V., Peralta, J. M., & Teixeira, L. M. (2019). Evolution of penicillin non-susceptibility among *Streptococcus pneumoniae* isolates recovered from asymptomatic carriage and invasive disease over 25 years in Brazil, 1990–2014. *Frontiers in Microbiology, 10*(486).

Podbielski, A., Beckert, S., Schattke, R., Leithäuser, F., Lestin, F., Goßler, B., & Kreikemeyer, B. (2003). Epidemiology and virulence gene expression of intracellular group A streptococci in tonsils of recurrently infected adults. *International Journal of Medical Microbiology, 293*(2), 179-190.

Prjibelski, A. D., Vasilinetc, I., Bankevich, A., Gurevich, A., Krivosheeva, T., Nurk, S., Pham, S., Korobeynikov, A., Lapidus, A., & Pevzner, P. A. (2014). ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics, 30*(12), i293-i301.

Proft, T., & Fraser, J. D. (2003). Bacterial superantigens. *Clinical and Experimental Immunology, 133*(3), 299-306.

Rasko, D. A., Rosovitz, M., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., & Chaudhuri, R. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology, 190*(20), 6881-6893.

Rasmussen, C. D., Haugaard, M. M., Petersen, M. R., Nielsen, J. M., Pedersen, H. G., & Bojesen, A. M. (2013). *Streptococcus equi* subsp. *zooepidemicus* isolates from equine infectious endometritis belong to a distinct genetic group. *Veterinary Research, 44*, 26.

Richards, V. P., Palmer, S. R., Bitar, P. D. P., Qin, X., Weinstock, G. M., Highlander, S. K., Town, C. D., Burne, R. A., & Stanhope, M. J. (2014). Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biology and Evolution, 6*(4), 741-753.

Riddle, W., LeBlanc, M., & Stromberg, A. (2007). Relationships between uterine culture, cytology and pregnancy rates in a Thoroughbred practice. *Theriogenology, 68*(3), 395-402.

Roberts, A. L., Connolly, K. L., Kirse, D. J., Evans, A. K., Poehling, K. A., Peters, T. R., & Reid, S. D. (2012). Detection of group A *Streptococcus* in tonsils from pediatric patients reveals high rate of asymptomatic streptococcal carriage. *BMC Pediatrics, 12*, 3-3.

Rohde, M., & Cleary, P. P. (2016). Adhesion and invasion of *Streptococcus pyogenes* into host cells and clinical relevance of intracellular streptococci. In *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet]*. University of Oklahoma Health Sciences Center. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK333420/.

Rouli, L., Merhej, V., Fournier, P. E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections, 7*, 72-85.

Roy, K., Bisgaard, M., Kyvsgaard, N. C., Christensen, J. P., Nielsen, O. L., Biswas, P. K., Pors, S. E., & Bojesen, A. M. (2013). Pathogenicity of wild-type and small-colony variants of *Streptococcus equi* subsp. *zooepidemicus* in layer chickens. *Avian Pathology, 42*(4), 316-322.

Ryu, S.-H., Koo, H. C., Lee, Y.-W., Park, Y. H., & Lee, C.-W. (2011). Etiologic and epidemiologic analysis of bacterial infectious upper respiratory disease in Thoroughbred horses at the Seoul Race Park. *Journal of Veterinary Science, 12*(2), 195-197.

Salasia, S. I. O., Wibawan, I. W. T., Pasaribu, F. H., Abdulmawjood, A., & Lammler, C. (2004). Persistent occurrence of a single *Streptococcus equi* subsp. *zooepidemicus* clone in the pig and monkey population in Indonesia. *Journal of Veterinary Science, 5*(3), 263-266.

Sanchez, C. J., Kumar, N., Lizcano, A., Shivshankar, P., Dunning Hotopp, J. C., Jorgensen, J. H., Tettelin, H., & Orihuela, C. J. (2011). *Streptococcus pneumoniae* in biofilms are unable to cause invasive disease due to altered virulence determinant production. *PLOS ONE, 6*(12), e28738.

Sayers, S., Li, L., Ong, E., Deng, S., Fu, G., Lin, Y., Yang, B., Zhang, S., Fa, Z., Zhao, B., Xiang, Z., Li, Y., Zhao, X.-M., Olszewski, M. A., Chen, L., & He, Y. (2019). Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Research, 47*(D1), D693-D700.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics, 27*(4), 592-3.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research, 34*(Database issue), D32-6.

Sitkiewicz, I., & Hryniewicz, W. (2010). Pyogenic streptococci--danger of re-emerging pathogens. *Polish Journal of Microbiology, 59*(4), 219-26.

Sitto, F., & Battistuzzi, F. U. (2019). Estimating pangenomes with Roary. *Molecular Biology and Evolution, 37*(3), 933-939.

Skive, B., Rohde, M., Molinari, G., Braunstein, T. H., & Bojesen, A. M. (2017). *Streptococcus equi* subsp. *zooepidemicus* Invades and Survives in Epithelial Cells. *Frontiers in Cellular and Infection Microbiology, 7*, 465.

Somvanshi, V. S., Sloup, R. E., Crawford, J. M., Martin, A. R., Heidt, A. J., Kim, K.-s., Clardy, J., & Ciche, T. A. (2012). A Single Promoter Inversion Switches *Photorhabdus* Between Pathogenic and Mutualistic States. *Science, 337*(6090), 88.

Song, L., Florea, L., & Langmead, B. (2014). Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biology, 15*(11), 509.

Speck, S., Höner, O. P., Wachter, B., & Fickel, J. (2008). Characterization of *Streptococcus equi* subsp. *ruminatorum* isolated from spotted hyenas (*Crocuta crocuta*) and plains zebras (*Equus burchelli*), and identification of a M-like protein (SrM) encoding gene. *Veterinary Microbiology, 128*(1), 148-159.

Stepanović, S., Vuković, D., Hola, V., Bonaventura, G. D., Djukić, S., Cirković, I., & Ruzicka, F. (2007). Quantification of biofilm in microtiter plates: overview of testing conditions and practical recommendations for assessment of biofilm production by staphylococci. *Acta Pathologica, Microbiologica, et Immunologica Scandinavica, 115*(8), 891-899.

Steward, K., Robinson, C., Maskell, D., Holden, M., Harris, S., Darby, A., & Waller, A. (2012). Pili of *Streptococcus equi* and *Streptococcus zooepidemicus*. *Journal of Equine Veterinary Science, 10*(32), S15.

Steward, K. F., Harrison, T., Robinson, C., Slater, J., Maskell, D. J., Harris, S. R., Holden, M. T. G., & Waller, A. S. (2015). PinR mediates the generation of reversible population diversity in *Streptococcus zooepidemicus*. *Microbiology, 161*(5), 1105-1112.

Steward, K. F., Robinson, C., Holden, M. T. G., Harris, S. R., Ros, A. F., Pérez, G. C., Baselga, R., & Waller, A. S. (2017). Diversity of *Streptococcus equi* subsp. *zooepidemicus* strains isolated from the Spanish sheep and goat population and the identification, function and prevalence of a novel arbutin utilisation system. *Veterinary Microbiology, 207*, 231-238.

Sweeney, C. R., Timoney, J. F., Newton, J. R., & Hines, M. T. (2005). *Streptococcus equi* infections in horses guidelines for treatment, control, and prevention of strangles. *Journal of Veterinary Internal Medicine, 19*, 123-134.

Taylor, S. D., & Wilson, W. D. (2006). *Streptococcus equi* subsp. *equi* (Strangles) Infection. *Clinical Techniques in Equine Practice, 5*(3), 211-217.

Tenenbaum, T., Bloier, C., Adam, R., Reinscheid, D. J., & Schroten, H. (2005). Adherence to and invasion of human brain microvascular endothelial cells are promoted by fibrinogen-binding protein FbsA of *Streptococcus agalactiae*. *Infection and Immunity, 73*(7), 4404-4409.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., & Durkin, A. S. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences, 102*(39), 13950-13955.

Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology, 11*(5), 472-477.

Thomas, P., Sekhar, A. C., Upreti, R., Mujawar, M. M., & Pasha, S. S. (2015). Optimization of single plate-serial dilution spotting (SP-SDS) with sample anchoring as an assured method for bacterial and yeast cfu enumeration and single colony isolation from diverse samples. *Biotechnology Reports, 8*, 45-55.

Timoney, J. F. (1993a). Patent No. U.S. Patent No. 5,183,659. Protection of equines against *Streptococcus equi*.

Timoney, J. F. (1993b). Strangles. *The Veterinary Clinics of North America: Equine Practice, 9*(2), 365-374.

Timoney, J. F. (2004). The pathogenic equine streptococci. *Veterinary Research, 35*(4), 397-409.

Tomida, J., Sauer, S., Morita, Y., Ezaki, T., & Kawamura, Y. (2011). *Streptococcus fryi* sp. *nov*., a novel species with Lancefield group M antigens. *FEMS Microbiology Letters, 314*(1), 95-100.

Toombs-Ruane, L. J., Riley, C. B., Kendall, A. T., Bolwell, C. F., Benschop, J., & Rosanowski, S. M. (2015). Antimicrobial susceptibilities of aerobic isolates from respiratory samples of young New Zealand horses. *Journal of Veterinary Internal Medicine, 29*(6), 1700-1706.

Tsang, A. K. L., Lee, H. H., Yiu, S.-M., Lau, S. K. P., & Woo, P. C. Y. (2017). Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. *Scientific Reports, 7*(1), 4536.

Urošević, M., Stojanović, D., Lako, B., Jajić, I., Miličić, Ž., Pušić, I., & Prodanov-Radulovic, J. (2011). Research on reproductive performance of mares in Serbia using bacteriological examination. *Biotechnology in Animal Husbandry, 27*(3), 883-892.

van Samkar, A., Brouwer, M. C., van der Ende, A., & Van de Beek, D. (2016). *Streptococcus equi* meningitis. *Clinical Microbiology and Infection, 22*(1), e3-e4.

van Spijk, J. N., Schmitt, S., Fürst, A., & Schoster, A. (2016). A retrospective study of bacterial pathogens in an equine hospital (1988–2014). *Schweizer Archiv für Tierheilkunde,, 158*(6), 423-431.

van Heel, A. J., de Jong, A., Song, C., Viel, J. H., Kok, J., & Kuipers, O. P. (2018). BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Research, 46*(W1), W278-W281.

Vasilinetc, I., Prjibelski, A. D., Gurevich, A., Korobeynikov, A., & Pevzner, P. A. (2015). Assembling short reads from jumping libraries with large insert sizes. *Bioinformatics, 31*(20), 3262-3268.

Velineni, S., Breathnach, C. C., & Timoney, J. F. (2014a). Evidence of lateral gene transfer among strains of *Streptococcus zooepidemicus* in weanling horses with respiratory disease. *Infection, Genetics and Evolution, 21*, 157-60.

Velineni, S., Desoutter, D., Perchec, A.-M., & Timoney, J. F. (2014b). Characterization of a mucoid clone of *Streptococcus zooepidemicus* from an epizootic of equine respiratory disease in New Caledonia. *The Veterinary Journal, 200*(1), 82-87.

Velineni, S., & Timoney, J. (2014). Capsular hyaluronic acid of equine isolates of *Streptococcus zooepidemicus* is up‐regulated at temperatures below 35° C. *Equine Veterinary Journal, 47*(3), 333-338.

Velineni, S., & Timoney, J. F. (2013a). Identification of novel immunoreactive proteins of *Streptococcus zooepidemicus* with potential as vaccine components. *Vaccine, 31*(38), 4129-4135.

Velineni, S., & Timoney, J. F. (2013b). Characterization and protective immunogenicity of the SzM protein of *Streptococcus zooepidemicus* NC78 from a clonal outbreak of equine respiratory disease. *Clinical and Vaccine Immunology, 20*(8), 1181-1188.

Velineni, S., Timoney, J. F., Russell, K., Hamlen, H. J., Pesavento, P., Fortney, W. D., & Crawford, P. C. (2014c). Clones of *Streptococcus zooepidemicus* from outbreaks of hemorrhagic canine pneumonia and associated immune responses. *Clinical and Vaccine Immunology, 21*(9), 1246-1252.

Von Dollen, K. A., Jones, M., Beachler, T., Harris, T. L., Papich, M. G., Lyle, S. K., & Bailey, C. S. (2019). Antimicrobial activity of ceftiofur and penicillin with gentamicin against *Escherichia coli* and *Streptococcus equi* subspecies *zooepidemicus* in an *ex vivo* model of equine postpartum uterine disease. *Journal of Equine Veterinary Science, 79*, 121-126.

Vyas, H. K. N., Proctor, E.-J., McArthur, J., Gorman, J., & Sanderson-Smith, M. (2019). Current Understanding of Group A Streptococcal Biofilms. *Current Drug Targets, 20*(9), 982-993.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE, 9*(11), e112963.

Walker, J. A., & Timoney, J. F. (1998). Molecular basis of variationin protective SzP proteins of *Streptococcus zooepidemicus*. *American Journal of Veterinary Research, 59*(9), 1129-1133.

Walker, J. A., & Timoney, J. F. (2002). Construction of a stable non-mucoid deletion mutant of the *Streptococcus equi* Pinnacle vaccine strain. *Veterinary Microbiology, 89*(4), 311-321.

Walker, R. L., & Runyan, C. A. (2003). Identification of variations in SzP proteins of *Streptococcus equi* subspecies *zooepidemicus* and the relationship between protein variants and clinical signs of infection in horses. *American Journal of Veterinary Research, 64*(8), 976-81.

Waller, A. S. (2014). Equine respiratory disease: a causal role for *Streptococcus zooepidemicus*. *Veterinary Journal 201*(1), 3-4.

Waller, A. S., Paillot, R., & Timoney, J. F. (2011). *Streptococcus equi*: a pathogen restricted to one host. *Journal of Medical Microbiology, 60*(9), 1231-1240.

Waller, A. S., & Robinson, C. (2013). *Streptococcus zooepidemicus* and *Streptococcus equi* evolution: the role of CRISPRs. *Biochemical Society Transactions, 41*(6), 1437-1443.

Walter, J., Neuberg, K., Failing, K., & Wehrend, A. (2012). Cytological diagnosis of endometritis in the mare: investigations of sampling techniques and relation to bacteriological results. *Animal Reproduction Science, 132*(3), 178-186.

Webb, K., Jolley, K. A., Mitchell, Z., Robinson, C., Newton, J. R., Maiden, M. C., & Waller, A. (2008). Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group. *Microbiology, 154*(10), 3016-3024.

Wei, Z., Fu, Q., Chen, Y., Li, M., Cong, P., Mo, D., & Liu, X. (2013). *Streptococcus equi* ssp. *zooepidemicus* C5a peptidase, a putative invasin, induces protective immune response in mice. *Research in Veterinary Science, 95*(2), 444-450.

Weinert, L. A., Chaudhuri, R. R., Wang, J., Peters, S. E., Corander, J., Jombart, T., Baig, A., Howell, K. J., Vehkala, M., Välimäki, N., Harris, D., Chieu, T. T. B., Van Vinh Chau, N., Campbell, J., Schultsz, C., Parkhill, J., Bentley, S. D., Langford, P. R., Rycroft, A. N., Wren, B. W., Farrar, J., Baker, S., Hoa, N. T., Holden, M. T. G., Tucker, A. W., Maskell, D. J., Bossé, J. T., Li, Y., Maglennon, G. A., Matthews, D., Cuccui, J., Terra, V., & Consortium, B. R. T. (2015). Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nature Communications, 6*(1), 6740.

Weinert, L. A., & Welch, J. J. (2017). Why might bacterial pathogens have small genomes? *Trends in Ecology & Evolution, 32*(12), 936-947.

Wen, Y., Behiels, E., & Devreese, B. (2014). Toxin–Antitoxin systems: their role in persistence, biofilm formation, and pathogenicity. *Pathogens and Disease, 70*(3), 240-249.

Whatmore, A. M., Engler, K. H., Gudmundsdottir, G., & Efstratiou, A. (2001). Identification of isolates of *Streptococcus canis* infecting humans. *Journal of Clinical Microbiology, 39*(11), 4196-4199.

Wood, J., Newton, J., Chanter, N., & Mumford, J. (2005a). Inflammatory airway disease, nasal discharge and respiratory infections in young British racehorses. *Equine Veterinary Journal, 37*(3), 236-242.

Wood, J., Newton, J., Chanter, N., & Mumford, J. (2005b). Association between respiratory disease and bacterial and viral infections in British racehorses. *Journal of Clinical Microbiology, 43*(1), 120-126.

Xu, B., Pei, X., Su, Y., Ma, Z., & Fan, H. (2016). Capsule of *Streptococcus equi* subsp. *zooepidemicus* hampers the adherence and invasion of epithelial and endothelial cells and is attenuated during internalization. *FEMS Microbiology Letters, 363*(16).

Xu, B., Zhang, P., Zhou, H., Sun, Y., Tang, J., & Fan, H. (2019). Identification of novel genes associated with anti-phagocytic functions in *Streptococcus equi* subsp. *zooepidemicus*. *Veterinary Microbiology, 233*, 28-38.

Yi, L., Wang, Y., Ma, Z., Lin, H.-X., Xu, B., Grenier, D., Fan, H.-J., & Lu, C.-P. (2016). Identification and characterization of a *Streptococcus equi* ssp. *zooepidemicus* immunogenic GroEL protein involved in biofilm formation. *Veterinary Research, 47*(1), 50.

Yi, L., Wang, Y., Ma, Z., Zhang, H., Li, Y., Zheng, J. X., Yang, Y. C., Fan, H. J., & Lu, C. P. (2014). Biofilm formation of *Streptococcus equi* ssp. *zooepidemicus* and comparative proteomic analysis of biofilm and planktonic cells. *Current Microbiology, 69*(3), 227-33.

Yi, L., Wang, Y., Ma, Z., Zhang, H., Xie, H., Yang, Y., Lu, C., & Fan, H. (2013). Identification of genes transcribed by *Streptococcus equi* ssp. *zooepidemicus* in infected porcine lung. *Microbial Pathogenesis, 59-60*, 7-12.

Young, C., Holder, R. C., Dubois, L., & Reid, S. D. (2016). *Streptococcus pyogenes* Biofilm. In *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet]*. University of Oklahoma Health Sciences Center. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK333419/.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., Wu, J., & Xiao, J. (2014). PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics, 30*(9), 1297-1299.

Zhou, H., Ma, Z., Yuan, J., & Fan, H.-j. (2013). Identification and characterization of putative virulent genes in *Streptococcus equi* ssp. *zooepidemicus*. *Journal of Integrative Agriculture, 12*(2), 327-333.

Supplementary Files

Chapter 2

Table S2. 1: Details of New Zealand *S. zooepidemicus* isolates from the University of Waikato collection. The columns for "# of isolates" and "# of szp types" have data in the first isolate row per sample ID and represent all isolates from that sample. Table is provided as an Excel file.

Table S2. 2: Szp strain sequence accession numbers assigned by NCBI. Table is provided as an Excel file.

Table S2. 3: MLST STs with complete profiles used in this current study. Table is provided as an Excel file.

Chapter 3

File S3. 1: Preliminary trials of invasion assay. This supplementary file is provided as a word document.

Chapter 4

Table S4. 1: Details of New Zealand horse *S. zooepidemicus* isolates genome sequencing. Table is provided as an Excel file.

Table S4. 2: Details of all *S. zooepidemicus* isolates used for global comparisons. Table is provided as an Excel file.

Table S4. 3: Details of the 31 bacteriocins identified in New Zealand *S. zooepidemicus* isolates obtained for the BAGEL4 webserver (http://bagel4.molgenrug.nl/). Table is provided as an Excel file.

References

Table S4. 4: NZ gene presence absence table of the 51 New Zealand *S. zooepidemicus* isolates created by the Roary pipeline (v3.8.2) with a minimum percentage identity of 95 % for sequence comparisons (Page *et al.*, 2015; Sitto & Battistuzzi, 2019). Table is provided as an Excel file.

Table S4. 5: Global gene presence absence table of the 99 global *S. zooepidemicus* isolates created by the Roary pipeline (v3.8.2) with a minimum percentage identity of 95 % for sequence comparisons (Page *et al.*, 2015; Sitto & Battistuzzi, 2019). Table is provided as an Excel file.

Dataset S4. 1: Pan-GWAS results, for the New Zealand *S. zooepidemicus* groupings, produced by the Scoary (v1.6.16) pipeline (Brynildsrud et al., 2016). The naïve Fisher's exact test with a with a Benjiamini-Hochberg false discovery rate adjusted p-value less than 0.05 was set to identify a gene of interest. Gene presence was then considered to have a significant association for the trait analysed with the following criteria: odds ratio > 1, sensitivity > 80 % and specificity > 80 %. Dataset is provided as an Excel file.

Dataset S4. 2: Pan-GWAS results, for the global *S. zooepidemicus* groupings, produced by the Scoary (v1.6.16) pipeline (Brynildsrud et al., 2016). The naïve Fisher's exact test with a with a Benjiamini-Hochberg false discovery rate adjusted p-value less than 0.05 was set to identify a gene of interest. Gene presence was then considered to have a significant association for the trait analysed with the following criteria: odds ratio > 1, sensitivity > 80 % and specificity > 80 %. Dataset is provided as an Excel file.