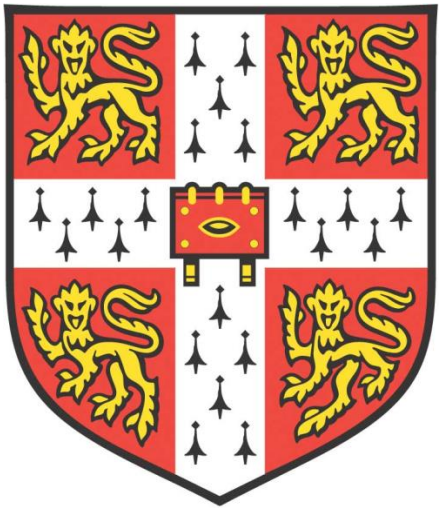
Barcoded Transposon Directed Insertion-site Sequencing (TraDIS); a tool to improve our understanding of the functional genomics of *Streptococcus equi* subsp. *equi*



Amelia Rose Louise Charbonneau

Department of Veterinary Medicine

University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy

Lucy Cavendish College

December 2018

Barcoded Transposon Directed Insertion-site Sequencing (TraDIS); a tool to improve our understanding of the functional genomics of *Streptococcus equi* subsp. *equi*

Amelia Rose Louise Charbonneau

Strangles, caused by *Streptococcus equi* (*S. equi*) remains the most frequently diagnosed infectious disease of horses and is a cause of significant welfare and economic cost. Vaccine research has been limited by the time taken to make mutations in individual genes to determine their role in the disease process. However, the development of transposon directed insertion-site sequencing (TraDIS) technologies provides an opportunity to simultaneously determine the importance of every gene in *S. equi* under disease relevant conditions, significantly enhancing the capacity to identify new vaccine targets.

In this project, a novel barcoded TraDIS technique was designed, which identified that 19.5 percent of the *S. equi* genome is essential to basic survival in rich medium *in vitro*, 73.4 percent of genes being non-essential, with the remainder either not defined or of an ambiguous assignment. Comparative analysis revealed that more than 83 percent of the essential gene set of *S. equi* was concordant with the essential genomes of *S. pyogenes* and *S. agalactiae*, highlighting the close genetic relationships between these important pathogenic bacteria.

Barcoded libraries were exposed to hydrogen peroxide (H_2O_2) and whole equine blood, to simulate the interaction with the equine immune system. Sequencing of surviving mutants enabled identification of genes important to *S. equi* under these conditions *in vitro*. Fifteen and 36 genes were implicated in the survival of *S. equi* in H_2O_2 and whole equine blood, respectively. Results were validated by generating deletion mutant strains in 4 of the genes (*pyrP*, *mnmE*, *addA* and *recG*). Mutant strains were exposed to H_2O_2 or whole equine blood and surviving bacteria measured over time. An additional 2 deletion strains in *eqbE* and *hasA*, generated prior to this project, were also utilised.

Barcoded TraDIS is proposed to reduce the effects of stochastic loss commonly seen in similar datasets, enhancing the ability to resolve differences in the fitness of mutants. To determine the *in vivo* capabilities of barcoded TraDIS, 12 Welsh mountain ponies were each infected with 2 of 3 barcoded libraries. Viable mutants were recovered and sequenced from the abscess material of infected lymph nodes and data analysed both exploiting (barcoded analysis; BC) and disregarding (per animal analysis; PA) the input library barcodes. Exploiting the barcodes enables output data to be combined on a per input library basis, as opposed to a per animal basis as is traditionally completed in

comparable *in vivo* transposon library studies. From the BC analysis, sequencing identified 368 genes required for fitness. Mutations in a further 85 genes conferred a fitness advantage *in vivo*. In the PA analysis, only 97 genes required for fitness were identified, which were all similarly identified in the barcoded analysis. No genes in which an insertion conferred a fitness advantage were identified in the PA analysis. To validate these results and confirm the benefit of applying a barcoded technique, 12 genes required for fitness were selected, plus 1 control gene where transposon insertions did not alter fitness, for tagged allelic replacement mutagenesis and repeat challenge *in vivo*. Seven genes required for fitness in both methods of analysis were selected, plus an additional 5 genes uniquely identified by the BC analysis. All deletion mutants appeared to be attenuated *in vivo*, however the control mutants and wild-type *S. equi* did not behave as expected, confounding statistical analysis.

Thirty-nine percent (14/36) and 60 percent (9/15) of fitness genes identified in the whole equine blood and H_2O_2 TraDIS screens, respectively, were also identified as being required for *in vivo* fitness. Nine consensus genes were identified as required in all 3 experiments. Comparison of the genes implicated in *in vivo* survival of *S. equi* to those in *S. pyogenes* in a non-human primate model of necrotising myositis and in a mouse model of subcutaneous infection, uncovered a set of 23 pan-species fitness genes. Eighteen genes were also commonly identified between the *S. equi in vivo* data and *S. pyogenes ex vivo* in human saliva, alluding to the potential genes required by *S. equi* to survive in the nasopharynx before translocation to the local lymph nodes.

The data presented in this thesis provide an unprecedented insight into the mechanisms employed by *S. equi* to cause disease in the natural host. The data also shed light on the pan-streptococcal pathways important for virulence that are likely to be important for future development of novel therapeutics and vaccines.

Declaration

I hereby declare that the contents of this dissertation are the result of my own work except where specific reference is made to the work of others and includes nothing which is the outcome of work done in collaboration except where specified in the text. This work has not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation contains fewer than 60,000 words excluding appendices, bibliography, tables and figures.

Signed:_____

Date:_____

Amelia Rose Louise Charbonneau

Acknowledgements

I would like to thank Dr Andrew Waller, for his outstanding guidance and encouragement throughout my time at the Animal Health Trust. He has been an incredibly supportive and considerate supervisor who has believed in me since day 1. His enthusiasm and optimism have made this PhD a rewarding and fulfilling experience that I will remember fondly. I would like to thank other members of the Animal Health Trust family, Carl Robinson for his advice and assistance in particular with generating the validation mutants used in this thesis and overall guidance when it came to troubleshooting experiments. Dr Oliver Forman assisted in the initial TraDIS runs and taught me most of what I know about sequencing. I am also grateful to Catriona Mitchell for her support in the laboratory but also for the emotional support and friendship that was completely invaluable, especially over the last year. All of the *in vivo* work completed in this PhD would not have run as smoothly as it did, without her assistance. I would also like to thank my Animal Health Trust mentor, Sally Ricketts, for keeping an eye on me and meeting up for chats about my progress.

I would also like to thank 2 people in particular at the Wellcome Sanger Institute, Dr Amy Cain and Dr Matthew Mayho for their interest in my project and for the help with experimental design. Sequencing of the *in vivo* samples at the Sanger would not have been possible without Dr Matthew Mayho. Dr Lars Barquist also provided invaluable advice regarding the analysis of the *in vivo* data and helped implement the TraDIS scripts. Prof. Duncan Maskell and Prof. James Leigh have also provided me with a wealth of advice and have spent many hours discussing data and experiments with me. Regular meetings with both Duncan and James always made me feel good about my research and spurred me on for the next experiment. I am incredibly grateful for their encouragement and interest in my project.

Last, but not least, I would like to acknowledge my amazing network of family and friends that have supported me thought my studies. In particular, my partner Richard Armstrong, has seen me through the good and the bad. He has always believed in me and has encouraged me through this PhD, to make it the best piece of research I could produce.

Abstract

Strangles, caused by *Streptococcus equi* (*S. equi*) remains the most frequently diagnosed infectious disease of horses and is a cause of significant welfare and economic cost. Vaccine research has been limited by the time taken to make mutations in individual genes to determine their role in the disease process. However, the development of transposon directed insertion-site sequencing (TraDIS) technologies provides an opportunity to simultaneously determine the importance of every gene in *S. equi* under disease relevant conditions, significantly enhancing the capacity to identify new vaccine targets.

In this project, a novel barcoded TraDIS technique was designed, which identified that 19.5 percent of the *S. equi* genome is essential to basic survival in rich medium *in vitro*, 73.4 percent of genes being non-essential, with the remainder either not defined or of an ambiguous assignment. Comparative analysis revealed that more than 83 percent of the essential gene set of *S. equi* was concordant with the essential genomes of *S. pyogenes* and *S. agalactiae*, highlighting the close genetic relationships between these important pathogenic bacteria.

Barcoded libraries were exposed to hydrogen peroxide (H_2O_2) and whole equine blood, to simulate the interaction with the equine immune system. Sequencing of surviving mutants enabled identification of genes important to *S. equi* under these conditions *in vitro*. Fifteen and 36 genes were implicated in the survival of *S. equi* in H_2O_2 and whole equine blood, respectively. Results were validated by generating deletion mutant strains in 4 of the genes (*pyrP*, *mnmE*, *addA* and *recG*). Mutant strains were exposed to H_2O_2 or whole equine blood and surviving bacteria measured over time. An additional 2 deletion strains in *eqbE* and *hasA*, generated prior to this project, were also utilised.

Barcoded TraDIS is proposed to reduce the effects of stochastic loss commonly seen in similar datasets, enhancing the ability to resolve differences in the fitness of mutants. To determine the *in vivo* capabilities of barcoded TraDIS, 12 Welsh mountain ponies were each infected with 2 of 3 barcoded libraries. Viable mutants were recovered and sequenced from the abscess material of infected lymph nodes and data analysed both exploiting (barcoded analysis; BC) and disregarding (per animal analysis; PA) the input library barcodes. Exploiting the barcodes enables output data to be combined on a per

input library basis, as opposed to a per animal basis as is traditionally completed in comparable *in vivo* transposon library studies. From the BC analysis, sequencing identified 368 genes required for fitness. Mutations in a further 85 genes conferred a fitness advantage *in vivo*. In the PA analysis, only 97 genes required for fitness were identified, which were all similarly identified in the barcoded analysis. No genes in which an insertion conferred a fitness advantage were identified in the PA analysis. To validate these results and confirm the benefit of applying a barcoded technique, 12 genes required for fitness were selected, plus 1 control gene where transposon insertions did not alter fitness, for tagged allelic replacement mutagenesis and repeat challenge *in vivo*. Seven genes required for fitness in both methods of analysis were selected, plus an additional 5 genes uniquely identified by the BC analysis. All deletion mutants appeared to be attenuated *in vivo*, however the control mutants and wild-type *S. equi* did not behave as expected, confounding statistical analysis.

Thirty-nine percent (14/36) and 60 percent (9/15) of fitness genes identified in the whole equine blood and H_2O_2 TraDIS screens, respectively, were also identified as being required for *in vivo* fitness. Nine consensus genes were identified as required in all 3 experiments. Comparison of the genes implicated in *in vivo* survival of *S. equi* to those in *S. pyogenes* in a non-human primate model of necrotising myositis and in a mouse model of subcutaneous infection, uncovered a set of 23 pan-species fitness genes. Eighteen genes were also commonly identified between the *S. equi in vivo* data and *S. pyogenes ex vivo* in human saliva, alluding to the potential genes required by *S. equi* to survive in the nasopharynx before translocation to the local lymph nodes.

The data presented in this thesis provide an unprecedented insight into the mechanisms employed by *S. equi* to cause disease in the natural host. The data also shed light on the pan-streptococcal pathways important for virulence that are likely to be important for future development of novel therapeutics and vaccines.

Table of contents

Abbreviations

3Rs	Replacement, reduction and refinement
BC	Barcoded analysis
CFU	Colony forming units
COG	Clusters of orthologous groups
DSB	Double strand break
FAD	Flavin adenine dinucleotide
GTP	Guanosine triphosphate
H_2O_2	Hydrogen peroxide
HITS	High-throughput insertion tracking sequencing
IC	Internal control
ICE	Integrative conjugative element
INSEQ	Insertion sequencing
IS*S1*	Insertion sequence *S1*
KEGG	Kyoto encyclopaedia of genes and genomes
LLR	Log_2 likelihood ratio
LRP	Left retropharyngeal lymph node
LSM	Left submandibular lymph node
MGE	Mobile genetic element
NAD	Nicotinamide adenine dinucleotide
NGS	Next-generation sequencing
PA	Per animal analysis
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
pGh	pG+ host
PIMMS	Pragmatic insertional mutation mapping system
qPCR	Quantitative polymerase chain reaction
RRP	Right retropharyngeal lymph node
RSM	Right submandibular lymph node
STM	Signature tagged mutagenesis
THA	Todd-hewitt agar
THAE	Todd-hewitt agar containing erythromycin
THB	Todd hewitt broth
THBE	Todd-hewitt broth containing erythromycin
Tn-seq	Transposon sequencing
TraDIS	Transposon directed insertion-site sequencing
TS	Temperature sensitive
WT	Wild-type

List of figures

List of tables

1 Introduction

Strangles, caused by *Streptococcus equi* subspecies *equi* (*S. equi*), is one of the most frequently diagnosed infectious diseases of Equids worldwide and is responsible for considerable economic and welfare cost to the horse industry. *S. equi* is a Gram positive bacterium and belongs to the Lancefield group C family of streptococci [1]. The first account of strangles was noted by Ruffus in 1251, although it is likely that the disease existed much before this.

1.1 *S. equi* and *S. zooepidemicus*

It has been known for some time that *S. equi* is likely to have evolved from an ancestral strain of *Streptococcus equi* subspecies *zooepidemicus* (*S. zooepidemicus*) [2]. *S. zooepidemicus* also causes disease in horses, but additionally infects other animals including humans. More recently, genome sequencing of both *S. equi* strain 4047 (*Se*4047) and *S. zooepidemicus* strain H70 (*Sz*H70) has provided information concerning the genetic events that have shaped this bacterial evolution. Comparatively, the genomes of *S. equi* and *S. zooepidemicus* were found to share over 97 percent genetic identity, in addition to both sharing 80 percent identity with the human pathogen, *Streptococcus pyogenes* (*S. pyogenes*) [3]. *Se*4047 has a slightly larger circular chromosome containing 2,253,793 base pairs (bp) compared to that of *Sz*H70 (2,149,866 bp) [3]. Despite their overall similarity, various genetic losses due to nonsense mutations and deletions, and gene gains by the integration of mobile genetic elements (MGEs) has led to the evolution of *S. equi*, its host-restriction and virulence [3].

1.2 Pathogenesis of strangles

The high infectivity of *S. equi* enables transmission directly, from horse to horse, and indirectly, most commonly via stable equipment, water and humans. In natural disease, following infection via the nasopharyngeal or oral routes, *S. equi* binds to and invades the mucosal epithelium before transitioning to the lymph nodes of the head and neck,

where it can be identified within 3 hours [4]. The exact mechanism in which *S. equi* translocates though to the lymph nodes is not yet defined. The presence of *S. equi* within lymph nodes induces substantial infiltration of polymorphonuclear leukocytes, leading to swelling and abscessation (Figure 1.1). The enlargement of infected lymph nodes may obstruct the airways, causing dysphagia and inspiratory difficulty, lending to this disease's common name of strangles [5]. Despite the induction of dramatic clinical signs in up to 100 percent of infected animals, mortality rates are relatively low at around 2 percent.

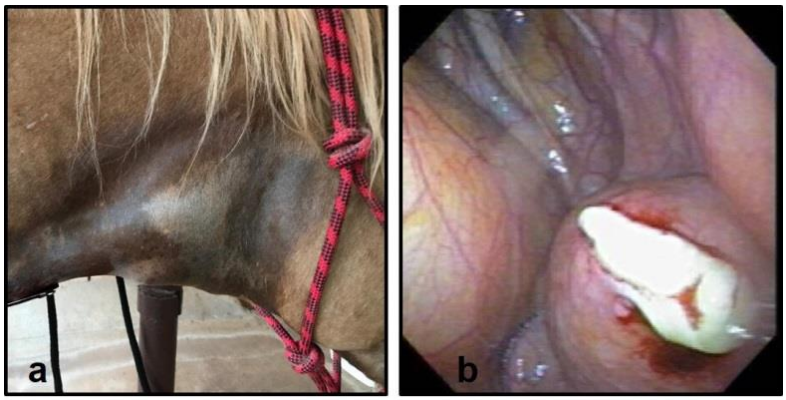


Figure 1.1. Horse with strangles.a) Retropharyngeal lymph node abscess caused by strangles infection. b) Endoscopy of the guttural pouch in the same horse as in image a, depicting the rupture of a retropharyngeal abscess. Images reproduced with permission by Amy Armentrout.

Abscess material drains into the nasal cavity leading to mucopurulent nasal discharge, assisting clearance of the infection. Usually, submandibular lymph node abscesses will rupture through the skin, and retropharyngeal lymph node abscess into the guttural pouch (Figure 1b). The guttural pouch is a bilateral air-filled sac that is an extension of the Eustachian tube (Figure 1.2).

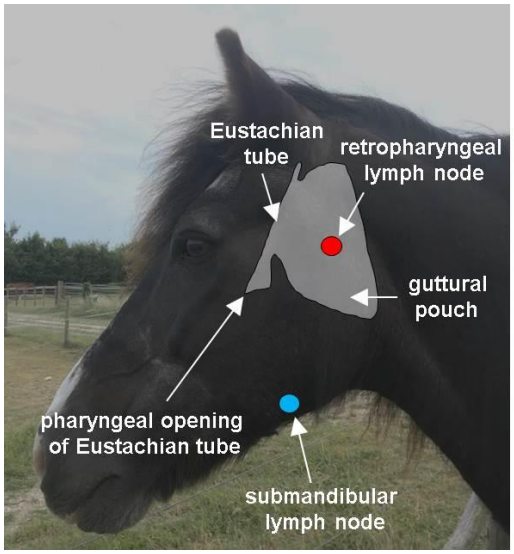


Figure 1.2. Schematic representation of the equine head anatomy highlighting structures relevant to infection with *Streptococcus equi*, causing strangles.

In an estimated 10 percent of cases, full drainage of retropharyngeal abscesses does not occur, leading to the retention of abscess material within the guttural pouch that can become dried, forming chondroids containing live *S. equi* [6]. Several chondroids can form and remain in the guttural pouch for several years without detection, creating sub-clinical carrier animals [6] (Figure 1.3).

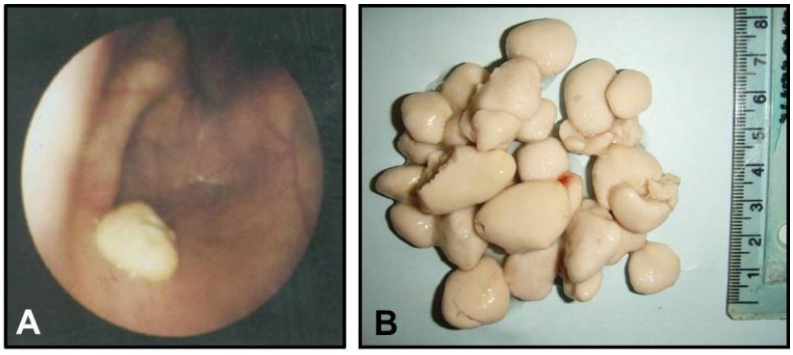


Figure 1.3. Chondroids associated with strangles A) Endoscopic view of a chondroid caused by an *S. equi* infection of the equine guttural pouch. Retention of abscess material in the guttural pouches from burst retropharyngeal lymph node abscesses leads to solidification of live *S. equi*, creating persistently infected carrier animals that often exhibit no clinical signs. B) Chondroids recovered from the guttural pouch of a Shetland pony. Taken from [7].

A carrier horse may continue to shed bacteria and the inadvertent mixing of seemingly 'healthy' carrier and naïve horses is likely to have contributed to the worldwide success of *S. equi*. In rare cases, abscesses may form in other organs of the body, mainly in the lungs, mesentery, spleen, liver, kidneys and the brain [8]. Rupture of these abscesses

can be fatal and is known as ‘bastard strangles’ [8]. A more common, and usually fatal, complication of strangles is purpura haemorrhagica [9]. Oedema of the limbs, eyelids and gums is seen and is associated with circulatory failure. An accumulation of circulating antibody complexes with the M-protein of *S. equi* within capillaries leads to this depletion of circulatory health [10].

1.3 *S. equi* virulence

1.3.1 Core Genome of *Se*4047

FimI and carbohydrate utilisation

In comparison to *S. zooepidemicus*, *S. equi* is impaired in terms of colonisation ability. *S. zooepidemicus* colonises the tonsillar tissue and mucosal surface of the nasopharynx, whereas a lack of nasopharyngeal detection of *S. equi* after infection suggests that *S. equi* does not colonise this area [4]. This reduced ability can be explained by the loss of sortase-processed cell surface proteins, which have been noted to enhance the adherence of streptococci to host tissues [11]. *S. zooepidemicus* was found to encode 39 of these proteins whereas *S. equi* encodes 29, with this loss of 10 including the FimII, FimIII and FimIV pilus loci [3]. *S. equi* does however contain a nonsense mutation within the gene encoding a transcriptional repressor of the FimI locus, which may result in deregulation and elongation of the pilus that may allow brief attachment to tonsillar tissue prior to invasion [3]. The loss of genes encoding key enzymes required for the utilisation of ribose, sorbitol and lactose may further restrict colonisation through a reduced ability to utilise available carbohydrates [3]. These alterations to the genome may have directed the selection of strains that can rapidly invade tonsillar tissue, in which rapid multiplication can then occur.

Hyaluronic acid capsule

A distinguishing feature of *S. equi* in comparison to most strains of *S. zooepidemicus* is the presence of a pronounced hyaluronic acid capsule. The enzymes required for capsule production are encoded by the *has* operon, which contains the genes *hasA*, *hasB* and *hasC* that encode hyaluronate synthase, UDP-glucose dehydrogenase and UDP-glucose pyrophosphorylase, respectively [12]. Compared to *Sz*H70, a small intra-replichore inversion exists in *Se*4047 between the 2 copies of *hasC*, causing rearrangement and potential alteration of gene regulation [3]. This inversion may be responsible for the more substantial capsule seen with *S. equi*. The enhanced capsule

may confer greater protection from the host immune system by evading complement activation, opsonisation and therefore, phagocytosis. The enhanced capsule may however limit movement, by reducing attachment to the trachea [3]. In support of this, a *S. equi* strain lacking *hasA*, and therefore the capsule, was significantly improved in its ability to attach to tracheal explants compared to the parental wild type strain [13]. The lack of dissemination of *S. equi* to the rest of the body, beyond the lymph nodes of the head and neck, may therefore be explained by the presence of its substantial capsule.

The presence of the enhanced capsule in *S. equi* could be explained by a lack of degradation or turnover, rather than increased production. *S. zooepidemicus* degrades its capsule through the production of hyaluronate lyases which are secreted enzymes that break down hyaluronic acid [3]. Capsule turnover may aid the progression of more widespread disease in *S. zooepidemicus* infections. *Se*4047 contains a 4 bp deletion within a gene encoding hyaluronate lyase (*hylA* (*SEQ1479*)), causing a frame shift and therefore a N-terminally truncated product. The truncated, N-terminally secreted protein, is hypothesised to be inactive against the capsule as it is depleted in the necessary substrate binding sites and catalytic residues required to degrade hyaluronic acid [14]. It is not known whether the adjacent C-terminal portion of this pseudogene is extracellularly active. Since truncation has removed the signal peptide required to mediate its extracellular transport, export is unlikely [14]. Hyaluronate lyases can also break down tissue, in addition to hyaluronic acid, and so this 4 bp deletion in *hylA* may also prevent *S. equi* from disseminating to the rest of the body.

S. equi may undergo genetic changes when adapting to its carrier state. The *has* locus is the most significantly affected area of the genome, namely through deletions and duplications. Seven unique deletion variants and 8 unique duplications flanked by the insertion element IS3 were observed in carrier isolates [15]. Single nucleotide polymorphisms (SNPs) and nonsense mutations were also identified. The strain JKS551, isolated from a persistently infected horse contained a deletion of *hasA* and a deletion in the 3' end of *hasB*, resulting in a lack of *hasA* transcription and therefore lack of capsule. It could be hypothesised that in the guttural pouch, *S. equi* no longer benefits from the protective function provided by the capsule, and therefore it may be favourable to cease production of this potentially metabolically costly locus. Another strain, 851, isolated from a persistently infected horse, contained an amplification of the *has* locus, enhancing capsule production [15]. The amplification may have been short lived however, as all 3 later isolates recovered from the same animal, including 1 recovered just 12 days later than 851, contained wild-type *has* locus. It is hypothesised that increased capsule production shortly after infection assists immune evasion, when the equine immune system may be most active, but as the immune response diminishes, the production of

the hyaluronic acid capsule may become dispensable to *S. equi*, progressing the bacteria to a persistent state. Harris et al., noted that it is also possible that this animal contained more than one *has* variant at each sampling date [15].

M-like proteins

M-like proteins enable the binding of fibrinogen to assist evasion of phagocytosis. These proteins are cell wall associated and are capable of preventing the accumulation of the complement component C3b onto the bacteria's surface, through the binding of fibrinogen to block C3b binding sites [16, 17]. An intact complement pathway is required for neutrophil chemotaxis with deficiencies in the pathway leading to impaired activation of phagocytes [18].

One M-like protein produced by *S. equi*, SeM, is unique to this bacterium, with another, *Sz*P*Se*, 85 percent homologous to the M-like protein found in *S. zooepidemicus* (*Sz*P) [19]. The absence of SeM in *S. zooepidemicus* is supported by the action of antiserum raised against SeM, as it increases opsonisation of *S. equi*, but not of *S. zooepidemicus* [19]. On the other hand, antiserum raised to *Sz*P*Se* increased the opsonisation of *S. zooepidemicus*, confirming the similarity of *Sz*P*Se* and *Sz*P [19]. In *S. equi* strain CF32, a SeM deficient mutant that expressed only 4 percent of the SeM of its wildtype counterpart, survival was reduced by 100-fold in equine blood, compared to the parental strain [16].

The SeM protein has been the focus of *S. equi* epidemiological studies [20] as its N-terminus has been shown to be highly variable between strains [19, 21]. Currently, 72 variants of SeM have been identified which have been collated on an online SeM database (https://pubmlst.org/databases/) (Access date: 24/06/18).

IgG endopeptidases

Two immunoglobulin G (IgG) endopeptidases, IdeE and IdeE2 exist in *S. equi* [22, 23]. These share amino acid homology with IdeZ and IdeZ2 of *S. zooepidemicus* [22] and provide protective immunity when included as recombinant proteins in a strangles vaccine [24, 25]. These enzymes cleave IgGs produced by the host, reducing their effectiveness, dampening the responsiveness of the immune response towards the bacteria due to the lack of opsonisation [22, 23]. The predominant role of IdeE is debated though as it is suggested that it also acts as a bactericidal agent against neutrophils [26].

Fibronectin-binding proteins

S. equi produces 4 fibronectin-binding proteins, 1 of which, FNE, is truncated. Before truncation, a LPxTG motif existed at the C-terminal end of the protein to enable anchorage to the cell wall [27]. The truncation removed this region, causing FNE to be secreted whilst the fibronectin-binding ability was retained [27]. FNE can bind to the gelatin-binding domain of human fibronectin via a thioester bond, tethering it to the extracellular matrix of host cells [28]. It has been speculated that a thioester bond in the structured region of FNE may be capable of simultaneously attaching to another compound of the extracellular matrix or to another fibronectin domain [28]. Binding of FNE to host cells in this way may cause contraction, creating space within the lymphoid tissue for *S. equi* to proliferate and generate the typical foci of infection seen in infected lymph nodes.

Streptolysin S

Streptolysin S (SLS) is an extracellular toxin produced by both *S. equi* and *S. zooepidemicus*. SLS is the cause of the characteristic clearing of β-haemolysis observed surrounding colonies grown on blood agar [29]. SLS degrades host cells and may contribute to immune evasion, and/or nutrient acquisition. In *S. equi* and *S. zooepidemicus* this toxin has high homology with that produced by *S. pyogenes*, which destroys host cells of many types [30]. A non-β-haemolytic strain of *S. pyogenes* was isolated from a case of human soft tissue infection, in which the bacteria showed alteration to the *sagC* gene within the SLS operon [31]. This mutant was found to contain a premature stop codon within *sagC*, resulting in the loss of haemolytic activity [31]. This case of infection was severe, which suggests that SLS is not required to cause this severity of disease in *S. pyogenes*, which may also be true of *S. equi*.

Superoxide dismutase

The majority of reactive oxygen species (ROS) bacteria are exposed to are generated endogenously, however, host cells such as neutrophils and macrophages produce ROS to aid in the killing of phagocytosed bacteria [32]. Some bacteria are able to resist the action of host ROS by neutralisation. *S. equi* and many other bacteria produce superoxide dismutase (SOD) which reduces the toxicity of ROS by converting it to hydrogen peroxide and oxygen [33]. The gene *sodA*, encoding SOD has been identified in *S. equi* [34], although more in-depth functional studies of SOD within this bacterium are lacking. In *S. pyogenes*, SOD has been identified on the bacterial surface and is secreted into growth medium, suggesting that it functions exogenously [35, 36]. Studies in *S. agalactiae*, have shown that a *ΔsodA* mutant was highly susceptible to oxidative

stress, supporting the putative anti-oxidant properties of SOD [37]. This mutant was also more susceptible to macrophage killing which reflects the dampened ability to neutralise the free radicals produced by phagocytes. The use of SOD by *S. equi* may assist its ability to survive within phagocytic cells [38].

SeCEP

The sortase-processed cell envelope proteinase (SeCEP) in *S. equi* shares 59 percent homology with SpyCEP of *S. pyogenes* [39]. SpyCEP inactivates the neutrophil chemoattractant interleukin 8 and other CXC chemokines [39, 40]. The inactivation of interleukin 8 occurs through the removal of its C-terminal α-helix, resulting in the interruption of phagocyte recruitment and therefore immune evasion [39]. The homology between SeCEP and SpyCEP may result in similar functions, yet the 41 percent dissimilarity may indicate differing efficacy of these proteinases. Strangles is characterised by a massive influx of neutrophils to the infected lymph nodes and so this suggests reduced efficacy or an alternative function of SeCEP in comparison to SpyCEP.

1.3.2 Mobile genetic elements of *Se*4047

The *S. equi* genome contains mobile genetic elements (MGE), such as prophages and integrative conjugative elements (ICE), which make up 16.4 percent of the total *Se*4047 genome [3] (Figure 1.4). *S. equi* is polylysogenic containing 4 prophages; ϕSeq1, ϕSeq2, ϕSeq3 and ϕSeq4 [3] (Figure 1.4). Prophages can carry cargo genes that can increase the survival and fitness of lysogens, which may have led to the enhanced niche adaption of *S. equi* [41].

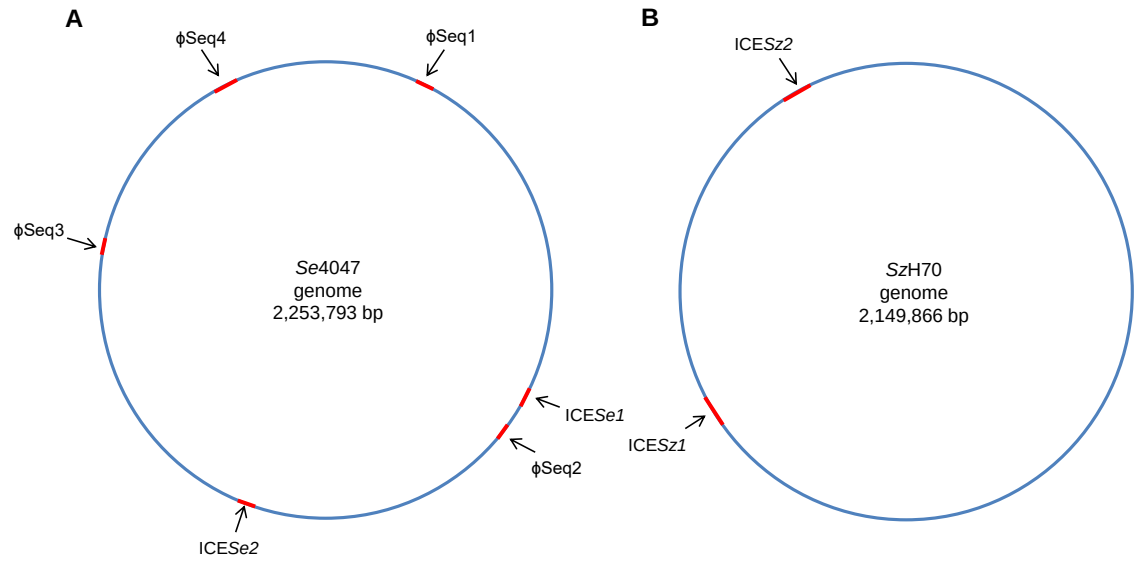


Figure 1.4. Mobile genetic elements in the genomes of *Se*4047 and *Sz*H70. The genome of *Se*4047 contains 4 prophage (ϕSeq1, ϕSeq2, ϕSeq3 and ϕSeq4) and 2 ICE elements (ICE*Se1* and ICE*Se2*). *Sz*H70 contains 2 ICE elements (ICE*Sz*1 and ICE*Sz*2), but does not contain any prophage. Adapted and redrawn [3].

As previously described, hyaluronate lyase in the core genome of *S. equi*, is truncated and likely inactive against the capsule. However, the integration of ϕSeq4, which carries the gene *SEQ2045*, produces a secreted hyaluronate lyase, as specific antibodies were identified in convalescent serum from naturally infected horses [14]. Another putative hyaluronate lyase is encoded on ϕSeq2, *SEQ0837*, but has not been investigated. It could be hypothesised that secretion of these enzymes may assist in phage penetration through the capsule of *S. equi*, allowing the acquisition of MGEs.

Phospholipase A2

Integration of ϕSeq2 into the *S. equi* genome resulted in the acquisition of *slaA*, which encodes a putative phospholipase A_2 toxin [3]. SlaA was present in all sequenced *S. equi* isolates and in 31 percent of *S. zooepidemicus* isolates [3]. SlaB, a second putative phospholipase A_2 toxin, however was encoded by all *S. equi* and *S. zooepidemicus* strains sequenced, being situated next to a phage remnant, suggesting that it was previously carried by a prophage [3]. Phospholipase A_2 toxins hydrolyse host fatty acids at the *sn*-2 position, generating a lysophospholipid and a free fatty acid which affect multiple signalling pathways and can mediate host inflammatory responses [42]. Lysophospholipids can bind to G-protein coupled receptors on the surface of immune cells, inducing cytoskeleton rearrangement, proliferation, differentiation and chemotaxis [43].

SlaA and SlaB share 98 and 70 percent amino acid identity with SlaA of *S. pyogenes* M3 MGAS315, respectively [3]. These toxins represent major virulence factors, with the acquisition of SlaA in *S. pyogenes* resulting in increased morbidity and mortality in humans, increased tissue destruction and dissemination in the murine model of infection [44, 45]. A Δ*slaA* deletion mutant in *S. pyogenes* serotype M3 was reduced in its ability to colonise the respiratory tract in a non-human primate model of pharyngitis [45]. In contrast, a *S. equi* Δ*slaAB* double deletion mutant was not significantly attenuated *in vivo*, although ponies did produce less nasal discharge, supporting a previously described link between phospholipase A_2 toxins in mucus formation in humans [46, 47].

Superantigens

Both ϕSeq3 and ϕSeq4 carry genes that encode superantigens [3]. The cargo of ϕSeq3 encodes SeeL and SeeM, and that of ϕSeq4 encodes SeeH and SeeI [3]. These superantigens initiate an inappropriate host immune response as they bind, as whole proteins, to regions of MHC class II molecules that are aside from the typical peptide binding site and to T-cell receptor Vβ chains [48]. This leads to the non-specific activation of approximately 5-20 percent of the host's T-cell population rather than an antigen-specific response [49]. Inappropriate activation of such a large non-specific immune response is thought to assist immune evasion.

Equibactin

ICE*Se*2 is an additional MGE in *S. equi* that carries cargo genes [3]. ICE*Se2* encodes a locus of 14 genes which transcribe a non-ribosomal peptide synthesis system for iron acquisition. Iron is essential to the growth of pathogenic bacteria [50]. This non-ribosomal peptide synthesis system is unique to *S. equi*, compared to other streptococci, and produces a secreted molecule, provisionally named equibactin, which has been shown to acquire iron *in vitro* [51] and be required for full virulence of *S. equi in vivo* [15]. Equibactin is likely to assist the acquisition of iron in a nutrient deficient environment, such as a lymph node (Figure 1.5).

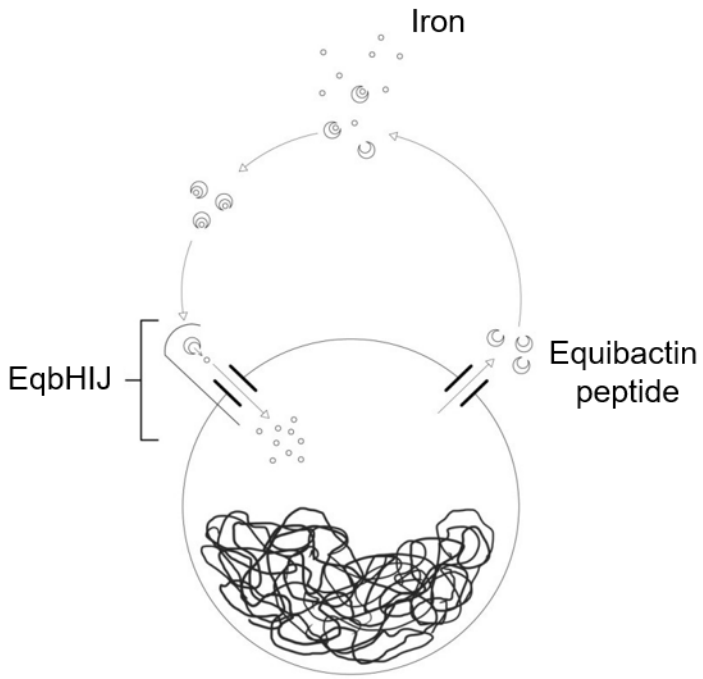


Figure 1.5. The iron acquisition system in *S. equi*. The equibactin locus encoded on the mobile genetic element ICE*Se2*, enables iron acquisition from the environment. *eqbH*, *eqbI* and *eqbJ* encode subunits of a putative energy coupling factor transporter [51]. Redrawn from [52].

The persistence of *S. equi* in carrier horses may be enhanced by the loss of the equibactin locus. Two isolates collected from a persistently infected horse contained a deletion of a 39.5 kb region in ICE*Se*2, which included the equibactin locus [15]. Deletions of varying extents in the equibactin locus were identified in 3 other persistently infected carrier horses involved in separate, unrelated strangles outbreaks, whereas none of the 78 acute isolates sequenced contained any deletions in this locus [15].

1.4 Current strangles vaccines

A strangles vaccine currently licensed for use in the USA, Canada and New Zealand, Pinnacle IN (Pfizer Animal Health), was created by random chemical mutagenesis to generate non-encapsulated mutants [53]. The vaccine is based on the strain CF32, which was originally isolated in 1981 from a horse in New York. The mutants were prone to back mutations and reversion to virulence, so further deletions to the 3' terminal end of *hasA* and the 5' terminal end of *hasB* were made to improve genetic stability [12]. When injected intramuscularly or in conjunction with other vaccines, adverse reactions are observed, such as the development of injection site or lymph node abscesses and nasal shedding [54]. Two horses, previously vaccinated with Pinnacle IN in fact

developed strangles caused by the Pinnacle IN *S. equi* strain, as confirmed by sequencing [55]. Pinnacle IN is not currently licensed for use in the UK due to these safety concerns.

5-enolpyruvylshikimate-3-phosphate synthase (*aroA*) is involved in the production of chorismic acid from shikimic acid, an important intermediate in the synthesis of aromatic amino acids [56]. A vaccine licensed in Europe, Equilis StrepE (MSD Animal Health), contains an *aroA* deletion mutant of the strain TW, which was originally isolated from a horse in the Netherlands in 1990 [21]. Administered intramuscularly as a vaccine, the *aroA* mutant conferred 100 percent protection although it caused severe injection site reactions, resulting in abscess formation [57]. The injection site was altered to the upper lip, leading to 100 percent and 50 percent protection of horses from lymph node abscess development in 2 independent experiments [57]. Attempts were made to improve this vaccine through further attenuation. SLS and capsule mutants were prepared but no benefits were observed either intramuscularly or intranasally [57]. A SLS/capsule double mutant was found to be strongly attenuated in mice yet caused strangles in yearling horses, which was confirmed by isolation of the mutant from the induced lymph node abscess (A. Jacobs, unpublished data). As with Pinnacle IN, adverse reactions were seen when administering alongside other vaccines [58]. In addition, it has been noted that injection of Equilis StrepE into a horse with diarrhoea caused submandibular lymph node abscesses [58]. These reactions reflect the caution required when using this vaccine. Often, stable yards will plan vaccinations around veterinarian visits to minimise call out fees and so horses may receive multiple vaccines in close succession. With Equilis StrepE, this vaccine should be administered to healthy horses and separately to any other vaccine to minimise the risk of adverse reactions.

The above vaccines do not permit the differentiation between vaccinated and infected horses (so-called DIVA). A live attenuated vaccine is currently in development at the Animal Health Trust, with initial studies proving promising. Six deletions were made to the *S. equi Se*4047 strain, removing the genes; *sagA*, *hasA*, *aroB*, *pyrC*, *seM* and *recA* [13]. Eighteen Welsh mountain ponies were utilised in this study, 9 ponies received the vaccine strain by intramuscular infection, and 9 control ponies received blank growth medium (Todd-Hewitt broth containing foetal calf serum) [13]. Adverse injection site reactions occurred in 4 ponies administered with the vaccine strain with up to 30 mls of abscess material recovered from the animals. Sequencing confirmed the vaccine strain as the cause. Two ponies were removed from the study over welfare concerns. Fifty-six days later, a booster vaccination was administered to the 7 remaining ponies. A small injection site reaction was evident in 1 pony not affected by the 1st vaccination. The 7 vaccinated ponies and the 9 control animals were challenged 52 days post 2nd

vaccination with wild-type parental *Se*4047 strain. Six of the control ponies exhibited clinical signs of disease, whilst all vaccinated ponies remained clinically healthy with no signs of disease. All 9 control ponies had at least 1 lymph node abscess and only 1 vaccinated pony had 1 lymph node abscess. Deletion of *seM* from the vaccine strain provides this experimental vaccine with DIVA potential, as SeM is 1 of the targets utilised in a diagnostic iELISA [59]. Development of a DIVA capable vaccine will aid the movement of horses as owners would be able to prove that their animals are vaccinated and not infected. A DIVA capable vaccine would also enable sub-clinical animals exposed to *S. equi* after vaccination to be identified. Vaccinated ponies developed antibody responses 2 weeks after the 2nd vaccination to SeeH, SeeI and SEQ2190, but not SeM as expected [13]. However, the antibody responses of vaccinated ponies did not increase post-challenge as would be expected after immune-priming [13]. Therefore, further research into this promising developmental vaccine is required to improve the duration of immunity and to prevent the formation of injection site reactions.

Another strangles vaccine with DIVA potential is Strangvac, formerly Septavac. Septavac contained 5 surface antigens (EAG, CNE, SclC, SEQ0256 and SEQ0402) and 2 secreted proteins; IdeE and IdeE2. Seven Welsh mountain ponies were vaccinated with Septavac which were significantly protected from challenge with *Se*4047, exhibiting reduced lymph node swelling and abscessation, fewer pyretic days, reduced pathology scores and lower bacterial loads within lymph nodes [24]. Seven unvaccinated control ponies became infected after challenge. Another 7 ponies were vaccinated with just the 5 surface antigens, omitting IdeE and IdeE2. Protection was reduced, demonstrating the immunogenic nature of these IgG endopeptidases [24]. To mitigate against the cost of manufacturing recombinant proteins individually, Septavac was developed further into Strangvac, where 8 antigens were combined into 1 single and 2 fusion recombinant proteins [25]. EAG, CNE, SclC, SclF and SclI were fused together to create CCE, SEQ0402 and SEQ0256 fused to make Eq85, with IdeE remaining unfused [25]. Sixteen Welsh mountain ponies were vaccinated with this formulation of Strangvac, 3 of which became pyretic after challenge with *Se*4047, whereas all 16 control animals exhibited pyrexia post-challenge [25]. Retropharyngeal lymph nodes from 8 vaccinated ponies and 15 control ponies contained *S. equi*, although bacterial loads in the vaccinated ponies was significantly lower than the controls [25]. No adverse reactions were observed in the course of this study and none of the proteins used are included in the diagnostic strangles iELISA, providing this vaccine with possible DIVA potential [25, 59]. Antibody levels were only measured 2 weeks post vaccination and therefore duration of immunity is not known for this vaccine. Another formulation of Strangvac contained the same proteins as above,

except that IdeE2 was also included. No significant differences in protection were seen between the formulations, warranting the removal of IdeE2 from further studies [25].

The 2 potential DIVA capable vaccines described, 1 live attenuated and 1 subunit, protect Welsh mountain ponies from strangles. Both vaccines however have issues that need further investigation before a protective, safe and efficacious vaccine can be brought to market. In the case of the live attenuated vaccine, further gene deletions that reduce the pathogenicity of *S. equi* further, should eliminate the formation of injection site reactions.

1.5 Transposable elements

Transposable elements (TEs) are found ubiquitously in nature, existing in eukaryotic and prokaryotic genomes. TEs are fragments of DNA ranging in size, which like MGEs, are mobile within a genome when active, enabling mutagenesis through transposition. It is not uncommon for TEs to be inactive as a result of deletion or mutation which prevents transposition [60]. The insertion of TEs has contributed to the evolution of genomes as this can either enhance or disrupt a gene, with protein coding genes as common targets [60]. TEs could be viewed as epigenetic regulators that can influence gene expression, improving or decreasing fitness, ultimately driving genomic evolution.

Transposons and insertion sequences (ISs) are both TEs that have received great attention from the scientific community. There are 2 classes of transposons, with that of class I representing retrotransposons. These TEs are mobile through a 'copy and paste' mechanism that is accomplished by RNA transcription of the transposon followed by DNA transcription using reverse transcriptase [61]. The 'copied' strand can then be inserted into a new site. Class II transposons follow a 'cut and paste' mechanism in which the transposon is excised from its current position and relocated into another [60].

ISs are short in length (700-1,800 base pairs) and encode 1 or 2 genes exclusively for transposition, with a transposase gene normally constituting the vast majority of the IS [62]. Transposase is an enzyme that recognises inverted terminal repeats that flank the IS (Figure 1.6), enabling the IS to be excised from its current position and integrated into a new site [60].

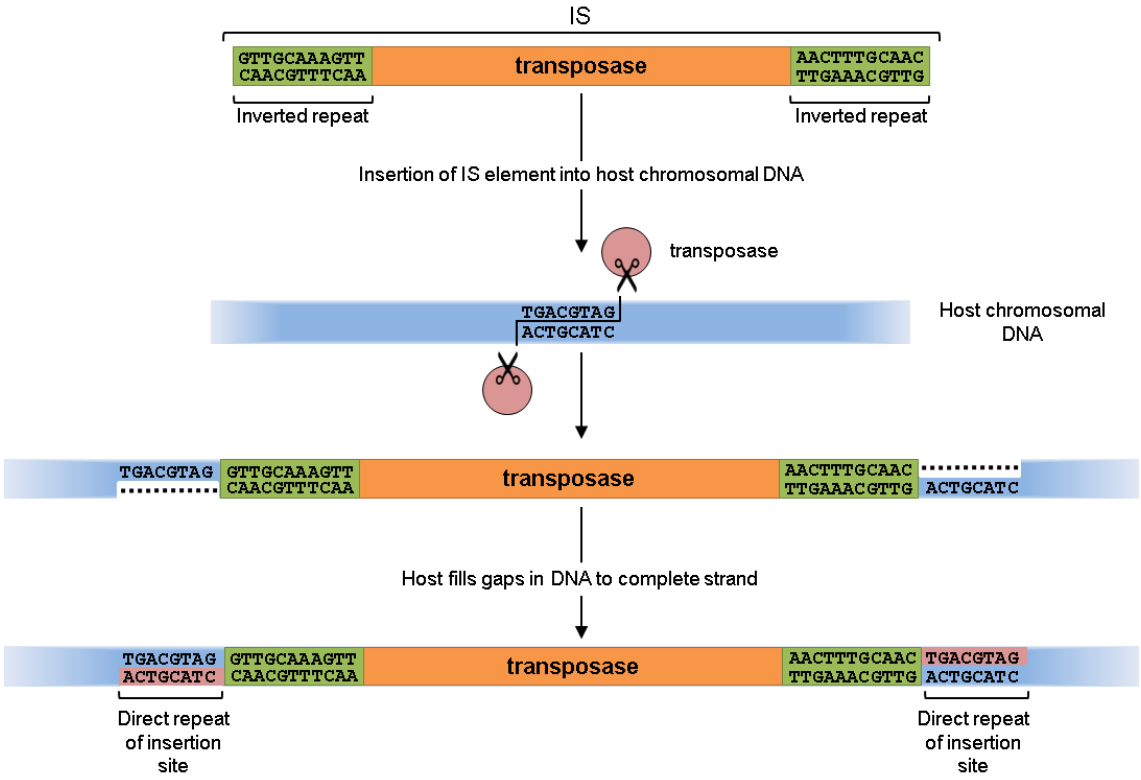


Figure 1.6. Transposition of class II insertion elements into host DNA.The transposase encoded by the transposon, cuts host DNA enabling insertion of the transposon into the host chromosome. Gaps in the DNA are filled using host mechanisms, generating direct repeats that flank the integrated transposon.

The inverted terminal repeats range in length from around 15-25 bp and are flanked by short direct repeats that mark the insertion site [63]. Upon synthesis or even during translation, transposase binds non-specifically in close proximity to the transposase gene from which it was transcribed [64]. Following this, the transposase will scan the DNA molecule for its target site, to which it binds, using one-dimensional diffusion [64-66].

Undeniably, IS elements have aided the evolution of *S. equi* from *S. zooepidemicus*. In *Sz*H70, 30 IS elements exist whereas *Se*4047 contains 73 IS elements, supporting their involvement in the speciation of *S. equi* [3]. Many families of IS exist with the family IS3 being that which has expanded the most within the *S. equi* genome. Within *S. zooepidemicus*, 4 IS3 sequences were identified compared to 40 found in *S. equi* [3]. The increased number of ISs is associated with the inactivation of more genes in *S. equi* compared to that in *S. zooepidemicus*, which has most likely directed its host-restriction.

1.6 Transposon mutagenesis

The discovery of TEs has enabled the generation of new research tools, particularly focused on the assignment of gene function. Experimentally inducing TE transposition has enabled researchers to create mutant bacteria for use in a range of studies. It can enable the production of mutant libraries in which the relevant bacteria is transformed with a TE, usually incorporated into a vector, which can lead to the generation of many thousands of mutant strains in 1 pool. In transposon libraries, viable mutants capable of replication contain TEs in so-called non-essential genes, with insertions into essential genes proving lethal.

1.6.1 Signature tagged mutagenesis

Signature tagged mutagenesis (STM) is one such technique, originally described in *S. typhimurium*, where each transposon mutant is tagged with a unique DNA sequence [67]. To generate the unique tags, a 40 bp variable region of DNA flanked by *Hind*III sites and 2 20 bp invariable regions were inserted into the mini-Tn*5* transposon [67]. The 2 invariable regions enable PCR amplification. These uniquely tagged transposons were individually transformed and randomly integrated into *S. typhimurium* DNA on a suicide vector. Each tagged transposon mutant was then transferred into a well of a microtiter plate for storage. Before experimental use, mutants were combined to form the 'input pool', a sample taken for DNA extraction and mutants exposed to a desired condition. Mutants were then recovered from the condition on agar plates and DNA extracted, representing the 'output pool'. DNA from the input and output pools were PCR amplified using primers specific to the invariable flaking regions of the transposon and a radiolabelled probe. PCR products were digested with *Hind*III to release the labelled products, which were subsequently hybridised onto colony blots from the individual tagged transposon mutants, one set for the input DNA and another for the output DNA. Hybridisation signals were compared between the input and output DNA bound blots to determine mutant survival. A loss of signal for a particular well in the output blot indicated a mutant reduced in virulence under the experimental condition tested. The limitations of this method are that the diversity of the combined mutant pool is restricted by the potential variations of unique tags and the time burden of generating so many tagged transposons. Additionally, STM does not accurately quantify mutant prevalence.

1.6.2 Transposon junction sequencing

The development of next-generation sequencing (NGS) technologies has improved on the STM technique, by enabling transposon mutants to be simultaneously sequenced without the need for unique transposon tags. The role of the STM tag is replaced by the identification of the transposon-genome junction alluding to the exact insertion site, which

by nature will be unique for each mutant. Negating the need for tagged transposons removes the limitations on library size, meaning that very dense mutant libraries can be generated and utilised to accurately identify essential genomes. Essential genes are evident from such data as no or few sequencing reads will be mapped to these genes. NGS has become extremely accessible and has fuelled the development of transposon-genome junction sequencing techniques such as TraDIS, Tn-seq, HITS, INSeq and PIMMS [68-72]. The precise details of these methods vary, yet all produce similar end-point data [73].

In recent years, a range of essential bacterial genomes have been published using transposon directed sequencing methods [68, 69, 72, 74-80]. Interrogating genomes in this way provides an unprecedented insight into genome-wide fitness, especially when libraries are subjected to disease relevant conditions. Exposure of dense mutant libraries to specific experimental conditions takes the power of these techniques a step further, enabling relative fitness and genome-wide conditional essentiality to be determined.

In vitro studies have proved insightful with a range of conditions applied to determine mutant fitness, such as pulmonary colonisation in *Haemophilus influenzae* [71], bile tolerance in *Salmonella enterica* (*S. enterica*) [68], sporulation in *Clostridium difficile* [74] and survival of *S. pyogenes* in human saliva and blood [81, 82]. *In vivo* application of transposon libraries has also significantly improved the acquisition of novel information regarding the fitness of mutants both in model systems and natural hosts. TraDIS has been used in *S. enterica* to measure intestinal colonisation in chickens, pigs and cattle of transposon mutants [83], in multiple tissue infection in mice [84] and in *Acinetobacter baumannii* (*A. baumannii*) to determine mutant fitness in the leukopenic murine model of blood stream infection [85]. Tn-seq has been applied to murine *S. pneumoniae* lung infection and nasopharyngeal colonisation models [86] and to *S. pyogenes* in a murine model of soft tissue infection [87].

TraDIS/Tn-seq like techniques also have the capability to direct subunit vaccine design. In *Streptococcus suis* (*S. suis*) TraDIS was used to determine targets of such a subunit vaccine, by the identification of essential surface associated proteins in the presence of pig epithelial cells [88]. Five selected proteins were developed into a subunit vaccine that provided protection against experimental challenge with *S. suis* [88].

1.7 Transposon sequencing (Tn-seq) and Transposon directed insertion-site sequencing (TraDIS)

1.7.1 *mariner* transposition system

The techniques used to generate mutant libraries varies between studies, as methods specific to the relevant transposon and microbe are used. There is however some continuity between Tn-seq studies in that they usually use *mariner* family components, a class II transposon first identified in *Drosophila* [89]. The *mariner*-based transposon is generally comprised of an antibiotic resistance gene and inverted terminal repeats recognised by the transposase, which contain *Mme*I recognition sites. These *Mme*I sites facilitate library DNA fragmentation as *Mme*I cuts 20 bp downstream of the recognition site, in the genomic DNA. *mariner* transposases however, only recognise the dinucleotide TA in target DNA, restricting potential mutant library size, particularly in genomes of high GC content [62, 90, 91]. After *Mme*I digestion, adaptors are ligated onto the cut ends and fragments PCR amplified with a transposon specific PCR primer and a reverse primer specific to the 3' adaptor, which can be indexed if required. Amplified fragments are then sequenced using Illumina technologies via a custom sequencing primer.

The first published Tn-seq study constructed transposon mutant libraries in *Streptococcus pneumoniae* (*S. pneumoniae*) using the *mariner* derivative, *magellan6*, mediated by the *Himar1* C9 transposase [69]. *magellan6* was transposed into *S. pneumoniae* DNA via '*in vitro* transposition' which involves integration of the transposon, into linear fragments of extracted genomic DNA. The staggered DNA ends at the insertion site were filled using T4 DNA ligase and dNTPs. *magellan6* containing fragments were transformed into competent *S. pneumoniae* and integrated by double cross-over homologous recombination. *megallan6* encodes an antibiotic resistance gene that was used to select 6 pools of 25,000 successfully transformed mutants, which equates to an insertion approximately every 86 bp in *S. pneumoniae*.

Mutant libraries in *S. pyogenes* were generated by transforming competent cells of the target strain with the vector pKRMIT, which carries the *Krmit mariner* family transposon and the *Himar1* C9 transposase [78]. Transforming with such a vector is termed '*in vivo* transposition' as the transposase gene is transcribed directly from the vector post-transformation. Twenty libraries were generated and measured for randomness by arbitrary-primed PCR. *Krmit* insertion randomness was between 35-95 percent, varying widely between pools. Four of the most random libraries were selected and passaged 4 times. Analysis by Tn-seq found that the libraries contained many intact vectors, which restricted the accurate identification of *S. pyogenes* insertion sites by sequencing.

Overnight growth of the libraries however negated this effect and allowed identification of between 24,000-90,000 unique insertion sites for individual libraries (insertion every 20-76 bp).

The major benefit of using a Tn-seq system is the enzymatic fragmentation of DNA, via *Mme*I, generating fragments of a consistent size. These fragments, however, only contain 20 bp of genomic DNA, which may confound the mapping quality of subsequent sequencing reads to the reference genome, especially where repetitive sequences are evident. Additionally, using *mariner*-based transposons limits the potential library size since it can only recognise TA dinucleotides as insertion sites. This limiting factor may confound a truly genome-wide measurement of fitness as some genes may not be represented and may therefore be inaccurately identified as essential.

1.7.2 EZ-Tn*5* transposition system

Beyond *mariner* based mutant libraries in Tn-seq, other transposons have successfully been used to generate dense and random mutant libraries. TraDIS tends to utilise Tn*5* transposon derivatives due to availability of the EZ-Tn*5* kit (Epicenter Biotechnologies) and the lack of apparent insertion recognition site. The first published work using TraDIS was designed using an *S. enterica* mutant library made with a Tn*5* derivative transposon, containing a kanamycin resistance gene. The Tn*5* derivative was amplified to add EZ-Tn*5* transposase recognition sites, and subsequently incubated with the EZ-Tn*5* transposase to form what is known as the EZ-Tn*5* transposome. This transposome complex is the transposase bound to the transposon recognition site, ready to randomly cleave the target DNA after transformation, by *in vitro* transposition. Using this system in *S. enterica* enabled the generation of 370,000 unique mutants, equating to an insertion every 13 bp, on average [68]. This density was however achieved by combining smaller pools of mutants generated in separate batches. Each batch contained 42,000-146,000 unique mutants which was the sum of at least 10 unique electro-transformations. Thirteen of these batches were combined to generate the final dense mutant library, equating to >130 electro-transformation events, highlighting the relatively inefficient transformation of this system in *S. enterica*. The same EZ-Tn*5* transposome system was however used to generate a mutant library in *A. baumannii*, which contained 109,000 unique mutants (insertion every 37 bp), from 1 transformation event [85]. In TraDIS, mutant library DNA is randomly fragmented mechanically or enzymatically, instead of by targeted digestion as in Tn-seq. The remainder of the library preparation protocol remains comparable to Tn-seq.

The major benefit of using a TraDIS based system over Tn-seq is the unlimited insertion of the EZ-Tn*5* transposon, which appears to have no bias towards particular sequences

and is therefore in theory able to insert at every base pair. Fragmentation of the library DNA by mechanical and enzymatic means, however, produces irregular size fragments that slightly differ between library preparations, since DNA cutting is random by both methods. A mixture of different sized DNA fragments does make library quantification and subsequent sequencing a little more variable since an average, and not an absolute, fragment size must be taken into account.

As the popularity of Tn-seq and TraDIS-like techniques grow, the differences between them becomes blurred, with no concrete rules on what transposon and fragmentation method is to be used in each technique.

1.8 pGh9:IS*S1* transposition system

The thermosensitive plasmid pG+ host (pGh), a derivative of pWV01 [92], has proven useful in a range of studies from allelic replacement mutagenesis and controlled gene expression technologies to mutant library generation [93-98]. pGh can be used to generate transposon mutant libraries through the delivery of the transposable element IS*S1* [94-97]. pGh9 has become the derivative of choice for use with IS*S1* (pGh9:IS*S1*) and contains a Gram-positive *repA*+ temperature sensitive (ts) origin of replication, allowing replication at 30 ºC but not 37 ºC, and an erythromycin resistance gene, *ermB* (Figure 1.7). Other derivatives of pGh, including pGh5 and pGh8, suffered from problems with tandem transposition in *L. lactis* (Figure 1.8B). IS*S1* undergoes replicative transposition, resulting in the integration of the plasmid between 2 IS*S1* copies. This is known as a monocopy insertion despite the presence of 2 IS*S1* copies (Figure 1.8A).

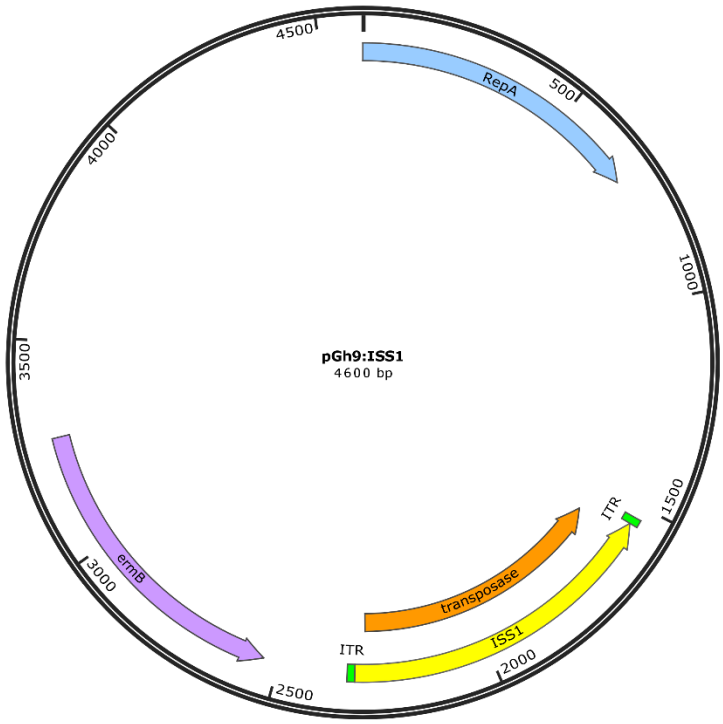


Figure 1.7. pGh9:IS*S1* map. pGh9 is a temperature sensitive plasmid used for the delivery of IS*S1*. The IS*S1* transposon is flanked by an 18 bp inverted terminal repeat (ITR) and an 8 bp direct repeat. pGh9:IS*S1* contains an erythromycin resistance gene (*ermB*) used to select for successful transposition. The Gram-positive temperature sensitive replicase, RepA is utilised by pGh9. Adapted from [94] using SnapGene®.

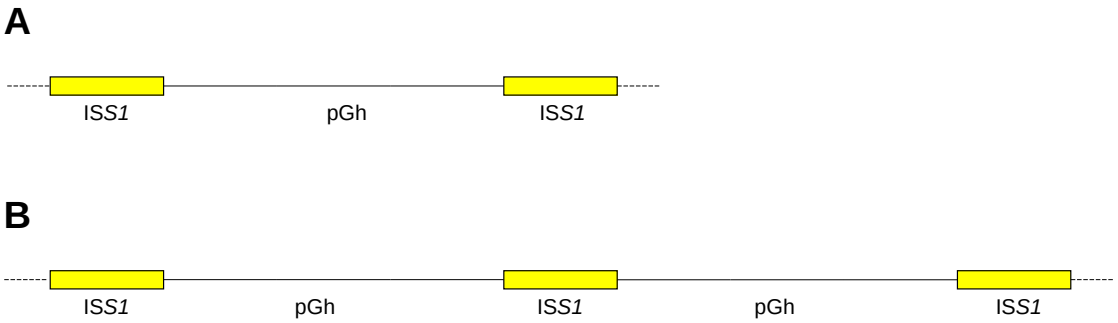


Figure 1.8. Transposition products of pGh:IS*S1*. (A) Monocopy transposition. The replicative transposition of pGh:IS*S1* most frequently results in monocopy insertion, being 1 copy of pGh flanked by 2 copies of IS*S1*. (B) Tandem transposition. IS*S1* insertion may also result in multiple copies of the transposed structure. Redrawn from [94].

Tandem transposition is hypothesised to occur as a result of poor antibiotic resistance marker expression, when inserted as a monocopy [94]. Another explanation is based on the production of linear plasmid multimers by pGh5:IS*S1* in *L. lactis* [94, 99], which is

associated with the presence of recombination hot spots referred to as Chi sites [100]. Following pGh re-design, pGh9:IS*S1* was improved to confer 80 percent monocopy transposition rate in *L. lactis* [94].

pGh9:IS*S1* has been successfully used to generate dense transposon libraries in *Streptococcus uberis* (*S*. *uberis*) without bias for any particular nucleotides [72]. A library containing 80,617 mutants was generated, equating to on average 31 unique insertions per gene. One-hundred and ninety-six genes contained no insertions, with 69 genes containing insertions that were limited to the last 10 percent of coding sequences. These 2 sets of genes, totalling 263, were identified as essential for the *in vitro* survival of *S. uberis* and contributed to known essential basic cellular functions such as protein and RNA metabolism and cell division/cell cycle. The unbiased, random and dense insertion of pGh9:IS*S1* into the *S. uberis* genome, suggests that libraries of the same quality could be generated in *S. equi*.

1.9 Transposon mutagenesis of *S. equi*

Transposon mutagenesis of *S. equi* has been attempted using the transposons Tn*916* and Tn*917*, with little success. When using Tn*916*, 3 insertion sites were identified in 1 clone of *S. equi* strain CF32, therefore it cannot be anticipated that single random mutagenesis events will occur, preventing insertion effects to be accurately measured [29]. Tn*917* inserted only once per clone, but preferentially inserted into a 15kb region of the *S. equi* genome [101]. This is recognised as the phenomenon termed 'hot spotting' in which a transposon preferentially inserts into a particular region(s) over others. Hot spotting makes transposon mutagenesis studies difficult as the data produced is biased towards the preferred region(s), with areas of the genome that the transposon inserts with less frequency being under represented.

Transposition in *S. equi* using the *Himar*1 mini-transposon and *Himar*1 C9 transposase in the vector pCAM45, has been attempted [102]. After electro-transformation, *S. equi* was recovered at 30 ºC as, like pGh, pCAM45 is under the control of *repA*+ ts. Transformants were maintained at 30 ºC for 2 days, followed by incubation at 37 ºC, the non-permissive temperature for pCAM45, in an attempt to cure *S. equi* of the plasmid. Curing transformants of the plasmid is thought to reduce the chance of transposon translocation as the *Himar*1 C9 transposase is removed with the plasmid, leaving only the integrated transposons within the chromosome. Analysis of the *S. equi* library generated with *Himar* 1, identified a single bp deletion within *repA*+ ts, which significantly decreased the number of unique mutants produced (2,500).

As a part of the MSc project preceding this PhD, a TraDIS method, adapted from that designed by Langridge et al., (2009), was applied to *S. equi* utilising pGh9:IS*S1* (Amelia Charbonneau MSc thesis, University of Aberystwyth, 2013). A *S. equi* mutant library of 115,951 unique mutants was generated, representing, on average, an insertion every 19 bp from 1 transposition event. These wide-spread insertions provide evidence for the random nature of insertion using the pGh9:IS*S1* system. The *S. equi* pGh9:IS*S1* system is further explored and developed in the course of this thesis.

1.10 Project outline

Using TraDIS to investigate the growth of mutant libraries under particular conditions has the power to greatly enhance knowledge of the functional genomics of bacteria, ultimately assisting future vaccine design. In *S. equi* many genes remain described as of unknown function. Even when putative function has been proposed based on sequence similarity, the importance of the protein encoded to the ability of *S. equi* to cause disease is rarely known. Therefore, the application of TraDIS could assist the identification and classification of genes of unknown function, in addition to the implication of other genes in certain cellular processes. Much remains to be uncovered regarding the interaction between *S. equi* and the host, warranting both *in vitro* and *in vivo* investigation to identify genes implicated in the antigenicity and pathogenesis of this economically important bacterium.

In this PhD project, a novel barcoded pGh9:IS*S1* TraDIS technique is described, in which a 2 bp barcode in the sequenced region of IS*S1* enables the combination of libraries, and their subsequent deconvolution from sequencing data. Barcoded libraries were initially used to determine the essential genome of *S. equi,* which was compared to that of *S. pyogenes* and *S. agalactiae* from published works. Three barcoded libraries were also exposed to hydrogen peroxide (H_2O_2) and whole equine blood, to simulate the interaction with the equine immune system. Sequencing of surviving mutants enabled the identification of genes important to *S. equi* under these conditions *in vitro*. To validate these results, 5 genes attenuated in H_2O_2 and/or whole blood, according to TraDIS, were deleted by allelic replacement mutagenesis, re-exposed to the conditions *in vitro* and survival measured.

TraDIS was additionally assessed in the susceptible natural host by the challenge of twelve Welsh mountain ponies with barcoded *S. equi* libraries. Each animal was infected with 2 of 3 barcoded libraries, abscess material recovered from retropharyngeal and submandibular lymph nodes, and sequenced by TraDIS to identify genes required for

infection. Applying the barcoded technique in this way reduced the number of animals required, whilst maximising the quality and robustness of the data obtained. To validate these data, twelve genes with attenuated fitness *in vivo* as a result of IS*S1* insertion were deleted by allelic replacement mutagenesis and used to challenge 5 additional ponies, in combination with the parental WT strain *Se*4047 and a negative control deletion mutant. All deletion mutants were designed to contain a short tag matching the sequencing primer binding site for TraDIS, to enable material recovered from ponies to be sequenced by TraDIS. The results of this *in vivo* study were compared to the *S. equi* whole equine blood and H_2O_2 data and to 3 *in/ex vivo S. pyogenes* studies to determine the overlap between *in vitro* and *in vivo* data and a potential pan-streptococcal gene set required for *in vivo* infection.

2 Defining the ABC of gene essentiality in streptococci

The following data was published in BMC Genomics in May 2017 [103]. Basic development of a non-barcoded TraDIS method was conducted in part for an MSc degree award, but all data presented in this chapter was generated during this PhD. The barcoding system and the plasmid depletion step of library preparation are new additions to the method, which have not previously been submitted for any degree award.

2.1 Introduction

S. equi is closely related to the group A *Streptococcus*, *Streptococcus pyogenes* (*S. pyogenes*) [3] and the group B *Streptococcus*, *Streptococcus agalactiae* (*S. agalactiae*) [104], both of which are important human pathogens. *S. pyogenes* causes impetigo, pharyngitis, scarlet fever and necrotising fasciitis [105-107] and *S. agalactiae* causes meningitis, pneumonia and sepsis in neonates [108], in addition to mastitis in cattle [109] and streptococcosis in fish [110]. Identifying genes required for the survival of these 3 streptococci will provide valuable information for defining the pan-streptococcal essential genome.

In this chapter, the development of a barcoded transposon directed insertion-site sequencing (TraDIS) system is described, which can be conducted using standard Illumina sequencer protocols. Dense mutant libraries utilising the plasmid pGh9 carrying the insertion element, or transposon, IS*S1* (pGh9:IS*S1*) [94], have previously been utilised with success in *S. uberis* [72]. In the study described in this chapter, the plasmid pGh9:IS*S1* was modified within the 5' terminal of IS*S1* to create 6 barcoded plasmids.

These 6 plasmids were used to generate 6 independent mutant libraries in *S. equi* strain 4047. The libraries were sequenced after growth in rich media, with data for each barcoded library compared and combined, providing a blue-print data set for the subsequent analysis of conditional fitness and gene essentiality assignment in *S. equi*.

The agreement of gene essentiality between the *S. equi* TraDIS data and Tn-Seq data from the close relatives *S. pyogenes* and *S. agalactiae* was determined. KEGG (Kyoto encyclopaedia of genes and genomes) pathways were attributed to the essential gene sets of *S. equi*, *S. pyogenes* and *S. agalactiae* to unveil the key biochemical pathways in which they are involved.

2.2 Materials and Methods

2.2.1 Bacterial strains, DNA isolation and primers

The *S. equi* strain *Se*4047 was used throughout this thesis, an isolate originally recovered from a New Forest pony with strangles in Hampshire in 1990 [3]. *S. equi* was grown at 37 ºC in a humidified atmosphere containing 5 percent CO_2, unless otherwise stated. The *Escherichia coli* (*E. coli*) strain TG1 *repA*+, supplied by Emmanuelle Maguin (Institut Nationale de la Recherche Agronomique, Jouy en Josas, France), was used for the replication of the plasmid pGh9:IS*S1* at 37°C. *S. equi* genomic DNA was extracted using GenElute spin columns (Sigma Aldrich) according to manufacturer's instructions, except that cells were lysed for 1 hour instead of 30 minutes. A table of all primers used in this study is available in Table A1.2 (Appendix 1).

2.2.2 Barcoding IS*S1*

In addition to the original pGh9:IS*S1* [94], another 5 barcoded variants were generated by mutating the 2 nucleotides (CA) located 3 and 4 bases downstream of the IS*S1* inverted repeat (Figure 2.1). The new plasmids: pGh9:IS*S1*:TC, pGh9:IS*S1*:AG, pGh9:IS*S1*:AC, pGh9:IS*S1*:CT and pGh9:IS*S1*:GA contained the alternative bases TC, AG, AC, CT or GA, respectively at these positions. For clarity, the original pGh9:IS*S1* will be referred to as pGh9:IS*S1*:CA.

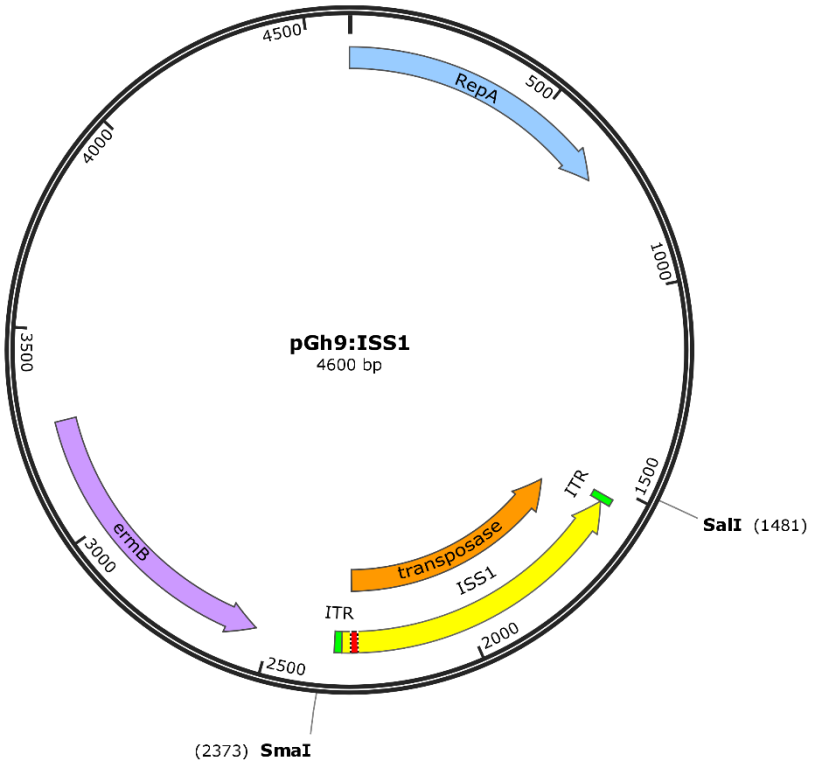


Figure 2.1. Barcoded pGh9:IS*S1* map. pGh9 is a temperature sensitive plasmid used for the delivery of IS*S1*. The IS*S1* transposase is flanked by an 18 bp inverted terminal repeat (ITR) and an 8 bp direct repeat. pGh9:IS*S1* contains an erythromycin resistance gene (*ermB*) used to select for transformation and transposition. The Gram-positive temperature sensitive replicase, *repA*+ ts is utilised by pGh9. A 2 bp barcode in IS*S1* is shown in red. Adapted from [94] using SnapGene®.

Plasmid digestion

One µg of pGh9:IS*S1* was enzymatically digested with the restriction enzyme *Sal*I, followed by *Sma*I (Figure 2.1), to remove IS*S1* according to the manufacturer's protocol for these restriction enzymes (New England Biolabs). Digestions were completed as single digests, with the first digestion cleaned up using the Qiagen nucleotide removal kit. The plasmid was treated with Antarctic phosphatase, according to the manufacturer's protocol (New England Biolabs) to prevent re-ligation. The phosphatased plasmid was electrophoresed on a 1 percent agarose gel alongside a DNA ladder (Bioline Hyperladder I) at 120 V for 30 minutes to separate the plasmid from IS*S1*. The plasmid band was excised from the gel and purified as per manufacturer's instructions (Qiagen gel purification kit).

Base substitution by PCR

Three Phusion (New England Biolabs) PCRs were completed to generate each barcoded IS*S1*. In PCR 1, the forward primer (P1) was used to amplify pGh9:IS*S1* from the *Sma*I site, with the reverse primer (P3) spanning the site of base substitution in IS*S1*, generating a 78 bp fragment (Figure 2.2). Primer P3 varied by 2 base pairs depending on the barcode to be introduced.

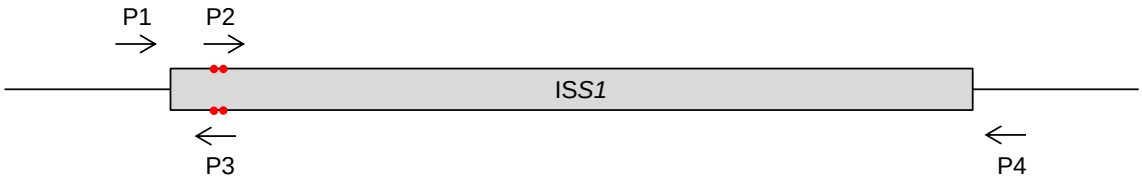


Figure 2.2. Primer binding sites for the generation of barcoded IS*S1*. Primers P1 and P3 and primers P2 and P4 are utilised in 2 separate PCR reactions to generate 2 fragments of the new barcoded IS*S1*. A recombinant PCR was conducted using the 2 products from the previous PCRs and primers P1 and P4 to generate the whole product. The red dots indicate the location of the 2 base pair barcode.

PCR 2 utilised a forward primer also spanning the base substitution site (P2), which varied by 2 base pairs depending on the barcode, and a reverse primer (P4) designed from the *Sal*I site within the plasmid, generating an 848 bp fragment (Figure 2.2). Both PCR products from PCR 1 and 2 were used in a recombinant PCR, PCR 3, using the primers P1 and P4, creating 1 product of 896 bp. All PCR products were purified using the Monarch® PCR & DNA cleanup kit (NEB). The recombinant PCR product was digested with *Sal*I, followed by *Sma*I as previously described and cloned back into the digested pGh9 overnight using T4 ligase according to the manufacturer's protocol (New England Biolabs) at a ratio of 8 times the recombinant PCR product to 1 digested pGh9.

Plasmid transformation

Barcoded plasmids were transformed into *E. coli* TG1 *repA*+, by mixing 20 µl of each ligation reaction with 80 µl *E. coli* cells on ice and incubated for 30 minutes. The reactions were incubated at 42 °C for 90 seconds before placing back on ice for 2 minutes. Nine hundred µl of Luria-Bertani broth (LB) was added to each reaction followed by incubation at 37 °C for 1 hour. One hundred µl of each reaction was spread on LB agar supplemented with 150 µg/ml erythromycin (LBE) to select for successful transformation as pGh9 contains an erythromycin resistance cassette, *ermB*. The remaining *E. coli* in LB was centrifuged at 10,000 rpm for 5 minutes, the supernatant removed and the pellet

resuspended in 100 µl LB. The 100 µl of resuspended pellet was also spread on LB erythromycin agar. All Petri dishes were incubated at 37 °C for 1-2 days.

PCR verification of ligation

Between 8 and 16 erythromycin resistant colonies were picked from each Petri dish and resuspended individually in 10 µl of double distilled water. Four µl of each colony suspension was amplified in a *taq* polymerase (Sigma Aldrich) PCR reaction using the primers P4 and 5'9, a pGh9 primer 86 bp downstream of IS*S1*, according to manufacturer's instructions (Sigma-Aldrich) to identify successful transformants. The PCR products were run on a 1 percent agarose gel at 120 V for 35 minutes alongside a DNA ladder (Bioline Hyperladder I). For successful transformants (band of 974 bp), the remaining colony suspension was added to 10 ml LB containing 150 µg/ml erythromycin. The cultures were grown for 16 hours in a shaking incubator (220 rpm) at 37 °C before centrifuging at 4500 xg for 10 minutes to generate a pellet. The plasmids were extracted from these pellets (Qiagen Plasmid Midi kit) followed by sequencing in-house on an ABI3100 DNA sequencer with BigDye fluorescent terminators, using the primers 5'9 and 3'9 (which span the cloning site utilised to ligate the barcoded IS*S1*), P1, P2, P3 and P4 at a concentration of 10 µM, in separate reactions.

2.2.3 Generation of IS*S1* libraries

Following confirmation of sequencing data, the plasmids pGh9:IS*S1*:CA, pGh9:IS*S1*:TC, pGh9:IS*S1*:AG, pGh9:IS*S1*:AC, pGh9:IS*S1*:CT and pGh9:IS*S1*:GA were used to generate mutant libraries in *S. equi*, which are herein referred to as CA, TC, AG, AC, CT and GA.

S. equi competent cells

*Se*4047 was grown on Todd-Hewitt agar (THA) for 16 hours. A single colony was picked into 5 ml Todd-Hewitt broth (THB) containing 0.03 µg/ml hyaluronidase and grown for 16 hours. An additional 45 ml of fresh THB containing 0.03 µg/ml hyaluronidase was pre-warmed and pre-gassed by incubation alongside the 5 ml overnight culture. After 16 hours, the 5 ml culture was transferred into the 45 ml pre-warmed TH. The diluted culture was incubated for 2-3 hours to ensure cells were in early log/log phase. The culture was centrifuged at 5 ºC at 10,000 xg until a pellet was formed. The supernatant was discarded and the pellet washed 3 times with 5 ml of cold 0.5 M sucrose, on ice. The pellet was resuspended in 100 µl of 0.5 M sucrose per 10 ml of culture and stored in 100 µl aliquots at -80 ºC.

Transformation of barcoded pGh9:IS*S1* into *S. equi*

*Se*4047 competent cells were transformed with the desired barcoded pGh9:IS*S1* plasmid by electroporation [98]. One hundred µl of *Se*4047 competent cells were defrosted on ice and combined with 4 µl of the desired barcoded pGh9:IS*S1*. A Gene Pulser electroporater (Bio-Rad, United Kingdom) set at 2.5 kV cm^{-1}, 200 Ω, and 25 µF, was used to induce transformation, which typically gave a pulse time of 4 ms. One ml of ice-cold THB was added to the electroporated cells which was incubated for 3 hours at 28 ºC, recovering the cells and allowing extrachromosomal plasmid replication. Transformants were grown on THA supplemented with 0.5 µg/ml erythromycin (THAE) for 3 days at 28 ºC. A colony of erythromycin resistant transformants was picked into THB supplemented with 0.5 µg/ml erythromycin (THBE) and grown for 16 hours at 28 ºC. Overnight cultures were heat shocked at 40 ºC for 3 hours resulting in random transposition of pGh9:IS*S1* into the bacterial chromosome. In order to calculate the transposition frequency of libraries, transposants were grown on THA and THAE for 16 hours. Frequencies were determined by counting the colony forming units per millilitre of transposants on THAE versus THA. Transposants were selected by overnight growth on 30 large (150 mm diameter) THAE Petri dishes supplemented with 0.03 µg/ml of hyaluronidase at a density of approximately 6,500 colonies per dish. Pools of random transposon mutants (transposon libraries) were harvested from the dishes by washing with THB containing 25 percent glycerol and the bacterial suspension stored at -20 ºC. Prior to sequencing, the transposon libraries were grown to an OD_{600nm} of 0.3 in THBE. Two and a half ml of the culture was centrifuged at 10,000 xg for 5 minutes and the bacterial pellet stored at -20 ºC in preparation for DNA extraction.

2.2.4 Stability of integrated pGh9:IS*S1*

Ninety-five colonies recovered from library CA were grown overnight in THBE, before they were combined to generate P0. The 95 mutant pool was passaged twice overnight under the same conditions to produce P1 and P2. Two and a half ml of each culture was centrifuged at 10,000 xg for 5 minutes, the supernatant removed and the bacterial pellet stored at -20 ºC in preparation for DNA extraction.

2.2.5 DNA preparation and sequencing by TraDIS

DNA was extracted from the 6 pelleted barcoded mutant libraries and the 3 stability library cell pellets using a GenElute column kit according to the manufacturer's instructions for Gram positive bacteria (Sigma-Aldrich). DNA was quantified using the Qubit dsDNA BR assay kit according to the manufacturer's instructions. One and a half µg DNA was fragmented by sonication using a Misonix XL 2020 Ultrasonic Liquid

Processor (cup horn arrangement) to produce fragments in the range of 200-800 bp, with 800 bp fragments being most prevalent. Y-adaptor was generated in-house using Illumina multiplexing adaptor sequences (Oligonucleotide sequences © 2007- 2012 Illumina, Inc. All rights reserved). To generate the Y- adaptor, 15 µl of both Adaptor primer 1 and Adaptor primer 2 were combined and incubated at 95 ºC for 2 minutes, followed by an incremental decrease in temperature by 0.1 ºC per second to 20 ºC. The reactions were chilled on ice before 70 µl of ice cold ultra-pure water was added to dilute the reaction to 15 µM. Y-adaptors were ligated to 1 µg of fragmented DNA using the NEBNext Ultra II DNA library prep kit for Illumina (New England Biolabs) according to the manufacturer's instructions for End Repair and Adaptor Ligation. Fragments were purified using AMPure XP beads (Agencourt, Beckman Coulter) with a bead to DNA ratio of 1:1, according to the manufacturer's instructions.

Incubation of adaptor ligated DNA with the restriction enzyme *Sma*I for 2 hours at 25 °C, according to the manufacturer's instructions, was used to cleave the pGh9:IS*S1* plasmid 33 bp upstream of the sequence encoding IS*S1* in order to minimise the amount of TraDIS reads mapping to plasmid. The Monarch® PCR & DNA cleanup kit (NEB) was used to purify digested DNA, according to manufacturer's instructions. The amount of DNA recovered was quantified using the Qubit dsDNA HS assay kit (Invitrogen) according to the manufacturer's instructions. As recommended by Langridge et al. [68], 100 ng of library DNA was PCR amplified for 20 cycles according to 1.4C of the NEBNext Ultra II DNA library prep kit protocol. Amplification utilised the specific IS*S1* primer and a unique indexing PCR primer per library, which facilitated the attachment of the resultant product to the sequencing flow cell. The regions that were amplified span the 5' end of IS*S1* and the site of transposition in the *S. equi* genome. The use of a Y-adaptor enabled amplification of IS*S1* containing fragments only, as reverse amplification could not occur until the specific IS*S1* primer had generated a complementary Y-adaptor sequence for the indexing PCR primer to bind (Figure 2.3)

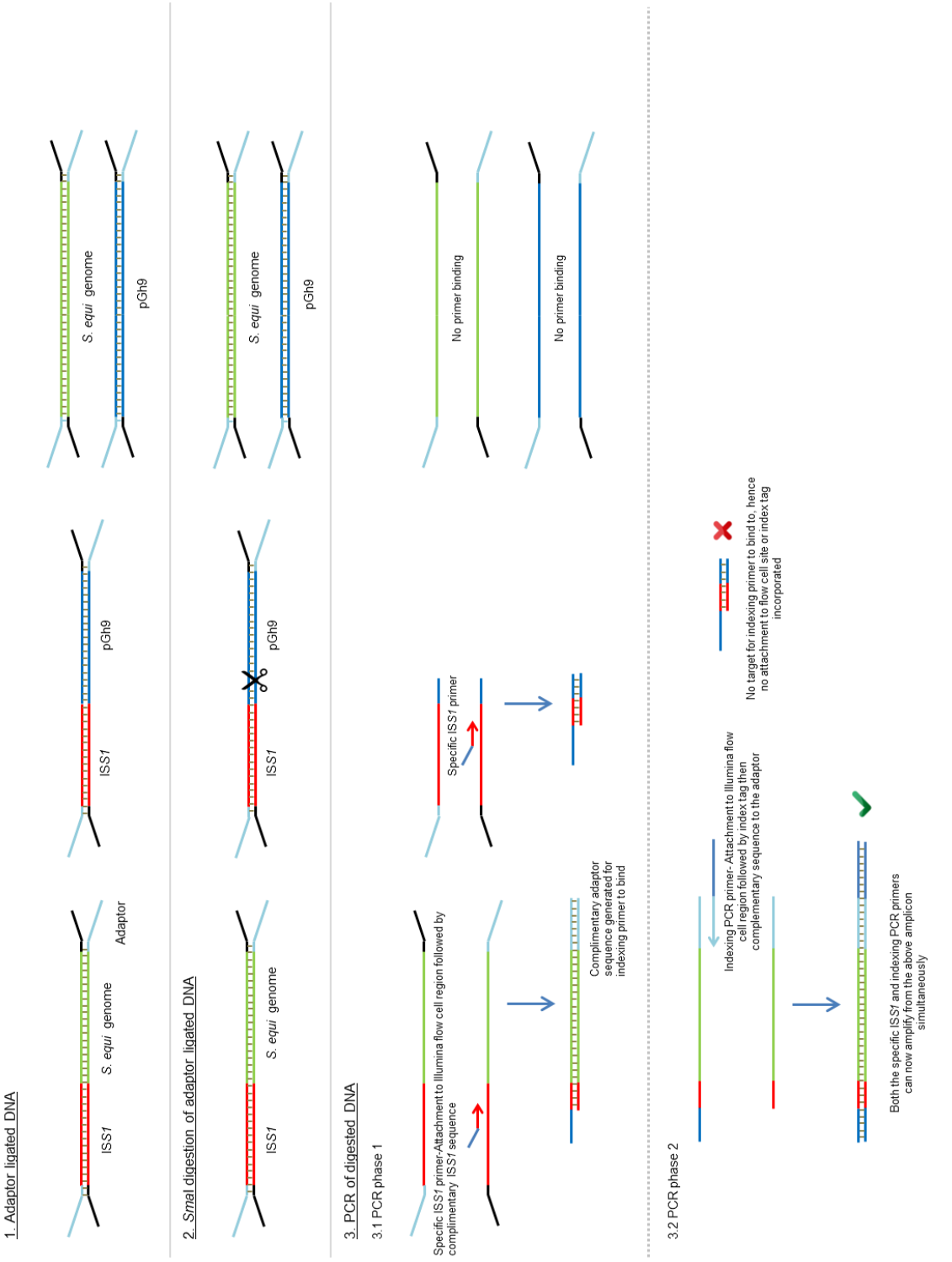


Figure 2.3. TraDIS PCR strategy. 1) Adaptor ligated DNA. Y-adaptors were ligated onto DNA fragments containing either the desired IS*S1*-*S. equi* genome junction, IS*S1*-plasmid (pGh9) junction, only *S. equi* genome or only pGh9 DNA. 2) *Sma*I digestion of adaptor ligated DNA. Undesirable IS*S1*-plasmid junction containing DNA is depleted by digesting all fragments with the restriction enzyme, *Sma*I. This enzyme cuts pGh9 at a restriction site 33 bp from the IS*S1*-plasmid junction, which is rare in the *S. equi* genome. 3) PCR of digested DNA. 3.1) PCR phase 1. A specific IS*S1* forward primer was designed to amplify from the 5' end IS*S1,* enriching for fragments containing an IS*S1* junction. Initial amplification with the specific IS*S1* primer generates an amplicon with a complementary adaptor sequence (shown in light blue). 3.2) PCR phase 2. The indexing PCR primer can now amplify from the complimentary adaptor sequence in the amplicon

generated by phase 1. After phase 2, both primers can simultaneously amplify the amplicon. This strategy ensures that no reverse indexing primer amplification can occur until the forward primer has specifically amplified from IS*S1*.

AMPure XP beads with a bead to DNA ratio of 0.8:1 were used to remove small PCR products, non-ligated adaptors and primer dimers. The concentrations of the libraries were calculated using the Kapa Biosystems library quantification kit, with average fragment sizes estimated by running the libraries on a 1 percent agarose gel at 120 V for 40 minutes alongside a DNA ladder (Bioline Hyperladder I). The amplified libraries were single-end sequenced using the Illumina MiSeq. Libraries CA, TC and AG were uniquely indexed and sequenced on 1 MiSeq run, as were libraries AC, CT and GA. The stability libraries were sequenced on 1 MiSeq run and were also uniquely indexed. All libraries to be sequenced on 1 MiSeq run were diluted to 2 nM and combined at equal concentrations to generate a pooled library for sequencing. Five µl of the pooled library at 2 nM and PhiX at 2 nM (Illumina) were individually combined with 5 µl of 0.1 N NaOH and incubated at room temperature for 5 minutes to denature the DNA. To neutralise the denatured DNA, 990 µl of HT1 buffer (Illumina) was added diluting the pooled library and PhiX to 10 pM, the final load concentration. The neutralised DNA was pulse vortexed and stored on ice until ready to load into the MiSeq cartridge. To generate the final load libraries for the 2 barcoded library MiSeq runs, 360 µl of the neutralised pooled library at 10 pM was combined with 240 µl denatured PhiX at 10 pM (PhiX contributing 40 percent of the run). For the stability libraries, neutralised DNA was combined with 90 percent PhiX to increase cluster diversity, since the stability libraries are largely homogenous. For each run, 3.4 µl of the custom Read 1 primer was added to the Read 1 primer mix of the MiSeq cartridge (Illumina) to enable sequencing of PhiX and to generate reads beginning with the barcoded IS*S1*. A custom Index Read primer was also loaded into the MiSeq cartridge according to the manufacturer's instructions. Fastq only files were generated according to the following settings; TruSeq LT, single-end sequencing, 1 index read, 76 cycles, adaptor trimming unchecked and custom indexing primer selected.

2.2.6 Analysis of sequencing data

Raw demultiplexed fastq files were analysed using the Bio-TraDIS scripts made available by the Wellcome Sanger Institute [111] (https://github.com/sanger-pathogens/Bio-Tradis). Descriptions of all scripts/programmes/online tools used in this thesis are available in Table A1.1 (Appendix 1). Initially, the single command pipeline script, bacteria_tradis, was utilised. The pipeline filtered and removed reads according to the transposon tag specified (e.g. CAGAAAACTTTGCAACAGAACC for library CA). After tag removal, the remaining 46 bp of *S. equi* DNA were mapped to the *Se*4047 reference

genome using SMALT short read mapper, producing a plot file of insertion sites for viewing in the Artemis genome browser [112], and for downstream analysis. The default transposon tag mismatch of 0 was maintained, however a mapping threshold of 100 percent was set (SMALT parameter y = 1) to improve accuracy and confidence in the assignment of insertion sites. Next, the plot files generated from bacteria_tradis were analysed by tradis_gene_insert_sites, generating a readable document of unique insertion sites, total read counts and insertion indices, per gene. Unique insertion sites represented by 2 or fewer reads were not included in the analysis. The output file from tradis_gene_insert_sites was used in tradis_essentiality to determine the essential genome of *S. equi*. Tradis_essentiality uses the empirically observed bimodal distribution of the insertion indices (essential and non-essential peaks) to fit gamma distributions. Insertion indices are calculated by dividing the number of unique insertion sites within a gene by the size of the gene in base pairs. Log2 likelihood ratios (LLR) are calculated between the gamma distributions, with genes assigned a LLR of less than -2 identified as essential, more than 2 as non-essential and between these values as ambiguous [111]. Essential and ambiguous changepoints were calculated from these LLRs to categorise genes into essential, ambiguous and non-essential groups. Essentialities of genes with multiple genomic copies were called as 'not defined' due to reduced confidence in read mapping. The fastq files from each library were combined, clipped of their first 2 bp to standardise the IS*S1* tag at the beginning of each read and re-analysed to generate a master library, from which final gene essentiality is reported in this study. To identify any insertion site bias, the master library mapped reads, with duplicates removed, were parsed through WebLogo, to determine the probability of each nucleotide occurring at positions 1-20 (the insertion site to 20 bp downstream) [113].

2.2.7 Comparative analysis of *S. equi* TraDIS to *S. pyogenes* and *S. agalactiae* Tn-Seq data

Gene essentiality calls of *S. pyogenes* strain M1T1 5448 and *S. agalactiae* strain A909 were retrieved from the supplementary information provided by Le Breton et al. and Hooven et al [77, 78]. In these studies, each gene of *S. pyogenes* and *S. agalactiae* was reported as essential, critical, non-essential or not defined/non-conclusive. KEGG pathway enrichment was completed on the essential and critical genes of *S. pyogenes* and *S. agalactiae* in addition to the essential and ambiguous genes of *S. equi*, using the gene set enrichment analysis available as an online tool on Genome 2D (http://genome2d.molgenrug.nl/index.php/gsea-pro-sh) [114]. The KEGG pathways attributed to the essential, critical and ambiguous genes were compared between the 3 bacteria. Gene orthologues were also identified between *Se*4047 and *S. pyogenes* strain MGAS5005 (reference strain used by Le Breton et al. for M1T1 5448), *Se*4047 and *S.*

agalactiae strain A909 and between *S. pyogenes* strain MGAS5005 and *S. agalactiae* strain A909 using the online tool OrthologeDB (http://www.pathogenomics.sfu.ca/orthologedb/) [115]. The essentiality calls of each orthologous gene pair were compared to determine concordance. All results generated from OrthologeDB were included in the analysis, except for duplicated calls where multiple copies of a gene exist in either bacterium or when gene essentiality is not defined or non-conclusive.

2.2.8 Effect of barcoded IS*S1* on library growth

Each of the 6 barcoded libraries were grown overnight in THBE alongside wild-type *Se*4047, which was grown in THB. Cultures were diluted to an initial OD_{600nm} of approximately 0.08 and incubated under the same conditions. The OD_{600nm} was measured every 30 minutes until stationary phase. The growth curves were completed in triplicate, with each replicate conducted on different days and from different stored aliquots. Doubling times were calculated from the mean exponential phase data for each library and *Se*4047. The mean doubling times of the libraries were tested for statistical significance using the Student's t-test.

2.3 Results

2.3.1 Insertion of barcoded pGh9:IS*S1* is random, stable and dense in *S. equi*

To generate 6 *S. equi* mutant libraries, 6 variant barcoded pGh9:IS*S1* plasmids were utilised. There were no significant differences in the mean doubling time of *Se*4047 relative to those of the 6 barcoded libraries (p = 0.48) (Figure. 2.4).

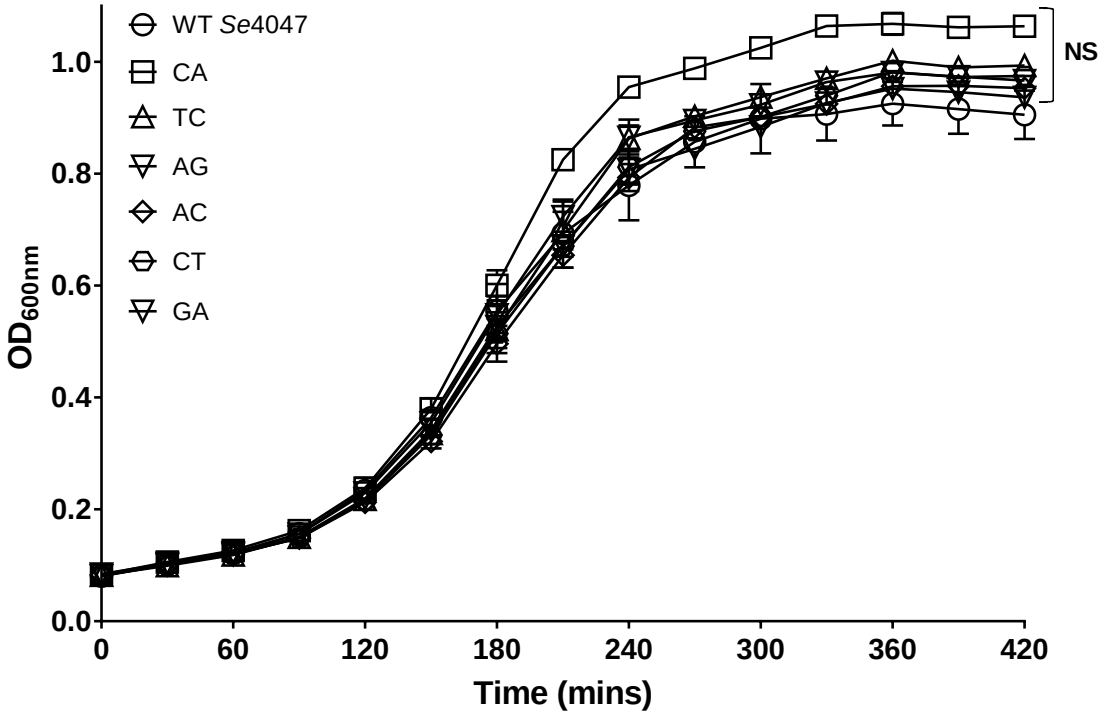


Figure 2.4. Average growth curves of 6 *S. equi* barcoded IS*S1* mutant libraries. The libraries were grown in triplicate, alongside *Se*4047, the strain from which the libraries were made. Error bars were calculated from standard deviations between the triplicate data per library at each timepoint. In some cases, the error bars lie within the point and are therefore not visible.

Transposition frequencies of between 3.5 x10^{-3} and 7.8 x10^{-3} were observed across the 6 barcoded libraries, which is comparable to the frequency of 4.9 x10^{-3} reported by Magiun et al. where pGh9:IS*S1* was transposed into *L. lactis* strain IL1403 [94]. The transposition frequency of pGh9:IS*S1* in *S. equi* was also comparable to that of the transposon, *Krmit*, in *S. pyogenes* (4 x10^{-3}) [78], but was higher than *Himar1*, a mini-transposon, in *S. agalactiae* (1 x10^{-4}–1 x10^{-6}) [77]. In common with previous studies that

identified IS*S1* transposition sites [72, 94], no specific sequence motif was observed at the transposition sites of IS*S1* in *S. equi* (Figure 2.5).

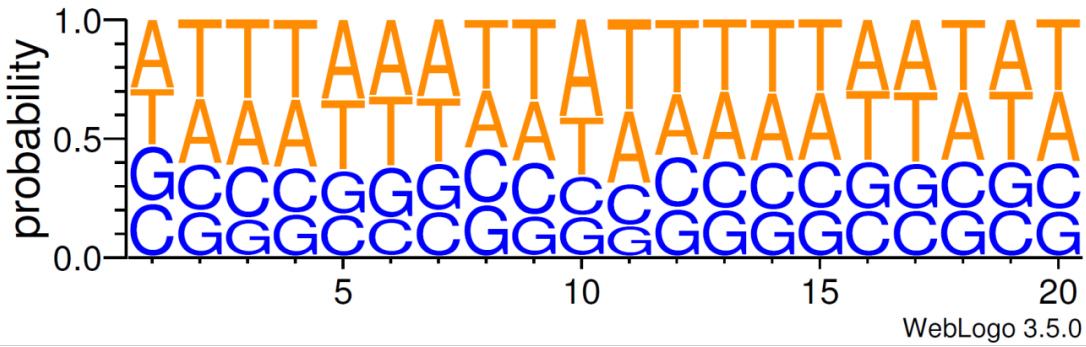


Figure 2.5. WebLogo of IS*S1* insertion sites in *S. equi*. Data from 6 barcoded IS*S1* mutant libraries in *S. equi* were combined to generate a master library. Unique sequence reads were isolated from the master library data set and parsed through WebLogo [113] to identify any insertion site bias between the insertion site and 20 bp downstream. No insertion site bias was found.

The probability of either an A or a T occuring at any position between the insertion site and 20 bp downstream, was between 54 percent to 70 percent per bp highlighting a modest preference of IS*S1* for AT rich regions, which is in agreement with the overall AT content of the *S. equi* genome (58.7 percent) [3].

To determine the stability of pGh9:IS*S1* transposition, 95 colonies from library CA were pooled (P0) and passaged twice. Sequencing of P0 identified 95 insertion sites, representing 84 genes. Ninety-five insertion sites were also identified in P1, in the same 84 genes, except that an additional mutant was identified in *SEQ1253* and a *SEQ0705* mutant was lost. For P2, 92 mutants were identified, representing 83 of the same genes. The *SEQ1253* mutant gained in P1 was lost, in addition to 2 other mutants in *SEQ1270* and *SEQ1697*. The gain then loss of a mutant in *SEQ1253* is likely due to sample preparation/sequencing differences with the remaining losses due to fitness effects following transposition of IS*S1*. These data support the stability of pGh9:IS*S1* in the *S. equi* genome and provide evidence that any onward translocation of pGh9:IS*S1* post-transposition occurs at an undetectable level.

This TraDIS technique for the generation of transposon libraries, in common with the PIMMS method utilised for the identification of IS*S1* insertion sites in *S. uberis* [72], does not attempt to eliminate the plasmid after transposition. IS*S1* duplicates on transposition generating a copy of pGh9, flanked on both sides by IS*S1*, resulting in the presence of undesirable IS*S1*-plasmid fragments in library DNA [94]. PIMMS employs an inverse

PCR of re-circularised DNA fragments to identify genomic sequences flanking IS*S1* insertion sites [72]. The TraDIS approach developed utilises Y-adapters to specifically amplify from IS*S1* generating both IS*S1*-plasmid and IS*S1*-genome fragments. Incubation of Y-adaptor ligated DNA with *Sma*I before PCR cleaved IS*S1*-plasmid fragments, such that these undesirable sequence reads accounted for only 5 to 10 percent of the final dataset. Thirteen *Sma*I restriction sites are present in the *Se*4047 genome and it is predicted that sequence reads mapping to the immediate regions surrounding these sites will similarly be lost from the final TraDIS data set.

The fastq files from each barcoded library were combined and reanalysed to generate a master library (Table 2.1). The master library represented sequencing data obtained from 2 MiSeq runs, from which 37.6 million reads were generated. Reads that contained the desired IS*S1* tag totalled 32.6 million of which 17.2 million (53 percent) mapped with 100 percent identity to *Se*4047 coding sequences. IS*S1*-plasmid reads accounted for some of the unmapped reads, however the majority are likely attributable to a combination of reads mapping in intergenic regions of DNA, reads with insufficient mapping quality using the high stringency criteria described above or through mapping to repetitive sequences within the *S. equi* genome [3].

Table 2.1. Summary of TraDIS data obtained from sequencing 6 barcoded IS*S1 S. equi* mutant libraries. Data from the 6 libraries were combined to generate the master library.

Library	Unique insertion sites in genes	Total reads in genes	Genes containing insertions (% of total genes)	Library saturation (insertion every n bp in genes)
CA	54,815	1,645,725	1,787 (87.6)	35
TC	51,827	2,162,710	1,804 (88.5)	37
AG	66,384	1,816,701	1,792 (87.9)	29
AC	35,592	3,290,822	1,797 (88.1)	54
CT	32,502	3,171,602	1,804 (88.5)	59
GA	44,761	2,650,678	1,815 (89)	43
master	**208,531**	**14,825,797**	**1,935 (94.9)**	**9**

On average, the master library contained an insertion every 9 bp in genes, representing a 79 percent increase in saturation when compared to insertions in the individual barcoded libraries. This considerable increase in library saturation did not greatly increase the number of genes represented in the master library, which was an average of 6.6 percent more than was found in the individual barcoded libraries. These data demonstrate that IS*S1* transposition occurred reproducibly across the *S. equi* genome regardless of the barcoded IS*S1* that was used.

The widespread distribution of IS*S1* transposition is evident from Figure 2.6A, which shows common regions of increased and decreased transposition (insertion index

(number of unique insertions/ size of the gene)) across the 6 libraries. A low insertion index was observed in genes encoding ribosomal proteins, with increased insertion indices evident in regions of low GC content for example in the integrative conjugative element ICE*Se*1 and ICE*Se*2 (Figure 2.6A). The pooling of data to generate the master library was supported by the increased interquartile range observed in Figure. 2.6B. Pooling the data elevated the lower quartile range increasing the robustness of the data set from which gene essentiality was determined.

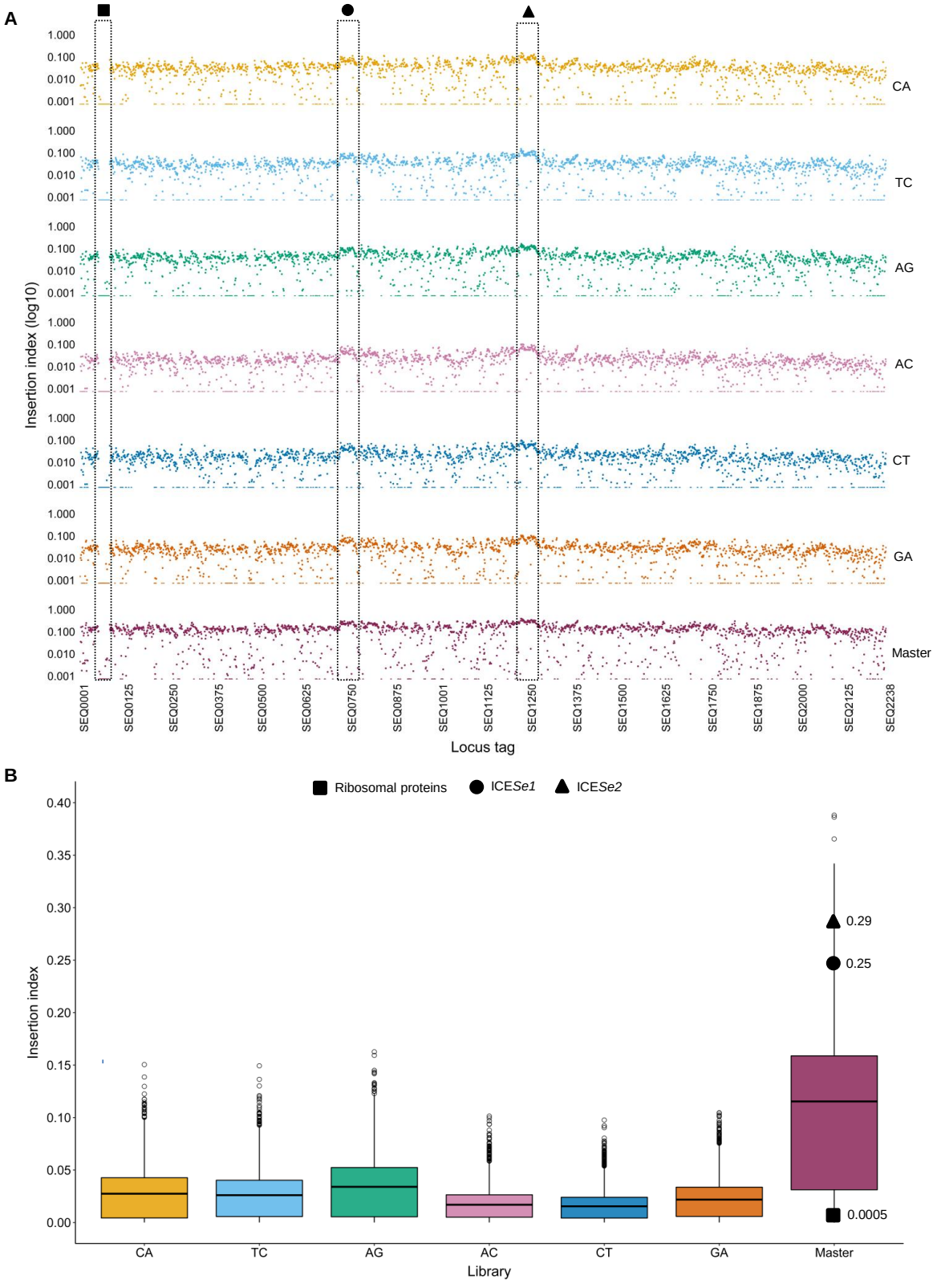


Figure 2.6. Insertion indices of *S. equi* genes disrupted by barcoded pGh9:IS*S1*. A) Insertion indices (log10) per gene is replicable between the 6 barcoded libraries. Each library is identified by its barcode on the right of the figure. The data was combined to generate a master library. Common peaks and troughs are evident; a decreased insertion index is clear in all libraries in a region of ribosomal proteins, with peaks in the integrative conjugative elements ICE*Se*1 and ICE*Se*2 visible. B) Box and whisker plot of the insertion indices of each barcoded library and the master library. The pooling of data to generate the master library was supported by the increased interquartile range and

the elevated lower quartile range, increasing the robustness of the data set from which gene essentiality was determined. Average insertion indices from master library data in a region of ribosomal proteins, ICE*Se*1 and ICE*Se*2 are shown.

2.3.2 The essential genome of *S. equi* is comparable to that of group A and B streptococci

Analysis of the master library with the tradis_essentiality TraDIS toolkit script [111] identified essential, ambiguous and non-essential genes based on the insertion index attributed to each gene. The tradis_essentiality script calculates the essential and ambiguous changepoints, from which gene essentiality is categorised. For the master data set, the essential and ambiguous changepoints were 0.0314 and 0.0408, respectively. Diagnostic plots produced by the script are available in Appendix 1, Figure A1.1. Using these thresholds, 19.5 percent of the *Se*4047 genome was found to be essential, 1.2 percent ambiguous, 73.4 percent non-essential and 5.8 percent not defined. The proportion of essential genes in *Se*4047 is similar to the 12 percent and 13.5 percent essential genes in *S. pyogenes* [78] and *S. agalactiae* [77], respectively. The essential gene set for *Se*4047 were compared to those reported for *S. pyogenes* M1T1 5448 [78] and *S. agalactiae* A909 [77]. There was 90.2 percent concordance of gene essentiality between *S. equi* and *S. pyogenes* (null= 0.17 percent, 2 genes); 89.4 percent between *S. equi* and *S. agalactiae* (null= 0.17 percent, 2 genes); 90.9 percent between *S. pyogenes* and *S. agalactiae* (null= 0.18 percent, 2 genes) and 83.7 percent between the 3 species (null= 0.31 percent, 3 genes) (Figure 2.7). These data highlight the similarities of the functional genomes of these different pathogens in support of previous studies that identified shared core and accessory genomes [3, 104].

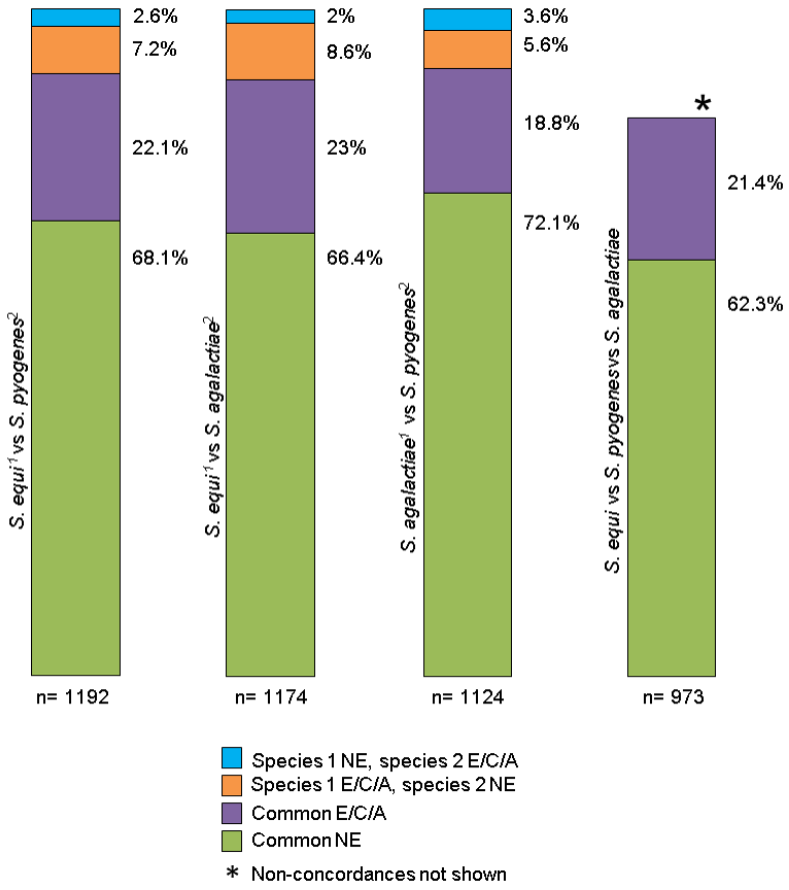


Figure 2.7. Gene essentiality concordance between a Group A, B and C streptococci. Essentiality between orthologous gene pairs in *S. equi*, *S. pyogenes* and *S. agalactiae* were compared. Orthologues were classified as either essential/critical/ambiguous concordant (E/C/A) or non-essential (NE) concordant. Non-concordances are shown for 2-species comparisons only.

In each species, libraries were generated using different transposons, prepared and analysed in different ways and yet identified common essential gene sets, illustrating the compatibility of these methodologies and the reproducibility of essentiality assignments across these streptococci.

The biosynthetic pathways attributed to each species' essential/critical/ambiguous gene set were identified by KEGG pathway analysis. This analysis revealed that the essential/critical/ambiguous genes of *S. equi*, *S. pyogenes* and *S. agalactiae* were attributed to 45, 41 and 41 KEGG categories, respectively, 39 of which were shared between the 3 species (Figure 2.8A). The 10 most prevalent essential/critical/ambiguous KEGG pathways in each species were compared (Figure 2.8B).

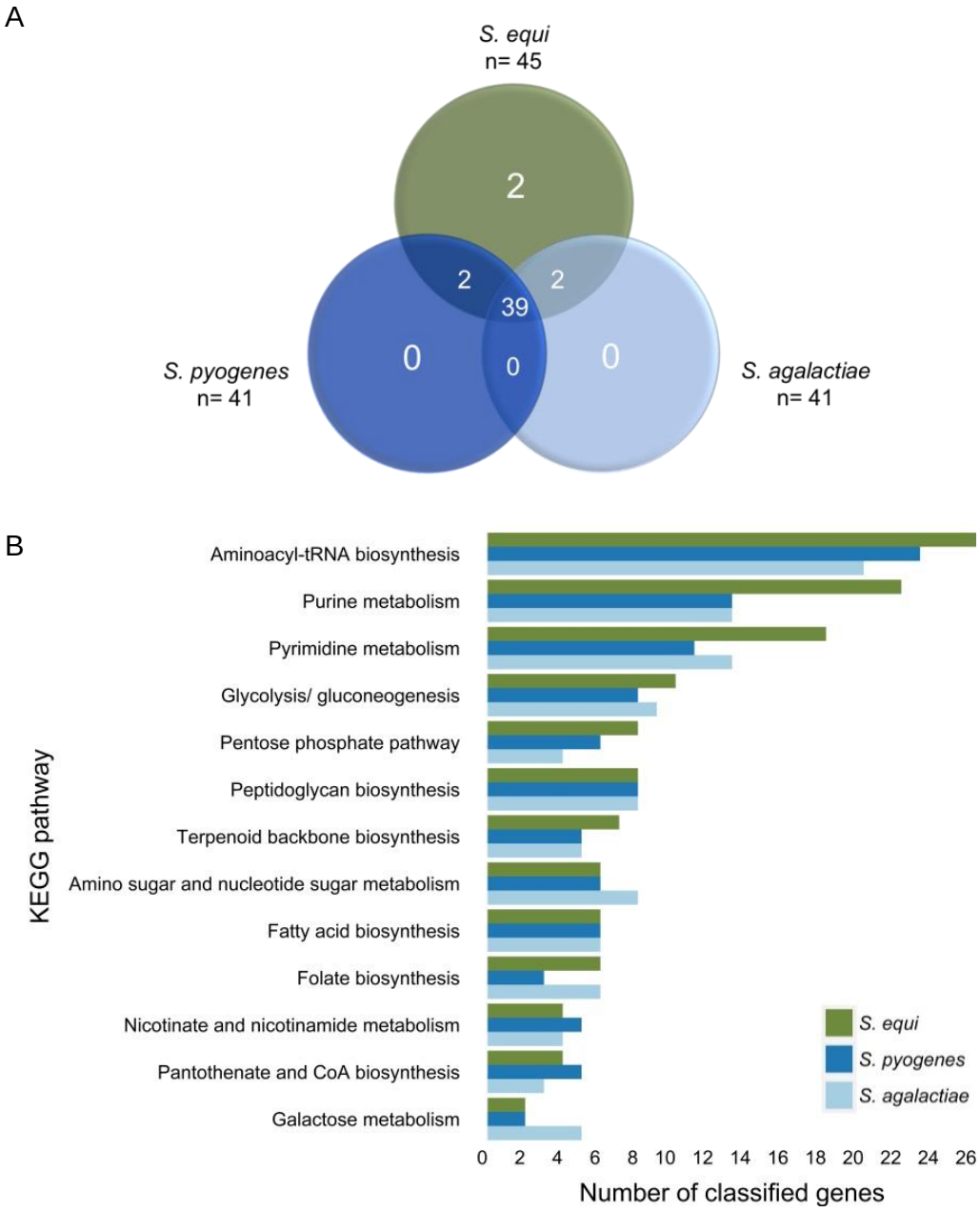


Figure 2.8. KEGG analysis of the essential/critical/ambiguous genes of Group A, B and C streptococci. A) Venn diagram showing the comparison of the KEGG categories assigned to the essential/critical/ambiguous genes of *S. equi*, *S. pyogenes* and *S. agalactiae*. The overlap of genes concludes that the essential pathways employed by the 3 different species are conserved. B) Barchart of the calls within most highly ranked KEGG pathways. The top KEGG categories in each species were consistent with one another.

The highest-ranked categories were involved in key cellular processes such as aminoacyl-tRNA biosynthesis, purine and pyrimidine metabolism, glycolysis and gluconeogenesis, the pentose phosphate pathway and peptidoglycan biosynthesis. The top KEGG categories in each species were consistent with one another. However, the *S. equi* essential genome contained noticeably more genes implicated in purine and

pyrimidine biosynthesis. This may reflect the larger essential gene set of *Se*4047 or may be attributed to the *in vitro* conditions in which the libraries were grown. *S. equi* libraries were grown in THB, whereas the *S. pyogenes* libraries were grown in THB supplemented with 0.2 percent yeast. The *S. agalactiae* libraries was grown in tryptic soy broth. The differences in media used between these studies is likely to have impacted the essential genes sets identified.

2.4 Discussion

The successful customisation of a barcoded TraDIS technique based on the original method developed by Langridge *et al*. [68] was described in this Chapter. Insertion of pGh9:IS*S1* into the *S. equi* genome is random, dense and stable, making it a highly useful tool for the progression of TraDIS studies in this important bacterium. TraDIS identified that 19.5 percent of the *S. equi* genome is essential to basic survival in rich medium, 73.4 percent of genes being non-essential, with the remainder either not defined or of an ambiguous assignment. Comparative analysis revealed that more than 83 percent of the essential gene set of *S. equi* was concordant with the essential genomes of *S. pyogenes* and *S. agalactiae*, highlighting the close genetic relationships between these important pathogenic bacteria. The pan-species essential genome and novel *S. equi* essential genes are explored in the remainder of this discussion.

2.4.1 Pan-species essential genes

Glycolysis

The 'glycolysis/gluconeogenesis' KEGG category is ranked highly in all 3 species' essential gene sets. This is unsurprising considering that the process of glycolysis is widespread in nature, being the basis for both aerobic and anaerobic respiration. Carbohydrates processed by glycolysis can be imported into bacteria via phosphotransferase (PTS) systems, which involves the transfer of phosphate from the glycolytic intermediate, phosphoenolpyruvate (PEP), to the enzymes PEP phosphotransferase (EI) encoded by *ptsI* and subsequently to the histidine-containing phosphocarrier protein (HPr), encoded by *ptsH*, causing rapid phosphorylation of the sugar in transport (Figure 2.9) [116]. It is likely that EI and HPr are employed in other sugar transport PTS systems and are not specific to glucose metabolism. Non-PTS import systems may also be utilised, such as the sugar uptake permease GlcU (Figure 2.9). Once imported, glucose is converted into glucose-6-phosphate (G-6-P) by phosphorylation where it can be directed into the glycolysis pathway.

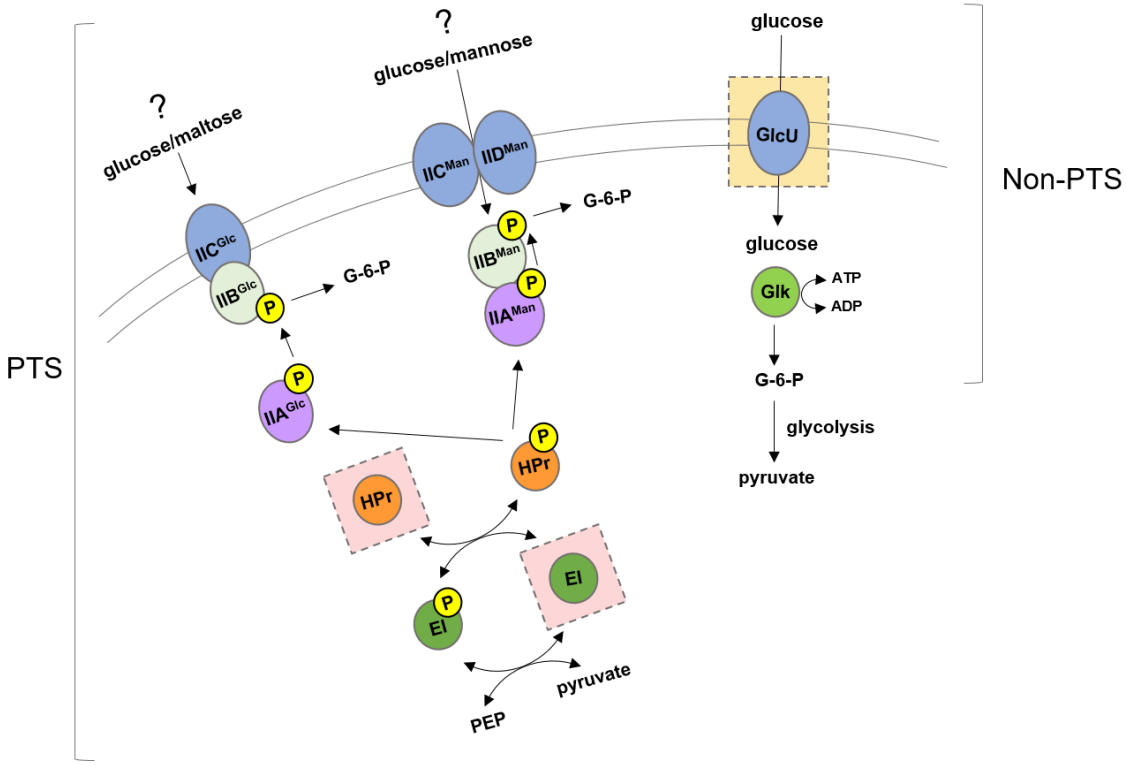


Figure 2.9. Schematic diagram of PTS and non-PTS mechanisms of carbohydrate import in bacteria. IIC^{glu}= putative membrane domain for glucose import, IIB^{glu}= putative C-terminal domain for glucose import and conversion to G-6-P by phosphorylation, G-6-P= glucose-6-phosphate, IIA^{glu}= phosphate donor to IIB^{glu}, P= phosphate, HPr= histidine-containing phosphocarrier protein, EI= PEP phosphotransferase, PEP= phosphoenolpyruvate, IIC^{Man} and IID^{Man} = putative membrane domain for glucose/mannose import, IIB^{glu}= putatively phosphorylates glucose into G-6-P, IIA^{glu}= phosphate donor to IIB^{Man}, GlcU= putative membrane domain for glucose import, Glk= glucokinase, ATP= adensine triphosphate, ADP= adenosine diphosphate. Pink dashed boxes= essential enzymes in *S. equi*, *S. pyogenes* and *S. agalactiae in vitro*. Orange dashed box= essential non-PTS importer in *S. equi*, but non-essential in *S. pyogenes* and *S. agalactiae in vitro*.

In Group A and C streptococci, it was hypothesised that glucose is transported by PTS, however it was discovered in *S. pyogenes* that IIBCGlu, a known PTS system transporter encoded by *ptsG*, in fact transports maltose and the gene has since been renamed as *malT* [117], suggesting glucose is transported by other means. The GlcU non-PTS pathway in *L. lactis* was shown to be driven by a proton-motive force to translocate glucose into the cell, but with low affinity [118]. The imported unphosphorylated glucose by GlcU is then phosphorylated by glucokinase (Glk) at the expense of ATP.

From *S. equi* TraDIS data, transposon insertion in *ptsG* (IIBCGlu) and *crr* (IIAGlu) do not incur a growth defect in *S. equi* and were hence identified as non-essential, suggesting that glucose import is conducted via another system. IIBCGlu and IIAGlu were also nonessential in the *S. pyogenes* and *S. agalactiae* Tn-seq data, implying that these

enzymes are not required for streptococcal sugar import when grown in rich medium, with other import systems likely compensating for the loss of their function. The enzymes EI and HPr were however identified as essential in *S. equi*, *S. pyogenes* and *S. agalactiae* (Figure 2.9, pink dashed boxes), which may result from their involvement in the phosphorylation of a range of imported sugars. The lack of growth on a range of PTS and non-PTS imported sugars in a Δ*ptsI* mutant of *S. pyogenes* strain MGAS5005 supports this hypothesis, highlighting its involvement in both PTS and non-PTS systems [119].

The non-PTS importer GlcU is non-essential in *S. pyogenes* and *S. agalactiae*, but essential in *S. equi* (Figure 2.9, yellow dashed box). It is possible that *S. equi* does not have any other glucose import systems, with no import occurring through either the 'glucose' or mannose PTS systems, agreeing with the reclassification of the 'glucose' system as a maltose PTS. *S. pyogenes* and *S. agalactiae* may be capable of compensating for the loss of GlcU function by the utilisation of other glucose importers. *S. equi* has undergone significant gene losses, refining its genome and driving its niche adaptation, so it is possible that it only has 1 glucose import system.

The importance of GlcU in *S. equi* is not, however, reflected in its associated phosphorylating enzyme, Glk. In *S. equi*, *S. pyogenes*, *S. agalactiae*, and *Streptococcus sanguinis* (*S. sanguinis*) [120], Glk (a.k.a NagC) is non-essential, suggesting that another enzyme is capable of phosphorylating glucose imported via GlcU. In *S. aureus*, a Δ*glk*Δ*ptsH* mutant could not grow on glucose, implying that the enzymes encoded by these genes are the only ways of phosphorylating imported glucose [121]. Therefore, HPr may be capable of compensating for a disrupted glucokinase, explaining its identification as non-essential in *S. equi*. A Δ*glcU*Δ*pstH* mutant in *S. aureus* however, was able to grow on glucose, albeit not as well as the WT strain [121], implying that *S. aureus* has other import systems capable of compensating for the loss of GlcU.

Once glucose has been phosphorylated into G-6-P, the remainder of the glycolysis pathway can be executed. Glycolysis involves the conversion of glucose into pyruvate via 10 enzymatic steps (Figure 2.10). All 10 enzymes, except glucokinase, are essential to the survival of *S. equi*, *S. pyogenes* and *S. agalactiae in vitro* from the TraDIS/Tn-seq data (Figure 2.10, red), allowing the conclusion that no other complementary enzymes can compensate for their lack of function, in these species.

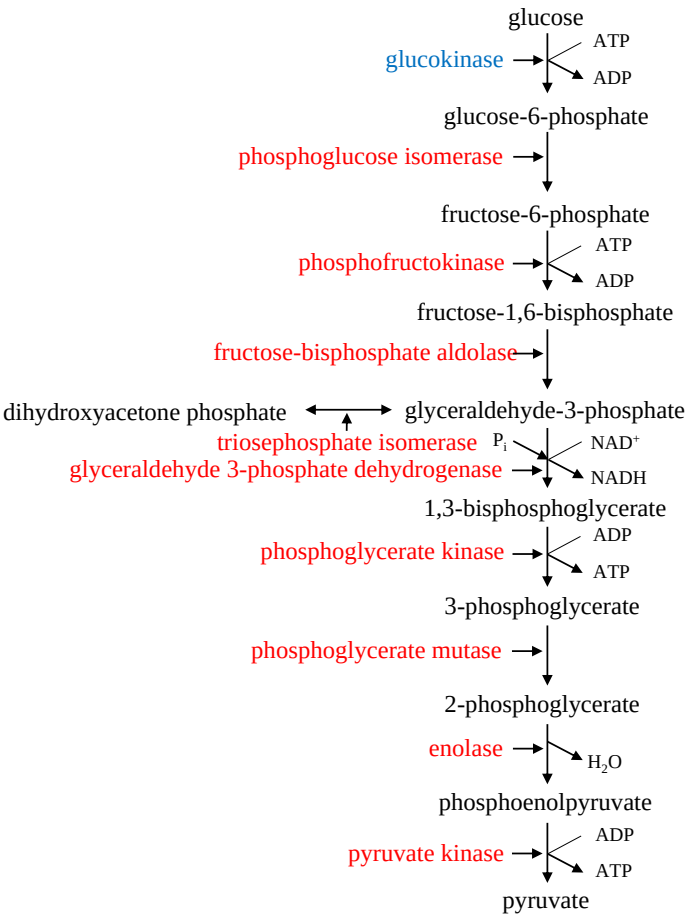


Figure 2.10. Schematic diagram of glycolysis. Enzymes in red= essential to *S. equi, S. pyogenes* and *S. agalactiae in vitro* as identified by TraDIS/Tn-seq. Blue enzyme= non-essential to *S. equi, S. pyogenes* and *S. agalactiae in vitro* as identified by TraDIS/Tn-seq.

Eight of the 9 essential glycolysis enzymes contain IS*S1* insertions which are strictly limited to the very 3' end of coding regions, which are likely to have no or little effect on the function of the transcribed product. Phosphofructokinase is the only enzyme to not have been disrupted whatsoever by IS*SI* insertion. The importance of these enzymes to the production of energy for the basic functioning of bacteria makes their identification as essential in these 3 species of streptococci unsurprising, but gives greater confidence in the accuracy of the TraDIS system.

Pentose phosphate pathway

The pentose phosphate pathway mainly serves to metabolise glucose-6-phosphate into 5-phosphoribosyl-1-pyrophosphate (PRPP) for downstream purine, pyrimidine and histidine metabolism. It is unsurprising that such an important pathway would largely be identified as essential in *S. equi*, *S. pyogenes* and *S. agalactiae*. However, there seems

to be an ‘essential route’ within the pentose phosphate pathway, in *S. equi* at least (Figure 2.11, red text, green shading).

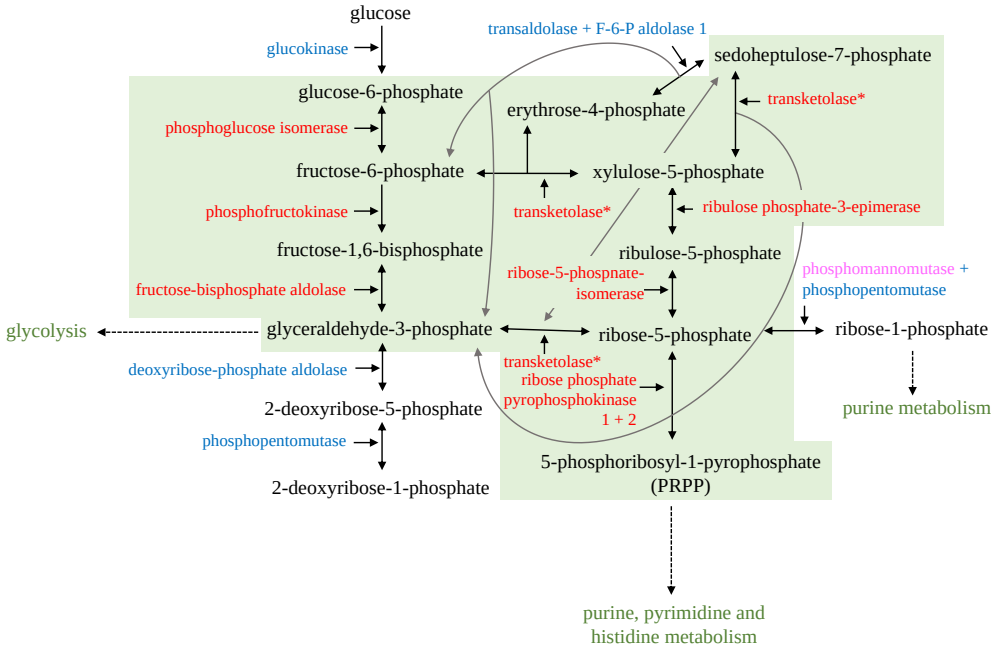


Figure 2.11. Schematic diagram of the pentose phosphate pathway. Red enzymes= essential to *S. equi in vitro*, blue enzymes= non-essential *in vitro* and pink enzymes= ambiguously defined essentiality. *= enzyme has multiple functions in pathway. Green shading= putative ‘essential pathway’ in *S. equi*.

Table 2.2. Essentiality of pentose phosphate pathway genes in *S. equi*, *S. pyogenes* and *S. agalactiae*, identified by TraDIS/Tn-seq [77, 78]. Critical and ambiguous categories both refer to important genes that contain fewer insertions than non-essential genes, but too many to be classed as essential.

		*S. equi Se*4047		*S. pyogenes* MGAS5005		*S. agalactiae* A909	
Gene	Function	Locus tag	Essentiality	Locus tag	Essentiality	Locus tag	Essentiality
deoC	deoxyribose-phosphate aldolase	SEQ1059	non-essential	M5005_Spy_1585	non-essential	SAK_2009	non-essential
deoB	phosphopentomutase	SEQ1355	non-essential	M5005_Spy_0696	non-essential	SAK_1269	non-essential
-	transketolase subunit	SEQ0125	non-essential	-	-	SAK_0262	non-essential
-	transketolase subunit	SEQ0126	non-essential	-	-	SAK_0263	non-essential
tkt	transketolase	SEQ1818	essential	M5005_Spy_1375	non-essential	SAK_1756	non-essential
-	transaldolase	SEQ1819	non-essential	M5005_Spy_1376	non-essential	SAK_1757	non-essential
fsaA	fructose-6-phosphate aldolase 1	SEQ2105	non-essential	M5005_Spy_1742	non-essential	SAK_0402	non-essential
rpe	ribulose-phosphate 3-epimerase	SEQ0334	essential	M5005_Spy_0224	critical	SAK_1798	non-essential
rpiA	ribose-5-phosphate isomerase	SEQ1356	essential	M5005_Spy_0695	critical	SAK_1270	non-essential
prsA1	ribose-phosphate pyrophosphokinase 1	SEQ0020	essential	M5005_Spy_0018	essential	SAK_0051	essential
prsA2	ribose-phosphate pyrophosphokinase 2	SEQ1141	essential	M5005_Spy_0845	non-essential	SAK_1182	non-essential
pgmA	phosphomannomutase	SEQ1067	ambiguous	M5005_Spy_0938	essential	SAK_1155	essential

The enzymes phosphoglucose isomerase, phosphofructokinase and fructose-bisphosphate aldolase are shared with the glycolysis pathway, as previously described. The other enzymes depicted in Figure 2.11 are likely unique to the pentose phosphate

pathway. Fructose-6-phosphate is ultimately converted into 5-phosphoribosyl-1-pyrophosphate (PRPP) for downstream purine, pyrimidine and histidine metabolism. In *S. equi*, the enzymes encoded by *tkt*, *rpe*, *rpiA*, *prsA1*, *prsA2* are essential to this conversion (Table 2.2). However, in *S. pyogenes*, only *rpe*, *rpiA* and *prsA1* were identified as essential. In *S. agalactiae*, the only essential enzymes in this pathway are *prsA1* and *pgmA*. The non-essentiality of the *prsA2* kinase in *S. pyogenes* and *S. agalactiae* suggests that in these species, PrsA1, or another kinase, can compensate for a lack of PrsA2. However, the essentiality of PrsA1 suggests that PrsA2 cannot compensate for its loss of function in transposon mutants.

Peptidoglycan synthesis

In ovococci, such as *S. equi*, a group of cell division proteins are required for the peripheral or septal synthesis of peptidoglycans (PGs) destined for the cell wall. The existence of both peripheral and septal systems ensures the cell is maintained in its ellipsoid shape. It is not currently known whether the 2 machineries exist as separate complexes in ovococci or whether they combine mid-cell and assist one another in cell wall PG synthesis. The peripheral PG machinery of ovococci consists of the genes *mreC*, *mreD*, *rodA* and *pbp2b*, with the septal machinery encoded by *ftsZ*, *ezrA*, *gpsB*, *pbp1a*, *pbp2x*, *ftsW*, *divIB*, *ftsL* and *divIC*, repectively [122-129] (Figure 2.12). *gpsB* and *pbp1a* are shuttled between the 2 machineries [129].

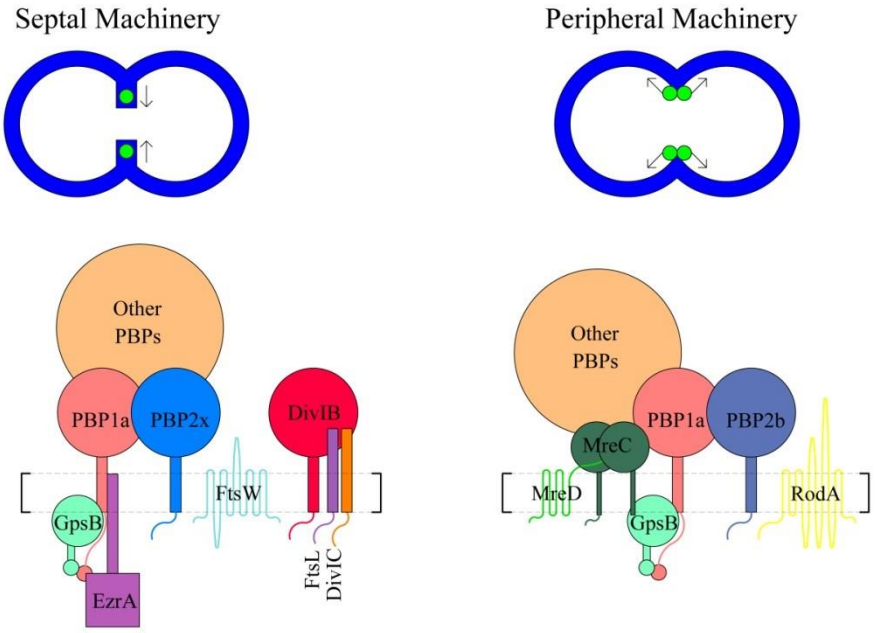


Figure 2.12. Schematic diagram of the peripheral and septal peptidoglycan synthesis machinery employed by ovococcal bacteria. Adapted and redrawn from [122].

Table 2.3. Essentiality of peripheral and septal peptidoglycan synthesis machinery genes in *S. equi*, *S. pyogenes* and *S. agalactiae,* identified by TraDIS/Tn-seq [77, 78].

			S. equi Se 4047		*S. pyogenes* MGAS5005		*S. agalactiae* A909	
	Gene	Function	Locus tag	Essentiality	Locus tag	Essentiality	Locus tag	Essentiality
peripheral	*mreC*	rod shape-determining protein	SEQ0017	non-essential	-	-	-	-
	mreD	putative membrane protein	SEQ0018	non-essential	-	-	-	-
	rodA	putative peptidoglycan biosynthesis protein	SEQ0892	non-essential	-	-	SAK_0706	non-essential
	pbp2b	penicillin-binding protein 2b	SEQ0660	non-essential	M5005_Spy_1160	non-essential	SAK_0890	non-essential
septal	*ftsZ*	cell division protein	SEQ0621	essential	M5005_Spy_1249	essential	SAK_0581	essential
	ezrA	septation ring formation regulator	SEQ0895	essential	M5005_Spy_0554	essential	SAK_0709	essential
	pbp2x	putative penicillin binding protein 2x	SEQ1803	essential	M5005_Spy_1366	essential	SAK_0359	essential
	ftsW	putative cell division protein	SEQ0777	essential	M5005_Spy_0506	essential	SAK_0886	non-essential
	divIB	putative cell division protein	SEQ0619	essential	M5005_Spy_1251	not defined	SAK_0579	non-essential
	ftsL	putative cell division protein	SEQ1804	essential	M5005_Spy_1367	essential	SAK_0358	non-essential
	divIC	putative septum formation initiator protein	SEQ0010	essential	M5005_Spy_0008	non-essential	SAK_0010	essential
	gpsB	cell division regulator	SEQ1787	essential	M5005_Spy_1352	not defined	SAK_0373	not defined
	pbp1a	putative penicillin-binding protein 1A	SEQ1790	essential	M5005_Spy_1355	essential	SAK_0370	non-essential

TraDIS of *S. equi* was able to identify that all septal genes are essential for basic survival in rich medium, as were *gpsB* and *pbp1a*, whereas all peripheral machinery genes were identified as non-essential (Table 2.3). This finding suggests that the septum machinery is able to function independently and is sufficient, without need for the peripheral machinery, for replication, but producing most likely unconventionally shaped cells. In the ovococcus, *S. pneumoniae*, deletion of *mreC* or *mreD* caused cell rounding and a reduction in chain length [123].

The essentiality of the septal machinery in *S. equi* was very similar to that of *S. pyogenes* and *S. agalactiae* (Table 2.3). *S. pyogenes* is less elongated than many other ovococci,

due to the loss of *mreCD* and *rodA* of the peripheral machinery, which is likely also true in *S. agalactiae* [124, 130]. The peripheral machinery genes remaining in *S. pyogenes* and *S. agalactiae* are non-essential (Table 2.3). The essentiality of 2 genes in the septal machinery of *S. pyogenes* were not defined, with another identified as non-essential. In *S. agalactiae*, several components of the septal machinery were identified as non-essential. Three of these non-essential genes encode some of the smaller complexes in the machinery (Figure 2.12), yet 1 component within this, *divIC*, remains essential. The importance of the septal machinery (including *gpsB* and *pbp1a*) and the dispensability of the peripheral machinery is also reflected in *S. mutans* [131].

Fatty acid biosynthesis

The *fab* genes for fatty acid biosynthesis are relatively conserved between bacterial species, although the arrangement of the operon differs. Identification of the 'fatty acid biosynthesis' KEGG category within the highest-ranking pathways supports this. In *S. equi*, the operon contains the genes *fabT*, *fabH*, *acpP*, *fabK*, *fabD*, *fabG*, *fabF*, *accB*, *fabZ*, *accC*, *accD* and *accA*. The operon in *S. pyogenes* reflects that in *S. equi* but is in the opposite orientation i.e. *accA* at the 5' end and *fabT* at the 3' end [132]. The orientation of the operon in *S. zooepidemicus* matches that in *S. equi* except that some strains e.g. MGCS10565 do not contain *accC* [3]. This implies that strains of *S. zooepidemicus* that contain differences in generally conserved gene sets share a more distant common ancestor with *S. equi* than other strains of *S. zooepidemicus*.

The fab genes encode various enzymes that participate in the conversion of acetyl-CoA and malonyl-CoA into long chain fatty acids, in particular for use as membrane phospholipids [133, 134]. All genes, except *fabT* were identified as essential or critical in *S. equi*, *S. pyogenes* and *S. agalactiae*. FabT acts as a repressor of fatty acid biosynthesis and has been identified as such in *S. pneumoniae* [135]. FabT is functionally dependent on the demand for fatty acids. Fatty acid biosynthesis comes at high energetic cost to the bacteria and so FabT ensures that energy is only expended as required [135]. The rich media that libraries in all 3 species were exposed to *in vitro* may negate the need for FabT, due to high nutrient availability.

Heat shock protein regulon

Under temperature stress, a heat shock regulon is employed in bacteria to promote continued growth. The heat shock regulon acts to prevent the aggregation of stress-denatured DNA by promoting DNA folding [136]. In many Gram positive bacteria a *groE* operon exists which contains *groEL* and groES along with a *dnaK* operon containing *hrcA*, *grpE*, *dnaK* and *dnaJ* [137, 138]. HrcA acts as a regulator of both these operons

through the binding to a highly conserved cis-acting element contained within their regulatory domains [137]. The *groE* and *dnaK* operons have also been shown to be involved in signal transduction pathways, through their influence on transcriptional regulator activity via the control of regulator and protein kinase stability [139, 140]. TraDIS identified that all genes in the *groE* and *dnaK* operons are essential to *S. equi*. The IS*S1* libraries were not subjected to a sudden temperature shift after library generation, so the involvement of these genes in other signal transduction pathways may incur their importance. *S. agalactiae* does not contain homologs of *groEL* and *groES,* however, all genes in the *dnaK* operon were critical or essential for survival by Tn-seq [77]. All genes in the *groE* and *dnaK* operon were critical or essential in *S. pyogenes* by Tn-seq [78].

Sec protein-translocation pathway

The proteins transcribed by the genes *secY*, *secE*, and *secG* form a membrane complex, SecYEG, which is involved in the Secretory (Sec) pathway and relies on *secA*, an ATPase to energise the complex [141]. This complex creates a channel for the translocation of newly synthesised proteins to the cell surface prior to final maturation [141-144] (Figure 2.13). SecY, SecG and SecA were identified as essential to *S. equi*, whereas SecE, was identified as ambiguous. The importance of SecY, G and A was shared by *S. pyogenes,* however in *S. agalactiae,* whilst SecY and G were essential, SecA was not. *SecE* was non-essential in *S. pyogenes,* but essential in *S. agalactiae.*

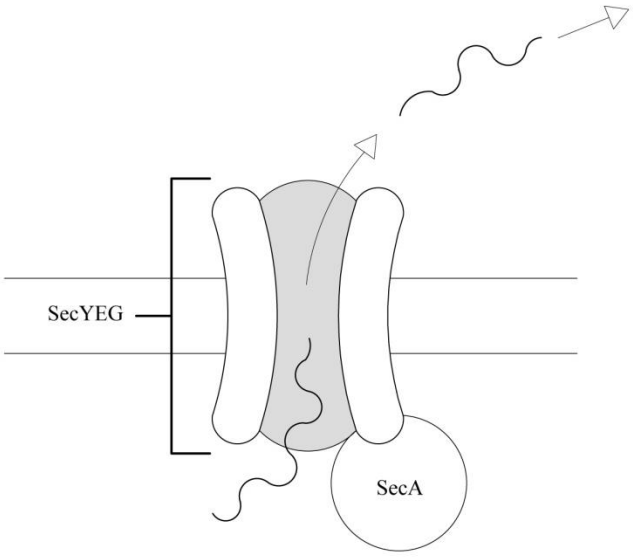


Figure 2.13. Protein translocation through SecYEG, a part of the secretory pathway. This pathway is considered central to the translocation of newly synthesised proteins to the cell surface prior to their final maturation. SecA (ATPase) provides energy for the translocation of proteins through SecYEG.

The genes encoding SecY, E, G and A are not located in an operon and are dispersed throughout the *S. equi* genome [3]. The ability of TraDIS to identify their importance regardless of their location instils confidence in other novel findings, particularly when a gene is flanked with non-essential genes, as is the case with *secG* (Figure 2.14).

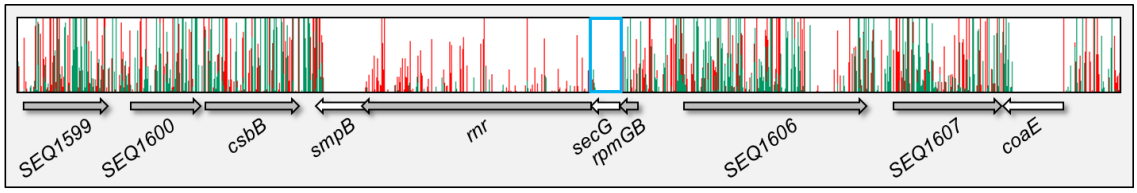


Figure 2.14. IS*S1* insertions sites in the *S. equi* genes located between *SEQ1599-SEQ1608*. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand, respectively. IS*S1* insertion is dense in the region, except in 3 distinct genes, *smpB*, *secG* and *coaE* (white arrows), identified as essential genes *in vitro*. Grey arrows indicate non-essential genes. SecG, along with SecY and SecA (located elsewhere on the genome) are essential components of the secretory pathway in *S. equi*. Blue box highlights the lack of reads in *secG*. Data is viewed in Artemis [112].

Streptolysin S associated membrane proteins

Streptolysin S (SLS), as previously described in more detail in Chapter 1, is an extracellular toxin produced by *S. equi* that destroys many types of host cells [30]. SLS degrades host cells and may contribute to immune evasion, and/or nutrient acquisition. Nine genes, *sagA-I*, contribute to SLS production and are located within a SLS-associated operon. *sagA-E* are concerned with the production of SLS, whereas *sagF-I* encode membrane proteins [3]. Three of these membrane proteins (SagGHI) have been identified in *S. equi* as ABC transporters, which allow the export of SLS into the extracellular environment [3]. All 3 of these ABC transporters were identified as essential to *S. equi* by TraDIS, with the rest of the operon identified as non-essential. It can be concluded that *S. equi* cannot tolerate an accumulation of SLS intracellularly and requires the ABC transporters to export the toxin. The essentiality of all 3 ABC transporters suggests that they are likely to work in conjunction with each other as subunits of equal importance.

SLS is not produced by *S. agalactiae*, so no comparison can be made with this pathogen, but in *S. pyogenes,* SagGHI were initially identified as non-essential *in vitro* [78, 81, 87, 145]. However, after 24 and 48 h passage in THB, *sagGHI* mutants of *S. pyogenes* were all identified as essential [87]. Incubation of a Δ*sagH* deletion mutant in THB confirmed the Tn-Seq findings [87]. The *S. equi* libraries used to determine essential genes in this

thesis were grown for approximately 3 hours after 16 h growth on TH agar. Slight differences in library growth may account for these inter-species differences.

2.4.2 Novel features of the *S. equi* essential genome

Although the majority of essential genes in *S. equi* were similarly important in *S. pyogenes* and *S. agalactiae*, analysis also identified some essential genes that were restricted to *Se*4047. *S. equi* produces a secreted molecule provisionally named equibactin, which aids the acquisition of iron *in vitro* [51] and is required for the full virulence of *S. equi* in a susceptible natural host [15]. Equibactin is synthesised by a non-ribosomal peptide synthesis system encoded in an operon (*eqbB* to *eqbN*) on the integrative conjugative element ICE*Se*2 (Figure 2.15A), which is regulated by the iron-dependent transcriptional repressor, EqbA [3, 51]. None of the genes *eqbB* to *N* were identified as essential in *S. equi*, in agreement with the free availability of iron in Todd-Hewitt medium. However, *eqbA* was essential for growth *in vitro* (Figure 2.15B). Our results concur with those of Heather *et al.* who found that deletion of *eqbA* led to a slow-growth phenotype that was caused by excessive import of iron following de-regulation of the equibactin operon [51].

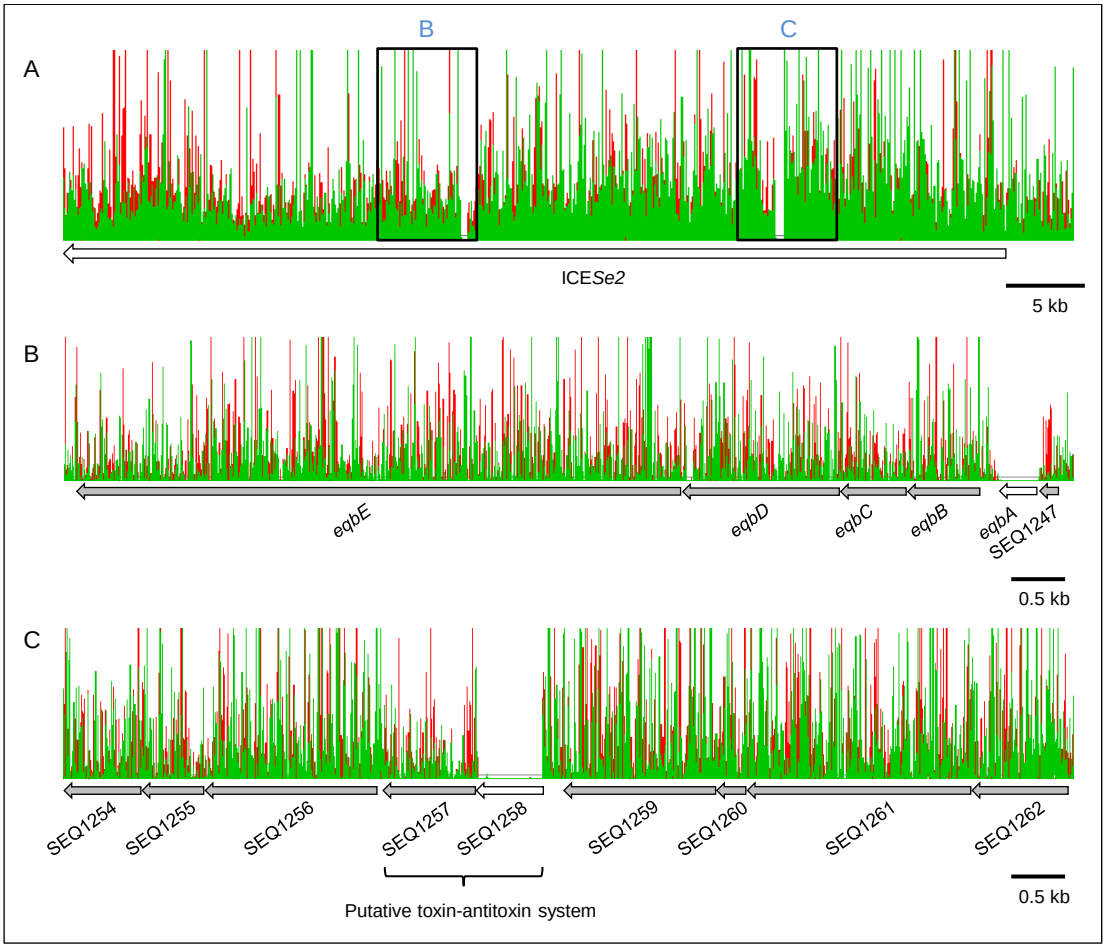


Figure 2.15. IS*S1* insertion sites in ICE*Se2* of *S. equi*. A) Overview of the integrative conjugative element, ICE*Se2*. Green and red peaks indicate reads mapping on the forward and reverse strand, respectively. IS*S1* insertion is dense in the region, except in 2 distinct genes, *eqbA* and *SEQ1258*. The labels indicate the areas zoomed into in B and C of the Figure. B) *eqbE* to *SEQ1247*. IS*S1* insertion is dense, except for in *eqbA*, the regulator of the equibactin locus. Equibactin aids the acquisition of iron, which if unregulated leads to excessive iron import and a slow growth phenotype. C) *SEQ1254* to *SEQ1262*. IS*S1* insertion is dense, except for in *SEQ1258*, a putative antitoxin, which may maintain the ICE in the bacterial genome. Both *eqbA* and *SEQ1258* were identified as essential genes. Data is viewed in Artemis [112].

ICE*Se2* also contained a second essential gene, *SEQ1258* (Figure 2.15B). *SEQ1258* and *SEQ1257* are predicted to encode a novel toxin-antitoxin system in *S. equi* [3]. Toxin-antitoxin systems comprise a stable toxin and a labile antitoxin, which promote the maintenance of the element on which they are encoded within the bacterial genome [146]. Our data suggest that *SEQ1258* encodes the antitoxin in this system (Figure 2.15C). The gene encoding the MosA antitoxin of the integrative conjugative element, SXT, of *Vibrio cholerae* was found to be essential, while *mosT*, encoding the toxin component could be deleted [146]. Recircularised extra-chromasomal copies of ICE*Se2* could not be recovered from *Se*4047 [51]. One possible explanation for this finding is that recircularisation of ICE*Se*2 halts the production of the labile antitoxin, which cannot

then neutralise the stable toxin still present in the cell. *S. equi* and *S. zooepidemicus* share over 97 percent genetic identity [3], yet ICE*Se*2 is not present in any strains of *S. zooepidemicus* studied to date [51, 147]. The maintenance of ICE*Se*2 by its toxin-antitoxin system may restrict it to *S. equi*. Interestingly, in some strains recovered from persistently infected horses, *SEQ1258*, in addition to the neighbouring equibactin locus, had been lost [15].

2.5 Conclusion

This Chapter has described the successful customisation of a barcoded TraDIS method based on the original method developed by Langridge *et al.* [68]. This barcoded technique will be of value to other researchers as it can be easily applied to other transposon systems for the study of a wide range of pathogenic bacteria. The shared essential gene set of group A, B and C streptococci provides further evidence of the close relationships of these important pathogenic bacteria. Therefore, this ABC of essential genes provides a solid foundation upon which to begin the process of reading the functional genomes of streptococci.

Defining the essential genome of *S. equi* has also set a foundation for further *in vitro* and *in vivo* experiments with the barcoded libraries, as prior identification of an essential gene set is required to enable condition-specific gene importance to be confidently assigned. In the following chapters, utilisation of the barcoded technique in immune-like conditions *in vitro* and in the natural equine host is described.

3 Genes required for survival in whole equine blood and H_2O_2

3.1 Introduction

Investigating the molecular mechanisms employed by *S. equi* under infection-like *in vitro* conditions may provide valuable insights into how this bacterium readily evades the equine immune system. To cause disease, *S. equi* has to disseminate from the nasopharynx or oral cavity through the mucosal epithelium, to eventually reach and infect the local lymph nodes. This process is not fully understood, but may be assisted by the ability of *S. equi* to survive phagocytosis within immune cells bound for the lymph nodes, which could provide a route to the target site. As previously described in Chapter 1, *S. equi* resists the equine immune system by producing known factors such as superoxide dismutase, streptolysin S, M-proteins, fibronectin binding proteins and a protective hyaluronic acid capsule [3, 12, 13, 15, 16, 19, 27, 30, 34, 38]. However, there are likely to be other factors that remain unidentified that are employed in the presence of the equine immune system.

Using TraDIS to investigate the survival of mutant libraries under immune-like conditions has the power to greatly enhance knowledge of the functional genomics of bacteria such as *S. equi* and shed light on the genes employed in these niches. In conditional TraDIS experiments, mutant libraries are exposed to the condition of choice and surviving mutants sequenced (Figure 3.1). The population recovered from the condition, or 'output' pools, are statistically compared to the population before exposure, termed 'input' pools (Figure 3.1). The changes seen in the sequenced populations are used to determine mutant fitness in the niche, as mutants that do not survive, or have decreased in number, contain transposon insertions in genes required for fitness.

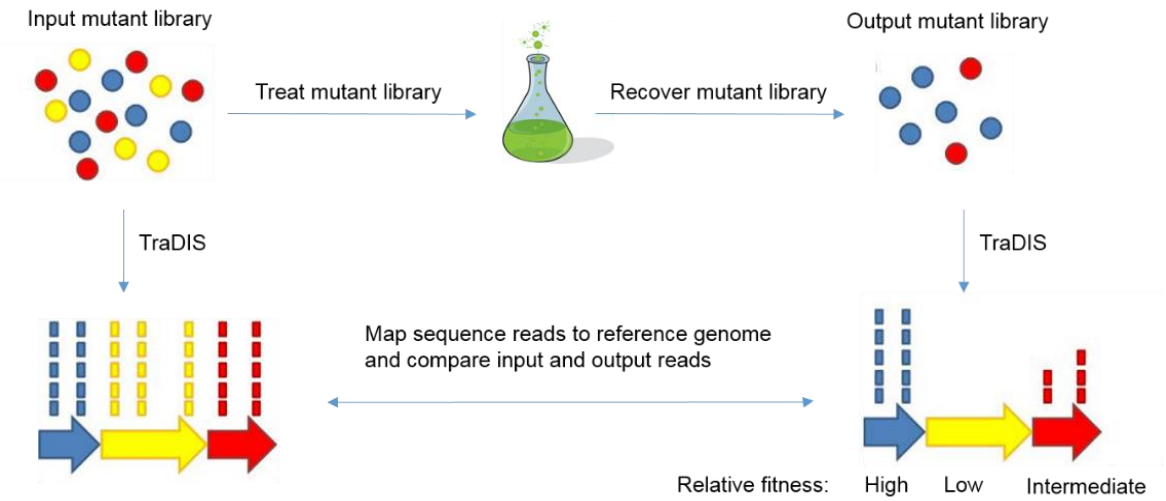


Figure 3.1. Schematic representation of conditional TraDIS experiments. TraDIS is used to measure the population of transposon mutants before and after exposure to the desired experimental condition, to identify specific gene fitness. Adapted from [83].

Conditional TraDIS/Tn-seq experiments have been conducted in many bacterial species with success. *S*. Typhimurium colonises the gall bladder where it must be able to tolerate bile. Incubation of a *S*. Typhimurium mutant library in nutrient broth containing 10 percent bile, followed by TraDIS, enabled 169 putatively genes required for survival in the gall bladder to be identified [68].

TraDIS was used to identify genes required for the survival and replication of the intracellular pathogen *Mycobacterium marinum* (*M. marinum*) [148]. This bacterium primarily inhabits phagocytic cells in a range of species and so transposon libraries of *M. marinum* were used to infect human, mouse, fish macrophage-like cells and 2 amoeba species. To validate the TraDIS data, transposon mutants in 2 genes conferring reduced fitness (*eccCb1* and *cpsA*) and 1 transposon mutant in a gene conferring enhanced fitness (*ppm1a*) were isolated and re-infected into human and fish macrophage-like cells [148]. These mutants were coinfected at a 1:1 ratio with a transposon mutant unaffected in these cells. Attenuation of the *eccCb1* and *cpsA* mutants was confirmed, but the enhanced fitness of the *ppm1a* mutant was not reproducible [148].

Incubation of *S. pyogenes* IS*S1* libraries in human saliva enabled the identification of 92 genes required for survival in this clinically relevant fluid [81]. Deletion mutants in 6 of these genes, involved in transport, pyrimidine and arginine synthesis, carbohydrate metabolism, amino acid metabolism and phosphate import, were generated and individually incubated in human saliva. All 6 deletion mutants (Δ*sptA*, Δ*sptC*, Δ*carB*, Δ*lacR.1*, Δ*nifS1* and Δ*pstS*) were significantly attenuated, confirming the TraDIS screen data. *S. pyogenes* transposon libraries have also been exposed to human blood. Eighty-

one genes were required for survival in this niche including the proven *S. pyogenes* blood fitness genes, *mga, perR* and *ralp3* [82]. Ten genes not previously implicated in *S. pyogenes* survival in blood, were selected for validation. Inactivated insertion mutants were generated by targeted plasmid integration. All 10 mutant strains grew significantly slower in human blood than the parental wild-type strain, an effect that was reversed upon plasmid curing [82].

To identify any novel genes involved in the survival of *S. equi* in the face of the equine immune system, *S. equi* mutant libraries were exposed to 2 conditions; whole equine blood and Todd-Hewitt broth containing hydrogen peroxide (H_2O_2), to simulate the equine immune response. *S. equi* does not typically cause bacteraemia, however whole equine blood was utilised to provide an *ex vivo* source of equine immune cells. To validate the findings, a panel of 6 allelic replacement mutants were exposed to whole equine blood and H_2O_2 and the impact on their viability measured.

3.2 Materials and methods

3.2.1 Bacterial strains, DNA isolation and primers

S. equi was grown at 37 ºC in a humidified atmosphere containing 5 percent CO_2, unless otherwise stated. The *E. coli* strain TG1 *repA*+ was used for the replication of the plasmid pGh9 at 37°C. *S. equi* genomic DNA was extracted using GenElute spin columns (Sigma Aldrich) according to manufacturer's instructions, except that cells were lysed for 1 hour instead of 30 minutes. A table of all primers used in this study is available in Table A1.3 (Appendix 1).

3.2.2 TraDIS in whole equine blood

One ml of stored *S. equi* barcoded libraries; AC, CT and GA, were each individually added to 39 ml of pre-warmed and pre-gassed THBE, resulting in cultures of approximately 0.05-0.08 OD_{600nm}. Cultures were grown for approximately 3 hours until OD_{600nm} 0.3 was reached (equates to approximately 2x 10^8 colony forming units (CFU)/ml). One hundred µl of each culture was added to 50 ml of freshly drawn whole equine blood (4x 10^5 CFU/ml; pony 0949 (naïve Welsh mountain pony with no history of strangles)) and incubated for 2 hours with rotation (30 rpm). Blood was collected under the auspices of a Home Office Project License and following ethical review and approval by the Animal Health Trust's Animal Welfare and Ethical Review Body (RPP 01_12). Five mls of the OD_{600nm} 0.3 cultures were also centrifuged at 10,000 rpm for 5 minutes, generating a pellet representing the input population of mutants. The supernatant was removed and pellet frozen for DNA extraction. Mutants surviving incubation with whole blood were recovered by plating 300 µl neat onto 20 large (150mm) THAE Petri dishes containing 0.03 µg/ml hyaluronidase and incubating overnight before mutants were washed off dishes using THB containing 50 percent glycerol for direct storage.

DNA was extracted from the input pellets and from 2 mls of each recovered library (3 input and 3 output) and sequenced by TraDIS, as previously described in Chapter 2 section 2.2.5. In brief, DNA was fragmented to approximately 600-800 bp, ends repaired, A-tailed and Y-adaptors ligated. DNA was digested with *Sma*I to reduce the incidence of plasmid sequencing reads and PCR amplified with a specific IS*S1* primer and a unique indexing PCR primer for each of the 6 samples (Indexing primers AHT 6, 7, 15, 16, 21 and 32 in Table A1.3, Appendix 1). PCR products were purified and size selected using AMPure XP beads as previously described. Libraries were quantified using the Kapa Biosystems library quantification kit and gel electrophoresis.

Each prepared library was diluted to 2 nM and combined in equal concentrations to form a pool of the 6 uniquely indexed samples. PhiX (Illumina) was also diluted to 2 nM. The

library pool and PhiX were denatured and neutralised as previously described in Chapter 2 section 2.2.5 and combined at 60 percent pooled library DNA and 40 percent PhiX, to generate the final load library. The MiSeq was loaded and run according to the instructions as previously described in Chapter 2 section 2.2.5.

Analysis of sequencing data

The 6 generated fastq files were analysed as previously described in Chapter 2 section 2.2.6 using the bacteria_tradis and tradis_insert_sites scripts. Five-hundred and seventy-five genes, previously identified as essential, ambiguous or not defined in Chapter 1 were removed from the analysis. Reads mapping to the final 10 percent of each gene were discounted as these were assumed to have little or no effect on the transcribed product. Three genes that were overrepresented in the input pools due to the prevalence of a few specific IS*S1* mutants were also removed. Read counts per gene were normalised between the input libraries to facilitate data comparison. Eighty-five genes that contained < 10 reads mapping to them, in any 1of the 3 normalised input libraries, were removed to ensure each gene was sufficiently represented to minimise the effects of stochastic loss. These criteria permitted the inclusion of 1,502 genes in the analysis, which represents 94.5 percent of non-essential genes in *S. equi*. All genes removed from the input data were similarly removed from the output data before the read counts per gene were normalised between the output libraries.

The script tradis_comparison [111] was used to compare the 3 output libraries to the 3 input libraries, on a sequencing reads per gene basis, generating a fitness value (log_2 fold change (FC)), *p* and *q* value for each of the 1,502 genes. Genes were considered as required for fitness upon exposure to whole blood if they exhibited a log_2 FC value of < -2 and a *q* value of < 0.05.

3.2.3 Validation of TraDIS whole equine blood results

To confirm the reduced fitness conferred by some genes reported by TraDIS, allelic replacement mutants in *Se*4047 were generated lacking the genes *pyrP* (SEQ1316), *mnmE* (SEQ1365), *addA* (SEQ0953) and *recG* (SEQ0454). Strains of *Se*4047 lacking *hasA* (SEQ0269) and *eqbE* (SEQ1242), were also utilised, both of which already existed in the Animal Health Trust strain collection [13, 51]. The Δ*hasA* strain was used as a positive control as it has been shown to exhibit strong attenuation in equine blood [13]. The Δ*eqbE* strain was used as a negative control as TraDIS data showed that fitness in whole blood was not altered upon IS*S1* insertion.

Construct generation

Deletion mutants were generated using an allelic replacement mutagenesis technique, as previously described for the generation of a Δ*prtM* mutant [98]. For each deletion, approximately 500 bp regions of *S. equi* DNA flanking either side of the target gene were amplified with Phusion polymerase (New England Biolabs (NEB)) according to the manufacturer's instructions, using the primers P1, P2, P3 and P4 (Figure 3.2). Primers relevant to each deletion are suffixed with the gene name in Appendix 1, Table A1.3. These primers were designed to incorporate restriction digestion sites to the ends of the resultant PCR products and an additional 5 bp to aid digestion of the PCR product ends (Figure 3.2).

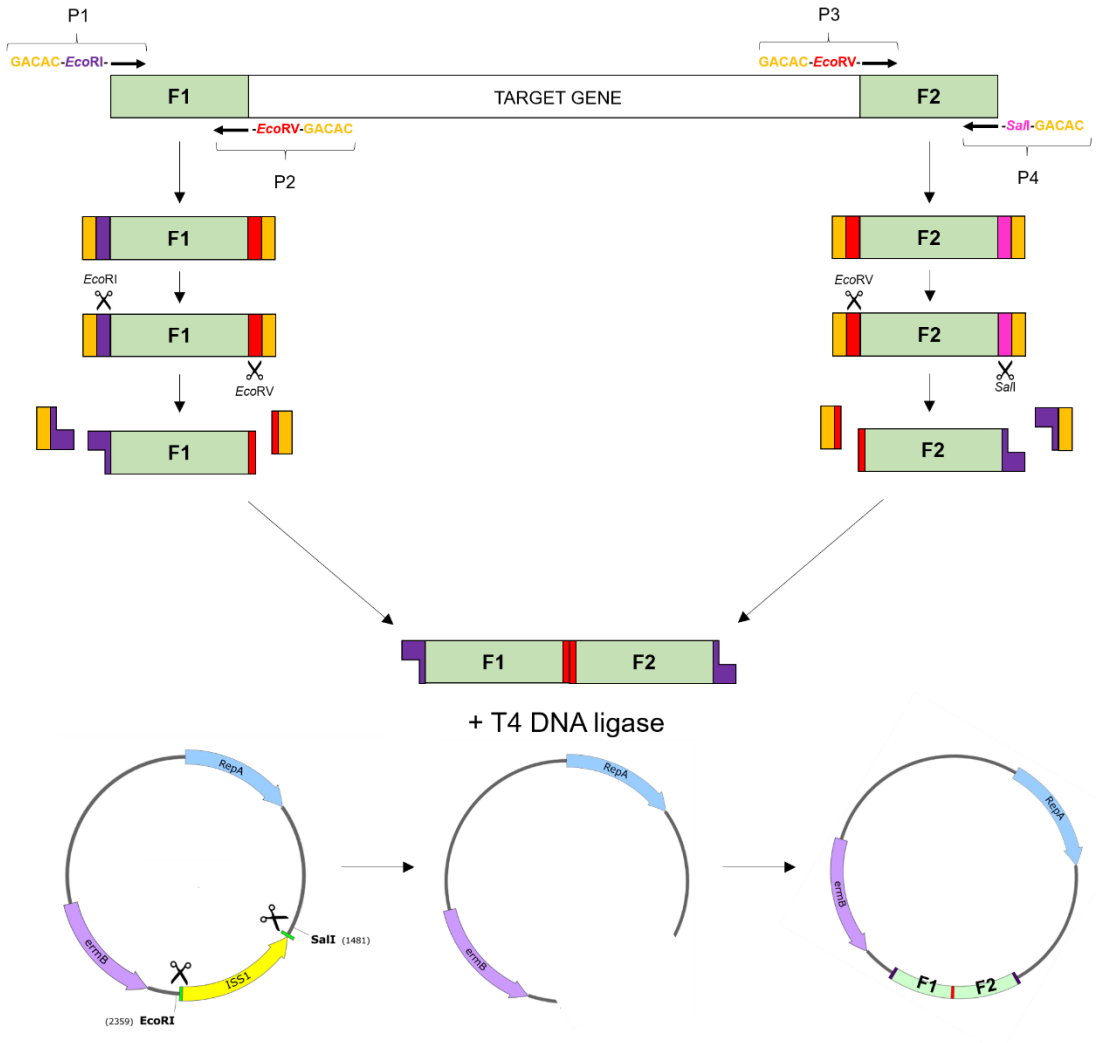


Figure 3.2. Schematic representation of deletion mutant construct generation. Approximately 500 bp regions of DNA flanking either side of the target gene are amplified using primers that incorporate restriction enzyme sites, relevant to the desired plasmid, to the ends of the products. An additional 5 bp of DNA are also present in the PCR primers, to aid efficient digestion of the products. In the case of this figure, *Eco*RI and *Eco*RV sites were incorporated into flank 1 (F1) by PCR and *Eco*RV and *Sal*I sites into flank 2 (F2). Both F1 and F2 were digested with the relevant restriction enzymes. *Eco*RI and *Sal*I were used to cut the plasmid, pGh9:IS*S1*, before the digested PCR products and plasmid were ligated together using T4 DNA ligase, forming a complete construct.

Each pair of relevant PCR products were digested according to the manufacturer's instructions using the appropriate enzymes; either *Eco*RI, *Eco*RV or *Sal*I. pGh9:IS*S1* was digested using *Eco*RI and *Sal*I and phosphorylated according to manufacturer's instructions, releasing IS*S1* from the plasmid as previously described in Chapter 2 section 2.2.2. pGh9 and both digested PCR products were ligated together in 1 reaction using T4 DNA ligase according to the manufacturer's instructions, at varying concentrations (Figure 3.2). Ligation reactions were completed at a ratio of 1:1 and 1:8 plasmid: digested PCR product, resulting in the 2 PCR products ligating together via the central *Eco*RV site and into pGh9 via the *Eco*RI and *Sal*I sites. Ligation reactions were incubated at room temperate overnight before they were transformed into *E. coli* TGI *repA*+ as previously described in Chapter 2 section 2.2.2.

Erythromycin resistant colonies were PCR screened as previously described in Chapter 2 section 2.2.2 using the primers 5'9 and 3'9 which amplify from pGh9 150 and 86 bp from the *Eco*RI and *Sal*I respectively, generating PCR products that span the cloned PCR products. The remaining 6 µl of colony suspensions, with PCR products of the correct size (~1200 bp), were added to 10 ml LB containing 150 µg/ml erythromycin (LBE) and grown overnight at 37 ºC to prepare plasmid DNA. Plasmid DNA was extracted and sequenced as previously described in Chapter 2 section 2.2.2.

Transformation into *S. equi*

Deletion constructs were transformed individually into competent *Se*4047 cells as previously described in Chapter 2 section 2.2.3. Briefly, competent *Se*4047 were electrotransformed, recovered in THBE for 1 hour at 28 ºC (plasmid permissive temperature), before plating onto THAE and incubating overnight at 28 ºC. Single colonies were inoculated into THBE and grown overnight at 28 ºC, diluted ten-fold from 1:10 to 1:10,000 and grown again overnight. Cultures closest to OD_{600nm} 0.3 were transferred to 37 ºC for 3 hours to induce chromosomal integration of the construct, which is the first cross-over event (Figure 3.3). Integrants were selected on THAE overnight at 37 ºC.

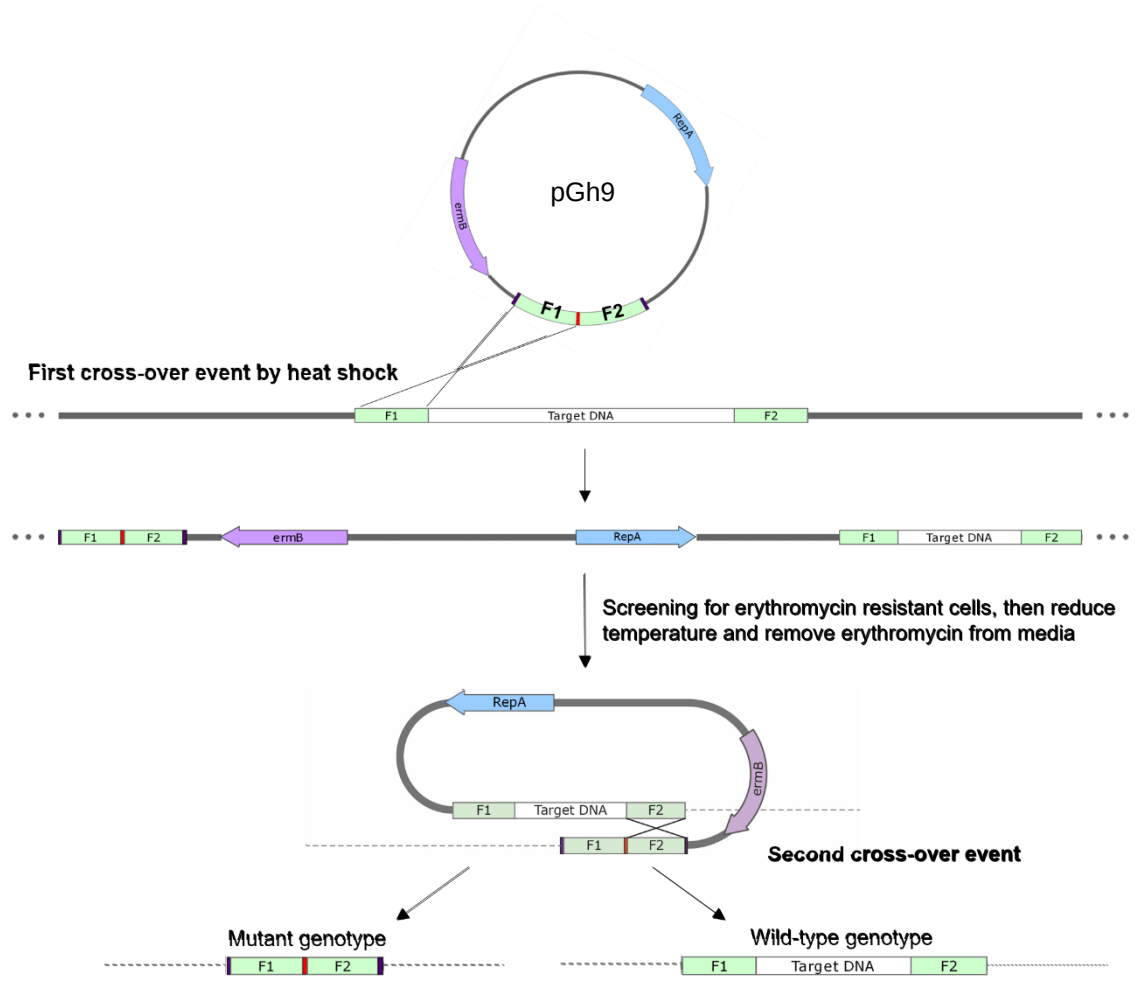


Figure 3.3. Allelic replacement in host chromosomal DNA. An allelic replacement construct (pGh9) containing 2 500 bp regions flanking the target gene for deletion (F1+F2) and an erythromycin resistance gene (*ermB*) (purple arrow) is transformed into the host cell. Transformants are heat shocked, causing integration of the construct into the genome at the target site (first cross-over event). Successful integrants are selected for by growth on agar containing erythromycin. Integrants are then excised of the pGh9 by reducing the temperature to that permissive of the plasmid and removing erythromycin from the medium. The flanks are retained, representing the second cross-over event and loss of the target gene.

Erythromycin resistant colonies were grown overnight at 37 ºC in 5 ml THBE. 2.5 ml of culture was centrifuged for DNA extraction with the remaining stored in 25 percent glycerol and frozen at -20 ºC. DNA was initially screened for successful integration using *taq* polymerase (Sigma Aldrich) and the primers 5'9 and 3'9. Successful integration is represented by a lack of PCR product as the linear integration of pGh9 renders these primers incompatible as they amplify away from each other (Figure 3.4). Any DNA not amplifying with these primers was additionally screened twice more to determine the orientation of the integration, with the relevant P1 primer and the 3'9 primer, and the relevant P4 primer with the 5'9 primer (Figure 3.4).

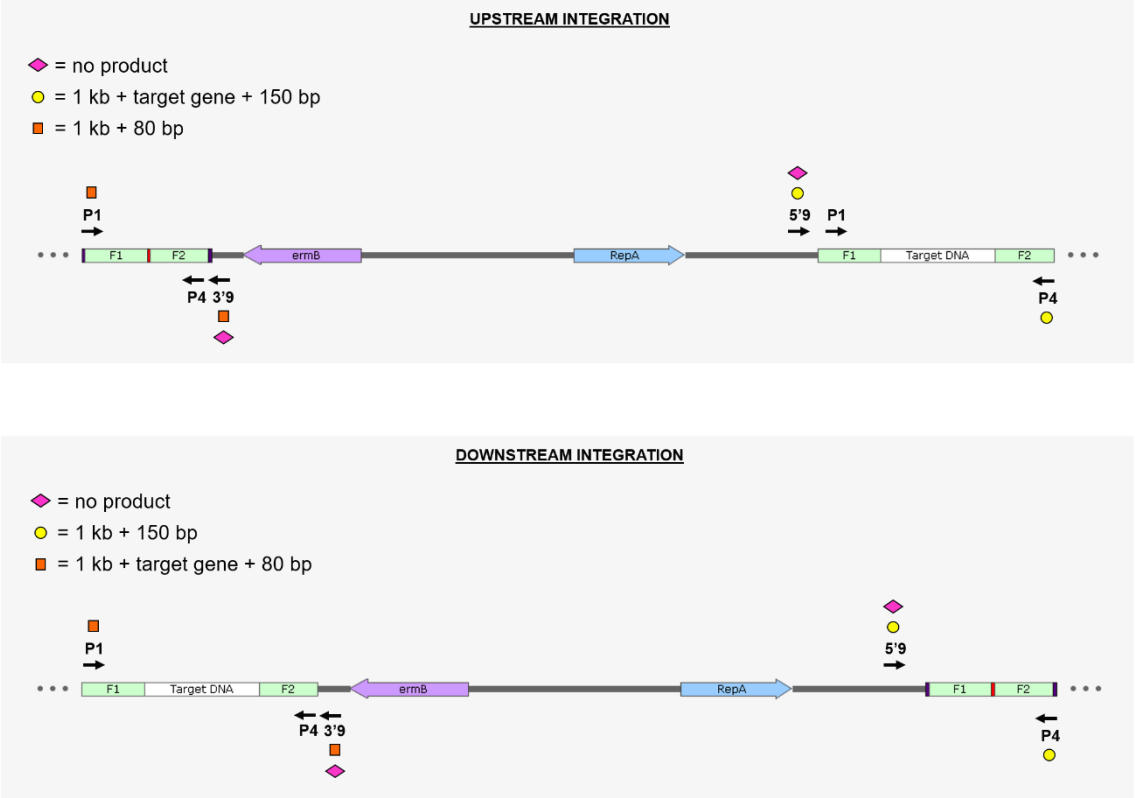


Figure 3.4. Identifying the orientation of deletion mutant construct integration into host chromosomal DNA by PCR. Different combinations of primers can be used to determine whether a deletion mutant construct has integrated upstream or downstream of the target gene. Initially, colonies are PCR screened for integration with the primer pairs indicated with pink diamonds. These primers produce no product if integration has been successful, as the primer binding sites are facing in opposite directions in the linearised, chromosomally integrated plasmid. A circular, non-integrated plasmid will produce a PCR product using these primers. Further PCR reactions with the primer pairs indicated with the yellow circle and orange square enables identification of the orientation of the construct in the integrant as different size products are generated.

The stored glycerols of 2 upstream and 2 downstream integrants were defrosted, 100 µl added to 5 ml THBE and grown overnight at 37 ºC. Cultures were diluted in THB in 10-fold dilutions from 1:10 to 1:10,000 and incubated at 28 ºC for 48 hours to excise pGh9 from the chromosome, but retain the flanks, representing the second cross-over event and loss of the target gene (Figure 3.3). The 1:1,000 and 1:10,000 dilutions were again diluted 10-fold to 1:10,000 in PBS and 100 µl of the 1:1,000 and 1:10,000 dilutions plated in triplicate on THA and grown overnight at 37 ºC to ensure free plasmid was lost. To confirm plasmid loss, and therefore loss of erythromycin resistance, colonies were re-streaked on both THAE and THA. Bacteria growing on THA but not THAE were grown overnight in 5 ml THB, 2.5 ml centrifuged at 13,000 xg to form a pellet for DNA extraction and 2.5 ml stored in 25 percent glycerol at -20 ºC. These bacteria were screened by PCR using *taq* polymerase (Sigma Aldrich) and the relevant P1 and P4 primers, to confirm the mutant allele (product of approximately 1,000 bp). PCR products with the mutant

allele were sequenced as previously described in Chapter 2 section 2.2.2 using the relevant P1 and P4 primers.

Growth curves of validation strains

Stored aliquots of the deletion mutants Δ*pyrP*, Δ*trmE*, Δ*addA*, Δ*recG*, Δ*hasA* and Δ*eqbE* and the parental WT strain, *Se*4047, were defrosted, streaked onto THA and grown for 16 hours. A single colony of each was inoculated into 10 ml THB in triplicate and grown for 16 hours. Cultures were diluted to approximately OD_{600nm} 0.08 in prewarmed and pregassed THB and grown until stationary phase was reached. OD_{600nm} was initially measured after 1 hour, with subsequent measurements taken every 30 minutes.

Whole equine blood validation assay

The fitness of each deletion strain in whole equine blood was measured in the following assay. Each deletion strain was tested on 1 day in triplicate alongside *Se*4047 in singlicate, so fitness of *Se*4047 was measured 6 times over the experiment. Blood was drawn from pony 0949 over the course of this validation assay, the same animal that was used in the preceding TraDIS study.

Stored aliquots of the deletion mutants Δ*pyrP*, Δ*trmE*, Δ*addA*, Δ*recG*, Δ*hasA* and Δ*eqbE* and the parental WT strain *Se*4047 were defrosted, streaked onto THA and grown for 16 hours. Three overnight cultures for each deletion strain and 1 overnight culture of *Se*4047 were generated by inoculating 10 ml THB with a single colony. Cultures were grown for 16 hours, diluted to OD_{600nm} 0.08 in pre-warmed and pre-gassed THB. Cultures were grown until OD_{600nm} 0.3 was reached when they were diluted 1 in 200 in THB, mixed by vortexing and diluted again 1 in 10 in THB to reach $1x\ 10^5$ CFU/ml. One hundred µl of diluted culture was added to 10 ml freshly drawn equine blood, equating to $1x\ 10^3$ CFU/ml, and incubated for 2 hours with rotation (30 rpm).

Immediately after adding the strains into the whole equine blood, 50 µl was removed and plated neat in triplicate onto Columbia CNA staph/strep selective agar (Oxoid) (time point 0 (T0)) to enumerate the initial concentration of *S. equi* cells. Surviving cells were enumerated at 1 hour (T1), 2 hours (T2) and 3 hours (T3). At T1, 50 µl was removed and plated neat in triplicate onto CNA agar. An additional 50 µl was diluted in 450 µl PBS, vortexed and 50 µl spread in triplicate on CNA agar, representing a 1:10 dilution. At T2 and T3, 1:10 and 1:100 dilutions were utilised to enumerate surviving cells. Petri dishes were incubated for 16 hours before colonies were counted to calculate bacteria present at each time point.

3.2.4 Minimum inhibitory concentration of hydrogen peroxide (H_2O_2)

To determine the concentration of H_2O_2 required to exert a selective pressure on *S. equi*, the minimum inhibitory concentration of H_2O_2 in THB was determined. A frozen glycerol of *Se*4047 was defrosted, streaked onto THA and the plate incubated for 16 hours. A single colony was inoculated into 10 ml THB and grown for 16 hours, alongside 30 ml of THB to prewarm and pregas the media. Five hundred µl of the overnight culture was diluted into 19.5 ml prewarmed and pregassed THB and incubated until OD_{600nm} 0.3 was reached. During this incubation period, serial doubling dilutions of 3 percent H_2O_2 (Sigma Aldrich) were made in a conical bottom 0.2 ml 96 well plate, in triplicate. One hundred µl of 3 percent H_2O_2 was added into wells A1-3, and 50 µl of THB added to another 45 wells, such that half the plate was filled. Fifty µl was removed from well A1-3 and added to wells B1-3, which contained 50 µl of THB, halving the concentration of H_2O_2 to 1.5 percent. Fifty µl was taken from wells B1-3 and added into wells C1-3 as previously described, diluting the concentration of H_2O_2 by half again. This method of dilution was continued until a concentration of 0.000092 percent H_2O_2 was reached. Once the *Se*4047 culture had reached OD_{600nm} 0.3, 40 µl was diluted in 10 ml prewarmed and pregassed THB, to reach 8x 10^5 CFU/ml. Fifty µl of the diluted culture was added into each of the 48 wells diluting the concentration of both the H_2O_2 in each well and the *Se*4047 by half, such that H_2O_2 concentrations of 1.5 percent to 0.000046 percent were tested, all containing 4x 10^5 CFU/ml *Se*4047. A control experiment was also conducted alongside, which was set up the same as described above, except 3 percent H_2O_2 was replaced with ddH_2O. Both 96 well plates were covered and incubated for 16 hours.

3.2.5 TraDIS in H_2O_2

TraDIS in H_2O_2 was conducted exactly as described for whole equine blood, except that the 50 ml of equine blood was substituted for 50 ml THB containing 0.0004 percent H_2O_2 (quarter of the MIC). Input pools of mutants and recovered mutants were sequenced as previously described and data again analysed using the bacteria_tradis and tradis_insertion_sites scripts. Genes previously identified as essential, ambiguous or not defined in Chapter 2 were removed from the analysis, as were the 3 overrepresented genes in the input pools due to the prevalence of a few specific IS*S1* mutants. Reads mapping to the final 10 percent of each gene were discounted and read counts per gene were normalised between the input libraries to facilitate data comparison. One-hundred and sixteen genes that contained < 10 reads mapping to them, in any 1 of the 3 normalised input libraries, were removed to ensure each gene was sufficiently represented, minimising the effects of stochastic loss. These criteria permitted the inclusion of 1,471 genes in the analysis, which represents 92.7 percent of non-essential genes in *S. equi*. All genes removed from the input data were similarly removed from the

output data before the read counts per gene were normalised between the output libraries.

The script tradis_comparison [111] was used as previously described to generate a fitness value (log_2 fold change (FC)), *p* and *q* value for each of the 1,471 genes. Genes were considered to be important for fitness upon exposure to H_2O_2 if they exhibited a log_2 FC value of < -2 and a *q* value of < 0.05.

Validation of TraDIS H_2O_2 results

The same panel of deletion mutants used to validate TraDIS in whole equine blood were also validated in H_2O_2. The assay was conducted exactly as the whole equine blood validation assay except that 10 ml of equine blood was replaced with THB containing 0.0004 percent H_2O_2.

3.2.6 Statistical analysis

Deletion mutant growth curves

The average OD_{600nm} across the 3 replicates of each strain and their standard errors were calculated. The doubling times of each replicate of each strain was calculated from exponential phase data, which was used to determine any significant differences in growth rates. The 3 doubling times calculated for each deletion strain were compared to WT *Se*4047 using a two-tailed student's *t*-test.

Whole equine blood validation assay

Colony counts for each set of triplicate Petri dishes were converted into an average CFU/ml for each timepoint for each replicate, considering the volume spread on each plate and the dilution if used. Average CFU/ml data from T1, T2 and T3 were transformed into a percent of T0 within each replicate to normalise the data, as the T0 CFU/ml varied slightly between experiments. The 3 values of transformed data per deletion strain at each timepoint were compared to the equivalent data for *Se*4047 using a two-tailed students *t*-test. A graph was plotted using the overall average percent of T0 data at each timepoint for each strain.

H_2O_2 validation assay

Data generated from the H_2O_2 validation assay was treated the same as that generated in the whole equine blood validation assay.

3.3 Results

The 3 barcoded IS*S1* libraries, designated AC, CT and GA, previously described in Chapter 2, were grown to an OD_{600nm} of 0.3 immediately before use in the whole equine blood and H_2O_2 experiments and resequenced to accurately identify input pool composition.

3.3.1 Genes that contribute to the fitness of *S. equi* in whole equine blood

Each input library represented between 80.6 and 81.7 percent of the 2,165 *S. equi* genes (Table 3.1) before data was filtered, to minimise the effects of any measurement error. Certain thresholds were imposed on the data as previously described in section 3.2.3, permitting the inclusion of 26,381 unique mutants in library AC, 24,353 unique mutants in library CT, and 28,128 in library GA, representing 69.4 percent of *S. equi* genes and 94.5 percent of the non-essential genes previously identified in Chapter 2.

Table 3.1. Composition of whole equine blood input libraries pre- and post-filtering.The number of genes containing insertions post-filtering is consistent between libraries, since filtering determines a consensus set of genes to be taken forward for analysis.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC-IN^{pre}	30,181	875,825	1,768 (81.7 : 100)
CT-IN^{pre}	27,764	886,661	1,744 (80.6 : 100)
GA-IN^{pre}	32,011	761,827	1,760 (81.3 : 100)
AC-IN^{post}	26,381	769,660	1,503 (69.4 : 94.5)
CT-IN^{post}	24,353	770,366	1,503 (69.4 : 94.5)
GA-IN^{post}	28,128	770,193	1,503 (69.4 : 94.5)

The composition of these barcoded libraries is described in detail in Chapter 2, where a higher unique mutant count was reported per library. Here, the input libraries were allocated a smaller proportion of a MiSeq run, approximately half of that described in Chapter 2, and so the libraries were sequenced to a lesser depth limiting the identification of unique mutants (Library AC; 15 percent fewer unique mutants were identified, Library CT; 14.6 percent fewer unique mutants were identified and Library GA; 28.5 percent fewer unique mutants were identified. The number of genes represented by each library is however comparable between the data sets as only 7.3 percent fewer genes were represented in the whole equine blood study. The 3 barcoded libraries recovered from whole equine blood contained, on average, 9.4 percent $\pm$ 3.5 (standard error of the mean (SEM)) fewer unique mutants than were present in the input libraries (Table 3.2).

Table 3.2. Composition of whole equine blood output libraries pre- and post-filtering. The number of genes containing insertions post-filtering is consistent between libraries, since filtering determines a consensus set of genes to be taken forward for analysis.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC-OUT^{pre}	24,985	689,276	1,676 (77.4 : 100)
CT-OUT^{pre}	25,593	744,624	1,666 (76.9 : 100)
GA-OUT^{pre}	30,404	721,703	1,687 (77.9 : 100)
AC-OUT^{post}	21,996	607,834	1,503 (69.4 : 94.5)
CT-OUT^{post}	22,555	607,834	1,503 (69.4 : 94.5)
GA-OUT^{post}	26,818	607,834	1,503 (69.4 : 94.5)

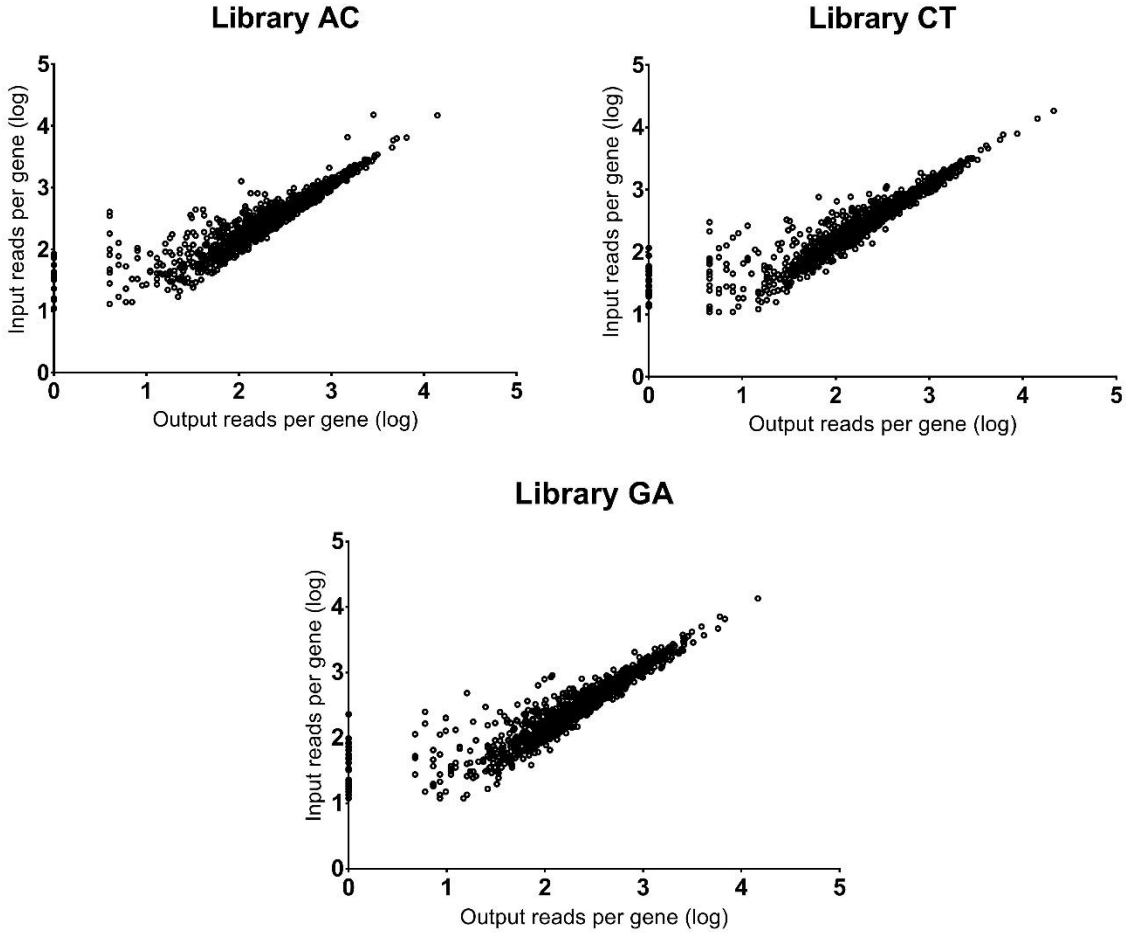


Figure 3.5. Read counts per gene in each of 3 *S. equi* barcoded IS*S1* libraries, pre- (input) and post- (output) exposure to whole equine blood.Genes represented by < 100 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Barcoded mutant libraries were each exposed to whole equine blood to identify genes contributing to fitness in this environment. TraDIS was used to identify any population changes in the recovered (output) mutants versus the input pools, alluding to gene fitness. In the majority of genes included in the analysis, the number of reads sequenced

per gene remained comparable between both the input and output libraries, highlighting genes that were not required for survival in whole equine blood (Figure 3.5). Some genes however were represented by fewer reads in the output pools compared to the input pools, alluding to reduced fitness. To quantify any fitness changes in the libraries, the log_2FC for all 1,502 genes permitted in the analysis was calculated (Figure 3.6). IS*S1* insertion into 36 genes significantly reduced the fitness of *S. equi* in the presence of whole equine blood (log_2FC < -2 and q < 0.05, Figure 3.6 blue and red dots, Table 3.3). The remaining 1,466 genes exhibited no growth defects in whole equine blood as a result of IS*S1* insertion (Figure 3.6, grey dots and green dot).

Cluster of orthologous groups (COG) analysis of the 36 fitness genes (Figure 3.6BC) identified that the most prevalent categories included genes involved in replication, recombination and repair (*n*=4, 4.6 percent of total in COG category), transcription (*n*=4, 3.9 percent of total in COG category) and energy production and conversion (*n*=4, 8.7 percent of total in COG category). Five genes (14 percent of fitness genes) did not belong to a defined COG category.

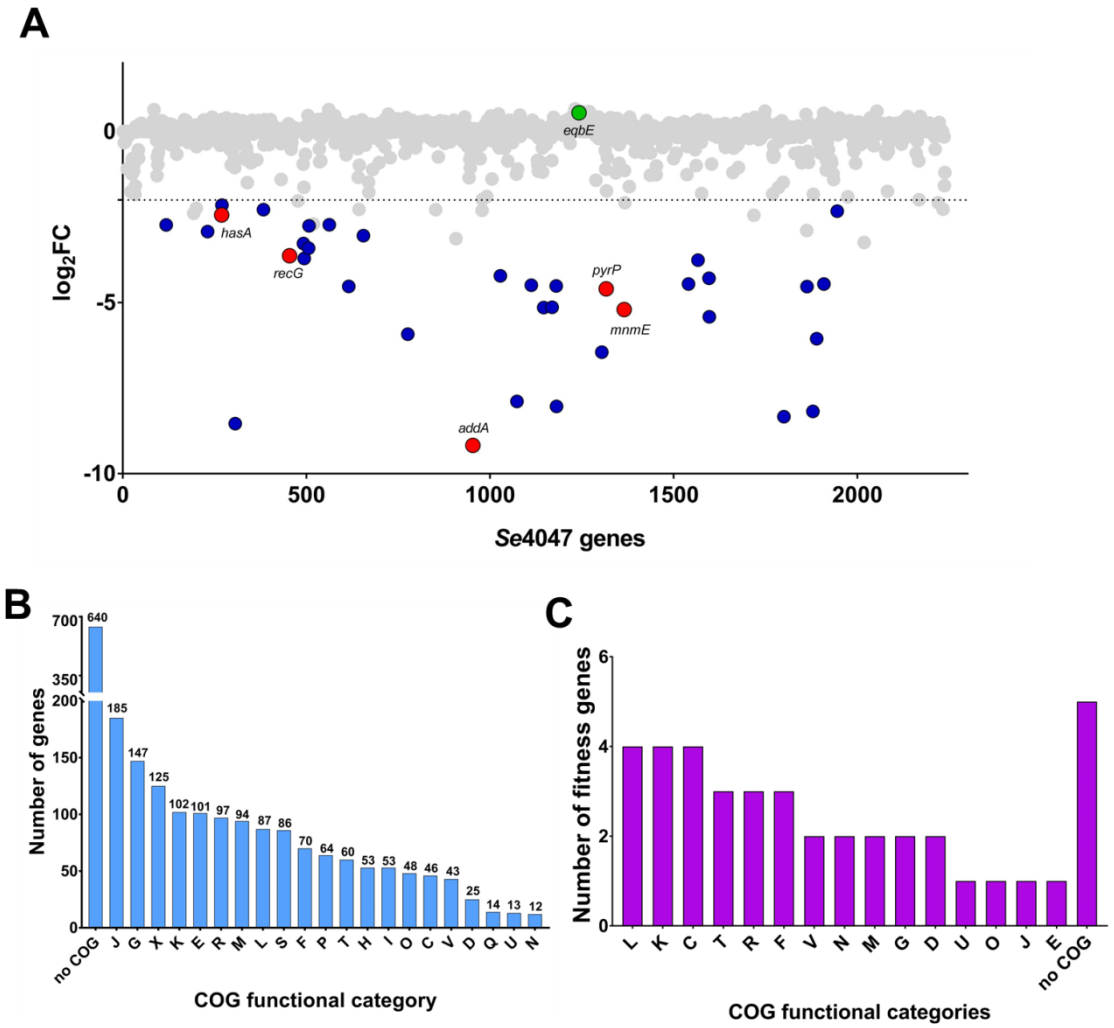


Figure 3.6. Fitness scores and COG categories of *S. equi* genes required for survival in whole equine blood. Fitness scores (log_2FC) per gene of *S. equi* IS*S1* mutants post incubation in whole equine blood, as determined by TraDIS. Blue dots= genes required for fitness (log_2FC < -2 and q < 0.05), red dots= genes required for fitness genes significantly reduced in fitness of which deletion mutants were made and retested to confirm TraDIS results, green dot= *eqbE*, exhibiting no fitness effect that was also used as a control for validation, grey dots= genes exhibiting no significant fitness effect B) Functional COG categories of all annotated *S. equi* genes C) Functional COG categories of the 36 fitness genes identified in whole equine blood. J: Translation, ribosomal structure and biogenesis, G: Carbohydrate transport and metabolism, X: Mobilome: prophages, transposons, K: Transcription, E: Amino acid transport and metabolism, R: General function prediction only, M: Cell wall/membrane/envelope biogenesis, L: Replication, recombination and repair, S: Function unknown, F: Nucleotide transport and metabolism, P: Inorganic ion transport and metabolism, T: Signal transduction mechanisms, H: Coenzyme transport and metabolism, I: Lipid transport and metabolism, O: Posttranslational modification, protein turnover, chaperones, C: Energy production and conversion, V: Defense mechanisms, D: Cell cycle control, cell division, chromosome partitioning, Q: Secondary metabolites biosynthesis, transport and catabolism, U: Intracellular trafficking, secretion, and vesicular transport, N: Cell motility

Table 3.3. *S. equi* genes with reduced fitness in equine whole blood as a result of IS*S1* insertion as identified by TraDIS. Genes highlighted in red and green were deleted by allelic replacement mutagenesis and incubated in whole equine blood to validate TraDIS results.

Gene	Locus tag	Function	Log_2FC	*q* value
ackA	SEQ0118	acetate kinase	-2.7	0.042
SEQ0231	SEQ0231	putative Mga-like regulatory protein	-2.9	<0.0005
hasA	SEQ0269	hyaluronan synthase	-2.4	0.046
hasB	SEQ0270	UDP-glucose 6-dehydrogenase	-2.2	<0.0005
SEQ0306	SEQ0306	putative ssDNA-binding protein	-8.5	<0.0005
pepX	SEQ0383	Xaa-Pro dipeptidyl-peptidase	-2.3	0.017
recG	SEQ0454	ATP-dependent DNA helicase	-3.6	0.001
SEQ0492	SEQ0492	putative mannose-specific phosphotransferase system (PTS), IID component	-3.3	0.042
SEQ0494	SEQ0494	putative mannose-specific phosphotransferase system (PTS), IIAB component	-3.7	0.017
pptA/ecsA	SEQ0506	ABC transporter ATP-binding protein	-3.4	0.021
pptB/ecsB	SEQ0507	ABC transporter protein	-2.8	0.002
SEQ0562	SEQ0562	exodeoxyribonuclease	-2.7	0.022
bipA/typA	SEQ0615	GTPase	-4.5	0.007
pyrD	SEQ0655	putative dihydroorotate dehydrogenase	-3.0	0.007
ppc	SEQ0776	putative phosphoenolpyruvate carboxylase	-5.9	<0.0005
addA	SEQ0953	putative ATP-dependent exonuclease subunit A	-9.2	<0.0005
SEQ1028	SEQ1028	GntR family regulatory protein	-4.2	0.004
SEQ1073	SEQ1073	putative phosphopantothenoylcysteine decarboxylase	-7.9	<0.0005
SEQ1112	SEQ1112	putative exported protein	-4.5	0.001
SEQ1146	SEQ1146	putative phosphate acetyltransferase	-5.1	<0.0005
ldh	SEQ1169	L-lactate dehydrogenase	-5.1	<0.0005
SEQ1180	SEQ1180	putative DNA-binding protein	-4.5	0.003
SEQ1181	SEQ1181	GntR family regulatory protein	-8.0	<0.0005
hupX	SEQ1304	pyridine nucleotide-disulphide oxidoreductase family protein	-6.4	<0.0005
pyrP	SEQ1316	uracil permease	-4.6	<0.0005
mnmE	SEQ1365	tRNA modification GTPase	-5.2	<0.0005
SEQ1540	SEQ1540	putative membrane protein	-4.5	0.003
smc	SEQ1566	putative chromosome partition protein	-3.8	<0.0005
ccpA	SEQ1596	catabolite control protein A	-4.3	0.011
pepQ	SEQ1597	putative Xaa-Pro dipeptidase	-5.4	<0.0005
SEQ1800	SEQ1800	putative exported protein	-8.3	<0.0005
scpA	SEQ1863	segregation and condensation protein A	-4.5	<0.0005
greA	SEQ1879	transcription elongation factor	-8.2	<0.0005
csrS	SEQ1889	sensor histidine kinase	-6.1	<0.0005
yqeK	SEQ1909	hydrolase, HD family	-4.5	0.002
pyrG	SEQ1945	putative CTP synthase	-2.3	<0.0005
eqbE	SEQ1242	equibactin nonribosomal peptide synthase protein	0.6	1

3.3.2 Validation of *S. equi* genes required for fitness in whole equine blood

To validate the TraDIS findings, allelic replacement mutants in 5 genes attenuated in whole equine blood as a result of IS*S1* insertion (*hasA*, *recG, addA*, *pyrP* and *mnmE*) were analysed in whole equine blood (Figure 3.6, red dots, Table 3.3, red). These genes, except *hasA*, were chosen because they have not previously been implicated in survival in the face of the equine immune response. An additional allelic replacement mutant was utilised as a negative control, Δ*eqbE*, as IS*S1* mutants in this gene exhibited no attenuation in whole equine blood (Figure 3.6, green dot, Table 3.3, green row).

Deletion strains were grown in THB with their optical densities measured over time to determine growth characteristics (Figure 3.7). All deletion strains, except Δ*eqbE* grew significantly differently to the parental *Se*4047 strain. Strains Δ*pyrP* and Δ*hasA* grew significantly faster than *Se*4047 ($p < 0.005$ and $p < 0.05$, respectively) and strains Δ*addA*, Δ*recG* and Δ*mnmE* grew significantly slower than *Se*4047 (all $p < 0.005$).

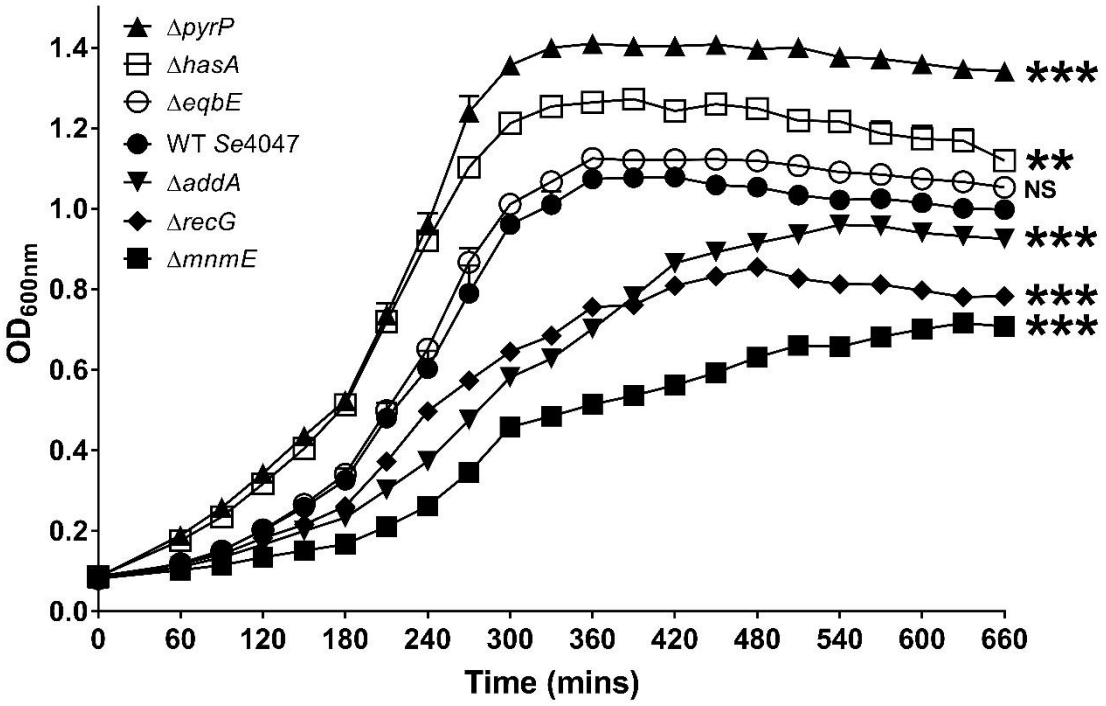


Figure 3.7. Growth curves of the parental *S. equi* strain 4047 and Δ*pyrP*, Δ*hasA,* Δ*eqbE,* Δ*addA*, Δ*recG* and Δ*mnmE* deletion mutants in Todd-Hewitt broth. ***= $p < 0.005$ ** = $p < 0.05$. In some cases, the error bars lie within the point and are therefore not visible.

The 6 deletions strains were incubated in whole equine blood with reduced bacterial loads compared to the preceding TraDIS assay. The CFU/ml of each deletion strain was reduced in the validation assays to more closely reflect the proportion of attenuated *S. equi* IS*S1* mutants present in the TraDIS assay. Validation assays were also incubated for an additional hour compared to the TraDIS assay as some attenuated effects may have been more detrimental when in competition with other neighbouring IS*S1* mutants.

The survival of each deletion strain in whole equine blood was measured over time and statistically compared to the survival of *Se*4047 (Figure 3.8). The Δ*hasA* strain was highly attenuated in whole equine blood at all time points (all $p < 0.01$, Figure 3.8A), which is reflected in published works describing this strain [13]. The Δ*addA* strain was also significantly attenuated at all time points (T1+T2, $p < 0.01$, T3, $p < 0.05$, Figure 3.8B). Survival of the Δ*recG* strain was not significantly different to *Se*4047 after 1 hour, but

was however, significant at 2 and 3 hours (T2, $p < 0.01$, T3 $p < 0.05$, Figure 3.8C). Survival of the Δ*pyrP* strain was not significantly attenuated, despite the apparent reduced survival observed at 3 hours ($p = 0.065$, Figure 3.8D). Neither the Δ*mnmE* or the Δ*eqbE* strains were attenuated at any timepoints and closely matched the survival of *Se*4047 in whole equine blood (Figure 3.8EF).

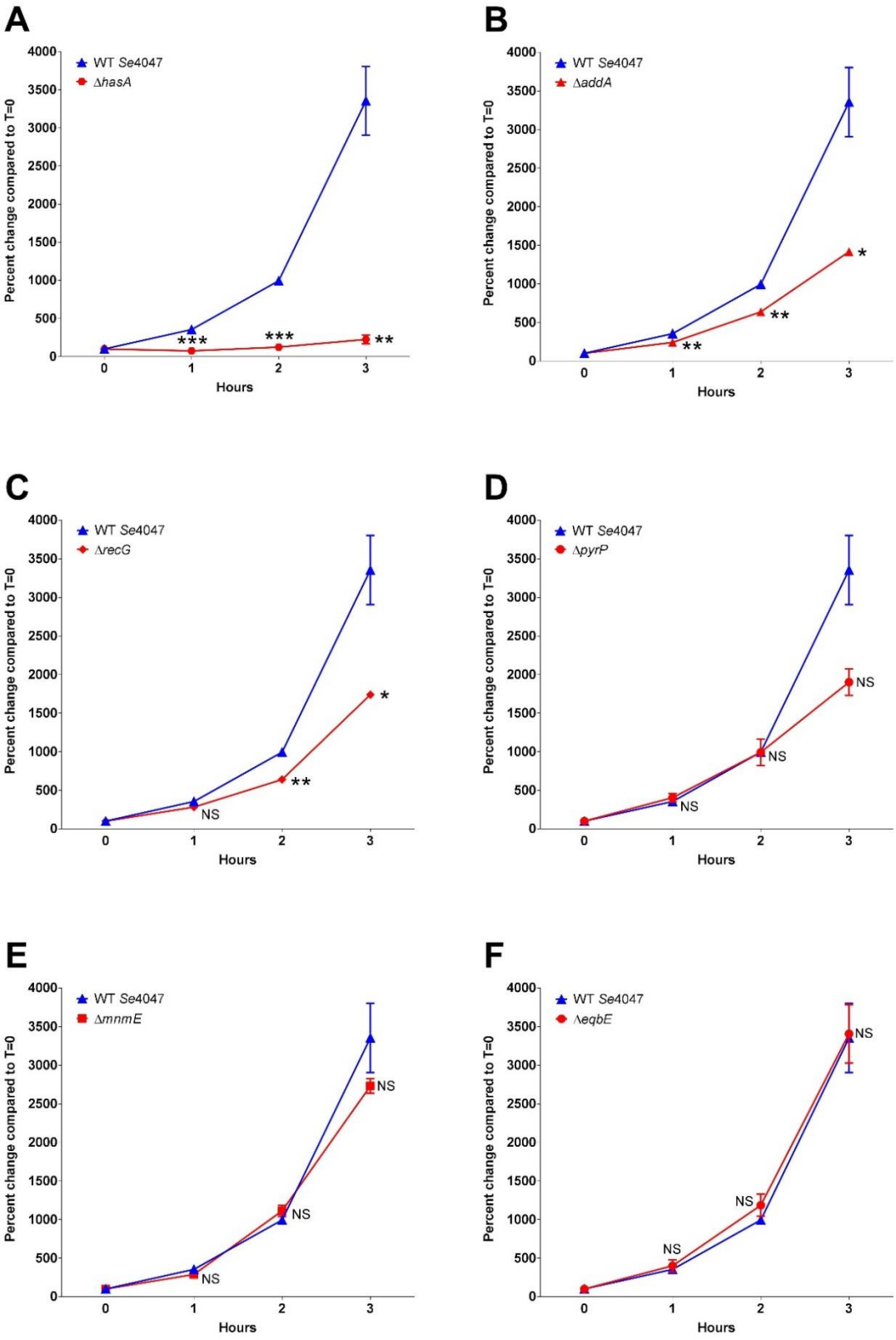


Figure 3.8. Validation of an *S. equi* TraDIS screen in whole equine blood. Deletion mutants in whole equine blood fitness genes, as identified by TraDIS, were incubated in blood for 3 hours and their survival measured each hour. A) Δ*hasA*, B) Δ*addA*, C) Δ*recG*, D) Δ*pyrP*, E) Δ*mnmE* and F) Δ*eqbE* deletion mutants compared to the wild-type parental strain, *Se*4047. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. In some cases, the error bars lie within the point and are therefore not visible.

3.3.3 *S. equi* minimum inhibitory concentration in hydrogen peroxide

S. equi was incubated overnight in doubling dilutions of THB containing 1.5 percent to 0.000092 percent H_2O_2 in a 96-well plate. Visual inspection of the plate concluded that the minimum inhibitory concentration (MIC) of *S. equi* in THB containing H_2O_2 is 0.0016 percent. A quarter of the MIC, 0.0004 percent, was used in the TraDIS experiments to exert an environmental pressure on the mutants, but not to prevent growth completely. A control plate was also prepared, which contained ddH_2O in place of H_2O_2. Consistent growth across the control plate was observed.

3.3.4 Genes that contribute to the fitness of *S. equi* in hydrogen peroxide

Each input library represented between 79.4 and 80.7 percent of the 2,165 *S. equi* genes (Table 3.4) before data was filtered. Certain thresholds were imposed on the data as previously described in section 3.2.3, permitting the inclusion of 24,372 unique mutants in library AC, 22,734 unique mutants in library CT, and 26,226 in library GA, representing 67.9 percent of *S. equi* genes and 92.5 percent of the non-essential genes previously identified in Chapter 2.

The 3 output libraries recovered from H_2O_2 contained on average, 2.1 percent ± 6.3 (SEM) fewer unique mutants than were present in the input libraries (Table 3.5). These values are skewed due to the over representation of the CT output library in the sequencing data. A higher concentration of DNA originating from this library was inadvertently sequenced compared to the other 5 DNA libraries sequenced on the same MiSeq run. To adjust for this, total reads counts per library were normalised between output pools, as is normally completed.

Table 3.4. Composition of hydrogen peroxide input libraries pre- and post-filtering.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC-IN^{pre}	27,920	647,044	1,747 (80.7 : 100)
CT-IN^{pre}	25,951	648,236	1,720 (79.4 : 100)
GA-IN^{pre}	29,973	631,309	1,737(80.2: 100)
AC-IN^{post}	24,372	567,560	1,471 (67.9 : 92.5)
CT-IN^{post}	22,734	569,090	1,471 (67.9 : 92.5)
GA-IN^{post}	26,226	568,628	1,471 (67.9 : 92.5)

Table 3.5. Composition of hydrogen peroxide output libraries pre- and post-filtering.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC-OUT^{pre}	25,239	718,526	1,680 (77.6: 100)
CT-OUT^{pre}	28,658	2,020,075	1,718 (79.4: 100)
GA-OUT^{pre}	27,436	551,893	1,674 (77.3 : 100)
AC-OUT^{post}	22,182	1,406,042	1,471 (67.9 : 92.5)
CT-OUT^{post}	25,138	1,406,042	1,471 (67.9 : 92.5)
GA-OUT^{post}	24,179	1,406,042	1,471 (67.9 : 92.5)

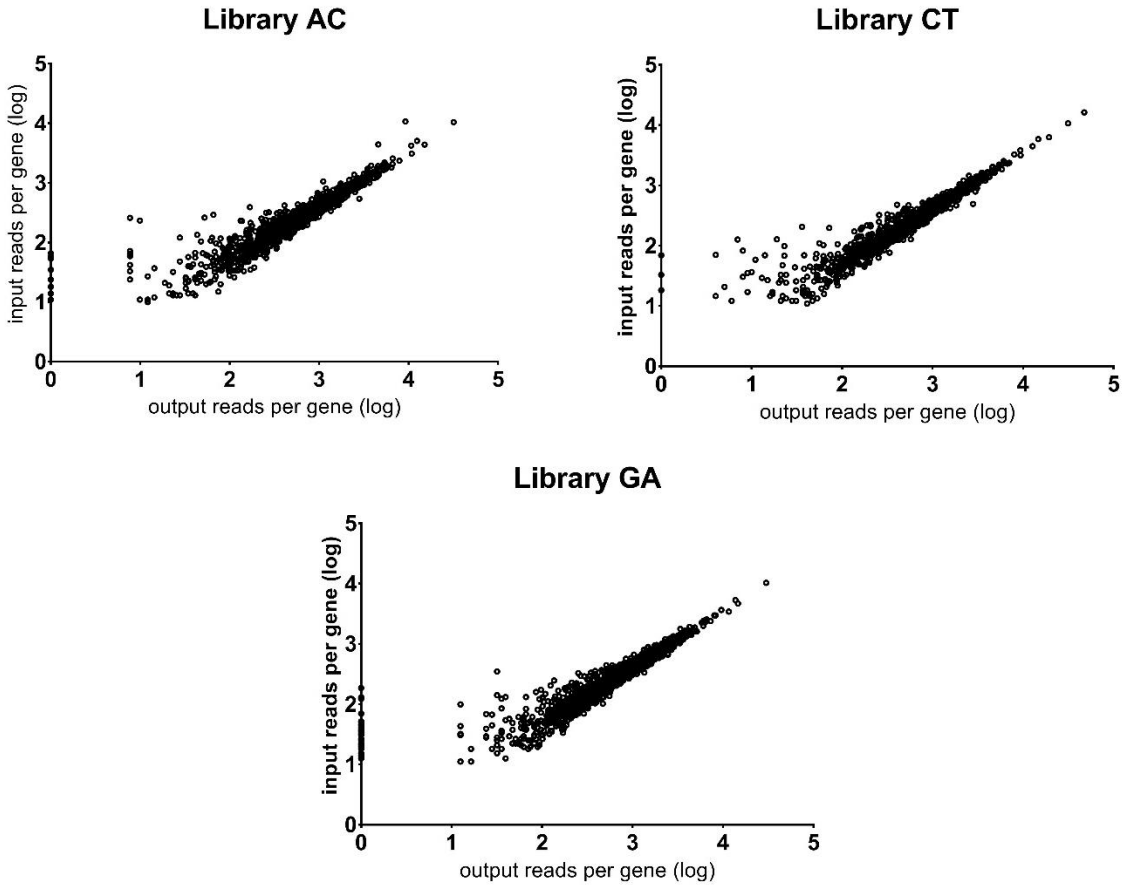


Figure 3.9. Read counts per gene in each of 3 S. equi barcoded ISS1 libraries, pre- (input) and post- (output) exposure to H2O2. Genes represented by < 100 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Barcoded mutant libraries were each exposed to H_2O_2 to identify genes contributing to fitness in this environment. TraDIS was used to identify any population changes in the recovered (output) mutants versus the input pools, alluding to gene fitness. In the majority of genes included in the analysis, the number of reads sequenced per gene remained comparable between both the input and output libraries, highlighting genes that were not required for survival in whole equine blood (Figure 3.9). Some genes

however were represented by fewer reads in the output pools compared to the input pools, alluding to reduced fitness. The effect of incubation with H_2O_2 on the fitness of IS*S1* mutants was determined by calculating the log_2FC for all 1,471 genes passing the inclusion criteria (Figure 3.10A). IS*S1* insertion in 15 genes significantly reduced the fitness of *S. equi* ($log_2FC < -2$ and $q < 0.05$, Table 3.6, Figure 3.10A, blue and red dots), with the remaining 1,456 genes exhibiting no growth effect in H_2O_2 (Figure 3.10A, grey and green dots).

Cluster of orthologous groups (COG) analysis of the 15 fitness genes (Figure 3.10B) identified that the most prevalent categories included genes involved in energy production and conversion (*n*=4, 8.7 percent of total in COG category (Figure 3.6B)) and replication, recombination and repair (*n*=3, 3.4 percent of total in COG category (Figure 3.6B)).

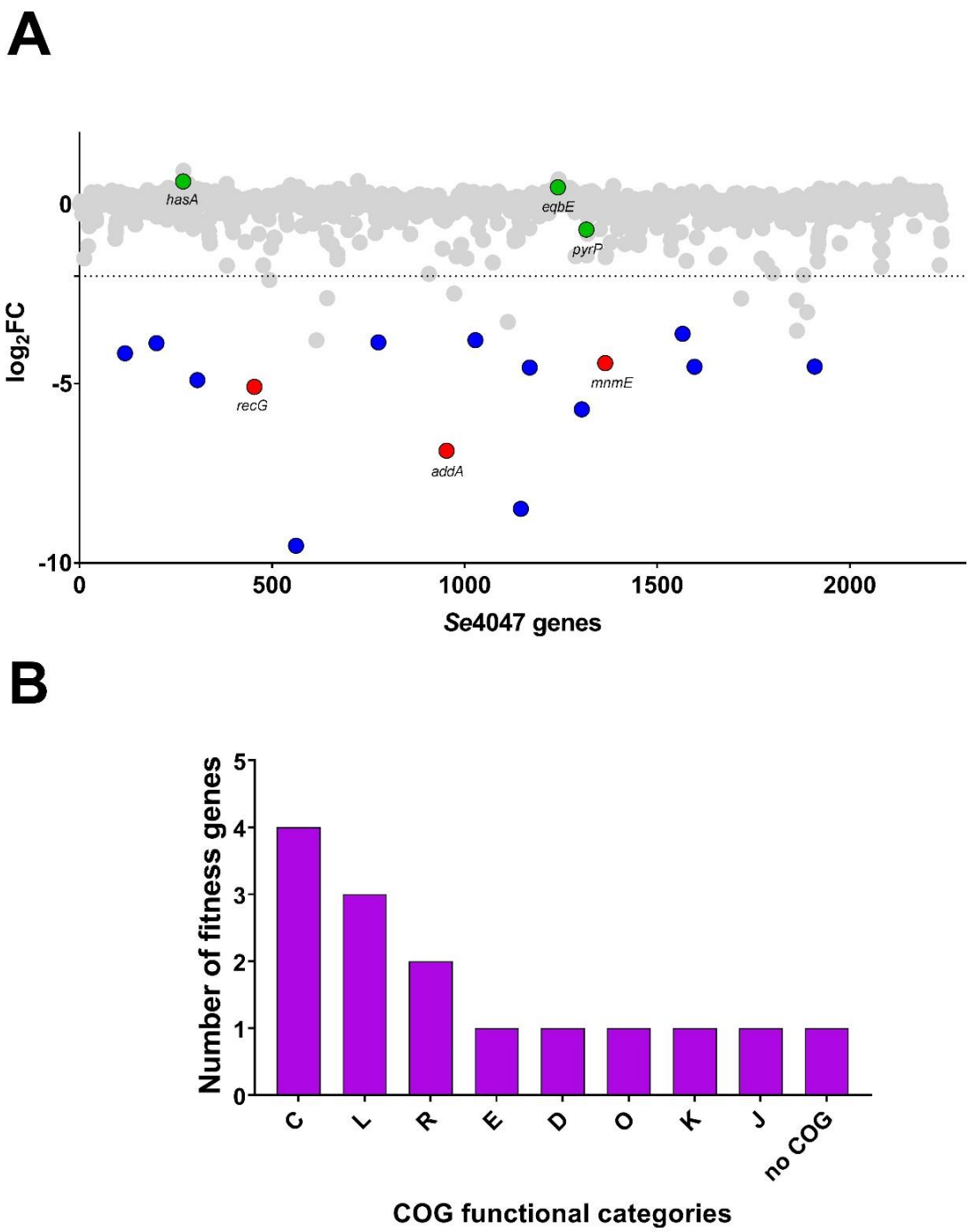


Figure 3.10. Fitness scores and COG categories of *S. equi* genes required for survival in hydrogen peroxide (H_2O_2). A) Fitness scores (log_2FC) per gene of *S. equi* IS*S1* mutants post incubation in H_2O_2, as determined by TraDIS. Blue dots= genes required for fitness ($log_2FC < -2$ and $q < 0.05$), red dots= genes required for fitness of which deletion mutants were made and retested to confirm TraDIS results, green dots= genes exhibiting no fitness effect that acted as negative controls for validation, grey dots= genes exhibiting no significant fitness effect. B) Functional COG categories of the fitness genes identified in H_2O_2. C: Energy production and conversion, L: Replication, recombination and repair, R: General function prediction only, E: Amino acid transport and metabolism, D: Cell cycle control, cell division, chromosome partitioning, O: Posttranslational modification, protein turnover, chaperones, K: Transcription, J: Translation, ribosomal structure and biogenesis.

Fourteen of the 15 genes identified in the H_2O_2 TraDIS screen, were also implicated in survival in whole equine blood (null= 1.3 genes) (Table 3.6, Figure 3.11). One gene that was uniquely identified in the H_2O_2 TraDIS screen, *ctsR*, is a negative transcriptional regulator involved in resisting environmental stresses, such as temperature, UV and acid [149-151] (Table 3.6, blue).

Table 3.6. *S. equi* genes with reduced fitness in the presence of hydrogen peroxide (H_2O_2) as a result of IS*S1* insertion, as identified by TraDIS. One gene highlighted in blue was uniquely identified in the presence of H_2O_2 when compared to genes identified as required for fitness in whole equine blood. The remaining genes were similarly identified as required in whole equine blood. The genes highlighted in red and green were deleted by allelic replacement mutagenesis and deletion strains incubated in Todd-Hewitt containing H_2O_2 to validate TraDIS results.

Gene	Locus tag	Function	Log_2FC	*q* value
SEQ0118	SEQ0118	acetate kinase	-4.1	0.0021
ctsR	SEQ0200	transcriptional regulator	-3.9	0.0221
SEQ0306	SEQ0306	putative ssDNA-binding protein	-4.9	<0.0005
recG	SEQ0454	ATP-dependent DNA helicase	-5.1	<0.0005
SEQ0562	SEQ0562	exodeoxyribonuclease	-9.5	<0.0005
ppc	SEQ0776	putative phosphoenolpyruvate carboxylase	-3.8	0.0221
addA	SEQ0953	putative ATP-dependent exonuclease subunit A	-6.9	<0.0005
SEQ1028	SEQ1028	GntR family regulatory protein	-3.8	0.0071
SEQ1146	SEQ1146	putative phosphate acetyltransferase	-8.5	<0.0005
ldh	SEQ1169	L-lactate dehydrogenase	-4.5	0.0015
hupX	SEQ1304	pyridine nucleotide-disulphide oxidoreductase family protein	-5.7	<0.0005
mnmE	SEQ1365	tRNA modification GTPase	-4.4	<0.0005
smc	SEQ1566	putative chromosome partition protein	-3.6	<0.0005
pepQ	SEQ1597	putative Xaa-Pro dipeptidase	-4.5	<0.0005
yqeK	SEQ1909	hydrolase, HD family	-4.5	0.0009
hasA	SEQ0269	hyaluronan synthase	0.6	1
pyrP	SEQ1316	uracil permease	-0.7	1
eqbE	SEQ1242	equibactin nonribosomal peptide synthase protein	0.5	1

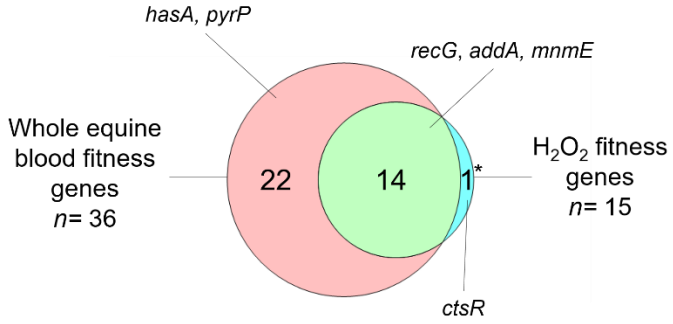


Figure 3.11. Venn diagram of the 36 genes required for the survival of *S. equi* in whole equine blood compared to the 15 genes required survival in hydrogen peroxide.

3.3.5 Validation of *S. equi* genes required for fitness in hydrogen peroxide

Of the 5 validation mutants predicted to be reduced in fitness in whole equine blood by TraDIS, 3 were similarly attenuated in H_2O_2 by TraDIS (Table 3.6, red, Figure 3.10A, red dots). IS*S1* insertion in *hasA* and *pyrP* did not confer any defects in H_2O_2 (Table 3.6, red, Figure 3.10A, green dots).

The survival of each deletion strain in H_2O_2 was measured over time and statistically compared to the survival of *Se*4047 (Figure 3.12). The Δ*hasA* strain was attenuated in H_2O_2 at T2 ($p < 0.01$), but not at the other time points as expected (Figure 3.12A). The Δ*addA* strain was significantly attenuated at all time points (T1, $p < 0.01$ and T2+ T3, $p < 0.001$, Figure 3.12B). Survival of the Δ*recG* strain was significantly different to *Se*4047 at all timepoints (T1, $p < 0.05$, T2, $p < 0.001$ and T3 $p < 0.01$, Figure 3.12C). The Δ*mnmE* strain was not attenuated at T1, but was significantly reduced in fitness at both later timepoints (T2, $p < 0.05$, T3, $p < 0.01$, Figure 3.12E). Neither the Δ*pyrP* or the Δ*eqbE* strains were attenuated at any timepoints, closely matching the survival of *Se*4047 in H_2O_2 and reflecting the TraDIS screen results (Figure 3.12DF).

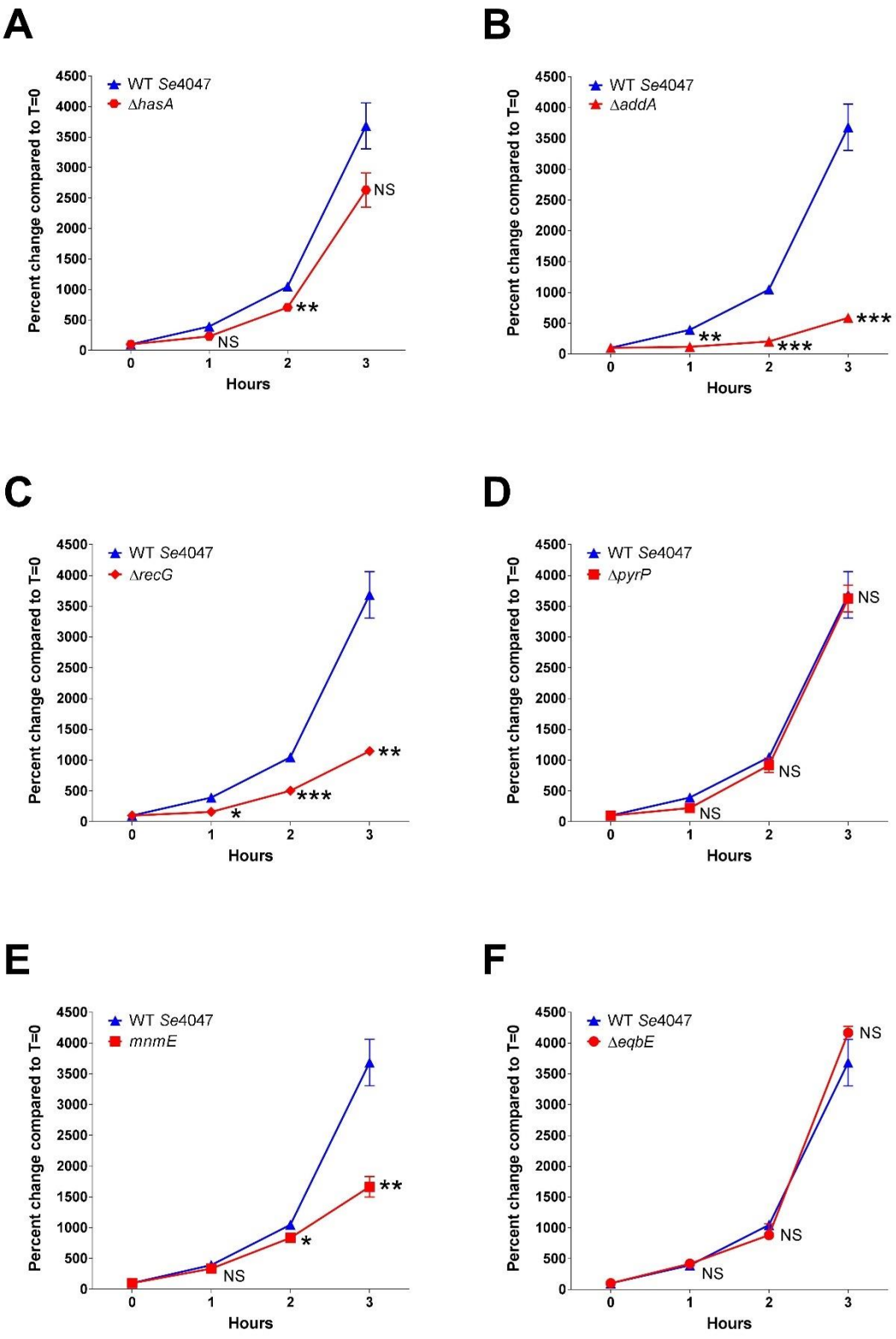


Figure 3.12. Validation of an *S. equi* TraDIS screen in Todd-Hewitt broth (THB) containing hydrogen peroxide (H_2O_2). Deletion mutants in H_2O_2 fitness genes, as identified by TraDIS, were incubated in THB containing H_2O_2 for 3 hours and their survival measured each hour. A) Δ*hasA*, B) Δ*addA*, C) Δ*recG*, D) Δ*pyrP*, E) Δ*mnmE* and F) Δ*eqbE* deletion mutants compared to the wild-type parental strain, *Se*4047. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. In some cases, the error bars lie within the point and are therefore not visible.

3.4 Discussion

The results of the experiments outlined in this Chapter describe the genome-wide identification of genes required by *S. equi* for survival in whole equine blood and H_2O_2, conditions that mimic an interaction with the equine immune response. IS*S1* mutants in 36 genes were significantly reduced in fitness upon exposure to whole equine blood, and 15 genes when exposed to THB containing H_2O_2. Fourteen genes were commonly identified between the 2 conditions, with 1 gene, *ctsR*, uniquely required for H_2O_2 resistance and 22 genes uniquely employed in the presence of whole equine blood. It is likely that fewer genes overall were identified in the H_2O_2 screen due to the presence of THB that provided a rich source of nutrients. Additionally, whole equine blood is a much more complex environment placing higher levels of selection on the mutant population.

3.4.1 Genes validated in whole equine blood and H_2O_2

Deletion mutants lacking 4 novel genes, Δ*addA,* Δ*recG*, Δ*pyrP* and Δ*mnmE*, identified as contributing to fitness in whole equine blood, were generated and re-tested in isolation to confirm the TraDIS results. Two control mutants were also tested, a capsule deletion mutant, Δ*hasA*, identified by TraDIS and proven to be attenuated in equine blood in other works, and a Δ*eqbE* mutant, for which fitness was not affected in the whole equine blood TraDIS screen. The Δ*eqbE* mutant survived comparably to the wild-type parental strain, confirming the TraDIS findings.

The susceptibility of streptococcal capsule mutants to killing in both *in vitro* and *in vivo*, has long been known [13, 152-157]. Disrupting the capsule, exposes the bacterial surface rendering the cells more susceptible to immune attack. A *S. equi* Δ*hasA* (hyaluronan synthase) mutant was previously shown to be highly susceptible to killing in equine blood, which was reflected in the TraDIS whole equine blood screen (Figure 3.13) and in the validation experiment.

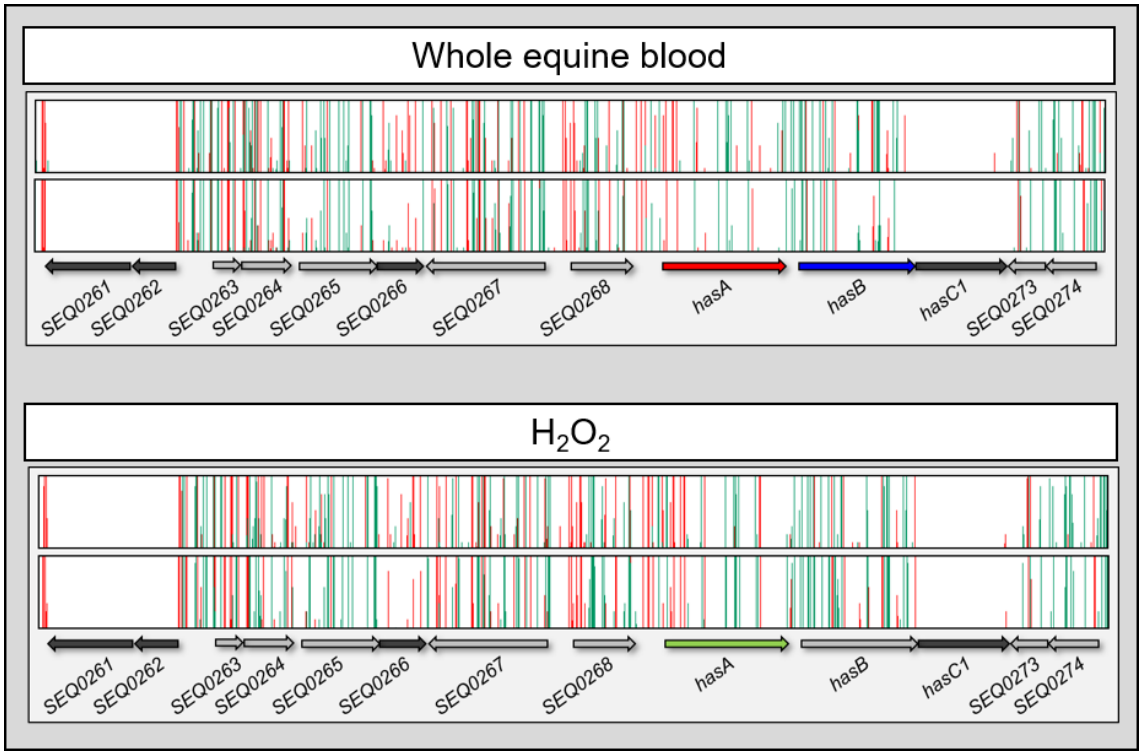


Figure 3.13. Prevalence of *S. equi* IS*S1* mutants in the genes *SEQ0261-SEQ0274* pre- and post-exposure to whole equine blood and H_2O_2. The top panels represent mutants present in the input pools, with the bottom panels representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. The capsule biosynthesis gene, *hasA*, is required for survival in whole equine blood by TraDIS, which is evident from a loss of reads within the recovered output population (red arrow). The requirement of *hasA* was successfully validated in whole equine blood using a gene deletion mutant. HasB, was also identified as important in whole equine blood (blue arrow). HasA was not essential for survival in H_2O_2 (green arrow). Light grey arrows indicate non-essential genes. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

Interestingly, the log_2FC for *hasA* determined in the TraDIS screen was only -2.4 (q= 0.046), close to the threshold of -2 used to determine attenuation, yet in isolation, the Δ*hasA* mutant was dramatically reduced in fitness in whole equine blood. It might be possible that acapsular mutants are able to benefit from the retained capsule of neighbouring mutants. This effect is reflected for *hasB* (UDP-glucose 6-dehyrdogenase), where fitness was also close to the threshold in the TraDIS screen (log_2FC= -2.2, q=<0.0005). Acapsular IS*S1* mutants were not attenuated when exposed to H_2O_2 (log_2FC= 0.6, q=1) (Figure 3.13), which was confirmed with the Δ*hasA* deletion mutant, suggesting that the capsule does not play a role in resisting killing by H_2O_2.

addAB (a.k.a *rexAB*) encode a major component of the homologous recombination process, acting to repair double strand breaks (DSBs), by catalysing the unwinding of DNA [158-160]. In *S. equi*, *addB* was essential for survival in THB (insertion index= 0.03, essential genes < 0.034, calculated from master library data in chapter 2), but cells could survive with a disrupted *addA.* Insertions were not dense in *addA* (Figure 3.12), however, equating to an insertion index of 0.042, which was considerably less than the average insertion index for non-essential genes (0.15) (Figure 3.14).

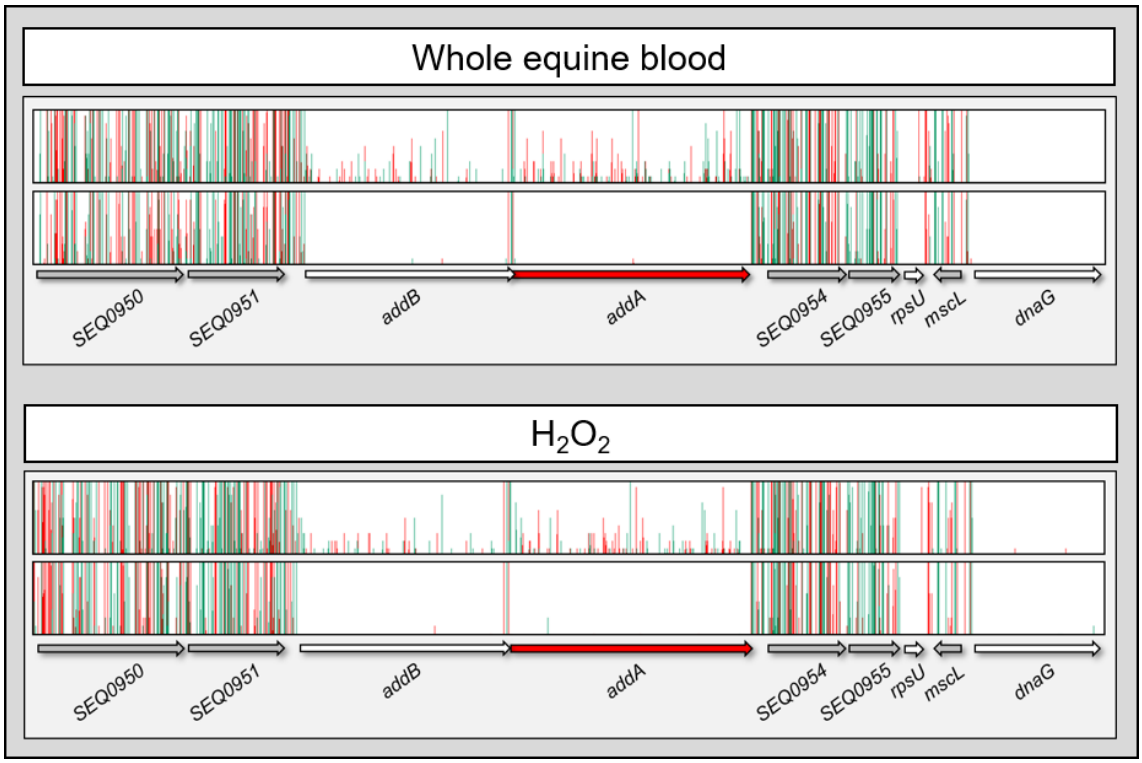


Figure 3.14. Prevalence of *S. equi* IS*S1* mutants in the genes *SEQ0950-SEQ0958* pre- and post-exposure to whole equine blood and H_2O_2. The top panels represent mutants present in the input pool, with the bottom panels representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. The double strand break repair gene, *addA*, is required for survival in whole equine blood and H_2O_2 by TraDIS, which is evident from a loss of reads within the recovered output populations (red arrows). The requirement of *addA* in both conditions was sucessfully validated using a whole deletion mutant. White arrows indicate essential genes in THB. Light grey arrows indicate non-essential genes. Data is viewed in Artemis [112].

It is likely that if the library was passaged, *addA* mutants would decrease in prevalence over time. In support of this, the Δ*addA* deletion mutant grew significantly slower than wild-type *Se*4047 in THB. This slow growth phenotype was also observed in *S. pneumoniae* Δ*addA* and Δ*addB* mutants [159]. Unsurprisingly, the *S. equi* Δ*addA* deletion mutant was significantly attenuated in whole equine blood and H_2O_2, a result

that is likely to be replicated under any stress conditions, especially when there is greater pressure to repair DSBs. In support of these findings, *S. pneumoniae* and *Helicobacter pylori* (*H. pylori*) Δ*addA* and Δ*addB* mutants were more susceptible to UV damage than the wild-type parental strains [159, 161]. These Δ*addA* and Δ*addB* mutants in *H. pylori* were also hypersensitive to the alkylating agent, mitocmycin C and the DNA gyrase inhibitor, ciprofloxacin and were less able to colonise the stomach in murine models of infection [161].

RecG is an ATP-dependent DNA helicase that is also critical for efficient recombination and DNA repair. RecG promotes the resolution of Holliday junctions by catalysing the conversion of junction intermediates to mature products by branch migration [162]. RecG is also thought to remove RNA from R-loops by unwinding the RNA-DNA hybrid [163, 164]. The insertion index of *recG* in the input libraries more closely reflected the average insertion index of all non-essential genes (0.092). The abundance of insertions is clear in Figure 3.15. As with Δ*addAB* mutants, Δ*recG* mutants in *E. coli* were more susceptible to UV light than the wild-type parental strain [165]. UV sensitivity was however greatly enhanced when additional *ruv* genes were deleted in the Δ*recG* background, suggesting that there is a functional overlap between these genes [166]. *S. equi* IS*S1* mutants in *recG* may therefore remain viable thorough the continued functioning of *ruv* genes.

recG IS*S1* mutants were more significantly attenuated in H_2O_2 compared to whole equine blood (H_2O_2; $\log_2$FC= -5.1, q= <0.0005, whole equine blood; $\log_2$FC= -3.6, q= 0.001), likely due to DNA degrading ability of H_2O_2, incurring a greater requirement for a functioning *recG* to repair damaged DNA. The *S. equi* Δ*recG* mutant was significantly attenuated in whole equine blood and H_2O_2, confirming the TraDIS screen results. This mutant, however, grew significantly slower in THB than *Se*4047, so as seen in the Δ*addA* mutant, these strains are likely to have a predisposed sensitivity to stress conditions.

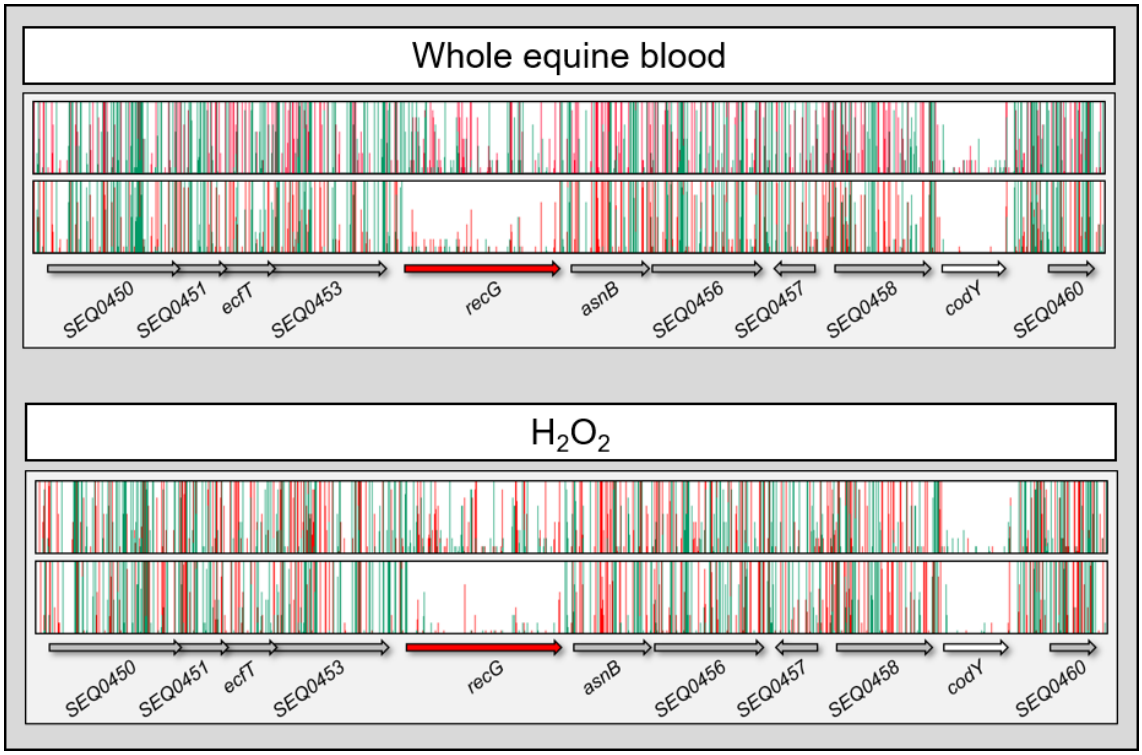


Figure 3.15. Prevalence of *S. equi* IS*S1* mutants in the genes *SEQ0450-SEQ0460* pre- and post-exposure to whole equine blood and H_2O_2. The top panels represent mutants present in the input pool, with the bottom panels representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. The ATP-dependent DNA helicase gene, *recG*, is required for survival in whole equine blood and H_2O_2 by TraDIS, which is evident from a loss of reads within the recovered output populations (red arrows). The requirement of *recG* was successfully validated in both conditions using a whole deletion mutant. White arrows indicate essential genes in THB. Light grey arrows indicate non-essential genes. Data is viewed in Artemis [112].

A membrane bound uracil permease, encoded by *pyrP,* scavenges uracil from the environment for pyrimidine biosynthesis [167]. Transcription of *pyrP* was downregulated in the transcriptome of *S. pneumoniae* upon switching from a colonising biofilm state, to that of invasive disease in a murine model of infection, suggesting that this gene is only required in colonising states in this streptococcal species [168]. In *S. pyogenes* serotype M28, transcription of *pyrP* was upregulated upon exposure to human amniotic fluid *ex vivo*, a niche in which this serotype is able to readily persist [169]. PyrP was required for fitness in whole equine blood, but not in H_2O_2, according to the TraDIS data (H_2O_2; log_2FC= -0.7, q= 1, whole equine blood; log_2FC= -4.6, q= <0.0005) (Figure 3.16). Interestingly, *pyrD* and *pyrG*, which are located elsewhere in the genome and are involved in the downstream biosynthetic pyrimidine pathway, were also required by *S. equi* in whole equine blood *in vitro*.

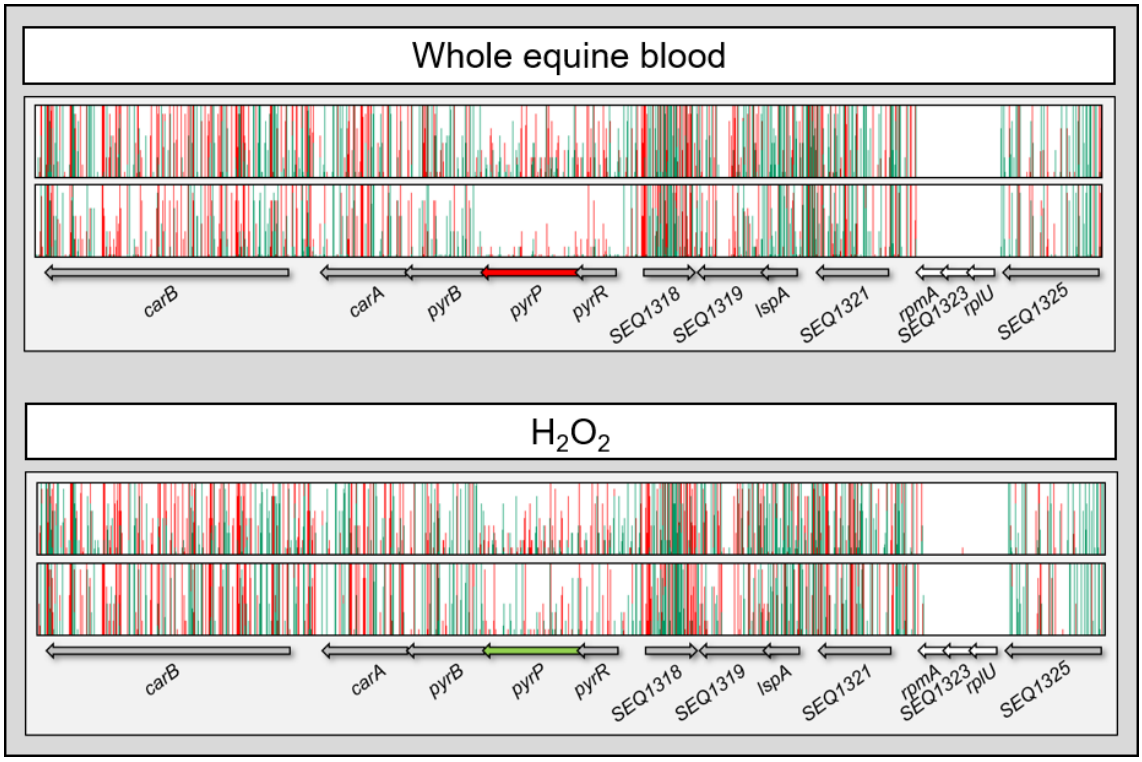


Figure 3.16. Prevalence of *S. equi* IS*S1* mutants in the genes *SEQ1313-SEQ1325* pre- and post-exposure to whole equine blood and H_2O_2. The top panels represent mutants present in the input pool, with the bottom panels representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. The uracil permease gene, *pyrP*, is required for survival in whole equine blood by TraDIS, which is evident from a loss of reads within the recovered output population (red arrow). The requirement of *pyrP* was not successfully validated in whole equine blood using a whole deletion mutant, as survival was comparable to the wild-type parental strain. PyrP was not essential for survival in H_2O_2 (green arrow). White arrows indicate essential genes in THB. Light grey arrows indicate non-essential genes. Data is viewed in Artemis [112].

Survival of the Δ*pyrP* deletion mutant in whole equine blood and H_2O_2 reflected that of the wild-type *Se*4047 strain. Interestingly, the Δ*pyrP* deletion mutant grew significantly faster in THB than the wild-type *Se*4047 strain. In THB, removal of this gene may represent an energy saving, where cells are still able to scavenge the required nutrients without importing uracil through the potentially energy costly PyrP. A *L. lactis* Δ*pyrP* deletion strain was unable to exploit uracil when provided at low concentrations, but at high concentrations, no effect was observed [167]. These data suggest that at high concentrations, such as in THB, uracil can be acquired by means not dependent on PyrP. The dispensability of PyrP in THB supports the H_2O_2 TraDIS and validation findings, since the vast majority of the medium was THB. The attenuation of the *pyrP* IS*S1* mutants in whole equine blood is reasonable, since the blood is likely to contain little uracil. The Δ*pyrP* deletion mutant appeared susceptible to whole equine blood after 3

hours of incubation, compared to the wild-type *Se*4047, but did not reach statistical significance (p= 0.06).

MnmE (a.k.a TrmE) is a tRNA modification enzyme that forms a heterotetrameric complex with MnmG (a.k.a GidA) [170, 171]. The MnmEG complex catalyses 2 different GTP and FAD dependent reactions, resulting in 5-aminomethyluridine and 5-carboxymethylaminomethyluridine, utilising ammonium and glycine as substrates, respectively [170]. GTP hydrolysis by MnmE causes structural rearrangements within the MnmEG complex, which is necessary for subsequent tRNA modification in *E. coli* [172]. In *S. equi*, MnmG is essential for survival in THB [103] and critical for survival in *S. pyogenes* and *S. agalactiae* [77, 78]. A Δ*mnmE* deletion mutant in *S. pyogenes* was reduced in expression of known virulence factors such as streptolysin O, M-protein, mitogenic factor and NAD-glycohydrolase, an effect that was reflected in a Δ*mnmG* deletion mutant [173]. In *S. mutans*, deletion of either *mnmE* or *mnmG* resulted in a 50 percent decrease in glucose-dependent biofilm formation [174]. Deletion of either *mnmE* or *mnmG* in *E. coli* caused translational errors, decreased growth rates and extreme sensitivity to acidic pH [175-178]. Growth rate was also significantly decreased in the *S. equi* Δ*mnmE* deletion mutant in THB in comparison to wild-type *Se*4047. IS*S1* mutants in *mnmE* were reduced in fitness in both whole equine blood and H_2O_2, which was only replicated with the Δ*mnmE* deletion mutant in H_2O_2 (whole equine blood; $\log_2$FC= -5.2, q= <0.0005, H_2O_2; $\log_2$FC= -4.4, q= <0.0005) (Figure 3.17).

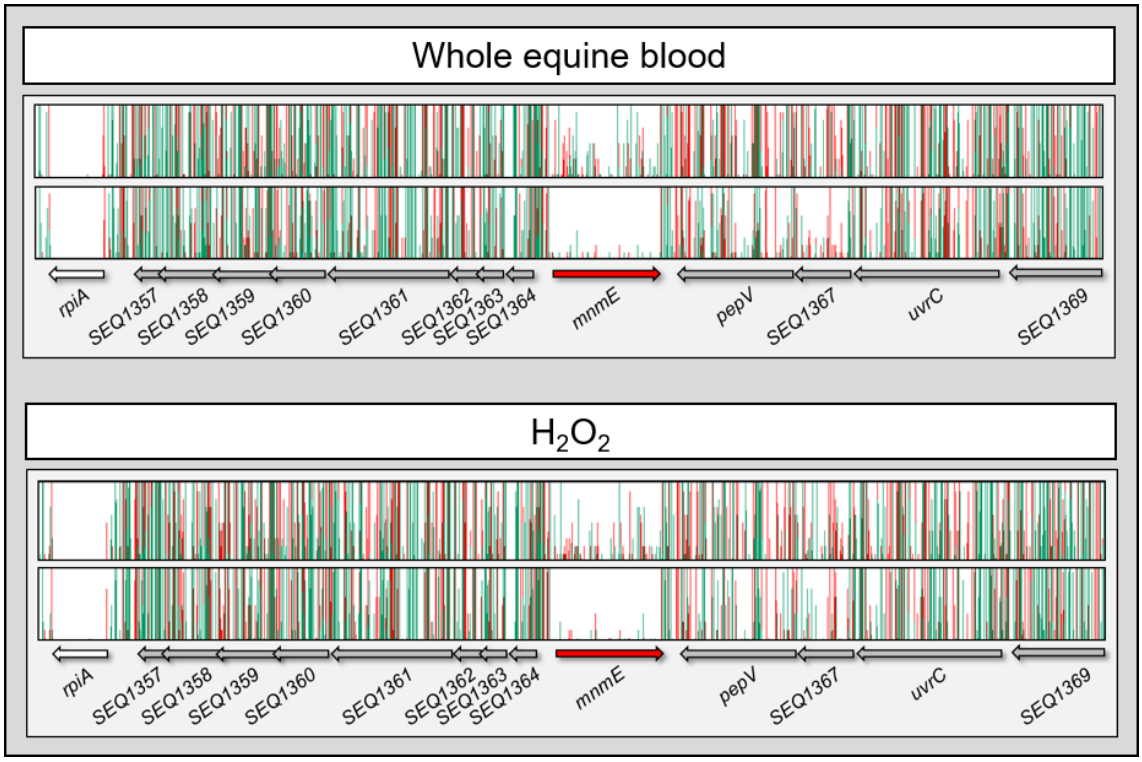


Figure 3.17. Prevalence of *S. equi* IS*S1* mutants in the genes *SEQ1356-SEQ1369* pre- and post-exposure to whole equine blood and H_2O_2. The top panels represent mutants present in the input pool, with the bottom panels representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. The tRNA modification gene, *mnmE*, is required for survival in whole equine blood and H_2O_2 by TraDIS, which is evident from a loss of sequencing reads within the recovered output populations (red arrows). The requirement of *mnmE* was only successfully validated in H_2O_2 using a gene deletion mutant. Survival of the *mnmE* mutant in whole equine blood was comparable to the wild-type parental strain. White arrows indicate essential genes in THB. Light grey arrows indicate non-essential genes. Data is viewed in Artemis [112].

In whole equine blood, the survival of the Δ*mnmE* deletion mutant and wild-type *Se*4047 were extremely similar, suggesting that an element of the whole equine blood reversed the growth phenotype seen in THB, and potentially that when in competition with other IS*S1* mutants, *mnmE* IS*S1* mutants are not competitive. Since the majority of the medium used in H_2O_2 experiments is THB, it is not surprising that both *mnmE* IS*S1* mutants and the Δ*mnmE* deletion mutant were significantly attenuated.

3.4.2 Other fitness genes

CsrR and CsrS comprise a two-component regulatory system, encoding the response regulator and sensor histidine kinase, respectively. In streptococci, CsrRS regulates virulence genes in reaction to environmental signals, with its mechanisms of action varying between species. In *S. pyogenes*, the CsrRS complex represses virulence genes, such as those for capsule synthesis and streptolysin S and exotoxin B production,

therefore is relatively inactive during infection. [179]. In *S. agalactiae,* CsrRS represses and activates virulence related genes, although the target genes are highly variable between strains. In *S. agalactiae* strain NEM316, capsule production was regulated by CsrRS [180], yet in strains 2603 and 515, which lack *csrR*, no change in capsule production was observed [181, 182]. This degree of variability may be a result of pathogen adaption to specific host niches. Such variability is unlikely to be seen between strains of *S. equi* due to its restriction to both its host and site of infection and the relative lack of population diversity [15, 183]. *CsrR* is essential in THB in *S. equi,* so was not included in the whole equine blood and H_2O_2 screens. CsrS, however, was required for fitness in whole equine blood, but not in H_2O_2 (whole equine blood; log_2FC= -6.1, *q*= <0.0005, H_2O_2; log_2FC= -3, *q*= 0.3), likely because of the presence of immune factors in blood. These results suggest that CsrRS activates *S. equi* genes, as occurs in some *S. agalactiae* strains*,* functioning in the opposite way to CsrRS in *S. pyogenes*. The genes targeted by CsrRS in *S. equi* are not yet known, but these data suggest that CsrRS does not repress virulence factors, such as capsule, which confer a fitness advantage against immune cells present in whole equine blood.

IS*S1* mutants in another potential capsule regulator, *ccpA* (catabolite control protein A), were significantly reduced in fitness in whole equine blood (log_2FC= -4.3, *q*= 0.011), reflecting the importance of *has* genes in the *S. equi* whole equine blood TraDIS screen. CcpA is a major regulator that is employed in metabolic adaption to different carbohydrate sources, so it is unsurprising that *ccpA* mutants would be decreased in fitness through its potential inability to adapt to the different energy sources in whole equine blood compared to THB. In *S. suis*, CcpA regulates many genes, primarily targeting those involved in carbohydrate metabolism and amino acid transporters, such as PTS uptake systems [184, 185] and so is important for maintaining virulence. In *S. equi*, 2 PTS genes specific for mannose import were identified as important for survival in whole equine blood (*SEQ0492;* log_2FC= -3.3, *q*= 0.042, *SEQ0494*; log_2FC= -3.7, *q*= 0.017), reflecting the results seen in *S. suis*. In addition, *S. equi* IS*S1* mutants in lactate dehydrogenase (*ldh*) were reduced in fitness in whole equine blood (log_2FC= -5.1, *q*= <0.0005) and in H_2O_2 (log_2FC= -4.5, *q*= 0.0015). Ldh was found to be regulated by CcpA in *S. suis* [184, 185].

In the *S. equi* whole equine blood TraDIS screen, 2 genes, *pptAB* (a.k.a *ecsAB*)*,* were required for survival (*pptA;* log_2FC= -3.4, *q*= <0.021, *pptB*; log_2FC= -2.8, *q*= 0.002). The *pptAB* genes have previously been implicated in Gram positive bacterial virulence and encode ABC transporter proteins that export the quorum sensing peptides, SHP2 and SHP3, into the extracellular environment [186]. In agreement with this finding, a *pptAB* deletion mutant of *Staphylococcus aureus* (*S. aureus*) behaved comparably to the wild-

type strain *in vitro* in rich medium, but in a murine model of arthritis, caused milder synovitis and reduced bone erosions [187]. The Δ*pptAB* strain was also significantly reduced in its ability to persist in the kidneys in later stages of infection [187]. The identification of *pptAB* in whole equine blood and not in H_2O_2, is likely due to the presence of immune cells in the blood and therefore greater pressure to resist immune attack. These 2 genes are explored further in Chapter 4 in the context of their *in vivo* essentiality.

The GTPase encoded by *bipA* (a.k.a *thyA*) regulates cell surface and virulence associated genes in enteropathogenic *E. coli* (EPEC) [188]. In EPECs, *bipA* regulates the expression of pathogenicity islands, resistance to antimicrobial peptides and capsule synthesis [188]. BipA is not well characterised in other bacteria, but in *P. aeruginosa,* Δ*bipA* mutants were more susceptible to killing by phagocytic amoebae and human macrophages. The mutant was also reduced in its ability to attach to surfaces, form biofilms and resist antibiotics [189]. *S. equi* IS*S1* mutants in *bipA* were reduced in fitness in whole equine blood, but not in H_2O_2 (whole equine blood; log_2FC= -4.5, q= 0.007, H_2O_2; log_2FC= -3.8, q= 0.15), which correlates with BipA's potential impact on capsule synthesis and the reduced fitness of IS*S1* mutants in capsule genes in whole equine blood only.

The transcriptional regulator, CtsR, was uniquely identified as important for survival in H_2O_2 (log_2FC= -3.9, q= 0.0221), missing significance in whole equine blood (log_2FC= -2.2, q=0.5). CtsR was identified as a negative regulator of the genes encoding the heat shock proteins ClpE, ClpP, ClpL, ClpC, GroE, GroS and GroL in *S. pneumoniae* [190]. These heat shock proteins function to maintain proper protein confirmation under cellular stress and so CtsR must be prevented from binding to DNA in these stress conditions [191-193]. In *B. subtilis*, it appears that CtsR is an intrinsic heat sensor that is inactivated upon detection of increased temperatures. A *ctsR* mutant, in *B. subtilis*, mutated specifically at amino acid 64, did not react to temperature increases and therefore, could not be heat-inactivated. The inability of the mutant to sense temperature meant that DNA binding of CtsR to the heat shock target genes continued, which is detrimental to the survival of cells under stress conditions. The sensitivity of *S. equi* IS*S1* mutants in *ctsR* suggests that this regulator may also react to other stresses, such as oxidative stress, to manage the transcription of these heat shock genes. However, after closer inspection, very few IS*S1* mutants existed in the 3 input libraries and were represented by few reads; in library AC, 2 mutants were represented by 14 reads, CT, 4 mutants represented by 17 reads and GA, 2 mutants represented by 22 reads (Figure 3.18). In support of this potential false positive, a Δ*ctsR* mutant in *Lactobacillus plantarum* was not significantly more susceptible to H_2O_2 than the wild-type parental strain [194].

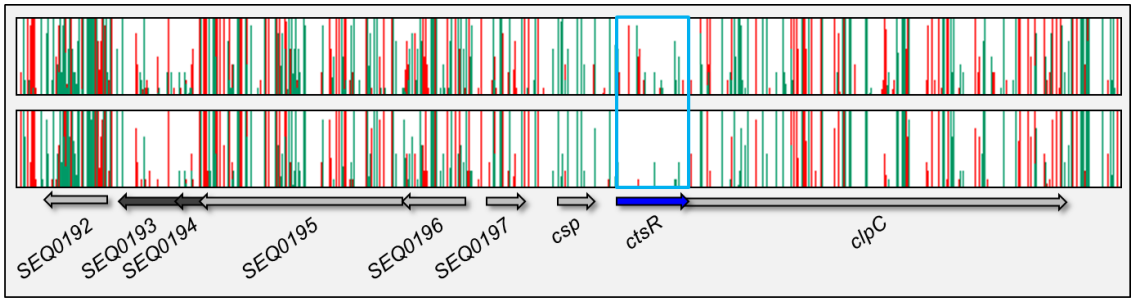


Figure 3.18. Prevalence of *S. equi* IS*S1* mutants in the genes *SEQ0192-SEQ0201* pre- and post-exposure to H_2O_2. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. The transcriptional regulator gene, *ctsR*, is required for survival in H_2O_2 by TraDIS, which is evident from a loss of sequencing reads within the recovered output population (blue arrow). Few mutants are evident in the input population and *ctsR* may represent a false positive result. Genes are included in the analysis if they are represented by more than 10 reads. *ctsR* contained 14, 17 and 22 reads in each of the input libraries. Light grey arrows indicate non-essential genes. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

3.4.3 Known virulence determinants not identified as fitness genes

Other than the capsule synthesis genes, according to TraDIS, no known virulence determinants were required by *S. equi* in whole equine blood and H_2O_2. The known immune defense factors in *S. equi* such as IdeE, IdeE2, superoxide dismutase and Se18.9 are likely all secreted, which may confound their identification by TraDIS. It is possible that IS*S1* mutants in these genes are able to benefit from neighbouring mutants that have retained the ability to produce these factors. To confirm this theory, incubation of a deletion mutant lacking 1 of these factors in whole equine blood and H_2O_2 should exhibit attenuation, despite its lack of identification by TraDIS.

It would be expected that SeM IS*S1* mutants would be attenuated in whole equine blood or H_2O_2, since a SeM deficient mutant which expressed only 4 percent of the SeM of its wildtype counterpart, was severely attenuated in equine blood, compared to the parental strain [16]. SeM was in fact removed from the analysis since it was identified as an essential gene in THB, with its insertion index equating to 0.03, just below the threshold for essentiality. None of the 4 fibronectin proteins were identified, in either condition, which suggests that they could be functionally redundant in this system, possibly complementing one another when 1 is non-functional.

The whole equine blood TraDIS data was compared to homologous genes identified in *S. pyogenes* as required for survival in human blood [82], but no similar genes were

identified. The lack of matches may in part be due to whole equine blood not being a typical niche for *S. equi* and interspecies differences in gene requirement.

3.5 Conclusions

Whilst the TraDIS screens have identified some interesting and novel genes, with many of them commonly identified between conditions, further investigation into genes additional to those selected for validation, is required to identify potential vaccine targets. Two of the genes included in the validation panel, *recG* and *addA* were attenuated in both conditions, but exhibited a slow growth phenotype compared to the wild-type parental strain, limiting their usefulness as future vaccine targets. The capsule mutant, *hasA*, confirmed the TraDIS screen results, but the reduced fitness of *mnmE* mutants in the TraDIS screen was only recapitulated in H_2O_2, with the *pyrP* mutant not attenuated in either condition. These results suggest that the presence of other mutants in the TraDIS screen, providing a competitive environment, can affect the results.

The incubation of transposon libraries in *in vitro* conditions is useful for the initial identification of target genes for further investigation, but does not necessarily relate to gene importance *in vivo*. Challenge of experimental animals with transposon libraries could potentially identify every gene required for infection providing a plethora of results to more accurately guide future vaccine design.

4 Genes required for the virulence of *S. equi* in the natural host by barcoded TraDIS

4.1 Introduction

The application of transposon libraries of *Salmonella* Typhimurium (*S*. Typhimurium) identified genes that play important roles during the infection of mice, chickens, pigs and cattle, albeit using relatively small mutant pools [83, 84]. However, the study of streptococcal pathogens has to date been restricted to *ex vivo* [81, 195] or rodent infection models [87, 131]. One recent study, describes the infection of non-human primates (NHP) with dense IS*S1 S. pyogenes* libraries as a model of necrotising myositis [145]. Whilst these studies have provided unparalleled insights into the genes that might be important to the disease-causing abilities of streptococcal pathogens, they do not necessarily reflect their importance for the infection of natural host species. The *S. pyogenes* NHP study did however measure the transcription of *S. pyogenes* genes in wild-type infected NHP muscle and in a human case of *S. pyogenes* infection and found that 6 genes which were identified by TraDIS as fitness genes and validated with whole deletion mutants, were transcribed in both samples.

Strangles presents as an ideal model of streptococcal disease as *in vivo* studies in the natural host are possible, with the infection process likely to reflect invasive disease caused by other streptococci. As previously described, the essential genome of *S. equi* is 83 percent similar to that of *S. pyogenes* and *S. agalactiae*, highlighting the close genetic relationships between these important pathogenic bacteria [77, 78, 103] and supporting the potential similarity between genes required for infection by *S. equi* to other streptococci.

In this Chapter, the results of a genome-wide barcoded TraDIS study of *S. equi*, conducted in a susceptible natural host are presented. Three barcoded IS*S1* libraries were used to co-infect 12 ponies. Each animal received 2 of the 3 barcoded libraries, reducing the total number of animals required, since each library can be deconvoluted back to its parental population after mutant recovery. The data was initially analysed taking each animal as a biological replicate, as is currently practiced in similar transposon library studies. The data was re-analysed considering the novel library barcodes. The data from the 2 analysis methods was compared and used to make a comprehensive measurement of genome-wide fitness of *S. equi* genes *in vivo*. Twelve genes identified as reduced in fitness as a result of IS*S1* insertion were selected for validation. Tagged whole deletion mutants were generated by allelic replacement mutagenesis and 5 ponies challenged with a mixture of these mutants alongside wild-type *Se*4047 and internal control deletion strains. The presence of the tag enabled next-generation sequencing of the challenge material and surviving mutants recovered from ponies.

Fitness genes identified in *S. equi in vivo* were compared to *S. pyogenes* fitness genes identified in a Tn-seq murine model of subcutaneous serotype M1 infection, and a TraDIS NHP model of necrotising myositis, that utilised both M1 and M28 strains [87, 145]. A consensus list of genes required for fitness *in vivo* were identified. Genes conferring an increased fitness upon insertion were also compared between studies. Identifying this set of pan-species fitness genes has uncovered unprecedented information for the design of novel therapeutics and vaccines against strangles and other streptococcal disease.

4.2 Materials and methods

4.2.1 Bacterial growth, DNA isolation and primers

The *S. equi* strain *Se*4047 was used throughout this study. All libraries and deletion strains were grown at 37 ºC in a humidified atmosphere containing 5 percent CO_2. Genomic DNA was extracted using GenElute spin columns (Sigma) according to manufacturer's instructions, except that lysis reactions were incubated for 1 hour. All primers used in this Chapter are described in Table A1.4 (Appendix 1).

4.2.2 *In vivo* challenge of ponies with barcoded IS*S1 S. equi* libraries

The barcoded IS*S1* mutant libraries of *S. equi*; AC, CT and GA, described in Chapter 2 were recovered in Todd-Hewitt broth containing 0.5 µg/ml erythromycin (THBE) and 10 percent fetal calf serum until OD_{600nm} 0.3 was reached (approx. 2x 10^8 CFU/ml). Challenge doses of 2.5 ml were aliquoted immediately from the OD_{600nm} 0.3 cultures (5 x10^8 CFU/ dose) along with 5 ml aliquots, which were stored in 25 percent glycerol at -80 ºC for processing with TraDIS as input pools. Twelve Welsh mountain ponies of approximately 1 year old were challenged intranasally with 2 libraries, 1 dose of 1 randomly assigned barcoded library per nostril, such that each library was administered to 8 animals, at a total dose of 1 x10^9 CFU/ml per animal (Table 4.1).

Table 4.1. Challenge pattern employed for the inoculation of Welsh mountain ponies with barcoded *S. equi* IS*S1* libraries. Each pony received 1 barcoded library per nostril in the pattern described, such that each library was tested 8 times over the 12 animals.

	Left nostril	Right nostril
Pony	Barcoded library	
477	GA	CT
2991	CT	AC
5867	AC	CT
5922	CT	GA
6061	CT	AC
6544	AC	GA
7454	AC	GA
7565	GA	CT
7616	AC	CT
7649	CT	GA
7799	GA	AC
7884	GA	AC

One spare challenge dose of each library was diluted 1 x10^4 fold and spread onto CNA to enumerate the dose pre-challenge, representing the minimum dose administered to the ponies. An additional spare dose was enumerated post-challenge to calculate the maximum possible dose received by ponies. This study was conducted under the auspices of a Home Office Project License and following ethical review and approval by the Animal Health Trust's Animal Welfare and Ethical Review Body (RPP 01_12).

Recovery of abscess material from ponies

Ponies were euthanised upon developing early clinical signs of disease; pyrexia and preference of haylage over dry pelleted food. All retropharyngeal (left and right (LRP, RRP)) and submandibular (left and right (LSM, RSM)) lymph nodes (bilateral, 1 on each side of the head) were recovered from all animals post-mortem. Abscess material was immediately recovered from 24 retropharyngeal and 14 submandibular diseased lymph nodes by sectioning the nodes and manually collecting the abscess material. For lymph nodes containing less than 1 ml, abscess material was collected before 5-10 pieces of tissue at approximately 1 cm^3 were each macerated in 1 ml PBS using a Qiagen tissue lyser for 15 minutes at 60 Hz. The tissue lysate was added to any recovered abscess material. Abscess material was stored in 25 percent glycerol at -80 °C.

Fifty µl of each lysate/abscess material was diluted in PBS to varying concentrations (between 1: 1x 10^1 and 1x 10^6), depending on the amount of abscess material recovered, and plated on THA and incubated overnight in triplicate to determine CFU/ml. Ten lymph nodes were not sufficiently diseased; 2 lymph nodes contained < 5 x10^3 CFU/ml and 8 contained no *S. equi*. The bacterial loads of retropharyngeal versus submandibular lymph nodes were statistically compared using a two-tailed Mann-Whitney U test.

Extraction directly from abscess material

Bacterial DNA extraction was initially attempted directly from abscess material. DNA was extracted from abscess material recovered from 1 lymph node (pony 5922, LRP), using GenElute spin columns (Sigma) as previously described in 4.2.1. DNA was prepared for sequencing as previously described in Chapter 2 section 2.2.5. As only 1 DNA library was sequenced, unique indexing was not utilised. A random indexing primer was chosen (AHT 6) and used during the PCR step of library preparation; however, the custom indexing primer was not loaded into the MiSeq cartridge for sequencing. The generated FASTQ file was analysed by identifying the number of reads containing IS*S1*. The FASTX barcode splitter (http://hannonlab.cshl.edu/ fastx_toolkit/) was used, inputting the FASTQ file and a text file containing the last 21 bp of IS*S1* (GAAAACTTTGCAACAGAACC, omitting the first 2 bp (barcode) of reads to capture any reads containing IS*S1*).

One hundred ng of the same library prepared DNA was re-amplified using a nested PCR method which involved initially amplifying with primers binding 49 bp upstream of the custom IS*S1* primer (nested IS*S1*) and to the 3' adaptor (nested adaptor). The same PCR conditions as used in the standard library preparation PCR were applied (Chapter 2 section 2.2.5). DNA was cleaned up using the Monarch® PCR & DNA cleanup kit (NEB), before 100 ng of DNA was amplified using the standard library preparation PCR using the custom IS*S1* primer and indexing primer as previously described (Chapter 2 section 2.2.5). DNA was sequenced and data analysed using bacteria_tradis and tradis_gene_insert_sites as previously described in Chapter 2 section 2.2.5.

In vitro recovery of IS*S1* mutants

IS*S1* mutants were initially recovered from 1 lymph node (pony 5922, LRP) on 100 large 150 mm THA Petri dishes supplemented with 0.5 µg/ml erythromycin and 0.03 µg/ml of hyaluronidase at 5x 10^5 CFU/ plate (200 µl/ plate), in batches of 10. Petri dishes were incubated overnight and mutants recovered by washing the dishes with THB containing 25 percent glycerol for storage at -80 ºC. Recovering mutants in batches was completed to determine the number of dishes required to accurately capture the mutant population.

DNA was extracted from 2 mls of recovered *S. equi* from each batch and processed by TraDIS, uniquely indexing each batch as previously described in Chapter 2. DNA was sequenced on 2 MiSeq runs (5 batches per run) as described in Chapter 2. Generated FASTQ files were individually analysed using the TraDIS toolkit scripts [111] as previously described to produce a readable document of insertion sites in each batch. FASTQ files from each batch were progressively combined to determine the point at which new mutant discovery plateaued. To achieve this, FASTQ files from each batch were combined using the example script below, which combines 3 sequencing files.

```
cat batch1.fastq batch2.fastq batch3.fastq > combined.fastq
```

The data was re-analysed using the TraDIS toolkit scripts [111] as previously described, splitting the data each time by barcode when using bacteria_tradis to identify the contribution of mutants by each library administered to different nostrils of pony 5922; CT and GA. No gene inclusion criteria were imposed on the data in this analysis.

It was concluded that recovering mutants on 30 large Petri dishes was most feasible taking into account the manual work required and the number of new mutants identified (data shown in section 4.3.2). Therefore, all abscess materials were recovered on 30 150 mm Petri dishes as described for 5922 LRP. Any abscess material not dense enough to plate at this concentration were spread neat and/or on as many Petri dishes as

recovered volume would allow. Colonies were washed off with THB containing 25 percent glycerol for storage at -80 ºC.

4.2.3 Sequencing of barcoded TraDIS libraries recovered on plates

DNA was extracted from stored abscess material/tissue lysate (Table 4.2) depending on their bacterial loads, for samples containing > 5x 10^3 CFU/ml (data shown in section 4.3.2). DNA was extracted from the stored 5 ml aliquots of the 3 input libraries.

Table 4.2. Volumes of stored abscess material from which DNA was extracted for sequencing by TraDIS. Abscess material was recovered from 12 ponies experimentally challenged with barcoded IS*S1* mutant libraries and stored in 25 percent glycerol before extraction.

	Pony	Volume of stored abscess extracted from
LRP	477	500 µl
	2991	500 µl
	5867	500 µl
	5922	500 µl
	6061	500 µl
	6544	500 µl
	7454	500 µl
	7565	500 µl
	7616	500 µl
	7649	500 µl
	7799	500 µl
	7884	500 µl
RRP	477	500 µl
	2991	500 µl
	5867	500 µl
	5922	500 µl
	6061	500 µl
	6544	500 µl
	7454	500 µl
	7565	500 µl
	7616	500 µl
	7649	500 µl
	7799	500 µl
	7884	500 µl
LSM	477	2 ml
	5867	2 ml
	5922	500 µl
	6544	500 µl
	7565	4 ml
	7616	1 ml
	7649	2 ml
	7884	4 ml
RSM	477	500 µl
	2991	4 ml
	5867	500 µl
	5922	500 µl
	6544	500 µl
	7799	4 ml

DNA was prepared for sequencing as described in Chapter 2 section 2.2.5. DNA libraries from abscess material containing > 2x 10^5 CFU/ml, and input pools, were uniquely indexed, diluted to 1.6 nM and combined for sequencing in varying concentrations to account for population diversity. Input libraries each contributed ~8.5 percent of the sequencing mix (26 percent total), DNA recovered from retropharyngeal abscesses each

contributed ~3 percent (64 percent total) with DNA recovered from submandibular abscesses contributing ~1 percent per node (10 percent total). Submandibular abscess material contained, on average, lower bacterial loads, which was used as a predictor for lower diversity. Due to the already homogenous nature of transposon-genome reads, submandibular abscesses were restricted in the sequencing mix and the whole mix combined with 40 percent PhiX DNA (Illumina) to maximise the final load heterogeneity and sequencing success.

The final load library was denatured, neutralised as previously described in Chapter 2 section 2.2.5, diluting the libraries to 8 pM, and sequenced on 5 HiSeq2500 Rapid Runs. One run was completed initially, from which total read counts per indexed DNA library was calculated. DNA libraries were adjusted to account for the read counts obtained. DNA libraries from abscess material of < 5x 10^5 CFU/ml were uniquely indexed and sequenced, in equal concentrations, on 1 MiSeq run as previously described in Chapter 2.

4.2.4 Per animal and barcoded analysis of recovered IS*S1* mutants

Input library analysis

The 3 FASTQ files generated from the 3 input libraries were combined using the FASTQ combination script described in section 4.2.2 before they were split according to barcode and reads mapped to the *S. equi* genome using the bacteria_tradis script [111] as described in Chapter 2. The script tradis_gene_insert_sites was used to produce readable documents of insertion sites for the 3 files, using the '-trim3 0.1' argument to discount reads mapping to in the last 10 percent of genes, as these insertions are assumed to have little or no effect on the transcribed product. To improve the robustness of the analysis and to minimise the effect of stochastic loss, certain filter thresholds were imposed on the input data. 575 genes, identified as essential, ambiguous or not defined, as determined in Chapter 2, were removed from the analysis. Three genes that were overrepresented in any 1 of the 3 input pools due to the prevalence of a few specific IS*S1* mutants were also removed (*SEQ0285*, *SEQ0882*, and *SEQ1147*). Genes were considered overrepresented when their read counts contributed over 2 percent of the total reads for the library. Read counts per gene were normalised between the input libraries to facilitate data comparison. Two-hundred and twenty-eight genes for which < 1,000 reads mapped to, in any 1 of the 3 normalised input libraries, were removed to ensure each gene was sufficiently represented to minimise the effects of stochastic loss. Enforcing these criteria permitted the inclusion of 1,359 genes in the following analysis.

Per animal analysis

Output data was initially processed on a per animal (PA) basis, as is currently practiced in similar *in vivo* transposon library studies, ignoring the library barcodes (Figure 4.1).

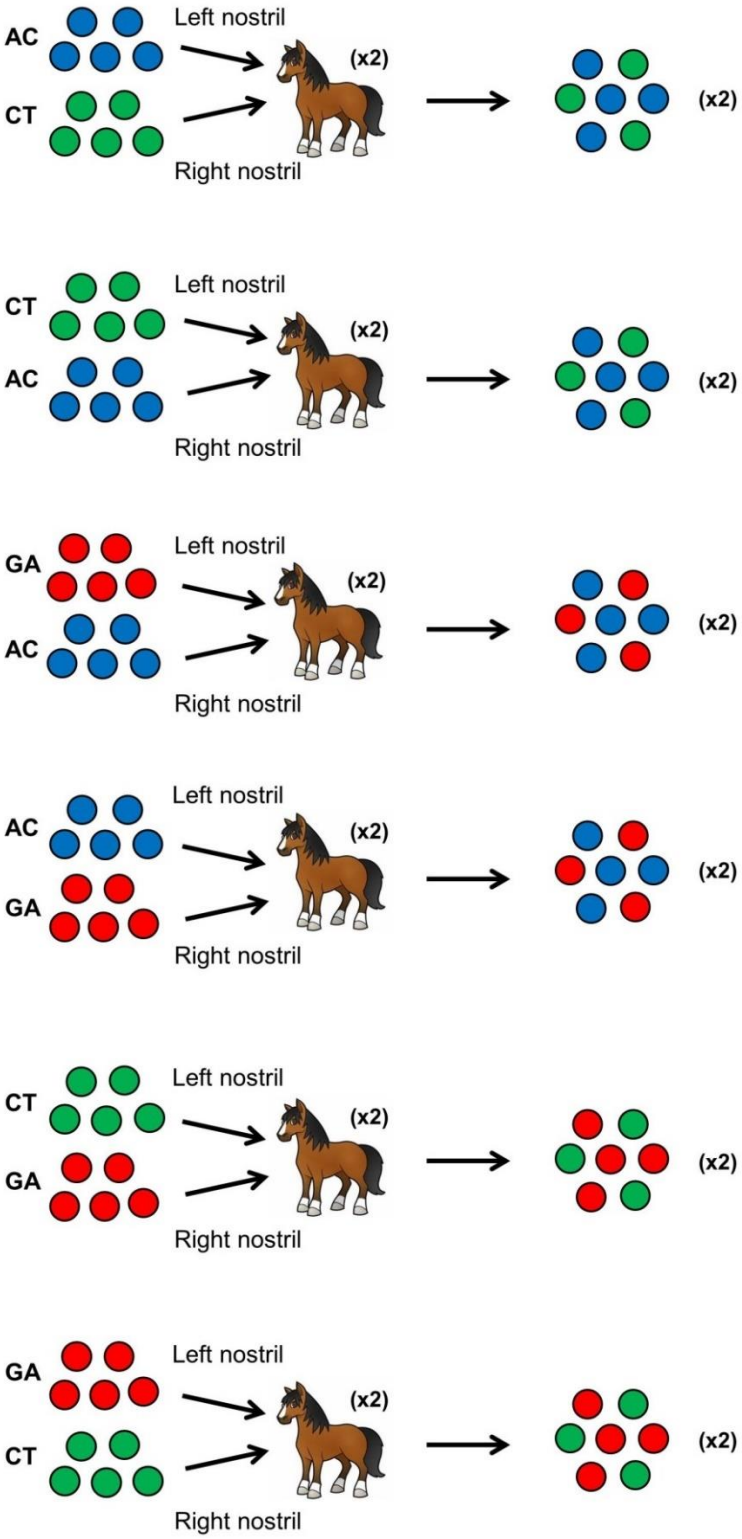


Figure 4.1. Schematic representation of per animal analysis of barcoded *S. equi* IS*S1* mutants able to cause disease in 12 Welsh mountain ponies. Traditionally in *in vivo* mutant library studies, mutants recovered from each animal are treated as single replicates, meaning any bottleneck effect and animal to animal variation can affect the quality of data obtained.

For the PA analysis, FASTQ files of DNA sequenced from each of the 38 infected lymph nodes were combined according to the animal from which they were recovered, using the FASTQ combination script described in section 4.2.2. Each of the 12 new FASTQ files (1 per animal) were next trimmed of the first 2 bp (barcode) using the FASTX-trimmer (http://hannonlab.cshl.edu/ fastx_toolkit/) to standardise the reads. All 12 files were processed using the TraDIS toolkit scripts [111]. The sequence GAAAACTTTGCAACAGAACC (sequenced end of IS*S1* without barcode) was used in bacteria_tradis to isolate and map transposon-genome junction reads for the 12 fastq files. The script tradis_gene_insert_sites was run on each file, again using the '-trim3 0.1' argument to discount reads mapping to in the last 10 percent of genes. Genes discounted from the input analysis as described above were similarly removed from these 12 PA files, and read counts per gene normalised between all 12 files to facilitate more accurate comparison.

The 12 PA output files were compared to the 3 input files using the tradis_comparison script [111] to generate a document containing fitness values (log_2 fold change (FC)) and statistical significance (*q* values) for the 1,359 genes meeting the inclusion criteria.

Barcoded analysis

For the BC analysis, all FASTQ files of sequenced abscess material were combined, using the FASTQ combination script as previously described, generating 1 file. As completed for the analysis of input libraries, the bacteria_tradis script was run 3 times, each using the appropriate barcode, e.g ACGAAAACTTTGCAACAGAACC for library AC, to generate 3 mapped files (Figure 4.2).

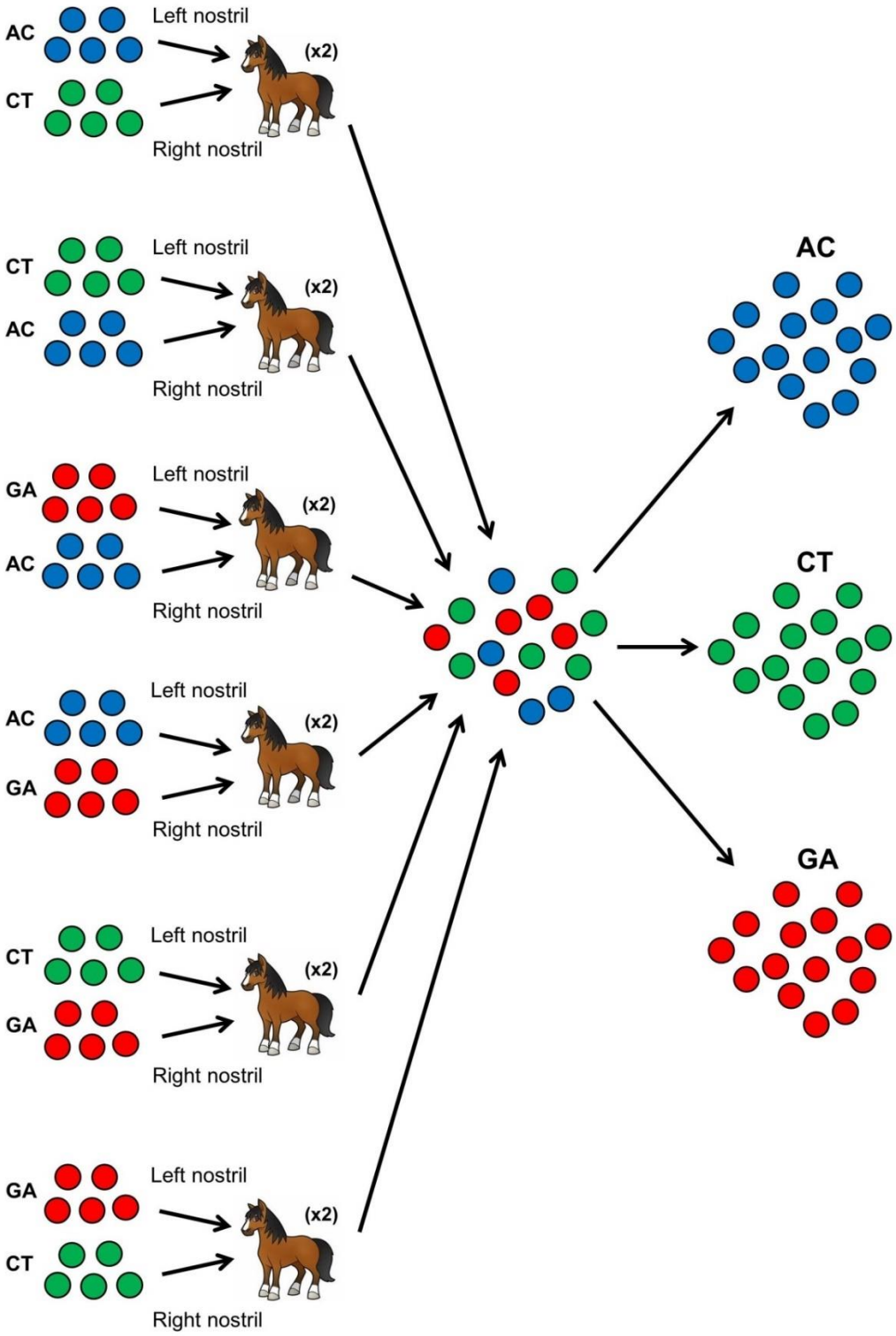


Figure 4.2. Schematic representation of barcoded analysis of barcoded *S. equi* IS*S1* mutants able to cause disease in 12 Welsh mountain ponies. Administration of 2 barcoded libraries to each pony, pooling of all surviving mutants and deconvolution according to parental library barcode, reduces the effect of a bottleneck and animal to animal variation of the data.

The script tradis_gene_insert_sites was used to produce readable documents of insertion sites for each of the 3 BC files, also using the '-trim3 0.1' argument. Genes

discounted from the input analysis as described above were similarly removed from these 3 BC files and read counts per gene were normalised between the 3 files.

The tradis_comparison script [111], was again run similarly to the PA analysis method, but comparing the 3 BC files to the 3 input files. The BC analysis identified 46 genes for which fitness values were confounded by an extensive increase in the number of reads for 1 unique mutant, in only 1 library recovered from ponies. These 46 genes were removed from the 3 input and 3 BC files and re-analysed, therefore reducing the number of genes analysed by the BC method to 1,311.

In both the PA and BC analysis, mutants in genes with a calculated $\log_2$FC value of < -2 and a *q* value of < 0.05 were deemed as significantly reduced in fitness due to IS*S1* insertion. Genes were considered as conferring an increased in fitness upon insertion if they had a $\log_2$FC value of > 2 and a *q* value of <0.05. For both the PA and BC data, cluster of orthologous genes (COG) analysis of genes required for fitness and of genes conferring a fitness advantage as a result of insertion, was conducted using the online tool, Integrated Microbial Genomes and Microbiomes (IMG/M) [196].

4.2.5 Translocation of IS*S1* mutants *in vivo*

Uniquely indexing each lymph node for sequencing also allowed the analysis on a per lymph node basis. Basic analysis to determine the rate of translocation of *S. equi* from the inoculation site on 1 side of the head, to the other side was performed by running the bacteria_tradis script twice on the same per lymph node FASTQ file, each using 1 of the 2 barcodes relevant to the libraries received by the corresponding animal. No gene inclusion criteria were imposed on the data in this analysis.

4.2.6 Validation of TraDIS *in vivo* results

Deletion mutant generation

Twelve attenuated genes were identified for validation *in vivo*, 7 of which were identified in both the BC and PA analysis, and a further 5 uniquely identified using the BC analysis technique (Table 4.3).

Table 4.3. Twelve *S. equi* genes selected for validation *in vivo* with tagged whole deletion mutants. The selected gene name, locus tag, function and tag used are described. Three tagged deletion mutants in an additional control gene, that was non-essential *in vivo*, *SEQ0751*, were generated (green row). Purple rows represent those identified similarly by the per animal and barcoded analysis. White rows indicate genes that were uniquely identified by the barcoded analysis.

Gene	Locus tag	Function	Tag
purN	SEQ0029	phosphoribosylglycinamide formyltransferase	B
SEQ0402	SEQ0402	putative cell surface-anchored protein	A
scfA	SEQ1551	putative permease	A
metP	SEQ1899	putative D-methionine transport system permease	B
sufC	SEQ1930	putative ABC transporter, ATP-binding protein	B
slaB	SEQ2155	putative exported protein	C
gacI	SEQ0969	putative glycosyl transferase	C
recG	SEQ0454	ATP-dependent DNA helicase	B
sptA	SEQ1312	putative exported protein	B
SEQ1410	SEQ1410	Branched-chain amino acid ABC transporter ATP-binding protein	C
dltB	SEQ1452	putative activated D-alanine transport protein	C
SEQ1536	SEQ1536	putative exported protein	A
SEQ0751	SEQ0751	Putative DNA-binding protein	A/B/C

Allelic replacement mutagenesis was performed, as previously described in Chapter 3 with some modifications, generating whole deletion tagged mutants in all genes selected, except for in *slaB*, as a construct for an internal deletion already existed in the AHT collection. One of 3 tags were introduced into the deletion mutants to enable next generation sequencing of the strains. Tag A is comprised of the last 3' 80 bp of IS*S1* starting from the IS*S1* sequencing primer binding site utilised in TraDIS (Table 4.4). The remaining 2 tags (B and C) contained the same IS*S1* sequencing primer binding site, but were instead followed by random sequence, to maximise diversity and successful sequencing (Table 4.4).

Table 4.4. Sequence of the tags inserted into allelic replacement deletion mutants to enable sequencing by TraDIS. Black nucleotides indicate the sequencing primer binding site, which is common between all 3 tags. Red nucleotides indicate the variable region of the tag.

Tag	Sequence 5'-3'
A	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTTTACACTAAAATAGACTTAT CAGAAAACTTTGCAACAGAACCC
B	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTTTACACTAAAATAGACTTAT GTTGACCCTATTGCAACTTGGAT
C	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTTTACACTAAAATAGACTTAT ACGTCTTCGAGTAATCTATCGTG

A further gene not effected by IS*S1 in vivo* was selected for the generation of 3 control strains (*SEQ0751*, herein referred to as internal control (IC)), each containing 1 of the 3 tags, to assess whether the tags influence the mutants *in vivo* (Table 4.3). The 3 IC strains, (ΔICtagA, ΔICtagB and ΔICtagC), Δ*slaB*tagC, Δ*recG*tagB, Δ*SEQ1410*tagC and Δ*dltB*tagC were generated by the author, with the remaining 8 strains generated by other

members of the Animal Health Trust bacteriology group using the allelic replacement method described in Chapter 3 section 3.2.3.

Briefly, 500 bp regions of *S. equi* DNA flanking the target gene were amplified, ligated together and cloned into the pGh9 plasmid [98]. The 3 80 bp tags (Table 4.4) were generated by annealing 2 primers together (Table A1.4, Appendix 1) at 95 ºC for 2 minutes. Each deletion construct was digested using the appropriate restriction enzyme for that used to ligate the 2 500 bp flanks together, according to manufacturer's instructions. One of the 3 tags were next blunt ligated into the digested construct at a ratio of 5:1 tag to construct using T4 DNA ligase according to manufacturer's instructions. Ligation reactions were transformed into *E. coli repA*+ cells as previously described in Chapter 3 section 3.2.3. Colonies were PCR screened as previously described in Chapter 3 section 3.2.3 using the corresponding P1 and P4 primers. Products containing the tag were 80 bp larger in size than those missing the tag. Colonies containing successfully tagged constructs were grown overnight, constructs extracted from the cultures and sequenced using the P1 and P4 primers as previously described in Chapter 3 section 3.2.3.

Correctly tagged constructs were used to generate the allelic replacement mutants. Briefly, constructs were individually transformed into competent *Se*4047 cells, grown on THAE at 28 ºC (plasmid permissive temperature), single colonies inoculated into THBE overnight at 28 ºC then transferred to 37 ºC for 3 hours to induce chromosomal integration of the construct. Integrants were selected on THAE overnight at 37 ºC. Integrants were grown overnight at 37 ºC in THBE, followed by dilution into TH broth and incubation at 28 ºC for 48 hours to excise pGh9 from the chromosome, but retaining the flanks containing the tag. Excised bacteria were spread on TH agar and grown overnight at 37 ºC to ensure free plasmid was lost. To confirm plasmid loss, and therefore loss of erythromycin resistance, deletion strains were spread on both THAE and THA. Mutant alleles were confirmed by PCR using the appropriate P1 and P4 primers (suffixed with gene name in Table A1.4, Appendix 1) and sequencing on an ABI3100 DNA sequencer using BigDye fluorescent terminators. deletion strains were stored in 25 percent glycerol at -80ºC.

In vivo challenge of ponies with validation mutants

Gycerols of all 15 deletion strains and *Se*4047 were streaked onto CNA Petri dishes and grown for 16 hours. Single colonies of each strain were individually grown in 10 ml THB containing 10 percent fetal calf serum (THBFCS) and grown for 16 hours. Cultures were diluted in THBFCS to OD_{600nm} 0.08 and grown until OD_{600nm} 0.3 was reached. The challenge inoculum was prepared by combining 0.66 ml of each of the 12 deletion strains

predicted to be attenuated *in vivo* with 6 ml of each tagged IC strain and 18 ml of *Se*4047. These volumes were chosen to reflect the proportion of genes required for fitness in the TraDIS screen. Challenge doses of 2.5 ml were aliquoted immediately from the inoculum (5 x10^8 CFU/ dose) along with a 5 ml aliquot, which was stored in 25 percent glycerol at -80 ºC for processing as the input pool. One spare challenge dose was diluted 1 x10^4 fold and spread onto CNA to enumerate the material pre-challenge, representing the minimum dose administered to the ponies. An additional spare dose was enumerated post-challenge to calculate the maximum possible dose received by ponies.

Five 2.5-year-old Welsh mountain ponies were challenged intranasally with the same mutant and WT *Se*4047 pool, at a total dose of 1 x10^9 CFU/ml per animal, reflecting the dosage used in the TraDIS study. This study was conducted under the auspices of a Home Office Project License and following ethical review and approval by the Animal Health Trust's Animal Welfare and Ethical Review Body (RPP 01_12). Two ponies were euthanised upon developing early clinical signs of disease on day 6 post-challenge; pyrexia and preference of haylage over dry pelleted food. However, the remaining 3 animals did not develop obvious clinical signs and were euthanised 10 days after challenge

At post-mortem, all retropharyngeal and submandibular lymph nodes were removed from the 5 ponies. Abscess material was recovered by sectioning the nodes and manually collecting the abscess material. For any lymph nodes that contained less than 1 ml abscess material, 5-10 pieces of tissue at approximately 1 cm^3 were each macerated in 1 ml of PBS, in a Qiagen tissue lyser at 60 Hz for 15 minutes to recover any surviving *S. equi*. All abscess material and lysed tissue was stored in 25 percent glycerol at -80 ºC.

Where > 1 ml abscess material was recovered, samples were enumerated by plating 10-fold dilutions up to 1 in 1x 10^5 on CNA Petri dishes. The CFU/ml calculated for each sample was used to recover *S. equi* on 5 large 150 mm THA Petri dishes at a density of 5x 10^5 CFU/plate with 200 µl spread onto each plate. Any abscess material/tissue lysate not dense enough to plate at this concentration were spread neat on as many Petri dishes as the recovered volume would allow, where 2 ml was spread onto each plate. Colonies were washed off with THB containing 25 percent glycerol for storage at -80 ºC. The bacterial loads of retropharyngeal versus submandibular lymph nodes were statistically compared within the validation study and to that from the TraDIS screen using a two-tailed Mann-Whitney U test.

Sequencing of recovered *S. equi*

DNA was extracted from varied volumes of stored abscess material/tissue lysate depending on their bacterial loads (Table 4.5).

Table 4.5. Volumes of stored abscess material from which DNA was extracted for sequencing by TraDIS. Abscess material was recovered from 5 ponies experimentally challenged with tagged deletion mutants and stored in 25 percent glycerol before extraction.

	Pony	Volume of glycerol extracted
LRP	4000	2 ml
	5822	1 ml
	7822	500 µl
	9219	500 µl
	9757	500 µl
RRP	4000	2 ml
	5822	1 ml
	7822	500 µl
	9219	500 µl
	9757	500 µl
LSM	4000	2 ml
	5822	2 ml
	7822	2 ml
	9219	2 ml
	9757	500 µl
RSM	4000	2 ml
	5822	2 ml
	7822	2 ml
	9219	2 ml
	9757	1 ml

DNA was extracted from the stored 5 ml aliquot of the input library. Recovered mutants were sequenced by TraDIS as previously described, except that DNA was fragmented using the NEBNext Ultra II FS DNA module (E7810), which additionally end repairs and A-tails DNA. Five-hundred µg of DNA was incubated with the fragmentation enzyme for 7.5 minutes according to manufacturer's instructions, generating fragments of approximately 600 bp ready for adaptor ligation as previously described. Additionally, the plasmid depletion step was not performed as no pGh9 exists in the deletion mutants.

All DNA libraries (1 input and 20 output) were indexed, diluted to 2 nM and combined in equal proportions before sequencing. Due to the already homogenous nature of the amplified fragments in the DNA libraries, the diversity of combined DNA libraries had to be improved to ensure successful sequencing. Therefore, the DNA was combined with 80 percent PhiX (Illumina), after both had been denatured and neutralised, as previously described, before sequencing on 1 MiSeq run. An Illumina software update removed the previously utilised Truseq LT function, and so the TruSeq DNA Single Indexes (A, B) setting was instead utilised, with all other settings remaining the same.

Analysis of *in vivo* validation data

The number of reads corresponding to each attenuated mutant and the 3 IC mutants were counted in each FASTQ file. First, the command 'grep' was used to isolate any reads matching the sequence provided with the command, since the sequence downstream of the tag in each mutant is known. For example, to isolate sequencing reads corresponding to the *purN* mutant from the LRP node from pony 4000, the following command line was used:

```
grep GTTGACCCTATTGCAACTTGGATATCGGCTTCATCACCACACTAGCAACT
4000LRP.fastq > purN_4000LRP.txt
```

The red text highlights the last 23 bp of tag, the blue text highlights the digested restriction enzyme site used to ligate the tag into the construct, and the green text highlights 24 bp of the following sequence flanking the deletion target site.

The number of lines within the new purN_4000LRP.txt file were then counted using the following command line to calculate the abundance of this mutant within the lymph node.

```
wc -l purN_4000LRP.txt
```

These 2 command lines were run on each of the 21 FASTQ files, for each of the 15 mutants (12 attenuated and 3 IC mutants) using the sequences in Table 4.6 to isolate reads matching to each mutant.

The tags ligated into the constructs in both the forward and reverse directions, therefore the sequence selected for use in the grep command line either included sequence upstream or downstream of the target gene.

Table 4.6. Sequences used to bioinformatically measure tagged mutant abundance in abscess material recovered from 5 experimentally challenged ponies. Direction in which the tag was cloned is indicated for each mutant. Red nucleotides indicate the variable end of each tag, blue nucleotides indicate half of the restricted site used to clone the tag into the deletion construct and the green nucleotides indicate the flanking genome to the target gene deletion site.

Strain	Tag direction	Sequence used with grep to isolate mutant reads
Δ*purN*tagB	reverse	GTTGACCCTATTGCAACTTGGATATCGGCTTCATCACCACACTAGCAACT
Δ*SEQ0402*tagA	reverse	CAGAAAACTTTGCAACAGAACCCATCTTCTCTCTCCTTTAATGATAGACA
Δ*scfA*tagA	reverse	CAGAAAACTTTGCAACAGAACCCATCTCACCGTATCCTTTCTATATGTTA
Δ*metP*tagB	reverse	GTTGACCCTATTGCAACTTGGATAACTTAAGCCCCTCTCTTTAAAATAGT
Δ*sufC*tagB	forward	GTTGACCCTATTGCAACTTGGATATCTAAGCTGCAAGGCTGTCTAAGGCT
Δ*slaB*tagC	forward	ACGTCTTCGAGTAATCTATCGTGAGCTTGAAACGGTAGGTGCTATTGGAT
Δ*gacI*tagC	forward	ACGTCTTCGAGTAATCTATCGTGATCGAGGATGTTTATTGGGTTACAGCT
Δ*recG*tagB	reverse	GTTGACCCTATTGCAACTTGGATATCGACCCTTCAAATTAGCAATCGAAC
Δ*sptA*tagB	forward	GTTGACCCTATTGCAACTTGGATATCTAGTGCCAGATGAGAAAAAAGAAT
Δ*SEQ1410*tagC	forward	ACGTCTTCGAGTAATCTATCGTGATCAATAAATACTCTAAAAGCCATTGG
Δ*dltB*tagC	forward	ACGTCTTCGAGTAATCTATCGTGAACAAAAGGAGAGTATAAAAATGTCTA
Δ*SEQ1536*tagA	forward	CAGAAAACTTTGCAACAGAACCCATCGTAATTTTTTTAAAACGTTGGTGA
ΔICtagA	forward	CAGAAAACTTTGCAACAGAACCCATCAATTAAGTTGCAAAACAAAGATTT
ΔICtagB	reverse	GTTGACCCTATTGCAACTTGGATATCCTGTTTATTTCACCACCTTTATTT
ΔICtagC	reverse	ACGTCTTCGAGTAATCTATCGTGATCCTGTTTATTTCACCACCTTTATTT

4.2.7 Comparative analysis of the genes implicated in *in vivo* infection in *S. equi* vs *S. pyogenes in vivo* and *ex vivo*

Gene fitness data concerning *S. pyogenes in vivo/ex vivo* was retrieved from the supplementary information of 3 Tn-seq/TraDIS studies; strain M1 MGAS5448 in a subcutaneous murine model of infection [87], strains M1 MGAS2221 and M28 MGAS27961 in a non-human primate (NHP) infection model [145] and strain M1 MGAS2221 in human saliva *ex vivo* [81]. The 2 latter TraDIS studies utilised pGh9:IS*S1* transposon libraries and describe collaborative works between the AHT and the Houston Methodist Research Institute undertaken during the course of this PhD.

Genes were included in comparative analyses where homologs existed between *S. pyogenes* and *S. equi*. Homologous genes were identified in Chapter 2 and are available in Additional File 5 [103]. Any genes not meeting the inclusion criteria imposed on the *S. equi in vivo* dataset were also not considered in the comparative analysis. Genes previously identified in *S. equi* as essential or ambiguous were included where homologs were identified as required for *in vivo* fitness in the *S. pyogenes* studies.

4.3 Results

4.3.1 Composition of input libraries *in vitro*

The 3 barcoded IS*S1* libraries, designated AC, CT and GA, in *S. equi* strain *Se*4047 have been described previously in Chapter 2. Libraries were grown to an OD_{600nm} of 0.3 immediately before use *in vivo* and resequenced to accurately identify input pool composition. Each input library represented between 89.1 and 90.3 percent of the 2,165 *S. equi* genes (Table 4.7) and therefore were representative of the *S. equi* genome.

Table 4.7. Composition of libraries used to experimentally challenge 12 Welsh mountain ponies pre- and post-filtering. The number of genes containing insertions post-filtering is consistent between libraries, since filtering determines a consensus set of genes to be taken forward for analysis.

Library	Unique insertion sites in genes	Library saturation (insertion every n bp in genes)	Genes containing insertions (% of total genes : % of non-essential genes)
AC^{pre}	42,964	45	1,929 (89.1 : 100)
CT^{pre}	39,333	49	1,956 (90.3 : 100)
GA^{pre}	57,338	34	1,937 (89.5 : 100)
Combined^{pre}	134,958	14	2,017 (93.2 : 100)
AC^{post}	35,533	54	1,359 (62.7 : 85.4)
CT^{post}	32,502	60	1,359 (62.7 : 85.4)
GA^{post}	48,008	40	1,359 (62.7 : 85.4)
Combined^{post}	122,338	16	1,359 (62.7 : 85.4)

To improve the robustness of the analysis and to minimise the effect of stochastic loss, certain filter thresholds were imposed on the input data as described in section 4.2.3. As part of this criteria, a set of 228 non-essential genes that contained < 1,000 mapped reads were removed from the analysis. These genes had an average length of 380 bp, compared to 975 bp for genes passing this criterion. Therefore, the analysis of shorter genes by TraDIS may be confounded as they are less likely to be represented by sufficient numbers of IS*S1* mutants. The criteria permitted the inclusion of 1,359 genes in the analysis, which represents 85.4 percent of non-essential genes in *S. equi* (Table 4.7).

4.3.2 *In vivo* infection of the natural host with barcoded *S. equi* libraries

Twelve Welsh mountain ponies were each challenged intranasally with 2 of 3 barcoded input libraries at a dose of $5x\ 10^8$ CFU administered in 2.5 ml of THBFCS per nostril. Using this method, each individual library was administered in 4 different combinations to 8 animals. The minimum (pre-challenge) and maximum (post-challenge) doses the animals received per nostril are described in Table 4.8.

Table 4.8. Minimum and maximum does of each *S. equi* barcoded IS*S1* library administered to Welsh mountain ponies.

Library	Minimum dose	Maximum dose
AC	5.08x 10^8 CFU	9.25x 10^8 CFU
CT	5.75x 10^8 CFU	8.42x 10^8 CFU
GA	7.83x 10^8 CFU	9.08x 10^8 CFU

Ponies were euthanised on the development of early clinical signs, which were pyrexia and a preference to eat hay and drink water over eating dry pelleted food. All animals were euthanised between 4 and 8 days post-challenge and post-mortem examinations conducted to remove the bilateral submandibular and retropharyngeal lymph nodes (Figure 4.3).

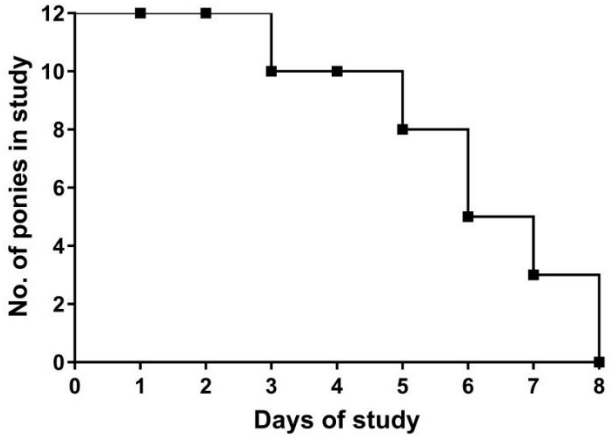


Figure 4.3. Kaplan-Meier curve of days post-challenge that Welsh mountain ponies were euthanised for post-mortem examination.

In total, 48 lymph nodes were recovered from the 12 animals, of which 24 retropharyngeal and 14 submandibular lymph nodes were suitable for analysis (bacterial load of > 5x 10^3 CFU/ml) (Figure 4.4, black and blue dots). Six of the 38 infected nodes contained < 4x 10^5 CFU/ml and were sequenced on a MiSeq as their diversity was predicted to be low and therefore did not require the higher sequencing capacity provided by the HiSeq (Figure 4.4, blue dots). Of the 10 insufficiently infected nodes, 8 did not contain any *S. equi* with the remaining 2 containing 667 CFU/ml and 4x 10^3 CFU/ml. Low bacterial load was used as a predictor of low diversity, which is explored further in section 4.3.4. The lymph nodes yielded, on average, 3.9x 10^7 (SEM ± 1x 10^7), 1.9x 10^7 (SEM ± 6.5x 10^6), 1.9x 10^6 (SEM ± 9.7x 10^5), and 2.8x 10^6 (SEM ± 1.1x 10^6) CFU/ ml in the LRP, RRP, LSM and RSM nodes, respectively (Figure 4.4, red lines). Statistical analysis revealed that the retropharyngeal lymph nodes yielded significantly higher bacterial loads than the submandibular lymph nodes (p < 0.00001).

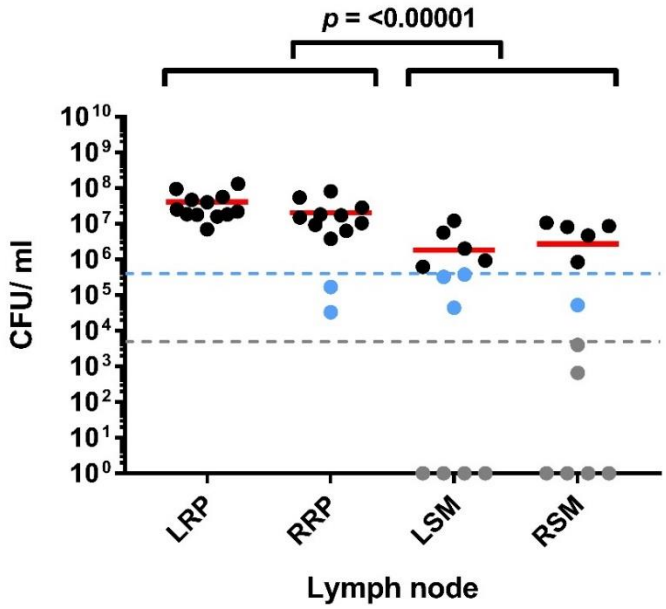


Figure 4.4. Bacterial loads of lymph nodes recovered from Welsh mountain ponies challenged with *S. equi* barcoded IS*S1* libraries. Mutant populations indicated by black dots were sequenced on an Illumina HiSeq, mutant populations represented by blue dots were sequenced on an Illumina MiSeq (< 4x 10^5 CFU/ml, blue dashed line) and those indicated by grey dots were not sequenced as the nodes were not sufficiently infected (> 5x 10^3 CFU/ml, grey dashed line). Red lines indicate the average bacterial load in each lymph node. The bacterial loads were significantly different between retropharyngeal and submandibular lymph nodes.

Initially, attempts were made to recover DNA for sequencing, directly from abscess material. However, as was similarly reported by Le Breton *et al.,* [78] attempts were unsuccessful as only 8 percent of sequencing reads contained IS*S1*. BLASTn searches of several random fragments identified the remaining reads as equine DNA. A nested PCR protocol was also attempted, which dramatically increased the proportion of IS*S1* containing reads to 89.7 percent, but this technique dramatically reduced output library diversity, revealing only 54 unique mutants that mapped to just 12 genes. In light of this, IS*S1* mutants were recovered from abscess materials by overnight growth on agar plates.

Abscess material from 1 lymph node (pony 5922, LRP) was initially grown on 100 large Petri dishes in batches of 10. The recovered mutants from each batch were sequenced by TraDIS and the data was progressively combined to identify the number of Petri dishes at which discovery of data in new genes plateaued (Figure 4.5) Data was split according to the 2 barcoded libraries administered to pony 5922; CT and GA and unique mutants identified. The percent of new mutants identified with each additional batch was calculated and it was concluded that recovering mutants on 30 Petri dishes provided a balance of practicality and mutant diversity (Figure 4.5).

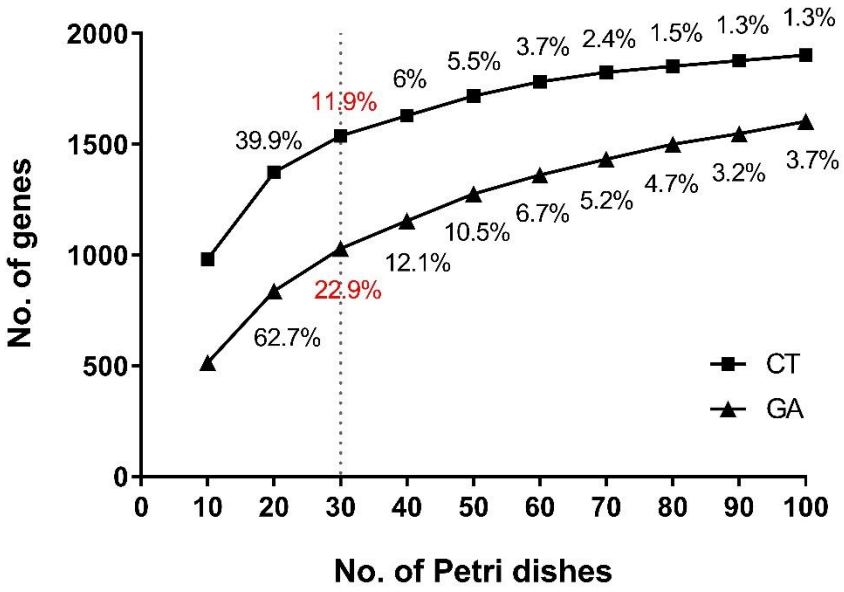


Figure 4.5. Discovery of new genes containing insertion sites with increased plating of recovered IS*S1* mutants from infected lymph nodes. Welsh mountain ponies were experimentally infected with 2 barcoded libraries. The animal used to generate this data was infected with the libraries referred to as CT and GA. Surviving mutants in the lymph nodes must be recovered on agar *in vitro* to separate them from the dense host abscess material before sequencing. Abscess material was plated on 10 batches of 10 plates, mutants sequenced by TraDIS and the data progressively combined and analysed to identify the optimum number of Petri dishes to sufficiently capture the population of mutants. The more Petri dishes used to recover mutants, the more new genes were identified, however, the discovery of new genes begins to slowly plateau around 30 dishes. Percentages indicate the number of new genes identified with each additional batch of 10 plates.

After overnight incubation, colonies were washed off the Petri dishes, DNA extracted alongside the stored input pool pellets, and DNA libraries were prepared. All DNA libraries were uniquely indexed before IS*S1*-genome junctions were sequenced by TraDIS as previously described. Indexing DNA from each lymph node facilitated analysis on a per animal basis, which is traditionally performed in similar *in vivo* transposon library studies, and also on a per barcoded library basis.

4.3.3 Barcoded and per animal analysis of TraDIS data

All sequencing data was trimmed of its first 2 bp to remove the barcodes, and output data combined according to the pony from which it originated, generating 12 datasets for per animal (PA) analysis (Table 4.9). In parallel, output data was combined on a per barcoded library basis, generating 3 output datasets, to determine if this methodology reduces the effects of stochastic loss (Table 4.10). In the 3 barcoded output libraries, 46 ‘jackpot genes’ were identified where each contained an extensive increase in the

number of reads for 1 unique mutant. One possible explanation for this increase in the amount of 1 mutant within a gene is if that mutant was the 'first past the post' and was able to populate a given lymph node before an innate immune response was triggered. These jackpot genes were removed and the data reanalysed to avoid bias. The jackpot genes may be more obvious in the barcoded (BC) analysis as fewer replicates were analysed.

Table 4.9. Composition of libraries recovered from 12 individual Welsh mountain ponies pre- and post-filtering. Mutants were recovered from up to 4 lymph nodes per animal, data combined on a per animal basis and analysed before determining gene fitness. The number of genes containing insertions post-filtering is consistent between animals, since filtering determines a consensus set of genes to be taken forward for analysis.

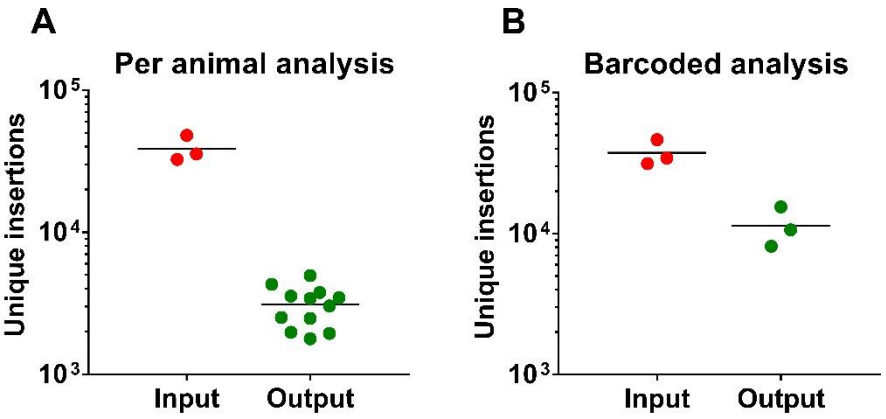
Pony	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes: % of non-essential genes)
477^{pre}	3,890	17,335,354	1,292 (59.7: 81.3)
2991^{pre}	2,807	13,547,528	1,116 (51.5: 70.2)
5867^{pre}	5,493	18,580,159	1,473 (68: 92.6)
5922^{pre}	4,861	15,167,789	1,457 (67.3: 91.6)
6061^{pre}	3,688	5,582,330	1,232 (56.9: 77.5)
6544^{pre}	4,179	17,252,877	1,329 (61.4: 83.6)
7454^{pre}	2,256	13,311,700	1,014 (46.8: 63.8)
7565^{pre}	3,460	15,950,820	1,254 (57.9: 78.9)
7616^{pre}	1,956	7,230,337	958 (44.2: 60.3)
7649^{pre}	3,787	16,055,851	1,289 (59.5: 81.1)
7799^{pre}	2,852	13,769,145	1,141 (52.7: 71.8)
7884^{pre}	2,260	16,694,637	994 (45.9: 62.5)
477^{post}	3,466	15,833,177	1,062 (49.1: 66.8)
2991^{post}	2,514	15,833,177	919 (42.4: 57.8)
5867^{post}	4,952	15,833,177	1,176 (54.3: 74)
5922^{post}	4,295	15,833,177	1,149 (53.1: 72.3)
6061^{post}	3,418	15,833,177	1,156 (53.4: 72.7)
6544^{post}	3,760	15,833,177	1,083 (50: 68.1)
7454^{post}	1,978	15,833,177	837 (38.7 52.6)
7565^{post}	3,031	15,833,177	1,007 (46.5: 63.3)
7616^{post}	1,778	15,833,177	822 (38: 51.7)
7649^{post}	3,351	15,833,177	1,041 (48.1: 65.5)
7799^{post}	2,477	15,833,177	922 (42.6: 58)
7884^{post}	1,941	15,833,177	811 (37.5: 51)

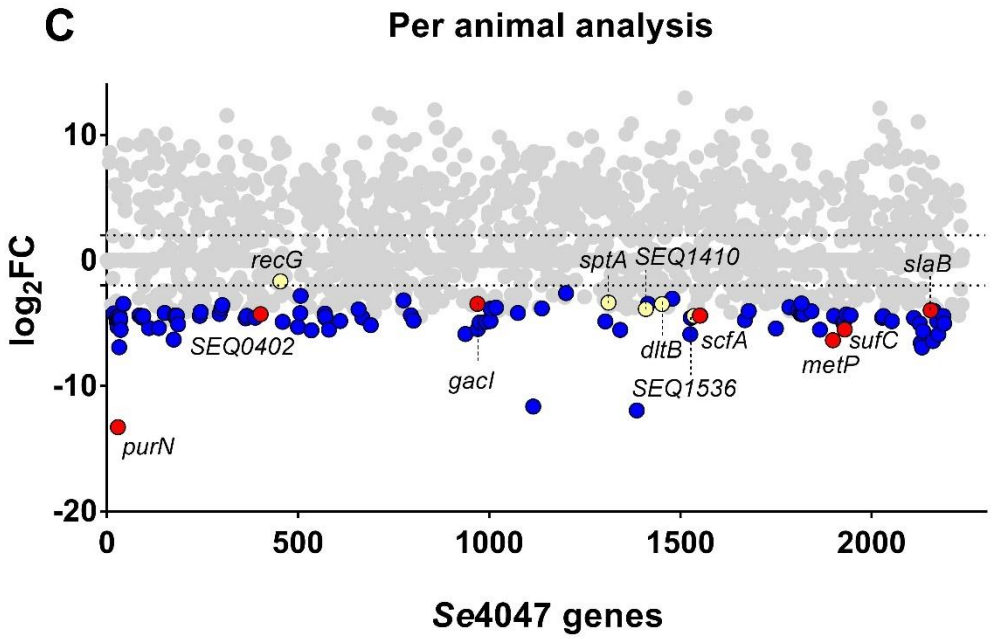
Table 4.10.Composition of barcoded libraries recovered from 12 Welsh mountain ponies pre- and post-filtering. Mutants were recovered from up to 4 lymph nodes per animal, data combined, split according to barcode and analysed before determining gene fitness. The number of genes containing insertions post-filtering is consistent between libraries, since filtering determines a consensus set of genes to be taken forward for analysis.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC^{pre}	13,792	50,477,885	1,829 (84.5 : 100)
CT^{pre}	19,922	64,419,073	1,894 (87.7 : 100)
GA^{pre}	11,024	69,425,389	1,764 (81.5 : 100)
AC^{post}	10,645	36,205,787	1,275 (58.9 : 80.2)
CT^{post}	15,449	36,205,787	1,294 (59.8 : 81.4)
GA^{post}	8,127	36,205,787	1,236 (57.1 : 77.7)

The per animal (PA) analysis of our TraDIS data identified, on average, 3,096 ± 286 (SEM) unique mutants per pony, equating to 8 percent recovery of the unique mutants

within the challenge inoculum (Figure 4.6A). The per animal output mutants were located within, on average, 74 percent ± 3 percent (SEM) of the 1,359 *S. equi* genes meeting the inclusion criteria. In contrast, the barcoded (BC) analysis identified 11,407 ± 2,148 (SEM) unique mutants per output library on average, representing 31 percent of the mutants within the challenge inoculum and 96.2 percent ± 1.3 percent (SEM) of 1,319 *S. equi* genes meeting the BC input pool inclusion criteria (Figure 4.6B and 4.7).


A
Per animal analysis
Unique insertions
10^5
10^4
10^3
Input
Output
B
Barcoded analysis
Unique insertions
10^5
10^4
10^3
Input
Output


C
Per animal analysis
log_2FC
10
0
-10
-20
recG
SEQ0402
purN
gacI
sptA
SEQ1410
dltB
scfA
SEQ1536
slaB
sufC
metP
0
500
1000
1500
2000
Se4047 genes

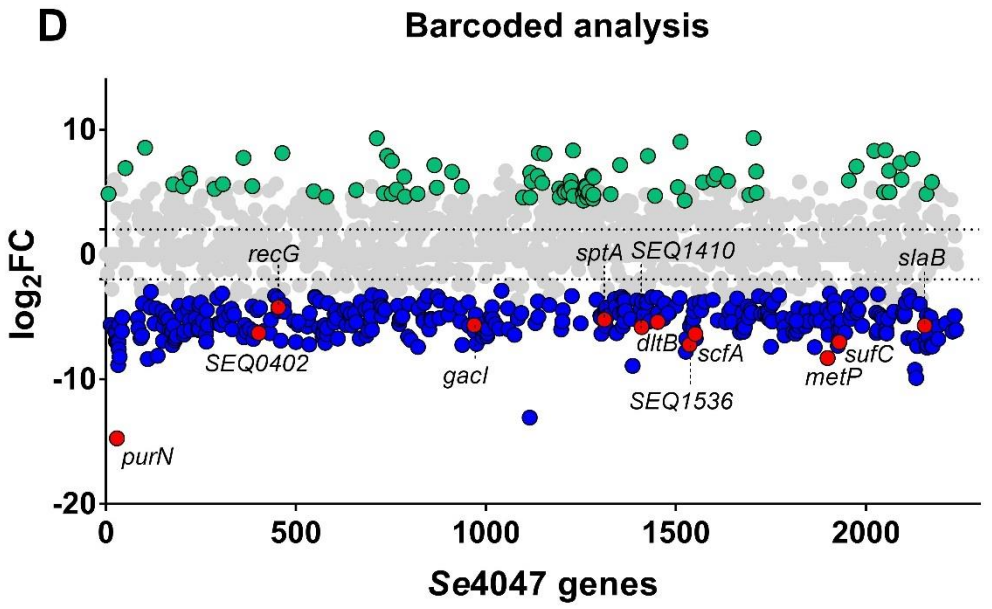

D
Barcoded analysis
log_2FC
10
0
-10
-20
recG
SEQ0402
purN
gacI
sptA
SEQ1410
dltB
scfA
SEQ1536
slaB
sufC
metP
0
500
1000
1500
2000
Se4047 genes

Figure 4.6. Number of unique mutants identified by per animal and barcoded analysis and genome-wide gene fitness assigned by the 2 analysis methods. The number of unique mutants in the input pools of mutants are equal between the per animal and barcoded analysis, but the number of unique mutants identified in the output pools depend on the analysis technique. A) On average, 3,096 ± 286 unique mutants were identified in the 12 output pools. B) On average, 11,407 ± 2,148 unique mutants were identified in the 3 output pools as instead of analysing data on a per animal basis, all recovered mutants are combined and split according to library barcode, generating only 3 samples. C) Genome-wide fitness of each gene as determined by the per animal technique. Blue dots indicate 97 genes required for fitness ($log_2FC < -2$, $q < 0.05$), red dots indicate a panel of genes required for fitness selected for validation, cream dots correspond to genes identified as contributing to fitness in the barcoded analysis, but not in the per animal analysis. Grey dots indicate genes non-essential to *in vivo* fitness. D) Genome-wide fitness of each gene as determined by the barcoded technique. Blue dots indicate 368 genes required for fitness ($log_2FC < -2$, $q < 0.05$), red dots indicate a panel of required fitness genes selected for validation. Green dots indicate 85 genes conferring a fitness advantage upon insertion and grey dots indicate genes non-essential to *in vivo* fitness.

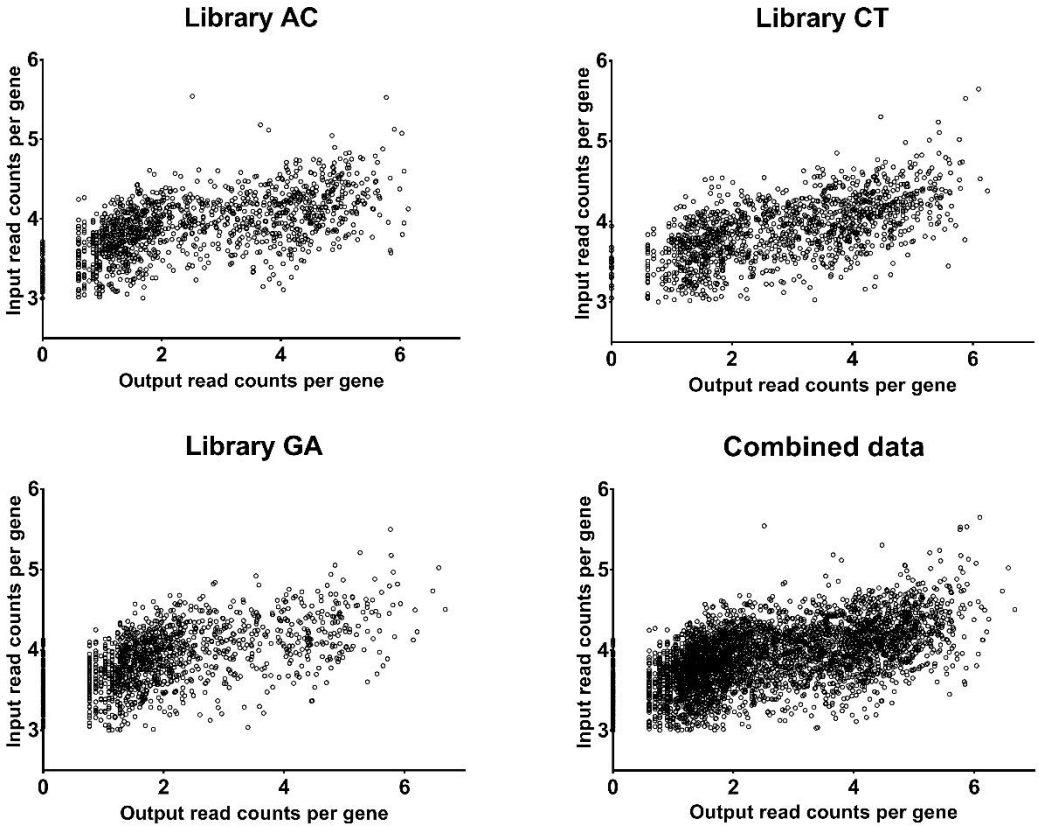


Figure 4.7. Read counts per gene in each of 3 *S. equi* barcoded IS*S1* libraries, pre- (input) and post- (output) infection of 12 Welsh mountain ponies. Genes represented by < 1,000 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Two populations of data can be roughly observed in Figure 4.7; genes that contain fewer reads in the output libraries compared to the input libraries, and genes that contain a similar number of reads in both the input and output libraries. Gene fitness was calculated by comparing the ratio (log_2FC) of read counts, per gene, in the PA and BC output pools to the 3 input pools. PA analysis identified 97 genes required for fitness ($log_2FC < -2$, $q < 0.05$) (Figure 4.6C, blue and red dots) and the BC analysis identified 368 genes required for fitness ($log_2FC < -2$, $q < 0.05$) (Figure 4.6D, blue and red dots, Table A3.1, Appendix 3). Further analysis identified 85 genes that conferred a significantly increased fitness following IS*S1* insertion in the BC analysis ($log_2FC > 2$, $q < 0.05$) (Figure 4.6D, green dots, Table A3.2, Appendix 3), however no genes conferring an increased fitness were identified by the PA analysis (Figure. 4.6C).

All 97 genes required for fitness that were identified in the PA analysis were also identified using the BC analysis (Figure 4.8A). These shared genes had an average log_2FC of -6.7 based on the BC analysis. The remaining 295 genes that were uniquely identified by the BC analysis had an average log_2FC of -4.9, highlighting the improved sensitivity of the BC method of data interpretation. Therefore, splitting the TraDIS data into the 3 barcoded libraries generated a 379 percent increase in the number of genes that were identified as conferring a fitness defect upon insertion, when compared to the PA analysis.

Clusters of orthologous groups (COG) enrichment of the genes required for fitness identified in the PA analysis found that 36 percent of these genes did not belong to a defined COG category, with an additional 9 percent of the data contributing to both the ‘function unknown’ and ‘general function prediction only’ categories (Figure. 4.8B). Other prevalent functional groups included nucleotide transport and metabolism (14 percent), posttranslational modification (6 percent) and amino acid transport and metabolism (6 percent).

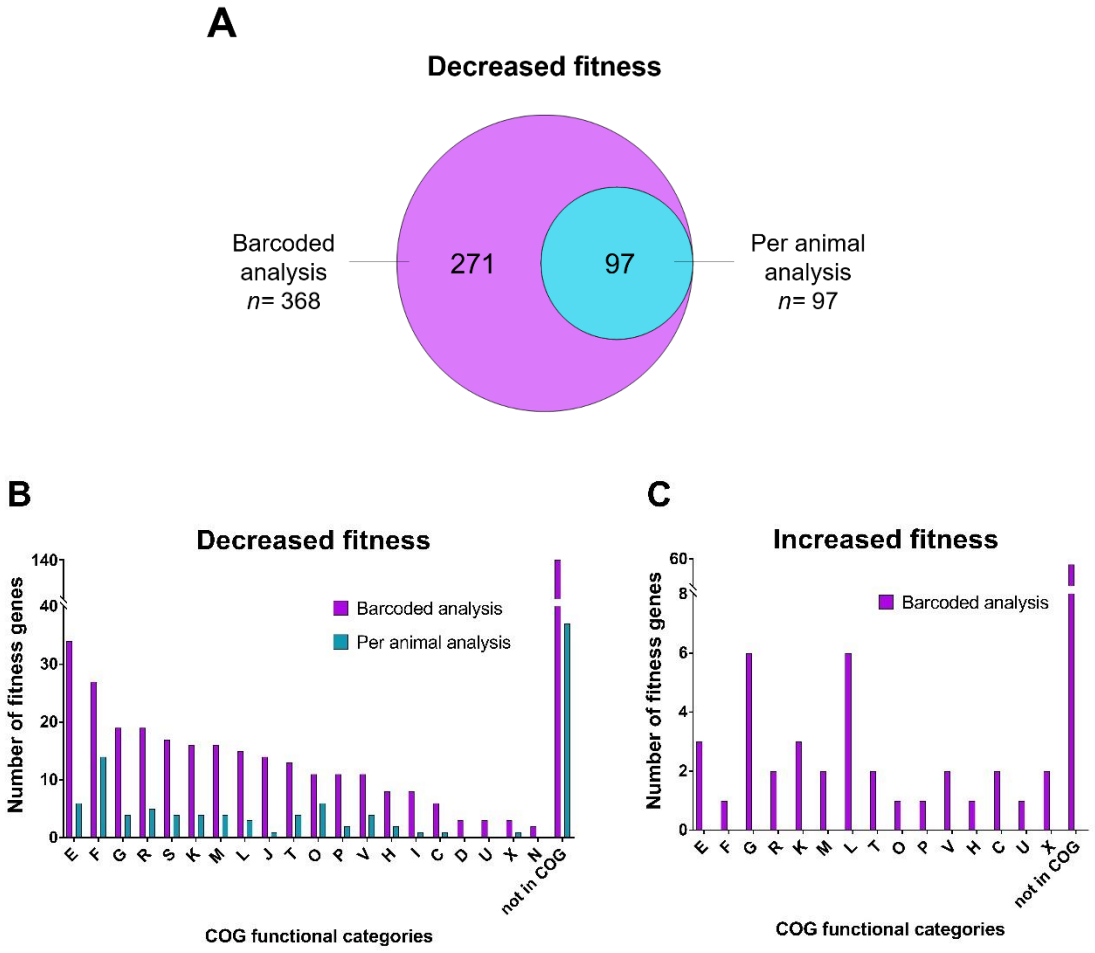


Figure 4.8. Comparison of the fitness genes identified in the per animal analysis to the barcoded analysis and the COG functional categories of the fitness genes. A) Venn diagram illustrating that all 97 fitness genes identified by the per animal analysis were also identified by the barcoded analysis technique. B) COG categories assigned to the genes required for fitness in both the per animal and barcoded analysis techniques. C) COG categories assigned to the genes conferring an enhanced fitness upon insertion identified in the barcoded analysis. C: Energy production and conversion, D: Cell cycle control, cell division, chromosome partitioning, E: Amino acid transport and metabolism, F: Nucleotide transport and metabolism, G: Carbohydrate transport and metabolism, H: Coenzyme transport and metabolism, I: Lipid transport and metabolism, J: Translation, ribosomal structure and biogenesis, K: Transcription, L: Replication, recombination and repair, M: Cell wall/membrane/envelope biogenesis, N: Cell motility, O: Posttranslational modification, protein turnover, chaperones, P: Inorganic ion transport and metabolism, R: General function prediction only, S: Function unknown, T: signal transduction mechanisms, U: Intracellular trafficking, secretion, and vesicular transport, V: Defence mechanisms, X: Mobilome- prophages, transposons.

COG enrichment of the BC analysis revealed that 35 percent and 60 percent of genes did not belong to a defined COG category, in the decreased and increased fitness groups, respectively (Figure. 4.7B, C). In the decreased fitness pool, an additional 9 percent of genes were identified as of unknown function or as a general function prediction only. Other principal COG categories defined in the decreased fitness data include genes associated with nucleotide (9 percent), amino acid (7 percent) and

carbohydrate (5 percent) transport and metabolism, transcription (4 percent), cell wall/membrane/envelope biogenesis (4 percent) and replication, recombination and repair (4 percent). Genes conferring an increased fitness upon insertion were mainly associated with carbohydrate (7 percent) and amino acid (3 percent) transport and metabolism, replication, recombination and repair (7 percent) and transcription (3 percent).

4.3.4 Measurement of mutant translocation *in vivo* post infection

Using barcoded libraries and uniquely indexing each lymph node for sequencing additionally enabled analysis on a per lymph node basis. Basic analysis to determine the rate of translocation of *S. equi* from an inoculation site on 1 side of the head, to lymph nodes on the opposing side was performed. The transitioning of *S. equi* within the head after challenge is not currently known. However, anecdotally, abscesses were predominantly located on the left side of the head when ponies were only challenged via their left nostril (Waller *et al*, unpublished data). Barcoded IS*S1* mutants progressing to the lymph nodes on the same side of the head as the nostril in which they were inoculated are referred to as the primary library within nodes, with the barcoded IS*S1* mutants transitioning from the opposing side referred to as secondary mutants. On average, 21 percent $\pm$ 4.3 (SEM) and 30 percent $\pm$ 4.3 of barcoded *S. equi* IS*S1* mutants recovered from the LRP and RRP lymph nodes, respectively, had been inoculated into the opposing nostril. For the LSM and RSM lymph nodes, 13 percent $\pm$ 3.6 and 11 percent $\pm$ 2.3 of recovered mutants had been inoculated into the opposing nostril (Figure 4.9).

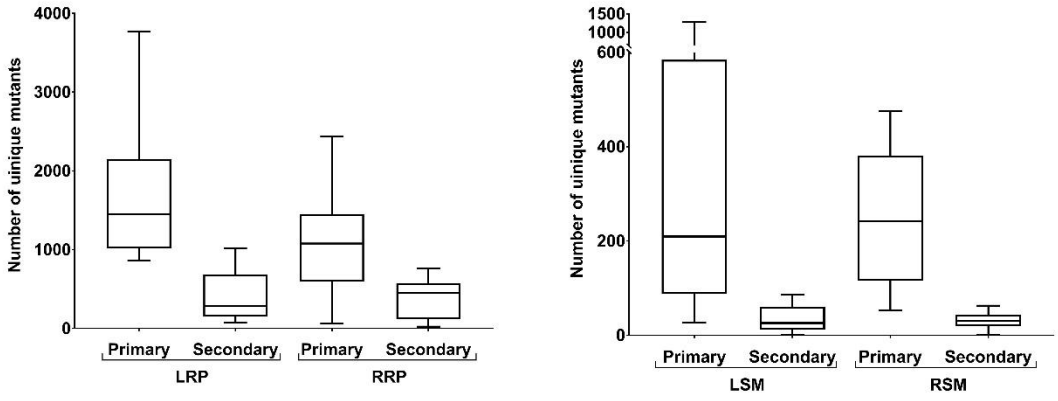


Figure 4.9. Unique mutants identified in each lymph node recovered from Welsh mountain ponies challenged with 2 barcoded *S. equi* IS*S1* libraries. Each pony was challenged with 1 unique barcoded library per nostril. Both barcoded libraries can be identified in each of the lymph nodes, providing evidence that libraries translocate to the opposing side of the head from which it was administered. The 'primary mutants' refer to the library administered to the same side of the head as the lymph node sampled. The 'secondary mutants' refer to mutants that have originated from the nostril on the opposing side of the head to the lymph node sampled. LRP= left retropharyngeal lymph node, RRP= right retropharyngeal lymph node, LSM= left submandibular lymph node and RSM= right submandibular lymph node.

4.3.5 Validation of *S. equi* genes required for fitness in the natural host

To validate our results and confirm the benefit of applying a barcoded technique, we selected 12 genes with a fitness defect, plus 1 control gene not affected by IS*S1* insertion, for tagged allelic replacement mutagenesis and repeat challenge *in vivo*. Seven genes required for fitness in both PA and BC analyses were selected, plus an additional 5 genes uniquely identified by the BC analysis (Table 4.11, Figure 4.10).

Table 4.11.Twelve genes with a fitness defect in the *S. equi in vivo* TraDIS screen selected for validation. One additional gene, *SEQ0751*, was non-essential *in vivo* and was included as an internal control (green). Purple rows indicate genes identified when data was analysed on a per animal basis and a per barcoded library basis. White rows indicate genes uniquely identified by the barcoded library analysis.

Gene	Locus tag	Function	Log_2FC	*q* value
purN	SEQ0029	phosphoribosylglycinamide formyltransferase	-14.7	3.2×10^{-7}
SEQ0402	SEQ0402	putative cell surface-anchored protein	-6.3	5.5×10^{-4}
scfA	SEQ1551	putative permease	-6.4	2.7×10^{-4}
metP	SEQ1899	putative D-methionine transport system permease	-8.3	6.5×10^{-5}
sufC	SEQ1930	putative ABC transporter, ATP-binding protein	-7	3.1×10^{-4}
slaB	SEQ2155	putative exported protein	-5.7	4.3×10^{-4}
gacI	SEQ0969	putative glycosyl transferase	-5.7	2.1×10^{-4}
recG	SEQ0454	ATP-dependent DNA helicase	-4.2	2.85×10^{-3}
sptA	SEQ1312	putative exported protein	-5.2	5.8×10^{-3}
SEQ1410	SEQ1410	ABC transporter ATP-binding protein	-5.8	4.96×10^{-4}
dltB	SEQ1452	putative activated D-alanine transport protein	-5.4	5.3×10^{-4}
SEQ1536	SEQ1536	putative exported protein	-7.2	8.5×10^{-5}
SEQ0751	SEQ0751	putative DNA-binding protein	0.3	0.9

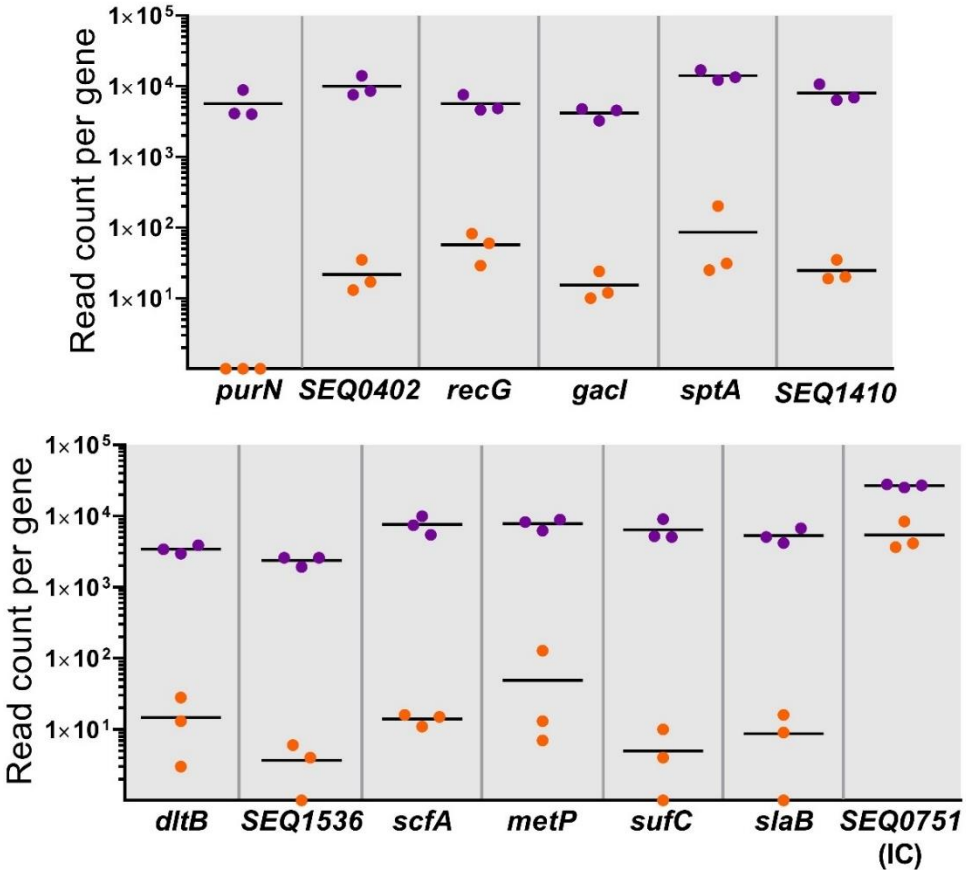


Figure 4.10. Read counts per gene in the input and output pools of *S. equi* genes selected for validation from an *in vivo* TraDIS screen. All genes with a fitness defect, according to TraDIS, were well represented in the input pools (purple dots), but significantly fewer reads were recovered in the output pools (orange dots). *SEQ0751* is a non-essential gene *in vivo* and was used an internal control (IC). Black horizontal lines indicate mean values.

When generating the mutant strains for validation, target genes were replaced by 1 of 3 80 bp tag sequences. The first 57 bp in all tags matched the IS*S1* sequencing primer binding site for TraDIS, allowing the pool of validation mutants to be individually measured by TraDIS. The remaining 23 bp of each tag varied to maximise DNA library diversity and improve cluster differentiation during next generation sequencing. Three internal control (IC) strains were generated by replacing *SEQ0751*, a gene unaffected by IS*S1 in vivo*, with each of the 3 tags. The 3 IC strains acted as experimental replicates of one another and allowed comparison to the respective tagged mutants.

Five Welsh mountain ponies were each challenged intranasally with the inoculum containing the validation and IC control mutants at a dose of 5 x 10^8 CFU, administered in 2.5 ml of THBFCS per nostril. The minimum and maximum doses the animals received were 6.9x 10^8 CFU/ml and 1.1x 10^9 CFU/ml, respectively. Twenty percent of the inoculum contained the 12 tagged validation deletion mutants, combined in equal proportions, 40

percent of the inoculum contained the 3 IC strains in equal proportions and 40 percent of the inoculum was WT *Se*4047. The proportion of the 12 tagged validation mutants in the total inoculum was chosen to replicate the relative amount of the genes pertaining to reduced fitness in the TraDIS screen. Equal concentrations of the combined IC mutants and WT *Se*4047 were included to provide a competitive environment for the tagged mutants and to determine if the tags had an effect on fitness by measuring the proportion of IC mutants relative to WT *Se*4047. Challenging ponies in this way reduced the number of animals required as the IC strains were validated against WT *Se*4047, within the same animal without the need for separate control ponies.

Two ponies were euthanised upon developing early clinical signs of disease on day 6 post-challenge; pyrexia and preference of haylage over dry pelleted food. However, the remaining 3 animals did not develop obvious clinical signs and were euthanised 10 days after challenge. Post-mortem examinations were conducted to remove the bilateral submandibular and retropharyngeal lymph nodes. Statistical analysis revealed that the retropharyngeal lymph nodes yielded significantly higher bacterial loads than the submandibular lymph nodes ($p < 0.005$), as in the *in vivo* TraDIS screen (Figure 4.11A). Three of the validation ponies did not show obvious clinical signs and were therefore predicted to contain less surviving *S. equi*. This hypothesis was supported by the significant difference in bacterial yields compared to the ponies euthanised on day 6 due to the presence of obvious clinical signs ($p = 0.04$).

Compared to the TraDIS screen, the total CFU/ml of *S. equi* present in the retropharyngeal lymph nodes was, on average, similar to that recovered in the TraDIS screen, however the SEMs of the validation data were large (validation data: LRP average = 4.5×10^7 CFU/ml SEM ± 2.4×10^7, RRP average = 4.5×10^7 CFU/ml SEM ± 2.8×10^7) (Figure 4.11B). Statistical analysis revealed that there was no significant difference in the bacterial loads recovered from the retropharyngeal lymph nodes in the TraDIS screen and validation studies ($p = 0.3$). The bacterial loads recovered from the submandibular lymph nodes were also not significantly different between the 2 studies ($p = 0.15$) (Figure 4.11C).

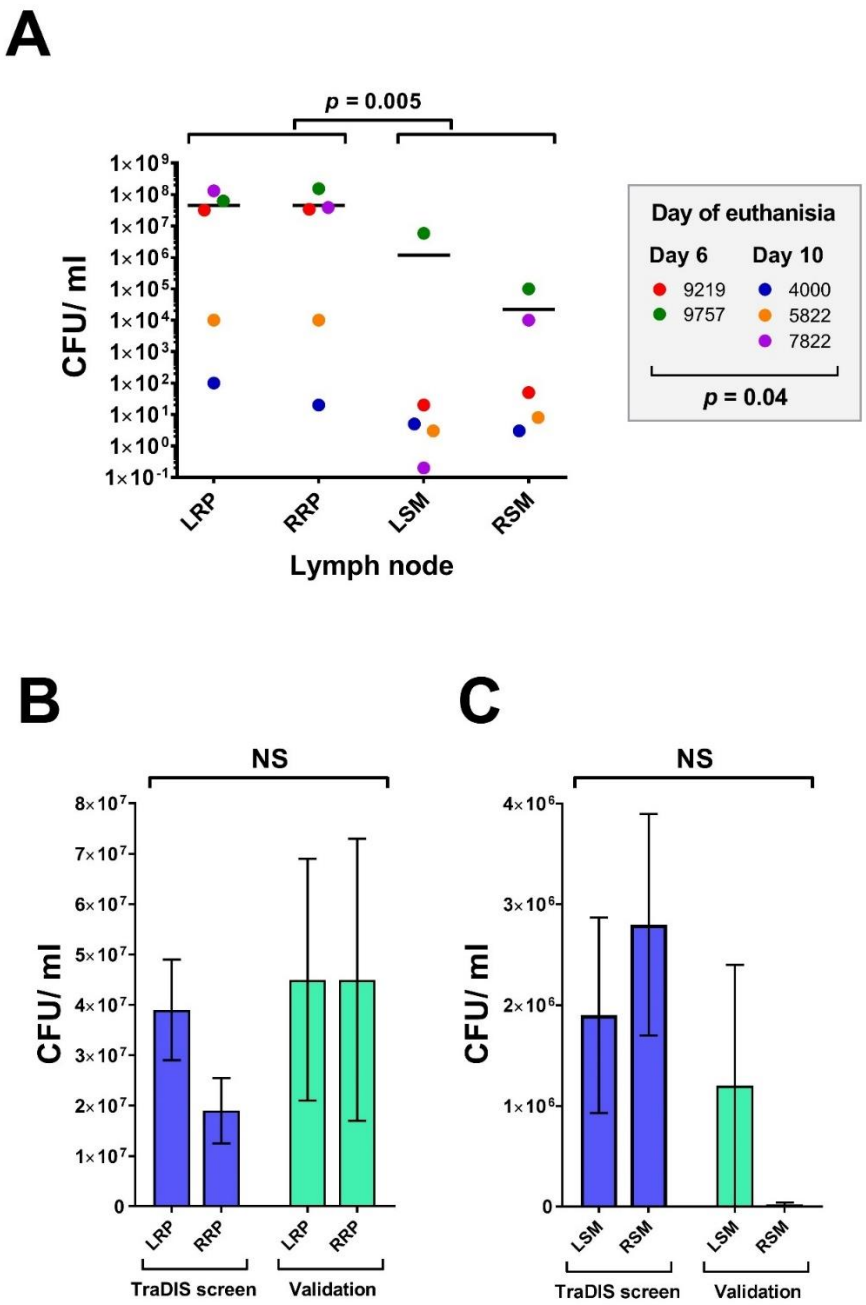


Figure 4.11. Bacterial loads recovered from the infected lymph nodes of Welsh mountain ponies challenged with a panel of *S. equi* tagged deletion mutants. A) Bacterial loads of each lymph node from all animals. Two animals, 9219 and 9757, showed obvious clinical signs on day 6 post-challenge and were euthanised for post-mortem recovery of infected lymph nodes. The remaining 3 animals did not show obvious clinical signs and were euthanised on day 10 post-challenge. The bacterial loads in the lymph nodes of the animals euthanised on day 6 post-challenge were significantly higher than those measured in the ponies showing no obvious clinical signs (p= 0.04). Considering all 5 animals, the bacterial loads of the retropharyngeal lymph nodes was higher than that recovered from the submandibular lymph nodes (p= 0.005). B) Bacterial loads of the retropharyngeal lymph nodes in the TraDIS screen and validation studies, which were not significantly different from one another. C) Bacterial loads of the submandibular lymph nodes in the TraDIS screen and validation studies, which were also not significantly different from one another. LRP= left retropharyngeal lymph node, RRP= right retropharyngeal lymph node, LSM= left submandibular lymph node and RSM= right submandibular lymph node.

Sequencing of recovered *S. equi*

TraDIS was utilised to measure the amount of each validation mutant and the IC strains in the inoculum and from the abscess materials recovered from the 5 ponies. No TraDIS reads were detected for the deletion mutants Δ*purN*tagB, Δ*SEQ0402*tagA, Δ*SEQ1536*tagA, Δ*scfA*tagA or Δ*sufC*tagB in the recovered abscess materials (Table 4.12, Figure 4.12). The strains Δ*recG*tagB, Δ*gacI*tagC, Δ*sptA*tagB, Δ*dltB*tagC and Δ*metP*tagB were detected at very low levels (≤ 11 reads each) across all animals (Table 4.12, Figure 4.12). The strains Δ*SEQ1410*tagC and Δ*slaB*tagC were detected at higher levels (1,330 and 4,133 reads, respectively). Reads corresponding to Δ*SEQ1410*tagC were sequenced from 2 animals and those corresponding to Δ*slaB*tagC were present in all 5 animals. For both these mutants, the vast majority of reads were attributable to 1 animal, however (Table 4.12, Figure 4.12).

Table 4.12. Sequencing reads corresponding to *S. equi* tagged validation mutants present in the inoculum and in the lymph nodes of 5 ponies challenged with the inoculum. Sequencing by TraDIS revealed that few reads corresponding to the mutants predicted to be attenuated *in vivo* were present in the infected lymph nodes of 5 experimentally challenged Welsh mountain ponies. Three tagged deletion mutants predicted to unaffected *in vivo* were not consistently recovered from ponies. Two ponies showed obvious clincal signs and were euthanised on day 6 post-challenge. The remaining 3 ponies did not show obvious clinical signs and were euthanised on day 10 post-challenge due to welfare concerns. Asterix indicate bacterial loads of the lymph nodes, **** > 1x 10^7 CFU/ml, *** >1x 10^5 CFU/ml, ** > 1x 10^3 CFU/ml, * < 1x 10^2 CFU/ml.

		Pony 9219 (obvious clinical signs)				Pony 9757 (obvious clinical signs)				Pony 4000 (**no** clinical signs)				Pony 5822 (**no** clinical signs)				Pony 7822 (**no** clinical signs)				
Validation mutant	Inoculum	LRP ****	RRP ****	LSM *	RSM *	LRP ****	RRP ****	LSM ***	RSM ***	LRP *	RRP *	LSM *	RSM *	LRP **	RRP **	LSM *	RSM *	LRP ****	RRP ****	LSM *	RRP **	Total reads per mutant
*ΔpurN*tagB	1,686	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*ΔSEQ0402*tagA	1,387	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*ΔrecG*tagB	1,498	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
*Δgacl*tagC	1,134	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
*ΔsptA*tagB	1,127	1	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
*ΔSEQ1410*tagC	915	1	1,328	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1,330
*ΔdltB*tagC	1,200	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
*ΔSEQ1536*tagA	1,102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*ΔscfA*tagA	1,085	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*ΔmetP*tagB	1,548	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
*ΔsufC*tagB	1,238	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
*ΔslaB*tagC	1,900	3,216	910	1	0	0	0	2	0	0	1	0	1	0	0	0	1	0	1	0	0	4,133
ΔICtagA	11,587	3,434	9,028	4	85	8,284	14,127	3	3	2	1	0	3	0	12,400	1	0	0	0	2	55,718	103,095
ΔICtagB	9,760	72,298	67,151	19	680	37,349	69,091	87,439	86,769	0	11	1	27	25,893	67,728	9	7	85,834	86,603	1	1	686,911
ΔICtagC	13,200	1	0	0	0	0	0	0	0	0	0	0	0	0	269	0	0	0	0	0	0	270
Combined ΔIC	34,547	75,733	76,179	23	765	45,633	83,218	87,442	86,772	2	12	1	30	25,893	80,397	10	7	85,834	86,603	3	55,719	790,276

Quantifying the 3 ΔIC strains individually revealed that they did not behave comparably to one another. The ΔICtagB strain was isolated from 19 of the 20 lymph nodes collected from the 5 animals, however the densities were not consistent across the nodes (Table 4.12). ΔICtagA was recovered from 15 lymph nodes, but the amount sequenced was considerably less than ΔICtagB. The ΔICtagC was only present in 2 of the nodes, at very low levels. Combining the number of ΔIC reads enabled some comparisons to be made, but no statistical analysis was possible using this data (Table 4.12, Figure 4.12).

Young (approx. 1 year old) animals, as used in the *in vivo* TraDIS screen, consistently develop infection after challenge with WT *Se*4047 [13, 25, 47]. 2.5-year old ponies were used in the validation study. Older ponies may have a more mature immune system able to mount a response to infecting *S. equi.* This is supported by the significant difference in bacterial loads recovered from the 2 validation ponies showing obvious clinical signs compared to the 3 that were seemingly 'healthy'. qPCR to quantify the presence of WT *Se*4047 in the inoculum and in the recovered abscess material remains to be conducted.

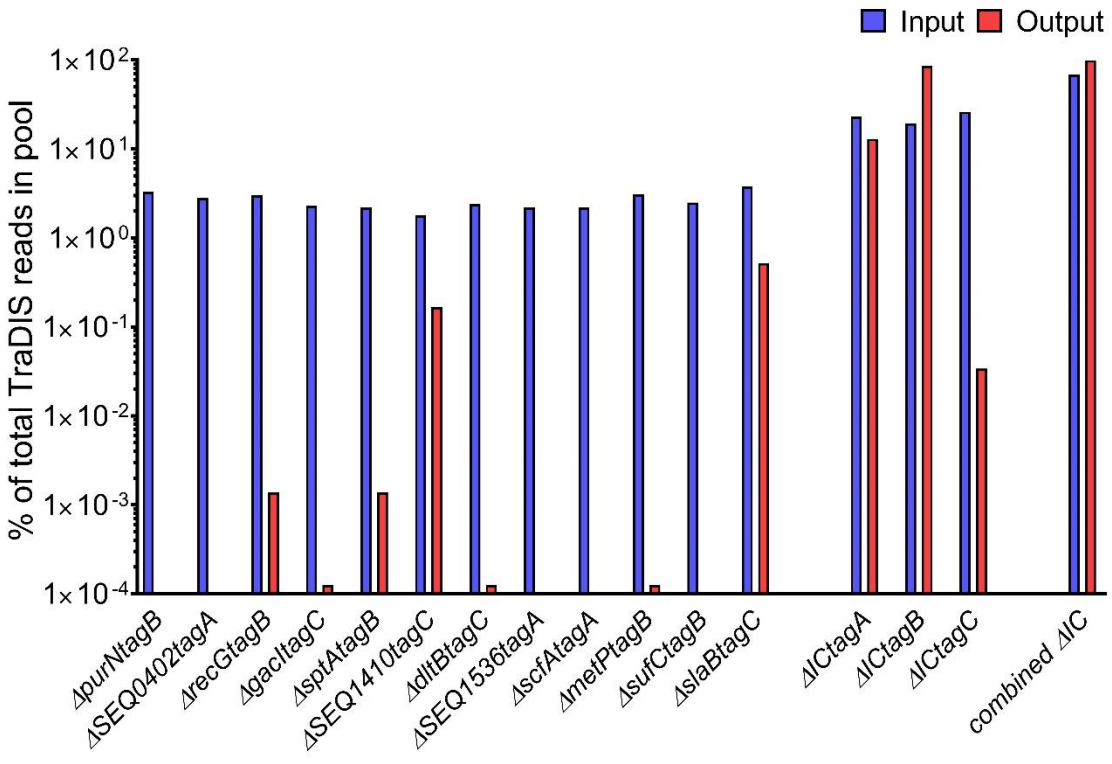


Figure 4.12. Percentage of total TraDIS reads contributed by each mutant, in the inoculum (blue bars) and recovered material (red bars) from ponies challenged with a panel of tagged deletion mutants. The internal control (IC) mutants contain a deletion in a gene not required for infection, so were predicted to behave as wild-type.

4.3.6 Comparison of *S. equi* genes required in whole equine blood and H_2O_2 *in vitro,* to *in vivo*

Comparison of the whole equine blood and H_2O_2 fitness genes to those identified *in vivo*, revealed a set of 9 genes required in all 3 conditions (null= 0.9 genes), equating to 2.4 percent of genes required for *in vivo* fitness, reflecting the complex nature of natural infection and the hurdles faced by the mutants (Figure 4.13). In whole equine blood and *in vivo*, 14 genes were similarly required (null= 1.3 genes), which equates to 39 percent of the 36 genes identified as contributing to fitness in whole equine blood. No genes were uniquely required in H_2O_2 and *in vivo*, since all but 1 gene was commonly essential in H_2O_2 and whole equine blood. Comparing the H_2O_2 data directly to the *in vivo* data, 60 percent of genes identified in H_2O_2 were also required in ponies (null= 0.08 genes).

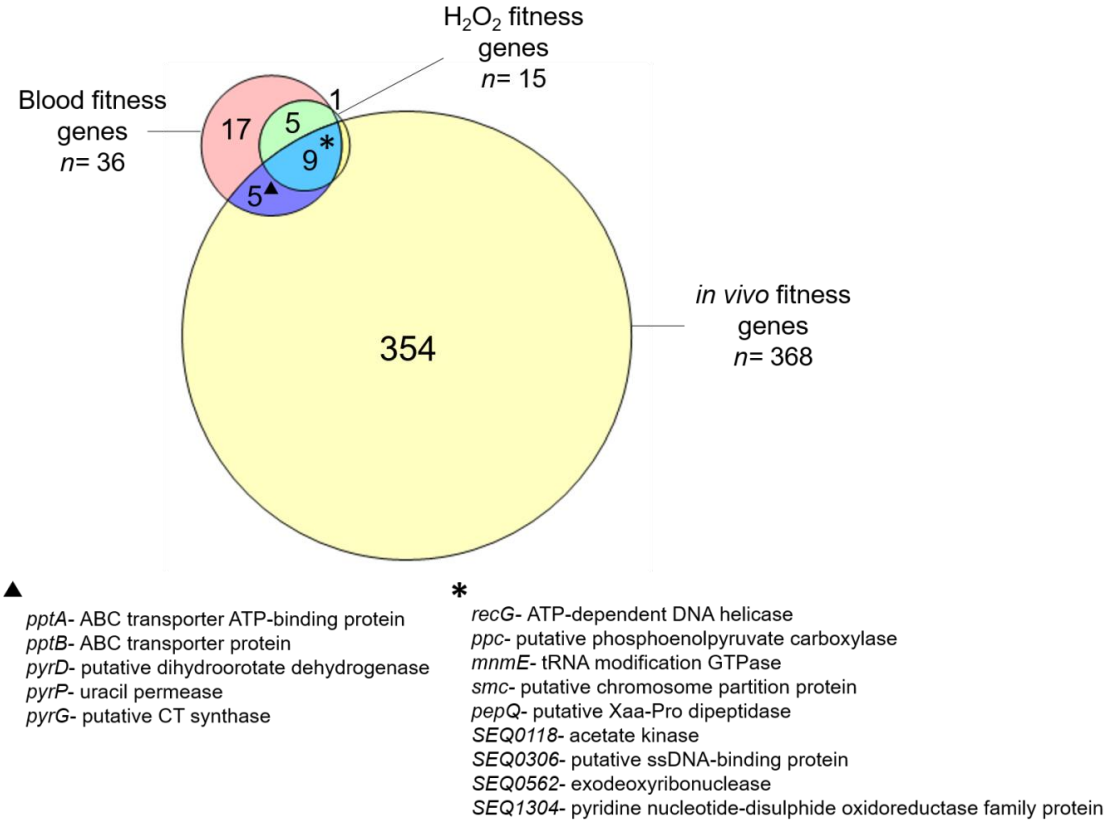


Figure 4.13. Venn diagram comparing the *S. equi* genes required for survival *in vivo* and *in vitro* in whole equine blood and in H_2O_2. Nine genes were commonly identified as important in all 3 niches (Asterix) and 5 genes were similarly important *in vivo* and in whole equine blood (triangle). Many more fitness genes were uniquely identified *in vivo*, owing to the complex nature of the host infection.

4.3.7 Comparative analysis of the genes implicated in *in vivo* infection in *S. equi* vs *S. pyogenes in vivo* and *ex vivo*

Genes required for fitness *in vivo*

Recent studies in *S. pyogenes* have utilised TraDIS/Tn-seq to measure gene fitness *in vivo* and *ex vivo*. Subcutaneous infection of mice with Tn-seq libraries in serotype M1 strain MGAS5448 identified 147 genes contributing to fitness *in vivo,* 101 had homologues in *S. equi* and could therefore be directly compared to the genes contributing to fitness in *S. equi* in the natural host. TraDIS libraries in *S. pyogenes* serotype M1 strain MGAS2221 and serotype M28 strain MGAS27961 were evaluated in a non-human primate (NHP) infection model, in which 72 consensus genes were identified as required for infection by both the M1 and M28 serotypes [145]. Sixty homologous genes could be included in the comparative analyses. Twenty-three genes were identified as important for survival *in vivo* in all 3 datasets (null= 0.7 genes), identifying a core set of genes required by both serotypes of *S. pyogenes* and *S. equi* for infection (Figure 4.14A, B, Table 4.13).

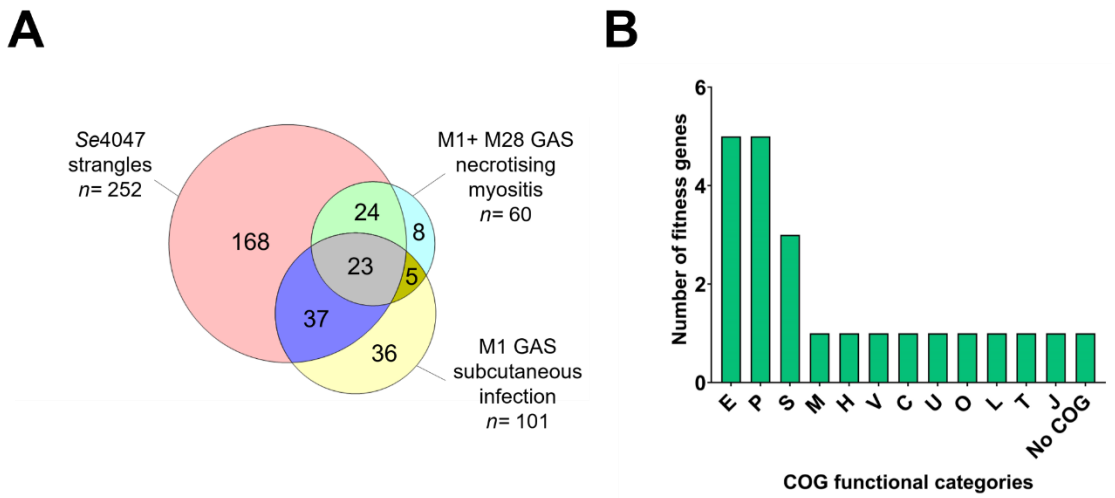


Figure 4.14. Comparison of homologous *S. equi* and *S. pyogenes in vivo* fitness genes and the functional COG categories of the consensus genes.A) Venn diagram comparing the *S. equi* genes required for survival *in vivo* and the *S. pyogenes* genes required for *in vivo* infection of serotype M1 in a murine model of subcutaneous infection and serotypes M1 and M28 in a non-human primate model of necrotising myositis. Twenty-three pan-species consensus genes are required for fitness by *S. equi* and *S. pyogenes* in all 3 niches. B) Functional COG categories assigned to the 23 consensus genes. E: Amino acid transport and metabolism, P: Inorganic ion transport and metabolism, S: Function unknown, M: Cell wall/membrane/envelope biogenesis, H: Coenzyme transport and metabolism, V: Defence mechanisms, C: Energy production and conversion, O: Posttranslational modification, protein turnover, chaperones, L: Replication, recombination and repair, T: Signal transduction mechanisms, J: Translation, ribosomal structure and biogenesis.

Table 4.13.Twenty-three consensus genes required for fitness *in vivo* in *S. pyogenes* serotype M1 in a subcutaneous murine model of infection, in *S. pyogenes* serotypes M1 and M28 strain in a necrotising myositis non-human primate model of infection and in *S. equi* in the natural equine host. Genes involved in transport (blue rows) and genes validated in *S. equi* in the natural equine host with allelic replacement mutants (bordered rows).

*Se*4047/MGAS5005 locus tag	Gene	Function
SEQ0005/Spy0004		putative GTP-binding protein
SEQ0095/Spy0079	*adcB*	ABC transporter permease protein
SEQ0255/Spy0161	*perR*	ferric uptake regulator family protein
SEQ0454/Spy1519	*recG*	ATP-dependent DNA helicase
SEQ0506/Spy1471	*pptA*	ABC transporter ATP-binding protein
SEQ0507/Spy1470	*pptB*	ABC transporter protein
SEQ0728/Spy1099		cell envelope-related transcriptional attenuator domain protein
SEQ0768/Spy0499		thiamine transporter
SEQ0776/Spy0505	*ppc*	putative phosphoenolpyruvate carboxylase
SEQ0853/Spy0537	*aspC*	aspartate aminotransferase
SEQ0969/Spy0610	*gacI*	putative glycosyl transferase
SEQ1304/Spy0657	*hupX*	pyridine nucleotide-disulphide oxidoreductase family protein
SEQ1551/Spy0478	*scfA*	putative permease
SEQ1552/Spy0477	*scfB*	putative membrane protein
SEQ1576/Spy0436	*vicK*	sensor histidine kinase
SEQ1659/Spy0369	*mtsB*	metal cation ABC transporter ATP-binding protein
SEQ1660/Spy0368	*mtsA*	metal ABC transporter substrate-binding lipoprotein precursor
SEQ1897/Spy0275	*sstT*	sodium:dicarboxylate symporter family protein
SEQ1898/Spy0274	*braB*	putative branched-chain amino acid transport system protein
SEQ1899/Spy0273	*metP*	putative D-methionine transport system permease protein
SEQ1900/Spy0272	*metN*	putative D-methionine transport system ATP-binding protein
SEQ1902/Spy0271	*metQ*	putative lipoprotein
SEQ2191/Spy1823		putative membrane protein

Fifty-seven percent (*n*= 13) of these consensus genes are involved in either proven or putative transport functions, such as uptake of amino acids, metal ions and vitamins (Table 4.13, blue). COG enrichment of the 23 consensus genes identified that the most prevalent categories included amino acid (*n*= 5) and inorganic ion transport and metabolism (*n*= 5) and function unknown (*n*= 3) (Figure 4.14B). The genes *scfA* and *scfB* were classified in the function unknown category, named for their essentiality in the subcutaneous mouse infection model (subcutaneous fitness) [87]. Requirement for these genes was confirmed *in vivo* by the out-competition of a *scfAB* mutant by WT *S. pyogenes* in mice [87].

S. equi must survive in the oropharynx before it disseminates to the lymph nodes in the head and neck. *S. equi* must therefore be able to survive in saliva, a niche which has not to date been investigated in relation to strangles. To predict the genes essential for surviving this niche, the *S. equi in vivo* fitness genes were compared to those identified by a TraDIS screen of *S. pyogenes* serotype M1 strain MGAS2221 in human saliva *ex vivo* [81], and to the genes required by the M1 serotype to cause necrotising myositis.

Comparing the data 3-ways identified 10 genes (null= 0.3 genes) required in all niches by *S. equi* and *S. pyogenes*, 18 genes which are likely employed in the face of equine saliva (null= 0.4 genes), and 51 genes likely required for survival when *S. equi* is in close contact with host tissues (null= 0.7 genes) (Figure 4.15A).

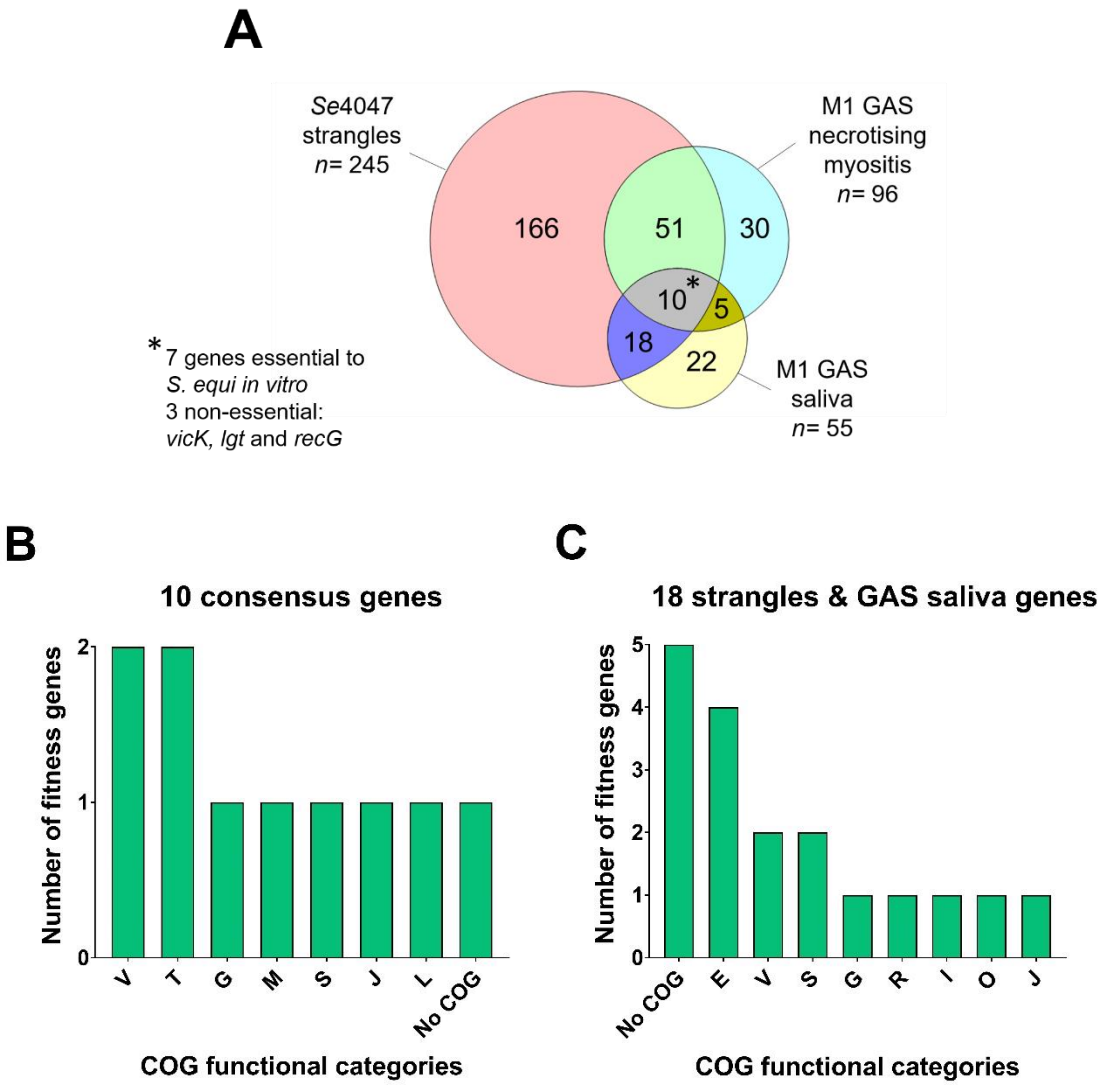


Figure 4.15. Comparison of homologous *S. equi* and *S. pyogenes in/ex vivo* fitness genes and the functional COG categories of the consensus genes.A) Venn diagram comparing the *S. equi* genes required for survival *in vivo* and the *S. pyogenes* genes required for *in vivo* infection of serotype M1 and M28 in a non-human primate model of necrotising myositis and *ex vivo* in human saliva. Ten pan-species consensus genes are required for fitness by *S. equi* and *S. pyogenes* in all 3 niches. Eighteen genes were specifically required by *S. equi in vivo* and by *S. pyogenes ex vivo* in human saliva B) Functional COG categories assigned to the 10 pan-species consensus genes. C) Functional COG categories assigned to the 18 saliva consensus genes. V: Defence mechanisms, T: Signal transduction mechanisms, G: Carbohydrate transport and metabolism, M: Cell wall/membrane/envelope biogenesis, S: Function unknown, J: Translation, ribosomal structure and biogenesis, L: Replication, recombination and repair, E: Amino acid transport and metabolism, R: General function prediction only, I: Lipid transport and metabolism, O: Posttranslational modification, protein turnover, chaperones.

Of the 10 consensus genes required in all 3 niches, 7 were identified as essential/ambiguous in *S. equi in vitro*, but non-essential in *S. pyogenes*, permitting their inclusion in the analysis (Figure 4.15A, Table 4.14). The remaining 3 genes, *vicK, lgt* and *recG* are involved in a 2-component signal transduction system, lipoprotein processing and DNA repair, respectively, with *recG* being confirmed as necessary for survival under stress conditions *in vitro* in *S. equi,* as previously described. COG enrichment of the 10 consensus genes identified that most prevalent categories are involved in defence mechanisms (*n*= 2) and signal transduction systems (*n*= 2) (Figure 4.15B).

COG enrichment analysis of the 18 genes potentially required for survival of *S. equi* and *S. pyogenes* in saliva, revealed that the most prevalent categories included 'no COG' (*n*= 5), highlighting the identification of novel information, and amino acid transport and metabolism (*n*= 4). Included in the 18 genes potentially required for survival in saliva for *S. equi* and *S. pyogenes*, are the genes *sptA* and *carAB* (Figure 4.15A, Table 4.15). *sptA* is a component of a putative ABC transporter system comprising of *sptABC*.

Essentiality of different components of the streptolysin S (*sag*) operon appear to depend on the environmental niche. The ABC transporter components (*sagGHI*) are essential in all 3 studies compared in Figure 4.15A (essential for *S. equi in vitro*), yet *sagEF* are additionally essential for survival in saliva, but non-essential for survival in the NHP model of necrotising myositis.

Table 4.14. Ten consensus genes required for fitness *in vivo* in *S. equi* in the natural equine host*, S. pyogenes* serotype M1 in a non-human primate model of infection and in human saliva *ex vivo.* Genes involved in transport (blue) and genes validated in *S. equi* in the natural equine host with allelic replacement mutants (bordered row).

*Se*4047/MGAS5005 locus tag	Gene	Function
SEQ0005/Spy0004		putative GTP-binding protein
SEQ0454/Spy1519	*recG*	ATP-dependent DNA helicase
SEQ0552/Spy0568	*sagG*	streptolysin S export ATP-binding protein
SEQ0553/Spy0569	*sagH*	streptolysin S export transmembrane protein
SEQ0554/Spy0570	*sagI*	streptolysin S export transmembrane protein
SEQ1537/Spy0485	*lgt*	prolipoprotein diacylglyceryl transferase
SEQ1576/Spy0436	*vicK*	sensor histidine kinase
SEQ1818/Spy1375	*tkt*	putative transketolase
SEQ2191/Spy1823		putative membrane protein
SEQ2205/Spy1837	*gdpP*	phosphoesterase, DHH family protein

Table 4.15.Eighteen consensus genes required for fitness *in vivo* in *S. equi* in the natural equine host and *S. pyogenes* serotype M1 strain MGAS2221 in human saliva *ex vivo.* Genes involved in transport (blue) and genes validated in *S. equi* in the natural equine host with allelic replacement mutants (bordered rows).

*Se*4047/MGAS5005 locus tag	Gene	Function
SEQ0550/Spy0566	*sagE*	CAAX amino terminal protease family protein
SEQ0551/Spy0567	*sagF*	streptolysin S biosynthesis protein
SEQ0628/Spy1242		conserved hypothetical protein
SEQ0647/Spy1226		conserved hypothetical protein
SEQ0683/Spy1139	*nagB*	glucosamine-6-phosphate isomerase
SEQ1003/Spy0993		putative membrane protein
SEQ1004/Spy0992		ABC transporter ATP-binding protein
SEQ1013/Spy0987	*sipC*	putative signal peptidase I
SEQ1025/Spy0973		conserved hypothetical protein
SEQ1137Spy0841		putative peptidase
SEQ1312/Spy0644	*sptA*	putative exported protein
SEQ1313/Spy0643	*carB*	carbamoyl-phosphate synthase large chain
SEQ1314/Spy0642	*carA*	carbamoyl-phosphate synthase small chain
SEQ1432/Spy0722	*miaA*	tRNA delta(2)-isopentenylpyrophosphate transferase
SEQ1536/Spy0486		putative exported protein
SEQ1640/Spy0382	*msrA2*	peptide methionine sulfoxide reductase
SEQ1870/Spy0301		putative membrane protein
SEQ1917/Spy0251	*oppC*	putative oligopeptide transporter permease protein

Genes conferring increased fitness when disrupted by IS*S1*

Of the 85 genes conferring increased fitness in the *S. equi* TraDIS screen, 28 had homologues in *S. pyogenes* serotype M1. When these 28 genes were compared to their homologues in *S. pyogenes* serotype M1 in a murine model of subcutaneous infection, 5 genes were similarly identified as enhancing fitness (null= 0.25 genes). These 5 consensus genes include a phosphoglycerate mutase (COG= carbohydrate transport and metabolism), an ATP-dependent protease subunit (*clpL*) (COG= posttranslational modification, protein turnover, chaperones) and 3 genes not belonging to a COG category; polysaccharide deacetylase, hyaluronoglucosaminidase (*hyl*) and a phage protein (Figure 4.16A). Insertion mutants in *sagB* were similarly identified as enhanced in fitness in both *S. equi* and *S. pyogenes* serotype M1 in an NHP model of necrotising myositis (null= 0.16 genes) (Figure 4.16B), however no consensus genes were identified when serotype M28, from the same NHP study, was considered (Figure 4.16C). Additionally, none of the *S. pyogenes* serotype M1 insertion mutants conferring increased fitness in human saliva were similarly identified in the *S. equi* TraDIS screen (Figure 4.16D).

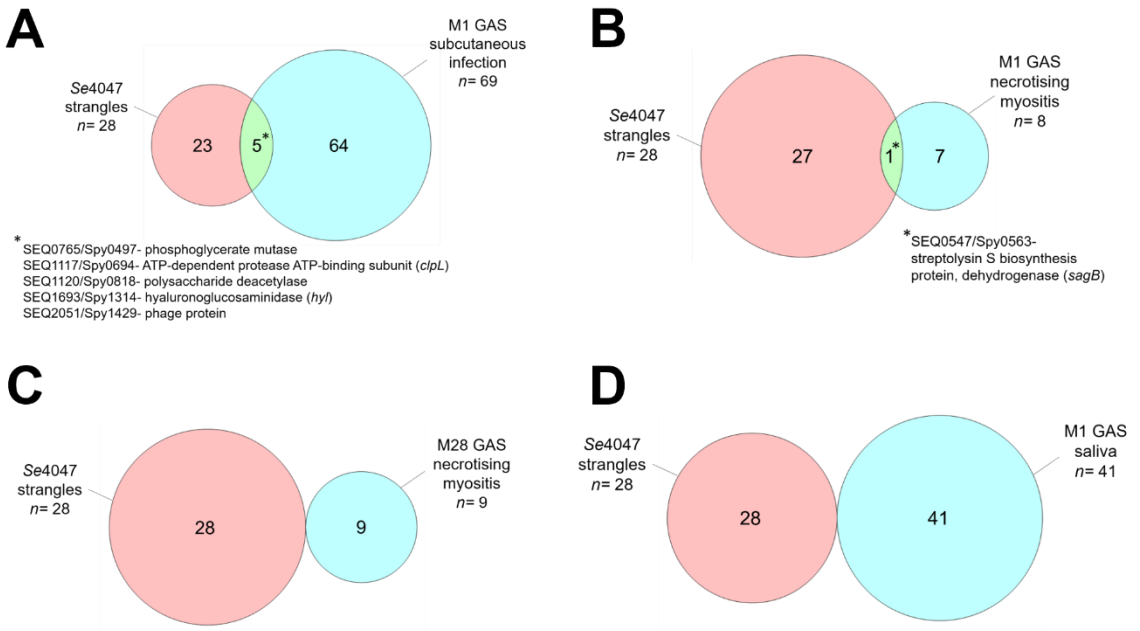


Figure 4.16. Comparison of *S. equi* and *S. pyogenes* genes enhanced in fitness *in vivo* as a result of transposon insertion. A) Venn diagram comparing genes conferring enhanced survival *in vivo* in *S. equi* in the natural host and in *S. pyogenes* serotype M1 in a murine model of subcutaneous infection. Five consensus genes were identified (Asterix). B) Venn diagram comparing genes conferring enhanced survival *in vivo* in *S. equi* in the natural host and in *S. pyogenes* serotype M1 in a non-human primate model of necrotising myositis. One consensus gene was identified (Asterix). C+D) Transposon insertion did not commonly confer enhanced fitness in any genes in *S. equi* compared to either *S. pyogenes* serotype M28 in a non-human primate model of necrotising myositis or to *S. pyogenes* serotype M1 in human saliva *ex vivo*.

4.4 Discussion

In this study, the use of barcoded TraDIS to evaluate the genome-wide fitness of *S. equi* in a susceptible natural host is described. Barcoded TraDIS identified 368 genes required for fitness *in vivo* and 85 genes that conferred a fitness advantage as a result of IS*S1* insertion. Using barcoded libraries also enabled the translocation of mutants from nostril to lymph node to be measured. Each nostril received 1 barcoded library, so the populations of mutants within lymph nodes could be traced back to the nostril in which it was inoculated. For left and right retropharyngeal lymph nodes, 21 and 30 percent of mutants had originated from the nostril on the opposing side of the head, respectively. For the left and right submandibular lymph nodes, 13 and 11 percent of mutants had originated from the opposing nostril. Anecdotally, it was known that *S. equi* forms larger abscesses on the same side of the head that it was inoculated (A. Waller, personal communication, 2018), but this had not been measured and would not have been possible without the library barcodes.

Twelve of the genes required for fitness as per the BC analysis were validated using allelic replacement deletion mutants *in vivo*. All deletion mutants appeared to exhibit the attenuated phenotype as suggested by TraDIS, however, the internal control mutants and wild-type *Se*4047 did not behave as expected, confounding robust statistical analysis. To investigate potential reasons for this, the internal control strains in the inoculum and in the recovered abscess material should be sequenced to confirm no mutations outside of the intended deletion have occurred, potentially causing the differences in fitness seen between the control strains.

Nine of the 12 genes in the *S. equi* deletion mutant validation panel were also found to be important in 1 or multiple of the *S. pyogenes* TraDIS/Tn-seq screens in NHPs, human saliva or mice [81, 87, 145]. Of the remaining 3 genes, 2 do not have homologues in *S. pyogenes* and 1 gene was uniquely attenuated in *S. equi*. Beyond the genes selected for validation, many other genes identified in the *S. equi* TraDIS screen were similarly important to *S. pyogenes in vivo* and *ex vivo.*

TraDIS/Tn-seq-like studies that utilise next-generation sequencing provide a means to simultaneously measure the fitness contribution of every gene in a bacterial genome under the condition tested. *In vivo*, the potential of these techniques is most apparent through the ability to achieve a wealth of data from a greatly reduced number of experimental animals, compared to traditional methods using single whole deletion mutants in isolation. TraDIS/Tn-seq-like tools fulfil the principles of the 3Rs; replacement, reduction and refinement [22], with the barcoded (BC) analysis and validation techniques described in this thesis taking the reduction in animal usage a step further.

BC analysis of the *S. equi* data significantly improved the ability to detect genes affected by IS*S1* insertion *in vivo*. Traditionally, similar studies analyse data on a per animal (PA) basis, treating each animal as a biological replicate. Comparing the BC analysis method to that traditionally used, 379 percent more genes were identified as required for fitness using this novel technique. The increased sensitivity of the barcoded technique can be attributed to the reduction in the impact of animal to animal variation and stochastic loss on the data. The potential effects of stochastic loss have long been a concern for researchers conducting such transposon library *in vivo* studies, especially when bacteria are required to overcome a significant bottleneck where the infection site is distant from the inoculation site. The data presented in this thesis should improve the design of similar studies in the future by mitigating against harsh bottlenecks.

All 97 genes required fitness genes identified in the PA analysis were also identified in the BC analysis. These 97 genes contribute 71 of the top 100 BC genes that exhibited the greatest fitness defects as a result of IS*S1* insertion. Included in these genes are *purN*, *SEQ0402, scfA, metP, sufC, slaB* and *gacI*, all of which were included in the *in vivo* validation panel.

Thirty-nine percent and 60 percent of genes identified in the whole equine blood and H_2O_2 TraDIS screens, respectively, were also identified as required for *in vivo* fitness. Nine consensus genes were identified as required in all 3 experiments. However, an additional 354 genes were uniquely required for *in vivo* fitness, highlighting the much more complex environment that *S. equi* encounters in the natural host and the presence of niche-specific genes.

Comparison of the genes implicated in *in vivo* survival of *S. pyogenes* in a NHP model of necrotising myositis and in a mouse model of subcutaneous infection and *S. equi* in the natural equine host, uncovered a set of 23 pan-species consensus genes that are potentially important for future development of novel therapeutics and vaccines. The *S. pyogenes* NHP model study, utilised 2 serotypes; M1 and M28, that have caused invasive infections in many countries. The M1 strain, MGAS2221 is genetically representative of a pandemic clone that became pandemic in the 1980, spreading globally and remains the most prevalent cause of severe infections worldwide [197-199]. The M28 strain, MGAS27961 is genetically representative of a clone that is prevalent in America and other countries [200]. The *S. pyogenes* murine model of subcutaneous infection also utilised a M1 strain, MGAS5448, which is genetically representative of the globally identified invasive M1 strains [201]. Genes conferring an enhanced as a result of insertion were also compared between these studies, but fewer similarities were identified.

The experimental methods used to induce infection in these studies are distinct from one another and so the environmental pressures exerted on the transposon mutants are varied. In the NHP model study, dense IS*S1* libraries in both strains were injected directly into the thighs of 6 cynomolgus macaques and recovered after 24 hours, therefore cells were not required to translocate to a distant site of infection [145]. Seventy-two fitness genes were commonly identified between the 2 serotypes of *S. pyogenes*. In the murine model of *S. pyogenes* subcutaneous infection, the MGAS5448 library was also injected directly into the infection site, but was left to form abscesses for either 24 or 48 hours [87]. At 24 and 48 hours, 75 and 106 genes were required for fitness, respectively, with insertions in 147 genes conferring reduced fitness at both or either timepoint [87].

S. equi IS*S1* mutants were sprayed intranasally into ponies, where mutants are required to translocate through the nasopharynx and into the local lymph nodes. Despite this pressure, the identification of a consensus set of 23 genes required by streptococci in all 3 varied niches is highly encouraging. A comparison between the M1 serotype NHP and the mouse model data identified 39 consensus genes required for *S. pyogenes* fitness *in vivo* [145]. All 39 genes had homologues in *S. equi* and therefore 59 percent of the M1 NHP and mouse consensus genes are similarly important to the genetically distinct and host-restricted *S. equi*.

The M1 serotype NHP data was also compared to genes required for survival of the IS*S1* mutants in the same *S. pyogenes* serotype M1 strain in human saliva *ex vivo* [81]. Only 19 genes were commonly identified, highlighting that fitness genes are niche specific. A comparison was made between the *S. equi in vivo* TraDIS screen data and these 2 M1 datasets, since *S. equi* has to survive in the presence of saliva before it can invade the nasopharynx. Ten genes were identified as important in all 3 datasets, with 18 genes specifically required in the *S. equi in vivo* and *S. pyogenes* saliva data. A further 51 genes were similarly identified between the M1 NHP data and the *S. equi* data, highlighting the genes required by both species to survive *in vivo* and form viable abscesses/lesions.

Genes conferring increased fitness as a result of transposon insertion were also compared. None of the genes conferring an enhanced fitness in the *ex vivo* human saliva study or for the M28 serotype library in NHPs, were similarly enhanced in the *S. equi* data. Five genes were similarly enhanced due to insertion in the M1 serotype in the murine subcutaneous model, compared to *S. equi* in ponies. Compared to the mutants enhanced in fitness in the M1 serotype in NHPs, only 1 gene was similarly enhanced in fitness in ponies.

The specifics of these consensus genes, whether transposon insertion conferred increased or decreased fitness are explored in more detail in the course of this section.

4.4.1 Genes required for fitness as identified by barcoded TraDIS

The importance of transport genes in streptococci

In NHPs, 25 percent of *S. pyogenes* genes necessary for infection were known or putative transporters [145]. In *S. equi*, 8.7 percent of genes conferring a decrease in fitness as a result of IS*S1* insertion, were involved in transport. Of the 23 pan-species consensus genes, 57 percent encoded proven or putative transporters. The abundance of transporter genes in these analyses suggest that utilising and communicating with the extracellular environment is extremely important for virulence in streptococci. The presence of ABC transporters concerned with the import of essential ions and cofactors suggests that 'scavenging' these compounds may be favoured in streptococci, over intracellular biosynthesis by means of conserving energy. Many live bacterial vaccines have focused on disrupting biosynthetic pathways to attenuate the target strain. Genes in the aromatic acid biosynthesis pathway have received considerable attention as vaccine targets in species such as *S.* Typhimurium, *Aeromonas salmonicida*, *Pasteurella multocida*, *E. coli, S. suis* and *S. equi* [57, 202-210]. The data presented in this Chapter suggests that deleting transporter genes may enhance the safety of live attenuated vaccines. Furthermore, many transporter systems utilise a surface-exposed component that could be targeted by multicomponent subunit vaccines. Targetting conserved transport systems could also be a productive avenue in the development of new antibacterial agents against antimicrobial resistant pathogens.

Acquisition of NAD precursors

COG enrichment of the BC analysis revealed that 35 percent of genes required for infection did not belong to a defined COG category, highlighting potentially novel information. The gene *niaX* (*SEQ0658*) was included in this category and is involved in the import of nicotinamide and nicotinic acid from the extracellular environment, for conversion into nicotinamide adenine dinucleotide (NAD). NAD is a necessary cofactor for all living cells and is utilised in respiration, in the conversion of aldehydes to alcohols and in DNA ligation and repair, amongst other basic cellular processes [211-214]. NAD synthesis is efficiently regulated, with 2 main pathways being used to source NAD, depending on the bacterial species. The important cofactor is either synthesised *de novo,* as is undertaken in species such as *Escherichia* [215], or by importing precursors, such as nicotinamide and nicotinic acid, from the environment and converting them to NAD via the salvage pathway, as is undertaken in streptococci [216, 217] (Figure 4.17). In *Streptococcus pneumoniae* (*S. pneumoniae*), a *niaX* deletion mutant was unable to import radio-labelled nicotinamide and nicotinic acid *in vitro*, a process which was restored upon complementation [218]. *S. pneumoniae* can additionally import other NAD precursors; nicotinamide mononucleotide and nicotinamide riboside, the latter (and

potentially the former) via the importer, PnuC [218]. Mice were intranasally infected with the *S. pneumoniae* Δ*niaX* mutant and a Δ*pnuC* mutant. The Δ*pnuC* strain was completely attenuated compared to wild-type *S. pneumoniae* strain and is therefore required for *in vivo* survival [218]. The Δ*niaX* strain, however, behaved as wild-type and was therefore dispensable *in vivo* [218].

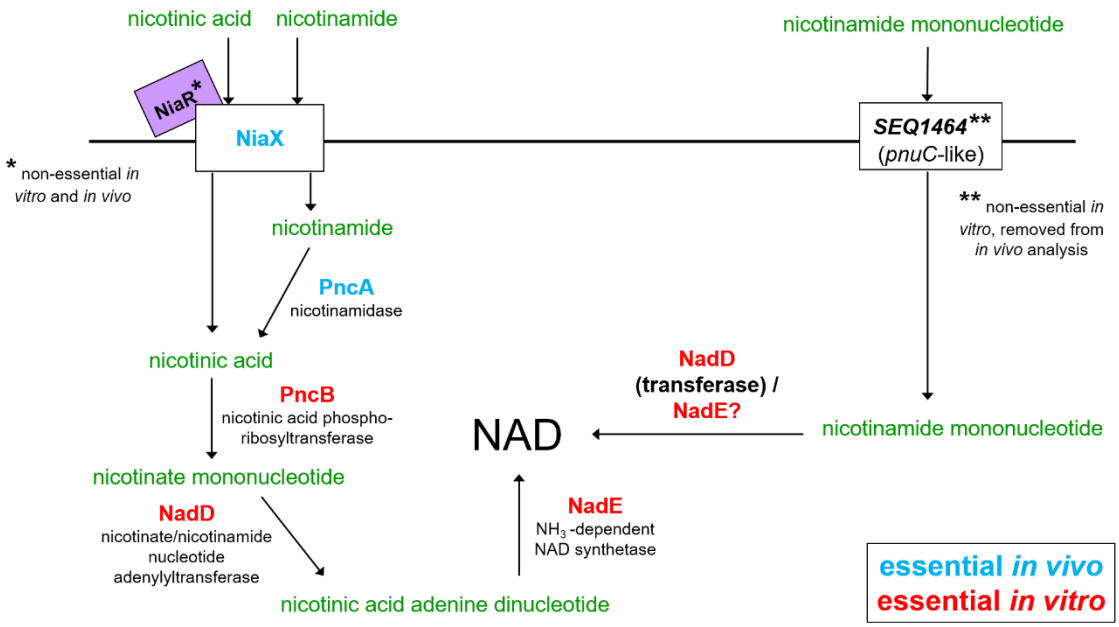


Figure 4.17. Schematic diagram of the putative import systems used by *S. equi* to acquire NAD precursors and convert them into NAD. Genes essential to *S. equi in vitro* in THB are indicated by red text and genes required during infection of the natural host are indicated in blue.

S. equi encodes a *pnuC*-like gene, but has no *pnuC* homologue, and may therefore rely on NiaX for NAD precursor scavenging. Downstream enzymes required for the conversion of the imported precursors into NAD are essential to *S. equi in vitro,* except for *pncA*, the necessity for which could be bypassed by utilising imported nicotinic acid (Figure 4.17). In *S. equi*, *niaX* was dispensable *in vitro*, but highly attenuated *in vivo* (log_2FC= -5.4, q = 0.0008). *In vitro*, *S. equi* may be able to readily acquire the NAD precursors from the rich THB medium by other means. *SEQ1464* encodes a *pnuC*-like transporter which imports nicotinamide mononucleotide, that is subsequently converted into NAD by NadD (Figure 4.17). No homologies to *SEQ1464* exist in *S. pneumoniae*, yet the species is able to import the same precursor, likely by other means. *SEQ1464* was removed from the *in vivo S. equi* TraDIS screen data as it contained too few reads to meet the stringent inclusion criteria, so no assessment of its essentiality in infection can be made. The importance of NiaX *in vivo* indicates that it is a vital importer of NAD precursors *in vivo* and that other importers may not be able to compensate for the lack

of NiaX. The importance of *niaX in vivo* is supported by the decreased fitness of transposon mutants in the *S. pyogenes* subcutaneous murine Tn-seq screen [87]. Interestingly, the transcriptional regulator of *niaX, niaR* (SEQ0657), located 395 bp upstream of *niaX* in *Se*4047, is non-essential *in vitro* and *in vivo*. *niaR* is a transcriptional repressor which acts on an operator site in the promoter region of *niaX* in *S. pneumoniae*, repressing *niaX* in the presence of niacin [217, 219]. The close proximity of these 2 genes in *S. equi* would suggest that they function similarly to that in *S. pneumoniae*. The dispensability of *niaR* would suggest that over-import of NAD precursors via *niaX* is well tolerated in *S. equi,* but does not elicit an increase in fitness. NiaR was similarly identified as non-essential *in vitro* or in the mouse subcutaneous model of infection with *S. pyogenes* [87].

Export of quorum sensing peptides

COG enrichment of the BC data revealed that 11 genes required for infection belonged in the 'defence mechanisms' category. The quorum sensing peptide transporter PptAB (a.k.a. EcsAB) was amongst these genes (*pptA:* $\log_2$FC= -5.6, *q*= 0.00031 and *pptB:* $\log_2$FC= -5.4, *q*= 0.00031). *pptAB* comprise 2 of the 23 consensus genes required for *S. equi* and *S. pyogenes* in all 3 niches compared. PptAB were also required for *S. equi* fitness in whole equine blood, as determined in Chapter 3. As previously described in Chapter 3 section 3.4.2, *pptAB* encode ABC transporter proteins that export the quorum sensing peptides, SHP2 and SHP3, into the extracellular environment [186]. A *S. aureus* Δ*pptAB* deletion mutant was attenuated in a murine model of arthritis, causing milder synovitis and reduced bone erosions [187]. The Δ*pptAB* strain was also significantly reduced in its ability to persist in the kidneys in later stages of infection [187]. Transcriptome analysis of genes expressed in NHP muscle tissue infected with wild-type *S. pyogenes* serotype M1 and in an infected human patient, detected transcription of *pptA*, confirming its expression *in vivo* [145].

Studies in *S. pyogenes* have shown that genes for 2 transcriptional regulators, Rgg2 and Rgg3, are adjacent to small open reading frames that encode the 2 quorum sensing peptides exported by PptAB, SHP2 and SHP3 [220, 221]. Rgg2 and Rgg3 were found to each control transcription of the promoters driving the *shp* genes, but control this with opposite effects. Rgg2 appears to be a transcriptional activator, being inactive and not bound to the *shp* promoters, until external SHPs are imported into the cell [220]. Rgg3 on the other hand, is a transcriptional repressor, binding to DNA to block transcription until SHPs are present [220]. *SEQ0653* of *S. equi*, simply annotated as a putative DNA-binding protein, has 32 percent amino acid identity with *rgg2* and 30 percent to *rgg3*. No other candidate *rgg* genes in *S. equi,* potentially involved in this quorum sensing pathway, seem to be present. *S. equi* IS*S1* insertion mutants in *SEQ0653* are non-

essential both *in vitro* and *in vivo*, but were trending towards being enhanced in fitness in ponies (log_2FC= 3.4, *q*= 0.1). In *Se*4047, *SEQ0653* may therefore behave more like the transcriptional repressor Rgg3, since disruption of this gene with IS*S1* has not negatively affected the quorum sensing pathway, yet disruption of the quorum sensing peptide transporter proteins, *pptAB*, have proven consistently detrimental in both *S. pyogenes* and *S. equi* in a number of environments. The essential nature of *pptAB* across all 3 TraDIS/Tn-seq screens compared in this thesis, and for *S. equi* in whole equine blood, provides valuable evidence of the repeatability and quality of data produced by such *in vitro/vivo* transposon studies and the potential of these 2 genes as vaccine targets. In support of the potential usefulness of *pptAB* in future vaccine design, a Δ*pptAB* remained viable in the NHP model of necrotising myositis, but lesion size was significantly reduced.

Import of extracellular zinc

The gene *adcB* comprises 1 of the 23 consensus genes required for *S. equi* and *S. pyogenes* infection across all 3 niches compared. *adcB* is located in an operon with *adcC* that encodes the inner membrane permease (AdcB) and the cytosolic ATPase (AdcC) that provides energy in the form of ATP, powering zinc import [222]. It is thought that this import system is the primary method of zinc import in GAS, which also employs a cell-surface zinc binding protein, AdcA, to assist the acquisition of extracellular zinc [222]. AdcB insertion mutants in *S. equi* were significantly attenuated in ponies (log_2FC= -6.5, *q*= 0.00017), with AdcC tending towards attenuation, just missing significance (log_2FC= -3, *q*= 0.08). In *S. equi* in ponies and *S. pyogenes* in NHPs, AdcA insertion mutants were unaffected *in vivo*, however, AdcA mutants were attenuated in *S. pyogenes* in mice [87]. Therefore, it appears that in *S. pyogenes* in a NHP model of necrotising myositis, *adcBC* can import zinc without a functioning *adcA*. The same is likely to be true of *S. equi* despite the non-significance obtained for *adcC* mutants. This notion is potentially supported by the arrangement of these genes; *adcBC* as an operon (*SEQ0095* and *SEQ0094*), with *adcA* (*SEQ0861*) encoded separately at a location considerably distant from *adcBC* [3].

In the dental plaque pathogen, *Streptococcus gordinii*, mutation of *adcB, adcC* and *acdR* significantly impaired its ability to form biofilms. An *adcA* mutant however, was able to generate comparable biofilms to the wild-type strain [223].

In *S. agalactiae*, the repressor protein, AcdR, regulates intracellular zinc homeostasis and is a member of the MarR family of regulators that use metal ions as co-repressors [224]. AdcR controls the adaptive responses to fluctuating zinc concentrations by, for example, enabling the transcription of zinc acquisition systems, such as that encoded by *adcABC* [224, 225]. When the intracellular concentration of zinc is adequate, AdcR is

bound to zinc, which causes a conformational change in AdcR enabling DNA binding, repressing transcription of target genes [224, 225]. When insufficient zinc is available, AdcR cannot bind to the target genes, causing redepression [224, 225]. *S. pyogenes* requires *adcR* for infection in NHPs [145]. Unfortunately, this gene was removed from the *S. equi* TraDIS screen analysis because it lacked sufficient reads in the input pools, which does somewhat imply that *S. equi* may require *adcR*. It is possible that *adcR* functions in a similar way to the equibactin regulator, *eqbA*, in *S. equi*, where unregulated import of the ion causes cell toxicity. In contrast, in the murine model, *acdR* transposon mutants in *S. pyogenes* were enhanced in fitness [87].

Export of streptolysin toxin

Streptolysin S (SLS) is an extracellular toxin produced by both *S. equi* and *S. pyogenes*, which causes the characteristic zone of β-haemolysis observed surrounding colonies grown on blood agar [29] and destroys host cells of many types [30]. SLS degrades host cells and may contribute to immune evasion, and/or nutrient acquisition. SLS biosynthesis and transport genes are encoded by the operon *sagA-I* (Figure 4.18). The ABC transporter proteins encoded by *sagGHI* are essential to *S. equi in vitro*, which is likely to be due to the toxic effects of retaining SLS within the bacterial cell, through lack of export. SagGHI, were identified as non-essential *in vitro* to *S. pyogenes* [78, 81, 87, 145]. However, after 24 and 48 h passage in THB, *sagGHI* mutants of *S. pyogenes* were all classified as essential [87]. Incubation of a Δ*sagH* deletion mutant in THB confirmed the Tn-Seq findings [87]. The *S. equi* libraries used to determine essential genes in this thesis were grown for approximately 3 hours after 16 h growth on TH agar. Slight differences in library growth may account for these differences.

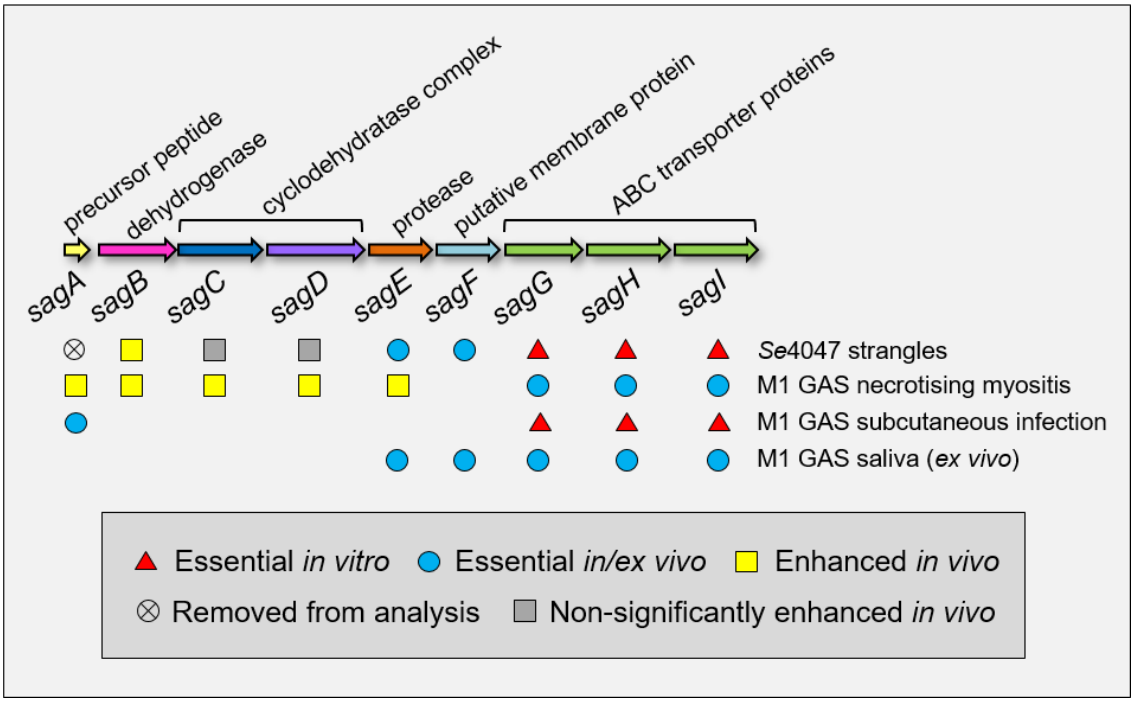


Figure 4.18. Essentiality of streptolysin S genes in *S. equi* and *S. pyogenes.* Transposon mutants in the streptolysin S biosynthesis genes *sagA-sagE* were enhanced in fitness in *S. pyogenes* serotype M1 in the non-human primate model of necrotising myositis. This was reflected in *S. equi* in the natural host for *sagB*, with mutants in *sagCD* missing significance, but trending towards increased fitness. *sagEF* were required for infection *in vivo* in *S. equi* and for survival of *S. pyogenes ex vivo* in human saliva. The streptolysin ABC transport genes *sagGHI*, are consistently identified as important, whether it is *in vitro* or *in vivo*, likely due to the toxic effects of retaining streptolysin S intracellularly.

In *S. equi*, *sagEF* IS*S1* mutants were reduced in fitness *in vivo* (*sagE;* $\log_2$FC= -3.4, *q*= 0.03, *sagF;* $\log_2$FC= -3.5, *q*= 0.03) (Figure 4.18). *S. pyogenes* serotype M1 also required *sagEF* along with *sagGHI* for survival in human saliva [81]. Both the M1 and M28 serotypes required *sagGHI* in NHPs [145] (Figure 4.18). In contrast, insertion into the *sagB-F* genes, did not have an effect on the *in vivo* fitness of *S. pyogenes* serotype M1 in the mouse model of subcutaneous infection [87] (Figure 4.18).

Interestingly, insertional mutagenesis of the *sag* biosynthesis genes, *sagABCDE*, enhanced the fitness of *S. pyogenes* in NHPs [145]. The same was evident for *sagB* in *S. equi in vivo* ($\log_2$FC= 5, *q*= 0.04) (Figure 4.18). Fitness of IS*S1* mutants in *sagCD* were also enhanced, but did not meet statistical significance in *S. equi* (*sagC;* $\log_2$FC= 3.7, *q*= 0.1, *sagD;* $\log_2$FC= 4.2, *q*= 0.08) (Figure 4.18). *sagA* was removed from the *S. equi* dataset as it contained too few reads to meet the inclusion criteria (Figure 4.18). In support of these fitness benefits, a non-β-haemolytic strain of *S. pyogenes* was isolated from a particularly severe case of human soft tissue infection, which contained a premature stop codon in *sagC* [31]. It is not clear why mutants defective in SLS biosynthesis may cause more severe disease in *S. equi*. It may be that inactivating SLS

enhances survival in phagocytes, improving migration to the lymph nodes. Conversely, mutants in *sag* biosynthesis genes may be able to benefit from neighbouring IS*S1* mutants that have retained their ability to produce SLS. 'Piggy backing' off neighbouring cells may negate the need to synthesise the potentially energetically expensive SLS, improving the fitness of *sag* biosynthesis mutants. The attenuation of *sag* deletion mutants in the majority of the current literature may support this idea, as this enhanced fitness effect may only be evident in a mixed population of mutants. IS*S1* mutants were not however enriched *in vitro*, suggesting that this potential 'piggy backing' effect may only be evident in nutrient deficient environments.

Known *S. equi* virulence factors not identified by TraDIS

As was described in Chapter 3 section 3.4, measuring fitness of mutants in a mixed population can confound the identification of genes that encode known virulence determinants, particularly if they are secreted. In this *in vivo* study, genes encoding the hyaluronic acid capsule, superoxide dismutase, equibactin, IdeE, IdeE2, SeCEP, superantigens and fibronectin binding proteins were non-essential. IS*S1* mutants in these genes are likely to retain their fitness in the presence of other mutants still capable of producing these virulence factors.

4.4.2 Genes required for fitness included in *S. equi in vivo* validation panel

Five 2.5-year-old Welsh mountain ponies were challenged with a mixed dose of 12 tagged deletion mutants predicted to be attenuated *in vivo*, 3 tagged internal control (IC) strains and wild-type *Se*4047. Tagged allelic replacement mutagenesis was utilised to generate the mutants, whereby 1 of 3 80 bp tags containing the TraDIS PCR and sequencing primer binding site, was integrated into the target gene deletion site. Presence of the tag enabled simultaneous measurement of all mutants by TraDIS. Each IC strain contained a deletion in *SEQ0751*, a gene unaffected by IS*S1* insertion in ponies and 1 of the 3 tags, to confirm that the tags did not alter fitness. It was predicted that the 3 IC strains would behave as wild-type, however, none of the IC strains or wild-type were consistently recovered from the animals.

Challenge studies are routinely conducted in ponies of approximately 1 year old, generating reliable and replicable results. The 5 2.5-year-old ponies used in the validation were from an excess stock of animals that were not suitable for rehoming, due to behavioural or congenital health concerns. To minimise unnecessary animal wastage, they were used in this validation study. Their age, however, may have confounded the findings of the experiment. Three of the 5 animals did not show obvious clinical signs of disease and were euthanised to minimise suffering, 4 days later than the 2 ponies

showing clear signs of disease. Lymph nodes were still recovered from 3 clinically healthy ponies and any surviving *S. equi* sequenced. The reduced severity of disease in these 3 ponies was reflected in the significantly lower bacterial loads recovered from the lymph nodes. A possible explanation for the inconsistency seen between animals could be that older animals may have a more mature immune system that is able to mount a response to *S. equi*. Older animals are also more likely to have been exposed to *S. zooepidemicus*, a commensal bacterium in horses for which animals may have seroconverted. Antibodies to *S. zooepidemicus* may be active against *S. equi* since the 2 species are so closely related.

Despite the issues faced in the validation study, all 12 deletion strains appeared to be attenuated, however, no robust statistical analysis comparing these mutants to the IC or wild-type strains could be completed. No TraDIS reads were detected for the deletion mutants *ΔpurN*tagB, *ΔSEQ0402*tagA, *ΔSEQ1536*tagA, *ΔscfA*tagA or *ΔsufC*tagB in the recovered abscess materials. The strains *ΔrecG*tagB, *ΔgacI*tagC, *ΔsptA*tagB, *ΔdltB*tagC and *ΔmetP*tagB were detected at very low levels, being represented by ≤ 11 reads each across all 5 animals. More reads were sequenced for the strains *ΔSEQ1410*tagC and *ΔslaB*tagC (1,330 and 4,133 reads, respectively), where the vast majority of reads were collected from only 1 animal (99.9 and 99.85 percent, respectively).

Basic cellular processes

Purine biosynthesis

No reads corresponding to the Δ*purN* mutant were sequenced in any recovered abscess material from the TraDIS validation study, inferring its importance in *S. equi* infection. *purN* encodes a phosphoribosylglycinamide formyltransferase that catalyses the transfer of a formyl group from 10-formyltetrahydrofolate to 5-phospho-ribosyl-glycinamide (GAR), resulting in 5-phopho-ribosyl-N-formylglycinamide (FGAR) and tetrahydrofolate (Figure 4.19). This reaction forms part of the IMP biosynthesis pathway, generating IMP which is the first compound to contain a complete purine ring system in the purine metabolism pathway. The potential inability of the Δ*purN*tagB mutant to cause disease supports the TraDIS screen findings, which identified *purN* IS*S1* mutants as those conferring the greatest fitness defect (log_2FC = -14.7, *q*= <0.00001, Figure 4.20, red arrow).

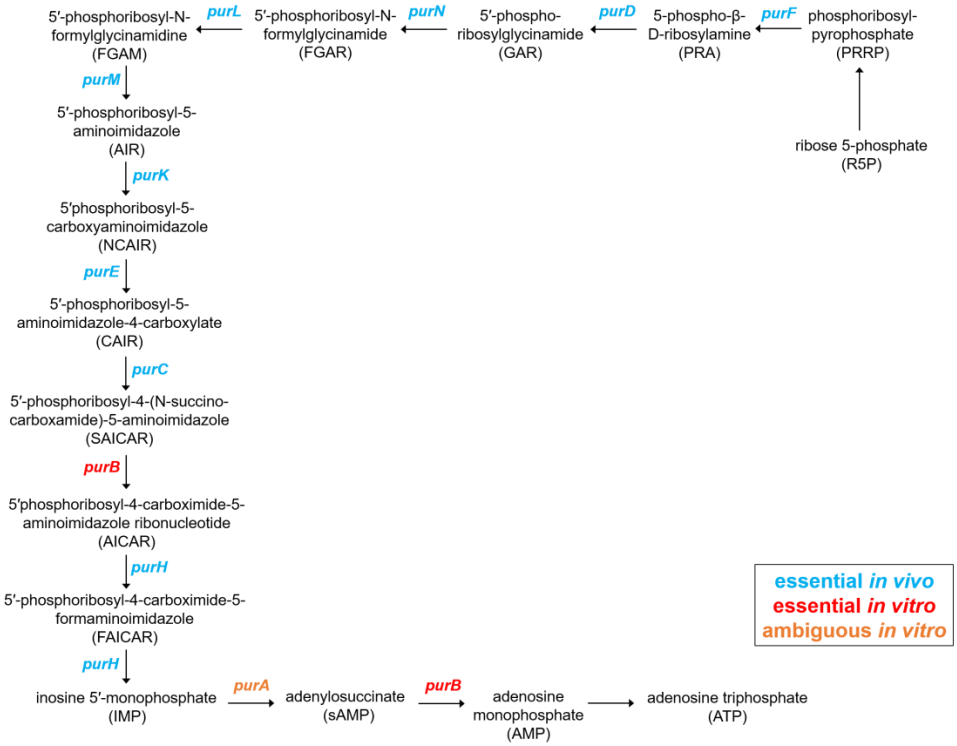


Figure 4.19. Schematic diagram of the putative functions of purine metabolism genes in *S. equi*. Genes essential to *S. equi in vitro* in THB are indicated by red text, genes with ambiguously defined essentiality *in vitro* in THB are indicated in orange and genes required during infection of the natural host are indicated in blue. Adapted and redrawn from [226].

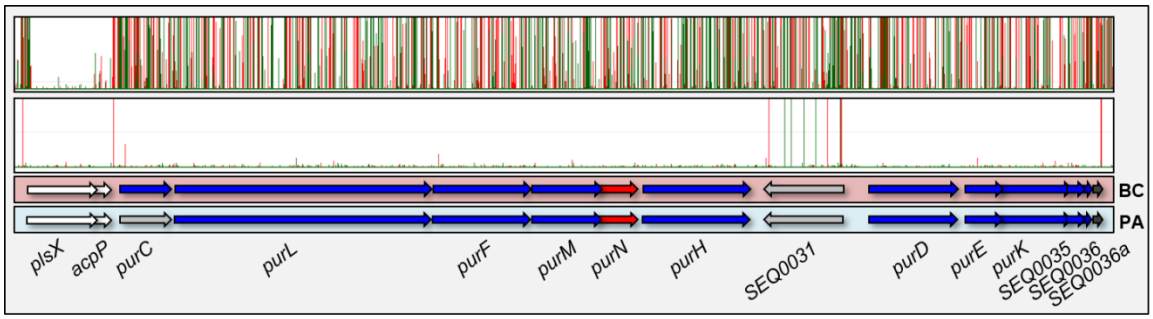


Figure 4.20. Prevalence of *S. equi* ISS*1* mutants in the purine locus pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. Essentiality assigned to these genes by the barcoded analysis is highlighted with a pink box with essentiality assigned by the per animal analysis highlighted by a blue box. The purine metabolism genes *purC-purK* are required for infection according to the barcoded analysis. The essentiality of *purC* is not reflected in the per animal analysis. White arrows indicate essential genes in THB *in vitro.* Blue arrows indicate genes in which insertion significantly reduced in fitness, red arrows indicate genes included in the validation panel, light grey arrows indicate non-essential genes *in vivo*. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

purN is a member of a locus containing several purine biosynthesis genes (Figure 4.20, *purB* (*SEQ0037*) was essential *in vitro* and is not included in the figure). Two other *pur* genes are located elsewhere in the genome; *purA* (*SEQ2113*), essentiality of which was ambiguous *in vitro*, and *purR* (*SEQ0338*), which was non-essential *in vitro*. All *pur* genes in Figure 4.20 contained many reads in the input pools, but were essential for *in vivo* infection in the BC data. In the PA analysis, *purC* was attenuated, but not significantly (log_2FC= -3.25, *q*= 0.2, Figure 4.20, grey arrow), demonstrating an example of when the BC analysis method improved the sensitivity of statistical analyses. The *pur* locus contains *SEQ0031*, a putative amidase between *purH* and *purD*, which was not attenuated *in vivo* as a result of IS*S1* insertion (Figure 4.20). The retained virulence of *SEQ0031* despite its position amongst genes essential for *in vivo* infection demonstrates the precision and sensitivity of TraDIS, as does the clear presence of TraDIS reads in the intergenic regions amongst some of these important genes.

The importance of *purN in vivo* is not well studied. Currently, the literature only describes 1 study investigating this critical gene in an animal model. A Δ*purN* deletion mutant of *S.* Typhimurium was evaluated in a murine model of systemic disease [227]. No *purN* mutants were recovered from mice when challenged in equal proportions with wild-type *S*. Typhimurium.

Attenuation of the *purN* IS*S1* mutants was unique to *S. equi in vivo*. In mice, only *purM* mutants in *S. pyogenes* were significantly attenuated [87]. In NHPs, *purA*, *purB* and *purR* mutants were reduced in fitness in both the M1 and M28 *S. pyogenes* serotypes [145]. *purA* and *purB* were ambiguous and essential to *S. equi in vitro*, respectively, but *purR*, the *pur* operon repressor, was non-essential *in vitro* and *in vivo*. Despite differences in specific gene essentiality within the *pur* genes, the importance of this locus is evident *in vivo* in streptococci. When grown *in vitro*, libraries are in a purine rich environment. The non-essentiality of the majority of *pur* genes suggests that cells are able to scavenge purines, negating the requirement for internal purine biosynthesis. The essentiality of *purAB, in vitro* in *S. equi* and *in vivo* in *S. pyogenes* in NHPs is supported by their involvement in the conversion of inosine 5'-monophosphate (IMP) to adenosine triphosphate (ATP) (Figure 4.19).

DNA replication and repair

The importance of the ATP-dependent DNA helicase, RecG, *in vivo*, is not well studied. In *S. equi,* RecG is required for survival in whole equine blood and H_2O_2 *in vitro* as described in Chapter 3, and in ponies. The Δ*recG* deletion mutant described in Chapter 3, however, grew significantly slower than the wild-type *Se*4047 strain in THB, suggesting that it is predisposed to identification by TraDIS in 'stressful' conditions. In

support of this, *recG* was also consistently identified as important in *S. pyogenes* in the murine model of subcutaneous infection, NHPs model of necrotising myositis and *ex vivo* in human saliva [81, 87, 145]. The PA analysis did not detect *recG* IS*S1* mutants as significantly reduced in fitness, however, highlighting the increased sensitivity of the BC analysis method (Figure 4.21).

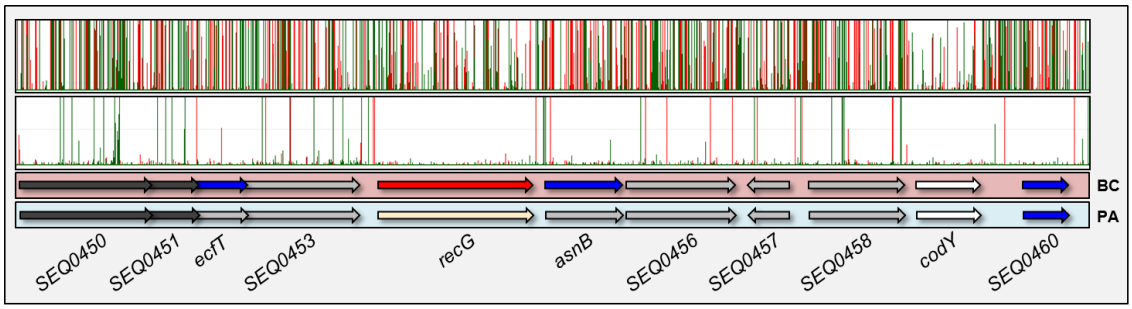


Figure 4.21. Prevalence of *S. equi* IS*S1* mutants in *SEQ0450-SEQ0460* pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand, respectively. Essentiality assigned to these genes by the barcoded analysis is highlighted with a pink box with essentiality assigned by the per animal analysis highlighted by a blue box. RecG is required for infection according to the barcoded analysis. The essentiality of *recG* is not reflected in the per animal analysis, indicatd by the cream arrow. White arrows indicate essential genes in THB *in vitro.* Blue arrows indicate genes in which insertion significantly reduced in fitness, red arrows indicate genes included in the validation panel, light grey arrows indicate non-essential genes *in vivo*. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

The Δ*recG*tagC deletion mutant in the TraDIS validation panel appeared to be attenuated in ponies, with sequencing reads corresponding to this mutant equating to only 0.0014 percent of all reads sequenced in the output pools. Despite this apparent attenuation, due to the slow growth phenotype of the Δ*recG* deletion mutant *in vitro*, RecG, does not present as an ideal target with which to attenuate live vaccines.

Iron-Sulphur immobilisation and transport

Iron-sulphur (Fe-S) proteins are required for a range of functions such as DNA synthesis and repair, electron transport, substrate binding and activation of dehydratases, and are therefore critical for the normal functioning of cells [228, 229]. Fe-S clusters are hypothesised to release iron and sulphur from storage within the cell, assemble them into the cluster and transfer them to the accepting protein [230, 231]. The genes encoding for these Fe-S cluster proteins are arranged in a *suf* operon (*sufCDSUB*) in most Gram positive bacteria [232, 233]. Little research has been conducted regarding

this locus in Gram positive bacteria, but in *E. coli,* the *suf* system is induced under stress conditions such as oxidative stress and iron starvation [234, 235]. No members of the *suf* operon, however, were required for survival in H_2O_2 in *S. equi*, yet the entire locus (SEQ1926-SEQ1930) was indispensable *in vivo* in the TraDIS screen (Figure 4.22), with a Δ*sufC* deletion mutant unable to cause disease in the TraDIS validation study (0 reads corresponding to Δ*sufC* sequenced from any animals, Figure 4.12, Table 4.12).

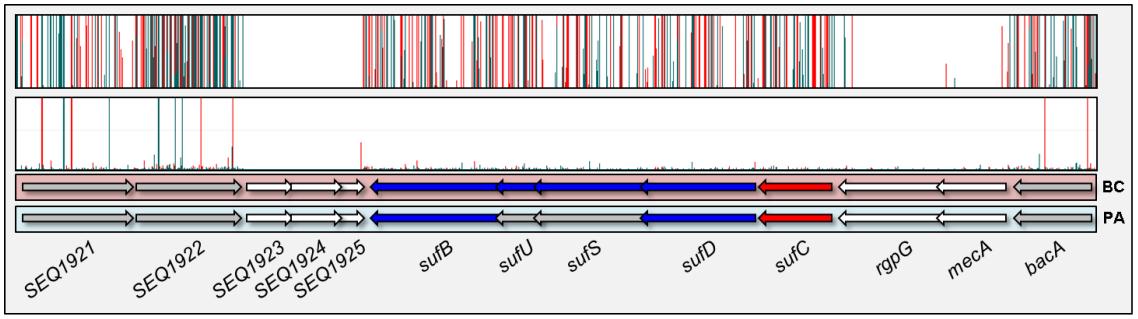


Figure 4.22. Prevalence of *S. equi* IS*S1* mutants in *SEQ1921-SEQ1933,* which includes the *suf* operon pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. Essentiality assigned to these genes by the barcoded analysis is highlighted with a pink box with essentiality assigned by the per animal analysis highlighted by a blue box. The whole *suf* operon is required for infection according to the barcoded analysis. The essentiality of *sufUS* was not identified in the per animal analysis. White arrows indicate essential genes in THB *in vitro.* Blue arrows indicate genes in which insertion significantly reduced in fitness, red arrows indicate genes included in the validation panel, light grey arrows indicate non-essential genes *in vivo*. Data is viewed in Artemis [112].

sufC was also required for subcutaneous *S. pyogenes* infection in mice, with transposon mutants in the remaining *suf* genes retaining virulence [87]. It is possible that the demand for iron is higher in *S. equi* than *S. pyogenes*, increasing the dependence on functioning Suf proteins able to release and utilise stored iron. The potential dispensability of these genes in *S. pyogenes* is supported by the survival of *suf* IS*S1* mutants in NHPs [145] and the presence of the unique equibactin locus in *S. equi*, which as previously described, is required for extracellular iron acquisition.

E. coli has a second Fe-S cluster system, encoded by *isc* genes, which serve as a more essential basic system required for the transfer of Fe-S to important enzymes [236, 237]. Deletion of the *isc* operon caused some growth defects. However, additional deletion of the *suf* operon proved lethal, suggesting that there is some functional redundancy between these operons in *E. coli* [235, 238]. It is likely that the Suf proteins in Gram positive bacteria are solely responsible for the maturation of Fe-S accepting proteins and

enzymes in the absence of an *isc* system. The non-essentiality of the *suf* operon *in vitro* suggests that this locus is, however, only essential in environmentally stressful conditions in *S. equi*.

Membrane/cell surface anchored proteins

Putative *S. equi* specific surface anchored protein

The gene *SEQ0402* has no homology with any other characterised proteins and is therefore unique to *S. equi*. *SEQ0402* is an antigenic LPXTG cell surface anchored protein, which contains an N-terminal non-repetitive domain. The N-terminal domain of *SEQ0402* was included as a component of the developmental strangles vaccine, Septavac, and was shown to elicit a significant IgA response in Welsh mountain ponies [24]. The Septavac vaccine was developed further, becoming Strangvac, where *SEQ0402* was fused to *SEQ0256*, forming Eq85. Other recombinant proteins were included in Strangvac, as outlined in Chapter 1 section 1.4. In ponies, 2 weeks after the 3rd vaccination with Strangvacc, specific IgG antibodies against recombinant SEQ0402 in serum and mucosal secretions were significantly increased compared to pre-vaccination [25]. *SEQ0402* IS*S1* mutants were significantly attenuated in ponies (log_2FC of -6.3, q = 0.0006), which is reflected in the validation study where no reads corresponding to *SEQ0402* were identified in the output pools (Figure 4.23).

These data suggest that *SEQ0402* plays an important role in infection and since it is present on the cell surface, may be required for attachment to host tissues and invasion. Its potential adhesive properties could also be required for biofilm formation.

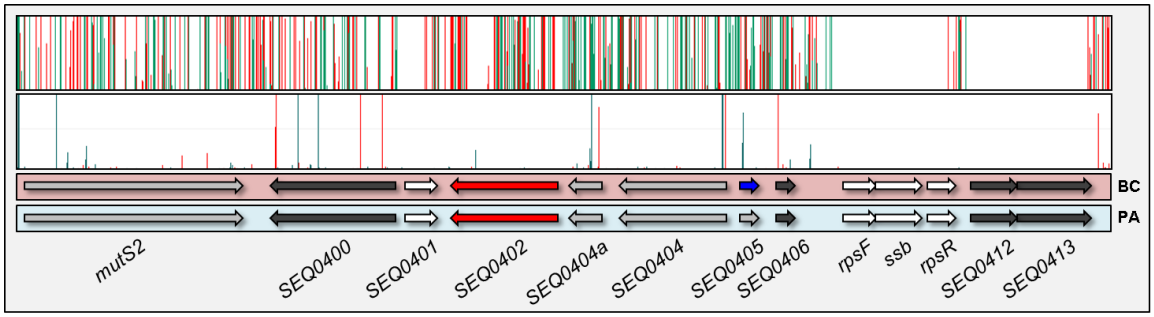


Figure 4.23. Prevalence of *S. equi* IS*S1* mutants in *SEQ0399-SEQ0413* pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. Essentiality assigned to these genes by the barcoded analysis is highlighted with a pink box with essentiality assigned by the per animal analysis highlighted by a blue box. *SEQ0402* is required for infection according to the both the barcoded and per animal analysis. White arrows indicate essential genes in THB *in vitro.* Blue arrows indicate genes in which insertion significantly reduced in fitness, red arrows indicate genes included in the validation panel, light grey arrows indicate non-essential genes *in vivo*. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

Surface polyrhamnose GlcNAc polymer processing

Gram positive bacteria contain a thick peptidoglycan cell wall that binds proteins and a range of carbohydrate polymers, particularly rhamnose [239]. In *S. pyogenes* and *Enterococcus faecalis*, these polymers are essential for maintaining and protecting bacterial cell envelopes and promoting pathogenesis [239]. These surface polymers differ between species of streptococci and are utilised to categorise the species into Lancefield group [240]. In *S. pyogenes,* the Group A carbohydrate (GAC), comprises around 40-60 percent of the cell wall [241]. The abundance of the Group C carbohydrate in *S. equi* is not known. GAC is covalently linked to the cell wall peptidoglycan and contains a polyrhamnose backbone with N-acetylglucosamine (GlcNAc) side chains [242, 243]. In *S. pyogenes*, GAC biosynthesis and transport genes are encoded as an operon (*gacA-L*) with an additional gene, *gacO,* located elsewhere in the genome (Figure 4.24). In *S. equi*, homologs of *gacABCDEFG* are followed by *gacI*, then a membrane protein (*SEQ0970*) with no homology to that of *S. pyogenes* encoded by *gacJ*. However, *SEQ0970* has homology with uncharacterised membrane proteins in other Group C streptococci; *S. zooepidemicus* and *Streptococcus dysgalactiae* (*S. dysgalactiae*) (Figure 4.24). SEQ0970 is therefore likely to encode a Group C-specific function. Following *SEQ0970,* is *gacK* and *gacL*, another gene *SEQ0973*, uncharacterised in *S. equi* and not present in *S. pyogenes*, and *gacH*. *SEQ0973* is predicted to be a glycosyltransferase and is 70 percent identical to *tuaC* in *S. dysgalactiae,* a homologue

of which in *Bacillus subtilis* (*B. subtilis*) was proven to be involved in teichuronic acid biosynthesis, that forms part of the cell wall [244].

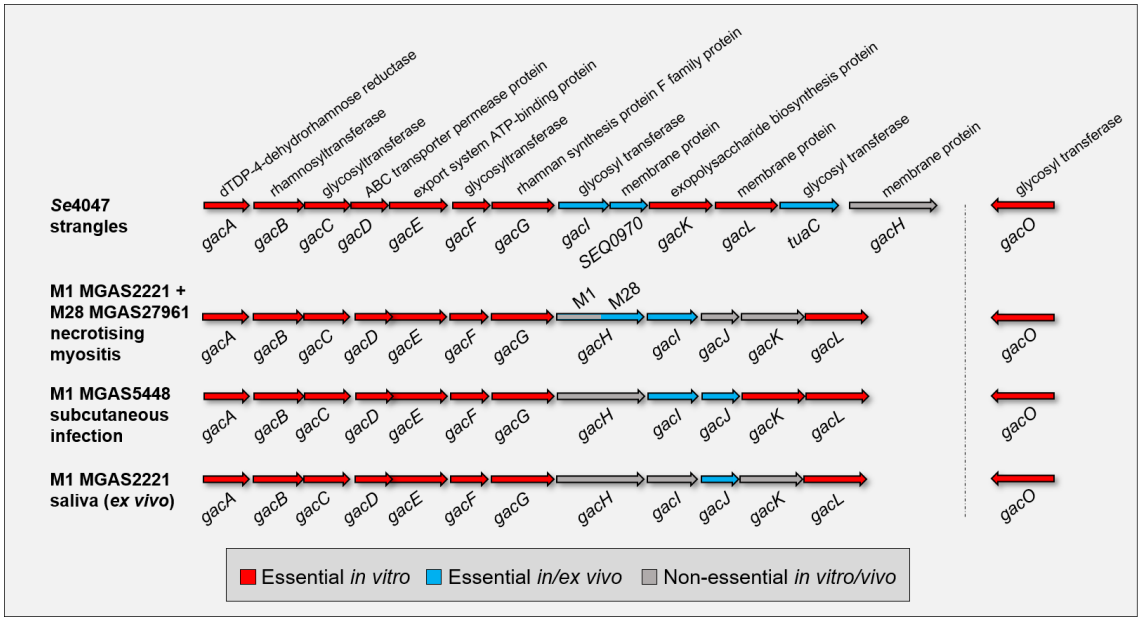


Figure 4.24. Essentiality of the *gac* operon genes in *S. equi* and *S. pyogenes in vitro* and *in vivo*. The results of 4 TraDIS/Tn-seq screens are summarised for the surface polyrhamnose GlcNAc polymer processing genes (*gac*) [81, 87, 145]. The *gac* biosynthesis and transport genes *gacA-G* are consistently required *in vitro* across the 2 species, as is *gacL* and *gacO*. GacI is required in all *in vivo* experiments and therefore represents a pan-species *in vivo* fitness gene. Red arrows indicate genes essential *in vitro*, blue arrows indicate genes essential *in/ex vivo* and grey arrows indicate genes non-essential *in vitro* and *in vivo*.

In *S. pyogenes*, it is hypothesised that GacO functions to transfer GlcNAc-phosphate from UDP-GlcNAc to Und-P, generating GlcNAc-pyrophosphorylundecaprenol (GlcNAc-P-P-Und), which initiates and acts as a membrane-anchored acceptor for polyrhamnose synthesis, catalysed by the rhamnosyltransferase synthesis system encoded by *gacA, gacB, gacC, gacF* and *gacG* [245] (Figure 4.25). Unsurprisingly, GacO is essential *in vitro* for *S. equi* and all *S. pyogenes* strains (Figure 4.24). GacDE encodes an ABC transporter that transfers the polymerised polyrhamnose to the outer membrane [246] (Figures 4.24 and 4.25). Within the inner membrane, GacI catalyses the formation of GlcNAc-P-Und which is stimulated by GacJ [245]. GlcNAc-P-Und is then diffused across the membrane by GacK [245] (Figure 4.25). GacL subsequently transfers the GlcNAc from GlcNAc-P-Und onto the exported polyrhamnose backbone [245] (Figure 4.25). Following this, GacH transfers membrane bound glycerol phosphate onto the GlcNAc side chains [247] (Figure 4.25). It is hypothesised that the resulting polysaccharide GAC is then attached to a cell wall peptidoglycan via a phosphate ester linkage [239] (Figure 4.25).

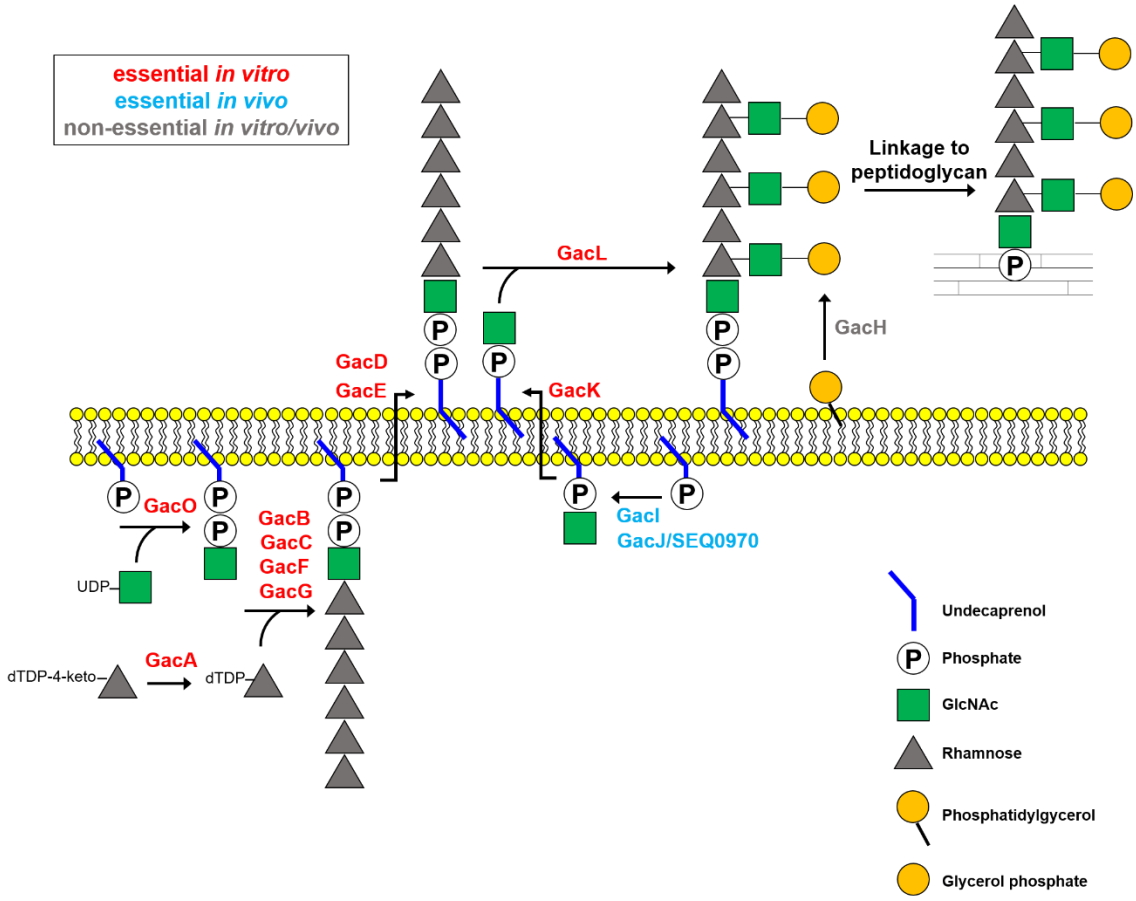


Figure 4.25. Schematic diagram of the putative surface polyrhamnose GlcNAc polymer processing system in *S. equi*. Based on findings in *S. pyogenes, gac* genes are likely to be involved in the biosynthesis and transport of the important rhamnose GlcNAc polymer onto the peptidoglycan layer. Genes essential to *S. equi* in THB *in vitro* are indicated in red, genes essential for *in vivo* infection of the natural host are indicated in blue and genes non-essential *in vitro* or *in vivo* are indicated in grey. Adapted and redrawn from [245, 247].

In both *S. equi* and *S. pyogenes*, TraDIS/Tn-seq screens consistently identified *gacA-G* and *gacL* as essential *in vitro* [78, 81] (Figure 4.24). Deletion mutants in *gacA-C* could not be made in *S. pyogenes*, confirming their essentiality [248]. However, deletion mutants lacking *gacD-G, gacL* and *gacH* could be generated in *S. pyogenes*, incurring no alterations to viability or ability to transfer the GlcNAc side chain to the rhamnose polysaccharide [248]. The dispensability of *gacH* is supported by the TraDIS/Tn-seq screens, except for in *S. pyogenes* serotype M28 in the NHP model of necrotising myositis [87, 145] (Figure 4.24).

S. pyogenes deletion mutants in *gacI, gacJ* and *gacK* were viable *in vitro*, but were necessary for GlcNAc side chain addition [248]. The loss of the GlcNAc side chain in a *gacI S. pyogenes* mutant resulted in increased susceptibility to killing by whole blood, neutrophils, cathelicidin and serum. In a rabbit model of pulmonary infection, the Δ*gacI*

mutant caused no fatalities, whereas the wild-type strain killed 89 percent of animals [248]. The Δ*gacI* mutant in a murine model of systemic infection caused significantly lower mortality and bacterial blood counts [248].

In contrast, Δ*gacI* and Δ*gacH* deletion mutants exhibited increased fitness in the presence of the bactericidal protein human Group IIA phospholipase A_2 (hGIIA), which was associated with the delayed penetration through the cell wall [249]. hGIIA is produced by the host in inflammation and during infection where it binds to negatively charged cells before penetrating through the peptidoglycan layer to access the phospholipid membrane, which is subsequently hydrolysed [250]. The increased fitness of these *gac* mutants suggest that the GlcNAc side chains are natural targets for hGIIA during infection. The attenuation of the Δ*gacI* mutant in other assays, as previously described, was concluded by the authors to be a result of using non-inflamed serum or plasma, in which hGIIA would be too low to cause killing. They also hypothesise that attenuation of Δ*gacI* mutants *in vivo* suggest that an intact GlcNAc side chain is potentially more beneficial than the ability to resist hGIIA. The increased fitness of *gacH* mutants in hGIIA assay is also not consistent with the previous finding in *S. pyogenes*, that GacH does not alter GlcNAc side chain formation, a result which remains unexplained [248].

S. equi gacI transposon mutants were attenuated in ponies (log_2FC= -5.7, q= 0.0002) and comprised 1 of the 23 consensus genes required for *S. equi in vivo* and *S. pyogenes in vivo* in mice and in NHPs [87, 145]. The *S. equi* Δ*gacI*tagC appeared to be attenuated in ponies, having recovered only 1 read from 1 lymph node corresponding to the mutant (Table 4.12). Together, these data suggest that the GlcNAc side chain in streptococci promotes virulence by increasing survival and resistance to host immune defences. GacI was not essential, however, *ex vivo* in human saliva [81] which may be explained by the lack of host immune cells, yet, *gacI* was also non-essential for *S. equi* in the presence of whole equine blood. The additional importance of *gacJ/SEQ0970* of *S. equi* and *S. pyogenes* in the TraDIS screens (except for in the NHP model), is not surprising given that GacJ stimulates GacI [87, 145].

The retained viability of Δ*gacI* mutants *in vitro*, the consistent identification of the gene as required for *in vivo* infection by TraDIS/Tn-seq, and the potential validation of a Δ*gacI S. equi* mutant in ponies makes this gene a prime target for attenuating live vaccines. The Group C carbohydrate, however, may be targeted by the host immune system and so deleting *gacI*, therefore removing the capability to process the Group C carbohydrate, may affect the generation of a protective host immune response. The Group C carbohydrate itself may, however, be an ideal subunit vaccine component.

Cell wall defence mechanism

Bacterial pathogens have to defend against a range of host immune factors such as cationic antimicrobial peptides (CAMPs) and bacteriolytic enzymes. These host factors are attracted to the negative charge of bacteria, for which some species have developed resistance by altering the charge of cell wall and membrane components. The *dlt* operon of Gram positive bacteria supplements existing teichoic acids in the cell wall with positively charged D-alanine esters, reducing the overall negative charge of the cell, therefore limiting the effectiveness of host CAMPs and bacteriolytic enzymes [250-254]. The *dlt* operon encodes 4 proteins, DltABCD; D-alanine--poly(phosphoribitol) ligase, D-alanine transport protein, D-alanyl carrier protein and D-alanyl-lipoteichoic acid biosynthesis protein, respectively. In the *S. equi* TraDIS screen, *dltC* was removed from the analysis since it contained too few reads to pass the inclusion criteria, owing to its small size. Despite this, the remaining 3 genes, *dltABD* were essential for virulence in ponies as determined by the BC analysis, of which the importance of *dltB* was reflected in the validation study. The PA analysis of the data failed to identify a significant decrease in reads corresponding to *dltB* and *dltD*. IS*S1* mutants in genes surrounding the *dlt* operon retained their virulence, as evident from presence of sequencing reads either side of the loci (Figure 4.26). This clear gap in sequencing data over the *dlt* operon illustrates the fine resolution of the *in vivo* TraDIS screen data and its ability to identify important genes with no ‘halo’ effect into the surrounding genes. *dltABD*, but not *dltC*, were also important in the murine model of subcutaneous infection screen, with the requirement of *dltA* being confirmed *in vivo* with a single gene deletion mutant [87] (Figure 4.26).

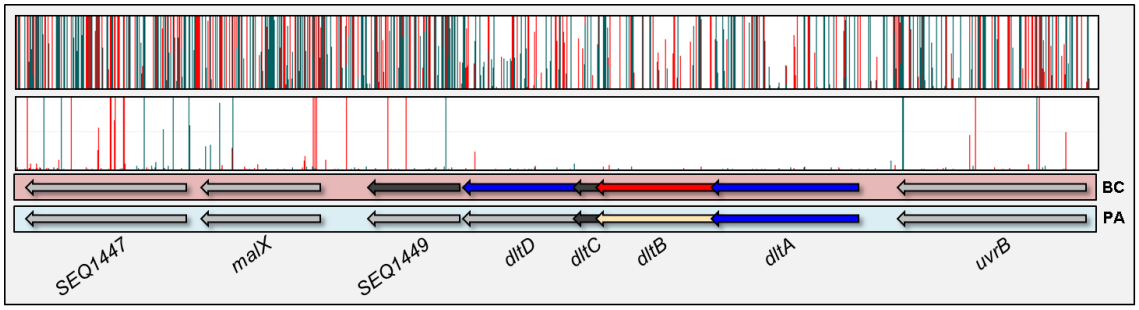


Figure 4.26. Prevalence of *S. equi* IS*S1* mutants in *SEQ1447-SEQ1454,* which includes the *dlt* operon, pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. Essentiality assigned to these genes by the barcoded analysis is highlighted with a pink box with essentiality assigned by the per animal analysis highlighted by a blue box. DltABD are required for infection according to the barcoded analysis, with the per animal analysis failing to identify *dltB* and *dltD* as *in vivo* fitness genes. Blue arrows indicate genes in which insertion significantly reduced in fitness, red arrows indicate genes included in the validation panel, the cream arrow indicates the lack of validation mutant identification in the per animal analysis and light grey arrows indicate non-essential genes *in vivo*. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

A *dlt* deficient strain of *L. monocytogenes* was severely attenuated in mice, yet its morphology and growth rate remained as wild-type [252]. However, adherence to macrophages and human epithelial cells was significantly reduced [252]. In a murine model of arthritis, a *S. aureus Δdlt* mutant caused significantly less sepsis and septic arthritis, resulting from the enhanced killing by neutrophils, rather than increased phagocytosis [255]. A *dltA* deficient *S. agalactiae* strain was more susceptible to killing by macrophages and neutrophils and was attenuated in mouse and neonatal rat infection models [256]. In *S. pyogenes*, a Δ*dltA* mutant was 45-fold more susceptible to hGIIA killing compared to wild-type [249].

The susceptibility of *dltABD* IS*S1* mutants and the Δ*dltB*tagC mutants in ponies suggest that equines can potentially produce CAMPs/hGIIA-like enzymes as a host defence mechanism to resist *S. equi* during infection. Genes of the Dlt operon represent potentially ideal targets for live attenuated vaccines and require further investigation.

Amino acid scavenging

Two putative proteins encoding a membrane protein and a permease, *scfAB*, were identified in the set of 23 consensus genes similarly required for streptococcal infection in ponies, NHPs and mice [87, 145]. A *scfA* mutant was unable to cause disease in the *S. equi* validation study (0 reads sequenced across all animals). These results are

reflected in the mouse model of subcutaneous infection where a *scfA, scfB* and a *scfAB* double deletion mutant were all significantly defective in lesion formation compared to the wild-type strain and were significantly reduced in their ability to disseminate to the spleen [87].

Homologues of *scfAB* were identified, consistently as a pair, in 21 *Streptococcus*, *Enterococcus* and *Bacillus* species [87]. The conservation of these genes in related species highlights their importance, particularly as genes surrounding *scfAB* in these species varies widely [87]. IS*S1* transposon mutants in *S. mutans* were screened on TH agar at pH 5.5, to identify mutants sensitive to acid stress. Inverse PCR was used to determine insertion sites in the sensitive mutants, identifying 7 sensitive strains, 4 of which contained insertions in, or very close to, SMU.746 (*scfA*) and SMU.747 (*scfB*) [257]. Single gene deletion mutants of *scfA* and *scfB* and a double deletion mutant, *scfAB*, in *S. mutans* were sensitive to low pH, confirming the IS*S1* screen results [257]. Biofilm formation at this low pH was hindered in the 3 deletion strains compared to wild-type [257]. At a neutral pH, the 3 *scf* deletion strains were capable of forming biofilms comparable to that of the wild-type strain [257]. The 2 mutants were also exposed to H_2O_2 and puromycin (causes premature chain termination during protein synthesis) to induce environmental stress responses. The fitness of none of the *scf* deletion mutants were affected [257], which is supported by the lack of identification of these genes in the *S. equi* H_2O_2 screen. Bacterial cells can uptake amino acids as single residues or as small peptides, by different transport mechanisms. The *scfAB* double deletion mutant grew similarly to wild-type in chemically defined medium (CDM) containing only peptone as an amino acid source, but grew poorly in CDM containing single amino acid residues, suggesting that *scfAB* are required for amino acid residue import [257]. Conducting these CDM experiments with late-stationary phase cells, starved of energy, worsened the poor growth characteristics in CDM containing amino acid residues, in comparison to cells from exponential phase growth [257]. These data suggest that the permease locus *scfAB* is especially important in energy deprived conditions, when cells need to scavenge vital compounds from the environment. Overall, the authors concluded that *scfAB* are required by *S. mutans* for survival in low pH, biofilm formation and amino acid import, all factors which ultimately affect virulence.

In *S. pyogenes*, a transposon mutant screen identified that *scfA* was required for survival in human blood [82]. However, this result was not replicated in the *S. equi* TraDIS screen in equine blood in this thesis. Nonetheless, the importance of yet another transporter, *scfAB,* to *S. equi*, *S. pyogenes* and *S. mutans* is evident and may represent potentially important future targets for the development of novel therapeutics and vaccines.

Putative and proven ABC transport systems

An ABC methionine transport system, encoded by *metQNP*, comprised 3 of the 23 consensus genes required for infection in ponies, NHPs and mice, highlighting yet another important transport mechanism necessary for fitness in streptococci. The *S. equi* Δ*metP*tagB mutant appeared to be completely attenuated in the TraDIS validation study, as no reads corresponding to the mutant were sequenced from any of the animals. A triple deletion mutant of *metQNP* in *S. pyogenes* grew as wild-type in THB, but had a severe growth defect in peptide-free CDM, confirming the importance of this locus in amino acid import [145]. Supplementing the growth medium with methionine restored the growth of the *ΔmetQNP* mutant to near wild-type levels, confirming this transporter's role in the acquisition of methionine [145]. The *ΔmetQNP* mutant caused significantly smaller lesions in the NHP model of necrotising myositis, in addition, lower bacterial loads were recovered from the inoculation site [145]. Transcriptome analysis of genes expressed in NHP muscle tissue infected with wild-type *S. pyogenes* serotype M1 and in an infected human patient, detected transcription of *metQ*, confirming its expression *in vivo* [145].

An uncharacterised putative ABC transporter system encoded by *SEQ1410-SEQ1412*, was required for infection in ponies (*SEQ1410*; log_2FC= -5.8, *q*= 0.0005, *SEQ1411*; log_2FC= -3.8, *q*= 0.03, *SEQ1412*; log_2FC= -5.3, *q*= 0.002) as concluded by the BC analysis. None of these genes were identified as required for fitness when analysed by the PA technique. Homologues of all 3 genes were also required by *S. pyogenes* in the murine model of subcutaneous infection [87]. BLAST searches of these genes identified them as a branched-chain amino acid ABC transport ATP-binding protein (*SEQ1410*), a branched-chain amino acid ABC transport permease (*SEQ1411*) and an ABC transport substrate-binding protein (*SEQ1412*). No homology to any characterised ABC systems was found alluding to the function of this system. Therefore, it remains unknown what amino acid this systems imports, until experiments utilising CDM containing different amino acids can be conducted. A Δ*SEQ1410* deletion mutant was included in the TraDIS validation panel, which appeared attenuated in ponies, contributing 0.17 percent (1,332 reads) of surviving *S. equi* recovered from ponies. All but 2 of these reads corresponding to the Δ*SEQ1410*tagC deletion mutant were sequenced from 1 lymph node, from 1 animal.

Components of another putative ABC transporter system (SEQ1310-SEQ1312) are important for survival *in vivo* for *S. equi*. Genes encoding this locus were termed *sptABC* due to their requirement in *S. pyogenes* for persistence in human saliva *ex vivo.* The requirement for *sptA* and *sptC* to *S. pyogenes* was confirmed using single gene deletion mutants in human saliva, where mutants were severely attenuated [81]. In ponies, *sptA* IS*S1* mutants were significantly attenuated in the BC analysis (log_2FC= -5.2, *q*= 0.001),

with *sptC* IS*S1* mutants just missing significance (log_2FC= -3.5, *q*= 0.066) (Figure 4.27). The PA analysis did not detect *sptA* IS*S1* mutants as significantly reduced in fitness, however, highlighting the increased sensitivity of the BC analysis method (Figure 4.27). Reads comprising 0.0014 percent of the total reads recovered from the validation ponies corresponded to the Δ*spt*AtagB deletion mutant, inferring its necessity *in vivo*.

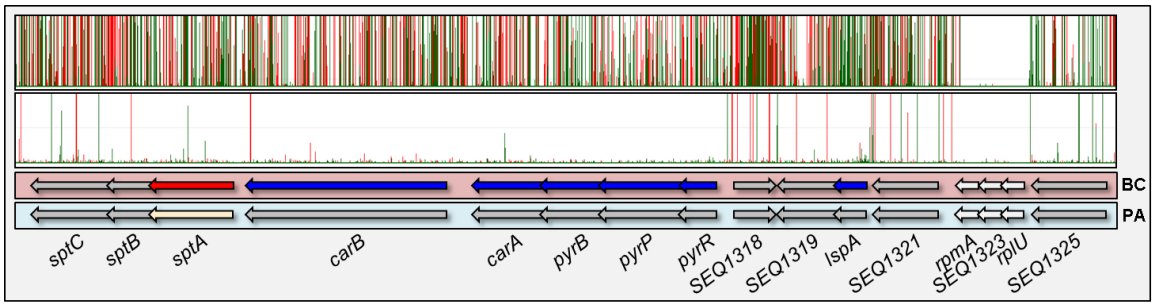


Figure 4.27. Prevalence of *S equi* IS*S1* mutants in *SEQ1310-SEQ1325,* which includes the *spt* and *car* operons, pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. Essentiality assigned to these genes by the barcoded analysis is highlighted with a pink box with essentiality assigned by the per animal analysis highlighted by a blue box. SptA and CarAB are required for infection according to the barcoded analysis, with the per animal analysis failing to identify any of these as *in vivo* fitness genes. Blue arrows indicate genes in which insertion significantly reduced in fitness, red arrows indicate genes included in the validation panel, the cream arrow indicates the lack of validation mutant identification in the per animal analysis and light grey arrows indicate non-essential genes *in vivo*. Data is viewed in Artemis [112].

In 7 pathogenic streptococci, including *S. equi*, *carAB* is situated directly upstream of *sptABC* [81] (Figure 4.27). The *carAB* locus is involved in pyrimidine and arginine synthesis; *carB* encodes the large subunit and *carA* the small subunit of a carbamoylphosphate synthase [258, 259], carbamoylphosphate being a precursor for pyrimidine and arginine synthesis [259]. A Δ*carB* deletion mutant of *S. pyogenes* was confirmed to be attenuated in human saliva [81] and in human blood [82]. IS*S1* mutants in both *carA* and *carB* were significantly attenuated in ponies in the BC analysis (*carA;* log_2FC= -4.3, *q*= 0.003, *carB;* log_2FC= -4.9, *q*= 0.004), however, the PA analysis failed to detect decreased reads in these genes as significant (Figure 4.27). The conserved nature of the *spt* and *car* loci in streptococci suggest that there could be a functional relationship between the 2 gene sets.

Exported proteins

Phospholipase A_2 toxins

S. equi encodes 2 phospholipase A_2 toxins as described in Chapter 1 section 1.3.2. SlaA and SlaB share 98 and 70 percent amino acid identity with SlaA of *S. pyogenes* M3 MGAS315, respectively [3]. Phospholipase A_2 toxins represent major virulence factors, with the acquisition of SlaA in *S. pyogenes* resulting in increased morbidity and mortality in humans, increased tissue destruction and dissemination in the murine model of infection [44, 45]. A Δ*slaA* deletion mutant in *S. pyogenes* serotype M3 was reduced in its ability to colonise the respiratory tract in the NHP model of pharyngitis [45]. SlaA seems to be restricted to primarily the M3 serotype, but can be found in some M4 and M28 strains [260-264].

Interestingly, *S. equi* IS*S1* mutants in *slaA* retained virulence in ponies, trending towards increased fitness (log_2FC= 4, q = 0.09), whereas *slaB* IS*S1* mutants were significantly attenuated (log_2FC= -5.7, q = 0.0004). A *S. equi* Δ*slaAB* double deletion mutant was not significantly attenuated *in vivo*, however, ponies produced less nasal discharge, which supports the previously described link between phospholipase A_2 toxins in mucus formation in humans [46, 47]. The opposing effects seen in the *S. equi slaA* and *slaB* IS*S1* mutants *in vivo*, may explain the lack of overall attenuation seen in ponies challenged with the Δ*slaAB* double deletion mutant. These opposing effects suggest that the 2 toxins serve different functions in *S. equi*. In the TraDIS validation study, Δ*slaB*tagC reads were recovered from all animals, but in very low numbers in 4 out of the 5 animals (1 or 2 reads). One animal that showed obvious clinical signs of disease in the validation study, contributed 4,126 Δ*slaB*tagC reads, equating to 0.52 percent of all the reads sequenced in the recovered materials. Δ*slaB*tagC represents the validation mutant recovered in highest abundance, yet the overall presence in comparison to the ΔIC strains was minimal.

Uncharacterised exported protein

SEQ1535 and *SEQ1536* encode 2 putative exported proteins, located in an operon. *SEQ1535* IS*S1* mutants were unaffected *in vivo*, yet *SEQ1536* mutants were highly attenuated in the *S. equi* TraDIS screen, when analysed using the BC technique, (log2FC= -7.2, q= 0.00009) and potentially in the validation study. The Δ*SEQ1536*tagA mutant was not present in any recovered abscess material. BLASTP searches of these genes identified them as ‘general stress proteins’, *SEQ1535* containing a YtxH domain and *SEQ1536* containing a YtxG domain. *ytxH* and *ytxG* were identified in *Bacillus* species, yet remain improperly categorised. The expression of *ytxH* and *ytxG* are upregulated in *Bacillus* during the ‘stringent response’ which is induced in stress conditions, such as amino acid limitation and heat shock [265, 266]. A *SEQ1536*

homolog (Spy0486) was also indispensable in the *S. pyogenes* murine model of subcutaneous infection and in human saliva *ex vivo* [81, 87], suggesting that this gene is employed in stress conditions in these streptococci. Further research into these genes is required to determine their function.

4.4.3 Genes conferring enhanced fitness as a result of insertion, identified by barcoded TraDIS

The majority of IS*S1* mutants increased in fitness were in genes unclassified by the COG grouping system. Five membrane proteins, 2 cell-surface anchored proteins, 4 exported proteins, 7 hypothetical proteins and 11 pseudogenes are included in this category. Pseudogenes contributed 13 percent of all genes conferring a fitness advantage *in vivo* upon insertion. The notable presence of pseudogenes is curious, since the nature of their annotation suggests that they are no longer functional. It is possible that insertion into these genes may confer a fitness advantage in comparison to other genes in the mutant pool that confer a reduced fitness. It may be that IS*S1* insertion into these genes does not necessarily confer a fitness advantage, but a more wild-type phenotype and would be difficult to replicate in isolation when not in a mixed mutant pool. This effect was evident in the mouse subcutaneous model Tn-seq screen, whereby only 1 of 7 selected enhanced fitness mutants (*covS*) was successfully validated *in vivo* using deletion mutants [87]. Spontaneous mutations in *covS*, aside from the transposon insertions, were identified in the recovered output pools, suggesting that the 6 ‘false positive’ genes potentially conferred a fitness advantage as a result of these *covS* mutations and not the transposon insertion [87]. After collection of abscess material, surviving mutants are isolated by out-growth on TH agar. This *in vitro* recovery step may also permit the expansion of some mutants that grow well on TH agar and may not have in fact been fitter *in vivo*.

Potential false positive may also be possible if some individual mutants are able to reach the lymph nodes, earlier than others, by chance. To mediate against this effect, genes were removed from the BC analysis that contained a large number of reads in only 1 of the 3 output libraries. Despite this measure, the data was analysed on a per gene read count basis, therefore it is not immediately apparent if the majority of reads were contributed by a single mutant within that gene.

Consensus transposon mutants enhanced in fitness

Despite the challenges faced with enhanced fitness mutants, 5 genes with a fitness advantage, as a result of transposon insertion, were similarly identified in *S. equi* in ponies and the *S. pyogenes* subcutaneous infection model. *S. equi* encodes 4 phosphoglycerate mutases, 1 of which, *gpmA*, is essential *in vitro* and is required for the

conversion of 3-phosphoglycerate to 2-phosphoglycerate, a step in the glycolysis pathway as described in Chapter 2, section 2.3.2. One of the other phosphoglycerate mutases, encoded by *SEQ0765* and its homolog in *S. pyogenes*, *Spy0497*, were enhanced in fitness as a result of transposon insertion [87]. Little research has been conducted into these additional phosphoglycerate mutases so it is not clear what their enzymatic functions are and why disrupting this gene would confer any fitness advantage. IS*S1* within the *SEQ0765* mutants are largely located at the start of the gene in *S. equi*, suggesting that enhanced fitness is a result of inactivation of the enzyme, or that insertion at this particular site has increased the functioning of the enzyme. Increases in fitness are less obvious than decreases in fitness, when visualising the data, and are often attributed to only a few mutants in the gene (Figure 4.28).

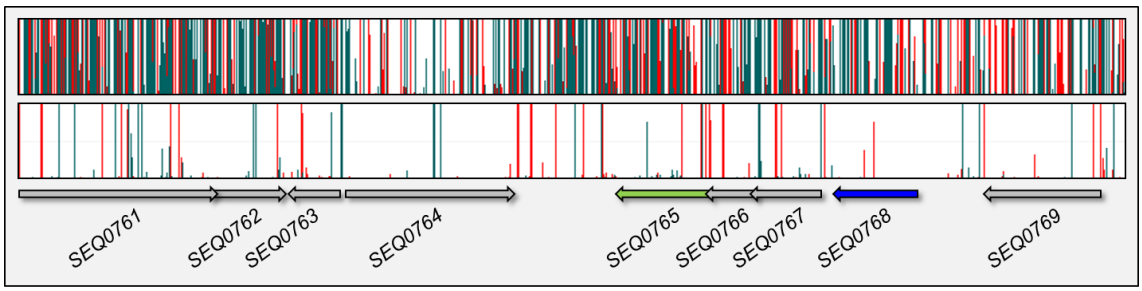


Figure 4.28. Prevalence of *S. equi* IS*S1* mutants in *SEQ0761-SEQ0769*, pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. IS*S1* mutants in SEQ0765, a phosphoglycerate mutase, confer a fitness advantage *in vivo* (green arrow). Blue arrows indicate genes in which insertion significantly reduced in fitness and light grey arrows indicate non-essential genes *in vivo*. Data is viewed in Artemis [112].

Transposon mutants in *clpL* were enhanced in fitness *in vivo* in both *S. equi* and *S. pyogenes* in the subcutaneous murine model of infection. CplL is an ATP-dependent protease ATP-binding subunit that functions as a chaperone, assisting the proper folding and translocation of proteins in reaction to altered environmental temperatures [191, 192, 267]. ClpL, and other heat shock ATPases, have been shown to associate with the proteolytic subunit, ClpP, to provide ATP-binding capabilities, regulating the proteolytic complex [268]. Upon Clp ATPase binding, the ClpP complex can degrade larger substrates [269, 270]. Expression of *clpP* has also been shown to induce virulence factor production, such as pneumolysin [267]. In *S. pneumoniae,* ClpL does not induce virulence factor expression independently of ClpP [267].

ClpL was not actively degraded in cultures of *S. pneumoniae* returned to normal conditions after heat shock [267]. The persistence of ClpL after initial induction is

hypothesised to enhance the stability of cells during infection [271] and promote the induction of virulence factors through its availability for ClpP binding. An *S. pneumoniae* Δ*clpL* deletion mutant in a murine model of disease, however, did not behave significantly differently to the wild-type strain, suggesting that ClpP can function without ClpL, likely through the binding of another ATPase instead [267]. A Δ*clpP* deletion mutant was attenuated in mice and was found to lack key virulence factor expression [267]. The enhanced fitness of *S. equi* IS*S1 clpL* mutants *in vivo* could be explained by an even further reduced rate of degradation post-heat shock, if the mutation causes some form of conformational change to the resultant protein. It may be possible that this structural change could also enhance its ability to bind ATP or to form a complex with ClpP. Further investigation into the location of the IS*S1* insertion sites in both *S. equi* and *S. pyogenes* may allude to this.

A polysaccharide acetylase encoded by *SEQ1120* and *Spy0818* conferred a fitness advantage when disrupted by transposon insertion. BLASTP searches of this gene match with 63 percent similarity to *pdi* of *Streptococcus iniae* (*S. iniae*). This enzyme has been implicated in cell wall modification by deacetylating GlcNAc in the peptidoglycan layer, protecting from host lysozyme activity [272]. Pdi is predicted to be localised in the membrane, via its signal peptide, therefore Pdi may have adhesive properties [272]. In *S. iniae*, a Δ*pdi* deletion mutant was sensitive to lysozyme, had a shortened chain length, reduced ability to survive in whole blood and to adhere and invade epithelial cells [272]. The authors hypothesised that the reduced chain length may increase the efficiency of host phagocytic ingestion, which could be beneficial to *S. equi*, in light of its potential transport within these cells as a means of migrating to local lymph nodes. The potential inactivation of *pdi* in the *S. equi* and *S. pyogenes* IS*S1* mutants may also reduce adherence, promoting translocation. The Δ*pdi* deletion mutant in *S. iniae* was, however, attenuated *in vivo* [272].

Transposon mutation of a β-*N*-acetylglucosaminidase encoded by *SEQ1693/Spy1314*, incurred a fitness advantage *in vivo* in *S. equi* and *S. pyogenes* in the murine subcutaneous infection model [87]. This gene was previous annotated as a hyaluronidase, but further research into the gene in *S. pyogenes*, proved that the enzyme had no activity against hyalronan and instead acts upon *N*-acetylglucosaminides [273]. The authors hypothesised that the enzyme removes GlcNAc from imported glycoconjugates, scavenged from the environment, and is therefore involved in carbohydrate metabolism [273]. Transcription of this β-*N*-acetylglucosaminidase was upregulated during phagocytosis, suggesting that the enzyme contributes to virulence and therefore may act to deglycosylate host oxygen (*O*-) linked GlcNAc, potentially disarming host cell machinery [273, 274]. It is unclear why transposon insertion in this

gene would incur a fitness advantage. It is possible that insertion enhanced the cleaving properties of the enzyme. Another explanation may be that *SEQ1693* is involved in the turnover of surface GlcNAc, therefore insertion mutants may have increased levels of surface GlcNAc, conferring an improved resistance to the host immune system.

The major tail phage protein, *SEQ2051*, is encoded on the prophage ϕSeq4. Transposon mutants in this gene, and in the *S. pyogenes* homolog, *Spy1429*, were enhanced in fitness *in vivo*. After bacterial infection, prophages can enter the lysogenic cycle, where the phage DNA is integrated into the host's genome. Prophage DNA is therefore replicated alongside that of the host. Cellular stresses such as DNA damage and resource limitation, can lead to the prophage entering the lytic cycle in an attempt to promote phage survival [275]. Upon entering the lytic cycle, the phage replicates and induces bacterial cell lysis, releasing the phage into the environment. It may be possible that mutations in the genes encoding the major tail proteins of *S. equi* and *S. pyogenes* prophage reduces host cell lysis resulting from cellular stress experienced *in vivo*.

As described previously in section 4.4.1, *sagB* was the only consensus gene, conferring an enhanced fitness *in vivo* as a result of IS*S1* insertion, in both *S. equi* and the *S. pyogenes* M1 serotype [145]. No genes conferring an enhanced fitness upon insertion were similarly identified in *S. equi* and the *S. pyogenes* serotype M28 data or the M1 in human saliva *ex vivo* data [145].

***S. equi* specific transposon mutants enhanced in fitness**

In *S. equi*, IS*S1* mutants in 1 gene grouped in the no COG category and 2 genes grouped into the signal transduction mechanisms COG category, *SEQ1704*, *SEQ1711* and *SEQ1712* were significantly increased in fitness *in vivo* (*SEQ1704*; $\log_2$FC= 9.3, *q*= 0.006, *SEQ1711*; $\log_2$FC= 6.6, *q*= 0.02, *SEQ1712*; $\log_2$FC= 4.9, *q*= 0.03) (Figure 4.27). *SEQ1704, SEQ1711* and *SEQ1712* encode a putative membrane protein of unknown function, a putative sensor histidine kinase and a putative response regulator protein, respectively. BLASTP searches, potentially identified them as a bacteriocin immunity protein, *blpH* and *blpR*, sharing homology with those of *S. pneumoniae*. No homologues to *blpH* and *blpR* exist in *S. pyogenes*.

The *blp* locus encodes a quorum sensing system that induces bacteriocin production, a system well described in *S. pneumoniae* [276-278] (Figure 4.29 and 4.30). Bacteriocins are small antimicrobial peptides that aid in inter- and intraspecies competition [276]. Interestingly, in *S. pneumoniae*, different strains can utilise specific bacteriocins to compete against other *S. pneumoniae* strains [276, 279, 280]. Bacteriocin production is controlled by the production of a pheromone, encoded by *blpC*, that is exported via the ABC transport system BlpAB [277, 281] (Figure 4.28A). BLASTP searches for a *S. equi*

BlpC homolog returned a currently unannotated coding region between *SEQ1710* and *SEQ1711*, which reflected the orientation of these genes in *S. pneumoniae* (Figure 4.29).

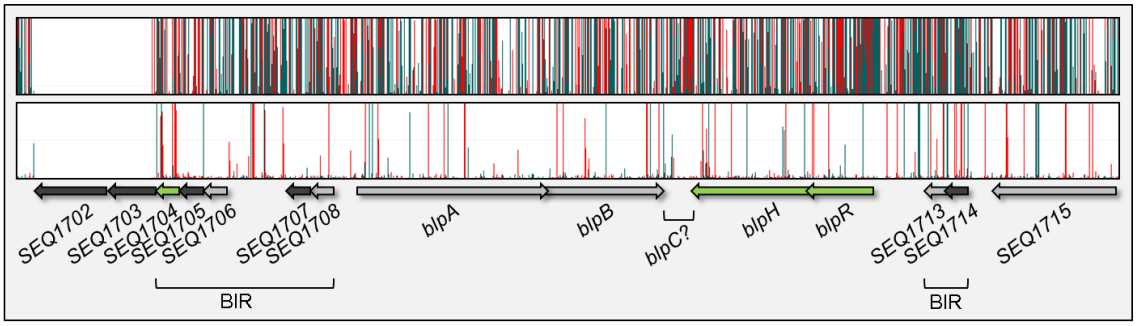


Figure 4.29. Prevalence of *S. equi* IS*S1* mutants in *SEQ1702-SEQ1715*, which includes a putative *blp* operon, pre- and post-infection of the natural equine host. The top panel represents mutants present in the input pools, with the bottom panel representing surviving mutants in the output pools. Data from the 3 input and 3 output libraries are combined for viewing purposes. Peaks indicate prevalence of each insertion mutant. Green and red peaks mapped on the forward and reverse strand of DNA, respectively. IS*S1* mutants in *SEQ1704* (putative bacteriocin immunity protein), *blpH* (sensor histidine kinase) and *blpR* (response regulator), confer a fitness advantage *in vivo* (green arrows). BIR indicates putative bacteriocin immunity regions. Light grey arrows indicate non-essential genes *in vivo*. Dark grey arrows indicate genes removed from the analysis because their essentiality in THB was not defined, or are non-essential, but contained too few reads in the input pool to meet the inclusion criteria. Data is viewed in Artemis [112].

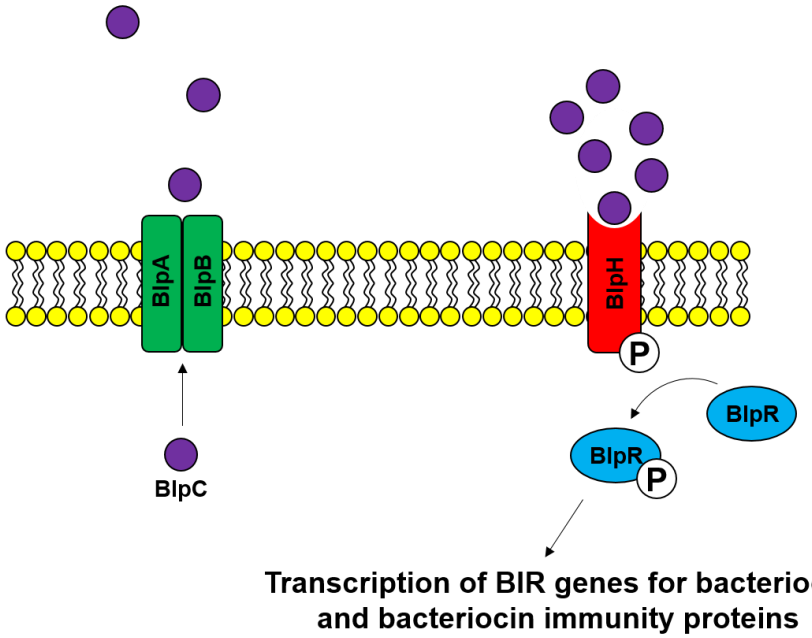


Figure 4.30. Schematic representation of a putative bacteriocin system in *S. equi*, based on homologues in *S. pneumoniae*. A previously unannotated region in *S. equi* may encode a pheromone, BlpC that is exported via BlpAB. When external concentrations of BlpC reach a threshold, BlpH is activated which in turn phosphorylates BlpR. P-BlpR then acts as a transcriptional regulator for genes in the bacteriocin immunity regions. The bacteriocins produced, which act on competing bacteria with toxic effect, may be exported by BlpAB. Redrawn and adapted from [278, 280, 282].

It is thought that bacteriocin production is an energetically costly process, and so the BlpC receptor, BlpH, is only stimulated when extracellular BlpC has sufficiently accumulated, indicating to high local cell densities [280]. Otherwise, if bacteriocin was produced when local cell density was low, it is unlikely that the extracellular bacteriocin concentration will be high enough to effect other competing bacteria, and would therefore be an unnecessary expenditure of energy [283]. Once BlpH, the sensor histidine kinase, has been stimulated, it in turn activates the response regulator, BlpR, by phosphorylation [276] (Figure 4.30). P-BlpR then binds to DNA at various loci including the bacteriocin immunity region (BIR), activating gene transcription [277]. In *S. pneumoniae*, genes in the BIR locus vary between strains, but exclusively encode for bacteriocins and bacteriocin immunity proteins [277]. Resultant bacteriocins are then exported via the *blpAB* transporter to act on other competing bacteria [276-278]. Bacteriocin immunity proteins are thought to protect from bacteriocin-mediated self-killing, so it is curious that *SEQ1704* IS*S1* mutants were acutely enhanced in fitness in *S. equi* (Figure 4.29, green arrow). *SEQ1704* was in fact 1 of the genes conferring the greatest enhanced fitness upon insertion *in vivo*. A second bacteriocin immunity protein in the *S. equi blp* locus was identified by BLASTP search; *SEQ1707*. This gene was removed from the *S. equi* TraDIS screen analysis because it did not contain enough reads in the input pools to meet the stringent criteria. All genes in the *blp* locus in *S. equi*, except *SEQ1704, blpH* and *blpR*, were either non-essential *in vivo*, or were removed from the analysis through lack of reads (Figure 4.29). These removed genes are very small (254 bp on average), hence are less likely to contain sufficient reads. The non-essentiality of the pheromone, BlpC, suggests that it is not required for survival and that cells may still react to external BlpC produced by neighbouring cells. The non-essentiality of *blpAB* may also highlight an effect of neighbouring cells, still able to produce and export BlpC. As the *blp* locus encodes several bacteriocins, it is likely that the non-essentiality of these is due to the compensatory nature of these genes.

The enhanced fitness of *blpH* and *blpR* IS*S1 S. equi* mutants may be a result of no bacteriocin production, which has been suggested to be an energetically costly process. In a mixed mutant population, as in transposon libraries, neighbouring mutants capable of producing bacteriocins, may compensate for the lack of production in *blpH* and *blpR* mutants, allowing *blpH* and *blpR* mutants to benefit from energy conservation.

Genes conferring an enhanced fitness as a result of insertion require a degree of caution when interpreting data. Except for the *covS* mutant validated in the subcutaneous mouse model of infection, no other genes conferring an increased fitness upon insertion have been successfully validated. Studies utilising TraDIS/Tn-seq generally validate genes required for fitness, due to their potential therapeutic uses. IS*S1* mutants increased in

fitness may have either been able to reach infection sites first, as would be possible in ponies where the challenge site is distinct from the infection site. In the NHP model of necrotising myositis and in the murine subcutaneous model, mutants are injected directly into the infection site and are not required to translocate to a distant site, so fitness increases seen here are less likely to be affected by this phenomenon. In the case of the *S. equi in vivo* dataset, it appears that few mutants or sometimes just 1 within the gene is responsible for the calculated enhanced fitness. This may suggest that mutation at 1 or a few specific sites causes increased functioning of the transcribed product, and therefore validating the result by re-testing a whole deletion mutant may not generate comparable results.

4.4.4 Other *in vivo* TraDIS/Tn-seq and validation studies

A small *Salmonella* Typhimurium transposon library, containing 8,550 mutants, was used to orally infect 18 pigs, 18 cattle and 90 chickens, all important production animals susceptible to *Salmonella* Typhimurium infection [83]. The pigs and cattle were each infected with a small pool of 475 unique transposon mutants and each chicken infected with 95 unique mutants. Surviving mutants were recovered from the intestinal tissue of each animal and sequenced by TraDIS. Sequencing revealed that overall, 91 percent of the input pool was still able to colonise chickens, 86.2 percent in pigs and 85 percent in calves [83]. In this study, data was analysed on a per mutant basis, as opposed to on a per gene basis. It is important to note however, that the percent recovery of input mutants quoted are also based on sequencing >1 read for each mutant and therefore the data analysis is less stringent than the *S. equi in vivo* TraDIS screen [83]. A total of 611 genes were commonly identified as required for *in vivo* infection across all 3 species, with smaller sets of species-specific fitness genes also identified [83].

Twelve genes were selected for *in vivo* validation using insertion mutants (kanamycin cassette inserted into target gene). Each insertion mutant was combined in a 1:1 ratio with the wild-type parental strain and used to infect 9 chickens [83]. Three chickens infected with each insertion mutant were terminated on day 4 post inoculation, 3 on day 6 and the remaining 3 on day 10. In total, 108 animals were used. Eight of the 12 insertion mutants were significantly attenuated on day 4, with another 2 mutants showing significant attenuated at later timepoints [83].

To study systemic disease caused by *S.* Typhimurium, small pools of transposon mutants, each containing 480 unique mutants, were intravenously administered to mice. Twenty-one C57/BL6 wild-type mice and 21 *gp91-/-phox* immunodeficient mice were inoculated in duplicate, equating to 42 mice of each genotype and 84 animals in total [84]. The 2 genotypes of mice were used in an attempt to identify potential vaccine

targets that would render a live vaccine attenuated and still able to colonise in a healthy host, but also remain attenuated in an immunocompromised host, as is potentially the case in human populations endemically effected by *Salmonella.*

Fitness was assigned to 9,356 mutants in total, providing data for 3,139 genes (68.3 percent of *S.* Typhimurium genes) [84]. In this TraDIS study, data was also analysed on a per mutant basis, as opposed to on a per gene basis. The reproducibility of read counts per mutant between the 2 wild-type mice infected with the same small pool of mutants, was not adequate. The authors state that a large number of mutants (approximately a third from interrogating the supplementary information) exhibited reduced fitness, that was not supported statistically with significance. It is likely that this high proportion of non-significant data is a result of random mutant dropout, highlighting problems with stochastic loss. Treating each mutant as an individual data point enhances the potential for problems such as these, especially when inoculating with such small pools per animal, as the amount of data required to reach statistical significance is increased. Challenging each animal with larger mutant pools and combining the data for all mutants mapping within 1 gene reduces these effects, maximising the potential for robust analysis to be conducted. It is possible that using larger libraries than those utilised in the *S. equi in vivo* TraDIS screen, would have improved the data further.

The fitness scores of mutants collected from the immunodeficient mice were however, more reliable between replicates [84]. Despite these challenges, 19 deletion mutant strains and the parental strain were generated based on unique mutant attenuation in the Tn-seq data and used to individually challenge an unknown number of C57/BL6 wild-type mice (likely to be approximately 60 mice if each strain was tested in triplicate). No statistical analysis was conducted on this data, but 6 mutants behaved as wild-type, permitting the euthanasia of mice on day 3 post-infection, due to the presence of clinical signs. Three mutants caused clinical signs on day 4 post-infection, 1 mutant on day 5 post-infection and 2 mutants on day 6. The remaining 7 mutants did not cause any clinical signs by the end point of the experiment (day 7 post-infection) [84]. Bacteria were isolated from all livers and spleens recovered post-mortem, with the mutants that did not cause any clinical signs colonising at approximately half the bacterial density of those causing clinical signs, in line with the parental wild-type *S.* Typhimurium strain.

In both *Salmonella* TraDIS/Tn-seq studies, the use of denser, barcoded libraries is likely to have increased the number of genes for which fitness data was calculated, improving genome-wide measurement of fitness, and reducing the number of animals required to generate robust, reproducible data. Utilising tagged deletion mutants, as in the *S. equi in vivo* validation study, also has the potential to reduce animal usage in follow up studies

as multiple mutants can be combined in 1 inoculum. The presence of the tag also enables all mutants to be quantified simultaneously in the inoculum and recovered materials, negating the need for time consuming culture or qPCR methods.

4.4.5 Conclusions

The genes implicated in the *in vivo* fitness of *S. equi* described above are just a fraction of those identified in this study. Many whole loci of important genes were identified, particularly in the barcoded analysis, where sensitivity for detecting less dramatic changes in fitness are apparent. Treating each animal as a biological replicate failed to identify 271 genes that were required for fitness as determined by the barcoded analysis. Combining all the mutants recovered from animals and splitting according to parental library population, reduces the number of biological replicates from 12 to 3, improving the robustness and statistical power of the data analysis. Five of the 12 genes included in the *S. equi* validation study were uniquely identified by the BC analysis and were therefore missed by the PA analysis. All of the deletion mutants corresponding to these 5 genes appeared to be attenuated *in vivo*, supporting the use of a BC technique. Not only does the BC technique improve the data analysis, but it reduces the number of animals required to achieve this quality of data, in accordance with the principles of the 3R's. Incorporating tags into the validation mutants further reduced the number of animals required since deletion mutants could be combined in 1 inoculum, along with control and wild-type strains. Simultaneous measurement of the mutants within the inoculum and in recovered materials by TraDIS negates the need for time consuming quantitative techniques such as culture and qPCR.

Thirty-nine percent and 60 percent of genes identified in the whole equine blood and H_2O_2 TraDIS screens, respectively, were also identified as required for *in vivo* fitness, highlighting a considerable overlap between the *in vitro* and *in vivo* genes. Despite this, an additional 354 genes were uniquely required for *in vivo* fitness, reflecting the much more complex environment that *S. equi* encounters in the natural host; an environment that cannot be replicated *in vitro*. Comparison of the *S. equi* TraDIS data to the *S. pyogenes* NHP and human saliva fitness genes identified 10 consensus genes, and 18 genes specifically required in the *S. equi* and *S. pyogenes* saliva data, potentially identifying genes required by *S. equi* for survival in equine saliva, a niche not yet explored for strangles. A further 51 genes were similarly identified between the M1 NHP data and the *S. equi* data, highlighting genes potentially required by both species to survive *in vivo* and form viable abscesses/lesions.

Twenty-three pan-species *in vivo* fitness genes were identified in *S. equi* and *S. pyogenes*, despite the differences in the experimental methods used to identify these

genes. Genes with proven involvement in infection were identified, along with novel genes not previously implicated in disease. Fifty-seven percent of the consensus genes encode products involved in the transport of amino acids, metal ions and pyrimidine precursors by scavenging and quorum sensing. These data suggest that transporter genes may present ideal future therapeutic and vaccine targets.

5 Additional analysis of *S. equi* ISS*1* libraries *in vivo*

5.1 Introduction

The barcoded *S. equi in vivo* data was reanalysed to investigate the effects of using a more stringent gene inclusion criteria on the data obtained. Using the 1,000 reads per gene cut-off as previously decribed in Chapter 4, may allow the identification of false positive data by enabling genes that are potentially not represented to a high enough degree in the input libraries to be included in the analysis. To investigate this, the barcoded data was reanalysed using a minimum read count of either 2,000 or 5,000 reads per gene. Only the barcoded data and not the per animal data was reanalysed for practicality.

To assess the validity of using a barcoded method, data from the 12 ponies was randomly combined into 3 pools of data and the sets of fitness genes identified by each analysis compared. The consensus data identified between these analyses should represent a robust gene set since genes were repeatedly identified regardless of whether data was randomly combined or combined by parental barcode.

5.2 Materials and methods

5.2.1 Minimum read count of 2,000 reads per gene

The 3 barcoded input libraries were analysed in the same way as described in Chapter 4 section 4.2.4, except that genes containing < 2,000 reads per gene in any 1 of the 3 libraries were removed from the analysis. The genes represented by < 2,000 reads in the input libraries were also removed from the barcoded output data but the output libraries were otherwise analysed as previously described in Chapter 4 section 4.2.4. Gene fitness was calculated using the same methods described in Chapter 4 section 4.2.4.

5.2.2 Minimum read count of 5,000 reads per gene

The data was also reanalysed using an even more stringent minimum read count per gene and was conducted as described above for the 2,000 reads per gene cut-off, except that a 5,000 reads minimum limit as applied. Gene fitness was calculated using the same methods described in Chapter 4 section 4.2.4.

5.2.3 Random combination of ponies analysis

The data sequenced from each of the 12 ponies was initially combined into 4 datasets depending on the number of unique mutants recovered from each animal, as previously defined in Chapter 4. Three ponies containing < 2,300 unique mutants comprised 1 dataset, 3 ponies containing 2,800-3,500 unique mutants, 3 ponies containing 3,600-3,900 unique mutants and 3 ponies containing 4,100-5,500 unique mutants comprised 3 additional datasets. One pony from each of these datasets was randomly selected and assigned to a new group, until 3 new groups were formed each containing 4 ponies with varied dentisties of unique mutants. Sorting and then randomly assigning the ponies into groups ensured that each group was represented by a consistent sequencing depth as libraries were previously allocated a proportion of the sequencing run in Chapter 4 depending on unique mutant density.

Fastq files from each pony were combined using the script below depending on their new group, to generate 3 output files for analysis.

```
cat 6544.fastq 7799.fastq 7616.fastq 6061.fastq > group1.fastq
```

The 3 resulting output files were analysed using bacteria_tradis and in tradis_comparison with the 3 previously analysed barcoded input files, as previously described in Chapter

4 section 4.2.4. The same gene inclusion criteria was applied as imposed on the barcoded analysis previously conducted.

5.3 Results

5.3.1 Genes implicated in the survival of *S. equi in vivo* using a gene inclusion criterion of 2,000 reads per gene minimum

Enforcing the 2,000 minimum reads per gene criteria permitted the exclusion of 405 genes previously identified as non-essential *in vitro*, an additional 177 genes based on the barcoded data presented in Chapter 4 (1,000 reads per gene minimum). Other inclusion criteria were also imposed on the data as previously described in Chapter 4 section 4.2.4, after which 1,131 genes remained for subsequent analysis. These 1,131 genes were represented by on average 35,597 ± 4,325 (SEM) unique mutants per input library (Table 5.1, Figure 5.1).

Table 5.1. Composition of libraries used to experimentally challenge 12 Welsh mountain ponies pre- and post-filtering. The number of genes containing insertions post-filtering is consistent between libraries, since filtering determines a consensus set of genes to be taken forward for analysis. Genes represented by < 2,000 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Library	Unique insertion sites in genes	Library saturation (insertion every n bp in genes)	Genes containing insertions (% of total genes : % of non-essential genes)
AC^{pre}	42,964	45	1,929 (89.1 : 100)
CT^{pre}	39,333	49	1,956 (90.3 : 100)
GA^{pre}	57,338	34	1,937 (89.5 : 100)
AC^{post}	32,741	69	1,131 (52.2 : 71.1)
CT^{post}	29,953	75	1,131 (52.2 : 71.1)
GA^{post}	44,097	51	1,131 (52.2 : 71.1)

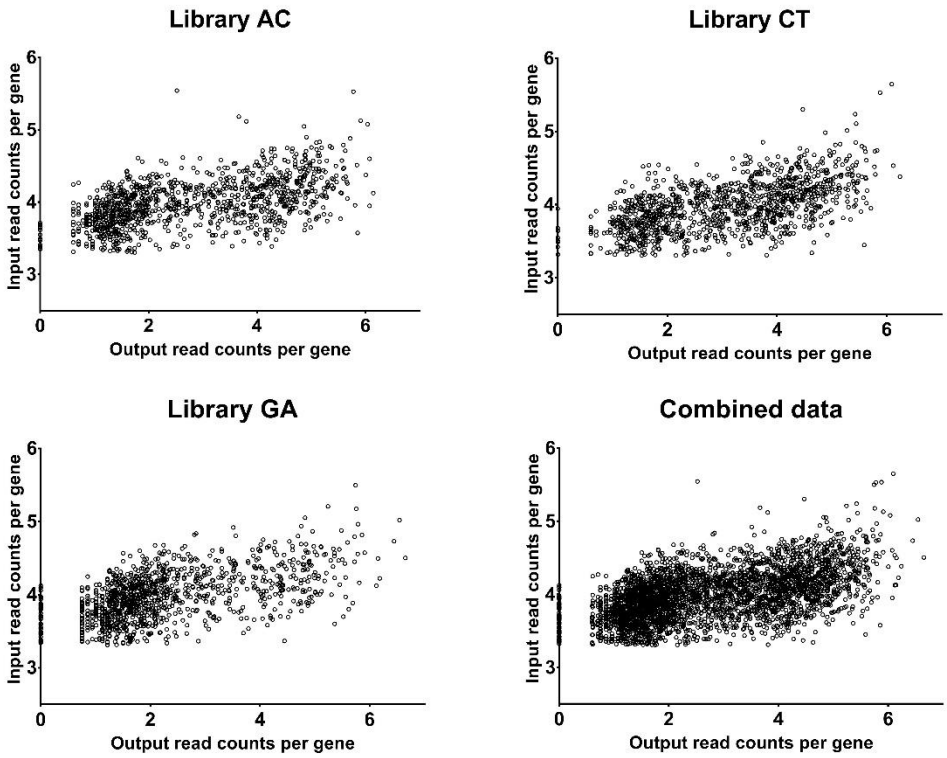


Figure 5.1. Read counts per gene in each of 3 *S. equi* barcoded IS*S1* libraries, pre- (input) and post- (output) infection of 12 Welsh mountain ponies. Genes represented by < 2,000 reads, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Analysis of the 1,131 permitted genes in the output libraries, identified on average 10,914 ± 2,048 (SEM) unique mutants per library (Table 5.2). These recovered mutants represented 30.1 percent of the mutants within the challenge inoculum and 99.7 percent ± 2.1 percent (SEM) of 1,131 *S. equi* genes meeting the input pool inclusion criteria (Table 5.2, Figure 5.1).

Table 5.2. Composition of barcoded libraries recovered from 12 Welsh mountain ponies pre- and post-filtering. Mutants were recovered from up to 4 lymph nodes per animal, data combined, split according to barcode and analysed before determining gene fitness. Genes represented by < 2,000 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC^{pre}	13,792	50,477,885	1,829 (84.5 : 100)
CT^{pre}	19,922	64,419,073	1,894 (87.7 : 100)
GA^{pre}	11,024	69,425,389	1,764 (81.5 : 100)
AC^{post}	10,207	34,293,471	1,117 (51.6 : 70.3)
CT^{post}	14,762	34,293,471	1,122 (51.8 : 70.6)
GA^{post}	7,774	34,293,471	1,091 (50.4 : 68.6)

Gene fitness was calculated by comparing the ratio (log_2FC) of read counts, per gene, in the 3 output pools to the 3 input pools. Analysis identified 297 genes required for fitness ($log_2FC < -2$, $q < 0.05$). Further analysis identified 74 genes in which IS*S1* insertion conferred a fitness advantage ($log_2FC > 2$, $q < 0.05$).

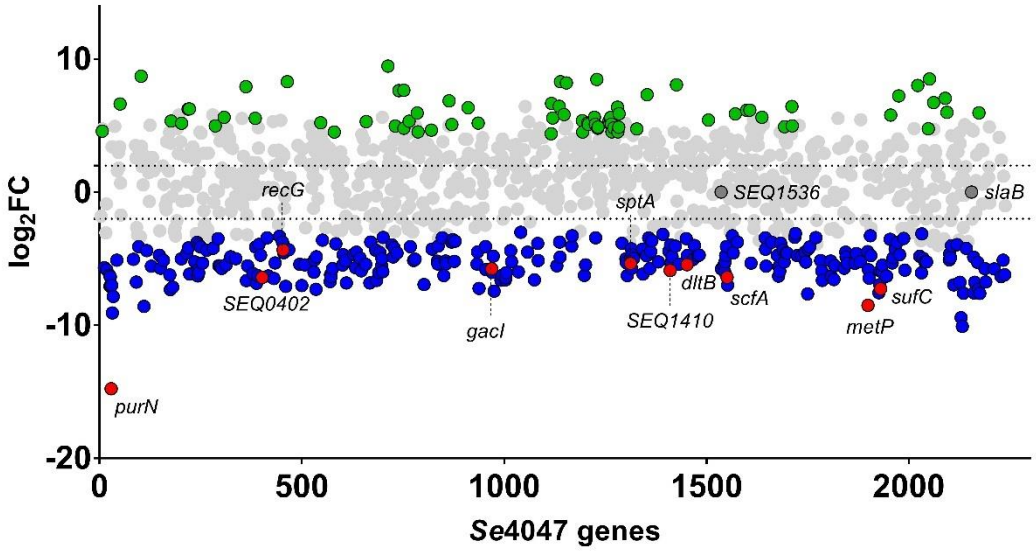


Figure 5.2. Genome-wide fitness of each *S. equi* gene *in vivo* determined by a barcoded technique. A 2,000 reads per gene inclusion criteria was imposed on the input libraries to limit false positive gene identification. Blue dots indicate 279 genes required for fitness ($log_2FC < -2$, $q < 0.05$), red and dark grey dots indicate a panel of genes required for fitness included in a validation study. Dark grey dots (*SEQ1536* and *slaB*) indicate validation mutants removed from the analysis when the gene inclusion criteria was increased from 1,000 reads per gene to 2,000 reads per gene. Green dots indicate 74 genes conferring an enhanced fitness upon insertion ($log_2FC > 2$, $q < 0.05$), and light grey dots indicate genes non-essential to *in vivo* fitness.

5.3.2 Genes implicated in the survival of *S. equi in vivo* using a gene inclusion criterion of 5,000 reads per gene minimum

Enforcing the 5,000 minimum reads per gene criteria permitted the exclusion of 820 genes previously identified as non-essential *in vitro*, an additional 592 genes based on the barcoded data presented in Chapter 4 (1,000 reads per gene minimum) and an additional 415 genes compared to the 2,000 reads per gene minimum criteria.

Other inclusion criteria were also imposed on the data as previously described in Chapter 4 section 4.2.4, after which 719 genes remained for subsequent analysis. These 719 genes were represented by on average 29,105 ± 3,483 (SEM) unique mutants per input library (Table 5.3, Figure 5.2).

Table 5.3. Composition of libraries used to experimentally challenge 12 Welsh mountain ponies pre- and post-filtering. The number of genes containing insertions post-filtering is consistent between libraries, since filtering determines a consensus set of genes to be taken forward for analysis. Genes represented by < 5,000 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Library	Unique insertion sites in genes	Library saturation (insertion every n bp in genes)	Genes containing insertions (% of total genes : % of non-essential genes)
AC^{pre}	42,964	45	1,929 (89.1 : 100)
CT^{pre}	39,333	49	1,956 (90.3 : 100)
GA^{pre}	57,338	34	1,937 (89.5 : 100)
AC^{post}	26,704	84	719 (33.2 : 45.2)
CT^{post}	24,641	86	719 (33.2 : 45.2)
GA^{post}	35,969	63	719 (33.2 : 45.2)

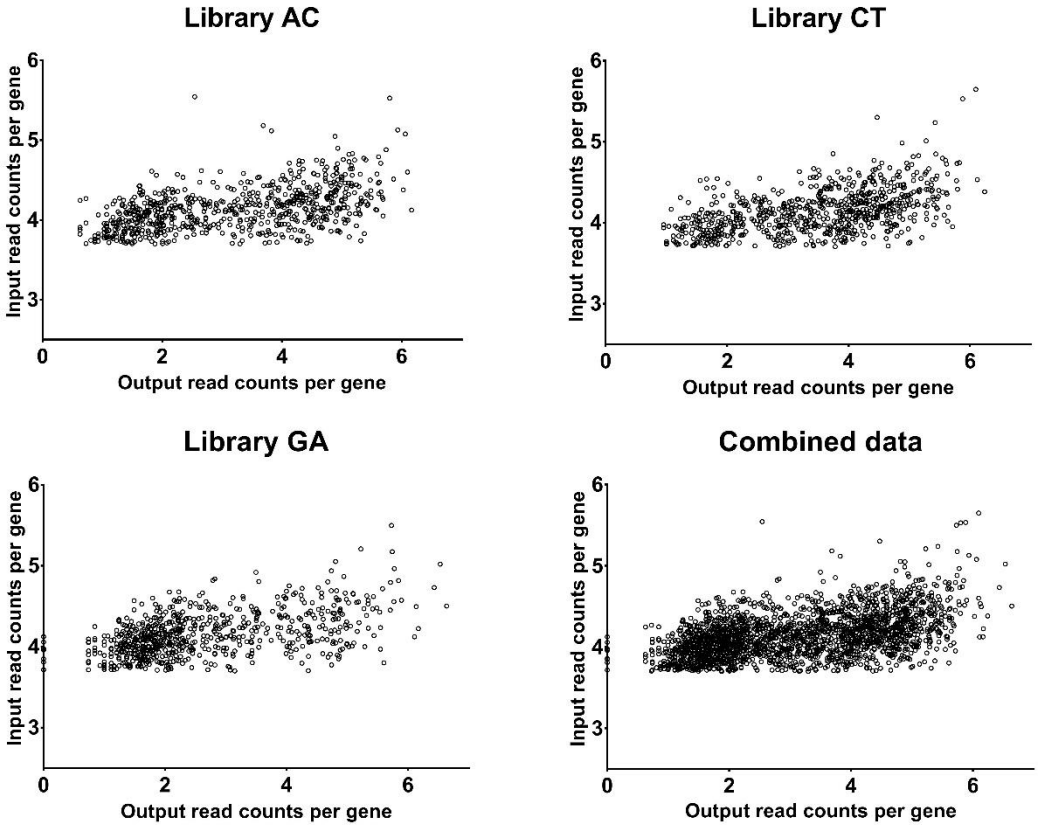


Figure 5.3. Read counts per gene in each of 3 *S. equi* barcoded IS*S1* libraries, pre- (input) and post- (output) infection of 12 Welsh mountain ponies. Genes represented by < 5,000 reads, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Using this more stringent inclusion criteria removed a large amount of data and is likely to confound the identification of true positive data. Analysis of the 719 permitted genes in the output libraries, however, identified on average 10,914 ± 2,048 (SEM) unique mutants per library (Table 5.4). These recovered mutants represented 30.1 percent of the mutants within the challenge inoculum and 99.7 percent ± 2.1 percent (SEM) of 1,131 *S. equi* genes meeting the input pool inclusion criteria (Table 5.4, Figure 5.1).

Table 5.4. Composition of barcoded libraries recovered from 12 Welsh mountain ponies pre- and post-filtering. Mutants were recovered from up to 4 lymph nodes per animal, data combined, split according to barcode and analysed before determining gene fitness. Genes represented by < 5,000 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
AC^{pre}	13,792	50,477,885	1,829 (84.5 : 100)
CT^{pre}	19,922	64,419,073	1,894 (87.7 : 100)
GA^{pre}	11,024	69,425,389	1,764 (81.5 : 100)
AC^{post}	8,405	30,496,451	719 (33.2 : 45.2)
CT^{post}	12,268	30,496,451	719 (33.2 : 45.2)
GA^{post}	6,394	30,496,451	711 (32.8 : 44.7)

Gene fitness was calculated by comparing the ratio (log_2FC) of read counts, per gene, in the 3 output pools to the 3 input pools. Analysis identified 152 genes required for fitness ($log_2FC < -2$, $q < 0.05$). Further analysis identified 21 genes in which IS*S1* insertion conferred a fitness advantage ($log_2FC > 2$, $q < 0.05$).

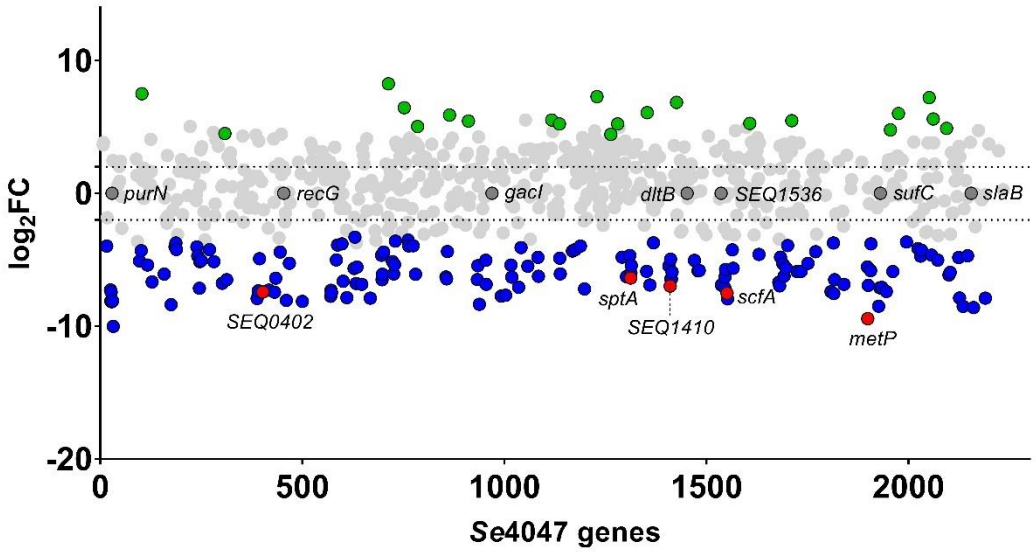


Figure 5.4. Genome-wide fitness of each *S. equi* gene *in vivo* determined by a barcoded technique. A 5,000 reads per gene inclusion criteria was imposed on the input libraries to limit false positive gene identification. Blue dots indicate 152 genes required for fitness ($log_2FC < -2$, $q < 0.05$), red and dark grey dots indicate a panel of required fitness genes included in a validation study. Dark grey dots indicate mutants that were removed from the analysis when the gene inclusion criteria was increased from 1,000 reads per gene to 5,000 reads per gene. Green dots indicate 21 genes conferring an enhanced fitness upon insertion ($log_2FC > 2$, $q < 0.05$), and light grey dots indicate genes non-essential to *in vivo* fitness.

5.3.3 Comparison of genes implicated in the survival of *S. equi* identified using 3 different gene inclusion stringencies

Using a 1,000 reads per gene inclusion criteria as described in Chapter 4 when analysing the barcoded data, 27.9 percent (368/ 1,319) of the genes analysed were identified as required for fitness. When considering genes conferring a fitness advantage upon insertion, 6.4 percent (85/ 1,319) of genes included in the analysis were identified. Increasing the stringency of inclusion to 2,000 reads per gene minimum did not dramatically alter the overall percent identification of genes required for fitness (26.3 percent, 279/1,131) or those conferring a fitness advantage 6.5 percent (74/ 1,131). Increasing the stringency even higher to 5,000 reads per gene, more dramatically reduced the number of genes included in the analysis and the proportion of these genes identified as required for fitness (21.1 percent, 152/ 719) and those conferring a fitness advantage upon insertion 2.9 percent (21/ 719).

As the minimum read counts per gene imposed on the input pools is increased, the number of small genes identified as required for fitness is decreased. Increasing the stringency from 1,000 reads to 2,000 reads in the input pools reduces the number of genes < 500 bp included in the analysis by 51.3 percent, however the proportion of these genes compared to the total number of required genes is only reduced by 11 percent (Table 5.5). Increasing the stringency again from 2,000 reads to 5,000 reads in the input pools reduces the number of shorter genes (< 500 bp) included in the analysis by 81.8 percent (Table 5.5). As the < 500 bp group of genes is most affected when the stringency is increased, it could be concluded that smaller genes are under represented in the input pools and may bias the calculated *in vivo* gene essentialities. Many of these excluded genes, however, may provide accurate and valuable data that is missed upon their exclusion.

Table 5.5. Size of *S. equi* genes identified as required for fitness *in vivo* by TraDIS using different minimum read count per gene stringencies imposed on the input pools.

Stringency	No. of genes < 500 bp (% of total)	No. of genes 500 bp-1 kb (% of total)	No. of genes > 1 kb (% of total)
< 1,000 reads per gene	113 (30.7)	157 (42.7)	98 (26.6)
< 2,000 reads per gene	55 (19.7)	126 (45.2)	98 (35.1)
< 5,000 reads per gene	10 (6.6)	64 (42.1)	78 (51.3)

Comparison of the genes required for fitness determined by the 3 analysis methods, identified a core set of 113 genes (null= 1.5 genes) (Figure 5.5, Table A3.1, Appendix 3). Comparison between the 1,000 reads per gene inclusion criteria and the 2,000 reads per gene criteria identified 226 consensus genes (null= 1.7 genes) (Table A3.1, Appendix

3), with just 13 of the genes identified by the 2,000 criteria not present in the 1,000 reads criteria dataset (Figure 5.5).

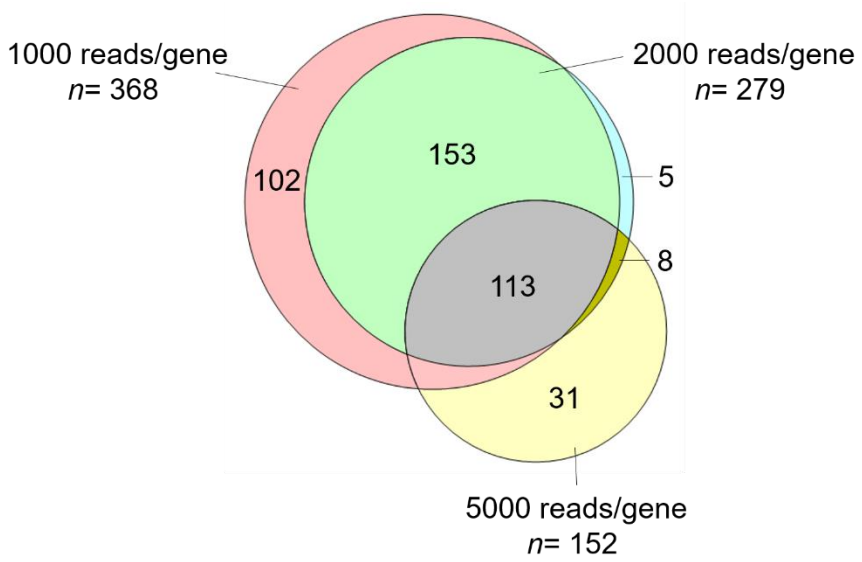


Figure 5.5. Venn diagram comparing the *S. equi* genes identified as required for fitness *in vivo* when a gene inclusion criterion of either 1,000, 2,000 or 5,000 reads per gene minimum are enforced on the input libraries. One hundred and thirteen consensus genes are required for fitness in all 3 analysis techniques.

Of the 279 fitness genes identified using the 2,000 reads per gene criteria, 10 of the 12 genes included in the validation experiment described in Chapter 4 were similarly defined as essential *in vivo*. The two genes that were not identified were *SEQ1536* and *slaB.* Some reads relating to the *slaB* mutant were recovered from the 5 ponies infected with the validation mutant pool (0.52 percent of all reads sequencing in output pools), but no reads corresponding to the *SEQ1536* mutant were recovered from any animals.

Comparison of the genes conferring a fitness advantage upon insertion revealed a core set of 21 genes between the 3 datsets (null= 1.5 genes) (Figure 5.6). All 21 genes conferring a fitness advantage upon insertion that were identified using the 5,000 reads per gene inclusion criteria were similarly identified by the other analyses (Figure 5.6). Only 1 gene was uniquely identified by the 2,000 reads per gene criteria compared to the 1,000 reads per gene criteria analysis (Figure 5.6).

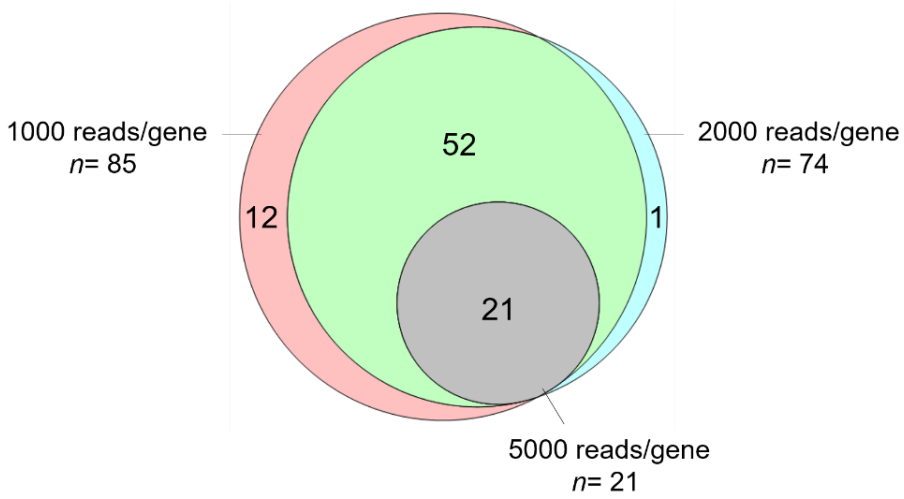


Figure 5.6. Venn diagram comparing the *S. equi* genes that enhance fitness *in vivo* upon transposon insertion, when a gene inclusion criterion of either 1,000, 2,000 or 5,000 reads per gene minimum are enforced on the input libraries. Twenty-one consensus genes enhance fitness when inserted into in all 3 analysis techniques.

5.3.4 Comparison of genes implicated in the survival of *S. equi* identified using 3 barcoded output libraries and 3 random pony groups

Sorting and then randomly assigning data from the 12 ponies into groups ensured that each group was represented by a consistent sequencing depth as libraries were previously allocated a proportion of the sequencing run in Chapter 4 depending on the density of unique mutants. Group 1 contained data from the ponies 6544, 7799, 7616 and 6061, Group 2; 5922, 2991, 7454 and 7649. Group 3; 5867, 7565, 7884 and 477.

Using the same inclusion criteria as outlined in Chapter 4 for the barcoded analysis, 1,319 genes were analysed for fitness in the 3 random groups of ponies. Analysis of these 1,319 genes identified on average 11,249 ± 549 (SEM) unique mutants per library (Table 5.6). These recovered mutants represented 29.1 percent of the mutants within the challenge inoculum (Table 4.7) and 96.4 percent ± 0.2 percent (SEM) of 1,319 *S. equi* genes meeting the input pool inclusion criteria (Table 5.6).

Table 5.6. Composition of barcoded libraries recovered from 12 Welsh mountain ponies randomly combined into 3 groups of ponies pre- and post-filtering. Mutants were recovered from up to 4 lymph nodes per animal, data from each animal sorted into 1 of 4 groups depending on mutant diversity, 1 pony randomly selected from each group and placed in a new group to form 3 datasets containing data pools of varying diversities. These 3 datasets were then analysed before determining gene fitness. Genes represented by < 1,000 reads in the input libraries, previously identified as essential *in vitro* or were over-represented in the input or output libraries, were removed from the analysis. Reads mapping in the last 10 percent of genes were also not considered.

Library	Unique insertion sites in genes	Total read count	Genes containing insertions (% of total genes : % of non-essential genes)
Group1^{pre}	12,120	43,839,909	1,770 (81.8 : 100)
Group2^{pre}	13,150	58,089,064	1,815 (83.8 : 100)
Group3^{pre}	14,469	68,567,535	1,817 (83.9 : 100)
Group1^{post}	10,402	36,629,746	1,269 (58.6 : 79.8)
Group2^{post}	11,068	36,629,746	1,268 (58.6 : 79.7)
Group3^{post}	12,276	36,629,746	1,278 (59 : 80.4)

Gene fitness was calculated by comparing the ratio (log_2FC) of read counts, per gene, in the 3 output pools to the 3 input pools. Analysis identified 378 genes required for fitness ($log_2FC < -2$, $q < 0.05$) (Table A3.1, Appendix 3). Further analysis identified 97 genes in which IS*S1* insertion conferred a fitness advantage ($log_2FC > 2$, $q < 0.05$) (Table A3.2, Appendix 3). No input/output plots could be drawn for this data since the 3 random pony groups could not be directly attributed to 1 of the 3 input libraries.

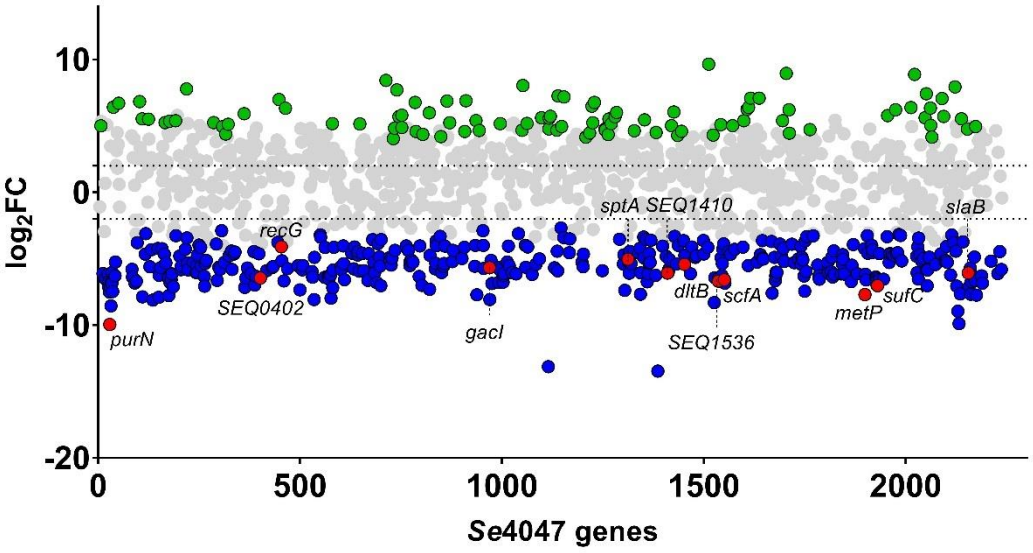


Figure 5.7. Genome-wide fitness of each *S. equi* gene *in vivo* determined by a random pony grouping technique. Blue dots indicate 378 genes required for fitness ($log_2FC < -2$, $q < 0.05$), red dots indicate a panel of required fitness genes included in a validation study. Green dots indicate 97 genes conferring an enhanced fitness upon insertion ($log_2FC > 2$, $q < 0.05$), and light grey dots indicate genes non-essential to *in vivo* fitness.

Comparison of the barcoded data to the random pony group data revealed that 357 genes required for fitness (null= 1.9 genes) (97 percent and 94 percent of the barcoded and random pony group fitness genes, respectively) were similarly identified (Figure 5.6). These findings reflect the continuity of data obtained regardless of how the data is combined to limit the effects of stochastic loss and animal variation. The 357 consensus genes represent those that can be considered with most confidence since their requirements for fitness were identified in the 2 independent analyses. All of the genes investigated in the validation study conducted in Chapter 4 were also identified as fitness genes in this 357-consensus gene set.

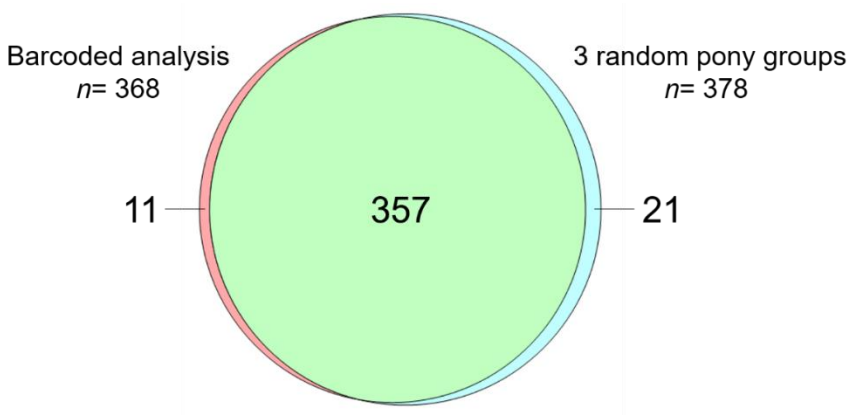


Figure 5.8. Venn diagram comparing the *S. equi* genes identified as required for fitness *in vivo* when output libraries are combined and deconvoluted on barcoded basis or when they are combined randomly into 3 groups of 4 ponies. Three hundred and fifty-seven genes are required for fitness in both analysis techniques.

Comparison of the genes conferring a fitness advantage upon insertion in the barcoded and random pony group analysis revealed a consensus set of 66 genes (null= 1.5 genes) (78 percent and 68 percent of the barcoded and random pony group fitness genes, respectively) (Figure 5.7). The proportion of consensus data is not as prominent as when considering genes required for fitness between these 2 datsets and is therefore likely impacted by the method used to combine the data. This suggests that large over representations of genes in the output pools is likely to skew the data and will reach statisitical significance depending on how the data is combined. Previous discussion (Chapter 4) regarding the questionable reliability of the enhanced fitness data is supported by these findings.

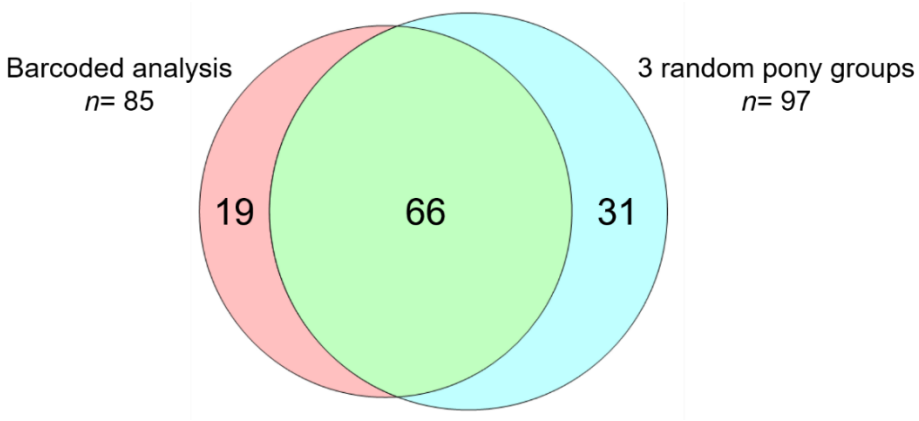


Figure 5.9. Venn diagram comparing the *S. equi* genes identified as enhancing fitness as a result of IS*S1 in vivo* when output libraries are combined and deconvoluted on barcoded basis or when they are combined randomly into 3 groups of 4 ponies. Sixty-six genes confer enhanced fitness in both analysis techniques.

5.4 Discussion

Reanalysis of the barcoded data was completed to investigate the impact of more stringent gene inclusion criteria on the identification of genes that contribute to survival of *S. equi* in the natural host. Increasing the minimum read count per gene in the input libraries to 2,000 from 1,000 only marginally reduced the total proportion of genes identified as required for infection by 1.6 percent (24.2 percent reduction in total number of required genes identified). A consensus set of 226 genes were identified between the 2 analyses, but this did not include 2 of the 12 genes investigated in the validation study described in Chapter 4, allelic replacement mutants for which appear to be reduced in fitness in ponies.

Using the less stringent minimum read count per gene of 1,000 may enhance the likelihood of identifying false positive data. However, increasing the minimum criteria to either 2,000 or 5,000 reads per gene limits the ability to measure the *in vivo* fitness of many non-essential *S. equi* genes, since many are excluded. TraDIS was designed as an intial screen to direct further experiments using single gene deletion mutants and the increased chance of detetcing false positives within the 1,000 read count inclusion is a reasonable compromise between losing potentially correct data and limiting the detection of false positivies. For example, the quorum sensing genes, *pptAB*, described in Chapter 4 section 4.4.1 and confirmed to be required for infection in the *S. pyogenes* NHP model of nectrotising myositis [145] were also required for infection in *S. equi* according to the analysis utilising the 1,000 read per gene minimum count. However, using the 2,000 reads per gene criteria, *pptA* was removed from the analysis and both of these genes were removed when the stringency was increased to 5,000 reads per gene.

To investigate the usefulness of a barcoded approach, data recovered from the 12 ponies was sorted into 4 groups depending on the diversity of mutants recovered from each animal. One pony from each group was then randomly selected and placed into one of three new groups. The three new groups formed in this way each contained 4 ponies with varying mutant diversities and therefore, more balanced sequencing depths than if ponies were selected at random before intial sorting. Comparison of the *in vivo* gene essentialities determined by the barcoded and the random pony groups data identified that 97 and 94 percent, respectively, of the genes required for fitness were similarly determined and included all 12 genes investigated in the validation study conducted in Chapter 4.

The high proportion of consensus genes observed between these datasets confirms that the barcoded analysis represents the overall population of mutants infecting the ponies well, since virtually the same data was obtained regardless of whether the output data

was split randomly into 3 pools or by barcode. The 357-consensus set of genes are those that are likely to be the most reliable. Repeating the random pony group analysis several times and comparing all the datasets obtained would enable a robust core set of essential genes, but it would be expected that the number of genes within this consensus set would not alter significantly to that reported in this Chapter.

Comparison of the genes conferring an increased fitness as a result of IS*S1* insertion determined by the barcoded and random pony group analysis, was not as consistent compared to the genes required for fitness. The consensus data within these datasets was reduced to 78 and 68 percent for the barcoded and random pony group data, respectively. This finding highlights the potential inconsistency between animals and likely jackpot effect where certain unique mutants are able to reach the lymph nodes first, by chance. Repeating the random pony group analysis is likely to identify new genes pertaining to enhanced fitness, reflecting the likely inconsistency between animals.

In conclusion, using a less stringent inclusion criteria on the input libraries is a reasonable balance between identifying truly essential genes *in vivo* and identifying false positives. TraDIS results taken forward for further investigation should be confirmed using deletion mutants. Imposing the inclusion criteria of 1,000 reads per gene minimum on the input pools goes some way to minimising the potential effects of false positives, but allows some potentially important genes to remain in the analysis.

Additionally, using a barcoded approach to identify genes required for fitness has proven to be consistent with using 3 random groups of ponies. This demonstrates that recovering all mutants able to cause disease in animals and splitting back into parental barcode represents the overall population of mutants infecting all 12 animals well and that the libraries have behaved consistently between these animals since the parental libraries were not considered when the random pony groups were generated.

6 Discussion

TraDIS and other transposon directed methods, represent a major advance in the study of gene function in bacteria. Utilising dense mutant libraries yields significant time and cost savings over the generation of traditional knockout strains, not only due to the speed at which saturated libraries can be generated, but also due to the ability to simultaneously identify conditionally essential genes. The use of barcoded pGh9:IS*S1* plasmids to generate mutant libraries of *S. equi* has provided a highly useful tool for the progression of TraDIS studies in this important bacterium. In particular, the ability to combine barcoded mutant libraries, challenge animals and then deconvolute the data generated minimises the effects of animal to animal variation, enhances data quality and reduces the total number of animals required in future studies in accordance with the principles of the 3Rs: replacement, reduction and refinement.

Identifying the essential genome of *S. equi* was a necessary task before conditional fitness genes could be identified. Without an essential gene set, genes pertaining to specific niche fitness cannot be confidently assigned. Fitness genes specific to survival in whole equine blood and H_2O_2 were identified, in which 94 percent of H_2O_2 fitness genes were similarly identified in whole equine blood. These *in vitro* fitness genes were compared to those identified *in vivo* in the natural host. Relatively high overlaps between the *in vitro* and *in vivo* conditions were observed, yet many more *in vivo* fitness genes were identified reflecting the complex nature of host infection and the difficulty faced in accurately replicating an *in vivo* condition *in vitro*.

Comparisons of TraDIS/Tn-seq data from different streptococcal species throughout this thesis reflects the close genetic relationships and the importance of collaborative works in these species, since many interesting genes were commonly identified. The pGh9:IS*S1* insertion system described in this thesis was transferred to the Houston Methodist Hospital for use in *S. pyogenes* to define fitness genes in human saliva and in

non-human primates. These collaborative works have yielded some extremely relevant data for the future development of both *S. equi* and *S. pyogenes* therapeutics and vaccines as it has provided a cross-species data set enabling comparative analysis and validation. Common identification of particular genes or operons that contribute to fitness instils confidence in the results identified, despite the distinct differences in the diseases these species cause to their respective hosts.

TraDIS screens are particularly useful for the identification of novel genes, which are annotated as hypothetical proteins or with a general function only. While TraDIS does not allude to the exact function of these genes, it does highlight their importance, promoting further investigations into their function. Two genes identified in this thesis as required for *in vivo* fitness, which would benefit from more in-depth functional studies, are *SEQ1410* and *SEQ1536*, encoding a putative branched-chain amino acid ABC transport ATP-binding protein and a putative exported protein, respectively.

Several potential vaccine targets were identified in this thesis; *pptAB*, *gacI*, *dltABD* and *scfAB*, that all function at the cell surface, either in defence/virulence mechanisms or in the transport of metabolism precursors scavenged from the environment. Making further deletions to the prototype *S. equi* live attenuated vaccine strain [13] may reduce the adverse injection site reactions seen after administration, improving the safety of this promising vaccine. The growth characteristics of the tagged deletion mutants generated in this thesis were not defined, so further investigation is required to further assess the viability of these genes as vaccine targets.

TraDIS is an incredibly useful technique, but isn't without limitations. Transposon mutants are measured in a pool where fitness effects may be due to the competitive nature of these mixed pools. This effect was evident in Chapter 3, where *pyrP* and *mnmE* deletion mutants were not consistently reduced in fitness as suggested by the TraDIS screens. TraDIS is also confounded by secreted elements as it is likely that mutants in important secreted factors can retain their virulence by benefiting from neighbouring cells still able to produce and secrete these elements. For example, none of the genes contributing to equibactin production and transport, a known virulence determinant *in vivo*, were identified as important. To test the confounding nature of secreted elements, a deletion mutant lacking a gene encoding a known secreted factor, not identified in a TraDIS screen, should be attenuated in isolation. *slaB*, however, which encodes a secreted phospholipase A_2 toxin, was identified by TraDIS and the Δ*slaB*tagC mutant was attenuated in ponies. These data suggest that SlaB may act more locally than the other potentially confounded secreted factors.

To recover surviving mutants from a TraDIS screen, whether it be *in vitro* or *in vivo*, mutants must be grown on agar. The effect of this on the subsequent data has not been investigated since growth on agar may select for mutants better at growing on this medium, skewing the final data. In future, accounting for 'agar fitness' in the data analysis may mediate against these potential artifacts.

In future, *in vitro* TraDIS experiments should be conducted by combining 3 or more libraries to maximise the potential of the library barcodes. In this thesis, libraries were treated individually and therefore the barcodes were not fully utilised. The *in vivo* TraDIS data was also not fully explored. The library barcodes additionally enable mutants to be tracked from the nostril in which it was administered to its final destination. Further analysis of the data could enable identification of mutants that are better at transitioning to and infecting particular lymph nodes, however only 4 animals had abscesses in all 4 lymph nodes which is likely to confound analysis.

This thesis has generated a wealth of data for future research in not only *S. equi*, but in related species. The barcoded TraDIS technique improves on published TraDIS methods as it enhances the statistical power of *in vivo* studies by mediating against bottleneck effects and animal to animal variation, problems currently faced by researchers using TraDIS or TraDIS-like techniques. Barcoded TraDIS also has significant welfare benefits by reducing the number of animals required to produce this quality of data, since multiple libraries can be administered to 1 animal. Overall, the data presented in this thesis provides an unprecedented insight into the mechanisms employed by *S. equi* to cause disease in the natural host. The data also sheds light on the pan-streptococcal pathways important for virulence that are likely important for future development of novel therapeutics and vaccines.

References

1. Farrow, J.A.E. and M.D. Collins, *Taxonomic studies on Streptococci of serological Groups C, G and L and possibly related taxa*. 1984, Systematic and Applied Microbiology. p. 483-493.
2. Jorm, L.R., et al., *Genetic structure of populations of beta-haemolytic Lancefield group C streptococci from horses and their association with disease.* Res Vet Sci, 1994. **57**(3): p. 292-9.
3. Holden, M.T., et al., *Genomic evidence for the evolution of* Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens.* PLoS Pathog, 2009. **5**(3): p. e1000346.
4. Timoney, J.F. and P. Kumar, *Early pathogenesis of equine* Streptococcus equi *infection (strangles).* Equine Vet J, 2008. **40**(7): p. 637-42.
5. Harrington, D.J., I.C. Sutcliffe, and N. Chanter, *The molecular basis of* Streptococcus equi *infection and disease.* Microbes Infect, 2002. **4**(4): p. 501-10.
6. Newton, J.R., et al., *Control of strangles outbreaks by isolation of guttural pouch carriers identified using PCR and culture of* Streptococcus equi. Equine Vet J, 2000. **32**(6): p. 515-26.
7. Newton, J.R., et al., *Naturally occurring persistent and asymptomatic infection of the guttural pouches of horses with* Streptococcus equi*.* Vet Rec, 1997. **140**(4): p. 84-90.
8. Ford, J. and M. Lokai, *Complications of* Streptococcus equi *infection*. 1980, E*quine Practice*. p. 41-44.
9. Sweeney, C.R., et al., *Complications associated with* Streptococcus equi *infection on a horse farm.* J Am Vet Med Assoc, 1987. **191**(11): p. 1446-8.
10. Galan, J.E. and J.F. Timoney, *Immune complexes in purpura hemorrhagica of the horse contain IgA and M antigen of* Streptococcus equi*.* J Immunol, 1985. **135**(5): p. 3134-7.
11. Pointon, J.A., et al., *A highly unusual thioester bond in a pilus adhesin is required for efficient host cell interaction.* J Biol Chem, 2010. **285**(44): p. 33858-66.
12. Walker, J.A. and J.F. Timoney, *Construction of a stable non-mucoid deletion mutant of the* Streptococcus equi *Pinnacle vaccine strain.* Vet Microbiol, 2002. **89**(4): p. 311-21.
13. Robinson, C., et al., *Vaccination with a live multi-gene deletion strain protects horses against virulent challenge with* Streptococcus equi*.* Vaccine, 2015. **33**(9): p. 1160-7.
14. Lindsay, A.M., et al., *The* Streptococcus equi *prophage-encoded protein SEQ2045 is a hyaluronan-specific hyaluronate lyase that is produced during equine infection.* Microbiology, 2009. **155**(Pt 2): p. 443-9.
15. Harris, S.R., Robinson, C., Steward, K.F., Webb, K.S., Paillot, R., Parkhill, J., Holden, M.T.G. & Waller, A.S., *Genome specialization and decay of the strangles pathogen,* Streptococcus equi*, is driven by persistent infection.* 2015, *Genome Res*. p. 1360–1371.

16. Boschwitz, J.S. and J.F. Timoney, *Characterization of the antiphagocytic activity of equine fibrinogen for* Streptococcus equi *subsp. equi.* Microb Pathog, 1994. **17**(2): p. 121-9.
17. Boschwitz, J.S. and J.F. Timoney, *Inhibition of C3 deposition on* Streptococcus equi *subsp. equi by M protein: a mechanism for survival in equine blood.* Infect Immun, 1994. **62**(8): p. 3515-20.
18. Muhktar, M.M. and J.F. Timoney, *Chemotactic response of equine polymorphonuclear leucocytes to* Streptococcus equi*.* Res Vet Sci, 1988. **45**(2): p. 225-9.
19. Timoney, J.F., S.C. Artiushin, and J.S. Boschwitz, *Comparison of the sequences and functions of* Streptococcus equi *M-like proteins SeM and SzPSe.* Infect Immun, 1997. **65**(9): p. 3600-5.
20. Parkinson, N.J., et al., *Molecular epidemiology of strangles outbreaks in the UK during 2010.* Vet Rec, 2011. **168**(25): p. 666.
21. Kelly, C., et al., *Sequence variation of the SeM gene of* Streptococcus equi *allows discrimination of the source of strangles outbreaks.* J Clin Microbiol, 2006. **44**(2): p. 480-6.
22. Hulting, G., et al., *Two novel IgG endopeptidases of* Streptococcus equi*.* FEMS Microbiol Lett, 2009. **298**(1): p. 44-50.
23. Lannergård, J. and B. Guss, *IdeE, an IgG-endopeptidase of* Streptococcus equi *ssp. equi.* FEMS Microbiol Lett, 2006. **262**(2): p. 230-5.
24. Guss, B., et al., *Getting to grips with strangles: an effective multi-component recombinant vaccine for the protection of horses from* Streptococcus equi *infection.* PLoS Pathog, 2009. **5**(9): p. e1000584.
25. Robinson, C., et al., *Strangvac: A recombinant fusion protein vaccine that protects against strangles, caused by* Streptococcus equi*.* Vaccine, 2018. **36**(11): p. 1484-1490.
26. Timoney, J.F., et al., *IdeE reduces the bactericidal activity of equine neutrophils for* Streptococcus equi*.* Vet Immunol Immunopathol, 2008. **122**(1-2): p. 76-82.
27. Lindmark, H., M. Nilsson, and B. Guss, *Comparison of the fibronectin-binding protein FNE from* Streptococcus equi *subspecies equi with FNZ from S. equi subspecies zooepidemicus reveals a major and conserved difference.* Infect Immun, 2001. **69**(5): p. 3159-63.
28. Tiouajni, M., et al., *Structural and functional analysis of the fibronectin-binding protein FNE from* Streptococcus equi *spp. equi.* FEBS J, 2014. **281**(24): p. 5513-31.
29. Flanagan, J., et al., *Characterization of the haemolytic activity of* Streptococcus equi*.* Microb Pathog, 1998. **24**(4): p. 211-21.
30. Wannamaker, L.W., *Streptococcal toxins.* Rev Infect Dis, 1983. **5 Suppl 4**: p. S723-32.
31. Jantsch, J., et al., *Severe soft tissue infection caused by a non-beta-hemolytic* Streptococcus pyogenes *strain harboring a premature stop mutation in the sagC gene.* J Clin Microbiol, 2013. **51**(6): p. 1962-5.
32. Sbarra, A.J. and M.L. Karnovsky, *The biochemical basis of phagocytosis. I. Metabolic changes during the ingestion of particles by polymorphonuclear leukocytes.* J Biol Chem, 1959. **234**(6): p. 1355-62.
33. Bannister, J.V., W.H. Bannister, and G. Rotilio, *Aspects of the structure, function, and applications of superoxide dismutase.* CRC Crit Rev Biochem, 1987. **22**(2): p. 111-80.
34. Poyart, C., et al., *Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase.* J Clin Microbiol, 1998. **36**(1): p. 41-7.
35. Gerlach, D., W. Reichardt, and S. Vettermann, *Extracellular superoxide dismutase from* Streptococcus pyogenes *type 12 strain is manganese-dependent.* FEMS Microbiol Lett, 1998. **160**(2): p. 217-24.
36. McMillan, D.J., et al., *Immune response to superoxide dismutase in group A streptococcal infection.* FEMS Immunol Med Microbiol, 2004. **40**(3): p. 249-56.

37. Poyart, C., et al., *Contribution of Mn-cofactored superoxide dismutase (SodA) to the virulence of* Streptococcus agalactiae*.* Infect Immun, 2001. **69**(8): p. 5098-106.
38. Anzai, T., et al., *In vivo pathogenicity and resistance to phagocytosis of* Streptococcus equi *strains with different levels of capsule expression.* Vet Microbiol, 1999. **67**(4): p. 277-86.
39. Turner, C.E., et al., *Emerging role of the interleukin-8 cleaving enzyme SpyCEP in clinical* Streptococcus pyogenes *infection.* J Infect Dis, 2009. **200**(4): p. 555-63.
40. Edwards, R.J., et al., *Specific C-terminal cleavage and inactivation of interleukin-8 by invasive disease isolates of* Streptococcus pyogenes*.* J Infect Dis, 2005. **192**(5): p. 783-90.
41. Brüssow, H., C. Canchaya, and W.D. Hardt, *Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion.* Microbiol Mol Biol Rev, 2004. **68**(3): p. 560-602.
42. Granata, F., et al., *Secretory phospholipases A2 as multivalent mediators of inflammatory and allergic disorders.* Int Arch Allergy Immunol, 2003. **131**(3): p. 153-63.
43. Gräler, M.H. and E.J. Goetzl, *Lysophospholipids and their G protein-coupled receptors in inflammation and immunity.* Biochim Biophys Acta, 2002. **1582**(1-3): p. 168-74.
44. Sitkiewicz, I., K.E. Stockbauer, and J.M. Musser, *Secreted bacterial phospholipase A2 enzymes: better living through phospholipolysis.* Trends Microbiol, 2007. **15**(2): p. 63-9.
45. Sitkiewicz, I., et al., *Emergence of a bacterial clone with enhanced virulence by acquisition of a phage encoding a secreted phospholipase A2.* Proc Natl Acad Sci U S A, 2006. **103**(43): p. 16009-14.
46. Rubin, B.K., et al., *Secretory hyperresponsiveness and pulmonary mucus hypersecretion.* Chest, 2014. **146**(2): p. 496-507.
47. López-Álvarez, M.R., et al., *Immunogenicity of phospholipase A.* Vet Microbiol, 2017. **204**: p. 15-19.
48. Li, H., et al., *The structural basis of T cell activation by superantigens.* Annu Rev Immunol, 1999. **17**: p. 435-66.
49. Llewelyn, M. and J. Cohen, *Superantigens: microbial agents that corrupt immunity.* Lancet Infect Dis, 2002. **2**(3): p. 156-62.
50. Brown, J.S. and D.W. Holden, *Iron acquisition by Gram-positive bacterial pathogens.* Microbes Infect, 2002. **4**(11): p. 1149-56.
51. Heather, Z., et al., *A novel streptococcal integrative conjugative element involved in iron acquisition.* Mol Microbiol, 2008. **70**(5): p. 1274-92.
52. Waller, A.S., R. Paillot, and J.F. Timoney, Streptococcus equi*: a pathogen restricted to one host.* J Med Microbiol, 2011. **60**(Pt 9): p. 1231-40.
53. Timoney, J.F. and J.E. Galan, *The protective response of the horse to an avirulent strain of* Streptococcus equi. Recent Advances in Streptococci and Streptococcal Diseases, ed. Y. Kimura, S. Kotami, and Y. Shiokowa. 1985, Bracknell, UK: Reedbooks.
54. Borst, L.B., et al., *Evaluation of a commercially available modified-live* Streptococcus equi *subsp equi vaccine in ponies.* Am J Vet Res, 2011. **72**(8): p. 1130-8.
55. Cursons, R., et al., *Strangles in horses can be caused by vaccination with Pinnacle I. N.* Vaccine, 2015. **33**(30): p. 3440-3.
56. Du, W., et al., *Characterization of* Streptococcus pneumoniae *5-enolpyruvylshikimate 3-phosphate synthase and its activation by univalent cations.* Eur J Biochem, 2000. **267**(1): p. 222-7.
57. Jacobs, A.A., et al., *Investigations towards an efficacious and safe strangles vaccine: submucosal vaccination with a live attenuated* Streptococcus equi*.* Vet Rec, 2000. **147**(20): p. 563-7.

58. Kemp-Symonds, J., T. Kemble, and A. Waller, *Modified live* Streptococcus equi *('strangles') vaccination followed by clinically adverse reactions associated with bacterial replication.* Equine Vet J, 2007. **39**(3): p. 284-6.
59. Robinson, C., et al., *Combining two serological assays optimises sensitivity and specificity for the identification of* Streptococcus equi *subsp. equi exposure.* Vet J, 2013. **197**(2): p. 188-91.
60. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome.* Nat Rev Genet, 2007. **8**(4): p. 272-85.
61. Boeke, J.D. and V.G. Corces, *Transcription and reverse transcription of retrotransposons.* Annu Rev Microbiol, 1989. **43**: p. 403-34.
62. Lampe, D.J., M.E. Churchill, and H.M. Robertson, *A purified mariner transposase is sufficient to mediate transposition in vitro.* EMBO J, 1996. **15**(19): p. 5470-9.
63. Romero, D.A. and T.R. Klaenhammer, *Characterization of insertion sequence IS946, an Iso-ISS1 element, isolated from the conjugative lactococcal plasmid pTR2030.* J Bacteriol, 1990. **172**(8): p. 4151-60.
64. Berg, O.G., R.B. Winter, and P.H. von Hippel, *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.* Biochemistry, 1981. **20**(24): p. 6929-48.
65. Morisato, D., et al., *Tn10 transposase acts preferentially on nearby transposon ends in vivo.* Cell, 1983. **32**(3): p. 799-807.
66. Kitts, P.A., et al., *Transposon-encoded site-specific recombination: nature of the Tn3 DNA sequences which constitute the recombination site res.* EMBO J, 1983. **2**(7): p. 1055-60.
67. Hensel, M., et al., *Simultaneous identification of bacterial virulence genes by negative selection.* Science, 1995. **269**(5222): p. 400-3.
68. Langridge, G.C., et al., *Simultaneous assay of every* Salmonella Typhi *gene using one million transposon mutants.* Genome Res, 2009. **19**(12): p. 2308-16.
69. van Opijnen, T., K.L. Bodi, and A. Camilli, *Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.* Nat Methods, 2009. **6**(10): p. 767-72.
70. Goodman, A.L., et al., *Identifying genetic determinants needed to establish a human gut symbiont in its habitat.* Cell Host Microbe, 2009. **6**(3): p. 279-89.
71. Gawronski, J.D., et al., *Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for* Haemophilus *genes required in the lung.* Proc Natl Acad Sci U S A, 2009. **106**(38): p. 16422-7.
72. Blanchard, A.M., et al., *PIMMS (Pragmatic Insertional Mutation Mapping System) Laboratory Methodology a Readily Accessible Tool for Identification of Essential Genes in* Streptococcus*.* Front Microbiol, 2016. **7**: p. 1645.
73. Barquist, L., C.J. Boinett, and A.K. Cain, *Approaches to querying bacterial genomes with transposon-insertion sequencing.* RNA Biol, 2013. **10**(7): p. 1161-9.
74. Dembek, M., et al., *High-throughput analysis of gene essentiality and sporulation in* Clostridium difficile*.* MBio, 2015. **6**(2): p. e02383.
75. Wong, Y.C., et al., *Candidate Essential Genes in* Burkholderia cenocepacia *J2315 Identified by Genome-Wide TraDIS.* Front Microbiol, 2016. **7**: p. 1288.
76. Pechter, K.B., et al., *Essential Genome of the Metabolically Versatile Alphaproteobacterium* Rhodopseudomonas palustris*.* J Bacteriol, 2015. **198**(5): p. 867-76.
77. Hooven, T.A., et al., *The essential genome of* Streptococcus agalactiae*.* BMC Genomics, 2016. **17**: p. 406.
78. Le Breton, Y., et al., *Essential Genes in the Core Genome of the Human Pathogen* Streptococcus pyogenes*.* Sci Rep, 2015. **5**: p. 9838.
79. Moule, M.G., et al., *Genome-wide saturation mutagenesis of* Burkholderia pseudomallei *K96243 predicts essential genes and novel targets for antimicrobial development.* MBio, 2014. **5**(1): p. e00926-13.

80. Luan, S.L., et al., *Generation of a Tn5 transposon library in* Haemophilus parasuis *and analysis by transposon-directed insertion-site sequencing (TraDIS).* Vet Microbiol, 2013. **166**(3-4): p. 558-66.
81. Zhu, L., et al., *Novel Genes Required for the Fitness of* Streptococcus pyogenes *in Human Saliva.* mSphere, 2017. **2**(6).
82. Le Breton, Y., et al., *Genome-wide identification of genes required for fitness of group A* Streptococcus *in human blood.* Infect Immun, 2013. **81**(3): p. 862-75.
83. Chaudhuri, R.R., et al., *Comprehensive assignment of roles for Salmonella typhimurium genes in intestinal colonization of food-producing animals.* PLoS Genet, 2013. **9**(4): p. e1003456.
84. Grant, A.J., et al., *Genes Required for the Fitness of* Salmonella enterica *Serovar* Typhimurium *during Infection of Immunodeficient gp91-/- phox Mice.* Infect Immun, 2016. **84**(4): p. 989-97.
85. Subashchandrabose, S., et al., Acinetobacter baumannii *Genes Required for Bacterial Survival during Bloodstream Infection.* mSphere, 2016. **1**(1).
86. van Opijnen, T. and A. Camilli, *A fine scale phenotype-genotype virulence map of a bacterial pathogen.* Genome Res, 2012. **22**(12): p. 2541-51.
87. Le Breton, Y., et al., *Genome-wide discovery of novel M1T1 group A streptococcal determinants important for fitness and virulence during soft-tissue infection.* PLoS Pathog, 2017. **13**(8): p. e1006584.
88. Brockmeier, S.L., et al., *Use of Proteins Identified through a Functional Genomic Screen To Develop a Protein Subunit Vaccine That Provides Significant Protection against Virulent* Streptococcus suis *in Pigs.* Infect Immun, 2018. **86**(3).
89. Jacobson, J.W., M.M. Medhora, and D.L. Hartl, *Molecular structure of a somatically unstable transposable element in* Drosophila*.* Proc Natl Acad Sci U S A, 1986. **83**(22): p. 8684-8.
90. Akerley, B.J., et al., *Systematic identification of essential genes by in vitro mariner mutagenesis.* Proc Natl Acad Sci U S A, 1998. **95**(15): p. 8927-32.
91. Rubin, E.J., et al., *In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria.* Proc Natl Acad Sci U S A, 1999. **96**(4): p. 1645-50.
92. Otto, R., W.M. de Vos, and J. Gavrieli, *Plasmid DNA in* Streptococcus cremoris *Wg2: Influence of pH on Selection in Chemostats of a Variant Lacking a Protease Plasmid.* Appl Environ Microbiol, 1982. **43**(6): p. 1272-7.
93. Henrich, B., et al., *Food-grade delivery system for controlled gene expression in* Lactococcus lactis*.* Appl Environ Microbiol, 2002. **68**(11): p. 5429-36.
94. Maguin, E., et al., *Efficient insertional mutagenesis in lactococci and other gram-positive bacteria.* J Bacteriol, 1996. **178**(3): p. 931-5.
95. Leigh, J.A., et al., *Sortase anchored proteins of* Streptococcus uberis *play major roles in the pathogenesis of bovine mastitis in dairy cattle.* Vet Res, 2010. **41**(5): p. 63.
96. Thibessard, A., et al., *Transposition of pGh9:ISS1 is random and efficient in* Streptococcus thermophilus *CNRZ368.* Can J Microbiol, 2002. **48**(5): p. 473-8.
97. Duwat, P., et al., *Characterization of* Lactococcus lactis *UV-sensitive mutants obtained by ISS1 transposition.* J Bacteriol, 1997. **179**(14): p. 4473-9.
98. Hamilton, A., et al., *Mutation of the maturase lipoprotein attenuates the virulence of* Streptococcus equi *to a greater extent than does loss of general lipoprotein lipidation.* Infect Immun, 2006. **74**(12): p. 6907-19.
99. Gruss, A. and S.D. Ehrlich, *Insertion of foreign DNA into plasmids from gram-positive bacteria induces formation of high-molecular-weight plasmid multimers.* J Bacteriol, 1988. **170**(3): p. 1183-90.
100. Dabert, P., S.D. Ehrlich, and A. Gruss, *Chi sequence protects against RecBCD degradation of DNA in vivo.* Proc Natl Acad Sci U S A, 1992. **89**(24): p. 12073-7.
101. Slater, J.D., et al., *Mutagenesis of* Streptococcus equi *and* Streptococcus suis *by transposon Tn917.* Vet Microbiol, 2003. **93**(3): p. 197-206.
102. May, J.P., et al., *Development of an in vivo Himar1 transposon mutagenesis system for use in* Streptococcus equi *subsp. equi.* FEMS Microbiol Lett, 2004. **238**(2): p. 401-9.

103. Charbonneau, A.R.L., et al., *Defining the ABC of gene essentiality in streptococci.* BMC Genomics, 2017. **18**(1): p. 426.
104. Lefébure, T., et al., *Gene repertoire evolution of* Streptococcus pyogenes *inferred from phylogenomic analysis with* Streptococcus canis *and* Streptococcus dysgalactiae*.* PLoS One, 2012. **7**(5): p. e37607.
105. Olsen, R.J. and J.M. Musser, *Molecular pathogenesis of necrotizing fasciitis.* Annu Rev Pathol, 2010. **5**: p. 1-31.
106. Cunningham, M.W., *Pathogenesis of group A streptococcal infections and their sequelae.* Adv Exp Med Biol, 2008. **609**: p. 29-42.
107. Katz, A.R. and D.M. Morens, *Severe streptococcal infections in historical perspective.* Clin Infect Dis, 1992. **14**(1): p. 298-307.
108. Gibbs, R.S., S. Schrag, and A. Schuchat, *Perinatal infections due to group B streptococci.* Obstet Gynecol, 2004. **104**(5 Pt 1): p. 1062-76.
109. Bohnsack, J.F., et al., *Serotype III* Streptococcus agalactiae *from bovine milk and human neonatal infections.* Emerg Infect Dis, 2004. **10**(8): p. 1412-9.
110. Amal, M.N., et al., *Molecular characterization of* Streptococcus agalactiae *strains isolated from fishes in Malaysia.* J Appl Microbiol, 2013. **115**(1): p. 20-9.
111. Barquist, L., et al., *The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries.* Bioinformatics, 2016. **32**(7): p. 1109-11.
112. Carver, T., et al., *Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.* Bioinformatics, 2012. **28**(4): p. 464-9.
113. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
114. Baerends, R.J., et al., *Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data.* Genome Biol, 2004. **5**(5): p. R37.
115. Whiteside, M.D., et al., *OrthologueDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis.* Nucleic Acids Res, 2013. **41**(Database issue): p. D366-76.
116. Saier, M.H. and J. Reizer, *Proposed uniform nomenclature for the proteins and protein domains of the bacterial phosphoenolpyruvate: sugar phosphotransferase system.* J Bacteriol, 1992. **174**(5): p. 1433-8.
117. Shelburne, S.A., et al., *Molecular characterization of group A Streptococcus maltodextrin catabolism and its role in pharyngitis.* Mol Microbiol, 2008. **69**(2): p. 436-52.
118. Castro, R., et al., *Characterization of the individual glucose uptake systems of* Lactococcus lactis*: mannose-PTS, cellobiose-PTS and the novel GlcU permease.* Mol Microbiol, 2009. **71**(3): p. 795-806.
119. Gera, K., et al., *The phosphoenolpyruvate phosphotransferase system in group A* Streptococcus *acts to reduce streptolysin S activity and lesion severity during soft tissue infection.* Infect Immun, 2014. **82**(3): p. 1192-204.
120. Xu, P., et al., *Genome-wide essential gene identification in* Streptococcus sanguinis*.* Sci Rep, 2011. **1**: p. 125.
121. Vitko, N.P., et al., *Expanded Glucose Import Capability Affords* Staphylococcus aureus *Optimized Glycolytic Flux during Infection.* MBio, 2016. **7**(3).
122. Sham, L.T., et al., *Recent advances in pneumococcal peptidoglycan biosynthesis suggest new vaccine and antimicrobial targets.* Curr Opin Microbiol, 2012. **15**(2): p. 194-203.
123. Land, A.D. and M.E. Winkler, *The requirement for pneumococcal MreC and MreD is relieved by inactivation of the gene encoding PBP1a.* J Bacteriol, 2011. **193**(16): p. 4166-79.
124. Zapun, A., T. Vernet, and M.G. Pinho, *The different shapes of cocci.* FEMS Microbiol Rev, 2008. **32**(2): p. 345-60.
125. Thibessard, A., et al., *Effects of rodA and pbp2b disruption on cell morphology and oxidative stress response of* Streptococcus thermophilus *CNRZ368.* J Bacteriol, 2002. **184**(10): p. 2821-6.

126. Morlot, C., et al., *Growth and division of* Streptococcus pneumoniae*: localization of the high molecular weight penicillin-binding proteins during the cell cycle.* Mol Microbiol, 2003. **50**(3): p. 845-55.
127. David, B., et al., *PBP2b plays a key role in both peripheral growth and septum positioning in* Lactococcus lactis*.* PLoS One, 2018. **13**(5): p. e0198014.
128. Young, K.D., *Bacterial shape: two-dimensional questions and possibilities.* Annu Rev Microbiol, 2010. **64**: p. 223-40.
129. Claessen, D., et al., *Control of the cell elongation-division cycle by shuttling of PBP1 protein in* Bacillus subtilis*.* Mol Microbiol, 2008. **68**(4): p. 1029-46.
130. Philippe, J., T. Vernet, and A. Zapun, *The elongation of ovococci.* Microb Drug Resist, 2014. **20**(3): p. 215-21.
131. Shields, R.C., et al., *Genomewide Identification of Essential Genes and Fitness Determinants of* Streptococcus mutans *UA159.* mSphere, 2018. **3**(1).
132. Fittipaldi, N., et al., *Full-genome dissection of an epidemic of severe invasive disease caused by a hypervirulent, recently emerged clone of group A* Streptococcus*.* Am J Pathol, 2012. **180**(4): p. 1522-34.
133. Campbell, J.W. and J.E. Cronan, *Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery.* Annu Rev Microbiol, 2001. **55**: p. 305-32.
134. Zhang, Y.M. and C.O. Rock, *Membrane lipid homeostasis in bacteria.* Nat Rev Microbiol, 2008. **6**(3): p. 222-33.
135. Jerga, A. and C.O. Rock, *Acyl-Acyl carrier protein regulates transcription of fatty acid biosynthetic genes via the FabT repressor in* Streptococcus pneumoniae*.* J Biol Chem, 2009. **284**(23): p. 15364-8.
136. Schlieker, C., B. Bukau, and A. Mogk, *Prevention and reversion of protein aggregation by molecular chaperones in the E. coli cytosol: implications for their applicability in biotechnology.* J Biotechnol, 2002. **96**(1): p. 13-21.
137. Schulz, A. and W. Schumann, *hrcA, the first gene of the* Bacillus subtilis *dnaK operon encodes a negative regulator of class I heat shock genes.* J Bacteriol, 1996. **178**(4): p. 1088-93.
138. Zuber, U. and W. Schumann, *CIRCE, a novel heat shock element involved in regulation of heat shock operon dnaK of* Bacillus subtilis*.* J Bacteriol, 1994. **176**(5): p. 1359-63.
139. Erbse, A., M.P. Mayer, and B. Bukau, *Mechanism of substrate recognition by Hsp70 chaperones.* Biochem Soc Trans, 2004. **32**(Pt 4): p. 617-21.
140. Hartl, F.U. and M. Hayer-Hartl, *Molecular chaperones in the cytosol: from nascent chain to folded protein.* Science, 2002. **295**(5561): p. 1852-8.
141. Economou, A., et al., *SecA membrane cycling at SecYEG is driven by distinct ATP binding and hydrolysis events and is regulated by SecD and SecF.* Cell, 1995. **83**(7): p. 1171-81.
142. Mori, H., et al., *Fluorescence resonance energy transfer analysis of protein translocase. SecYE from* Thermus thermophilus *HB8 forms a constitutive oligomer in membranes.* J Biol Chem, 2003. **278**(16): p. 14257-64.
143. Mori, T., et al., *Molecular mechanisms underlying the early stage of protein translocation through the Sec translocon.* Biochemistry, 2010. **49**(5): p. 945-50.
144. Mori, H. and K. Ito, *The Sec protein-translocation pathway.* Trends Microbiol, 2001. **9**(10): p. 494-500.
145. Zhu, L., et al., *Gene fitness landscape of group A* streptococcus *during necrotizing myositis*. 2018, Journal of Clinical Investigation, In Press.
146. Wozniak, R.A. and M.K. Waldor, *A toxin-antitoxin system promotes the maintenance of an integrative conjugative element.* PLoS Genet, 2009. **5**(3): p. e1000439.
147. Webb, K., et al., *Detection of* Streptococcus equi *subspecies equi using a triplex qPCR assay.* Vet J, 2013. **195**(3): p. 300-4.
148. Weerdenburg, E.M., et al., *Genome-wide transposon mutagenesis indicates that* Mycobacterium marinum *customizes its virulence mechanisms for survival and replication in different hosts.* Infect Immun, 2015. **83**(5): p. 1778-88.

149. Varmanen, P., et al., *ClpE from* Lactococcus lactis *promotes repression of CtsR-dependent gene expression.* J Bacteriol, 2003. **185**(17): p. 5117-24.
150. Hartke, A., et al., *Survival of* Enterococcus faecalis *in an oligotrophic microcosm: changes in morphology, development of general stress resistance, and analysis of protein synthesis.* Appl Environ Microbiol, 1998. **64**(11): p. 4238-45.
151. Frees, D. and H. Ingmer, *ClpP participates in the degradation of misfolded protein in* Lactococcus lactis*.* Mol Microbiol, 1999. **31**(1): p. 79-87.
152. Dinkla, K., et al., *Upregulation of capsule enables* Streptococcus pyogenes *to evade immune recognition by antigen-specific antibodies directed to the G-related alpha2-macroglobulin-binding protein GRAB located on the bacterial surface.* Microbes Infect, 2007. **9**(8): p. 922-31.
153. Rubens, C.E., et al., *Transposon mutagenesis of type III group B* Streptococcus*: correlation of capsule expression with virulence.* Proc Natl Acad Sci U S A, 1987. **84**(20): p. 7208-12.
154. Wessels, M.R. and M.S. Bronze, *Critical role of the group A streptococcal capsule in pharyngeal colonization and infection in mice.* Proc Natl Acad Sci U S A, 1994. **91**(25): p. 12238-42.
155. Wessels, M.R., et al., *Effects on virulence of mutations in a locus essential for hyaluronic acid capsule expression in group A streptococci.* Infect Immun, 1994. **62**(2): p. 433-41.
156. Roche, A.M., S.J. King, and J.N. Weiser, *Live attenuated* Streptococcus pneumoniae *strains induce serotype-independent mucosal and systemic protection in mice.* Infect Immun, 2007. **75**(5): p. 2469-75.
157. Ashbaugh, C.D., et al., *Molecular analysis of the role of the group A streptococcal cysteine protease, hyaluronic acid capsule, and M protein in a murine model of human invasive soft-tissue infection.* J Clin Invest, 1998. **102**(3): p. 550-60.
158. Chédin, F. and S.C. Kowalczykowski, *A novel family of regulated helicases/nucleases from Gram-positive bacteria: insights into the initiation of DNA recombination.* Mol Microbiol, 2002. **43**(4): p. 823-34.
159. Halpern, D., et al., *rexAB mutants in* Streptococcus pneumoniae*.* Microbiology, 2004. **150**(Pt 7): p. 2409-14.
160. Yeeles, J.T., et al., *The AddAB helicase-nuclease catalyses rapid and processive DNA unwinding using a single Superfamily 1A motor domain.* Nucleic Acids Res, 2011. **39**(6): p. 2271-85.
161. Amundsen, S.K., et al., *Helicobacter pylori AddAB helicase-nuclease and RecA promote recombination-related DNA repair and survival during stomach colonization.* Mol Microbiol, 2008. **69**(4): p. 994-1007.
162. Whitby, M.C., S.D. Vincent, and R.G. Lloyd, *Branch migration of Holliday junctions: identification of RecG protein as a junction specific DNA helicase.* EMBO J, 1994. **13**(21): p. 5220-8.
163. Hong, X., G.W. Cadwell, and T. Kogoma, *Escherichia coli RecG and RecA proteins in R-loop formation.* EMBO J, 1995. **14**(10): p. 2385-92.
164. Vincent, S.D., A.A. Mahdi, and R.G. Lloyd, *The RecG branch migration protein of* Escherichia coli *dissociates R-loops.* J Mol Biol, 1996. **264**(4): p. 713-21.
165. Lloyd, R.G. and C. Buckman, *Genetic analysis of the recG locus of Escherichia coli K-12 and of its role in recombination and DNA repair.* J Bacteriol, 1991. **173**(3): p. 1004-11.
166. Lloyd, R.G., *Conjugational recombination in resolvase-deficient ruvC mutants of* Escherichia coli *K-12 depends on recG.* J Bacteriol, 1991. **173**(17): p. 5414-8.
167. Martinussen, J., et al., *The pyrimidine operon pyrRPB-carA from* Lactococcus lactis*.* J Bacteriol, 2001. **183**(9): p. 2785-94.
168. Pettigrew, M.M., et al., *Dynamic changes in the* Streptococcus pneumoniae *transcriptome during transition from biofilm formation to invasive disease upon influenza A virus infection.* Infect Immun, 2014. **82**(11): p. 4607-19.
169. Sitkiewicz, I., et al., *Adaptation of group A* Streptococcus *to human amniotic fluid.* PLoS One, 2010. **5**(3): p. e9785.

170. Moukadiri, I., et al., *Evolutionarily conserved proteins MnmE and GidA catalyze the formation of two methyluridine derivatives at tRNA wobble positions.* Nucleic Acids Res, 2009. **37**(21): p. 7177-93.
171. Yim, L., et al., *Further insights into the tRNA modification process controlled by proteins MnmE and GidA of* Escherichia coli*.* Nucleic Acids Res, 2006. **34**(20): p. 5892-905.
172. Prado, S., et al., *The tRNA-modifying function of MnmE is controlled by post-hydrolysis steps of its GTPase cycle.* Nucleic Acids Res, 2013. **41**(12): p. 6190-208.
173. Cho, K.H. and M.G. Caparon, *tRNA modification by GidA/MnmE is necessary for* Streptococcus pyogenes *virulence: a new strategy to make live attenuated strains.* Infect Immun, 2008. **76**(7): p. 3176-86.
174. Li, D., et al., *A novel gene involved in the survival of* Streptococcus mutans *under stress conditions.* Appl Environ Microbiol, 2014. **80**(1): p. 97-103.
175. Cabedo, H., et al., *The* Escherichia coli *trmE (mnmE) gene, involved in tRNA modification, codes for an evolutionarily conserved GTPase with unusual biochemical properties.* EMBO J, 1999. **18**(24): p. 7063-76.
176. Martínez-Vicente, M., et al., *Effects of mutagenesis in the switch I region and conserved arginines of* Escherichia coli *MnmE protein, a GTPase involved in tRNA modification.* J Biol Chem, 2005. **280**(35): p. 30660-70.
177. Elseviers, D., L.A. Petrullo, and P.J. Gallagher, *Novel* E. coli *mutants deficient in biosynthesis of 5-methylaminomethyl-2-thiouridine.* Nucleic Acids Res, 1984. **12**(8): p. 3521-34.
178. Gong, S., Z. Ma, and J.W. Foster, *The Era-like GTPase TrmE conditionally activates gadE and glutamate-dependent acid resistance in* Escherichia coli*.* Mol Microbiol, 2004. **54**(4): p. 948-61.
179. Heath, A., et al., *A two-component regulatory system, CsrR-CsrS, represses expression of three* Streptococcus pyogenes *virulence factors, hyaluronic acid capsule, streptolysin S, and pyrogenic exotoxin B.* Infect Immun, 1999. **67**(10): p. 5298-305.
180. Lamy, M.C., et al., *CovS/CovR of group B streptococcus: a two-component global regulatory system involved in virulence.* Mol Microbiol, 2004. **54**(5): p. 1250-68.
181. Jiang, S.M., et al., *Regulation of virulence by a two-component system in group B* streptococcus*.* J Bacteriol, 2005. **187**(3): p. 1105-13.
182. Jiang, S.M., et al., *Variation in the group B* Streptococcus *CsrRS regulon and effects on pathogenicity.* J Bacteriol, 2008. **190**(6): p. 1956-65.
183. Webb, K., et al., *Development of an unambiguous and discriminatory multilocus sequence typing scheme for the* Streptococcus zooepidemicus *group.* Microbiology, 2008. **154**(Pt 10): p. 3016-24.
184. Willenborg, J., et al., *Role of glucose and CcpA in capsule expression and virulence of* Streptococcus suis*.* Microbiology, 2011. **157**(Pt 6): p. 1823-33.
185. Willenborg, J., et al., *The CcpA regulon of* Streptococcus suis *reveals novel insights into the regulation of the streptococcal central carbon metabolism by binding of CcpA to two distinct binding motifs.* Mol Microbiol, 2014. **92**(1): p. 61-83.
186. Chang, J.C. and M.J. Federle, *PptAB Exports Rgg Quorum-Sensing Peptides in* Streptococcus*.* PLoS One, 2016. **11**(12): p. e0168461.
187. Jonsson, I.M., et al., *Inactivation of the Ecs ABC transporter of* Staphylococcus aureus *attenuates virulence by altering composition and function of bacterial wall.* PLoS One, 2010. **5**(12): p. e14209.
188. Grant, A.J., et al., *Co-ordination of pathogenicity island expression by the BipA GTPase in enteropathogenic* Escherichia coli *(EPEC).* Mol Microbiol, 2003. **48**(2): p. 507-21.
189. Neidig, A., et al., *TypA is involved in virulence, antimicrobial resistance and biofilm formation in* Pseudomonas aeruginosa*.* BMC Microbiol, 2013. **13**: p. 77.

190. Chastanet, A., et al., *Regulation of* Streptococcus pneumoniae *clp genes and their role in competence development and stress survival.* J Bacteriol, 2001. **183**(24): p. 7295-307.
191. Craig, E.A., B.D. Gambill, and R.J. Nelson, *Heat shock proteins: molecular chaperones of protein biogenesis.* Microbiol Rev, 1993. **57**(2): p. 402-14.
192. Hendrick, J.P. and F.U. Hartl, *Molecular chaperone functions of heat-shock proteins.* Annu Rev Biochem, 1993. **62**: p. 349-84.
193. Elsholz, A.K., et al., *CtsR, the Gram-positive master regulator of protein quality control, feels the heat.* EMBO J, 2010. **29**(21): p. 3621-9.
194. Van Bokhorst-van de Veen, H., et al., *Transcriptome signatures of class I and III stress response deregulation in* Lactobacillus plantarum *reveal pleiotropic adaptation.* Microb Cell Fact, 2013. **12**: p. 112.
195. Hooven, T.A., et al., *The* Streptococcus agalactiae *Stringent Response Enhances Virulence and Persistence in Human Blood.* Infect Immun, 2018. **86**(1).
196. Chen, I.A., et al., *IMG/M: integrated genome and metagenome comparative data analysis system.* Nucleic Acids Res, 2017. **45**(D1): p. D507-D516.
197. Nasser, W., et al., *Evolutionary pathway to increased virulence and epidemic group A* Streptococcus *disease derived from 3,615 genome sequences.* Proc Natl Acad Sci U S A, 2014. **111**(17): p. E1768-76.
198. Zhu, L., et al., *A molecular trigger for intercontinental epidemics of group A* Streptococcus*.* J Clin Invest, 2015. **125**(9): p. 3545-59.
199. Sumby, P., et al., *Evolutionary origin and emergence of a highly successful clone of serotype M1 group a* Streptococcus *involved multiple horizontal gene transfer events.* J Infect Dis, 2005. **192**(5): p. 771-82.
200. Smit, P.W., et al., *Epidemiology and emm types of invasive group A streptococcal infections in Finland, 2008-2013.* Eur J Clin Microbiol Infect Dis, 2015. **34**(10): p. 2131-6.
201. Chatellier, S., et al., *Genetic relatedness and superantigen expression in group A* streptococcus *serotype M1 isolates from patients with severe and nonsevere invasive diseases.* Infect Immun, 2000. **68**(6): p. 3523-34.
202. Hoiseth, S.K. and B.A. Stocker, *Aromatic-dependent* Salmonella typhimurium *are non-virulent and effective as live vaccines.* Nature, 1981. **291**(5812): p. 238-9.
203. Marsden, M.J., et al., *A live (delta aroA)* Aeromonas salmonicida *vaccine for furunculosis preferentially stimulates T-cell responses relative to B-cell responses in rainbow trout (*Oncorhynchus mykiss*).* Infect Immun, 1996. **64**(9): p. 3863-9.
204. Stocker, B.A., S.K. Hoiseth, and B.P. Smith, *Aromatic-dependent "*Salmonella *sp." as live vaccine in mice and calves.* Dev Biol Stand, 1983. **53**: p. 47-54.
205. Vaughan, L.M., P.R. Smith, and T.J. Foster, *An aromatic-dependent mutant of the fish pathogen* Aeromonas salmonicida *is attenuated in fish and is effective as a live vaccine against the salmonid disease furunculosis.* Infect Immun, 1993. **61**(5): p. 2172-81.
206. Scott, P.C., J.F. Markham, and K.G. Whithear, *Safety and efficacy of two live* Pasteurella multocida *aro-A mutant vaccines in chickens.* Avian Dis, 1999. **43**(1): p. 83-8.
207. Homchampa, P., R.A. Strugnell, and B. Adler, *Molecular analysis of the aroA gene of* Pasteurella multocida *and vaccine potential of a constructed aroA mutant.* Mol Microbiol, 1992. **6**(23): p. 3585-93.
208. La Ragione, R.M., et al., *Efficacy of a live attenuated* Escherichia coli *O78:K80 vaccine in chickens and turkeys.* Avian Dis, 2013. **57**(2): p. 273-9.
209. Mombarg, M., et al., *Safety and efficacy of an aroA-deleted live vaccine against avian colibacillosis in a multicentre field trial in broilers in Morocco.* Avian Pathol, 2014. **43**(3): p. 276-81.
210. Fittipaldi, N., et al., *Potential use of an unencapsulated and aromatic amino acid-auxotrophic* Streptococcus suis *mutant as a live attenuated vaccine in swine.* Vaccine, 2007. **25**(18): p. 3524-35.

211. Jurtshuk P., J., *Bacterial Metabolism* in *In Medical Microbiology*, E.b.S. Baron., Editor. 1996: Galveston, TX: University of Texas Medical Branch.
212. Luong, T.T., et al., *Ethanol-induced alcohol dehydrogenase E (AdhE) potentiates pneumolysin in* Streptococcus pneumoniae*.* Infect Immun, 2015. **83**(1): p. 108-19.
213. Satoh, M.S. and T. Lindahl, *Role of poly(ADP-ribose) formation in DNA repair.* Nature, 1992. **356**(6367): p. 356-8.
214. Wilkinson, A., J. Day, and R. Bowater, *Bacterial DNA ligases.* Mol Microbiol, 2001. **40**(6): p. 1241-8.
215. Chandler, J.L. and R.K. Gholson, *De novo biosynthesis of nicotinamide adenine dinucleotide in* Escherichia coli: *excretion of quinolinic acid by mutants lacking quinolinate phosphoribosyl transferase.* J Bacteriol, 1972. **111**(1): p. 98-102.
216. Sorci, L., et al., *Quinolinate salvage and insights for targeting NAD biosynthesis in group A streptococci.* J Bacteriol, 2013. **195**(4): p. 726-32.
217. Rodionov, D.A., et al., *Transcriptional regulation of NAD metabolism in bacteria: genomic reconstruction of NiaR (YrxA) regulon.* Nucleic Acids Res, 2008. **36**(6): p. 2032-46.
218. Johnson, M.D., et al., *Characterization of NAD salvage pathways and their role in virulence in* Streptococcus pneumoniae*.* Microbiology, 2015. **161**(11): p. 2127-36.
219. Afzal, M., O.P. Kuipers, and S. Shafeeq, *Niacin-mediated Gene Expression and Role of NiaR as a Transcriptional Repressor of.* Front Cell Infect Microbiol, 2017. **7**: p. 70.
220. Chang, J.C., et al., *Two group A streptococcal peptide pheromones act through opposing Rgg regulators to control biofilm development.* PLoS Pathog, 2011. **7**(8): p. e1002190.
221. Ibrahim, M., et al., *A genome-wide survey of short coding sequences in streptococci.* Microbiology, 2007. **153**(Pt 11): p. 3631-44.
222. Makthal, N., et al., *A Critical Role of Zinc Importer AdcABC in Group A Streptococcus-Host Interactions During Infection and Its Implications for Vaccine Development.* EBioMedicine, 2017. **21**: p. 131-141.
223. Mitrakul, K., et al., *Mutational analysis of the adcCBA genes in* Streptococcus gordonii *biofilm formation.* Oral Microbiol Immunol, 2005. **20**(2): p. 122-7.
224. Reyes-Caballero, H., et al., *The metalloregulatory zinc site in* Streptococcus pneumoniae *AdcR, a zinc-activated MarR family repressor.* J Mol Biol, 2010. **403**(2): p. 197-216.
225. Guerra, A.J., C.E. Dann, and D.P. Giedroc, *Crystal structure of the zinc-dependent MarR family transcriptional regulator AdcR in the Zn(II)-bound state.* J Am Chem Soc, 2011. **133**(49): p. 19614-7.
226. Rajagopal, L., et al., *Regulation of purine biosynthesis by a eukaryotic-type kinase in* Streptococcus agalactiae*.* Mol Microbiol, 2005. **56**(5): p. 1329-46.
227. Jelsbak, L., et al., *The In Vitro Redundant Enzymes PurN and PurT Are Both Essential for Systemic Infection of Mice in* Salmonella enterica *Serovar* Typhimurium*.* Infect Immun, 2016. **84**(7): p. 2076-2085.
228. Beinert, H., R.H. Holm, and E. Münck, *Iron-sulfur clusters: nature's modular, multipurpose structures.* Science, 1997. **277**(5326): p. 653-9.
229. FJ, R. and B. H, *The soluble "high potential" type iron-sulfur protein from mitochondria is aconitase.* 1978, J Biol Chem. p. 2514-7.
230. Zheng, L., et al., *Cysteine desulfurase activity indicates a role for NIFS in metallocluster biosynthesis.* Proc Natl Acad Sci U S A, 1993. **90**(7): p. 2754-8.
231. Ayala-Castro, C., A. Saini, and F.W. Outten, *Fe-S cluster assembly pathways in bacteria.* Microbiol Mol Biol Rev, 2008. **72**(1): p. 110-25.
232. Riboldi, G.P., H. Verli, and J. Frazzon, *Structural studies of the* Enterococcus faecalis *SufU [Fe-S] cluster protein.* BMC Biochem, 2009. **10**: p. 3.
233. Riboldi, G.P., J.S. de Oliveira, and J. Frazzon, Enterococcus faecalis *SufU scaffold protein enhances SufS desulfurase activity by acquiring sulfur from its cysteine-153.* Biochim Biophys Acta, 2011. **1814**(12): p. 1910-8.

234. Outten, F.W., O. Djaman, and G. Storz, *A suf operon requirement for Fe-S cluster assembly during iron starvation in* Escherichia coli. Mol Microbiol, 2004. **52**(3): p. 861-72.
235. Takahashi, Y. and U. Tokumoto, *A third bacterial system for the assembly of iron-sulfur clusters with homologs in archaea and plastids.* J Biol Chem, 2002. **277**(32): p. 28380-3.
236. Santos, J.A., et al., *The unique regulation of iron-sulfur cluster biogenesis in a Gram-positive bacterium.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. E2251-60.
237. Fontecave, M., et al., *Mechanisms of iron-sulfur cluster assembly: the SUF machinery.* J Biol Inorg Chem, 2005. **10**(7): p. 713-21.
238. Takahashi, Y. and M. Nakamura, *Functional assignment of the ORF2-iscS-iscU-iscA-hscB-hscA-fdx-ORF3 gene cluster involved in the assembly of Fe-S clusters in* Escherichia coli*.* J Biochem, 1999. **126**(5): p. 917-26.
239. Mistou, M.Y., I.C. Sutcliffe, and N.M. van Sorge, *Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria.* FEMS Microbiol Rev, 2016. **40**(4): p. 464-79.
240. Lancefield, R.C., *A serological differentiation of human and other groups of hemolytic streptococci.* J Exp Med, 1933. **57**(4): p. 571-95.
241. McCarty, M., *The lysis of group A hemolytic streptococci by extracellular enzymes of* Streptomyces albus*. II. Nature of the cellular substrate attacked by the lytic enzymes.* J Exp Med, 1952. **96**(6): p. 569-80.
242. McCarty, M., *Variation in the group-specific carbohydrate of group A streptococci. II. Studies on the chemical basis for serological specificity of the carbohydrates.* J Exp Med, 1956. **104**(5): p. 629-43.
243. Coligan, J.E., T.J. Kindt, and R.M. Krause, *Structure of the streptococcal groups A, A-variant and C carbohydrates.* Immunochemistry, 1978. **15**(10-11): p. 755-60.
244. Xie, S., et al., *Transcriptome profiling of* Bacillus subtilis *OKB105 in response to rice seedlings.* BMC Microbiol, 2015. **15**: p. 21.
245. Rush, J.S., et al., *The molecular mechanism of N-acetylglucosamine side-chain attachment to the Lancefield group A carbohydrate in* Streptococcus pyogenes*.* J Biol Chem, 2017. **292**(47): p. 19441-19457.
246. Shibata, Y., et al., *Expression and characterization of streptococcal rgp genes required for rhamnan synthesis in Escherichia coli.* Infect Immun, 2002. **70**(6): p. 2891-8.
247. Edgar, R., et al., *Genetic insight into zinc and antimicrobial toxicity uncovers a glycerol phosphate modification on streptococcal rhamnose polysaccharides*. 2018, BioRxiv.
248. van Sorge, N.M., et al., *The classical lancefield antigen of group a* Streptococcus *is a virulence determinant with implications for vaccine design.* Cell Host Microbe, 2014. **15**(6): p. 729-740.
249. van Hensbergen, V.P., et al., *Streptococcal Lancefield polysaccharides are critical cell wall determinants for human Group IIA secreted phospholipase A2 to exert its bactericidal effects.* PLoS Pathog, 2018. **14**(10): p. e1007348.
250. Koprivnjak, T., et al., *Role of charge properties of bacterial envelope in bactericidal action of human group IIA phospholipase A2 against* Staphylococcus aureus*.* J Biol Chem, 2002. **277**(49): p. 47636-44.
251. Neuhaus, F.C. and J. Baddiley, *A continuum of anionic charge: structures and functions of D-alanyl-teichoic acids in gram-positive bacteria.* Microbiol Mol Biol Rev, 2003. **67**(4): p. 686-723.
252. Abachin, E., et al., *Formation of D-alanyl-lipoteichoic acid is required for adhesion and virulence of* Listeria monocytogenes*.* Mol Microbiol, 2002. **43**(1): p. 1-14.
253. Wartha, F., et al., *Capsule and D-alanylated lipoteichoic acids protect* Streptococcus pneumoniae *against neutrophil extracellular traps.* Cell Microbiol, 2007. **9**(5): p. 1162-71.
254. Weidenmaier, C., et al., *DltABCD- and MprF-mediated cell envelope modifications of* Staphylococcus aureus *confer resistance to platelet microbicidal*

proteins and contribute to virulence in a rabbit endocarditis model. Infect Immun, 2005. **73**(12): p. 8033-8.
255. Collins, L.V., et al., Staphylococcus aureus *strains lacking D-alanine modifications of teichoic acids are highly susceptible to human neutrophil killing and are virulence attenuated in mice.* J Infect Dis, 2002. **186**(2): p. 214-9.
256. Poyart, C., et al., *Attenuated virulence of* Streptococcus agalactiae *deficient in D-alanyl-lipoteichoic acid is due to an increased susceptibility to defensins and phagocytic cells.* Mol Microbiol, 2003. **49**(6): p. 1615-25.
257. Król, J.E., et al., *SMU.746-SMU.747, a putative membrane permease complex, is involved in aciduricity, acidogenesis, and biofilm formation in* Streptococcus mutans*.* J Bacteriol, 2014. **196**(1): p. 129-39.
258. Nyunoya, H. and C.J. Lusty, *The carB gene of* Escherichia coli*: a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase.* Proc Natl Acad Sci U S A, 1983. **80**(15): p. 4629-33.
259. Arioli, S., et al., *Carbamoylphosphate synthetase activity is essential for the optimal growth of* Streptococcus thermophilus *in milk.* J Appl Microbiol, 2009. **107**(1): p. 348-54.
260. Oda, M., et al., Streptococcus pyogenes *Phospholipase A2 Induces the Expression of Adhesion Molecules on Human Umbilical Vein Endothelial Cells and Aorta of Mice.* Front Cell Infect Microbiol, 2017. **7**: p. 300.
261. Banks, D.J., et al., *Progress toward characterization of the group A* Streptococcus *metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain.* J Infect Dis, 2004. **190**(4): p. 727-38.
262. Beres, S.B., et al., *Genome sequence of a serotype M3 strain of group A* Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 10078-83.
263. Beres, S.B., et al., *Genome-wide molecular dissection of serotype M3 group A* Streptococcus *strains causing two epidemics of invasive infections.* Proc Natl Acad Sci U S A, 2004. **101**(32): p. 11833-8.
264. Nagiec, M.J., et al., *Analysis of a novel prophage-encoded group A* Streptococcus *extracellular phospholipase A(2).* J Biol Chem, 2004. **279**(44): p. 45909-18.
265. Eymann, C., G. Mittenhuber, and M. Hecker, *The stringent response, sigmaH-dependent gene expression and sporulation in* Bacillus subtilis*.* Mol Gen Genet, 2001. **264**(6): p. 913-23.
266. Eymann, C., et al., Bacillus subtilis *functional genomics: global characterization of the stringent response by proteome and transcriptome analysis.* J Bacteriol, 2002. **184**(9): p. 2500-20.
267. Kwon, H.Y., et al., *Effect of heat shock and mutations in ClpL and ClpP on virulence gene expression in Streptococcus pneumoniae.* Infect Immun, 2003. **71**(7): p. 3757-65.
268. Wawrzynow, A., B. Banecki, and M. Zylicz, *The Clp ATPases define a novel class of molecular chaperones.* Mol Microbiol, 1996. **21**(5): p. 895-9.
269. Gottesman, S., et al., *The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system.* Genes Dev, 1998. **12**(9): p. 1338-47.
270. Makovets, S., A.J. Titheradge, and N.E. Murray, *ClpX and ClpP are essential for the efficient acquisition of genes specifying type IA and IB restriction systems.* Mol Microbiol, 1998. **28**(1): p. 25-35.
271. Buchmeier, N.A. and F. Heffron, *Induction of* Salmonella *stress proteins upon infection of macrophages.* Science, 1990. **248**(4956): p. 730-2.
272. Milani, C.J., et al., *The novel polysaccharide deacetylase homologue Pdi contributes to virulence of the aquatic pathogen* Streptococcus iniae. Microbiology, 2010. **156**(Pt 2): p. 543-54.
273. Sheldon, W.L., et al., *Functional analysis of a group A streptococcal glycoside hydrolase Spy1600 from family 84 reveals it is a beta-N-acetylglucosaminidase and not a hyaluronidase.* Biochem J, 2006. **399**(2): p. 241-7.

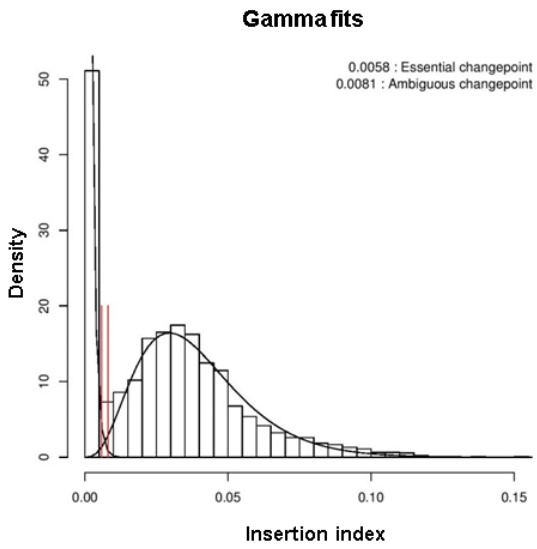
274. Voyich, J.M., et al., *Genome-wide protective response used by group A* Streptococcus *to evade destruction by human polymorphonuclear leukocytes.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1996-2001.
275. Stewart, F.M. and B.R. Levin, *The population biology of bacterial viruses: why be temperate.* Theor Popul Biol, 1984. **26**(1): p. 93-117.
276. Dawid, S., A.M. Roche, and J.N. Weiser, *The blp bacteriocins of* Streptococcus pneumoniae *mediate intraspecies competition both in vitro and in vivo.* Infect Immun, 2007. **75**(1): p. 443-51.
277. de Saizieu, A., et al., *Microarray-based identification of a novel* Streptococcus pneumoniae *regulon controlled by an autoinduced peptide.* J Bacteriol, 2000. **182**(17): p. 4696-703.
278. Wholey, W.Y., et al., *Coordinated Bacteriocin Expression and Competence in* Streptococcus pneumoniae *Contributes to Genetic Adaptation through Neighbor Predation.* PLoS Pathog, 2016. **12**(2): p. e1005413.
279. Valente, C., et al., *The blp Locus of* Streptococcus pneumoniae *Plays a Limited Role in the Selection of Strains That Can Cocolonize the Human Nasopharynx.* Appl Environ Microbiol, 2016. **82**(17): p. 5206-15.
280. Pinchas, M.D., N.C. LaCross, and S. Dawid, *An electrostatic interaction between BlpC and BlpH dictates pheromone specificity in the control of bacteriocin production and immunity in* Streptococcus pneumoniae. J Bacteriol, 2015. **197**(7): p. 1236-48.
281. Håvarstein, L.S., D.B. Diep, and I.F. Nes, *A family of bacteriocin ABC transporters carry out proteolytic processing of their substrates concomitant with export.* Mol Microbiol, 1995. **16**(2): p. 229-40.
282. Kjos, M., et al., *Expression of* Streptococcus pneumoniae *Bacteriocins Is Induced by Antibiotics via Regulatory Interplay with the Competence System.* PLoS Pathog, 2016. **12**(2): p. e1005422.
283. Redfield, R.J., *Is quorum sensing a side effect of diffusion sensing?* Trends Microbiol, 2002. **10**(8): p. 365-70.

Appendix 1

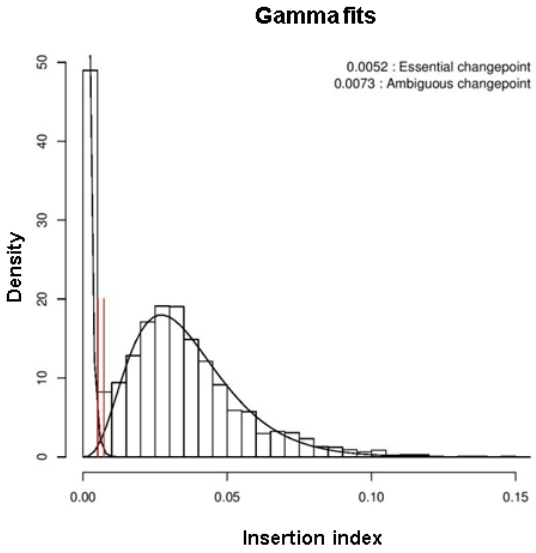
Table A1.1. Description of bioinformatic scripts, programmes and online tools utilised

Script/programme/tool	Description	Reference
bacteria_tradis	**Input files**: FASTQ file, transposon tag and FASTA reference file **Function**: filters data for tag, trims tag from tag matching reads and maps to genome	[111]
tradis_gene_insert_sites	**Input files**: plot file produced by bacteria_tradis and reference embl file **Function**: generates readable csv file containing read counts and insertion sites per gene	[111]
tradis_essentiality	**Input files**: csv file from tradis_gene_insert_sites **Function**: determines gene essentiality from in 1 library using empirically observed bimodal distribution of insertion indices to fit gamma distributions	[111]
tradis_comparison	**Input files**: csv files from tradis_gene_insert_sites; 3 input libraries (controls) and 3 output libraries (conditions) **Function**: determines gene fitness by comparing the read counts per gene of 3 libraries exposed to an experimental condition, to the 3 libraries before exposure	[111]
Weblogo	**Input files**: txt file of unique reads **Function**: identifies insertion sites bias across genome by calculating the probability of each nucleotide occurring at each position downstream of the insertion site. Produces stacked probability plot.	[113]
Genome2D	Webserver for analysis and visualization of bacterial genomes Used as a database for KEGG categories assigned to *S. equi* genes	[114]
OrthologueDB	Orthology predictions between bacterial genomes Used to determine orthologous genes between *S. equi*, *S. pyogenes* and *S. agalactiae*	[115]
IGM/M	Onine database for analysis and annotation of genomes Used to determine COG categories of *S. equi* genes	[196]

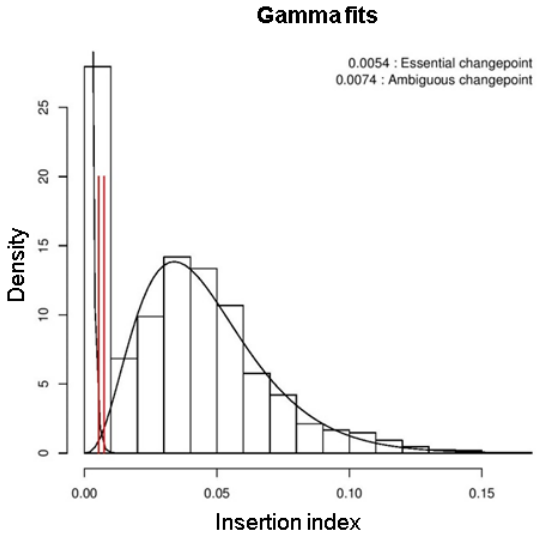
Library CA



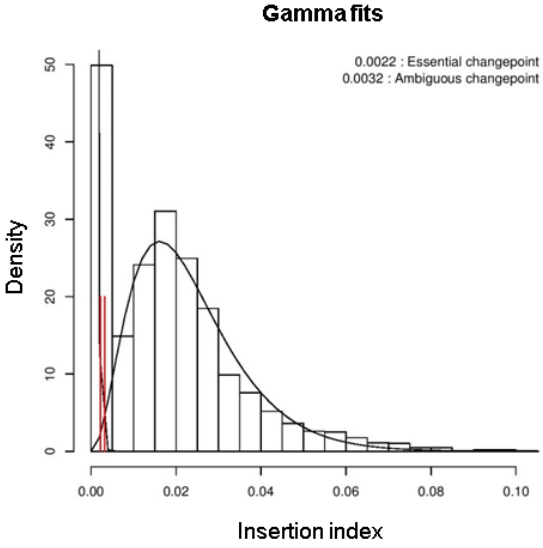
Library TC



Library AG



Library AC



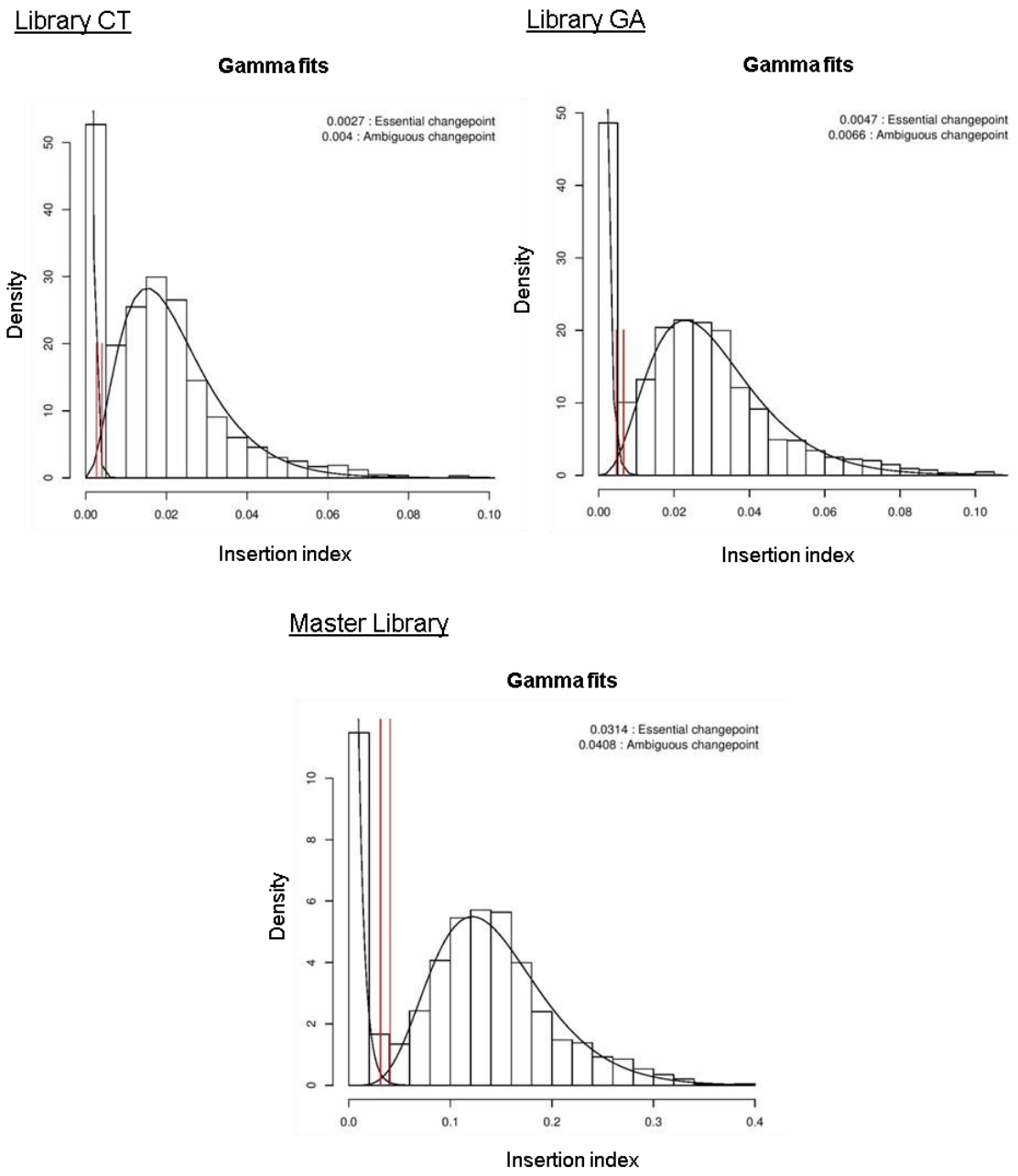


Figure A1.1. Diagnostic plots for each barcoded library and the master library generated by the tradis_essentiality script [111].

Table A1.2. Primers used in Chapter 2. * Barcode is underlined, ** Index sequence is underlined, × = Phosphorothioate bond, P = phosphorylation, Double underline = complementary sequence between adaptor primers, Δ = Oligonucleotide sequences © 2007- 2012 Illumina, Inc. All rights reserved

Primer name	Sequence (5'-3')	Concentration
P1	GACACGTCGACGGTATCGATAAGCTTG	100 µM
P4	GACACCCCGGGCTGCAGGAA	100 µM
P2 TC*	GCAAAGTTTTCGAATAAGTCTATTTTAGTG	100 µM
P3 TC*	CACTAAAATAGACTTATTCGAAAACTTTGC	100 µM
P2 AG*	GCAAAGTTTTCCTATAAGTCTATTTTAGTG	100 µM
P3 AG*	CACTAAAATAGACTTATAGGAAAACTTTGC	100 µM
P2 AC*	GCAAAGTTTTCGTATAAGTCTATTTTAGTG	100 µM
P3 AC*	CACTAAAATAGACTTATACGAAAACTTTGC	100 µM
P2 CT*	GCAAAGTTTTCAGATAAGTCTATTTTAGTG	100 µM
P3 CT*	CACTAAAATAGACTTATCTGAAAACTTTGC	100 µM
P2 GA*	GCAAAGTTTTCTCATAAGTCTATTTTAGTG	100 µM
P3 GA*	CACTAAAATAGACTTATGAGAAAACTTTGC	100 µM
5'9	CTGGAACATCTGTGGTATGG	100 µM
3'9	GCGTACCTTGGATATTCACC	100 µM
Adaptor primer 1Δ	P-GATCGGAAGAGCACACGTCT	100 µM
Adaptor primer 2Δ	ACACTCTTTCCCTACACGACGCTCTTCCGATC×T	100 µM
Specific IS*S1* primer	AATGATACGGCGACCACCGAGATCTACACGTTCATTGA TATATCCTCGCTG	25 µM
Indexing PCR primer 1**	CAAGCAGAAGACGGCATACGAGATCGGTTCGCCTTAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing PCR primer 2**	CAAGCAGAAGACGGCATACGAGATCGGTCTAGTACGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing PCR primer 4**	CAAGCAGAAGACGGCATACGAGATCGGTGCTCAGGAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Custom read 1 primer	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTTTA CACTAAAATAGACTTAT	100 µM
Custom Index Read primer	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	100 µM

Table A1.3. Primers used in Chapter 3. × = Phosphorothioate bond, P = phosphorylation, Double underline = complementary sequence between adaptor primers, Δ = Oligonucleotide sequences © 2007- 2012 Illumina, Inc. All rights reserved

Primer name	Sequence (5'-3')	Concentration
Adaptor primer 1Δ	P-GATCGGAAGAGCACACGTCT	100 µM
Adaptor primer 2Δ	ACACTCTTTCCCTACACGACGCTCTTCCGATC×T	100 µM
Specific IS*S1* primer	AATGATACGGCGACCACCGAGATCTACACGTTCATT GATATATCCTCGCTG	25 µM
Indexing primer AHT 6	CAAGCAGAAGACGGCATACGAGATCCTTACCATAAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 7	CAAGCAGAAGACGGCATACGAGATTGATATCTCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 15	CAAGCAGAAGACGGCATACGAGATGATAGAGACAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 16	CAAGCAGAAGACGGCATACGAGATATCATAGACGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 21	CAAGCAGAAGACGGCATACGAGATCGCTGCAGTAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 32	CAAGCAGAAGACGGCATACGAGATTACACTCATGAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Custom read 1 primer	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTT TACACTAAAATAGACTTAT	100 µM
Custom Index Read primer	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	100 µM
P1 *mnmE*	GACACGAATTCCGTGGTTGAAAAAGAAGCC	100 µM
P2 *mnmE*	GACACGATATCCAATGCCAATAGCTCCTTCAC	100 µM
P3 *mnmE*	GACACGATATCCCTAGGAGAAATCACAGGCG	100 µM
P4 *mnmE*	GACACGTCGACCTTTGGACGCTTGCTTGAG	100 µM
P1 *pyrP*	GACACGAATTCCATGAAGCGTGCGATCAC	100 µM
P2 *pyrP*	GACACGATATCGCCTTTGGCACTTCTTCTAC	100 µM
P3 *pyrP*	GACACGATATCCAATGGCCTTCAGATTTCG	100 µM
P4 *pyrP*	GACACGTCGACCCTCCATTAACGATAGAGGC	100 µM
P1 *addA*	GACACGAATTCGCTTGAGTCCTCAGCTTGTGAC	100 µM
P2 *addA*	GACACGATATCCTCCTGCTGCAAACGAGC	100 µM
P3 *addA*	GACACGATATCGGGTGGATCACAGCTAGAAG	100 µM
P4 *addA*	GACACGTCGACCAGGAGAGCCTTCTATCCAG	100 µM
P1 *recG*	GACACGAATTCCTTCTAGACAAGCACCCTGCC	100 µM
P2 *recG*	GACACGATATCGACCCTTCAAATTAGCAATCG	100 µM
P3 *recG*	GACACGATATCGGCAAGACGAGTCGCTGCT	100 µM
P4 *recG*	GACACGTCGACGCTGAGCCAAGGGTTCGCTT	100 µM
5'9	CTGGAACATCTGTGGTATGG	100 µM
3'9	GCGTACCTTGGATATTCACC	100 µM

Table A1.4. Primers used in Chapter 4. × = Phosphorothioate bond, P = phosphorylation, Double underline = complementary sequence between adaptor primers, Δ = Oligonucleotide sequences © 2007- 2012 Illumina, Inc. All rights reserved

Primer name	Sequence (5'-3')	Concentration
Adaptor primer 1Δ	P-GATCGGAAGAGCACACGTCT	100 µM
Adaptor primer 2Δ	ACACTCTTTCCCTACACGACGCTCTTCCGATC×T	100 µM
Specific IS*S1* primer	AATGATACGGCGACCACCGAGATCTACACGTTCATT GATATATCCTCGCTG	25 µM
Custom read 1 primer	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTT TACACTAAAATAGACTTAT	100 µM
Custom Index Read primer	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	100 µM
5'9	CTGGAACATCTGTGGTATGG	100 µM
3'9	GCGTACCTTGGATATTCACC	100 µM
Nested IS*S1*	CAACAGCGACAATAATCACATCT	100 µM
Nested adaptor	ACACTCTTTCCCTCACGACG	100 µM
HiSeq sequencing		
Indexing primer AHT1	CAAGCAGAAGACGGCATACGAGATTACCACAACAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT2	CAAGCAGAAGACGGCATACGAGATTAGACACACTA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT3	CAAGCAGAAGACGGCATACGAGATGATGTGACAAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT4	CAAGCAGAAGACGGCATACGAGATGTCTACTGTCA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT5	CAAGCAGAAGACGGCATACGAGATTAGCCTCCAGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT6	CAAGCAGAAGACGGCATACGAGATCCTTACCATAAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT7	CAAGCAGAAGACGGCATACGAGATTGATATCTCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT8	CAAGCAGAAGACGGCATACGAGATGACATATATCAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT9	CAAGCAGAAGACGGCATACGAGATTCAGACATGTA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT10	CAAGCAGAAGACGGCATACGAGATATGTCTGGACA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT11	CAAGCAGAAGACGGCATACGAGATCCTCAATCCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT12	CAAGCAGAAGACGGCATACGAGATCTATCGAACAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT13	CAAGCAGAAGACGGCATACGAGATTGACAGCTGCA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT14	CAAGCAGAAGACGGCATACGAGATGCAAGGACAAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT15	CAAGCAGAAGACGGCATACGAGATGATAGAGACAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT16	CAAGCAGAAGACGGCATACGAGATATCATAGACGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT21	CAAGCAGAAGACGGCATACGAGATCGCTGCAGTAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT22	CAAGCAGAAGACGGCATACGAGATACGTACAGTCA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT23	CAAGCAGAAGACGGCATACGAGATATATGACTGTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT24	CAAGCAGAAGACGGCATACGAGATGAGATATGATA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM

Indexing primer AHT25	CAAGCAGAAGACGGCATACGAGATGATCACAGCCA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT26	CAAGCAGAAGACGGCATACGAGATATGTCGTAGAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT28	CAAGCAGAAGACGGCATACGAGATAGCACGCTCAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT30	CAAGCAGAAGACGGCATACGAGATACAGCACTCGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT32	CAAGCAGAAGACGGCATACGAGATTACACTCATGAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT35	CAAGCAGAAGACGGCATACGAGATCAGCGATGTAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT36	CAAGCAGAAGACGGCATACGAGATCATAGCGTGAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT37	CAAGCAGAAGACGGCATACGAGATTACGAAGAACA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT38	CAAGCAGAAGACGGCATACGAGATAGTGAAGCTAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT39	CAAGCAGAAGACGGCATACGAGATTCCACCATATAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT42	CAAGCAGAAGACGGCATACGAGATCTGACTAGTCA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT44	CAAGCAGAAGACGGCATACGAGATACGCTCGATTA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT45	CAAGCAGAAGACGGCATACGAGATTCTTGGCGTTA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT49	CAAGCAGAAGACGGCATACGAGATACGTAGTACGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
Indexing primer AHT50	CAAGCAGAAGACGGCATACGAGATAGCTCTCCACA CACTCTTTCCCTACACGACGCTCTTCCGATCT	100 µM
MiSeq sequencing		
Indexing primer AHT 6	CAAGCAGAAGACGGCATACGAGATCCTTACCATAAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 7	CAAGCAGAAGACGGCATACGAGATTGATATCTCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 15	CAAGCAGAAGACGGCATACGAGATGATAGAGACAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 16	CAAGCAGAAGACGGCATACGAGATATCATAGACGA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 21	CAAGCAGAAGACGGCATACGAGATCGCTGCAGTAA CACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Indexing primer AHT 32	CAAGCAGAAGACGGCATACGAGATTACACTCATGAC ACTCTTTCCCTACACGACGCTCTTCCGATCT	25 µM
Validation mutants		
P1 *purN*	CAACAAGAATTCGCTGAGGGAGATCATCTCCTAGG	100 µM
P2 *purN*	CAACAAGATATCGGCTTCATCACCACACTAGCAACT GTTC	100 µM
P3 *purN*	CAACAAGATATCTTGTGTAAGATTAATAACTATTTTC CCTATTGTGGC	100 µM
P4 *purN*	CAACAAGTCGACGACATCATAAGTGACATCTGGACG G	100 µM
P1 *SEQ0402*	CAACAAGAATTCGAGGAGAAAACACAATGAAGCTG	100 µM
P2 *SEQ0402*	CAACAAGATATCTTCTCTCTCCTTTAATGATAGAC	100 µM
P3 *SEQ0402*	CAACAAGATATCACAAATTAACAATCCCCATATTTCA AC	100 µM
P4 *SEQ0402*	CAACAAGTCGACGGACTTTTGGGCAACCTGGTGCG G	100 µM
P1 *sufC*	CAACAAGAATTCGGTAGCCGTTGTTACACCGG	100 µM
P2 *sufC*	CAACAAGATATCCATTTCTATTAGGCTCTTTC	100 µM
P3 *sufC*	CAACAAGATATCTAAGCTGCAAGGCTGTCTAAGGC	100 µM

P4 *sufC*	CAACAAGTCGACGGAACATATAGCACAGCCGCACTG	100 µM
P1 *sptA*	CAACAAGAATTCCTGAGAAAAGGGCAGATTCAAGC	100 µM
P2 *sptA*	CAACAAGATATCTTCTGTCCTCCCCTTTTGTAC	100 µM
P3 *sptA*	CAACAAGATATCTAGTGCCAGATGAGAAAAAAGAATTAATAC	100 µM
P4 *sptA*	CAACAAGTCGACGGTTATTAACCAAGGCTCTTG	100 µM
P1 *gacI*	CAACAAGAATTCGTTGACGCTTGGAATGAGCACC	100 µM
P2 *gacI*	CAACAAGATATCATCGTCAAATATTCTCTCTATTC	100 µM
P3 *gacI*	CAACAAGATATCGAGGATGTTTATTGGGTTACAGC	100 µM
P4 *gacI*	CAACAAGTCGACAGCATAGGCAAAGCTAAAGGTATC	100 µM
P1 *metP*	CAACAAGAATTCTCGCTCTTTTGCAGGAATTAAACCGC	100 µM
P2 *metP*	CAACAAGTTAACTTAAGCCCCTCTCTTTAAAATAGTG	100 µM
P3 *metP*	CAACAAGTTAACTAAAACTGCTCAGAGCTTTATTAGC	100 µM
P4 *metP*	CAACAAGTCGACACCGGTCTTATCACCTCAAGTG	100 µM
P1 *SEQ1536*	CAACAACTGCAGTCTTTGTTTATTGCTACTATAAGGTGC	100 µM
P2 *SEQ1536*	CAACAAGATATCAAACGTCTCCTCTTTCTATCCC	100 µM
P3 *SEQ1536*	CAACAAGATATCGTAATTTTTTTAAAACGTTGGTGATTGG	100 µM
P4 *SEQ1536*	CAACAAGTCGACGCCTATTTTCTTGAATAAACGAG	100 µM
P1 *scfA*	CAACAAGAATTCCTCTTTCAGCTAATGCCAAGGCAG	100 µM
P2 *scfA*	CAACAAGATATCTCACCGTATCCTTTCTATATG	100 µM
P3 *scfA*	CAACAAGATATCTGATTCGCTTTTTAATTTTAGCAGGC	100 µM
P4 *scfA*	CAACAAGTCGACTAATTCCATTACCTCCATATAATTG	100 µM
P1 *SEQ1410*	CAACAAGAATTCGCTGACACCAATGTGTTGGTAGCTGG	100 µM
P2 *SEQ1410*	CAACAAGATATCCGTGATAAGCTAGCCTCCTTAATAATTGC	100 µM
P3 *SEQ1410*	CAACAAGATATCAATAAATACTCTAAAAGCCATTGGAATG	100 µM
P4 *SEQ1410*	CAACAAGTCGACCTCGATGGTCAATTCTGAACC	100 µM
P1 *dltB*	CAACAAGAATTCCAACAGGGAGAGATTATTGTAACGGG	100 µM
P2 *dltB*	CAACAAGTTAACTGATGATTAACCTCGTTAATCAAG	100 µM
P3 *dltB*	CAACAAGTTAACAAAAGGAGAGTATAAAAATGTCTAC	100 µM
P4 *dltB*	CAACAAGTCGACGATGGGTGCATGCTATCCATGCG	100 µM
P1 *slaB*	GACACGAATTCGGGGACCATAGTACTTAACTG	100 µM
P2 *slaB*	GACACAAGCTTAACATCAATAACAGGTAAAGTAAAATG	100 µM
P3 *slaB*	GACACAAGCTTGAAACGGTAGGTGCTATTGG	100 µM
P4 *slaB*	GACACGTCGACCCAAGAATGAGAAAGCAATGCTC	100 µM
P1 *SEQ0751* (IC)	CAACAAGAATTCTTTAAGGGATTTGTAGAAAGAG	100 µM
P2 *SEQ0751* (IC)	CAACAAGATATCCTGTTTATTTCACCACCTTTATTTC	100 µM
P3 *SEQ0751* (IC)	CAACAAGATATCAATTAAGTTGCAAAACAAAGATTTTTATAAATAAGAGGG	100 µM
P4 *SEQ0751* (IC)	CAACAAGTCGACCAAAGATTGACCTCATTGACATCC	100 µM
Tag A 1	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTTTACACTAAAATAGACTTATCAGAAAACTTTGCAACAGAACCC	100 µM

Tag A 2	GGGTTCTGTTGCAAAGTTTTCTGATAAGTCTATTTTA GTGTAAAATGAATAAAAATGACAGCGAGGATATATC AATGAAC	100 µM
Tag B 1	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTT TACACTAAAATAGACTTATGTTGACCCTATTGCAACT TGGAT	100 µM
Tag B 2	ATCCAAGTTGCAATAGGGTCAACATAAGTCTATTTTA GTGTAAAATGAATAAAAATGACAGCGAGGATATATC AATGAAC	100 µM
Tag C 1	GTTCATTGATATATCCTCGCTGTCATTTTTATTCATTT TACACTAAAATAGACTTATACGTCTTCGAGTAATCTA TCGTG	100 µM
Tag C 2	CACGATAGATTACTCGAAGACGTATAAGTCTATTTTA GTGTAAAATGAATAAAAATGACAGCGAGGATATATC AATGAAC	100 µM

Appendix 2

Table A2.1. *S.equi* genes required for fitness in the presence of whole equine blood determined using barcoded IS*S1* libraries. Read counts (reads) and number of unique insertions (insertions) per fitness gene in each barcoded library (input) and post (output) incubation with whole equine blood are presented.

		Input libraries						Output libraries					
Gene	Locus tag	AC reads	AC insertions	CT reads	CT insertions	GA reads	GA insertions	AC reads	AC insertions	CT reads	CT insertions	GA reads	GA insertions
ackA	SEQ0118	54	8	66	8	52	6	12	3	3	1	4	1
SEQ0231	SEQ0231	1,269	23	761	14	848	18	105	15	65	6	114	10
hasA	SEQ0269	156	3	249	4	634	8	26	3	34	2	84	4
hasB	SEQ0270	807	6	771	10	788	8	158	5	143	7	98	5
SEQ0306	SEQ0306	84	12	56	4	19	3	0	0	0	0	0	0
pepX	SEQ0383	295	17	313	21	322	12	61	11	31	7	51	3
recG	SEQ0454	440	24	170	18	199	20	33	7	7	1	9	2
SEQ0492	SEQ0492	72	8	22	5	49	3	7	2	0	0	4	1
SEQ0494	SEQ0494	41	8	28	5	46	5	0	0	0	0	6	1
pptA/ecsA	SEQ0506	123	8	87	7	111	6	15	2	0	0	8	1
pptB/ecsB	SEQ0507	355	15	330	8	248	11	41	7	29	3	34	5
SEQ0562	SEQ0562	105	8	109	10	72	6	7	1	13	3	13	1
bipA/typA	SEQ0615	77	10	42	8	13	2	4	1	0	0	0	0
pyrD	SEQ0655	85	10	83	12	64	7	10	2	5	1	6	1
ppc	SEQ0776	403	33	301	27	250	26	3	1	3	1	5	1

		Input libraries						Output libraries					
Gene	Locus tag	AC reads	AC insertions	CT reads	CT insertions	GA reads	GA insertions	AC reads	AC insertions	CT reads	CT insertions	GA reads	GA insertions
addA	SEQ0953	82	12	117	17	47	9	0	0	0	0	0	0
SEQ1028	SEQ1028	43	5	29	3	22	2	0	0	3	1	0	0
SEQ1073	SEQ1073	35	8	44	7	22	4	0	0	0	0	0	0
SEQ1112	SEQ1112	43	3	38	3	33	4	0	0	3	1	0	0
SEQ1146	SEQ1146	42	7	64	9	72	9	0	0	3	1	0	0
ldh	SEQ1169	36	5	76	7	64	8	0	0	3	1	0	0
SEQ1180	SEQ1180	56	3	44	4	16	2	0	0	3	1	0	0
SEQ1181	SEQ1181	73	9	21	4	19	3	0	0	0	0	0	0
SEQ1304	SEQ1304	348	17	215	17	229	16	3	1	3	1	0	0
pyrP	SEQ1316	77	11	127	16	113	13	0	0	6	1	4	1
mnmE	SEQ1365	127	19	53	10	31	5	4	1	0	0	0	0
SEQ1540	SEQ1540	100	8	115	8	16	1	3	1	5	1	0	0
smc	SEQ1566	439	28	265	20	484	39	40	7	10	3	15	3
ccpA	SEQ1596	47	7	28	4	11	2	3	1	0	0	0	0
pepQ	SEQ1597	79	10	59	8	165	15	0	0	0	0	5	1
SEQ1800	SEQ1800	56	3	49	3	33	1	0	0	0	0	0	0
scpA	SEQ1863	81	5	79	6	49	5	3	1	3	1	0	0
greA	SEQ1879	83	6	24	4	17	3	0	0	0	0	0	0
csrS	SEQ1889	180	17	34	6	82	6	3	1	0	0	0	0
yqeK	SEQ1909	56	5	29	5	41	3	4	1	0	0	0	0
pyrG	SEQ1945	809	24	652	16	903	30	133	13	102	9	118	15
eqbE	SEQ1242	6,463	272	7,903	232	4,690	238	6,483	233	8,716	239	5,778	272

Table A2.2. *S. equi* genes required for fitness in the presence of hydrogen peroxide determined using barcoded IS*S1* libraries. Read counts (reads) and number of unique insertions (insertions) per fitness gene in each barcoded library (input) and post (output) incubation with hydrogen peroxide are presented.

		Input libraries						Output libraries					
Gene	Locus tag	AC reads	AC insertions	CT reads	CT insertions	GA reads	GA insertions	AC reads	AC insertions	CT reads	CT insertions	GA reads	GA insertions
SEQ0118	SEQ0118	59	6	69	11	25	6	7	1	13	4	0	0
ctsR	SEQ0200	14	2	17	4	22	2	0	0	8	2	0	0
SEQ0306	SEQ0306	126	3	128	2	663	9	292	4	573	4	2,455	9
recG	SEQ0454	35	7	30	4	31	4	0	0	7	1	0	0
SEQ0562	SEQ0562	233	18	99	13	130	16	9	1	22	5	0	0
ppc	SEQ0776	57	5	69	9	38	3	0	0	0	0	0	0
addA	SEQ0953	292	30	206	23	123	15	65	7	35	5	0	0
SEQ1028	SEQ1028	54	8	71	10	52	8	0	0	3	1	0	0
SEQ1146	SEQ1146	33	7	21	3	13	2	7	1	4	1	0	0
ldh	SEQ1169	35	5	33	4	13	2	0	0	0	0	0	0
SEQ1304	SEQ1304	24	5	36	6	35	6	0	0	9	2	0	0
mnmE	SEQ1365	259	15	128	12	186	21	7	1	18	4	0	0
smc	SEQ1566	23	7	82	13	68	9	33	2	150	18	65	5
pepQ	SEQ1597	65	9	35	6	40	5	7	1	8	2	0	0
yqeK	SEQ1909	262	22	197	17	349	35	51	5	71	12	31	2
hasA	SEQ0269	72	10	83	9	99	11	7	1	7	1	12	1
pyrP	SEQ1316	42	2	15	3	45	3	7	1	3	1	0	0
eqbE	SEQ1242	4,397	243	6,294	211	3,429	213	15,062	239	19,305	265	11,574	234

Appendix 3

Table A3.1. *S. equi* genes required for fitness determined by infection of Welsh mountain ponies with barcoded *S. equi* IS*S1* libraries. Read counts (reads) and number of unique insertions (ins) per fitness gene in each barcoded library (input) and post (output) infection. Data presented originates from analysis using a minimum input read count per gene of 1,000. Data was additionally analysed using a minimum input read count per gene of 2,000 and 5,000. Comparative analysis conducted between the genes identified as required for infection between the 3 stringencies analysis identified 113 consensus genes (stringency analysis column). Other results from the comparisons are highlighted in the stringency analysis column, however, only genes called in the 1,000 reads per gene stringency analysis are presented. Data was also analysed on a random pony group basis, by separating the output data per animal into 3 random groups of 4 ponies. Data was then compared to the barcoded analysis to assess its validity. Comparative analysis identified 357 consensus genes which are shown in the random pony groups analysis column. Only genes identified in the barcoded analysis are presented.

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0025	5,312	25	7,074	18	6,509	27	6	2	40	2	10	2	113 consensus	357 consensus
SEQ0026	40,835	67	34,990	60	37,054	97	62	12	68	18	33	6	113 consensus	357 consensus
SEQ0027	15,695	28	8,754	21	9,393	40	19	5	31	4	21	3	113 consensus	357 consensus
SEQ0028	11,586	11	7,990	21	14,205	37	7	2	17	3	16	2	113 consensus	357 consensus
SEQ0030	11,874	28	20,454	30	14,390	47	23	6	16	4	20	3	113 consensus	357 consensus
SEQ0032	18,536	34	5,766	17	9,470	39	4	1	13	3	0	0	113 consensus	357 consensus
SEQ0097	6,513	21	10,583	31	7,393	37	61	6	277	10	33	4	113 consensus	357 consensus
SEQ0117	10,611	19	5,627	16	5,774	24	148	6	20	4	28	5	113 consensus	357 consensus
SEQ0128	26,466	49	18,023	35	27,614	57	37	6	162	21	78	9	113 consensus	357 consensus
SEQ0158	11,343	18	9,060	17	8,377	23	75	7	134	7	8	1	113 consensus	357 consensus
SEQ0175	8,100	9	14,706	10	5,801	12	10	2	11	3	10	2	113 consensus	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0245	7,963	16	6,986	15	14,521	25	24	5	39	10	20	3	113 consensus	357 consensus
SEQ0246	7,538	12	8,822	16	8,320	23	3	1	74	11	122	5	113 consensus	357 consensus
SEQ0250	12,952	22	5,214	9	8,865	21	264	7	97	11	5	1	113 consensus	357 consensus
SEQ0302	16,281	33	7,519	19	10,269	33	47	8	48	10	26	5	113 consensus	357 consensus
SEQ0313	5,274	19	7,075	13	12,182	19	28	4	89	5	15	3	113 consensus	357 consensus
SEQ0388	6,014	15	9,504	19	7,808	24	10	3	18	4	10	2	113 consensus	357 consensus
SEQ0391	16,183	32	30,677	48	33,671	52	58	11	102	15	44	7	113 consensus	357 consensus
SEQ0402	13,971	24	8,547	18	7,561	30	35	3	17	5	13	2	113 consensus	357 consensus
SEQ0417	9,222	24	8,404	24	8,957	36	35	8	35	7	7	1	113 consensus	357 consensus
SEQ0431	349,27	50	22,865	41	30,379	56	326	9	700	15	81	9	113 consensus	357 consensus
SEQ0434	8,785	14	8,654	16	9,669	16	66	8	14	3	31	4	113 consensus	357 consensus
SEQ0445	7,582	14	6,384	20	6,452	21	128	4	60	8	116	5	113 consensus	357 consensus
SEQ0460	9,324	20	12,154	19	7,970	23	17	5	29	6	5	1	113 consensus	357 consensus
SEQ0500	17,602	25	10,682	16	18,712	43	3	1	39	11	21	4	113 consensus	357 consensus
SEQ0570	21,007	31	16,851	31	20,518	52	42	11	55	12	18	3	113 consensus	357 consensus
SEQ0571	13,139	17	6,864	12	8,660	21	26	5	12	3	21	3	113 consensus	357 consensus
SEQ0572	36,336	53	27,205	52	39,854	78	74	15	59	13	67	6	113 consensus	357 consensus
SEQ0584	6,262	17	8,921	16	7,605	22	7	2	151	12	98	3	113 consensus	357 consensus
SEQ0601	9,343	28	5,539	20	18,650	39	24	6	45	9	46	7	113 consensus	357 consensus
SEQ0610	9,361	23	26,803	22	7,053	32	14	2	45	6	18	2	113 consensus	357 consensus
SEQ0628	7,566	12	7,247	12	8,244	11	89	4	120	3	13	2	113 consensus	357 consensus
SEQ0633	7,174	14	10,968	16	11,532	35	20	4	45	10	33	4	113 consensus	357 consensus
SEQ0635	8,493	16	11,125	19	15,511	18	97	10	186	6	50	6	113 consensus	357 consensus
SEQ0647	13,787	18	11,519	26	6,326	24	91	4	19	4	5	1	113 consensus	357 consensus
SEQ0668	9,030	28	5,287	14	11,586	36	9	2	27	6	10	2	113 consensus	357 consensus
SEQ0697	12,604	29	7,403	26	12,077	52	56	11	123	16	7	1	113 consensus	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0698	7,087	21	11,432	30	10,182	41	73	4	22	4	46	5	113 consensus	357 consensus
SEQ0701	6,105	15	5,698	17	5,794	28	240	6	32	4	42	3	113 consensus	357 consensus
SEQ0723	10,461	19	13,318	20	8,706	27	19	4	57	12	164	14	113 consensus	357 consensus
SEQ0726	6,807	20	5,506	16	13,525	33	15	3	29	7	65	8	113 consensus	357 consensus
SEQ0728	6,658	20	6,798	24	12,256	23	14	4	70	10	103	10	113 consensus	357 consensus
SEQ0780	10,811	24	8,433	19	7,182	29	13	4	126	11	34	3	113 consensus	357 consensus
SEQ0855	6,240	19	7,324	17	10,134	29	9	2	69	7	37	7	113 consensus	357 consensus
SEQ0857	7,373	18	6,284	23	10,263	28	36	7	43	8	26	3	113 consensus	357 consensus
SEQ0933	9,746	25	13,107	29	16,180	37	58	13	48	10	47	6	113 consensus	357 consensus
SEQ0934	6,974	26	15,556	30	5,822	28	9	3	359	8	18	3	113 consensus	357 consensus
SEQ0938	15,864	66	33,289	57	28,646	63	13	3	35	8	33	3	113 consensus	357 consensus
SEQ0954	16,841	28	16,335	29	14,007	41	425	10	192	7	28	5	113 consensus	357 consensus
SEQ0955	13,667	10	6,577	16	5,349	20	38	9	33	6	16	3	113 consensus	357 consensus
SEQ0992	18,751	28	10,200	25	20,500	38	17	4	34	7	29	4	113 consensus	357 consensus
SEQ1003	9,833	24	6,881	22	6,258	28	12	3	15	4	13	2	113 consensus	357 consensus
SEQ1016	6,661	18	7,776	22	7,582	28	41	6	203	9	24	3	113 consensus	357 consensus
SEQ1018	12,284	31	13,789	22	16,973	35	36	6	127	12	55	8	113 consensus	357 consensus
SEQ1035	5,772	17	5,273	12	5,792	25	6	2	28	6	15	2	113 consensus	357 consensus
SEQ1041	6,510	18	5,188	14	10,045	25	157	5	287	11	101	4	113 consensus	357 consensus
SEQ1058	6,704	22	12,404	25	8,868	22	7	2	342	10	20	3	113 consensus	357 consensus
SEQ1084	5,973	12	5,943	14	5,543	22	9	2	52	7	28	4	113 consensus	357 consensus
SEQ1137	14,223	37	9,915	35	18,089	49	147	10	114	15	190	16	113 consensus	357 consensus
SEQ1138	6,087	32	11,216	31	7,957	47	10	2	16	5	70	10	113 consensus	357 consensus
SEQ1198	6,779	17	6,255	13	7,180	21	21	4	55	8	0	0	113 consensus	357 consensus
SEQ1291	9,099	22	11,139	22	9,055	26	133	10	403	10	24	2	113 consensus	357 consensus
SEQ1302	11,090	20	14,293	29	13,619	40	47	6	229	16	10	2	113 consensus	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1312	12,128	27	16,871	33	13,392	43	25	6	201	19	31	4	113 consensus	357 consensus
SEQ1313	15,336	67	21,657	83	21,119	85	152	25	84	22	88	12	113 consensus	357 consensus
SEQ1314	7,228	25	8,042	22	10,479	37	10	3	52	11	104	6	113 consensus	357 consensus
SEQ1352	5,899	26	8,313	26	7,715	35	57	5	26	6	42	7	113 consensus	357 consensus
SEQ1360	6,086	19	9,077	24	11,628	41	15	5	20	6	34	5	113 consensus	357 consensus
SEQ1407	19,100	51	21,131	63	12,819	71	252	22	61	15	103	9	113 consensus	357 consensus
SEQ1410	6,357	22	6,884	11	10,657	30	19	5	35	6	20	3	113 consensus	357 consensus
SEQ1412	7,656	22	12,029	22	9,566	25	14	4	43	7	50	5	113 consensus	357 consensus
SEQ1413	14,670	49	16,480	52	26,726	77	29	7	115	20	88	9	113 consensus	357 consensus
SEQ1415	21,992	45	13,008	33	13,014	45	38	11	147	10	98	7	113 consensus	357 consensus
SEQ1470	7,435	28	7,005	25	5,365	27	42	7	167	8	50	8	113 consensus	357 consensus
SEQ1479	21,651	42	35,326	61	20,340	70	96	11	342	18	127	16	113 consensus	357 consensus
SEQ1482	6,860	18	5,418	13	8,251	25	35	6	69	10	37	3	113 consensus	357 consensus
SEQ1537	5,521	21	5,214	19	10,826	32	15	4	26	6	23	4	113 consensus	357 consensus
SEQ1545	9,706	21	8,547	17	8,548	24	16	5	33	8	23	3	113 consensus	357 consensus
SEQ1546	5,187	14	11,718	23	8,269	26	11	3	104	8	29	4	113 consensus	357 consensus
SEQ1547	9,247	20	10,681	21	9,577	32	33	8	271	16	13	2	113 consensus	357 consensus
SEQ1551	7,458	17	5,433	15	9,941	32	11	3	15	4	16	3	113 consensus	357 consensus
SEQ1552	12,683	15	5,641	14	13,303	28	30	6	35	5	0	0	113 consensus	357 consensus
SEQ1566	9,637	43	6,906	31	12,233	67	77	14	100	24	50	8	113 consensus	357 consensus
SEQ1679	12,108	18	13,052	21	13,874	39	49	11	34	7	44	6	113 consensus	357 consensus
SEQ1681	16,411	20	34,678	17	8,214	28	86	9	45	9	42	6	113 consensus	357 consensus
SEQ1689	10,172	14	10,366	15	7,233	31	17	5	355	9	24	3	113 consensus	357 consensus
SEQ1692	10,000	23	9,958	21	7,536	38	92	12	54	10	13	2	113 consensus	357 consensus
SEQ1697	5,336	16	5,651	16	10,779	25	31	7	157	5	28	4	113 consensus	357 consensus
SEQ1724	5,732	14	5,395	18	6,998	32	39	8	14	3	41	3	113 consensus	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1734	11,538	22	8,270	25	8,262	32	44	9	80	9	52	6	113 consensus	357 consensus
SEQ1751	7,341	21	7,997	21	13,587	35	11	3	446	8	11	2	113 consensus	357 consensus
SEQ1808	8,125	14	8,381	16	21,208	24	15	4	68	9	18	3	113 consensus	357 consensus
SEQ1815	11,729	21	11,326	19	12,456	32	10	2	93	10	5	1	113 consensus	357 consensus
SEQ1817	6,208	13	8,912	17	8,201	25	13	3	128	6	7	1	113 consensus	357 consensus
SEQ1840	7,714	14	7,433	11	5,706	24	9	2	25	5	26	4	113 consensus	357 consensus
SEQ1898	7,656	26	17,907	24	18,125	48	182	5	66	13	86	8	113 consensus	357 consensus
SEQ1899	5,065	13	9,037	16	5,192	13	4	1	10	3	0	0	113 consensus	357 consensus
SEQ1900	17,632	26	8,312	18	12,258	38	20	5	61	9	36	5	113 consensus	357 consensus
SEQ1906	22,922	33	10,786	24	12,749	40	100	15	140	9	75	7	113 consensus	357 consensus
SEQ1926	6,317	17	13,064	21	11,404	36	6	2	45	8	0	0	113 consensus	357 consensus
SEQ1929	9,162	18	5,851	15	5,927	19	4	1	9	3	28	3	113 consensus	357 consensus
SEQ1934	10,267	28	17,791	23	10,146	27	36	8	22	5	34	6	113 consensus	357 consensus
SEQ1945	15,126	32	12,300	22	18,271	45	19	4	62	7	28	4	113 consensus	357 consensus
SEQ1961	5,057	12	8,071	20	5,190	27	94	4	40	7	5	1	113 consensus	357 consensus
SEQ2032	6,802	17	7,064	13	7,705	20	28	7	175	8	163	3	113 consensus	357 consensus
SEQ2100	22,347	36	16,370	35	33,515	48	142	10	219	13	75	7	113 consensus	357 consensus
SEQ2103	7,164	10	5,451	16	8,500	22	98	6	9	2	20	3	113 consensus	357 consensus
SEQ2126	5,536	13	8,953	16	9,269	21	6	2	18	5	13	2	113 consensus	357 consensus
SEQ2135	9,784	14	8,696	12	6,441	18	22	6	8	2	0	0	113 consensus	357 consensus
SEQ2161	12,681	18	9,469	11	9,657	25	16	4	8	2	7	1	113 consensus	357 consensus
SEQ2190	8,093	18	6,680	18	10,833	27	3	1	17	5	16	2	113 consensus	357 consensus
SEQ0033	5,710	7	1,506	5	4,872	15	3	1	20	3	0	0	1,000 only	357 consensus
SEQ0035	4,338	6	1,930	2	2,224	5	7	2	11	2	10	2	1,000 only	357 consensus
SEQ0036	1,886	2	1,469	1	1,902	6	3	1	0	0	0	0	1,000 only	357 consensus
SEQ0042	1,421	8	1,998	8	1,093	7	9	3	3	1	0	0	1,000 only	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0085	1,242	10	3,244	5	2,391	8	3	1	6	2	8	1	1,000 only	357 consensus
SEQ0095	1,678	6	6,061	8	3,411	13	3	1	13	3	7	1	1,000 only	357 consensus
SEQ0118	1,460	18	1,704	20	1,285	24	29	4	56	10	21	3	1,000 only	357 consensus
SEQ0132	1,663	8	1,397	7	1,335	18	5	1	32	3	8	1	1,000 only	357 consensus
SEQ0136	1,677	6	3,467	6	4,391	7	3	1	4	1	0	0	1,000 only	357 consensus
SEQ0150	2,870	8	1,799	10	1,945	11	0	0	3	1	5	1	1,000 only	357 consensus
SEQ0155	4,100	14	1,817	7	4,835	11	10	3	19	5	20	3	1,000 only	357 consensus
SEQ0164	4,860	15	1,675	10	4,596	15	14	3	36	6	10	2	1,000 only	357 consensus
SEQ0168	2,218	5	1,587	5	2,114	9	0	0	24	4	0	0	1,000 only	357 consensus
SEQ0178	1,003	8	3,228	10	1,867	13	0	0	3	1	11	2	1,000 only	357 consensus
SEQ0183	1,586	4	1,263	3	2,934	5	0	0	14	3	0	0	1,000 only	357 consensus
SEQ0186	2,632	6	5,114	10	1,505	8	8	2	3	1	5	1	1,000 only	357 consensus
SEQ0193	1,043	11	2,308	13	1,436	8	3	1	56	4	7	1	1,000 only	357 consensus
SEQ0205	1,928	7	3,732	8	4,788	9	7	1	22	5	0	0	1,000 only	357 consensus
SEQ0206	3,630	5	1,617	3	3,167	9	7	2	21	4	0	0	1,000 only	357 consensus
SEQ0221	4,022	13	1,919	10	7,376	22	8	2	29	3	20	3	1,000 only	357 consensus
SEQ0263	4,006	3	5,715	4	1,317	7	12	2	18	2	0	0	1,000 only	357 consensus
SEQ0264	3,871	9	1,909	4	5,128	16	0	0	11	3	5	1	1,000 only	357 consensus
SEQ0268	2,528	9	2,150	5	1,230	10	0	0	20	4	5	1	1,000 only	357 consensus
SEQ0306	1,526	17	1,093	11	1,028	14	13	3	53	12	16	2	1,000 only	357 consensus
SEQ0330	2,344	13	1,715	6	1,658	11	7	2	20	5	18	2	1,000 only	357 consensus
SEQ0332	6,173	14	1,706	11	4,539	18	15	3	17	5	5	1	1,000 only	357 consensus
SEQ0370b	1,962	3	2,916	6	1,685	2	14	3	3	1	0	0	1,000 only	357 consensus
SEQ0469	3,715	17	1,598	5	2,237	14	8	2	0	0	28	4	1,000 only	357 consensus
SEQ0506	1,297	9	1,115	13	1,130	11	3	1	3	1	5	1	1,000 only	357 consensus
SEQ0550	2,536	7	1,427	5	3,927	15	15	4	172	4	5	1	1,000 only	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0561	1,194	7	3,319	5	5,000	9	0	0	14	4	11	2	1,000 only	357 consensus
SEQ0562	1,579	13	2,323	19	1,484	15	32	4	10	3	20	3	1,000 only	357 consensus
SEQ0577	2,786	3	1,122	4	1,614	5	7	2	0	0	0	0	1,000 only	357 consensus
SEQ0581	1,110	2	7,056	8	2,954	11	8	2	4	1	8	1	1,000 only	357 consensus
SEQ0582	4,072	9	1,302	3	2,605	8	13	4	16	4	0	0	1,000 only	357 consensus
SEQ0607	1,111	2	2,660	2	2,240	8	4	1	70	2	0	0	1,000 only	357 consensus
SEQ0643	2,142	9	1,809	11	3,799	20	18	4	17	5	10	2	1,000 only	357 consensus
SEQ0655	1,902	22	1,770	28	1,772	23	11	3	23	7	7	1	1,000 only	357 consensus
SEQ0722	3,040	12	1,974	8	4,349	9	18	4	6	1	76	2	1,000 only	357 consensus
SEQ0772	4,656	6	1,842	8	1,781	6	13	4	28	5	0	0	1,000 only	357 consensus
SEQ0773	7,587	12	6,289	12	1,609	10	14	4	49	7	18	2	1,000 only	357 consensus
SEQ0795	1,870	6	3,176	5	1,756	11	7	2	8	2	0	0	1,000 only	357 consensus
SEQ0797	5,013	14	1,680	8	3,609	17	28	6	80	7	23	4	1,000 only	357 consensus
SEQ0821	1,825	4	1,154	4	2,082	12	0	0	7	1	0	0	1,000 only	357 consensus
SEQ0901	3,481	13	2,602	6	1,890	9	26	2	20	4	7	1	1,000 only	357 consensus
SEQ0907	4,110	28	1,545	14	1,205	15	7	2	43	9	24	2	1,000 only	357 consensus
SEQ0991	1,764	15	2,891	15	3,413	21	3	1	12	2	11	2	1,000 only	357 consensus
SEQ1013	4,316	17	1,838	11	4,714	23	23	6	17	3	16	1	1,000 only	357 consensus
SEQ1025	3,482	11	1,410	9	4,391	11	7	2	12	3	11	2	1,000 only	357 consensus
SEQ1059	1,302	6	6,236	11	2,240	12	14	3	21	4	0	0	1,000 only	357 consensus
SEQ1093	3,215	11	6,288	11	1,952	9	34	3	24	2	18	2	1,000 only	357 consensus
SEQ1115	1,255	6	2,019	7	2,198	9	0	0	0	0	0	0	1,000 only	357 consensus
SEQ1202	3,561	11	3,224	8	1,788	11	64	5	10	3	0	0	1,000 only	357 consensus
SEQ1249	7,095	20	1,623	8	6,098	17	11	3	24	5	18	3	1,000 only	357 consensus
SEQ1317	2,565	11	2,483	13	1,959	15	26	5	30	8	11	2	1,000 only	357 consensus
SEQ1339	4,695	16	1,122	8	5,991	13	10	2	32	8	81	3	1,000 only	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1344	1,317	8	5,140	13	4,094	14	8	2	24	5	0	0	1,000 only	357 consensus
SEQ1364	2,576	8	1,730	8	6,489	10	7	2	42	3	7	1	1,000 only	357 consensus
SEQ1367	1,838	11	1,550	14	2,157	19	4	1	14	3	18	2	1,000 only	357 consensus
SEQ1386	2,959	11	1,792	8	2,117	11	0	0	3	1	0	0	1,000 only	357 consensus
SEQ1398	1,677	4	3,863	12	5,705	19	29	3	47	9	55	7	1,000 only	357 consensus
SEQ1437	12,281	45	3,802	20	9,062	48	51	10	55	10	186	4	1,000 only	357 consensus
SEQ1500	1,982	9	1,606	7	3,938	9	6	2	25	5	31	5	1,000 only	357 consensus
SEQ1510	1,482	11	1,721	7	1,473	10	12	3	10	3	33	3	1,000 only	357 consensus
SEQ1526	3,278	8	1,170	2	2,294	8	3	1	3	1	0	0	1,000 only	357 consensus
SEQ1527	2,647	7	1,415	5	1,333	5	3	1	8	2	0	0	1,000 only	357 consensus
SEQ1536	1,929	5	2,581	6	2,598	17	4	1	6	2	0	0	1,000 only	357 consensus
SEQ1562	1,990	8	2,326	6	3,545	10	6	2	19	4	46	1	1,000 only	357 consensus
SEQ1597	1,779	18	1,544	13	2,642	29	34	5	13	3	24	3	1,000 only	357 consensus
SEQ1632	1,957	8	1,078	5	5,587	15	0	0	41	1	5	1	1,000 only	357 consensus
SEQ1640	5,304	6	3,967	9	1,662	9	33	7	25	2	20	3	1,000 only	357 consensus
SEQ1669	3,067	6	1,782	6	6,625	10	3	1	3	1	13	2	1,000 only	357 consensus
SEQ1677a	1,343	5	2,231	5	1,427	8	7	2	0	0	5	1	1,000 only	357 consensus
SEQ1740	2,086	6	1,244	7	2,119	12	6	2	34	4	29	5	1,000 only	357 consensus
SEQ1741	1,253	9	5,836	7	4,635	14	15	3	37	9	20	3	1,000 only	357 consensus
SEQ1748	2,459	4	1,002	4	3,281	8	3	1	5	1	18	1	1,000 only	357 consensus
SEQ1753	3,569	8	4,670	10	1,847	14	9	1	7	2	18	2	1,000 only	357 consensus
SEQ1771	4,516	14	3,186	22	1,927	19	15	3	155	7	39	5	1,000 only	357 consensus
SEQ1819	4,566	10	1,778	7	7,758	17	7	2	31	5	5	1	1,000 only	357 consensus
SEQ1822	1,698	6	2,390	9	2,972	18	25	3	41	5	7	1	1,000 only	357 consensus
SEQ1823	3,112	10	4,563	10	1,440	12	0	0	10	3	15	2	1,000 only	357 consensus
SEQ1843	3,694	12	3,104	11	1,742	14	9	2	19	5	10	1	1,000 only	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1865	1,692	4	1,437	5	3,272	12	0	0	9	3	0	0	1,000 only	357 consensus
SEQ1916	1,421	17	1,411	15	3,799	41	8	2	51	8	15	2	1,000 only	357 consensus
SEQ1917	2,742	20	1,609	17	3,141	25	3	1	18	5	50	4	1,000 only	357 consensus
SEQ1918	1,039	13	3,650	21	6,881	32	8	2	87	11	21	2	1,000 only	357 consensus
SEQ1927	7,820	12	3,164	7	1,288	8	14	3	0	0	21	3	1,000 only	357 consensus
SEQ1980	1,391	5	2,798	5	1,758	7	50	8	0	0	0	0	1,000 only	357 consensus
SEQ2028	2,037	6	1,286	3	6,543	6	0	0	21	4	5	1	1,000 only	357 consensus
SEQ2052	3,794	7	5,342	7	1,988	8	0	0	28	6	5	1	1,000 only	357 consensus
SEQ2053	3,737	6	3,394	8	1,536	8	3	1	0	0	13	1	1,000 only	357 consensus
SEQ2054	4,971	8	2,426	6	1,098	8	3	1	24	4	0	0	1,000 only	357 consensus
SEQ2058	2,843	7	3,635	6	1,692	11	8	2	39	3	5	1	1,000 only	357 consensus
SEQ2074	4,488	9	2,194	9	1,783	9	16	3	4	1	26	3	1,000 only	357 consensus
SEQ2097	1,840	12	1,952	6	3,077	8	3	1	16	4	20	1	1,000 only	357 consensus
SEQ2114	3,797	9	5,886	12	1,153	6	129	2	60	5	20	3	1,000 only	357 consensus
SEQ2155	7,004	6	1,851	8	5,484	13	11	3	15	4	16	2	1,000 only	357 consensus
SEQ2157	1,486	4	6,211	4	1,114	9	0	0	9	3	5	1	1,000 only	357 consensus
SEQ2162	1,393	4	2,179	5	4,203	7	4	1	7	2	0	0	1,000 only	357 consensus
SEQ0012	3,491	16	4,084	17	4,125	27	16	4	6	2	13	2	1,000/2,000	357 consensus
SEQ0018	7,009	22	6,888	17	3,303	34	11	3	3	1	23	4	1,000/2,000	357 consensus
SEQ0029	4,117	8	8,822	12	4,025	14	0	0	0	0	0	0	1,000/2,000	357 consensus
SEQ0034	11,439	19	10,263	21	4,673	22	10	3	22	4	0	0	1,000/2,000	357 consensus
SEQ0043	12,284	40	3,199	26	4,112	50	20	6	68	10	24	3	1,000/2,000	357 consensus
SEQ0084	13,568	24	4,049	12	14,342	34	30	5	30	8	65	3	1,000/2,000	357 consensus
SEQ0091	5,332	17	6,712	9	3,498	10	3	1	16	4	8	1	1,000/2,000	357 consensus
SEQ0110	3,905	9	3,795	9	5,564	13	4	1	4	1	0	0	1,000/2,000	357 consensus
SEQ0148	4,710	12	3,491	12	5,409	29	46	3	76	3	7	1	1,000/2,000	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0151	4,871	11	4,135	6	2,626	15	7	2	12	2	15	2	1,000/2,000	357 consensus
SEQ0174	4,913	8	5,619	7	5,121	11	6	2	27	4	10	2	1,000/2,000	357 consensus
SEQ0181	2,595	13	3,353	7	6,771	16	4	1	22	7	0	0	1,000/2,000	357 consensus
SEQ0202	3,157	8	3,971	11	4,552	15	42	5	21	4	10	2	1,000/2,000	357 consensus
SEQ0216	4,009	11	7,107	7	3,913	14	4	1	149	2	16	3	1,000/2,000	357 consensus
SEQ0220	2,381	16	2,017	9	5,187	18	22	2	142	9	0	0	1,000/2,000	357 consensus
SEQ0223	3,873	14	11,593	15	7,105	18	24	5	29	6	11	2	1,000/2,000	357 consensus
SEQ0236	4,520	10	2,175	9	4,678	24	12	3	35	4	13	2	1,000/2,000	357 consensus
SEQ0243	2,295	5	3,517	7	3,578	13	12	3	20	4	0	0	1,000/2,000	357 consensus
SEQ0244	3,523	10	7,591	12	16,473	12	3	1	99	4	16	3	1,000/2,000	357 consensus
SEQ0255	7,814	18	2,575	11	2,749	9	16	4	20	6	18	3	1,000/2,000	357 consensus
SEQ0265	3,127	6	3,094	9	10,066	17	7	2	17	4	63	3	1,000/2,000	357 consensus
SEQ0293	5,896	26	4,571	22	7,738	24	46	4	144	12	88	3	1,000/2,000	357 consensus
SEQ0295	4,418	9	2,636	6	4,715	11	17	4	0	0	16	3	1,000/2,000	357 consensus
SEQ0299	6,541	10	2,930	7	9,602	11	19	5	35	6	20	3	1,000/2,000	357 consensus
SEQ0324	6,535	8	3,348	7	2,278	12	13	3	39	7	23	4	1,000/2,000	357 consensus
SEQ0337	3,747	10	3,311	11	2,290	16	36	5	24	4	5	1	1,000/2,000	357 consensus
SEQ0363	2,842	6	4,274	7	3,897	10	4	1	4	1	8	1	1,000/2,000	357 consensus
SEQ0366	4,847	13	5,929	11	8,513	20	0	0	14	4	34	5	1,000/2,000	357 consensus
SEQ0379	4,244	14	4,959	10	5,521	21	13	4	28	7	7	1	1,000/2,000	357 consensus
SEQ0386	4,615	9	4,755	7	7,087	11	6	1	85	2	60	7	1,000/2,000	357 consensus
SEQ0398	3,286	14	2,191	8	5,557	16	84	3	7	2	0	0	1,000/2,000	357 consensus
SEQ0454	7,566	37	4,833	44	4,601	50	60	14	82	17	29	5	1,000/2,000	357 consensus
SEQ0455	9,991	19	4,818	20	13,304	25	74	10	30	7	68	7	1,000/2,000	357 consensus
SEQ0495	4,861	20	6,129	13	15,303	25	55	10	24	5	26	3	1,000/2,000	357 consensus
SEQ0507	3,851	22	4,809	20	3,895	28	17	3	18	5	15	2	1,000/2,000	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0530	2,905	5	2,168	3	2,961	2	24	2	3	1	0	0	1,000/2,000	357 consensus
SEQ0531	10,913	15	2,775	6	3,238	9	15	4	3	1	37	4	1,000/2,000	357 consensus
SEQ0535	9,094	15	4,161	6	2,754	11	6	2	8	2	5	1	1,000/2,000	357 consensus
SEQ0536	6,088	11	3,691	8	3,982	17	51	9	34	2	13	2	1,000/2,000	357 consensus
SEQ0551	5,062	5	3,090	9	2,963	18	4	1	206	3	26	4	1,000/2,000	357 consensus
SEQ0576	4,029	8	5,329	7	4,260	12	14	3	18	2	13	2	1,000/2,000	357 consensus
SEQ0602	3,738	12	6,638	17	7,637	30	23	4	12	2	50	5	1,000/2,000	357 consensus
SEQ0608	4,114	14	7,028	19	10,281	28	26	6	44	5	33	4	1,000/2,000	357 consensus
SEQ0644	2,305	12	3,535	10	3,365	14	54	5	0	0	5	1	1,000/2,000	357 consensus
SEQ0658	6,437	12	2,806	6	10,708	25	23	4	17	5	28	4	1,000/2,000	357 consensus
SEQ0670	2,710	24	3,473	26	5,619	23	23	5	36	6	60	10	1,000/2,000	357 consensus
SEQ0683	9,924	16	3,721	13	6,393	25	38	7	30	7	63	8	1,000/2,000	357 consensus
SEQ0685b	2,265	8	2,442	7	7,719	18	0	0	24	5	5	1	1,000/2,000	357 consensus
SEQ0690	4,821	13	4,522	7	3,080	14	3	1	8	2	15	2	1,000/2,000	357 consensus
SEQ0693	5,899	17	3,138	13	2,886	23	20	6	142	9	0	0	1,000/2,000	357 consensus
SEQ0700	4,548	7	2,236	9	3,319	9	0	0	8	2	13	1	1,000/2,000	357 consensus
SEQ0735	8,842	22	4,993	13	10,936	38	61	5	165	6	24	4	1,000/2,000	357 consensus
SEQ0768	6,552	14	10,608	22	4,811	25	86	3	10	3	80	4	1,000/2,000	357 consensus
SEQ0776	8,751	59	6,081	52	4,875	56	25	7	55	13	26	5	1,000/2,000	357 consensus
SEQ0802	2,775	5	4,415	10	7,440	16	3	1	13	3	7	1	1,000/2,000	357 consensus
SEQ0822	5,775	12	3,474	12	2,260	8	9	2	25	4	44	4	1,000/2,000	357 consensus
SEQ0835	5,240	20	4,786	17	4,198	27	14	3	15	4	26	3	1,000/2,000	357 consensus
SEQ0836	4,850	20	3,907	12	5,673	21	7	2	34	7	20	3	1,000/2,000	357 consensus
SEQ0837	5,057	10	4,333	9	4,440	11	6	1	192	11	23	4	1,000/2,000	357 consensus
SEQ0847	4,731	20	7,647	23	5,126	29	15	3	299	5	63	7	1,000/2,000	357 consensus
SEQ0851	2,839	8	5,890	11	2,249	10	24	5	10	3	0	0	1,000/2,000	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0877	6,625	16	6,208	18	4,203	20	38	5	26	5	20	2	1,000/2,000	357 consensus
SEQ0879	2,219	5	5,508	8	2,452	11	25	6	52	9	49	3	1,000/2,000	357 consensus
SEQ0969	4,781	16	3,246	15	4,532	30	12	3	24	6	10	2	1,000/2,000	357 consensus
SEQ0970	3,609	15	3,201	8	14,000	22	3	1	0	0	28	3	1,000/2,000	357 consensus
SEQ0973	8,880	27	3,871	27	3,906	39	12	4	43	9	37	5	1,000/2,000	357 consensus
SEQ0975	3,370	19	22,513	21	12,495	41	21	5	43	11	0	0	1,000/2,000	357 consensus
SEQ0995	8,314	13	4,306	9	7,852	18	20	4	34	4	7	1	1,000/2,000	357 consensus
SEQ1004	6,753	23	2,686	11	6,062	26	4	1	30	6	7	1	1,000/2,000	357 consensus
SEQ1005	2,660	11	2,073	7	3,052	11	8	2	10	3	5	1	1,000/2,000	357 consensus
SEQ1075	6,542	22	9,913	28	3,879	27	13	3	28	7	7	1	1,000/2,000	357 consensus
SEQ1130	4,954	23	5,085	20	7,655	29	12	3	33	4	21	3	1,000/2,000	357 consensus
SEQ1167	5,127	21	4,489	16	2,415	12	16	4	59	4	44	7	1,000/2,000	357 consensus
SEQ1201	8,421	20	6,621	15	3,756	17	28	5	36	2	18	1	1,000/2,000	357 consensus
SEQ1301	5,967	15	2,660	10	4,206	22	39	4	13	4	16	3	1,000/2,000	357 consensus
SEQ1304	6,738	26	4,791	24	5,395	38	6	2	0	0	18	3	1,000/2,000	357 consensus
SEQ1315	2,278	14	5,198	18	11,228	29	8	2	23	6	23	3	1,000/2,000	357 consensus
SEQ1316	2,122	22	4,713	32	3,610	36	25	6	39	10	16	3	1,000/2,000	357 consensus
SEQ1320	2,573	14	15,280	22	4,530	20	64	5	15	4	36	4	1,000/2,000	357 consensus
SEQ1342	4,632	27	4,056	26	6,361	34	20	4	116	21	31	5	1,000/2,000	357 consensus
SEQ1343	3,494	10	3,086	8	8,075	15	3	1	0	0	18	3	1,000/2,000	357 consensus
SEQ1357	2,187	6	6,626	12	5,493	13	13	4	172	9	70	2	1,000/2,000	357 consensus
SEQ1362	3,359	3	8,132	11	3,564	14	7	2	4	1	59	1	1,000/2,000	357 consensus
SEQ1365	2,834	40	2,263	27	2,338	44	13	4	19	5	44	6	1,000/2,000	357 consensus
SEQ1381	3,622	15	5,282	15	4,667	15	27	7	15	2	18	3	1,000/2,000	357 consensus
SEQ1432	3,384	13	3,742	12	4,352	24	13	4	24	5	13	2	1,000/2,000	357 consensus
SEQ1450	2,965	18	3,454	16	2,781	34	29	8	70	12	44	5	1,000/2,000	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1452	3,901	16	2,955	14	3,409	21	3	1	28	6	13	2	1,000/2,000	357 consensus
SEQ1453	5,943	16	2,102	13	4,998	39	26	6	38	9	23	4	1,000/2,000	357 consensus
SEQ1467	6,854	14	3,461	10	4,827	21	8	2	15	4	34	3	1,000/2,000	357 consensus
SEQ1543	4,997	11	2,495	7	3,897	14	18	4	55	8	0	0	1,000/2,000	357 consensus
SEQ1549	4,093	11	3,378	13	3,278	13	54	4	65	9	18	3	1,000/2,000	357 consensus
SEQ1576	2,106	17	7,094	30	4,594	30	18	5	56	10	72	10	1,000/2,000	357 consensus
SEQ1610	17,360	13	4,740	15	6,888	21	57	9	32	7	33	5	1,000/2,000	357 consensus
SEQ1646	5,616	9	5,930	14	3,918	19	41	4	18	5	50	6	1,000/2,000	357 consensus
SEQ1647	5,135	12	4,096	10	8,877	30	20	5	61	6	10	1	1,000/2,000	357 consensus
SEQ1675	5,023	12	2,784	5	3,797	14	29	4	34	8	18	3	1,000/2,000	357 consensus
SEQ1678	4,243	18	4,752	17	13,845	26	24	6	17	5	11	2	1,000/2,000	357 consensus
SEQ1699	4,233	10	4,902	13	3,372	15	23	5	3	1	37	2	1,000/2,000	357 consensus
SEQ1718	4,114	15	2,439	21	2,144	26	73	6	16	4	41	6	1,000/2,000	357 consensus
SEQ1743	3,184	7	2,319	6	5,310	10	12	3	22	3	33	2	1,000/2,000	357 consensus
SEQ1745	3,742	7	4,632	13	3,601	15	6	2	44	4	15	2	1,000/2,000	357 consensus
SEQ1750	3,632	11	15,487	5	2,664	12	15	4	13	3	0	0	1,000/2,000	357 consensus
SEQ1755	2,570	9	3,674	13	4,014	15	0	0	184	8	5	1	1,000/2,000	357 consensus
SEQ1782	2,303	8	4,839	10	3,157	14	14	3	14	2	18	3	1,000/2,000	357 consensus
SEQ1785	3,635	18	6,102	15	4,733	29	19	6	11	3	11	2	1,000/2,000	357 consensus
SEQ1786	2,056	9	5,644	11	5,205	16	3	1	22	5	5	1	1,000/2,000	357 consensus
SEQ1807	7,381	21	10,985	10	3,705	14	58	6	17	3	28	3	1,000/2,000	357 consensus
SEQ1844	7,425	14	4,759	10	7,968	27	26	7	70	7	50	2	1,000/2,000	357 consensus
SEQ1845	5,930	11	4,017	10	2,253	9	74	4	3	1	7	1	1,000/2,000	357 consensus
SEQ1848	3,803	8	4,183	8	6,114	16	12	2	16	4	23	2	1,000/2,000	357 consensus
SEQ1867	3,563	12	2,452	6	5,053	17	19	4	15	4	13	1	1,000/2,000	357 consensus
SEQ1870	5,022	8	4,246	7	2,199	8	7	2	150	4	0	0	1,000/2,000	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1880	7,169	29	4,881	28	7,922	41	4	1	22	5	41	5	1,000/2,000	357 consensus
SEQ1902	9,629	19	12,996	18	4,464	17	9	2	26	7	18	3	1,000/2,000	357 consensus
SEQ1905	4,536	10	3,902	10	13,521	22	195	6	188	5	72	7	1,000/2,000	357 consensus
SEQ1928	3,981	13	10,374	23	5,916	22	11	1	21	5	5	1	1,000/2,000	357 consensus
SEQ1930	4,177	8	6,737	15	5,085	16	16	5	9	2	0	0	1,000/2,000	357 consensus
SEQ1935	4,710	11	7,247	11	3,787	17	5	1	77	5	10	2	1,000/2,000	357 consensus
SEQ1954	3,170	13	6,512	17	6,726	28	31	7	264	8	46	4	1,000/2,000	357 consensus
SEQ1959	9,157	25	4,842	13	16,183	42	42	9	103	5	23	4	1,000/2,000	357 consensus
SEQ1966	2,324	7	2,986	7	6,068	12	42	4	31	6	28	4	1,000/2,000	357 consensus
SEQ1971	2,427	8	2,081	8	2,571	27	0	0	17	2	42	2	1,000/2,000	357 consensus
SEQ1977	5,959	28	2,916	18	5,137	34	10	3	57	9	29	4	1,000/2,000	357 consensus
SEQ1987	5,034	11	8,006	10	2,498	12	211	5	135	7	5	1	1,000/2,000	357 consensus
SEQ1990	9,799	15	2,158	8	7,988	14	16	5	53	4	10	2	1,000/2,000	357 consensus
SEQ2027	6,011	11	4,788	6	3,196	8	31	4	72	3	13	2	1,000/2,000	357 consensus
SEQ2031	6,657	13	3,458	6	5,646	11	6	1	12	3	16	3	1,000/2,000	357 consensus
SEQ2035	3,861	16	4,526	13	6,676	26	21	5	87	1	11	1	1,000/2,000	357 consensus
SEQ2040	2,154	5	2,535	6	3,123	4	40	2	5	1	0	0	1,000/2,000	357 consensus
SEQ2044	12,523	19	9,542	24	4,580	29	35	9	88	13	26	4	1,000/2,000	357 consensus
SEQ2101	2,887	8	3,909	11	6,132	16	8	2	146	3	5	1	1,000/2,000	357 consensus
SEQ2111	3,929	12	9,685	14	6,559	13	14	2	32	2	0	0	1,000/2,000	357 consensus
SEQ2119	9,351	16	4,160	16	12,808	25	226	7	101	12	33	5	1,000/2,000	357 consensus
SEQ2129	3,978	6	4,376	8	2,313	8	3	1	0	0	0	0	1,000/2,000	357 consensus
SEQ2132	4,424	7	4,254	6	4,749	4	0	0	3	1	0	0	1,000/2,000	357 consensus
SEQ2142	10,407	8	4,049	7	4,343	11	11	2	302	3	7	1	1,000/2,000	357 consensus
SEQ2160	4,802	7	4,485	6	2,578	12	19	5	18	2	24	3	1,000/2,000	357 consensus
SEQ2163	3,281	7	6,163	7	6,253	6	18	4	19	5	0	0	1,000/2,000	357 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ2170	3,307	7	4,667	9	3,424	17	11	3	82	4	0	0	1,000/2,000	357 consensus
SEQ2172	2,250	9	5,087	15	4,829	16	0	0	10	3	5	1	1,000/2,000	357 consensus
SEQ2175	9,307	15	4,637	13	8,605	27	22	5	3	1	0	0	1,000/2,000	357 consensus
SEQ2188	6,060	14	3,132	7	7,409	17	6	2	18	4	21	3	1,000/2,000	357 consensus
SEQ2211	3,681	8	2,381	7	4,973	16	23	4	30	6	0	0	1,000/2,000	357 consensus
SEQ2228	5,512	8	7,994	12	3,342	9	6	2	43	1	5	1	1,000/2,000	357 consensus
SEQ2233	5,554	19	2,482	11	2,333	17	9	2	68	11	5	1	1,000/2,000	357 consensus
SEQ2237	5,669	29	5,143	13	4,589	26	9	3	38	8	5	1	1,000/2,000	357 consensus
SEQ0468	12,515	16	12,884	15	12,628	30	22	6	597	9	7	1	113 consensus	barcoded only
SEQ1411	10,822	24	6,133	22	12,153	31	275	5	47	11	42	7	113 consensus	barcoded only
SEQ0310	1,389	7	6,419	11	13,370	17	0	0	145	9	34	4	1,000 only	barcoded only
SEQ0405	1,747	5	1,482	3	1,562	4	0	0	69	1	0	0	1,000 only	barcoded only
SEQ1363	1,311	7	3,588	11	8,502	20	4	1	247	2	24	4	1,000 only	barcoded only
SEQ0452	2,238	13	5,778	11	7,180	20	27	7	310	8	10	1	1,000/2,000	barcoded only
SEQ0873	9,561	15	5,840	12	3,705	16	16	4	448	8	21	3	1,000/2,000	barcoded only
SEQ1226	7,447	22	3,229	9	5,427	26	46	4	221	6	57	4	1,000/2,000	barcoded only
SEQ1423	6,595	15	3,210	9	5,192	15	9	3	242	4	13	1	1,000/2,000	barcoded only
SEQ1661	9,454	28	9,451	26	4,388	25	322	6	24	3	18	3	1,000/2,000	barcoded only
SEQ2105	9,595	12	9,745	14	2,062	14	6	2	356	2	24	2	1,000/2,000	barcoded only

Table A3.2. *S. equi* genes conferring enhanced fitness upon insertion determined by infection of Welsh mountain ponies with barcoded *S. equi* IS*S1* libraries. Read counts (reads) and number of unique insertions (ins) per gene in each barcoded library (input) and post (output) infection. Data presented originates from analysis using a minimum input read count per gene of 1,000. Data was additionally analysed using a minimum input read count per gene of 2,000 and 5,000. Comparative analysis conducted between the genes identified as conferring an enhanced fitness during infection between the 3 stringencies analysis identified 21 consensus genes (stringency analysis column). Other results from the comparisons are highlighted in the stringency analysis column, however, only genes called in the 1,000 reads per gene stringency analysis are presented. Data was also analysed on a random pony group basis, by separating the output data per animal into 3 random groups of 4 ponies. Data was then compared to the barcoded analysis to assess its validity. Comparative analysis identified 66 consensus genes which are shown in the random pony groups analysis column. Only genes identified in the barcoded analysis are presented.

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0103	14,186	29	16,771	46	29,340	20	13,775	33	117,355	9	1,612,984	11	21 consensus	66 consensus
SEQ0308	18,984	43	37,233	55	71,418	25	29,099	39	288,794	14	240,619	18	21 consensus	66 consensus
SEQ0713	26,184	36	31,789	65	12,463	23	18,415	38	180,480	25	4,730,717	64	21 consensus	66 consensus
SEQ0752b	46,033	92	53,922	122	115,848	39	65,650	80	52,408	31	2,947,764	34	21 consensus	66 consensus
SEQ0785	13,498	16	6,126	22	42,869	8	7,903	12	344,835	6	622	9	21 consensus	66 consensus
SEQ0864	13,336	19	29,803	30	56,045	35	13,284	14	1,372,601	29	9,064	10	21 consensus	66 consensus
SEQ0911	12,105	27	9,578	42	1,319,539	55	34,092	32	186,180	14	1,128	6	21 consensus	66 consensus
SEQ1117	134,206	73	31,281	93	25,347	27	38,806	73	803,527	49	1,492,508	40	21 consensus	66 consensus
SEQ1136	19,565	47	36,080	65	10,147	18	29,983	54	54,578	12	639,328	23	21 consensus	66 consensus
SEQ1229	15,269	58	13,276	84	54,047	32	19,130	65	247,908	21	1,398,533	34	21 consensus	66 consensus
SEQ1263	67,975	161	65,607	195	273,185	71	70,913	149	140,934	42	763,679	53	21 consensus	66 consensus
SEQ1280	22,427	55	28,333	63	426,036	35	40,372	38	39,148	15	558,848	17	21 consensus	66 consensus
SEQ1353	11,294	26	37,538	28	43,129	10	12,768	31	8,614	10	855,348	19	21 consensus	66 consensus
SEQ1426	6,193	8	6,338	13	8,970	5	5,981	9	1,507	4	438,286	9	21 consensus	66 consensus
SEQ1607	11,209	37	5,167	28	41,901	15	14,043	33	456,492	12	1,292	8	21 consensus	66 consensus
SEQ1711	39,856	65	20,201	79	1,549	22	15,701	44	1,168,754	35	106,408	16	21 consensus	66 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ1955	32,606	50	35,838	54	240,005	28	34,087	38	685,527	28	166,577	21	21 consensus	66 consensus
SEQ1975	7,570	24	9,347	41	13,071	16	11,872	26	521	5	383,919	12	21 consensus	66 consensus
SEQ2051	6,637	14	104,812	22	41,568	7	6,388	11	4,022	2	3,697,996	18	21 consensus	66 consensus
SEQ2061	18,849	36	13,755	49	192,603	16	13,053	42	196,141	15	300,995	12	21 consensus	66 consensus
SEQ2094	9,086	36	7,110	33	3,796	10	5,543	24	106,451	7	78,765	5	21 consensus	66 consensus
SEQ1098	23,913	56	22,354	73	354,160	27	23,875	42	74,905	21	3,472	16	1,000 only	66 consensus
SEQ1217	15,492	43	16,662	48	331	9	12,344	43	168,811	16	47,606	6	1,000 only	66 consensus
SEQ1255	8,432	26	13,740	43	36,038	13	13,793	31	164,009	11	1,297	9	1,000 only	66 consensus
SEQ1256	61,772	121	49,214	160	92,049	64	47,063	118	384,526	49	72,679	27	1,000 only	66 consensus
SEQ1445	17,342	39	15,772	53	47,344	21	17,442	35	201,605	13	2,902	13	1,000 only	66 consensus
SEQ1512	4,013	14	1,630	15	52	7	1,722	7	687,683	9	4,585	1	1,000 only	66 consensus
SEQ1523	17,325	40	17,500	45	18,226	18	12,973	29	91,661	14	38,040	12	1,000 only	66 consensus
SEQ1704	6,349	9	1,084	7	219	8	2,086	7	1,098,300	8	2,510	3	1,000 only	66 consensus
SEQ2062	16,046	30	9,155	39	41,921	17	12,248	29	190,534	15	308	5	1,000 only	66 consensus
SEQ2122	6,679	10	1,557	16	733,725	11	5,946	10	60,648	6	400	4	1,000 only	66 consensus
SEQ0007	13,949	42	12,842	68	251,862	18	19,514	51	84,539	11	907	6	1,000/2,000	66 consensus
SEQ0051	5,885	22	4,390	32	43,791	6	6,711	26	342,753	8	1,042	3	1,000/2,000	66 consensus
SEQ0177	5,811	11	3,021	14	6,015	5	5,149	10	120,810	6	155	3	1,000/2,000	66 consensus
SEQ0219	6,542	15	6,258	24	389,747	10	2,805	12	8,276	6	350	2	1,000/2,000	66 consensus
SEQ0286	21,322	44	24,732	60	93,693	44	19,026	35	387,169	23	950	10	1,000/2,000	66 consensus
SEQ0362	13,581	20	7,715	31	378	16	3,652	10	27,550	10	520,996	5	1,000/2,000	66 consensus
SEQ0464	6,157	18	14,561	30	5,835	12	4,057	16	917	7	682,482	9	1,000/2,000	66 consensus
SEQ0580	56,130	118	45,639	145	670,423	63	55,198	102	35,326	35	136,017	24	1,000/2,000	66 consensus
SEQ0731	13,151	34	10,945	45	37,120	14	23,405	36	42,644	12	108,150	8	1,000/2,000	66 consensus
SEQ0740	3,750	14	5,859	22	86,413	11	7,891	12	704,929	22	3,501	7	1,000/2,000	66 consensus
SEQ0752a	119,752	300	149,437	398	266,292	126	172,652	278	1,075,626	114	600,757	85	1,000/2,000	66 consensus

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0787	20,908	66	22,208	86	181,134	28	39,823	59	121,429	28	77,980	18	1,000/2,000	66 consensus
SEQ0820	27,226	72	22,589	79	599,695	54	33,022	74	5,995	26	31,831	20	1,000/2,000	66 consensus
SEQ0871	13,690	15	10,030	33	96,339	21	8,406	16	177,389	9	218	7	1,000/2,000	66 consensus
SEQ0936	19,131	44	23,513	71	97,302	22	21,952	44	445,512	24	3,555	7	1,000/2,000	66 consensus
SEQ1116	75,973	135	73,759	172	353,041	56	62,747	135	514,138	81	95,839	37	1,000/2,000	66 consensus
SEQ1120	5,480	21	5,520	24	22,128	11	8,765	15	193,744	19	298	11	1,000/2,000	66 consensus
SEQ1139	4,168	18	5,011	22	8,515	6	3,400	17	448	3	349,019	8	1,000/2,000	66 consensus
SEQ1148	8,374	21	16,644	39	182	7	3,590	19	50,995	6	121,756	67	1,000/2,000	66 consensus
SEQ1154	3,155	15	3,993	12	217	6	2,308	11	15,054	4	242,308	16	1,000/2,000	66 consensus
SEQ1208	23,985	45	42,486	52	85,182	21	21,572	36	33,867	12	229,237	13	1,000/2,000	66 consensus
SEQ1223	18,151	42	13,399	51	13,050	19	12,235	42	457,757	18	2,972	12	1,000/2,000	66 consensus
SEQ1225	22,493	43	13,071	50	56,188	26	25,088	49	422,430	27	1,094	9	1,000/2,000	66 consensus
SEQ1265	6,743	28	14,292	44	69,350	20	13,945	33	222,850	13	316	7	1,000/2,000	66 consensus
SEQ1271	12,441	32	63,355	35	517,052	21	8,946	32	61,441	14	52,109	15	1,000/2,000	66 consensus
SEQ1284	23,799	47	34,207	61	148,969	28	25,644	48	994,408	22	1,286	17	1,000/2,000	66 consensus
SEQ1328	28,746	106	38,573	161	232,882	40	24,584	92	134,362	29	105,525	22	1,000/2,000	66 consensus
SEQ1571	26,621	83	26,749	116	6,939	36	28,225	73	138,182	29	373,051	26	1,000/2,000	66 consensus
SEQ1599	10,815	22	10,009	36	52,558	16	14,596	32	245	7	212,361	8	1,000/2,000	66 consensus
SEQ1637	11,805	13	3,055	12	339,313	14	8,053	16	35,068	10	293	6	1,000/2,000	66 consensus
SEQ1712	16,468	41	7,881	42	20,311	13	13,068	33	63,155	18	74,811	9	1,000/2,000	66 consensus
SEQ2022	8,872	14	4,794	19	10,888	9	7,166	18	1,170,588	26	6,025	5	1,000/2,000	66 consensus
SEQ2048	51,373	91	46,258	112	610,168	44	53,890	86	418,739	43	41,237	29	1,000/2,000	66 consensus
SEQ2090	9,062	28	4,895	39	33,726	15	8,476	26	633,909	15	1,486	4	1,000/2,000	66 consensus
SEQ2173	7,223	17	11,107	27	9,998	11	7,612	21	2,027	4	139,882	4	1,000/2,000	66 consensus
SEQ1269	12,009	37	9,240	40	18,903	24	10,980	37	158,252	8	568	7	1,000 only	barcoded only
SEQ2159	9,443	13	2,692	14	82,869	3	3,855	4	31,439	4	311	3	1,000 only	barcoded only

	Input libraries						Output libraries							
locus tag	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	AC reads	AC ins	CT reads	CT ins	GA reads	GA ins	stringency analysis	random pony groups analysis
SEQ0203	5,602	15	4,637	21	6,468	6	2,036	5	91,575	4	130	5	1,000/2,000	barcoded only
SEQ0222	29,760	36	17,223	49	9,962	10	10,841	25	394	9	384,820	19	1,000/2,000	barcoded only
SEQ0385	19,640	31	15,468	43	414	10	13,767	20	106,403	13	159,128	11	1,000/2,000	barcoded only
SEQ0547	7,973	15	7,637	26	8,116	10	3,565	14	4,248	6	58,549	9	1,000/2,000	barcoded only
SEQ0659	6,873	16	9,393	33	18,474	6	11,252	20	4,363	6	90,937	9	1,000/2,000	barcoded only
SEQ0765	12,296	29	10,196	36	56,395	15	17,760	21	688	12	130,963	7	1,000/2,000	barcoded only
SEQ1193	10,679	31	10,687	49	14,993	21	13,329	34	45,169	10	106,423	13	1,000/2,000	barcoded only
SEQ1194	5,127	26	10,144	30	44,859	11	5,179	19	24,179	6	20,074	8	1,000/2,000	barcoded only
SEQ1206	56,518	78	51,958	126	14,480	40	41,148	82	153,507	31	416,341	19	1,000/2,000	barcoded only
SEQ1231	15,974	42	11,956	46	7,298	14	11,745	43	29,807	12	84,538	7	1,000/2,000	barcoded only
SEQ1259	44,754	110	62,203	164	26,214	65	58,914	107	74,298	34	562,621	27	1,000/2,000	barcoded only
SEQ1261	79,211	188	92,373	230	109,776	80	52,332	157	80,201	53	647,865	43	1,000/2,000	barcoded only
SEQ1266	14,301	55	25,803	84	26,636	38	16,348	46	102,365	22	67,541	17	1,000/2,000	barcoded only
SEQ1281	41,607	70	35,303	82	189,918	36	25,845	57	12,895	26	155,871	13	1,000/2,000	barcoded only
SEQ1283	37,237	63	29,502	75	37,751	28	31,784	60	106,674	27	205,086	24	1,000/2,000	barcoded only
SEQ1505	31,686	65	29,315	99	1,546	40	42,855	68	279,784	36	285,433	22	1,000/2,000	barcoded only
SEQ1693	19,729	31	17,979	48	25,424	23	11,593	21	3,569	13	123,436	6	1,000/2,000	barcoded only