USING WHOLE GENOME SEQUENCING TO STUDY THE EPIDEMIOLOGY AND COMPARATIVE GENOMICS OF STREPTOCOCCUS EQUI FROM THE UNITED STATES

A Dissertation

by

ELLEN RUTH C. ALEXANDER

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Noah D. Cohen
Committee Members,	Michelle Coleman
	Ivan Ivanov
	Sara D. Lawhon
	Canaan Whitfield-Cargile
Head of Department,	Susan Eades

August 2021

Major Subject: Biomedical Sciences

ABSTRACT

Streptococcus equi subsp. *equi* (SEE) is the bacterium that causes the equine respiratory disease known as strangles. Strangles is endemic worldwide among horses. Despite its apparent prevalence and costs to equine agriculture, limited data exist regarding the molecular epidemiology of SEE from the United States (US). Thus, we conducted a series of genomic studies of SEE isolates from the US. First, we showed that mutations are rare in the genomes of SEE from an outbreak, and that some US isolates are closely related to SEE strains from other countries. Collectively, these data improved our understanding of phenotypic and genotypic variation of isolates within an outbreak, and the international distribution of SEE. Next, we compared the genomes and methylomes of US isolates of SEE with its multi-host ancestor *Streptococcus equi* subsp. *zooepidemicus* (SEZ) to identify a molecular basis for the host-specificity of SEE. We identified mobile genetic elements and methylation of genes that differed between SEE and SEZ, and are thus candidates for further investigation for their role in host-specificity of SEE. Because SEE does not survive in the environment for an extended period and has no known biological vectors, and because most horses develop prolonged immunity following recovery from disease, the persistence of strangles must be attributable to survival in horses that shed the bacterium without showing clinical signs (a.k.a. carrier horses). Thus, we examined the genomes of SEE isolates from carrier horses from the US and Europe. Whole genome sequencing of carrier and clinical SEE isolates from Pennsylvania and Sweden revealed neither significant nor consistent differences in the genomes or methylomes between carrier and clinical strains, and RNA sequencing of SEE isolates from Pennsylvania demonstrated no differentially expressed genes between clinical and carrier isolates of SEE. These results indicate that pathogen-adaptations of SEE are unlikely to explain the carrier state. Together, our findings indicate that genetic changes occur among isolates within outbreaks and within individual hosts, and that host factors are most likely to drive the carrier state. The host-specificity of SEE might have arisen from acquisition of mobile genetic elements or differential methylation of specific genes.

DEDICATION

I would like to dedicate this dissertation to my husband and my family: my mom, dad, sister, and brother. Thank you for your unending love and support while I pursued my dream.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Noah Cohen, and my committee members, Dr. Michelle Coleman, Dr. Ivan Ivanov, Dr. Sara Lawhon, and Dr. Canaan Whitfield, for their guidance and support throughout the course of this research.

Thank you, Dr. Noah Cohen for your encouragement, your patience, for sharing your knowledge, and for always believing in me. Without your belief, I would not be where I am today.

I would like to thank everyone who I had the pleasure of working with from the Equine Infectious Disease Laboratory, Dr. Angela Bordin, Mrs. Josy Bray, Mrs. Susie Kahn, Dr. Becky Legere, Dr. Bibiana Petri da Silveira, and Ms. Sophia Cortez and all the student workers, for their help with projects and for their friendship.

I would like to thank Drs. Susan Eades and Wesley Bissett for their support while Department Head of VLCS.

I would like to thank Drs. John Pringle, Miia Riihimäki, Ashley Boyle, Craig Carter, Erdal Erol, Sara Lawhon, Jing Wu, Amy Swinford, and Mrs. Sonia Lingsweiler for your generosity in sharing isolates and collaboration.

Thank you, Drs. Coleman, Ivanov, Lawhon, and Whitfield for helping with the design of the experiments.

Finally, thanks to my mother, father, and sister for their encouragement and understanding, and to my husband for his patience and love.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Noah Cohen (chair of committee) of the Department of Large Animal Clinical Sciences, and Dr. Michelle Coleman of the Department of Large Animal Clinical Sciences, Dr. Ivan Ivanov of the Department of Veterinary Physiology and Pharmacology, Dr. Sara Lawhon of the Department of Veterinary Pathobiology, and Dr. Canaan Whitfield-Cargile of the Department of Large Animal Clinical Sciences. Dr. Sara Lawhon of the Department of Veterinary Pathobiology, Dr. Amy Swinford and Mrs. Sonia Lingsweiler of the Texas A&M Veterinary Medical Diagnostic Laboratory, Drs. Craig Carter and Erdal Erol of the University of Kentucky Veterinary Diagnostic Laboratory, Drs. John Pringle and Miia Riihimäki of Department of the Clinical Sciences, Swedish University of Agricultural Sciences, and Dr. Ashley Boyle of the Department of Clinical Studies, School of Veterinary Medicine, University of Pennsylvania provided the *Streptococcus equi* isolates.

The data analyzed for Chapter 2 were conducted in part by Dr. Noah Cohen of the Department of Large Animal Clinical Sciences and were published in 2020 in Veterinary Microbiology (Morris ERA *et al.,* Comparison of whole genome sequences of *Streptococcus equi* subsp. *equi* from an outbreak in Texas with isolates from within the region, Kentucky, USA, and other countries. Vet Microbiol. 2020 Apr;243:108638. doi: 10.1016/j.vetmic.2020.108638.). The data analyzed in Chapter 4 were conducted in part by Dr. Noah Cohen of the Department of Large Animal Clinical Sciences, and Dr. Ivan Ivanov of the Department of Veterinary Physiology and Pharmacology and were submitted in 2021 to PLoS One. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by an assistantship from Texas A&M University, and by a grant from the Morris Animal Foundation (MAF) (https://www.morrisanimalfoundation.org/grants) and Grayson-Jockey Club Research Foundation (https://www.grayson-jockeyclub.org/). Additional support was provided by the Link Equine Research Endowment, Texas A&M University, a grant from USDA-NIFA through AgriLife Formula Animal Health funds, and grants from the Department of Large Animal Clinical Sciences, College of Veterinary Medicine & Biomedical Sciences, Texas A&M University. The funders had no role in study design, data collection and interpretations, or the decision to submit the work for publication.

NOMENCLATURE

SEE	*Streptococcus equi* subsp. *equi*
US	United States
SeM	M-like protein
ELISA	Enzyme-linked immunosorbent assay
IgG	Immunoglobulin G
PCR	Polymerase chain reaction
SEZ	*Streptococcus equi* subsp. *zooepidemicus*
MLST	Multilocus sequencing typing
SNP	Single-nucleotide polymorphism
VCF	Variance call format
MGE	Mobile genetic elements
S. pyogenes	*Streptococcus pyogenes*
ICE	Integrative conjugative element
CDS	Coding sequences
slaA	Phospholipase A_2
NRPS	Non-ribosomal peptide synthetase
bp	Base-pairs
TX	Texas
KY	Kentucky
CVM	College of Veterinary Medicine & Biomedical Sciences

IN	Intranasal
pbp2x	Penicillin-binding protein 2x
TSA	Tryptic soy agar
MIC	Minimum inhibitory concentration
SRA	Sequence Read Archive
S. pneumoniae	*Streptococcus pneumoniae*
AGEs	Accessory genome elements
m6A	N6-methyl-adenosine
m4c	N4-methyl-cytosine
m5c	C5-methyl-cytosine
SMRT	Single molecule, real-time
GO	Gene ontology
REBSE	Restriction Enzyme Database
IFN-γ	gamma interferon
RNA-Seq	RNA sequencing
PA-USA	Pennsylvania
GT	Genomic tips
TIGSS	Texas A&M Institute for Genome Sciences and Society
HPRC	Texas A&M High Performance Research Computing
FDR	False discovery rate
logFC	Log$_2$-fold change
GEO	Gene Expression Omnibus

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

CONTRIBUTORS AND FUNDING SOURCES...vi

NOMENCLATURE... viii

TABLE OF CONTENTS x

LIST OF FIGURES... xiii

LIST OF TABLES xvi

1. REVIEW OF *STREPTOCOCCUS EQUI* SUBSPECIES *EQUI* 1

1.1. Background and pathogenesis of strangles 1
1.2. Epidemiology of SEE and Strangles 3
1.3. Molecular epidemiology of *Streptococcus equi* 6
1.4. Comparison of SEE and SEZ 9
1.5. Limitations of current knowledge 12

2. COMPARISON OF WHOLE GENOME SEQUENCES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM AN OUTBREAK IN TEXAS WITH ISOLATES FROM WITHIN THE REGION, KENTUCKY, USA, AND OTHER COUNTRIES....14

2.1. Introduction 14
2.2. Materials and methods 15
2.2.1. *Streptococcus equi* subspecies *equi* isolates..15
2.2.2. Bacterial DNA extraction and sequencing 19
2.2.3. Assembly and computational analysis 20
2.2.4. Colony morphology...21
2.2.5. Susceptibility testing 22
2.2.6. Database accession numbers 23
2.3. Results 23
2.4. Discussion 30

3. DIFFERENCES IN THE ACCESSORY GENOMES AND METHYLOMES OF STRAINS OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* AND OF *STREPTOCOCCUS EQUI* SUBSP. *ZOOEPIDEMICUS* OBTAINED FROM THE RESPIRATORY TRACT OF HORSES FROM TEXAS 35

3.1. Introduction 35
3.2. Materials and methods 37
3.2.1. *Streptococcus equi* isolates 37
3.2.2. Bacterial DNA extraction and whole genome sequencing 38
3.2.3. Bioinformatic analysis 39
3.3. Results 41
3.4. Discussion 54

4. DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS 64

4.1. Introduction 64
4.2. Materials and methods 67
4.2.1. *Streptococcus equi* subsp. *equi* isolates 67
4.2.2. Bacterial DNA extraction and whole genome sequencing 69
4.2.3. Bacterial RNA extraction and RNA sequencing 71
4.2.4. Bioinformatic analysis 73
4.2.5. Accession numbers 75
4.3. Results 75
4.4. Discussion 84

5. SUMMARY AND FUTURE DIRECTIONS 91

5.1. Summary of results 91
5.2. Future directions 94
5.2.1. Molecular epidemiology of other regions of the US 94
5.2.2. Potential to improve SEE diagnostics 94
5.2.3. Functional differences in methylomes of SEE and SEZ 96
5.2.4. Carrier SEE and host-adaptions 97
5.3. Final remarks 98

REFERENCES 99

APPENDIX A SUPPLEMENTARY MATERIAL: COMPARISON OF WHOLE GENOME SEQUENCES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM AN OUTBREAK IN TEXAS WITH ISOLATES FROM WITHIN THE REGION, KENTUCKY, USA, AND OTHER COUNTRIES 109

APPENDIX B SUPPLEMENTARY MATERIAL: DIFFERENCES IN THE ACCESSORY GENOMES AND METHYLOMES OS STRAINS OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* AND OF *STREPTOCOCCUS EQUI* SUBSP. *ZOOEPIDEMICUS* OBTAINED FROM THE RESPIRATORY TRACT OF HORSES FROM TEXAS 124

APPENDIX C LINUX AND R CODE: DIFFERENCES IN THE ACCESSORY GENOMES AND METHYLOMES OS STRAINS OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* AND OF *STREPTOCOCCUS EQUI* SUBSP. *ZOOEPIDEMICUS* OBTAINED FROM THE RESPIRATORY TRACT OF HORSES FROM TEXAS ...141

APPENDIX D SUPPLEMENTARY FIGURES AND TABLES: DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS 148

APPENDIX E LINUX AND R CODE: DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS......173

APPENDIX F CLUSTAL OMEGA MULTIPLE SEQUENCE ALIGNMENT: DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS 179

LIST OF FIGURES

Page

Fig 1-1. Transmission and persistence of SEE. Persistence of strangles cannot be attributed to other mammalian hosts due to host-specificity, limited environmental persistence, biological vectors, or horses that have recovered from infection. Transmission of SEE from inapparent carrier horses to susceptible horses is the reason strangles persists and continues to infect horses world-wide. Created with BioRender.com 4

Fig 2-1. Phenotypic colony morphology of bacterial colonies.[80] Top row: Visual representation of bacterial colony form from the overhead perspective. Middle row: Visual representation of bacterial colony form regarding the elevation. Bottom row: Representation of the margins of bacterial colonies..22

Fig 2-2. Phylogenetic tree of SEE isolates from United States (Texas [TX] and Kentucky [KY]). Phylogenetic comparisons of 54 SEE isolates from Texas and Kentucky. Isolates collected from the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM) outbreak clustered together, along with the strains associated with the 2017 and 2018 vaccine project. 25

Fig 2-3. Colony morphology types of 54 SEE isolates from US. A) Colonies are observed to as circular, umbonate, entire, and white in color. B) Colonies were observed as circular, convex, entire, and salmon in color. C) Colonies found to have the penicillin-binding protein 2x SNP were circular, raised, entire, and white in color. 28

Fig 2-4. Phylogenetic tree of the 54 United States and 230 publicly available SEE isolates. The majority of SEE isolates cluster based on continent of origin, whereas some isolates from Kentucky and Texas are clustered with isolates originating from Europe. 29

Fig 3-1. Comparison of accessory genome elements (AGE) of SEE (n = 50) and SEZ (n = 50) genomes. The outer ring shows the ClustAGE bins that are ≥ 200 base-pairs in size these are ordered clockwise from the largest bin to the smallest bin, and are differentiated by orange and green to define bin borders. The concentric inner bands show the distribution of AGE within each individual isolate. Bands that are blue represents SEE isolates, and bands that are red represent SEZ isolates. The central ruler of the figure indicates the cumulative size of the AGE in kilobases...................................42

Fig 3-2. Gene ontology (GO) terms and KEGG pathways (annotated in ClueGO) in the accessory genome elements identified in all SEE (n = 50) genomes.

Circle size represents the degree of the positive relationship between the GO terms, and the term's adjusted P-value. The related terms are grouped and presented in the same color. 47

Fig 3-3. Gene ontology (GO) terms and KEGG pathways (annotated in ClueGO) in the accessory genome elements identified in all SEZ (n = 50) genomes. Circle size represents the degree of the positive relationship between the GO terms, and the term's adjusted P-value. The related terms are grouped and presented in the same color. 49

Fig 3-4. Gene ontology (GO) terms and KEGG pathways (annotated in ClueGO) on homologous proteins where methylation is present in SEE (n = 24) genomes, but absent in SEZ (n = 24) genomes. Circle size represents the degree of the positive relationship between the GO terms, and the term's adjusted P-value. The related terms are grouped and presented in the same color. 54

Fig 4-1. Comparison of accessory genome elements (AGE) of inapparent carrier SEE (n = 8), and acute clinical SEE (n = 6) genomes from Sweden. The outer ring shows the ClustAGE bins that are ≥ 200 base-pairs in size these are ordered clockwise from the largest bin to the smallest bin, and are differentiated by orange and green to define bin borders. The concentric inner bands show the distribution of AGE within each individual isolate. Bands that are blue represents inapparent carrier isolates, and bands that are red represent acute clinical isolates. The central ruler of the figure indicates the cumulative size of the AGE in kilobases. 77

Fig 4-2. Comparison of accessory genome elements (AGE) of inapparent carrier SEE (n = 11), and acute clinical SEE (n = 10) genomes from Pennsylvania. The outer ring shows the ClustAGE bins that are ≥ 200 base-pairs in size these are ordered clockwise from the largest bin to the smallest bin, and are differentiated by orange and green to define bin borders. The concentric inner bands show the distribution of AGE within each individual isolate. Bands that are blue represents inapparent carrier isolates, and bands that are red represent acute clinical isolates. The central ruler of the figure indicates the cumulative size of the AGE in kilobases. 78

Fig 4-3. Methylation locations and motifs from SEE isolates from Sweden. (A) Depiction of whether methylation occurred at a specified genomic location. Genomic locations are indicated along the x-axis, and whether methylation occurred is indicated on the y-axis as yes (Y) or no (N), by SEE isolates. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates. (B) Sites of methylation (x-axis), by the methylation motif (y-axis). The type of methylation and exact position in the genome are indicated by different colors. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates..80

Fig 4-4. Methylation locations and motifs from SEE isolates from Pennsylvania. (A) Depiction of whether methylation occurred at a specified genomic location. Genomic locations are indicated along the x-axis, and whether methylation occurred is indicated on the y-axis as yes (Y) or no (N) by SEE isolates. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates. (B) Sites of methylation (x-axis), by the methylation motif (y-axis). The type of methylation and exact position in the genome correspond to the different colors. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates..80

Fig 4-5. Volcano plot of Pennsylvania SEE RNA-Seq genes counts. The $\log_2$ fold-change (logFC) is represented along the x-axis, and the $\log_{10}$-transformed false discovery rate (FDR) is represented along the y-axis. Gray points represent genes that were not identified as significantly differentially expressed ($FDR \leq 0.05$), and green points represent genes whose expression had a $logFC \leq -1$ or ≥ 1. No genes met the criteria for interest of having an $FDR \leq 0.05$ and a $logFC \leq -1$ or ≥ 1...83

LIST OF TABLES

Page

Table 2-1. Description 22 SEE isolates sequence that were associated with the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM) outbreak. 17

Table 2-2. Description of 32 SEE isolates sequences from Texas (TX) and Kentucky (KY). 18

Table 2-3. SEE isolates with penicillin-binding protein 2x (*Pbp2x*) SNP. 27

Table 3-1. Accessory genome elements identified in all 50 SEE genomes. 43

Table 3-2. Accessory genome elements identified in all 50 SEZ genomes. 48

Table 3-3. Novel motif sequences from SEE (n = 24) and SEZ (n = 24) genomes. 50

Table 3-4. Methylation location, type and motif in 24 SEE genomes. 52

Table 4-1. Description of the 14 SEE isolates from Sweden and the 21 SEE isolates from Pennsylvania. 68

Table 4-2. The summary of the methylation motif sequences, modification types, and modification percentage for all study SEE isolates from Sweden and Pennsylvania. 81

1. REVIEW OF *STREPTOCOCCUS EQUI* SUBSPECIES *EQUI*

1.1. Background and pathogenesis of strangles

Streptococcus equi subspecies *equi* (SEE) is the causative agent of the equine infectious disease known as strangles. Strangles is one of the most commonly diagnosed infectious disease of horses world-wide.[1] A host-specific bacterium, it causes considerable economic losses for the equine industry in the United States (US) and abroad.[2,3] This ancient disease of the equine upper respiratory tract is characterized by the classic clinical signs of pyrexia, purulent nasal discharge, inflammation of the pharynx, and abscessation of the submandibular and retropharyngeal lymph nodes.[1,3] Strangles outbreaks occur commonly. It is a reportable disease in the United Kingdom where > 600 outbreaks have been reported in some years.[4] In the US, reporting of strangles is voluntary, and < 100 outbreaks were reported in 2020 according to the Equine Disease Communication Center (https://www.equinediseasecc.org/); however, many US outbreaks go unreported. This under-reporting of strangles in the US hampers efforts to understand the frequency and distribution of strangles in the US that might be used to improve methods for controlling this disease.

Following exposure to SEE, the bacterium attaches to cells within the lingual and palatine tonsils, and epithelium of the pharyngeal and tubal tonsils.[5] A few hours after infection the organism is found within the epithelial cells or tonsillar follicles, and then migrates to the submandibular and retropharyngeal lymph nodes. Abscessation of the lymph nodes is not usually noted until 3 to 5 days after infection.[6] Abscessed lymph nodes generally progress to rupture either externally or internally into the mouth, pharynx, or guttural pouches.[1] Rupture of a lymph node into a guttural pouch can result in guttural pouch empyema and purulent nasal discharge.[1] In addition, disseminated infection to other body systems and immune-mediated sequelae such as myositis or vasculitis can occur as a result of infection with SEE.[1,7-9] Virulence factors such as the M-like (SeM) protein, hyaluronic acid capsule, and the factor H binding protein Se18.9 contribute to its ability evade phagocytosis by neutrophils and cause disease.[6,10] The SeM protein is a factor that helps to prevent uptake and killing of SEE by neutrophils through the ability to bind fibrinogen.[11,12] Similarly, the hyaluronic acid capsule (encoded by the genes *hasA, hasB,* and *hasC*) also helps to facilitate evasion of phagocytosis.[13,14] The factor H binding protein Se18.9 is secreted by SEE to bind the complement protein, factor H, decreasing deposition of complement component 3 onto the surface of SEE, thereby protecting the SEE against complement-mediated phagocytosis or destruction.[15]

SEE can be shed from the nasal passage 2 to 3 days after fever onset, and infection can persist for 2 to 3 weeks in most animals.[1] Systemic and mucosal immune responses are commonly evident 2 to 3 weeks following infection and correspond with clearance of the SEE from the mucosa.[16] When antibiotics are not used, the majority (75%) of horses develop prolonged immunity to strangles; however, use of antibiotics such as penicillin can disrupt the development of immunity to strangles.[17,18] For example, Pringle *et al.* found that horses treated with penicillin within 11 day of onset of fever and with a treatment duration of 11 days (mean value) were significantly less likely to remain seropositive 4 and 7 weeks after diagnosis of the index case.[19] Additionally, foals with maternal antibodies, and vaccinated horses are usually less susceptible to infection.[1]

1.2. Epidemiology of SEE and Strangles

SEE is highly contagious, and is spread to susceptible horses of any age through direct and indirect transmission.[1] Direct transmission mainly occurs through nose-to-nose contact of horses. Indirect transmission happens via shared water sources, or by use of contaminated personnel or equipment such as buckets, halters, or brushes.[1] Most frequently these modes of transmission occur after exposure to the purulent discharge from active or recovering cases of strangles. A host-restricted pathogen, SEE is described as a poor colonizer,[20] and is rarely described as the source of infection of other mammalian species.[21-24]

Current knowledge indicates that SEE does not survive in the external environment for extended periods of time. SEE can persist approximately 2 days outside its host on wood, metal, or rubber.[25] Longer survival times of 1 to 4 weeks in a wet environment were dependent upon the season, with winter yielding an increased duration of survival.[26] There are no known biological vectors of SEE,[1] and horses that have recovered from the disease usually develop prolonged immunity.[1,27] Among weanlings diagnosed with strangles 6 months earlier (n = 12), only 2 developed strangles again within 10 days after a second exposure to SEE.[17] Immunity to strangles was further demonstrated by the finding that weanlings that did not develop strangles at either the initial exposure or the second exposure had significantly higher enzyme-linked immunosorbent assay (ELISA) values for mucosal SeM protein-specific immunoglobulin G (IgG) than the weanlings that developed strangles.[17] Similarly, horses with prior exposure to strangles were used as controls in a vaccine study, and following experimental challenge via comingling none of these control horses developed clinical disease.[18] Consequently, the most likely source of spread and persistence of SEE is apparently healthy horses that shed SEE undetected (so-called carrier horses);[1,28,29] these carriers transmit SEE to susceptible horses, perpetuating the disease in nature (Fig 1-1).[30,31]

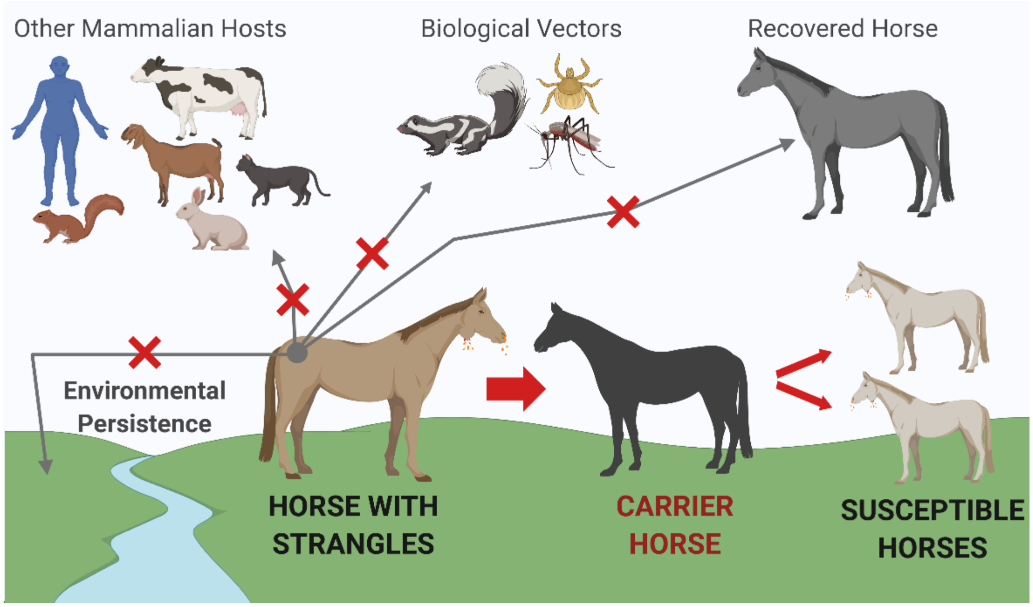


Fig 1-1. Transmission and persistence of SEE. Persistence of strangles cannot be attributed to other mammalian hosts due to host-specificity, limited environmental persistence, biological vectors, or horses that have recovered from infection. Transmission of SEE from inapparent carrier horses to susceptible horses is the reason strangles persists and continues to infect horses world-wide. Created with BioRender.com

The mechanism by which SEE persist in the host is poorly understood, and is the crux the persistence of this disease. After resolution of clinical signs, approximately 20% of horses will continue to shed SEE in nasal secretions for at least 4 weeks.[1,27,32] However, some horses that are inapparent carriers can shed the organism intermittently over extended periods of time but never exhibit the typical signs of strangles and appear outwardly healthy.[32] The carrier state is widely considered to be attributable to the presence of chondroids (*i.e.*, concretions of inspissated pus) or empyema in the guttural pouches.[28,30] Chondroids can remain in horses for years, and SEE can intermittently be shed resulting in the continued spread of infection.[30,31,33] Some carriers, however, have neither chondroids nor empyema yet shed SEE as identified by polymerase chain reaction (PCR) or culture and can transmit infection to health horses.[32,34]

Identification of inapparent carriers of SEE is challenging. Diagnosis of strangles is often reliant upon the observation of clinical signs, detection of SEE by culture or PCR, or based on known exposure.[1,31,35] Detection of horses with subclinical chondroids or empyema of the guttural pouches requires endoscopy.[1,27] Moreover, as noted above, not all SEE carriers have chondroids or empyema. Detection of carriers by diagnostic testing of lavage fluid of guttural pouches, nasopharyngeal lavage, or both is complicated by the finding that some carriers will be intermittently test positive for SEE by PCR or culture, and results can vary between samples (*i.e.*, guttural pouch lavage, nasopharyngeal lavage, or nasal swab).[32,35,36] Moreover, serological testing by ELISA is not reliable for detecting carrier horses.[35,37] Detection of SEE carrier horses is also hampered by poor patient histories and client compliance with diagnostic and biosecurity processes.

Because carrier horses appear to be critical to controlling strangles, further understanding the factors that drive the carrier state are crucial. Some evidence exists that carrier strains might differ from strains that cause clinical disease.[29,38,39] Truncation of the SeM protein has been suggested to contribute to the ability of SEE to remain in the host undetected.[38] Another factor that has been proposed to contribute to carriage of SEE in horses without clinical signs is its equibactin locus (*eqbA* to *eqbN*), a novel iron acquisition element present on ICE*Se2*.[29,39] More efficient iron acquisition is theorized to facilitate survival in the host and thus promote long-term carriage of SEE.[39] This evidence, however, is limited and excludes data from the U.S. Thus, there is great need to better determine whether strains that cause clinical disease differ from carrier strains. Molecular techniques have proven to be of great importance for understanding the epidemiology of infectious diseases.[40,41] Conceivably, molecular characterization of carrier and clinical strains of SEE could elucidate the role of adaptation of the bacterium to the host in establishing the SEE carrier state.

1.3. Molecular epidemiology of *Streptococcus equi*

The molecular epidemiology of SEE has been investigated. Genome sequencing indicates that SEE appears to have evolved from *Streptococcus equi* subsp. *zooepidemicus* (SEZ).[10,42,43] Whereas SEE is a host-specific pathogen, SEZ is a commensal of the upper respiratory tract of horses and is known to infect a variety of mammalian hosts, including humans.[13,44-48] Historically, differences in carbohydrate fermentation have been used to distinguish SEE from SEZ in clinical laboratories:[49] SEZ has the ability to ferment lactose and can inconsistently ferment sorbitol or ribose, whereas SEE does not have the capacity to ferment lactose, sorbitol, or ribose.[10,49,50] Inconsistencies of results of sugar fermentation for SEE and SEZ, and time required to use sugar fermentation profiles to differentiate SEE from SEZ led to the development of PCR tests based on differences in genes between SEE and SEZ.[31,50,51] Accurate diagnosis of infection with SEE is crucial for understanding the epidemiology and clinical spectrum of strangles, and evidence exists that there are limitations to existing PCR tests for SEE.[52]

Multilocus sequencing typing (MLST) is a method used to genotype bacteria, including SEE.[43,53] The MLST scheme for SEZ is a public database that uses MLST to generate sequence types (STs) for isolates of SEE and SEZ.[43,53] Over 400 STs have been documented in the MLST scheme for SEZ[43] (https://pubmlst.org/organisms/streptococcus-zooepidemicus; accessed Feb. 6, 2021) the majority of which are attributed to SEZ genomes, whereas only a few STs are reported for SEE strains.[29] Another molecular epidemiological method to characterize SEE is through targeted sequencing a portion of the SeM protein.[54] Sequencing the N-terminal region of the SeM protein (also known as the variable region) has been used to differentiate strains and trace the source of an outbreak.[55-58] In China, a novel SeM type (SeM 136) that caused a multi-farm outbreak of strangles in donkeys was identified using this PCR-based approach.[57]

The next generation sequencing (NGS) technology of whole genome sequencing (WGS) is a more powerful tool than genotyping methods such as MLST or SeM typing to study the molecular epidemiology of bacteria.[41] The genomes of SEE that are publicly-available[29] (n = 244) have been predominately from European cases and outbreaks; only a few genomes from the US (including the genomes of the live, attenuated vaccine strain licensed in the US known as Pinnacle IN™) and other countries such as Australia and Saudi Arabia were publicly available (prior to the work reported in this dissertation). Illumina short-read, WGS of this population of 224 isolates revealed a low sequence diversity with only 3,109 sites of single-nucleotide polymorphisms (SNPs), or insertion and deletions that differed from the reference genome, SEE 4047 (2.5 million base-pairs).[10,29] The majority of these sites were in the mobile genetic elements (MGEs) of the SEE genome. Four different clusters of SEE were derived based on a Bayesian method[29,59] for dividing populations based on sequence similarity. This small number of clusters further demonstrates the low diversity of the SEE isolates. Although these genomes of SEE are described as having low overall diversity, variations within SEE isolates are nevertheless reported. For example, the *has* operon (*hasA, hasB,* and *hasC*), encoding for the hyaluronic acid capsule has been identified as the region having the most deletions or duplications in the SEE genome.[29] Moreover, carrier isolates of SEE were found to have varying deletions in the equibactin locus (*eqbA* to *eqbN*), indicating that this locus is not necessary for the inapparent carrier state.[29] Details such as these mutations within these SEE genomes would not have been characterized without the use of WGS, thereby demonstrating the importance and power of using this tool for molecular epidemiological investigations.

Beyond the aforementioned study in which only 3 US isolates of SEE were represented, not much is known about the molecular epidemiology of SEE strains from the US. Because of the magnitude and influence of the US horse industry and the frequent international transportation of horses associated with equine breeding and competition, there is critical need to better understand the molecular epidemiology of US isolates of SEE and how they relate to SEE from other countries. The goal of our work in Chapter 2 of this dissertation was to help fill this knowledge-gap. In addition to explaining spread and transmission of SEE, molecular genetic studies can also help shed light on the evolution of SEE. Importantly, US isolates of SEE are also poorly represented in molecular studies of the evolution of SEE.

1.4. Comparison of SEE and SEZ

Through a reduction in genetic variability described as an evolutionary bottleneck, SEE is thought to have evolved from SEZ.[10] SEZ is a common opportunistic pathogen of many mammalian hosts, including humans and horses, and is often recovered from the upper respiratory tract of horses as a commensal.[13,44-48,60] A number of targeted approaches have been used to characterize genetic differences between SEE and SEZ. Of note, differences have been described in the superantigens of SEE and SEZ. The superantigens *seeH*, *seeI*, *seeL*, and *seeM* of SEE have been described to share > 96% homology with the superantigens SpeL, SpeM, SpeH, and SpeI of *Streptococcus pyogenes* (*S. pyogenes*).[10,61] Interestingly, *seeL* and *seeM* have been detected in 4 of 140

SEZ isolates (ST 120).[10,61] Alternative novel superantigens (*szeF*, *szeN*, *szeP*) identified in SEZ strains share lower levels of homology (34% to 59%) with SpeH, SpeM, and SpeL of *S. pyogenes*.[62] At least 1 of these novel SEZ superantigens was identified in approximately half (49%; 81/165) SEZ isolates screened.[62] Despite SEE and SEZ having evolved to express different superantigen proteins the superantigens of both SEE and SEZ serve similar functions to stimulate gamma interferon production and proliferation of equine peripheral blood mononuclear cells.[61,62] Both SEE and SEZ produce proteins that bind fibronectin.[63] In SEZ, the fibronectin binding protein FNZ (encoded by the *fnz* gene) is anchored to the surface of the bacterium.[10,63] In contrast, the fibronectin-binding protein FNE of SEE is secreted from the bacterium because a conserved base-pair deletion in the *fne* gene leads to the loss of a surface anchor.[10]

Prior to the work reported in this dissertation, WGS had been used only to compare a single SEE strain (4047) with a single SEZ strain (H70).[10] To validate major differences between these individual strains of SEE and SEZ identified by WGS, real-time PCR of targeted genes was performed using 26 SEE strains and 140 SEZ strains.[10] Ninety-five (95) STs were represented among the 140 SEZ isolates included. The SEE strain 4047 was found to have 4 prophages (φSeq1 – φSeq4) and 2 integrative conjugative element (ICE; ICE*Se1*, ICE*Se2*) regions as MGEs making up 16% of the total genome, whereas SEZ H70 had only 2 ICE (ICE*Sz1*, ICE*Sz2*) making up 7% of the genome. Among the predicted coding sequences (CDS) in both genomes, 1,671 CDS were found to have orthologs in both strains of SEE and SEZ. The number of functional classes of these CDS was similar for both SEE and SEZ, except that SEE had greater numbers of CDS with the functional classes identified as protective responses or adaptions and laterally-acquired elements.[10] The association of SEE with laterally-acquired elements was not surprising considering the increased proportion of MGEs identified in SEE 4047.

The study comparing SEE and SEZ[10] also demonstrated a homolog of the gene encoding for phospholipase A_2 (*slaA*, a virulence factor of *S. pyogene*s[64]) in all 26 strains of SEE on φSeq2 but only in a minority of SEZ isolates (44 of 140). However, a second putative phospholipase A_2 toxin, *slaB* was determined to be present in all SEE and SEZ strains. Not surprisingly, the genes *lacE*, *sorD*, and *rbsD* were deleted from SEE, resulting in the inability to ferment lactose, sorbitol, or ribose.[10] Sixteen of the 140 SEZ isolates were unable to ferment ribose or sorbitol.[10] The hyaluronic acid capsule is thicker in SEE than in SEZ.[10] This increase in hyaluronic capsules is likely the result of a 4 base-pair deletion in SEQ_1479 in SEE, and a second copy of the gene (SEQ_2045) acquired on a prophage which yields reduced hyaluronate lyase activity compared to SEZ. Finally, the equibactin locus comprised of 14 CDS (*eqbA* to *eqbN*) located on ICE*Se2* in all SEE was not identified in any of the 140 SEZ isolates.[10] The equibactin locus, a novel non-ribosomal peptide synthetase (NRPS) system is described to be a yersiniabactin-like NRPS system. Yersiniabactin is a ferric iron siderophore of *Yersinia* species,[65] and similarly equibactin has been described to aid in iron acquisition for SEE.[39] In summary, the combination of targeted (real-time PCR of 26 SEE and 140 SEZ) and untargeted (WGS of 1 strain of SEE and 1 strain of SEZ) comparison of SEE and SEZ provided important insights into the evolution and host-restriction of SEE.

However, WGS of only 2 isolates and including strains almost exclusively from Europe were limitations of this seminal work.

1.5. Limitations of current knowledge

Review of current understanding of the molecular epidemiology of strangles reveals a great under-representation of data from the US, despite the major impact of the US equine industry and the apparent prevalence of strangles in the US.[10,29,66] Thus, we included strains of a local outbreak of strangles to understand the dynamics of SEE strain variation within an outbreak, as well as to contrast the genomes of a convenience sample of isolates of SEE from different regions of the US (Chapter 2). Understanding why SEE is host-specific while its close relative SEZ has a more promiscuous host range will help to better understand pathogenesis and epidemiology of SEE. To date, comparison of the whole genomes of only a single isolate each of SEE and SEZ have been reported.[10] Thus, we sought to compare the genomes of a larger number of SEE and SEZ from the US, including both disease-associated and commensal strains of SEE (Chapter 3). Finally, the genomes of carrier strains of SEE isolates are poorly characterized.[29,32] This is important because it is unclear if the SEE carrier state in horses is predominately driven by host responses to the pathogen, adaptions of the pathogen to the host, or both. In an attempt to clarify the relative role of the pathogen, we compared the genomes, methylomes, and transcriptomes of carrier and clinical strains of SEE (Chapter 4). Collectively, the work comprising this dissertation sheds important light on the molecular epidemiology and pathogenesis of SEE and identifies areas in need of further investigation. This work also provided essential training to prepare the student for a career in bioinformatics and computational biology with an emphasis on microbial genomics.

2. COMPARISON OF WHOLE GENOME SEQUENCES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM AN OUTBREAK IN TEXAS WITH ISOLATES FROM WITHIN THE REGION, KENTUCKY, USA, AND OTHER COUNTRIES[*]

2.1. Introduction

Strangles is caused by the equine-specific bacterium, *Streptococcus equi* subspecies *equi* (SEE).[1,20,29] Strangles is highly contagious and remains one of the most commonly diagnosed infectious diseases in horses worldwide. Outbreaks result in a large financial burden to the equine industry and horse-owners and raise concerns for the health and welfare of horses.[1] Consequently, efforts to improve diagnosis and prevention are of paramount importance to the equine industry. Epidemiological questions such as tracing a foodborne illness or an infectious disease outbreak are greatly aided by application of molecular biological methods.[67] Increasingly, next generation sequencing (NGS) technologies for whole microbial genome sequencing have become an affordable strategy for molecular epidemiological investigations.[68] Whole genome sequencing (WGS) using Illumina® generates short-reads of DNA sequence (< 300 base-pairs [bp]) yielding in-draft bacterial genomes that can be used to characterize the genetic composition of a large number of bacterial isolates. Results of WGS can aid in control and prevention of outbreaks by expediting understanding of the dynamics and

[*] Reprinted with permission from "Comparison of whole genome sequences of *Streptococcus equi* subsp. *equi* from an outbreak in Texas with isolates from within the region, Kentucky, USA, and other countries" by Morris, E.R.A., Hillhouse, A.E., Konganti, K., Wu, J., Lawhon, S.D., Bordin, A.I., Cohen, N.D., 2020. *Vet Microbiol*, 243, 108638. https://doi.org/10.1016/j.vetmic.2020.108638, Copyright 2020 by Elsevier B.V.

dissemination of infections. To the authors' knowledge, there are limited publicly-available data regarding the sequences of SEE isolates recovered in the United States (US), whereas there is a robust array of isolates from other countries on other continents. Thus, the primary objective of this study was to utilize WGS to compare the genomes of SEE isolates from an outbreak of strangles in a herd of horses used for teaching and research in Texas (TX) with isolates from other regions of TX and central Kentucky (KY) and with publicly-available sequences of SEE strains from other continents.

2.2. Materials and methods

2.2.1. *Streptococcus equi* subspecies *equi* isolates

A total of 54 SEE isolates from the US were selected for sequencing. Sixteen SEE isolates were collected from an outbreak that occurred during late 2017 to early 2018 among horses in the teaching herd at the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM). This outbreak occurred approximately 5 months after the conclusion of the first phase of a SEE vaccine trial that entailed infecting a different group of yearling horses with SEE intranasally (IN). Isolates from the 2017 SEE vaccine study with IN infection of individual horses (n = 2), isolates from a subsequent 2018 SEE vaccine study using infection by direct contact with horses that developed strangles during the CVM outbreak (n = 4) were included for sequencing (Table 2-1). To characterize regional differences, isolates from the same county (Brazos County) of TX as Texas A&M University (n = 5), isolates from individual horses from other regions of TX collected in 2014 (n = 8), isolates from a 2011 outbreak at a ranch in north TX (n = 8), and isolates from central KY (n = 9) also were sequenced. The strain used for IN infection in 2017 was from the 2011 outbreak at the ranch in north TX. The remaining SEE strains (n = 2) sequenced were included for sequencing quality control, and comparative reference to the Zoetis Pinnacle® vaccine (Table 2-2). Additionally, publicly available SEE genomes that were sequenced from horses around the world were retrieved from PATRIC,[69] including isolates from Europe (n = 217), Asia (n = 2), Australia (n = 2), and North America (n = 9; A-1 Table). We also included the sequence from the reference strain SEE ATCC 39506 (SEE 39506) obtained from NCBI GenBank.

Table 2-1. Description 22 SEE isolates sequence that were associated with the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM) outbreak.

Isolate ID	State	Outbreak	Subclinical	SeM Type
17-007	TX	2017 Strangles Project	N	39
17-008	TX	2017 Strangles Project	N	39
17-003	TX	CVM Outbreak	N	39
17-004	TX	CVM Outbreak	N	NA
18-001	TX	CVM Outbreak	N	39
18-002	TX	CVM Outbreak	Y	39
18-003	TX	CVM Outbreak	Y	39
18-004	TX	CVM Outbreak	Y	39
18-006	TX	CVM Outbreak	N	39
18-011	TX	CVM Outbreak	N	39
18-012	TX	CVM Outbreak	Y	39
18-013	TX	CVM Outbreak	Y	39
18-014	TX	CVM Outbreak	N	39
18-015	TX	CVM Outbreak	Y	39
18-018	TX	CVM Outbreak	N	39
18-021	TX	CVM Outbreak	N	39
18-022	TX	CVM Outbreak	N	39
18-024	TX	CVM Outbreak	Y	39
18-037	TX	2018 Strangles Project	Y	39
18-039	TX	2018 Strangles Project	Y	39
18-078	TX	2018 Strangles Project	Y	39
18-079	TX	2018 Strangles Project	Y	39

Table 2-2. Description of 32 SEE isolates sequences from Texas (TX) and Kentucky (KY).

Isolate ID	State	Outbreak	Subclinical	SeM Type
11-002	TX	North TX Outbreak	N	39
11-004	TX	North TX Outbreak	N	39
11-006	TX	North TX Outbreak	N	39
11-008	TX	North TX Outbreak	N	39
11-010	TX	North TX Outbreak	N	39
11-014	TX	North TX Outbreak	N	39
11-017	TX	North TX Outbreak	N	39
11-018	TX	North TX Outbreak	N	39
14-052	KY	2014 KY	N	148
14-057	KY	2014 KY	N	2
14-061	KY	2014 KY	N	2
14-066	KY	2014 KY	N	20
14-071	KY	2014 KY	N	20
14-073	KY	2014 KY	N	20
14-080	KY	2014 KY	N	20
14-082	KY	2014 KY	N	20
14-092	KY	2014 KY	N	28
14-105	TX	2014 TX	N	2
14-112	TX	2014 TX	N	39
14-125	TX	2014 TX	N	55
14-133	TX	2014 TX	N	2
14-140	TX	2014 TX	N	55
14-146	TX	2014 TX	N	55
14-148	TX	2014 TX	N	28
14-150	TX	2014 TX	N	157
17-009	TX	Brazos County	N	28
18-008	TX	Brazos County	N	39
18-009	TX	Brazos County	N	28
18-025	USA	Pinnacle vaccine	N	2
18-026	TX	Brazos County	N	2
18-027	TX	Brazos County	N	2
18-028	USA	Quality control	N	39

2.2.2. Bacterial DNA extraction and sequencing

SEE isolates from frozen stocks were grown overnight in the incubator in duplicates in 5 ml of Todd Hewitt (HIMEDIA®, Mumbai, India) broth at 37°C and 5% CO_2. Isolates were centrifuged twice at 3,000 g for 5 minutes with 1X PBS and then resuspended in 250 µl 1X PBS (LONZA, Basel, Switzerland) and stored at -20°C until DNA extraction. DNA was extracted using the Macherey-Nagel NucleoMag Tissue extraction kit (Düren, Germany), following manufacturer instructions. Library preparation and WGS were performed at the Texas A&M Institute for Genome Sciences and Society (TIGSS) molecular genomics core laboratory. Briefly, DNA was quantified using the Qubit fluorometric dsDNA (Thermo Fisher Scientific, Waltham, MA, USA) assay for normalization before library preparation. Libraries were prepared using NextFlex Rapid DNA kit (Bioo Scientific, Austin, Texas, US) following manufacturer instructions and each isolate was identified with a unique 12-bp barcode. For verification of library preparation, the Agilent Tapestation (Santa Clara, CA, US) was used with D1000 tape, and quantification with the Qubit fluorometric dsDNA for the normalization of the concentration. Samples were then pooled, and sequencing was performed using the Illumina MiSeq (San Diego, CA, US) at the TIGSS core laboratory. The library pool was run twice with the MiSeq v3, 300 × 300-bp paired-end sequence run. Each sequencing run yielded approximately about 25 million sequencing reads, and each sample had a range of 30X to 100X sequencing coverage for each individual sequencing run.

2.2.3. Assembly and computational analysis

Following sequencing, computational analysis was completed using the Texas A&M High Performance Research Computing cluster. Sequence quality was verified using FastQC (v0.11.6; www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequences were then filtered and trimmed using Trimmomatic (v0.36)[70] with the parameters of removing the first 10 bases, using a 5-bp slide-window and trimming when the average quality score was below 20 and with removal of bases at the end of the read that were below a quality score of 25. Trimmed sequences were assembled *de novo* using SPAdes (v3.11.1).[71] Assembled genomes were aligned by the core genome and a core genome single nucleotide polymorphism (SNP) phylogenetic tree was built using ParSnp (v1.2).[72] The ParSnp output was viewed with gingr (v1.2), and HarvestTools (v1.2) was used to create a variant call format (VCF) file. Phylogenetic tree outputs from ParSnp were viewed and edited using Microreact (v5.123.1).[73] The VCF file outputs were a binary matrix of variants from the SEE genomes relative to the reference genome. Percentage of core genome variance was determined by adding up the total number of variants in the VCF file binary matrix for each isolates, and subsequently those totals were divided by the length of the ParSnp defined core genome for each isolates using R (v3.5.2),[74] and graphs were generated with ggplot2 (v3.1.0).[75]

Additionally, we performed local alignment on the SPAdes assembled genomes using BWA (v0.7.17),[76] against reference SEE 39506. These genomes were converted to a Bam file format with BWA. All Bam files were sorted by genome position using SAMtools (v1.8).[77] Sorted Bam files of all genomes were then combined using

BCFtools (v1.8)[78] mpileup and call functions to create a VCF file, using SEE 39506 as the reference. SnpEff (v4.3T)[79] on web-based platform Galaxy[80] was used to annotate and quantify variants from outputs using default settings. Prior to annotation, a SnpEff database was built using the build function for SEE 39506 with the required Fasta and Genbank files. Graphs were generated utilizing the SnpEff output show the frequency of transition and transversion mutations which were visualized using ggplot2. All 54 US SEE isolates were checked for M protein (SeM) identification using the PubMLST *Streptococcus equi* subsp. *zooepidemicus* (SEZ) database (accessed Dec. 18, 2018).[54]

2.2.4. Colony morphology

The association between colony morphology and the presence of the SNP found in penicillin-binding protein 2x (*pbp2x*; SE071780_01907), which has been associated with the cell wall and cell division, was examined. All 54 SEE isolates from the study were plated for isolation onto tryptic soy agar (TSA) plates with 5% sheep blood (Hardy Diagnostics, Santa Maria, California, US) to observe the colony morphology and color. Bacterial colonies were evaluated on the basis of their color, form, elevation, and margin (Fig 2-1).[81]

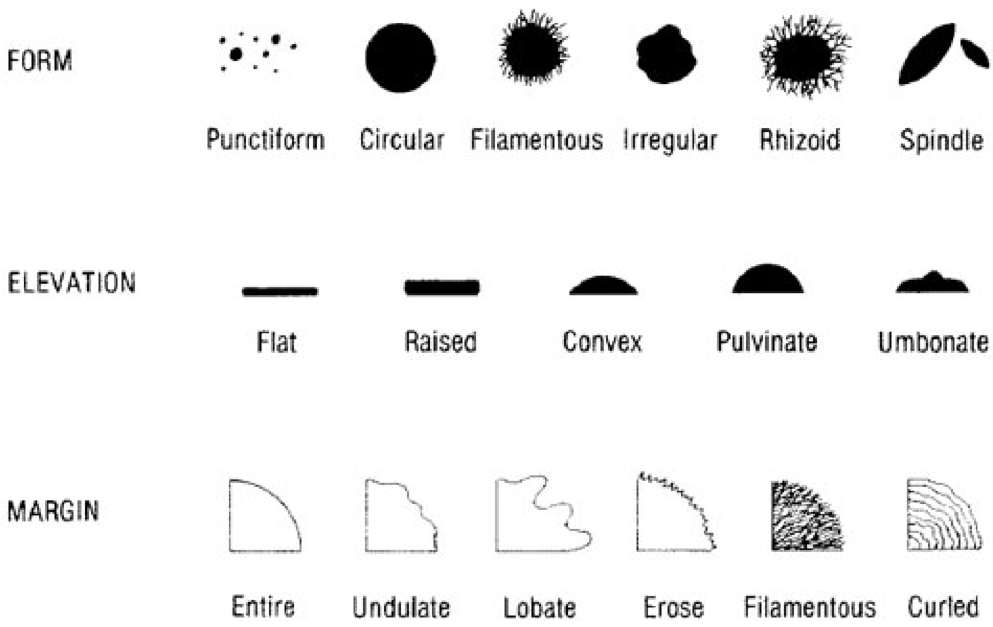


Fig 2-1. Phenotypic colony morphology of bacterial colonies.[80] Top row: Visual representation of bacterial colony form from the overhead perspective. Middle row: Visual representation of bacterial colony form regarding the elevation. Bottom row: Representation of the margins of bacterial colonies.

2.2.5. Susceptibility testing

Minimum inhibitory concentration (MIC) tests were performed with the penicillin E®-test strip (bioMérieux, Marcy-l'Étoile, France) to evaluate the effects of the SNP found in *pbp2x* on penicillin resistance using several (n = 6) SEE isolates from the CVM outbreak. Isolates were struck for isolation on TSA plates with 5% sheep blood, and then submitted to the CVM Clinical Microbiology Laboratory for MIC testing using E-strips, following the manufacturer’s instructions. Briefly, colonies selected from an overnight plate are placed in to sterile saline until a McFarland standard of 0.5 was reached. Using a cotton swab, a Mueller-Hinton agar plate with 5% sheep blood was completely covered with the 0.5 McFarland turbid solution. An E-test strip was placed onto the blood agar plate, and incubated for 24 hours at 37°C to allow SEE lawn growth. Following the 24-hour incubation, the MIC level was determined by viewing the lowest penicillin concentration on the E-test strip where the SEE growth was observed to have been inhibited.

2.2.6. Database accession numbers

All of the Fastq data from this WGS project were submitted to the NCBI Sequence Read Archive (SRA) and are accessible through the SRA with submission number SUB6350545. Assembled genomes for each isolate were submitted to GenBank, submission number SUB6350566. Individual sequence and assembled genomes accession numbers are found in A-3 Table.

2.3. Results

The 54 SEE isolates collected from the CVM outbreak (A-1 File), horses in the SEE vaccine contact-challenge study, and a convenience sample of isolates from KY and other regions of TX (Table 2-1) generated draft bacterial genomes comprised of an average of 178 contigs (range, 131 to 565 contigs). SeM protein identification of the isolates from the CVM outbreak, a north TX outbreak, and the vaccine study were identified as type 39. Other isolates were identified with SeM types 2, 20, 28, 57, and 2 newly described variants, submitted to the pubMLST SEZ database with identified as SeM numbers 148 and 147 (Table 2-1). Of the 2 new variants, SeM 148 from strain 14-052 was similar to SeM 43, but differed in 3 bp at positions 45 (A → T), 206 (T → C), and 318 (T→ G), whereas, SeM 147 from strain 14-150 was similar to SeM 137, but differed by 1 bp at position 318 (T → G).

Phylogenetic comparisons of the 54 SEE isolated revealed a high degree of similarity between the infection strain (11-017), isolates from the north TX ranch from which 11-017 was derived, isolates from 2017 from the SEE vaccine studies, and isolates collected from horses that were a part of the 2017/ 2018 CVM outbreak (Fig 2-2). Isolates from the north Texas ranch were closely clustered with isolates from the 2017 SEE vaccine study, along with the isolate from the index case (17-004) from the CVM outbreak (Fig 2-2). Isolates from KY and from TX that were not associated with either the outbreaks at the CVM or the north TX ranch were grouped separately from the outbreak strains with the following exceptions: 1) an isolate collected in the same county (Brazos, TX) as the CVM outbreak (18-008); and, 2) 2 isolates from TX collected in 2014 (14-112, 14-146). Not all of the SEE isolates were phylogenetically grouped by location of origin. A clinical case from Brazos County, TX attributed to vaccination with the Pinnacle® (18-027) was closely clustered with the Pinnacle® vaccine strain (18-025), as well as 3 other SEE strains from KY and TX collected in 2014 (14-061, 14-105, and 14-133) that had not been identified as suspected to be attributable to the Pinnacle vaccine.

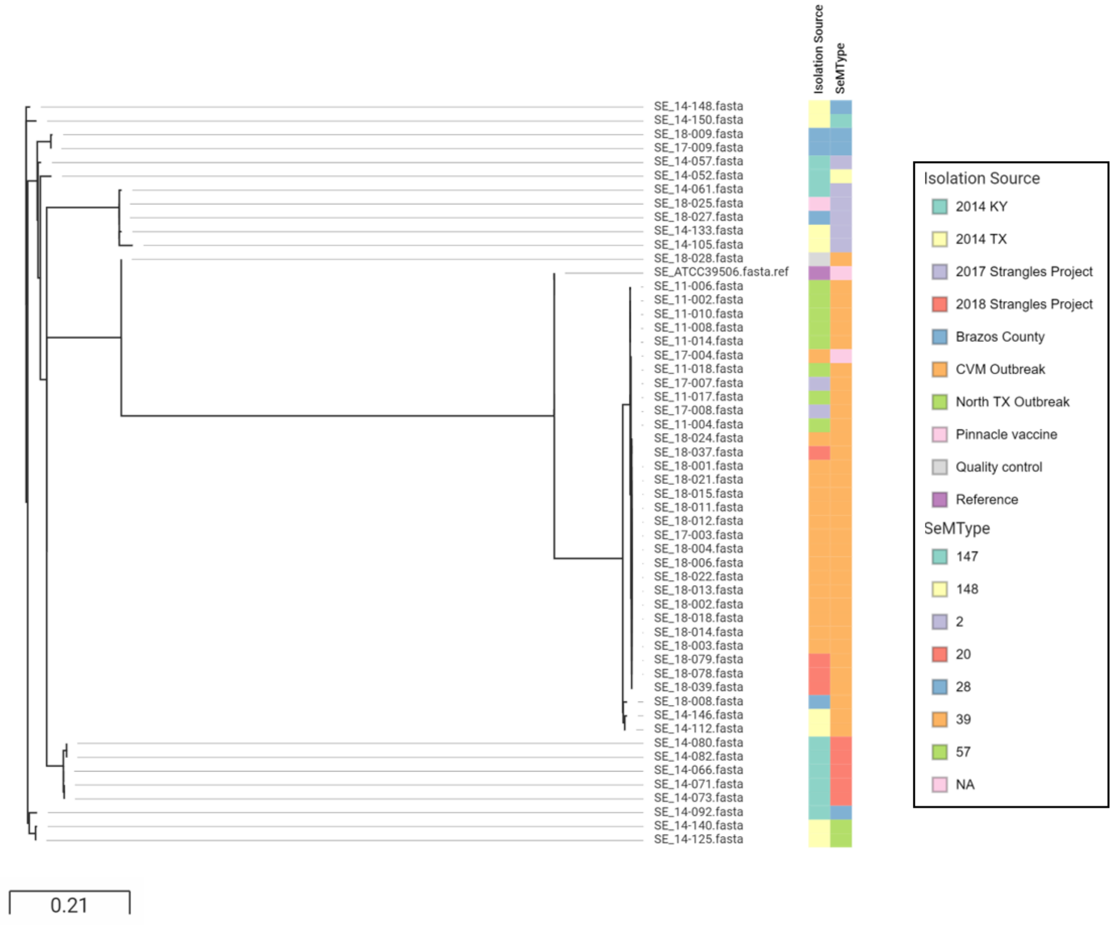


Fig 2-2. Phylogenetic tree of SEE isolates from United States (Texas [TX] and Kentucky [KY]). Phylogenetic comparisons of 54 SEE isolates from Texas and Kentucky. Isolates collected from the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM) outbreak clustered together, along with the strains associated with the 2017 and 2018 vaccine project.

Using SEE 39506 as a reference strain, isolates from the US (KY and TX) had an average core genome variation of 0.0167% (range, 0.0043% to 0.0265%) relative to the reference genome, whereas the variation seen among the isolates from the CVM outbreak was on average 0.0046% (A-1 Fig). A SNP was identified as a variant among the CVM outbreak isolates. This SNP was in the *pbp2x* (SE071780_01907), and it was predicted to result in an amino acid change at position 591 changing from a valine to an alanine (Table 2-3). This SNP occurred in all of the CVM outbreak isolates with the exception of 17-004. Notably, 17-004 was the index case for the CVM outbreak. In contrast, the infection strain (strain 11-017) used for the challenge of individual horses, and other isolates collected from individually challenged horses (strains 17-007, 17-008) lacked the *pbp2x* SNP. Considering the entire genomes of 54 SEE from the US, there were more transition mutations than transversion mutations relative to the reference genome (A-2 Fig). Genomes that were more highly related had similar numbers of these mutations, as demonstrated by those SEE strains that originated from the same outbreak. Isolates from KY or parts of TX not associated with the outbreaks had a greater number of these mutations.

Table 2-3. SEE isolates with penicillin-binding protein 2x (*Pbp2x*) SNP.

Isolate ID	***Pbp2x* SNP**
17-003	Y
18-001	Y
18-002	Y
18-003	Y
18-004	Y
18-006	Y
18-011	Y
18-012	Y
18-013	Y
18-014	Y
18-015	Y
18-018	Y
18-021	Y
18-022	Y
18-024	Y
18-037	Y
18-039	Y
18-078	Y
18-079	Y

Colony morphology of several of the SEE strains from the CVM outbreak appeared to be influenced by the presence of the SNP in *pbp2x*. Isolates with the SNP (A-2 Table) had a raised elevation structure with a white coloring, while those without the SNP were either umbonate in elevation structure with off-white coloring or had a convex structure with salmon coloring that was more mucoid in appearance (Fig 2-3).[81] MIC testing of penicillin with the E-test strip of representative isolates of SEE (n = 6) yielded no evidence of resistance to penicillin with the presence of the *pbp2x*SNP.

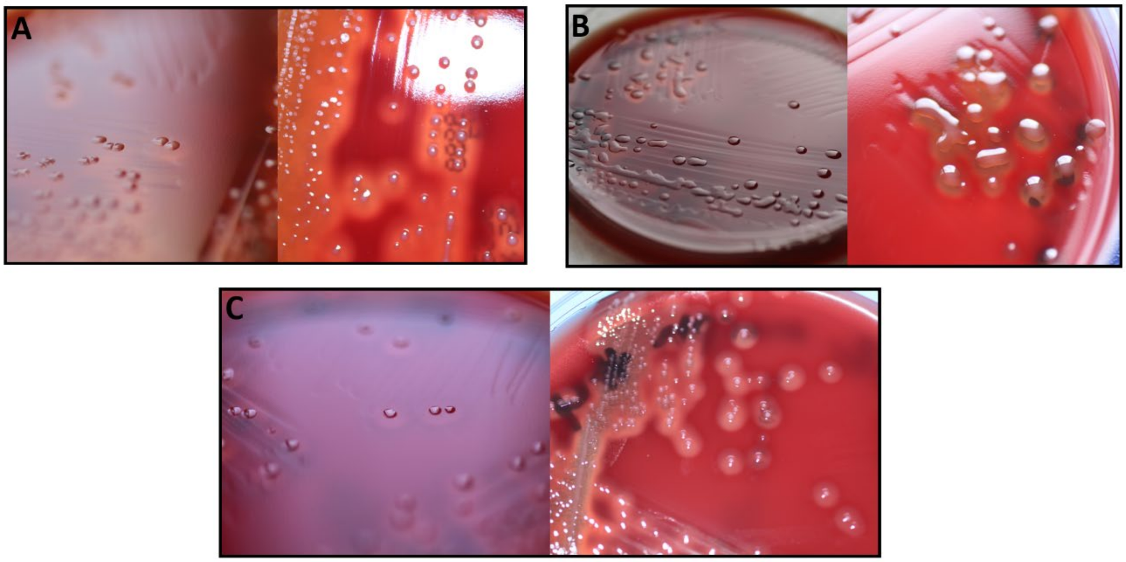


Fig 2-3. Colony morphology types of 54 SEE isolates from US. A) Colonies are observed to as circular, umbonate, entire, and white in color. B) Colonies were observed as circular, convex, entire, and salmon in color. C) Colonies found to have the penicillin-binding protein 2x SNP were circular, raised, entire, and white in color.

A phylogenetic representation was utilized to compare the 54 TX and KY isolates and publicly-available genome sequences from isolates in Europe (n=217), Asia (n=2), Australia (n=2), and other parts of North America (n = 9). The phylogenetic tree revealed large clusters of isolates from Europe, a central cluster comprising isolates from multiple countries and continents, and a grouping of the SEE isolates from the CVM outbreak, the north TX outbreak, and our vaccine challenge study (Fig 2-4). Variation of all 284 SEE isolates yielded an average core genome variation percentage of 0.0117% (range, 0.0008% to 0.0513%) relative to the reference genome, and the isolates from the CVM and north TX outbreak had an average variation of 0.0009% (A-3 Fig). Mutations resulting from changes in bp's observed among the entire genomes 284 SEE isolates relative to the reference genome were most frequently transition mutations, with transversion mutations occurring much less frequently (A-4 Fig).

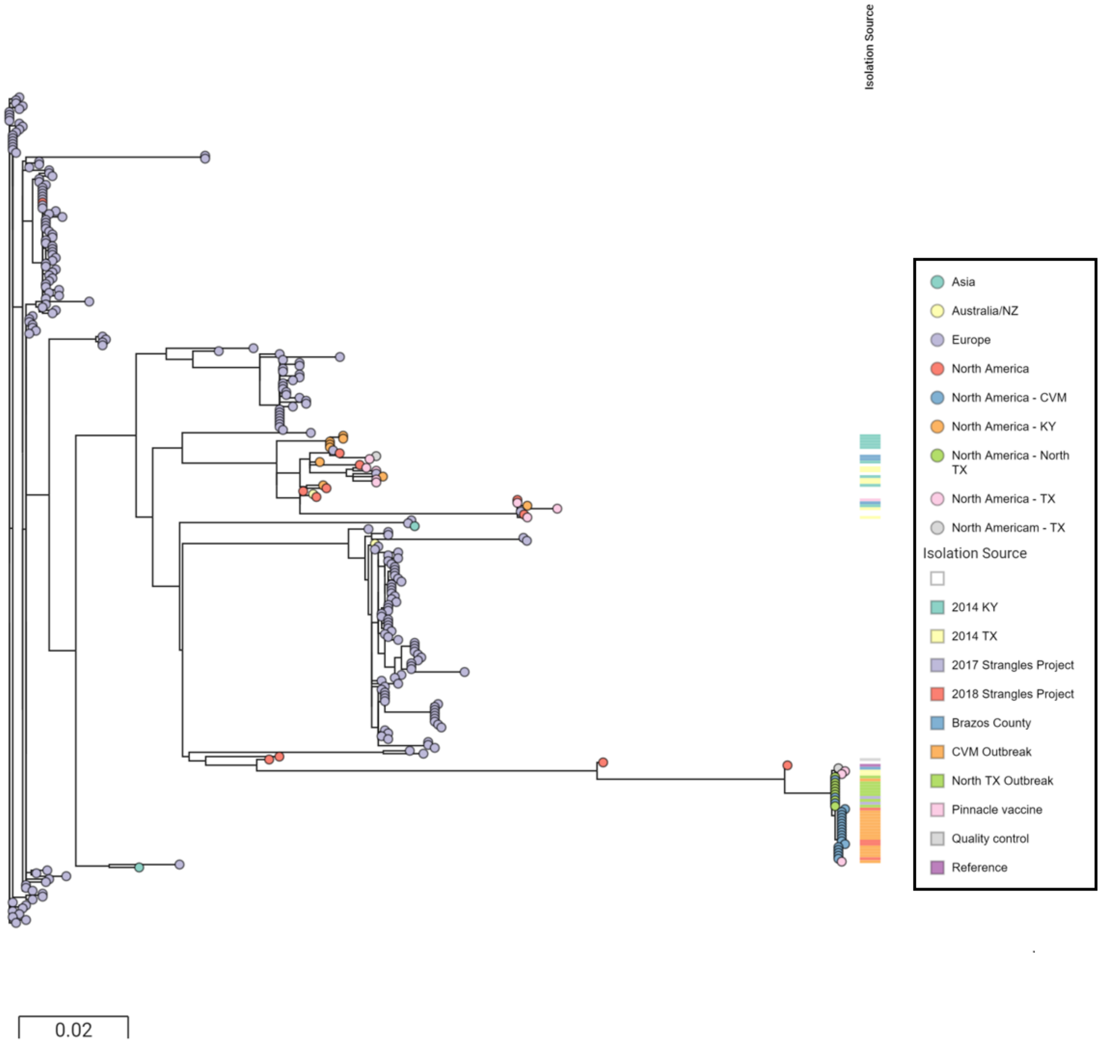


Fig 2-4. Phylogenetic tree of the 54 United States and 230 publicly available SEE isolates. The majority of SEE isolates cluster based on continent of origin, whereas some isolates from Kentucky and Texas are clustered with isolates originating from Europe.

2.4. Discussion

Our initial motivation for this project was to understand an outbreak of strangles that occurred >5 months after a vaccine study involving experimental infection of yearling horses with a strain of SEE recovered from an infected horse from a prior outbreak at a ranch in TX. Although the index horse for the CVM outbreak had been housed with horses previously infected with the infectious challenge strain, these horses and the index case had been co-mingled for several months prior to the onset of clinical signs. The horses with which the index horse was housed had neither clinical signs nor visible abnormalities in their guttural pouches on multiple sequential endoscopic examinations (at least 3 separate guttural pouch examinations and samplings at intervals of 2 weeks), and had culture-negative results of guttural pouch lavage fluid prior to being comingled. Persistent carriers of SEE represent an important source of perpetuating strangles.[20,28,29] Our results indicate that a persistent carrier from the first phase of our vaccine study was a likely source of the CVM strangles outbreak based on WGS results. Using qPCR testing in combination with culture-based diagnostic tests on guttural pouch lavage fluid is more likely to identify persistent carriers than using culture alone.[35,82] Additionally, use of nasopharyngeal lavage[83] is likely superior to culture of the guttural pouches alone[31] because carriage of SEE in sinuses or nasopharynx could be missed by sampling only the guttural pouches. The index case (17-004) was housed in a paddock with 5 horses used in the first phase of our 2017 vaccine study, and had no direct contact with other horses. The observed similarity between the index case, infection strain (11-017), and isolates from the 2017 vaccine study (17-007, 17-008) suggests that the index case was not the persistent shedder. However, this does strongly suggest that the persistent carrier was one of the horses that had previously been experimentally infected, during the 2017 vaccine project, sharing the paddock with the index case. If there was a carrier in the paddock, it is clear that repeated endoscopy and culture of lavage of the guttural pouches was not adequate to detect carriers. Indeed, even if the carrier was another horse from the vaccine project not housed in a paddock with this horse, all the recovered horses from the vaccine project had neither evidence of empyema, chondroids, or inflammation of the guttural pouches detected by endoscopy and were culture-negative for SEE in guttural pouch lavage fluid collected on at least 4 occasions. This also underscores advantages of collecting nasopharyngeal lavage fluid for testing rather than guttural pouch lavage fluid. Relying on culture-based identification of guttural pouch fluid alone as a criterion for releasing horses from isolation likely contributed to our failure to identify the persistent shedder.

An interesting finding was that the SNP in the *pbp2x* was not found in the horses from the vaccine challenge that preceded the outbreak. Thus, it seems probable that this mutation occurred in the index case, or the carrier that infected the index case. Previous studies of *Streptococcus pneumoniae* (*S. pneumoniae*) described *pbp2x* mutations associated with development of resistance to penicillin, and changes in cell division or cell wall development.[84,85] Furthermore, the *pbp2x* gene was identified as essential for survival of SEE using TraDIS.[86] To date, we have only demonstrated that presence of the *pbp2x* SNP influenced colony morphology of SEE isolates, similar to what has been observed for *S. pneumoniae*, and that this SNP was not associated with penicillin resistance. Notably, none of the horses in the outbreak was treated with penicillin.

The comparison of the 230 SEE strains from the public repository PATRIC with 54 US (TX and KY) isolates was made using a phylogenetic tree. Isolates from the CVM outbreak and north TX outbreak clustered separately from the isolates from other countries. Nevertheless, there were isolates from KY and TX from 2014 that were similar to isolates from continents other than North America such as Europe, suggesting evidence of possible transmission of infections from the US to Europe that is likely attributable to international horse transport despite procedures for biosecurity.[20] These findings help fill a knowledge gap of representation of US isolates in studies of the worldwide distribution and dissemination of SEE infection. This molecular epidemiological study represents the largest number of US SEE isolates reported using WGS. In a study conducted by Harris *et al.* of >200 SEE isolates, the majority of isolates were from countries in Europe, and a few were from countries in Asia and Australia. While that study included 3 US isolates, 2 were the modified-live SEE strains from the Pinnacle® vaccine available in the US, and the other was collected from a horse in the US in 1981.[29] The addition of 54 SEE US strains improves our understanding of changes in SEE isolates over time and by geographical region, in order to have a clearer picture of what variances occur around the world. We noted that the variation among isolates of SEE from around the world was estimated to be about 0.0117%.

We also provide new information regarding variation in the SeM protein based on those characterized in the pubMLST SEZ database. We identified a novel variant that differs by 3 bp from SeM 43 in strain 14-052, newly identified as SeM 148, and another that only differs by 1 bp from SeM 137 in strain 14-150, now identified as SeM 147. Moreover, from our index case (17-004) we found a 190-bp deletion (position 1 – 191) in the variable region of the SeM gene, likely causing the loss of protein function as previously described.[54,87] It is unclear whether this deletion of the SeM gene reflected a host-adaptation in either the index case or in the silent carrier that was transmitted to the index case. In the phylogenetic tree of the 54 US SEE most isolates grouped based on their SeM type, but a few isolates did not. Specifically, strain 14-092 with SeM 28 grouped with strains 14-125 and 14-140 which are identified as SeM 57, and strain 14-148 with SeM 28 was grouped with strain 14-150 with the newly described SeM 146. Possible explanations for these unexpected findings include recombination events in these isolates, or misclassification resulting as an artifact of some of the genome assemblies having many contigs.

Another interesting finding from our study was identification of a cluster of 4 clinical isolates that appeared to be the Pinnacle vaccine strain. Although the ability of this vaccine to cause disease is known,[88] only 1 of the 4 clinical isolates was from a horse that was considered to have been infected by the vaccine. It is possible that horses vaccinated during an outbreak that develop clinical signs develop disease from the vaccine rather than the natural infection and the vaccine goes unrecognized as the cause of disease. Using WGS can help distinguish the source and genomic diversity of Pinnacle vaccine-associated strangles.

Our study has a number of limitations. The principal limitation is the use of a convenience sample which limits our ability to extrapolate results beyond our isolates. For example, we only included isolates from 2 of the 50 states of the US. We also had relatively limited variation because so many of our isolates were from the 2 outbreaks in TX. Despite these limitations, we think our results have a number of important findings. First, they indicate that serial (n = 3) microbiologic culture of guttural pouch fluid alone is not adequate for identifying chronic carriers, and chronic carriers can occur in the absence of gross abnormalities observed in the guttural pouches. The role of the SNP in the *pbp2x* gene that appeared to arise between infecting individual horses with the challenge strain and the outbreak in a horse co-mingled with long-recovered (>5 months) horses merits further investigation. Although it could be a random mutation, it is possible that this change reflected a mechanism of adaptation to the host or modulated virulence or transmissibility. Some cases of Pinnacle vaccine-associated strangles appear to go undetected, possibly because they are associated with vaccination during an outbreak. Although isolates tend to cluster by place of origin, some isolates will cluster with isolates from other countries reflecting the global dissemination of SEE in horse populations. In addition, we identified 2 new SeM types. It is clear that much remains to be learned regarding the molecular epidemiology of SEE.

3. DIFFERENCES IN THE ACCESSORY GENOMES AND METHYLOMES OF STRAINS OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* AND OF *STREPTOCOCCUS EQUI* SUBSP. *ZOOEPIDEMICUS* OBTAINED FROM THE RESPIRATORY TRACT OF HORSES FROM TEXAS

3.1. Introduction

Streptococcus equi subspecies *equi* (SEE) is the causative agent of the infectious disease strangles. An ancient and highly contagious upper respiratory disease of horses, SEE is a host-restricted pathogen.[1,20,27,89,90] Strangles is characterized by swollen lymph nodes, purulent nasal discharge, guttural pouch empyema, lethargy, and fever.[1,27] SEE is thought to have evolved from an ancient strain of *Streptococcus equi* subspecies *zooepidemicus* (SEZ) through a proposed evolutionary bottleneck.[10,29,43] Generally, SEZ is an opportunistic pathogen of horses,[60] and is commonly recovered from the respiratory tract as a commensal bacterium,[13] however, strains of SEZ are known to cause outbreaks of upper respiratory tract disease in horses that resembles strangles.[91,92] SEZ is also a pathogen of other mammalian species, including livestock and humans.[44-48]

Published reports of genomic comparisons of SEE and SEZ are exiguous. Differences between the strains SEE 4047 and SEZ H70 were associated with the acquisition of mobile genetic elements, such as integrative conjugative elements (ICE) and prophages.[10] Specifically, SEZ H70 was described to have 2 ICEs and no acquired prophage, whereas 2 ICE (ICE*Se1*, ICE*Se2)* and 4 prophages (φSeq1 - φSeq4) were found in SEE 4047. That study further indicated that evolution of SEE from SEZ was associated with reduced genetic diversity in SEE as determined by using quantitative PCR to compare genes identified in either the SEE 4047 or SEZ H70 genome with additional isolates of SEZ and SEE.[10] Although SEE has relatively reduced genetic diversity, greater genetic variation has been described for isolates of SEZ.[45,93] The variability among isolates of SEZ is also demonstrated by the multilocus sequencing typing (MSLT) database of SEZ in which over 400 sequence types (ST) have been described, whereas only 2 primary ST profiles have been described for SEE (accessed Feb. 6, 2021).[43] This greater genetic diversity of SEZ might explain its ability to adapt to many mammalian hosts. Much remains to be learned, however, about the differences between SEE and SEZ, and about how SEE evolved to be host-restricted.

Data from untargeted sequencing methods such as whole genome sequencing (WGS) comparing strains of SEZ from the respiratory tract of horses with clinical isolates of SEE from horses to validate the existing observations are very limited.[10] One untargeted approach for studying bacterial species is to define and compare the core and accessory genomes of the individual species. This tack has been described either for studying a single bacterial species or for the comparisons of several species of streptococcal organisms.[94,95] The core genome elements for subspecies are defined as those found in the genomes of both subspecies, and the accessory genome elements (AGEs) for subspecies are those that are not found among the subspecies core genome elements. Furthermore, it is possible with PacBio WGS to characterize the complete methylome of prokaryotes.[96] Traditionally, the presence of methylation of bacterial DNA has been recognized as a means by which bacteria are protected against bacteriophages or other foreign DNA. Methyl groups present on the same sequence motifs protect against enzymatic degradation, whereas the DNA lacking the same methylation is recognized as foreign by bacterial endonucleases and results in cleavage at these unmethylated motifs.[97,98] Methylation, however, can also alter gene expression, alter virulence in some bacteria,[99-101] and even result in adaptive evolution.[102] Methylated bacterial DNA is most commonly recognized as residues of N6-methyl-adenosine (m6A), N4-methyl-cytosine (m4C), or C5-methyl-cytosine (m5C).[97,98] Thus, we used the WGS technology of PacBio® single molecule, real-time (SMRT) to characterize the core genome and accessory genomes, and to compared the methylomes of SEE and SEZ to identify potential differences that might help elucidate how SEE evolved to be a host-specific pathogen.

3.2. Materials and methods

3.2.1. *Streptococcus equi* isolates

Fifty SEE and 50 SEZ were selected to be included in this study (B-1 Table). The SEE isolates were collected from horses from various regions of Texas during multiple years (2012 – 2019), aiming for a more representative and geographically diverse population of isolates. The 50 SEZ isolates were selected from the respiratory tract of horses from various regions of Texas, from multiple years (2010 – 2020), and were representative of the differing disease states recognized for SEZ in horses (*i.e.,* commensal and virulent isolates).

3.2.2. Bacterial DNA extraction and whole genome sequencing

The *Streptococcus* isolates were cultured overnight in 3 ml of Todd Hewitt medium (HIMEDIA®, West Chester, PA, USA) in 5% CO_2 at 37°C. Following incubation overnight, the isolates were centrifuged at 3,000 × g for 10 minutes to create a pellet. The supernatants were discarded, and DNA extractions were performed using the DNeasy® UltraClean® Microbial kit (Qiagen®, Hilden, Germany), following the manufacturers' instructions with slight modifications. Briefly, the bacteria pellets were resuspended in 300 µl of PowerBead solution, and transferred into PowerBead tubes. Fifty µl of solution SL was added, and the PowerBead tubes were incubated at 70°C for 10 minutes, followed by horizontal vortexing for an additional 10 minutes. Then, the PowerBead tubes were centrifuged and the supernatants were transferred to new tubes. One hundred (100) µl of solution IRS were added to the supernatants, incubated for 15 minutes at 4°C, and then centrifuged. The supernatants were transferred to new tubes without disturbing the pellet, 900 µl of solution SB were added and mixed thoroughly. Seven hundred (700) µl of this solution was transferred to MB spin column tubes, centrifuged, and the flow-through was discarded, then this step was repeated. Additionally, 300 µl of solution CB was added to the columns and centrifuged. Then, another centrifuge step was performed to remove any excess fluid, and the MB spin columns were transferred to new collection tubes. Finally, 50 µl of the solution EB was added to the columns and centrifuged. The DNA quality and concentrations were measured using the NanoDrop spectrophotometer (ND-1000, Thermo Fisher Scientific, Waltham, MA, USA), and sent to the Duke Center for Genomic and Computational Biology (GCB) for WGS on the PacBio® Sequel platform.

3.2.3. Bioinformatic analysis

After the completion of WGS at GCB, raw subreads were assembled into genomes *de novo* using CANU (v7.0)[103] on the HPRC computing cluster. The assembled genomes were confirmed to be SEE or SEZ through ribosomal MLST.[53] The genomes were then annotated with RASTtk (v2.0),[104] using the web-based server. Following annotation, the genomes were input into Spine (v0.3.2)[94] to define the core genome (*i.e.*, elements found in all genomes) of both *Streptococcus equi* subspecies. Using the core genome output from Spine, the accessory genomes (*i.e.,* elements present in some genomes but absent from others) for each isolate were identified using AGEnt (v0.3.1).[94] Finally, ClustAGE (v0.8)[105] was used to identify and group the AGEs into bins for the SEE and SEZ genomes. The graphical representation of bins with clustered AGEs by each individual genome was performed with the ClustAGE plot (http://vfsmspineagent.fsm.northwestern.edu/cgi-bin/clustage_plot.cgi). Using a custom R script (v4.0.3) (C-1 Appendix), bins were identified with AGEs specific to either all SEE (n = 50) or all SEZ (n = 50). The genes of the AGEs within the selected bins with ≥ 95% of the protein identified were included, and were compared to their respective reference genomes (SEE 4047 or SEZ H70). Using the Cytoscape (v.3.8.2)[106] plug-in, ClueGO (v2.5.7)[107] the Gene Ontology (GO) terms and pathway interactions for the AGEs of SEE and SEZ were evaluated using default parameters, and the localization of the protein within the cell was determined using PSORTb (v3.0).[108]

The complete methylation profiles of a subset of SEE (n = 24) and SEZ (n = 24) genomes were characterized; these isolates were selected to be representative of distribution across the phylogenetic tree (B-1 – B-3 Figs). The complete methylomes were characterized with the BaseMod (https://github.com/ben-lerch/BaseMod-3.0) pipeline in the PacBio® SMRT Link (v8.0) command line tools. Briefly, pbmm2 was used to align the raw BAM files to the appropriate reference genome (*i.e.*, SEE 4047 or SEZ H70). Using the aligned BAM files, the kineticTools function *ipdSummary* was implemented to generate GFF and CSV files with the base modification information. Next, the MotifMaker *find* function was used to generate a second set of CSV files that identified consensus motifs. Finally, the execution of the MotifMaker *reprocess* function generated GFF files with all of the modifications that were part of the motifs. Using R (v4.0.3), the motif GFF files were filtered based on having the presence of a known methylation type (m4C or m6A), and having a QV score (*i.e.,* a quality measure of the detection event) of ≥ 30. These filtered GFF files of SEE or SEZ genomes were then annotated by either the SEE 4047 or SEZ H70 reference genome, respectively, using the BedTools[109] *annotate* function. Annotated outputs were then compared across the SEE and SEZ genomes for the presence or absence of methylation of homologous proteins using custom scripts in R (C-1 Appendix). A list of homologous proteins (≥ 99% identity) from SEE 4047 and SEZ H70 was generated using the PATRIC proteome comparison. Identified motifs were then compared to the SEE 4047 and SEZ H70 genomes using the Restriction Enzyme Database (REBASE).[110] The Cytoscape (v.3.8.2)[106] plug-in, ClueGO (v2.5.7)[107] was implemented using default parameters to assess the GO terms and pathway interactions for the different sites of methylation among the SEE and SEZ genomes. The Linux and R codes for this work are provided in the supplementary materials (C-1 Appendix).

3.3. Results

Comparisons of the accessory genome of the 50 SEE and 50 SEZ isolates were performed using the Spine, AGEnt, and ClustAGE pipeline (Fig 3-1) to generate the AGEs identified among these isolates (Fig 3-1). The AGEs found only in the 50 SEE isolates were primarily associated with 1 of the 2 ICEs or 1 of the 4 acquired prophages described for SEE 4047,[10] and a total of 85 coding sequences (CDS) within the SEE 4047 genome were identified: 4 of the 85 elements were within the region of the ICE*Se1* elements (SEQ_0756 – SEQ_0758; SEQ_0761) and 36 of the 85 elements were associated with ICE*Se2* (Table 3-1). Of the 85 CDS, none (0) AGEs was located on prophage φSeq1, 17 were part of φSeq2, 20 were from φSeq3, and 7 were on φSeq4 (Table 3-1). Finally, SEQ_1102 was identified as part of the AGEs and was not found on either of the 2 ICEs or 4 prophages, but was rather associated with an insertion element in SEE 4047. Interestingly, all of the CDS that form each of the described ICE or prophage from SEE 4047 were not found in all our 50 SEE isolates. The functions of the identified AGE were primarily associated with those of the acquired prophages and of hypothetical proteins. Additionally, the CDS that comprise the equibactin locus (SEQ_1233 – SEQ_1246) and 3 of the 4 superantigens, *seeH* (SEQ_2036), *seeI* (SEQ_2037), and *seeL* (SEQ_1728), were also identified as part of the AGE of SEE relative to SEZ. The GO functions and pathway interactions of the 85 CDS identified in the AGEs from SEE were assessed using ClueGO, and 23 CDS were characterized (Fig 3-2, B-2 Table). The primary GO functions identified were DNA modification, endonuclease activity, and ATPase activity. Also noted were the KEGG pathways of biosynthesis of the siderophore group nonribosomal peptides, and *Staphylococcus aureus* infection.

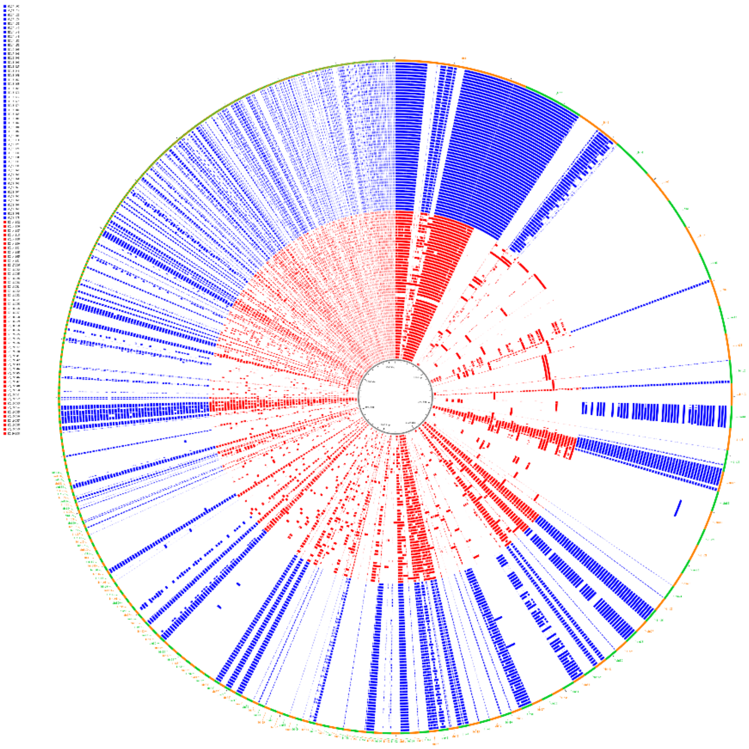


Fig 3-1. Comparison of accessory genome elements (AGE) of SEE (n = 50) and SEZ (n = 50) genomes. The outer ring shows the ClustAGE bins that are ≥ 200 base-pairs in size these are ordered clockwise from the largest bin to the smallest bin, and are differentiated by orange and green to define bin borders. The concentric inner bands show the distribution of AGE within each individual isolate. Bands that are blue represents SEE isolates, and bands that are red represent SEZ isolates. The central ruler of the figure indicates the cumulative size of the AGE in kilobases.

Table 3-1. Accessory genome elements identified in all 50 SEE genomes.

RefSeq_4047	Gene Name	Region	Psortb	Protein
SEQ_0756		ICESe1	Cytoplasmic	Transcriptional regulator
SEQ_0757		ICESe1	Cytoplasmic	Modification methylase PstI (EC 2.1.1.72)
SEQ_0758		ICESe1	Cytoplasmic	Type II site-specific deoxyribonuclease
SEQ_0761		ICESe1	Cytoplasmic	USG protein
SEQ_0787		Prophage Seq2	Unknown	Phage integrase: site-specific recombinase
SEQ_0816		Prophage Seq2	Unknown	Phage protein
SEQ_0817		Prophage Seq2	Unknown	Phage protein
SEQ_0818		Prophage Seq2	Unknown	Phage endonuclease
SEQ_0819		Prophage Seq2	Cytoplasmic	Phage terminase
SEQ_0823		Prophage Seq2	Cytoplasmic	Phage portal protein
SEQ_0824		Prophage Seq2	Cytoplasmic	Prophage Clp protease-like protein
SEQ_0825		Prophage Seq2	Cytoplasmic	Phage capsid protein
SEQ_0826		Prophage Seq2	Cytoplasmic	Putative capsid protein (ACLAME 311)
SEQ_0827		Prophage Seq2	Cytoplasmic	DNA packaging protein
SEQ_0828		Prophage Seq2	Unknown	Phage protein
SEQ_0829		Prophage Seq2	Cytoplasmic	Phage protein
SEQ_0830		Prophage Seq2	Cytoplasmic	Phage protein
SEQ_0831		Prophage Seq2	Cytoplasmic	Phage major tail protein
SEQ_0832		Prophage Seq2	Unknown	Phage protein
SEQ_0833		Prophage Seq2	Unknown	Phage protein
SEQ_0835		Prophage Seq2	Unknown	Phage-related protein
SEQ_1102		Insertion Element	Cytoplasmic	Site-specific recombinase

Table 3-1. Continued.

RefSeq_4047	Gene Name	Region	Psortb	Protein
SEQ_1231		ICESe2	Cytoplasmic	hypothetical protein
SEQ_1233	eqbN	ICESe2	Unknown	hypothetical protein
SEQ_1234	eqbM	ICESe2	Unknown	hypothetical protein
SEQ_1235	eqbL	ICESe2	CytoplasmicMembrane	Heterodimeric efflux ABC transporter
SEQ_1236	eqbK	ICESe2	CytoplasmicMembrane	Heterodimeric efflux ABC transporter
SEQ_1237	eqbJ	ICESe2	CytoplasmicMembrane	Duplicated ATPase component BL0693 of energizing module of predicted ECF transporter
SEQ_1238	eqbI	ICESe2	CytoplasmicMembrane	Transmembrane component BL0694 of energizing module of predicted ECF transporter
SEQ_1239	eqbH	ICESe2	Cytoplasmic Membrane	Substrate-specific component BL0695 of predicted ECF transporter
SEQ_1240	eqbG	ICESe2	Cytoplasmic	hypothetical protein
SEQ_1241	eqbF	ICESe2	Cytoplasmic	hypothetical protein
SEQ_1242	eqbE	ICESe2	Cytoplasmic	Polyketide synthase modules and related proteins
SEQ_1243	eqbD	ICESe2	Cytoplasmic	2,3-dihydroxybenzoate-AMP ligase (EC 2.7.7.58) of siderophore biosynthesis
SEQ_1244	eqbC	ICESe2	Cytoplasmic	4'-phosphopantetheinyl transferase (EC 2.7.8.-)
SEQ_1245	eqbB	ICESe2	Cytoplasmic	Iron aquisition yersiniabactin synthesis enzyme YbtT @ Thioesterase in siderophore biosynthesis gene cluster
SEQ_1246	eqbA	ICESe2	Cytoplasmic	Iron-dependent repressor
SEQ_1249		ICESe2	Unknown	hypothetical protein
SEQ_1250		ICESe2	Cytoplasmic	hypothetical protein
SEQ_1252		ICESe2	Cytoplasmic	hypothetical protein
SEQ_1253		ICESe2	Cell wall/Extracellular	Superfamily II DNA and RNA helicase
SEQ_1254		ICESe2	Cytoplasmic	hypothetical protein
SEQ_1257		ICESe2	Cytoplasmic	FIG00645039: hypothetical protein with HTH-domain

Table 3-1. Continued.

RefSeq_4047	Gene Name	Region	Psortb	Protein
SEQ_1258		ICESe2	Cytoplasmic	abortive infection protein AbiGI
SEQ_1260		ICESe2	Unknown	hypothetical protein
SEQ_1261		ICESe2	Unknown	NLP/P60 family protein
SEQ_1262		ICESe2	Cytoplasmic	Modification methylase Cfr9I (EC 2.1.1.113)
SEQ_1263		ICESe2	Unknown	TrsE-like protein
SEQ_1264		ICESe2	CytoplasmicMembrane	hypothetical protein
SEQ_1265		ICESe2	Cytoplasmic	hypothetical protein
SEQ_1266		ICESe2	CytoplasmicMembrane	hypothetical protein
SEQ_1267		ICESe2	CytoplasmicMembrane	Maff2 family protein
SEQ_1268		ICESe2	CytoplasmicMembrane	hypothetical protein
SEQ_1269		ICESe2	CytoplasmicMembrane	ABC-type antimicrobial peptide transport system
SEQ_1270		ICESe2	CytoplasmicMembrane	hypothetical protein
SEQ_1271		ICESe2	CytoplasmicMembrane	hypothetical protein
SEQ_1274		ICESe2	Cytoplasmic	Chromosome (plasmid) partitioning protein ParB
SEQ_1275		ICESe2	CytoplasmicMembrane	Chromosome (plasmid) partitioning protein ParA
SEQ_1728	seeL	Prophage Seq3	Unknown	Streptococcal pyrogenic exotoxin K (SpeK)
SEQ_1739		Prophage Seq3	CytoplasmicMembrane	Phage tail length tape-measure protein
SEQ_1740		Prophage Seq3	Unknown	conserved hypothetical protein - phage associated
SEQ_1741		Prophage Seq3	Cytoplasmic	conserved hypothetical protein - phage associated
SEQ_1742		Prophage Seq3	Unknown	Phage major tail protein
SEQ_1743		Prophage Seq3	Cytoplasmic	Phage major tail protein
SEQ_1744		Prophage Seq3	CytoplasmicMembrane	Structural protein
SEQ_1745		Prophage Seq3	Unknown	Phage protein

Table 3-1. Continued.

RefSeq_4047	Gene Name	Region	Psortb	Protein
SEQ_1746		Prophage Seq3	Unknown	Phage protein
SEQ_1747		Prophage Seq3	Cytoplasmic	Phage protein
SEQ_1748		Prophage Seq3	Cytoplasmic	hypothetical phage protein
SEQ_1749		Prophage Seq3	Unknown	Phage major capsid protein
SEQ_1750		Prophage Seq3	Cytoplasmic	Phage major capsid protein
SEQ_1751		Prophage Seq3	Unknown	FIG01114710: hypothetical protein
SEQ_1755		Prophage Seq3	Cytoplasmic	Guanosine-3'
SEQ_1756		Prophage Seq3	Unknown	hypothetical protein
SEQ_1757		Prophage Seq3	Cytoplasmic	Phi Mu50B-like protein
SEQ_1758		Prophage Seq3	Cytoplasmic	Phage portal protein
SEQ_1762		Prophage Seq3	Unknown	Pleiotropic regulator of exopolysaccharide synthesis
SEQ_1763		Prophage Seq3	Cytoplasmic	Chromosome segregation ATPase
SEQ_2036	seeH	Prophage Seq4	Extracellular	Streptococcal pyrogenic exotoxin H (SpeH)
SEQ_2037	seeI	Prophage Seq4	Extracellular	Exotoxin
SEQ_2038		Prophage Seq4	Unknown	Phage lysin
SEQ_2040		Prophage Seq4	CytoplasmicMembrane	Phage holing
SEQ_2041		Prophage Seq4	Unknown	Phage holing
SEQ_2042		Prophage Seq4	Cytoplasmic	Phage protein
SEQ_2043		Prophage Seq4	Unknown	hypothetical protein

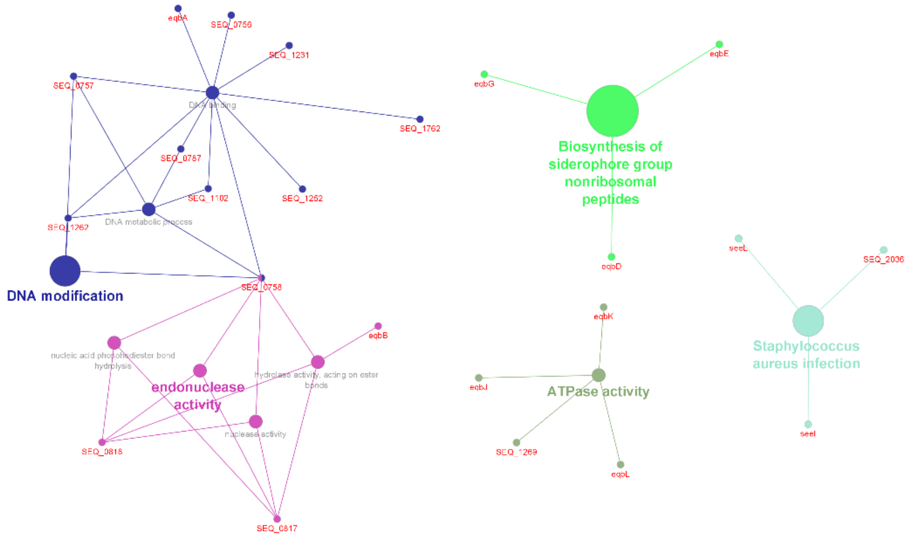


Fig 3-2. Gene ontology (GO) terms and KEGG pathways (annotated in ClueGO) in the accessory genome elements identified in all SEE (n = 50) genomes. Circle size represents the degree of the positive relationship between the GO terms, and the term's adjusted P-value. The related terms are grouped and presented in the same color.

Next, elements that were specific to all 50 SEZ genomes were considered, and only 15 CDS from the H70 genome were identified (Table 3-2). Of the 15 CDS, 8 had been previously described to be deleted from the SEE 4047 genome,[10] in agreement with our findings. These elements were found throughout the SEZ genomes and were not primarily localized to any ICE, unlike the SEE-specific AGEs. Localization of the 15 SEZ-specific AGEs were found primarily to be part of the cytoplasm (n = 5) or the cytoplasm membrane (n = 7) in the bacterium, a single hypothetical protein was extracellular, and the location of the remaining hypothetical proteins (n = 2) were unknown. The apparent function of these AGEs largely points to differences in fermentation of the carbohydrate lactose (SZO_15220 – SZO_15250) and sorbitol (SZO_01750) (Table 3-2). The functions of the 15 CDS were evaluated using ClueGO, and a function was identified for only 3 CDS (Fig 3-3, B-3 Table). Unsurprisingly, galactose metabolism was the only GO term described from the 3 CDS (*lacE*, *lacF*, and *lacG*).

Table 3-2. Accessory genome elements identified in all 50 SEZ genomes.

RefSeq_H70	Gene Name	Region	Psortb	Protein
SZO_01750	sorD	deleted in 4047	Cytoplasmic	Sorbitol-6-phosphate 2-dehydrogenase (EC 1.1.1.140)
SZO_14750			Cytoplasmic	Transcriptional regulator
SZO_15220	lacG	deleted in 4047	Cytoplasmic	6-phospho-beta-galactosidase (EC 3.2.1.85)
SZO_15240	lacF	deleted in 4047	Cytoplasmic	PTS system
SZO_15250	lacT	deleted in 4047	Cytoplasmic	Beta-glucoside bgl operon antiterminator
SZO_05610		deleted in 4047	CytoplasmicMembrane	ABC transporter ATP-binding protein
SZO_05620		deleted in 4047	CytoplasmicMembrane	Daunorubicin resistance transmembrane protein
SZO_05630		deleted in 4047	CytoplasmicMembrane	Efflux ABC transporter
SZO_14690		ESAT-6-like	CytoplasmicMembrane	Branched-chain amino acid transport system carrier protein
SZO_14730	comB		CytoplasmicMembrane	Competence-stimulating peptide ABC transporter permease protein ComB
SZO_14744			CytoplasmicMembrane	Competence-stimulating peptide ABC transporter ATP-binding protein ComA
SZO_15230	lacE	deleted in 4047	CytoplasmicMembrane	PTS system
SZO_14742			Extracellular	FIG01116836: hypothetical protein
SZO_10380			Unknown	FIG01117834: hypothetical protein
SZO_14743			Unknown	FIG01120711: hypothetical protein

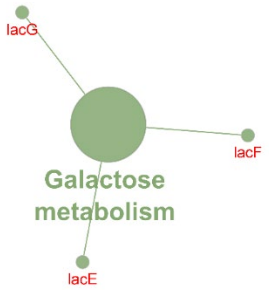


Fig 3-3. Gene ontology (GO) terms and KEGG pathways (annotated in ClueGO) in the accessory genome elements identified in all SEZ (n = 50) genomes. Circle size represents the degree of the positive relationship between the GO terms, and the term's adjusted P-value. The related terms are grouped and presented in the same color.

The PacBio SMRT WGS permits characterization of methylation patterns of bacterial genomes through the implementation of the BaseMod pipeline developed by PacBio®.[96] Using REBASE, the methylation motifs of a representative subset of 24 isolates each of SEE and SEZ were compared to the reference genomes SEE 4047 and SEZ H70. The methylation motifs identified in the 24 SEE genomes were more consistent than those identified in the 24 SEZ genomes (B-4 Table). All 24 SEE genomes had the motif sequence, CTGCAG with methylation occurring at approximately 95% of each sequencing occurrence. An additional methylation motif, CATCC not identified in REBASE but was noted in 13 of 24 SEE isolates, and a single novel methylation motif (GGATGNND) was found in the SEE isolate 18-074 originating from Salado, Texas (Table 3-3). However, the partnered methylation motif sequences, GGATG and CATCC, described in REBASE were only found in 12 of 24 SEZ isolates. Furthermore, the majority of the methylation motif sequences recognized in the 24 SEZ were not commonly seen in all these isolates, and the majority were novel motifs (Table 3-3).

Table 3-3. Novel motif sequences from SEE (n = 24) and SEZ (n = 24) genomes.

Motif Sequences	Genome ID	Subsp	Center Position	Modification Type
GGATGNND	18-074	equi	3	m6A
ACCNNNNNTCTT/AAGANNNNNGGT	19-050	zoo	4	m6A
ACAYNNNNNRGG	14-006	zoo	3	m6A
ACCCA	19-052	zoo	5	m6A
AGTNNNNNNGTC/GACNNNNNNACT	19-044	zoo	1	m6A
AGTNNNNNNGTC/GACNNNNNNACT	19-050	zoo	1	m6A
CCANNNNNNNNNTAC/GTANNNNNNNNNTGG	18-066	zoo	3	m6A
TCANNNNNNTGG/CCANNNNNNTGA	14-151	zoo	3	m6A
TCANNNNNNTGG/CCANNNNNNTGA	19-048	zoo	3	m6A
CTCCAG/CTGGAG	18-059	zoo	5	m6A
CTCCAG/CTGGAG	19-043	zoo	5	m6A
CTCCAG/CTGGAG	19-044	zoo	5	m6A
GACNNNNNTARG/CYTANNNNNGTC	19-047	zoo	4	m6A
GACNNNNNTARG	19-041	zoo	2	m6A
GCANNNNNNNNTTC/GAANNNNNNNNTGC	19-038	zoo	3	m6A
GACNNNNNTARG	19-047	zoo	2	m6A
GATC	19-058	zoo	2	m6A
GATGC/GCATC	19-056	zoo	2	m6A
GCTANAC	19-045	zoo	6	m6A
TCANNNNNGTTY/RAACNNNNNTGA	18-058	zoo	3	m6A
RGATCY	14-007	zoo	5	m4C
RGATCY	18-055	zoo	5	m4C
TCCAG	17-006	zoo	4	m6A
TCCAG	19-036	zoo	4	m6A
YACNNNNNGTR	19-058	zoo	2	m6A

Homologous proteins of the reference genomes of SEE (4047) and SEZ (H70) with a similarity of ≥ 99% were selected as targets to compare the presence or absence of methylation between the *Streptococcus equi* subspecies. In considering sites where methylation occurred in the 24 SEE genomes but not in the 24 SEZ genomes on homologous proteins, 37 CDS were identified. This was determined from a pool of 89 CDS with methylation present in the SEE genomes, and 231 CDS in which SEZ had no methylation present. The presence of methylation was found on the motif sequence, CTGCAG at 70 different locations within the 37 CDS, and was identified as the methylation type m6A (Table 3-4). To evaluate the GO terms and functions of these 37 CDS, ClueGo was implemented using default parameters. We noted the functions of exopeptidase activity, transition metal ion binding, transmembrane transport, quorum sensing, and propanoate metabolism (Fig 3-4, B-5 Table). Homologous proteins sites where methylation was found in all 24 SEZ but was absent in the 24 SEE genomes were reviewed. Likely due to the variability of SEZ genomes, only 10 potential CDS were identified on homologous proteins (B-6 Table). However, the location of the methylation and type (m6A or m4A) were not consistent among all 24 SEZ genomes.

Table 3-4. Methylation location, type and motif in 24 SEE genomes.

CDS	Location	Type	Motif
SEQ_0045	56855	m6A	CTGCAG
SEQ_0067	74697	m6A	CTGCAG
SEQ_0067	74700	m6A	CTGCAG
SEQ_0070	76364	m6A	CTGCAG
SEQ_0251	230695	m6A	CTGCAG
SEQ_0300	285354	m6A	CTGCAG
SEQ_0300	285357	m6A	CTGCAG
SEQ_0302	288740	m6A	CTGCAG
SEQ_0340	323013	m6A	CTGCAG
SEQ_0340	323016	m6A	CTGCAG
SEQ_0435	417164	m6A	CTGCAG
SEQ_0435	417167	m6A	CTGCAG
SEQ_0474	460537	m6A	CTGCAG
SEQ_0474	461395	m6A	CTGCAG
SEQ_0497	482852	m6A	CTGCAG
SEQ_0497	482855	m6A	CTGCAG
SEQ_0596	580039	m6A	CTGCAG
SEQ_0596	580042	m6A	CTGCAG
SEQ_0721	712040	m6A	CTGCAG
SEQ_0769	763220	m6A	CTGCAG
SEQ_0769	763223	m6A	CTGCAG
SEQ_0898	873274	m6A	CTGCAG
SEQ_0898	873277	m6A	CTGCAG
SEQ_0976	967141	m6A	CTGCAG
SEQ_1129	1118411	m6A	CTGCAG
SEQ_1277	1274166	m6A	CTGCAG
SEQ_1277	1274169	m6A	CTGCAG
SEQ_1278	1276130	m6A	CTGCAG
SEQ_1299	1296130	m6A	CTGCAG
SEQ_1299	1296133	m6A	CTGCAG
SEQ_1318	1318299	m6A	CTGCAG
SEQ_1318	1318302	m6A	CTGCAG
SEQ_1407	1406622	m6A	CTGCAG
SEQ_1407	1406625	m6A	CTGCAG
SEQ_1407	1408183	m6A	CTGCAG

Table 3-4. Continued.

CDS	Location	Type	Motif
SEQ_1410	1411644	m6A	CTGCAG
SEQ_1410	1411647	m6A	CTGCAG
SEQ_1439	1442999	m6A	CTGCAG
SEQ_1439	1443002	m6A	CTGCAG
SEQ_1448	1453017	m6A	CTGCAG
SEQ_1448	1453020	m6A	CTGCAG
SEQ_1597	1602240	m6A	CTGCAG
SEQ_1597	1602243	m6A	CTGCAG
SEQ_1597	1602706	m6A	CTGCAG
SEQ_1615	1626084	m6A	CTGCAG
SEQ_1625	1634867	m6A	CTGCAG
SEQ_1625	1634870	m6A	CTGCAG
SEQ_1627	1636655	m6A	CTGCAG
SEQ_1651	1658796	m6A	CTGCAG
SEQ_1651	1658799	m6A	CTGCAG
SEQ_1895	1896398	m6A	CTGCAG
SEQ_1895	1896401	m6A	CTGCAG
SEQ_1981	1925057	m6A	CTGCAG
SEQ_1981	1925060	m6A	CTGCAG
SEQ_1920	1928487	m6A	CTGCAG
SEQ_1920	1928490	m6A	CTGCAG
SEQ_1937	1945433	m6A	CTGCAG
SEQ_1937	1945436	m6A	CTGCAG
SEQ_1937	1945842	m6A	CTGCAG
SEQ_2009	2033472	m6A	CTGCAG
SEQ_2152	2161140	m6A	CTGCAG
SEQ_2152	2161143	m6A	CTGCAG
SEQ_2161	2171880	m6A	CTGCAG
SEQ_2161	2171883	m6A	CTGCAG
SEQ_2161	2172181	m6A	CTGCAG
SEQ_2161	2172184	m6A	CTGCAG
SEQ_2161	2173113	m6A	CTGCAG
SEQ_2161	2173116	m6A	CTGCAG
SEQ_2210	2224386	m6A	CTGCAG
SEQ_2210	2224389	m6A	CTGCAG

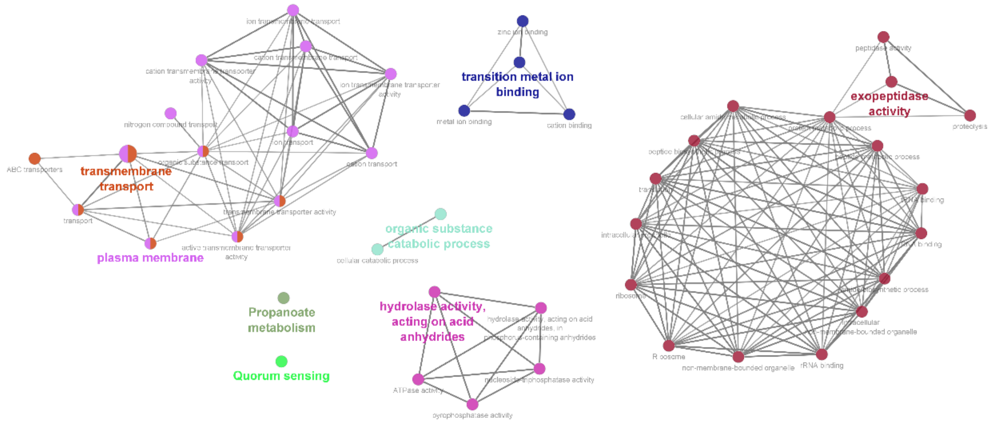

ion transmembrane transport
cation transmembrane transport
cation transmembrane transporter activity
ion transmembrane transporter activity
nitrogen compound transport
ion transport
organic substance transport
cation transport
ABC transporters
transmembrane transport
transport
transmembrane transporter activity
active transmembrane transporter activity
plasma membrane
zinc ion binding
transition metal ion binding
metal ion binding
cation binding
organic substance catabolic process
cellular catabolic process
Propanoate metabolism
Quorum sensing
hydrolase activity, acting on acid anhydrides
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
ATPase activity
nucleoside-triphosphatase activity
pyrophosphatase activity
peptidase activity
exopeptidase activity
proteolysis
cellular amide metabolic process
protein metabolic process
peptide biosynthetic process
peptide metabolic process
translation
tRNA binding
intracellular organelle
mRNA binding
ribosome
amide biosynthetic process
intracellular non-membrane-bounded organelle
Ribosome
non-membrane-bounded organelle
rRNA binding

Fig 3-4. Gene ontology (GO) terms and KEGG pathways (annotated in ClueGO) on homologous proteins where methylation is present in SEE (n = 24) genomes, but absent in SEZ (n = 24) genomes. Circle size represents the degree of the positive relationship between the GO terms, and the term's adjusted P-value. The related terms are grouped and presented in the same color.

3.4. Discussion

Comparisons of AGEs among isolates has been used to understand differences within the same bacterial species or across genera.[94,95] Our study was designed to help understand which genomic attributes contribute to host-specificity of SEE by comparing the AGEs of SEE (n = 50) and SEZ (n = 50) collected from the respiratory tract of horses from Texas. Through the AGEs analysis more SEE-specific CDS were noted than compared to the SEZ-specific CDS, and demonstrates the greater level of homozygosity (*i.e.,* reduced genetic diversity) in SEE isolates with an untargeted approach. This observation has been described before using the targeted approach of quantitative PCR in isolates from the United Kingdom.[10] The AGEs of the SEE isolates were primarily noted to be a part of the prophages (φSeq2 – φSeq4), and the 2 ICE

(ICE*Se1*, ICE*Se2*) described for the SEE 4047 genome.[10] However, no elements of the prophage φSeq1 were consistently found in the 50 SEE isolates used in our study. This finding is consistent with comparison of the accessory genome of SEE isolates by Harris *et al.*[29] The elements found on the prophage φSeq2 were primarily proteins characterized as phage elements and located in the cytoplasm. The superantigens *seeL* (SEQ_1728), *seeH* (SEQ_2036), and *seeI* (SEQ_2027) located on prophages, φSeq3 and φSeq4, were found among all SEE isolates. In contrast, *seeM* was not identified among our AGEs, which is consistent with evidence of its absence in some strains of SEE,[29] *seeM* also has been identified in a small number of strains of SEZ.[10] These superantigens have been show *in vitro* to induce increased production of gamma interferon (IFN-γ) from CD5$^+$ CD4$^+$ T-lymphocytes.[61] Similarly, superantigens in *Streptococcus pyogenes* (*S. pyogenes*) are described to cause the suppression of antibody production in part to the production of IFN -γ by overactivated CD4-positive T cells.[111,112] The conserved elements on ICE*Se1* (n = 4) were proteins noted as a transcriptional regulator, modification methylase, type II site-specific deoxyribonuclease, and a USG protein. These first 3 proteins are part of a type II restriction modification system according to REBASE,[110] and the USG protein function is unknown but is a member of the SIR protein family.[113] The elements from ICE*Se2* (n = 36) were hypothetical proteins, transport proteins, the equibactin locus (*eqbA* – *eqbN*) and chromosome partitioning proteins, and this was the most conversed (36 of 85 CDS) of the SEE mobile genetic element, similar to previous findings.[29] The equibactin locus (*eqbA* – *eqbN*), a novel iron acquisition element, was identified among all of the SEE isolates, although other studies have noted the partial or entire deletion of this locus in SEE isolates from the United Kingdom.[29,39] Interestingly, none of the ICE or prophages were identified in their entirety among all 50 SEE isolates (Table 3-1). This finding could be because none of the 50 SEE genomes are fully contiguous, due to more differences in acquired mobile genetic elements in SEE than initially thought, or because the absent portions of these acquired mobile genetic elements are similar to other CDS in the SEZ genomes. Nevertheless, this suggests that there is more variability seen among the CDS found within the ICE and prophages than described for SEE 4047. The primary functions described for the 23 CDS from the ClueGO analysis were DNA modification and binding, endonuclease activity, ATPase activity, and the KEGG pathways of *Staphylococcus aureus* infection and biosynthesis of siderophore group nonribosomal peptides. Thus, these functions reflect functions specific to SEE. The DNA binding, endonuclease activity, ATPase activity, and biosynthesis of siderophore group nonribosomal peptides pathway are all functions related to the novel iron acquisition of the equibactin locus.[39] Thus, enhanced iron acquisition might be a mechanism by which SEE is able to survive in the equine host, although it is unclear whether this function is somehow specific to the equine host (*i.e.,* enhances iron acquisition specifically or optimally in the equine respiratory tract) or whether genes in the equibactin locus serve functions other than iron acquisition that might confer host-specificity. Finally, the superantigens (*seeI*, *seeL*, *seeH*) all were a part of the *Staphylococcus aureus* infection pathway, which shares similarities in pathogenesis with SEE and its close relative *S. pyogenes.*[61,114] These superantigens activate multiple T-cells populations, and the production of antibodies can be suppressed through IFN-γ production by overactivated CD4 T cells.[111,112] Similarly, the over production of the proinflammatory cytokine tumor necrosis factor-α by immune cells that have recognized these superantigens results in suppression of phagocytic cell recruitment to the sites of infection.[115] These 2 functions divert the host's immune responses of antibody-complement opsonization and killing of the pathogens by phagocytes.[116]

Identification of AGEs in all 50 SEZ isolates from the same anatomic location of the same host-species from the same geographic region demonstrated the relatively high variability of this bacterial subspecies. Only 15 CDS were identified in all SEZ isolates that were also absent from all of the SEE isolates (Table 3-2). These elements were annotated to functions attributed to fermentation of lactose and sorbitol. Lactose and sorbitol are commonly known to be fermented almost exclusively by SEZ but not by SEE,[10] although alternative fermentation profiles have been described.[42,50] Another major difference between SEE and SEZ was in the components of the cytoplasmic membrane, 2 of which were related to competence stimulation (*ComA*, *ComB*). However, because of the highly variable genome of SEZ it was not possible to identify consistent differences between SEZ isolates and SEE. This variability in the genome of SEZ might explain its ability to adapt to new hosts and environments, whereas SEE might have evolved to more specifically infect horses (possibly by more efficiently scavenging iron when it is restricted).

The global methylomes of 24 SEE and 24 SEZ isolates were considered by using PacBio® SMRT sequencing and the BaseMod pipeline,[96] and sites of methylation on homologous proteins of the 2 subspecies were targeted. We elected to compare methylation of the homologous proteins of SEE and SEZ because of the high degree of similarity in the genomes of SEE and SEZ.[10] The important role of methylation has been described for *S. pyogenes*, the closest relative of SEE and SEZ, wherein the absence of methylation at a prominent motif was demonstrated to alter gene expression that resulted in decreased virulence and altered the bacterium's ability to thrive in neutrophils.[99] The global methylomes of the 24 SEE isolates were more consistent than those of the 24 SEZ isolates, commensurate with the greater variability of AGEs of the SEZ isolates studied here. Numerous methylation motifs were identified in the SEE and SEZ isolates, including several novel motifs, primarily among the SEZ isolates (Table 3-3). However, a single novel motif sequence (GGATGNND) was identified in an SEE isolate from Salado, Texas with a methylation frequency of 16%. The motif types that were identified in the SEZ isolates were mostly associated with methylation type m6A, although 1 motif (RGATCY) found in 2 SEZ isolates was the m4C methylation type. While many novel motifs were described, 12 of the 24 SEZ isolates had the motif (CATCC/GGATG) that has been identified in REBASE (B-4 Table), and this motif was also found in 13 of the 24 SEE isolates. These partnered motifs are both associated with type II restriction modification and methyltransferases in the SEZ H70 genome according to REBASE. Very little is known about the functions of these previously described restriction modification systems. These type II systems have been described as an immune system for the bacteria by protecting against invasion and modification by foreign DNA or bacteriophage by the presence of methylation at the motif sequence.[97,98,117] The presence of methylation at each occurrence of this motif sequence (CATCC/GGATG) was much higher (~ 97%; range [95% - 98%]) in the SEE isolates than in the SEZ (~ 68%; range [48% - 90%]). It is possible that this modification reflects a method of adaptation to protect against a bacteriophage that predominates in the respiratory tract of horses that targets SEE or SEZ, whereas strains of SEZ are far more diverse and adapt to many hosts or sites for opportunistic infection. It is also possible that this motif sequence (CATCC/GGATG) is an example of changes in the methyltransferase activity through acquisition of mobile genetic elements by horizontal gene transfer.[97,118,119] Although we observed more consistent methylation patterns in SEE isolates, this is likely to be explained in part by the consistency or maintenance in the acquired mobile genetic elements described for SEE, whereas different strains of SEZ are likely to able to acquire a greater variety of mobile genetic elements, thereby resulting in more variability of methyltransferase activity and methylation patterns.

By selecting homologous proteins with methylation present in all SEE (n = 24) but absent in all SEZ (n = 24) isolates, we identified 37 CDS. All 37 CDS in the SEE genomes had m6A type modification with the motif sequence, CTGCAG (Table 3-4). REBASE indicates that this modification and methylation motif is from a type II methyltransferase and restriction modification system that has been described in the strain SEE 4047. The presence of methylation at each occurrence of the motif sequence CTGCAG was highly prevalent (~ 95%) in all of the SEE genomes (B-4 Table). However, this particular motif sequence was not found among any of the SEZ isolates. The functions of these 37 CDS were assessed using ClueGO, and several GO terms and

KEGG pathways were noted (B-5 Table). Absence or alteration of methylation at motifs has been shown to alter gene expression in several different bacterial species.[99-101] Thus, we hypothesize that the absence of methylation at these homologous proteins in SEZ results in altered gene expression of these CDS resulting in functional differences. Several of the GO functions and KEGG pathways associated with these differentially methylated CDS have not been studied in either of the *Streptococcus equi* subspecies. These noted differences could be important to the pathogenesis and microbe-host interactions for SEE relative to SEZ. The exopeptidase activity was linked to 3 CDS (SEQ_0976, SEQ_1597, SEQ_1920) all of which had GO biological process term associated with peptidase activity (B-5 Table). Conceivably, this exopeptidase activity could contribute to host-specificity or pathogenesis of SEE infection in horses and warrants further investigation. Additionally, many of the differentially methylated genes were linked to proteins that functioned in transmembrane transport and transport of compounds, and further examination of whether these specific functions influence microbe-host interactions specifically in horses merits investigation. The quorum sensing pathway was connected with 3 CDS (SEQ_1918, SEQ_2009, SEQ_0435). Quorum sensing has been described in *S. pyogenes* to play a role in establishing disease in the host and evading the host's immune system,[120] and was recently described in SEZ to influence capsule polysaccharide production and biofilm formation.[121] Interestingly, an increased capsule depth has been noted in the SEE 4047 genome in comparison with SEZ H70 due to an inversion in genes involved in hyaluronate production,[10] but the observed methylation differences could also contribute to the thicker capsule of SEE which is thought to contribute to the pathogenesis of strangles. Although differences in bacterial function cannot be inferred based on methylation patterns, we believe our results indicate targets for further investigation regarding the host specificity and virulence of SEE.

Comparisons of methylation sites on homologous proteins in the 24 SEZ isolates but that were absent in the 24 SEE isolates demonstrated a greater degree of variability than was observed for those present in SEE but absent in SEZ. The methylation summaries of the SEZ isolates yielded far more inconsistent methylation motif sequences, and even absence of specific methylation motif sequences in 3 SEZ isolates (19-005, 19-051, 19-053; B-4 Table). Perhaps this variability in methylation contributes to the ability of SEZ to infect or colonize a wider number of hosts,[44-48] whereas a far more restricted methylation repertoire was found in the single-host pathogen SEE. Initially, 10 potential CDS were identified in the SEZ isolates where methylation was present but were absent on the homologous protein in the SEE isolates. However, upon further investigation it was determined that the presence of methylation did not occur at the same location within the CDS, did not have the same motif sequence, and sometimes differed in the type of methylation present (m6A or m4C; B-6 Table). Therefore, it is difficult to draw conclusions about the impact of the presence of methylation at these 10 homologous proteins.

This study has a number of limitations. The first limitation of this study is the incomplete genomes of the SEE and SEZ isolates. Although the use of PacBio sequencing allows for more contiguous draft genome (*i.e.,* fewer number of contigs) than using short-read technologies, gaps in the genome remained. The PacBio SMRT sequencing, however, enabled us to study the complete methylomes of these bacteria. Another limitation of our study was the necessity to utilize reference genomes for the characterization of the AGEs and methylation patterns of both SEE and SEZ genomes. The reference genome selected creates bias, and this was especially apparent for SEZ where we identified marked variability of the genomic elements both in the accessory genome and in the methylation patterns. However, the information derived from these 50 SEZ genomes and their methylation pattern will increase the publicly available genomic data. An important limitation was that we did not assess the function of the methylated CDS described on the homologous proteins in SEE and SEZ. Unfortunately, that work was beyond the scope of funding resources available to our laboratory. Further investigation of the function of these proteins is planned, and we hope our findings will stimulate other investigators to pursue these lines of inquiry as well. Even though function was not considered, the aim of this study was to characterize and describe the global methylation of SEE and SEZ. To the authors' knowledge this is also the first comparison of the global methylomes in SEE and SEZ isolates from the USA.

In this study we described the differences in the accessory genome (*i.e.,* elements that are not present in all isolates of the bacterial subspecies) and complete methylation patterns of SEE and SEZ isolates from Texas. We described that the majority AGEs found in all 50 SEE isolates were attributed to the mobile genetic elements (ICE and prophages) described in the reference SEE 4047. Fewer AGEs were found in all 50 SEZ isolates and were involved in lactose and sorbitol fermentation, but we also identified genes related to competence-stimulation that were not identified in SEE. Global methylomes were characterized for 24 SEE and 24 SEZ isolates, and more consistent patterns of methylation were noted in the SEE isolates in comparison with the SEZ isolates. We identified 19 novel methylation motifs primarily among the SEZ isolates. Importantly, methylation of homologous proteins in SEE but absent in SEZ was identified. Even though the effects of methylation at the homologous proteins of SEE and SEZ on bacterial function or host-specificity were not determined, further evaluation of these proteins is warranted to investigate the host-specificity and pathogenesis of SEE. Finally, we were unable to consistently identify sites of methylation in SEZ but absent in SEE on homologous proteins. Much remains to be learned about the impact of methylation on the differences in SEE and SEZ. In summary, the finding that comparison of the genomes and methylomes did not readily identify differences that explain the host-specificity of SEE indicates that it will be necessary to evaluate host-microbe interactions to unravel what drives specificity of SEE for infecting horses, using both *in vitro* and *in vivo* systems.

4. DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS*

4.1. Introduction

Streptococcus equi subsp. *equi* (SEE) is a host-specific bacterial pathogen that causes the infectious disease of horses known as strangles.[1,20,27,89,90] Infection with SEE occurs primarily in the upper respiratory tract, and is very contagious with a high rate of morbidity in naïve horses.[1] Typically, infection results in lethargy, pyrexia, swollen lymph nodes, guttural pouch empyema, and nasal discharge.[1,27] Other clinical signs of disease can be observed, including dissemination of infection to other organs and immune-mediated sequelae such as vasculitis and myositis.[1,90] Strangles is an ancient disease that is prevalent among horses worldwide.[10,89] The persistence of the disease appears to be attributable to the ability of SEE to survive in horses that are infected but do not show clinical signs. SEE cannot survive in the external environment for extended periods of time: SEE can persist approximately 2 days on surfaces outside its host,[25] and from 1 to 4 weeks in a wet environment, dependent upon the season.[26] There are no known biological or mechanical vectors of SEE,[1] and horses that have recovered from

*Reprinted with permission from "Differences in Genome, Methylome, and Transcriptome Do Not Differentiate Isolates of *Streptococcus equi* subsp. *equi* from Horses with Acute Clinical Signs from Isolates of Inapparent Carriers" by Morris, E.R.A., Boyle, A.G., Riihimäki, M., Aspán, A., Anis, E., Hillhouse, A.E., Ivanov, I., Bordin, A.I., Pringle, J. Cohen, N.D, 2021. PLoS ONE 16(6): e0252804. https://doi.org/10.1371/journal.pone.0252804. Copyright 2021 by CC-BY.

the disease usually develop prolonged immunity.[1,27] Consequently, the most likely source of spread and persistence of SEE is horses that appear healthy but shed SEE undetected (so-called **inapparent carrier horses**),[1,28,29] these carriers transmit SEE to susceptible horses, thereby perpetuating the disease in nature.[30,31]

Several host- and pathogen-associated adaptations have been suggested to give rise to the capacity for SEE to evade the immune system and persist within the host. The ability for some strains to be carried by apparently healthy horses has been attributed to the presence of chondroids (*i.e.*, concretions of inspissated pus) or empyema in the guttural pouches of infected horses recovered from strangles.[28,30] However, cases have been documented in which no clinical signs or vestiges of inflammation or niduses of infection (such as chondroids) were noted from clinically inapparent carriers of SEE.[32,34] Truncation of the N-terminus of the M-like protein (SeM) has been hypothesized to contribute to the ability of SEE to remain undetected in the host.[38] Another factor that has been proposed to contribute to inapparent carriage of SEE in horses is its equibactin locus (*eqbA – eqbN*), the novel iron acquisition element present on the integrative and conjugative element (ICE), ICE*Se2*.[29,39] More efficient iron acquisition is theorized to aid in the ability of SEE to better survive in the host without inducing clinical signs.[39] Despite these proposed characteristics of carrier isolates, none has been documented to be identified consistently among isolates of SEE from inapparent carriers. Consequently, it is unclear whether the inapparent carrier state of SEE is attributable to agent factors (adaptations to the host), host factors (such as immunity), or both.

Next-generation sequencing (NGS) technologies such as whole genome sequencing (WGS) or RNA sequencing (RNA-Seq) of SEE can be employed to investigate agent-associated adaptions within the bacterial genome or transcriptome that contribute to inapparent carriage. Using WGS, the bacterial genome can be defined by elements that make up the core or accessory genome.[94,122] Core genome elements are those found in the genomes of most isolates of the same bacterial species, whereas accessory genome elements (AGE) are elements that are not found in all isolates of the same bacterial species. Comparison of the AGE has been used to identify differences among isolates from the same bacterial species collected from different environments.[94] Additionally, using PacBio single molecule, real-time (SMRT) WGS allows for characterization of the methylation patterns of bacterial genomes.[96] Methylation of their DNA protects bacteria against bacteriophage or other foreign DNA; methyl groups sharing the same sequence motif as the bacteria's own DNA protect against enzymatic degradation, whereas the DNA lacking the same methylation is recognized as foreign by endonucleases that cleave at these unmethylated motifs.[97,98] Methylation can also alter gene expression and even alter virulence in some bacteria.[99-101] Methylated bacterial DNA is most commonly recognized as residues of N6-methyl-adenosine (m6A), N4-methyl-cytosine (m4C), or C5-methyl-cytosine (m5C).[97,98] In addition to the genome and the methylome, assessing the transcriptome through RNA-Seq can be used to characterize changes in gene expression that influence phenotype of the organism. For example, RNA-Seq revealed that differing regulation of gene expression resulted in a change in SEE colony morphology.[123] Thus, RNA-Seq might distinguish strains of SEE that result in inapparent carriage from isolates obtained from horses with acute clinical signs.

To our knowledge, however, potential differences in the genome, methylome, and transcriptome of inapparent carrier and acute clinical strains of SEE has not been investigated. Thus, we aimed to compare the AGE, methylomes, and transcriptomes of strains of SEE recovered from horses from within the same geographical regions that recovered from SEE without clinical signs (inapparent carriers) with strains of SEE from those with acute clinical signs of strangles, including some isolates collected by sequential sampling of individual horses. The purpose of these comparisons was to identify evidence of any adaptions of the pathogen to its host. We showed that there were no consistent differences between the 2 phenotypes of SEE strains for the AGE, methylome, or transcriptome that might explain persistence in the host. These findings indicate that pathogen-associated adaptions are highly improbable as an explanation for the ability of SEE to go undetected and persist within its host.

4.2. Materials and methods

4.2.1. *Streptococcus equi* subsp. *equi* isolates

Carrier and clinical SEE isolates from Pennsylvania (PA-USA) were provided by a co-author (AGB), and sequence data of Swedish isolates of SEE predominately from acute clinical cases and their isolates after progression to inapparent carriers were provided by 2 other co-authors (MR and JP) (Table 4-1). For the purposes of our study, inapparent carriers were defined as horses either recovered from strangles or exposed to strangles cases that were absent of clinical signs for ≥ 6 weeks prior to collection of the isolate. Swedish isolates of SEE (n = 14) were from a single outbreak at an individual farm in Sweden previously described[32] comprised of 8 isolates from inapparent carriers and 6 isolates from those with clinical disease; 5 horses from this herd contributed isolates during both acute disease and the inapparent carrier state. Isolates of SEE from PA-USA (n = 21) were from 11 inapparent carriers and 10 acute clinical cases located in a similar geographical area of the state, and isolates spanned different years (2014 to 2017).

Table 4-1. Description of the 14 SEE isolates from Sweden and the 21 SEE isolates from Pennsylvania.

Genome ID	Location	Status	Horse ID	Collection Source	Collection Date	Duration From Resolution of Clinical Signs	ST	SeM
470_007	Sweden	Carrier	H1	NL	11/11/2015	20 weeks	179	72
470_006	Sweden	Acute	H2	NL	5/21/2015	NA	179	72
470_003	Sweden	Carrier	H2	NL	8/26/2015	12 weeks	179	72
470_002	Sweden	Acute	H3	NL	5/21/2015	NA	179	72
489_007	Sweden	Carrier	H3	NL	11/11/2015	24 weeks	179	72
470_001	Sweden	Acute	H4	NL	5/21/2015	NA	179	72
489_004	Sweden	Acute	H5	NL	6/6/2015	NA	179	72
470_008	Sweden	Carrier	H5	NL	11/11/2015	20 weeks	179	72
489_006	Sweden	Carrier	H5	NL	11/11/2015	20 weeks	179	150[a]
489_003	Sweden	Acute	H7	NL	5/21/2015	NA	179	72
489_010	Sweden	Carrier	H7	GPL	3/3/2016	50 weeks	179	152[a]
489_002	Sweden	Acute	H8	NL	5/21/2015	NA	179	72
489_005	Sweden	Carrier	H8	NL	8/26/2015	12 weeks	179	72
489_009	Sweden	Carrier	H8	GPL	3/3/2016	50 weeks	179	151
20-080	PA	Carrier	PA1	GPL	7/15/2014	6 weeks	179	39
20-081	PA	Carrier	PA2	GPL	8/20/2014	12 weeks	179	39
20-082	PA	Carrier	PA3	GPL	11/26/2014	20 weeks	179	39
20-083	PA	Carrier	PA4	GPL	12/3/2014	20 weeks	179	39
20-084	PA	Carrier	PA5	NL	7/27/2016	16 weeks	179	28
20-085	PA	Carrier	PA6	NL	12/5/2016	None	179	147

Table 4-1. Continued.

Genome ID	Location	Status	Horse ID	Collection Source	Collection Date	Duration From Resolution of Clinical Signs	ST	SeM
20-086	PA	Carrier	PA7	NL	7/27/2016	8 weeks	179	39
20-087	PA	Carrier	PA8	NL	1/11/2017	8 weeks	179	224
20-088	PA	Carrier	PA9	GPL	4/4/2017	12 weeks	179	147
20-089	PA	Carrier	PA10	GPL	5/17/2017	None	179	225
20-090	PA	Carrier	PA11	GPL	8/8/2017	7 weeks	179	226
20-091	PA	Acute	PA12	GPL	6/6/2014	NA	179	28
20-092	PA	Acute	PA13	GPL	4/24/2014	NA	179	227
20-093	PA	Acute	PA14	NL	2/16/2017	NA	179	224
20-094	PA	Acute	PA15	NL	8/27/2014	NA	179	28
20-095	PA	Acute	PA16	NL	2/1/2016	NA	179	39
20-096	PA	Acute	PA17	NL	3/10/2014	NA	179	228
20-097	PA	Acute	PA18	NL	3/17/2014	NA	179	28
20-098	PA	Acute	PA19	NL	2/17/2016	NA	179	28
20-099	PA	Acute	PA20	NL	3/4/2016	NA	179	28
20-100	PA	Acute	PA21	NL	3/24/2016	NA	179	28

ST, Sequence type; SeM, M-like protein; PA, Pennsylvania; NL, Nasopharyngeal lavage; GPL, Guttural pouch lavage; NA, Not applicable.
[a]Truncation noted in SeM protein.

4.2.2. Bacterial DNA extraction and whole genome sequencing

The PA-USA SEE isolates were cultured overnight in 3 ml of Todd Hewitt broth (THB; HIMEDIA®, West Chester, PA, USA) in 5% CO_2 at 37°C. Following incubation overnight, bacterial isolates were centrifuged at 3,000 x g for 10 minutes to create a pellet. The supernatants were discarded, and DNA extractions were performed using the DNeasy® UltraClean® Microbial kit (Qiagen®, Hilden, Germany), following the manufacturer's instructions with some modifications. Briefly, the bacterial pellets were resuspended in 300 µl of PowerBead solution, and transferred into PowerBead tubes. Fifty µl of solution SL was added, and the PowerBead tubes were incubated at 70°C for

10 minutes, followed by horizontal vortexing for an additional 10 minutes. Then, the PowerBead tubes were centrifuged and the supernatants were transferred to new tubes. One hundred µl of solution IRS was added to the supernatants, incubated for 15 minutes at 4°C, and then centrifuged. The supernatants were transferred to another tube without disturbing the pellet, and 900 µl of solution SB was added. Seven hundred µl of this solution was transferred to MB spin column tubes, centrifuged, and, after the flow-through was discarded, this step was repeated. Additionally, 300 µl of solution CB was added to the columns and centrifuged. Another centrifuge step was performed to remove any excess fluid, and the MB spin columns were transferred to new collection tubes. Finally, 50 µl of the solution EB was added to the columns and centrifuged. The quality and concentration of the DNAs were assessed using a NanoDrop spectrophotometer (ND-1000, Thermo Fisher Scientific, Waltham, MA, USA), and sent to the Duke Center for Genomic and Computational Biology (GCB) for WGS using the PacBio Sequel platform.

The Swedish SEE isolates, cultured from horses during a strangles outbreak as described by Riihimäki *et al.*,[32] were retrieved from storage at -70°C, subculture was performed, and then grown overnight on 15-cm-diameter blood agar plates (SVA, Uppsala, Sweden) in 5% CO_2 at 37°C. DNAs were extracted by the Genomic-tip 100/G kit (GT) (Qiagen, Hilden, Germany) according to the manufacturer's protocol, but bacterial lysis was performed prior to extraction to obtain high molecular weight DNA. Briefly, SEE growth from the agar plates were harvested by a 10-µl loop into a 2-ml tube and thereafter lysed in 200 µl of 50 mM EDTA pH 8.0 supplemented with 20 µl

(100 mg/ml) lysozyme. After incubation on a thermomixer for 4 hours at 37ºC / 400 x g, 400 µl GT buffer B1 (provided by the manufacturer of the kit) and 20 µl proteinase K were added, and samples were mixed by inverting the tubes 10 times. This was followed by a further incubation for 4 hours, at 54ºC / 400 x g. Samples were frozen at -80ºC overnight, flash-thawed at 50ºC, and 300 µl of GT buffer B2 was added. Again, samples were mixed by inverting the tubes 10 times. Five µl of RNase was added and after 10 minutes at room temperature, samples were mixed for 30 minutes at 50ºC / 400 x g, before DNA extraction. After DNA extraction, the DNA quality was assessed using a NanoDrop spectrophotometer (ND-8000, Thermo Fisher Scientific, Waltham, MA, USA), and concentrations were determined using a Qubit® 2.0 fluorometer (Invitrogen, Carlsbad, CA, USA). The DNA from the 14 Swedish SEE isolates were then sent to the SciLifeLab (https://www.scilifelab.se/) for PacBio sequencing.

4.2.3. Bacterial RNA extraction and RNA sequencing

Carrier and clinical PA-USA SEE isolates were grown in THB for 4 hours (exponential phase growth) at 37°C in 5% CO_2. Following the 4-hour incubation, liquid cultures were centrifuged at 3,000 x g for 10 minutes to pellet the bacterium and the supernatants were discarded. The bacterial RNAs were then extracted using the RiboPure™ RNA Purification kit (Ambion® RiboPure™-Bacteria Kit; Invitrogen™, Carlsbad, CA, USA) following the manufacturer's instructions. Briefly, the SEE pellets were resuspended in 350 µl of the RNAWIZ solution, and then transferred to tubes with Zirconia beads. The tubes were placed on a horizontal vortex adaptor, beat for 10 minutes at maximum speed, and then centrifuged at 13,000 x g for 5 minutes at 4°C.

The supernatants containing the lysed bacteria were transferred to fresh tubes, 0.2 volumes of chloroform were added, and samples were incubated for 10 minutes at room temperature. To separate the organic and aqueous phases, tubes were centrifuged for 5 minutes at 4°C. The aqueous phases were transferred to new tubes, 0.5 volumes of 100% ethanol were added, mixed thoroughly, and transferred to filter cartridges in 2-ml tubes. The filter cartridge tubes were then centrifuged for 1 minute, the flow-through discarded, and the filters were washed by the addition of 700 µl of Wash Solution 1. A second and third wash steps were performed with the addition of Wash Solution 2/3. After the third wash step, the filter cartridges were transferred to new tubes. Finally, the RNA was eluted by 50 µl of Elution Solution, and a DNase treatment was performed. The quality and purity of the RNAs were assessed using the NanoDrop (ND-1000, Thermo Fisher Scientific, Waltham, MA, USA).

At the Texas A&M Institute for Genome Sciences and Society (TIGSS) molecular genomics laboratory, RNA extracted from the 21 PA-USA SEE isolates were quantified using the Qubit fluorometric RNA (Thermo Fisher Scientific, Waltham, MA, USA) assay for normalization prior to library preparation. RNA libraries were prepared using the Stranded Total RNA Preparation kit (Illumina$^{©}$, San Diego, CA, USA) following the manufacturer's instructions, in which each isolate received a unique barcode. The 21 isolates were pooled, and RNA-Seq was performed on the NovaSeq 6000 (Illumina$^{©}$, San Diego, CA, USA) instrument that generated 150-base-pair, paired-end sequences. The sequencing run produced approximately 6 million reads per sample and resulted in ~200 X coverage for each sample.

4.2.4. Bioinformatic analysis

Following WGS of the PA-USA and Swedish isolates, the Texas A&M High Performance Research Computing (HPRC) clusters were used to assemble genomes *de novo* using CANU (v1.7),[103] with the parameters of increased coverage (*corOutCoverage* = 100) and increased assembly sensitivity (*corMhapSensitivity* = high). Assembled genomes were confirmed to be SEE through the ribosomal multilocus sequence types database,[53] and StrainSeeker.[124] The ST- and SeM-type of each of the assembled genomes of SEE were determined using the PubMLST *Streptococcus zooepidemicus* database.[43,54] Then, assembled genomes were annotated using RASTtk (v2.0)[104] via the web-based server. Following annotation, the annotated genomes were inputted into Spine (v0.3.2)[94] to define the core genome (*i.e.,* elements found in all genomes) of SEE. Using the core genome output from Spine, the AGE (*i.e.,* elements found present in some genomes but absent from others) were identified using AGEnt (v0.3.1).[94] Finally, ClustAGE (v0.8)[105] was implemented to identify and group the AGE that differ within the carrier and clinical SEE isolates. A graphical representation of clustered AGE for each individual genome was generated with the ClustAGE plot (http://vfsmspineagent.fsm.northwestern.edu/cgi-bin/clustage_plot.cgi). AGE were only included if ≥ 95% of the protein was identified. Comparisons of the AGE of carrier and clinical SEE were performed using custom R scripts (E-1 Appendix). We conducted separate AGE analyses for SEE isolates from Sweden and PA-USA to avoid potential confounding effects by geographical location. A phylogenetic tree was built to assess the relatedness of the Swedish SEE isolates using PATRIC (v3.6.9) with default parameters.[125] Multiple sequence alignment of the SeM nucleotide sequences was performed using Clustal Omega (v1.2.4) at EMBL-EBI.[126,127]

The complete methylation profiles of carrier and clinical SEE genomes were characterized with the BaseMod (https://github.com/ben-lerch/BaseMod-3.0) pipeline in the PacBio SMRT Link (v8.0) command line tools. Briefly, pbmm2 was used to align the raw sequence read BAM files to the reference genome (SEE 4047). Using the aligned BAM file outputs, the kineticTools function *ipdSummary* was implemented to generate a GFF and CSV files with base-modification information. Next, the MotifMaker *find* function was used to generate a second set of CSV files with identified consensus motifs. Finally, the execution of the MotifMaker *reprocess* function generated GFF files with all the modifications that were associated with motifs. Using R (v4.0.3), the motif GFF files were filtered based on the presence of a known methylation types (m4C or m6A), and a having QV score (a quality score for the detection event) of ≥ 30. The filtered GFF files of carrier and clinical SEE genomes were annotated by the SEE 4047 reference genome with the BedTools (v2.29.2)[109] *annotate* function. The annotated outputs for both carrier and clinical SEE were then compared by looking for the presence or absence of methylation on proteins throughout the genomes using custom R scripts (E-1 Appendix). Identified motifs were then compared to the SEE 4047 genome using the Restriction Enzyme Database (REBASE).[110]

Following sequencing of the RNA of PA-USA SEE isolates at TIGSS, using the HPRC clusters raw RNA reads had their quality checked using FastQC (v0.11.6; www.bioinformatics.babraham.ac.uk/projects/fastqc/), and low quality reads were trimmed using Trimmomatic (v0.36).[70] These filtered reads were then aligned and quantified against the reference genome SEE 4047 using Salmon (v1.3.0).[128] The transcriptomes of all carrier SEE isolates were compared to clinical SEE isolates with edgeR (v3.30.3)[129] to identify any significantly (false discovery rate [FDR] ≤ 0.05) differently expressed genes with a $\log_2$-fold change (logFC) of ≤ -1 or ≥ 1 using a quasi-likelihood negative binomial generalized log-linear model (E-1 Appendix).[130]

4.2.5. Accession numbers

Genomes and raw sequence files were submitted to NCBI's GenBank and Sequence Read Archive under BioProject PRJNA704656. The RNA-Seq transcripts were deposited to NCBI's Gene Expression Omnibus (GEO) and are accessible through the GEO Series accession number GSE167862 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167862).[131] The specific accession numbers for each genome can be found in Supplementary Table 4-1 (D-1 Table).

4.3. Results

Initially, the WGS data were used to define the AGE of the SEE isolates. The AGE were examined to identify genetic elements that differed between SEE isolates collected from acute and inapparent carrier cases. No consistent or significant differences in AGE were observed between the carrier (n = 8) and acute clinical (n = 6) isolates of SEE from the Swedish outbreak (Fig 4-1). Similarly, there were no differences identified between AGE of the carrier (n = 11) and acute (n = 10) SEE isolates from PA-USA (Fig 4-2). Many components identified in the AGE were associated with acquired genetic elements. Markedly fewer AGE elements were identified in the Swedish SEE isolates (D-2 Table) than in the PA-USA SEE isolates (D-3 Table). The phylogenetic assessment of the Swedish SEE isolates demonstrated that there were minor genomic differences between isolates recovered from either clinical or carrier state from the same individual horses, but these adaptations were not consistent among individuals (D-1 Fig). For example, 2 carrier isolates (489_006 [H5], 489_010 [H7]) from Sweden were noted to have a truncated SeM protein (Table 4-1). Although neither horse had this truncation identified during acute clinical infection, in 1 horse (H5) the truncated isolate was collected via nasopharyngeal lavage simultaneously with a non-truncated isolate. These truncations were found at the beginning of the SeM protein, but ended at nucleotide base 318 and 333 in isolate 489_006 and 489_010, respectively (F-1 Appendix). Furthermore, no other truncation in the SeM proteins were identified in the remaining isolates of SEE from Sweden or PA-USA.

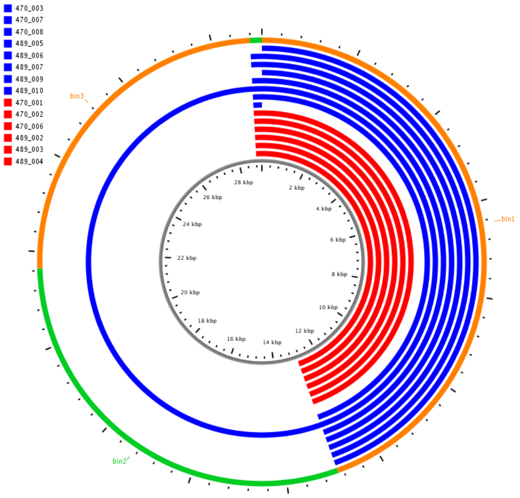


Fig 4-1. Comparison of accessory genome elements (AGE) of inapparent carrier SEE (n = 8), and acute clinical SEE (n = 6) genomes from Sweden. The outer ring shows the ClustAGE bins that are ≥ 200 base-pairs in size these are ordered clockwise from the largest bin to the smallest bin, and are differentiated by orange and green to define bin borders. The concentric inner bands show the distribution of AGE within each individual isolate. Bands that are blue represents inapparent carrier isolates, and bands that are red represent acute clinical isolates. The central ruler of the figure indicates the cumulative size of the AGE in kilobases.

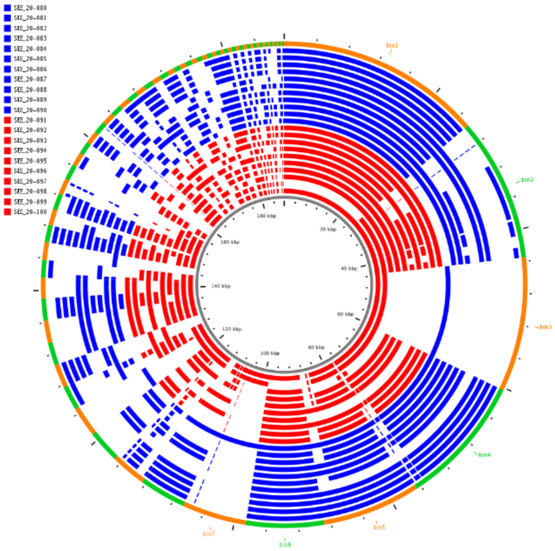


Fig 4-2. Comparison of accessory genome elements (AGE) of inapparent carrier SEE (n = 11), and acute clinical SEE (n = 10) genomes from Pennsylvania**.** The outer ring shows the ClustAGE bins that are ≥ 200 base-pairs in size these are ordered clockwise from the largest bin to the smallest bin, and are differentiated by orange and green to define bin borders. The concentric inner bands show the distribution of AGE within each individual isolate. Bands that are blue represents inapparent carrier isolates, and bands that are red represent acute clinical isolates. The central ruler of the figure indicates the cumulative size of the AGE in kilobases.

Because some methylation events have been described to influence gene expression in prokaryotes,[99-101] we performed additional characterization of these bacterial genomes by examining the methylomes of the carrier and acute clinical strains of SEE. As done for the AGE sequence data, separate analyses of the global methylation patterns determined from PacBio WGS were performed for the Swedish and PA-USA isolates of SEE. In both Swedish and PA-USA SEE isolates, no differences in methylation patterns were observed that consistently differed between the carrier and acute clinical isolates of SEE (Figs 4-3A and 4-4A). Using REBASE, we performed comparisons of the identified motifs to those in the reference genome, SEE 4047, in which REBASE used with the GenBank data for SEE 4047 to predict restriction enzyme and DNA methyltransferase genes.[110] We identified novel methylation motifs from the complete methylomes of the Swedish SEE isolates. The first new motif (ANNNGANCGNNNAATNNT) was associated with the m6A modification found in a clinical and a carrier SEE isolate, 470_001 and 470_008, respectively (Table 4-2). The second new motif (DNRTGCAGB) was observed in 4 carrier SEE isolates at 3 locations with the m6A type modification (Fig 4-3B); although we found other sites with this motif, we were unable to determine whether they were either m6A or m4C methylation (*i.e.,* Modified Base; Table 4-2). The most common motif seen among all SEE isolates regardless of location was CTGCAG (Table 4-2), which was associated with a type II restriction enzyme and methyltransferase according to REBASE. We also observed that the motif CATCC was found among all Swedish SEE isolates, but only in 12 of the 21 PA-USA SEE isolates (Table 4-2). Specific methylation sites that occurred in at least half the isolates for either disease status were considered. Six sites were identified that fit this criterion in the SEE isolates from Sweden; 2 of these were identified in acute clinical isolates and the remaining 4 were identified in carrier isolates (Fig 4-3A, D-4 Table). Within the genomes of the PA-USA SEE isolates, only 3 sites of methylation occurred in half of the carrier group (Fig 4-4A), and these sites all had m4C type modification with an unknown motif (Fig 4-4B, D-5 Table).

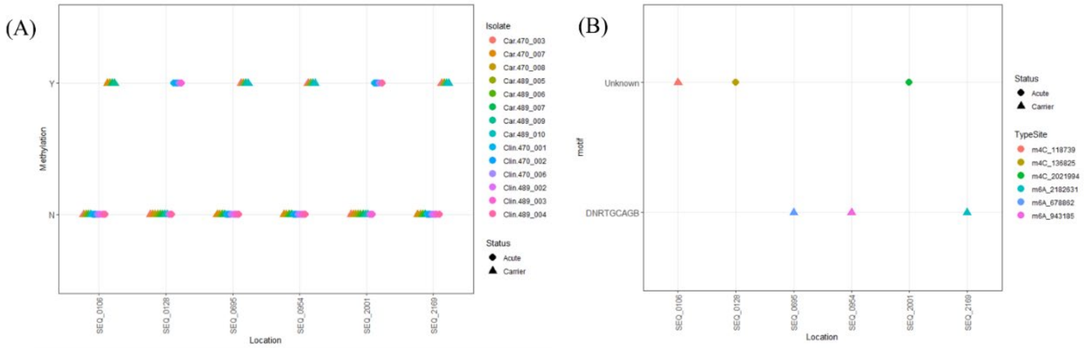


Fig 4-3. Methylation locations and motifs from SEE isolates from Sweden. (A) Depiction of whether methylation occurred at a specified genomic location. Genomic locations are indicated along the x-axis, and whether methylation occurred is indicated on the y-axis as yes (Y) or no (N), by SEE isolates. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates. (B) Sites of methylation (x-axis), by the methylation motif (y-axis). The type of methylation and exact position in the genome are indicated by different colors. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates.

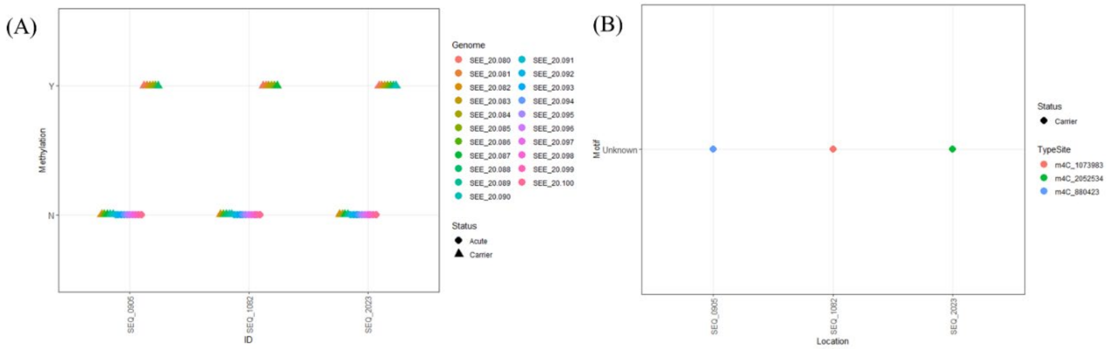


Fig 4-4. Methylation locations and motifs from SEE isolates from Pennsylvania. (A) Depiction of whether methylation occurred at a specified genomic location. Genomic locations are indicated along the x-axis, and whether methylation occurred is indicated on the y-axis as yes (Y) or no (N) by SEE isolates. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates. (B) Sites of methylation (x-axis), by the methylation motif (y-axis). The type of methylation and exact position in the genome correspond to the different colors. Circles represent acute clinical isolates, and triangles represent inapparent carrier isolates.

Table 4-2. The summary of the methylation motif sequences, modification types, and modification percentage for all study SEE isolates from Sweden and Pennsylvania.

Genome ID	Location	Status	Motif Sequence	Modification Type	Percent Modification
470_001	Sweden	Acute	ANNNGANCGNNNAATNNT	m6A	0.86
470_001	Sweden	Acute	CATCC	m6A	0.99
470_001	Sweden	Acute	CTGCAG	m6A	0.97
470_002	Sweden	Acute	CATCC	m6A	0.98
470_002	Sweden	Acute	CTGCAG	m6A	0.97
470_003	Sweden	Carrier	CATCC	m6A	0.98
470_003	Sweden	Carrier	CTGCAG	m6A	0.97
470_003	Sweden	Carrier	DNRTGCAGB	Modified Base	0.42
470_006	Sweden	Acute	CATCC	m6A	0.98
470_006	Sweden	Acute	CTGCAG	m6A	0.97
470_007	Sweden	Carrier	CATCC	m6A	0.99
470_007	Sweden	Carrier	CTGCAG	m6A	0.97
470_008	Sweden	Carrier	ANNNGANCGNNNAATNNT	m6A	0.85
470_008	Sweden	Carrier	CATCC	m6A	0.98
470_008	Sweden	Carrier	CTGCAG	m6A	0.97
489_001	Sweden	Acute	CATCC	m6A	0.99
489_001	Sweden	Acute	CTGCAG	m6A	0.97
489_002	Sweden	Acute	CATCC	m6A	0.99
489_002	Sweden	Acute	CTGCAG	m6A	0.97
489_003	Sweden	Acute	CATCC	m6A	0.99
489_003	Sweden	Acute	CTGCAG	m6A	0.97
489_004	Sweden	Acute	CATCC	m6A	0.99
489_004	Sweden	Acute	CTGCAG	m6A	0.97
489_005	Sweden	Carrier	CATCC	m6A	0.99
489_005	Sweden	Carrier	CTGCAG	m6A	0.97
489_005	Sweden	Carrier	DNRTGCAGB	Modified Base	0.51
489_006	Sweden	Carrier	CATCC	m6A	0.99
489_006	Sweden	Carrier	CTGCAG	m6A	0.97
489_007	Sweden	Carrier	CATCC	m6A	0.99
489_007	Sweden	Carrier	CTGCAG	m6A	0.97
489_009	Sweden	Carrier	CATCC	m6A	0.99
489_009	Sweden	Carrier	CTGCAG	m6A	0.97
489_009	Sweden	Carrier	DNRTGCAGB	Modified Base	0.57
489_010	Sweden	Carrier	CATCC	m6A	0.98
489_010	Sweden	Carrier	CTGCAG	m6A	0.96
489_010	Sweden	Carrier	DNRTGCAGB	Modified Base	0.49

Table 4-2. Continued.

Genome ID	Location	Status	Motif Sequence	Modification Type	Percent Modification
20-080	PA	Carrier	CTGCAG	m6A	0.95
20-081	PA	Carrier	CTGCAG	m6A	0.95
20-082	PA	Carrier	CTGCAG	m6A	0.95
20-083	PA	Carrier	CTGCAG	m6A	0.95
20-084	PA	Carrier	CATCC	m6A	0.98
20-084	PA	Carrier	CTGCAG	m6A	0.97
20-085	PA	Carrier	CATCC	m6A	0.98
20-085	PA	Carrier	CTGCAG	m6A	0.97
20-086	PA	Carrier	CTGCAG	m6A	0.95
20-087	PA	Carrier	CTGCAG	m6A	0.95
20-088	PA	Carrier	CATCC	m6A	0.98
20-088	PA	Carrier	CTGCAG	m6A	0.97
20-089	PA	Carrier	CATCC	m6A	0.98
20-089	PA	Carrier	CTGCAG	m6A	0.97
20-090	PA	Carrier	CTGCAG	m6A	0.95
20-091	PA	Acute	CATCC	m6A	0.97
20-091	PA	Acute	CTGCAG	m6A	0.95
20-092	PA	Acute	CATCC	m6A	0.97
20-092	PA	Acute	CTGCAG	m6A	0.96
20-093	PA	Acute	CTGCAG	m6A	0.95
20-094	PA	Acute	CATCC	m6A	0.98
20-094	PA	Acute	CTGCAG	m6A	0.97
20-095	PA	Acute	CTGCAG	m6A	0.95
20-096	PA	Acute	CATCC	m6A	0.98
20-096	PA	Acute	CTGCAG	m6A	0.97
20-097	PA	Acute	CATCC	m6A	0.98
20-097	PA	Acute	CTGCAG	m6A	0.97
20-098	PA	Acute	CATCC	m6A	0.98
20-098	PA	Acute	CTGCAG	m6A	0.97
20-099	PA	Acute	CATCC	m6A	0.98
20-099	PA	Acute	CTGCAG	m6A	0.97
20-099	PA	Acute	GGATGH	m6A	0.21
20-100	PA	Acute	CATCC	m6A	0.98
20-100	PA	Acute	CTGCAG	m6A	0.97

PA, Pennsylvania; m6A, N6-methyl-adenosine.

To assess differences in gene expression determined by RNA-Seq between acute clinical and inapparent carrier strains of SEE from PA-USA, we used a similar untargeted approach as was used to analyze the SEE isolate exhibiting phenotype switching among colonies.[123] Our differential gene expression analysis with edgeR did not identify any genes that were significantly (FDR ≤ 0.05, logFC ≤ -1 or ≥ 1) differentially expressed (Fig 4-5). Two genes (SEQ_0823, SEQ_0834) that were closest to being significant and that had a logFC < -1 and an FDR of 0.055 (D-2 Fig, D-6 Table) were associated with phage elements of the SEE genome found in the prophage φSeq2 in SEE 4047, but these elements have not been further studied.

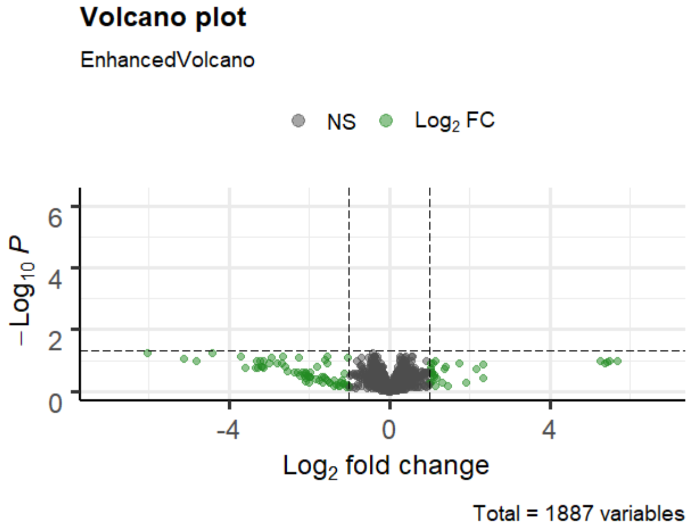


Fig 4-5. Volcano plot of Pennsylvania SEE RNA-Seq genes counts. The $\log_2$ fold-change (logFC) is represented along the x-axis, and the $\log_{10}$-transformed false discovery rate (FDR) is represented along the y-axis. Gray points represent genes that were not identified as significantly differentially expressed (FDR ≤ 0.05), and green points represent genes whose expression had a logFC ≤ -1 or ≥ 1. No genes met the criteria for interest of having an FDR ≤ 0.05 and a logFC ≤ -1 or ≥ 1.

4.4. Discussion

Comparing the AGE, methylomes, and transcriptomes of SEE isolates from horses either with acute clinical signs or that were inapparently-infected and that were derived from outbreaks in 2 different continents, we could not identify significant differences between strains of SEE from acute clinical and inapparent carrier strains. The genomic analysis of the AGE of the PA-USA and Swedish strains were considered separately to avoid confounding effects of geographical origin on any observed genomic differences between acute and carrier strains, because geographical clustering has previously been identified.[29,34] We defined the core genome for all isolates within a region, which differs from the approach taken in another study where the core genome was delineated by removing prophages, and the ICEs from the SEE 4047 genome, and any regions of other SEE genomes > 200 base-pairs that did not match the core genome were considered as part of the accessory genome.[29] Harris *et al.* demonstrated that the prophages φSeq2 – 4, and ICE*Se1* and ICE*Se2* were highly conserved among SEE isolates.[29] This finding led us to adapt our AGE definition to include all elements found present in some genomes but absent from others, therefore allowing these prophages and ICEs sequences to be considered as elements of the core genome.

The Swedish SEE isolates (n = 14) were collected throughout a single outbreak that occurred among Icelandic horses as previously reported.[32] Of note, 5 of the horses had isolates from both acute disease and after becoming inapparent carriers. Our 2nd population of SEE isolates (n = 21) were collected from a single region of PA-USA, but were not from a single outbreak. Among both sets of isolates, we observed no consistent differences in the AGE from isolates collected from inapparent carrier horses when compared to those collected from individuals exhibiting acute clinical signs. These finding are consistent with those previously described, despite the aforementioned difference between studies in how the accessory genome was defined.[29] Many of the observed AGE from these isolates were related to phages, ICEs, and hypothetical proteins that have not been characterized or identified in the reference genome SEE 4047 (D-3 Table). These findings are important because they indicate that strains of SEE from inapparent carriers cannot be distinguished by the presence or absence of any specific genes. However, the AGE from SEE isolates collected from different regions of the world do differ,[29] regardless of disease status, which illustrates the importance of accounting for geographical effects when comparing genomes of SEE. It should be noted that most (13/19) of the carrier isolates from this study - both from PA-USA (6/11 carriers) and Sweden (7/8 carriers)[32,132] - were collected from individuals that were healthy and lacked evidence of abnormalities including chondroids in their guttural pouches, findings that have been identified in some inapparent carriers.[28,30] Although other studies have described potential pathogen-associated genetic changes that could result in a carrier state of SEE,[29,38,39] we did not identify truncation of the SeM protein in any of the carrier strains from PA-USA (n = 11) and only truncation in 2 of 8 Swedish carrier strains (489_006, 489_010) as previously described,[32] and the equibactin locus was found in all SEE strains from both locations (*i.e.*, acute clinical and carrier strains). Although the reasons for the discrepancy between our findings and prior studies is unknown, it is possibly attributable to either geographical differences or clinical phenotypic differences (*e.g.,* isolates obtained from horses with chondroids or guttural pouch empyema[38]) between isolates in our study and isolates from previous studies. The variations within the Swedish SEE isolates collected from the same horse, such as the truncated SeM protein observed in 2 carrier isolates (D-1 Fig), likely reflect adaptation of SEE to its host over time. These variations were noted between the 2 disease states (acute clinical or inapparent carrier) and source (guttural pouch or nasopharyngeal lavage; Table 4-1; D-1 Fig). Nevertheless, these pathogen-adaptions were not consistent among the SEE isolates from Sweden from within the same horse or from other inapparent carrier isolates of SEE from Sweden or PA-USA.

PacBio WGS results also were used to describe the methylome of the PA-USA and Swedish SEE isolates. Methylation in prokaryotes has primarily been described as a mechanism of defense against invading bacteriophages and other foreign DNA.[97,98] Lack of methylation at a particular motif that occurs throughout the genome has been shown to produce modifications in the gene expression of microbes,[99,101] even contributing to the virulence of some pathogens.[99] To the authors' knowledge, this is the first detailed comparison of the global methylomes of inapparent carrier and acute clinical isolates of SEE. Despite the more comprehensive scrutiny of genetic elements of our approach, we failed to identify any changes in methylation that differentiated between inapparent carrier and acute clinical isolates of SEE (Figs 4-3 and 4-4). As we observed for the AGE analyses, methylation patterns differed between the geographical regions. Among the methylation observed in the SEE strains from Sweden, we identified 2 novel motifs that have not been described previously in SEE isolates. The lower frequency of methylation events associated with the novel motif (DNRTGCAGB) increases the likelihood that the absence of methylation at this motif could influence the gene expression in these isolates. Decreased methylation of target motifs has been reported to inhibit *Streptococcus pyogenes* from surviving in human neutrophils and to reduce expression of genes involved in immune evasion and adherence,[99] to alter the ability of *Borrelia burgdorferi* to colonize the host,[101] and to alter the expression of genes associated with metabolic pathways of *Mycobacterium tuberculosis.*[100] To evaluate sites with higher prevalence of methylation, we considered sites in which methylation occurred in at least half the isolates from either the carrier group or the acute clinical disease group. Six methylation sites were identified as occurring in at least half of the Swedish isolates of SEE, and only 3 methylation sites were identified in at least half of the PA-USA isolates of SEE (Figs 4-3 and 4-4). Of the 9 sites that had more frequent methylation in both geographical locations of SEE isolates, none were common among isolates from both locations. This further demonstrates geographical differences in the methylome of SEE, but methylation patterns did not differ between the 2 different phenotypes. Little can be inferred about the biological effects of the observed differences in methylation between geographical areas without further investigations, but our objective was to determine whether methylation patterns differed consistently between isolates of SEE from inapparent carriers and acute clinical disease.

RNA-Seq has been previously used to assess gene expression in SEE isolates.[29,123] Changes in transcription identified using untargeted RNA-Seq of an SEE isolate were associated with a difference in the phenotype of colonies of the isolate.[123] A targeted approach to gene expression (*viz.*, quantitative PCR) was used to evaluate gene expression of the *has* operon which regulated levels of hyaluronic acid capsule expression in SEE isolates where deletions in the *has* operon were identified.[29] We performed untargeted RNA-Seq on inapparent carrier and acute clinical SEE isolates from the same region of PA-USA. No significantly (FDR ≤ 0.05) differentially expressed genes were identified between the acute and carrier SEE isolates (D-6 Table). We did identify, however, 2 CDS, SEQ_0823 and SEQ_0843, that were closest to fitting our defined criteria for significance and magnitude of effect, and both were associated with mobile genetic elements found in the prophage φSeq2 of the SEE 4047 genome. For both genes, the magnitude of expression was only highly elevated in 3/10 of the acute SEE isolates from PA-USA (D-2 Fig). Besides being identified as a putative phage portal protein (SEQ_0823) and putative phage tail protein (SEQ_0843), not much is known about either of these genes. Homologs proteins of SEQ_0823 were found in *Streptococcus pyogenes*, *Streptococcus dysagalactiae*, and *Streptococcus agalactiae* with a similarity of 96%, and for SEQ_0843 in *Streptococcus equi* subsp. *zooepidemicus* with a similarity of 100%.

This study has a number of limitations. First, the definition of inapparent carriers can be highly variable.[28,31,32,34] The inapparent carrier strains of SEE from Sweden were recovered from horses between 12 and 50 weeks after resolution of their clinical signs, whereas the PA-USA inapparent carrier horses were collected between 6 and 20 weeks after resolution of clinical signs, or had no clinical signs observed (Table 4-1). Nevertheless, we found no evidence from horses from either location of any consistent differences in the genome or methylome of these isolates indicating any specific adaptations to the host environment, even among the Swedish strains representing isolates of acute disease and inapparent carrier phenotypes in the same animal. The PA-USA isolates were not all from the same outbreak or the same year, but even among isolates from within farm and year, there were no consistencies observed. Moreover, none of the PA-USA samples were derived from the same animal, and the number of isolates studied was modest. Nonetheless, this is the first comprehensive analysis of the genomes, methylomes, and transcriptomes of inapparent carrier and acute clinical strains of SEE. We only had isolates of PA-USA available to evaluate using RNA-Seq. Another limitation of the RNA-Seq approach was that SEE were grown in liquid media and this might not reflect transcription within the host.[133] However, an effective approach for studying transcription of SEE within its host's cells remains limited; to the authors' knowledge, this study provides comparisons of untargeted gene expression of carrier and acute SEE isolates from the USA that has not been previously available.

The most important finding from this study is that we failed to identify any consistent or specific pathogen-associated changes between the inapparent carrier strains and the acute clinical disease strains of SEE using a few NGS techniques. Although genomic differences were observed between the 2 geographical regions, no changes in the genome, methylome, or transcriptome were identified that could be interpreted as reflecting a consistent mechanism of adaptation of SEE to the host resulting in inapparent carriage. These findings indicate that host-associated differences are a more likely explanation of the bacterium's ability to persist in horses without resulting in either clinical signs or a robust immune response (*i.e.,* the presentation of clinical disease).[37] Thus, further evaluation of host immune responses to SEE is warranted to elucidate how to identify and eliminate chronic carriers of SEE to control and prevent this important equine infectious disease.

5. SUMMARY AND FUTURE DIRECTIONS

5.1. Summary of results

During late 2017 and early 2018, an outbreak of strangles occurred in the herd of horses maintained for teaching and research at the College of Veterinary Medicine & Biomedical Sciences (CVMBS), Texas A&M University. The molecular epidemiological studies of *Streptococcus equi* subspecies *equi* (SEE) that comprise this dissertation stem from this event. We initially wanted to understand the source of this strangles outbreak, and how it spread so rapidly among the herd. The outbreak occurred over 6 months after the conclusion of a study of a strangles vaccine that included experimental infection of horses with SEE. From our research, we realized there was a scarcity of genomic data for SEE strains from the United States (US). Therefore, we expanded the scope of our investigation beyond the initial investigation of an outbreak. Our new aims included learning how variable SEE isolates were between regions, years, and individual outbreaks or years. Our selection of US SEE isolates was a sample of convenience from our laboratory repository, representing various regions of Texas and central Kentucky. From this initial study (Chapter 2), we learned the CVMBS strangles outbreak was the result of an undetected silent carrier in the herd that infected a susceptible yearling in the teaching herd. Secondly, while only minimal genetic variation of isolates of SEE within the outbreak was observed, one mutation observed in the CVMBS outbreak was a single nucleotide polymorphism (SNP) of the gene encoding the penicillin binding protein 2x (*pbp2x*). We observed that the colony morphology differed between the SEE isolates with and without this *pbp2x* SNP: colonies with the SNP were circular in form, raised in elevation, had entire margins, and white in color for their morphology, while colonies had various other morphologies; there was no overlap in morphologies between isolates with and without the SNP. Despite the difference in colony morphologies, we suspect this SNP was simply a random mutation event because we found no evidence of penicillin resistance associated with the presence of the *pbp2x* SNP, nor did any of the horses in the outbreak receive penicillin. By comparing the genomes of isolates of SEE from Kentucky and Texas with publicly-available genomes from other countries, we also learned that several US SEE isolates were clustered genetically with those from other countries outside the US (predominately Europe), demonstrating the international transmission of this horse-restricted pathogen. Notably, isolates from both Kentucky and Texas were represented among the US isolates that clustered with some of the European SEE isolates.

Our initial study provided us with knowledge of SEE genomics, and demonstrated the power of WGS for addressing questions regarding the epidemiology of SEE. We were stimulated by our findings to address 2 other important questions regarding the molecular epidemiology of US SEE strains. First, we wanted to determine if we could understand the host-specificity of SEE by comparing genomes of SEE to the genomes of multi-host SEZ, from which SEE is believed to have originated (Chapter 3). Second, we wanted to understand how inapparent carriers of SEE arise (Chapter 4). Specifically, we wanted to understand whether the carrier state arose from adaptions of

SEE to the host (pathogen driven), responses of the host to infection with SEE (host driven), or were both pathogen and host play a role?

In Chapter 3, we substantiated previously described differences between the genomes of a single strain each SEE and SEZ[10] and that the acquired mobile genetic elements are the most likely factors explaining host-specify of SEE. Furthermore, we characterized for the first time the global methylome and its differences of SEE and SEZ using the whole genome sequencing (WGS) technology developed by PacBio® known single molecule, real-time (SMRT) sequencing.[96] We identified genes shared by SEE and SEZ that were differentially methylated that represent targets for further study as determinants of host-specificity of SEE or pathogenesis of SEE. Specifically, proteins with the presence of methylation in SEE but absence in SEZ were associated with gene ontology functions of exopeptidase activity and KEGG pathways such as quorum sensing. Future studies should address the repeatability of these findings and document differences in gene expression and function associated with these specific biological processes. If these findings are consistent and represent functional changes, further investigation of their role in host-specificity and virulence should be investigated.

In Chapter 4, we revealed no consistent or statistically significant differences in either the genome or methylome of clinical or carrier strains of SEE from 2 countries. Although some genetic changes were noted between carrier and clinical SEE isolates collected from the same host over time, none of these mutations were consistent, indicating that specific genes or gene regulators characterize carrier strains. Moreover, no significant differences were identified in the transcriptome of carrier strains versus clinical strains from the US. Collectively, these findings indicate that host responses to SEE are more likely to contribute to the carrier state.

5.2. Future directions

5.2.1. Molecular epidemiology of other regions of the US

While sequencing SEE isolates from Texas and Kentucky is a step in the right direction to understanding the epidemiology of SEE both within the US and in a more global context by comparing results to publicly-available SEE genomes, many other regions of the US also have large populations of horses, such as California, Florida, Oklahoma, Pennsylvania, and New York. In addition to Texas and Kentucky, these areas will also play an important role in the spread of SEE, including internationally. Through collaborations with state veterinary diagnostic laboratories like the Texas A&M Veterinary Medical Diagnostic Laboratory (TVMDL), SEE isolates could be collected to create a repository of US isolates. Application of WGS and analyses with ParSnp[72] (described in Chapter 2) to study the molecular epidemiology of SEE could be used as a pipeline for analysis of isolates of SEE both within the US and world-wide.

5.2.2. Potential to improve SEE diagnostics

For the work comprising this dissertation, 100 SEE and SEZ isolates were sequenced using PacBio® WGS. These sequences yielded draft genomes that were highly contiguous, and provide invaluable data to improve current diagnostic tests for SEE. Many polymerase chain reaction (PCR) assays targeting genes to identify or differentiate SEE and SEZ have limitations that contribute to false-negative and false positive results. For example, a non-accredited laboratory that is currently used widely by large animal clinical faculty at Texas A&M University targets the M-like (SeM) protein.[134] This approach has 2 important limitations: 1) the SeM protein in SEE is known to have truncations resulting in the missing the presence or previous exposure to SEE, such that this test can yield false negative results;[54] and, 2) we have identified that the M protein of some strains of SEZ (SzM) have homology with the SeM protein (unpublished data), such that this test can hypothetically result in misidentification of SEZ as SEE using this test (*i.e.*, yield false-positive results). Anecdotally, at the time of writing this chapter a horse whose clinical signs (unilateral nasal discharge, absence of lymph node abscessation, absence of other affected horses in the herd), microbiologic culture results (only SEZ and *Streptococcus dysgalactiae* subsp. *equisimilis* isolated by microbiologic culture), and alternative PCR testing (*sodA* and *seeI* genes at TVMDL) strongly suggest that the horse is not infected with SEE tested positive using an SeM-based PCR at a laboratory commonly used by some Texas A&M University clinicians because of the convenience of a single test encompassing a panel of pathogens. Analysis of the genomes of our US strains of SEE and SEZ allowed us to identify a gene specific to SEE and another specific to SEZ that have not been previously reported; alignment of these genes with publicly-available genomes from other regions confirmed the specificity of the genes for SEE and SEZ (unpublished data). Preliminary results indicate that under simulated co-infection (co-cultivation), both SEE and SEZ can be simultaneously identified by a duplex PCR at varying levels of presence of each individual organism. We plan to publish these findings, and we are collaborating with TVMDL to establish the diagnostic utility of primers and probes for real-time PCR tests for these genes. Although a disclosure of invention has been filed, we intend to make these data available publicly to improve equine health.

5.2.3. Functional differences in methylomes of SEE and SEZ

We characterized for the first time the global methylomes of SEE and SEZ. The high level of genetic homology between SEE and SEZ facilitated comparison of the global methylomes, specifically among homologous proteins of SEE and SEZ to gain further insights about gene regulation that determines the single-host restriction of SEE. Canonically, the presence of methylation in prokaryotes is considered to be a line of defense against invading bacteriophages or foreign DNA.[98] Nevertheless, the presence or absence of methylation has been demonstrated to affect either gene expression or virulence in prokaryotes.[99-101] Thus, it is plausible that differential methylation could alter the virulence or other phenotype of SEE and SEZ.

The global methylomes of SEE strains were more homologous in the occurrence and specific types of methylation compared to SEZ. Genes associated with exopeptidase activity (n = 3) and quorum sensing (n = 3) were found to be methylated in all SEE genomes but were not methylated in any SEZ isolates. Conceivably, exopeptidase activity may be imperative for SEE to help evading phagocytosis similar to the endopeptidase activity of EndoSe and its degradation of immunoglobulin G.[135] Although quorum sensing has been described in group A *Streptococcus* and in SEZ, it has not yet been investigated in SEE; however, components of quorum sensing were deemed necessary for the fitness of SEE in whole blood.[121,136,137] These hypothesized differences in quorum sensing could provide an additional explanation for the reported differences in the thickness of the capsules of SEE and SEZ.[10] The recently described *rgg*/*shp* quorum sensing system in SEZ was noted to influence the thickness of hyaluronic acid capsule production in SEZ: in an SEZ *shp* gene deletion mutant (Δshp) and in an SEZ *rgg*-*shp* double-deletion mutant (Δrgg-shp), production of hyaluronic acid was reduced about 20% when compared to the wild-type SEZ.[121] Performing reverse transcription real-time PCR of the 6 differentially methylated genes could be used to determine the influence of the observed methylation on the expression of these genes. Furthermore, observing the phenotypic effects of knocking out the genes found to be differentially methylated (with reverse complementation) would provide further evidence of the biological significance of our results. For example, the 3 genes associated with exopeptidase activity could be knocked out, and then survival of these deletion mutants could be assessed by comparing the colony forming units of bacteria before and after an incubation for 1 hour with equine neutrophils.

5.2.4. Carrier SEE and host-adaptions

For the first time, we used multiple "omics" technologies were used to compare the genomes of SEE isolated from clinical cases with SEE genomes isolated from inapparent carriers. No consistent or significant pathogen-associated adaptions were identified in the genome, global methylome, or transcriptome of carrier SEE isolates relative to clinical isolates. These data suggest that host-associated adaptions may be the key to the inapparent carrier of SEE. To date, studying the host's response to SEE has been difficult because of the absence of a challenge model in species other than horses that replicates natural disease,[1] and culture of appropriate equine cell-line for an *in vitro* approach has not been reported.

Current advances in our laboratory such as the successful culture of equine guttural pouch epithelial cells could provide an *in vitro* method for understanding host-agent interactions using cellular and molecular biological methods at the level of the respiratory epithelium, including comparing interactions of clinical strains with carrier strains and cells from immune and susceptible hosts. It remains unclear, however, in which cells SEE persists in hosts. Current advances in single-cell sequencing technologies with dual single-cell RNA sequencing (scDual-Seq)[138] of host and pathogen transcripts using an *in vitro* (guttural pouch epithelium) or *in vivo* approach could be applied to equine respiratory epithelial cells infected with SEE. An *in vivo* approach for understanding the relative virulence could be performed using horses infected with either carrier strains of SEE or clinical strains of SEE. Our results, however, dampen enthusiasm for this approach given the evidence that consistent differences between carrier and clinical strains of SEE were not observed.

5.3. Final remarks

While we were able to elucidate many aspects of the molecular epidemiology of SEE isolates from the US, much work remains. Ultimately, understanding the source of host-restriction for SEE and host-factors that result in the inapparent carrier presentation of SEE will help us to better control and prevent strangles outbreaks and infections. A byproduct of this research was the development of improved PCR targets for diagnosis of infection with SEE and SEZ.

REFERENCES

1 Boyle, A. G. *et al. Streptococcus equi* infections in horses: Guidelines for treatment, control, and prevention of strangles-Revised consensus statement. *J Vet Intern Med* **32**, 633-647, doi:10.1111/jvim.15043 (2018).

2 Rendle, D. *et al. Streptococcus equi* infections: Current best practice in the diagnosis and management of ‘strangles’. *UK-Vet Equine* **5**, S3-S15, doi:10.12968/ukve.2021.5.2.s.3 (2021).

3 Waller, A. S. New perspectives for the diagnosis, control, treatment, and prevention of strangles in horses. *Vet Clin North Am Equine Pract* **30**, 591-607, doi:10.1016/j.cveq.2014.08.007 (2014).

4 Waller, A. *Streptococcus equi*: breaking its strangles-hold. *Vet Rec* **182**, 316-318, doi:10.1136/vr.k1231 (2018).

5 Timoney, J. F. & Kumar, P. Early pathogenesis of equine *Streptococcus equi* infection (strangles). *Equine Vet J* **40**, 637-642, doi:10.2746/042516408x322120 (2008).

6 Muhktar, M. M. & Timoney, J. F. Chemotactic response of equine polymorphonuclear leucocytes to *Streptococcus equi*. *Res Vet Sci* **45**, 225-229 (1988).

7 Pusterla, N. *et al.* Purpura haemorrhagica in 53 horses. *Vet Rec* **153**, 118-121, doi:10.1136/vr.153.4.118 (2003).

8 Kaese, H. J. *et al.* Infarctive purpura hemorrhagica in five horses. *J Am Vet Med Assoc* **226**, 1893-1898, 1845, doi:10.2460/javma.2005.226.1893 (2005).

9 Sponseller, B. T. *et al.* Severe acute rhabdomyolysis associated with *Streptococcus equi i*nfection in four horses. *J Am Vet Med Assoc* **227**, 1800-1807, 1753-1804, doi:10.2460/javma.2005.227.1800 (2005).

10 Holden, M. T. G. *et al.* Genomic evidence for the evolution of *Streptococcus equi*: Host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog* **5**, e1000346, doi:10.1371/journal.ppat.1000346 (2009).

11 Anzai, T. *et al.* In vivo pathogenicity and resistance to phagocytosis of *Streptococcus equi* strains with different levels of capsule expression. *Vet Microbiol* **67**, 277-286, doi:10.1016/s0378-1135(99)00051-6 (1999).

12 Timoney, J. F., Artiushin, S. C. & Boschwitz, J. S. Comparison of the sequences and functions of *Streptococcus equi* M-like proteins SeM and SzPSe. *Infect Immun* **65**, 3600-3605 (1997).

13 Timoney, J. F. The pathogenic equine streptococci. *Vet Res* **35**, 397-409, doi:10.1051/vetres:2004025 (2004).

14 Timoney, J. F., Suther, P., Velineni, S. & Artiushin, S. C. The antiphagocytic activity of SeM of *Streptococcus equi* requires capsule. *J Equine Sci* **25**, 53-56, doi:10.1294/jes.25.53 (2014).

15 Tiwari, R., Qin, A., Artiushin, S. & Timoney, J. F. Se18.9, an anti-phagocytic factor H binding protein of *Streptococcus equi*. *Vet Microbiol* **121**, 105-115, doi:10.1016/j.vetmic.2006.11.023 (2007).

16 Galan, J. E. & Timoney, J. F. Mucosal nasopharyngeal immune responses of horses to protein antigens of *Streptococcus equi*. *Infect Immun* **47**, 623-628, doi:10.1128/iai.47.3.623-628.1985 (1985).

17 Hamlen, H. J., Timoney, J. F. & Bell, R. J. Epidemiologic and immunologic characteristics of *Streptococcus equi* infection in foals. *J Am Vet Med Assoc* **204**, 768-775 (1994).

18 Timoney, J. F., Qin, A., Muthupalani, S. & Artiushin, S. Vaccine potential of novel surface exposed and secreted proteins of *Streptococcus equi*. *Vaccine* **25**, 5583-5590, doi:10.1016/j.vaccine.2007.02.040 (2007).

19 Pringle, J., Storm, E., Waller, A. & Riihimäki, M. Influence of penicillin treatment of horses with strangles on seropositivity to *Streptococcus equi* ssp. *equi*-specific antibodies. *J Vet Intern Med* **34**, 294-299, doi:10.1111/jvim.15668 (2020).

20 Waller, A. S., Paillot, R. & Timoney, J. F. *Streptococcus equi*: a pathogen restricted to one host. *J Med Microbiol* **60**, 1231-1240, doi:10.1099/jmm.0.028233-0 (2011).

21 Torpiano, P., Nestorova, N. & Vella, C. *Streptococcus equi* subsp. *equi* meningitis, septicemia and subdural empyema in a child. *IDCases* **21**, e00808, doi:10.1016/j.idcr.2020.e00808 (2020).

22 Ladlow, J., Scase, T. & Waller, A. Canine strangles case reveals a new host susceptible to infection with *Streptococcus equi*. *J Clin Microbiol* **44**, 2664-2665, doi:10.1128/jcm.00571-06 (2006).

23 Elsayed, S., Hammerberg, O., Massey, V. & Hussain, Z. S*treptococcus equi* subspecies *equi* (Lancefield group C) meningitis in a child. *Clin Microbiol Infect* **9**, 869-872, doi:10.1046/j.1469-0691.2003.00663.x (2003).

24 Parmar, J., Winterbottom, A., Cooke, F., Lever, A. M. L. & Gaunt, M. Endovascular aortic stent graft infection with Streptococcus equi: the first documented case. *Vascular* **21**, 14-16, doi:10.1258/vasc.2010.cr0258 (2013).

25 Weese, J. S., Jarlot, C. & Morley, P. S. Survival of Streptococcus equi on surfaces in an outdoor environment. *Can Vet J* **50**, 968-970 (2009).

26 Durham, A. E., Hall, Y. S., Kulp, L. & Underwood, C. A study of the environmental survival of *Streptococcus equi* subspecies *equi*. *Equine Vet J* **50**, 861-864, doi:10.1111/evj.12840 (2018).

27 Sweeney, C. R., Timoney, J. F., Newton, J. R. & Hines, M. T. *Streptococcus equi* infections in horses: guidelines for treatment, control, and prevention of strangles. *J Vet Intern Med* **19**, 123-134 (2005).

28 Newton, J. R., Wood, J. L., Dunn, K. A., DeBrauwere, M. N. & Chanter, N. Naturally occurring persistent and asymptomatic infection of the guttural pouches of horses with *Streptococcus equi*. *Vet Rec* **140**, 84-90, doi:10.1136/vr.140.4.84 (1997).

29 Harris, S. R. *et al.* Genome specialization and decay of the strangles pathogen, *Streptococcus equi*, is driven by persistent infection. *Genome Res* **25**, 1360-1371, doi:10.1101/gr.189803.115 (2015).

30 Verheyen, K., Newton, J. R., Talbot, N. C., de Brauwere, M. N. & Chanter, N. Elimination of guttural pouch infection and inflammation in asymptomatic carriers of *Streptococcus equi*. *Equine Vet J* **32**, 527-532 (2000).

31 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* **32**, 515-526 (2000).

32 Riihimaki, M., Aspan, A., Ljung, H. & Pringle, J. Long term dynamics of a *Streptococcus equi* ssp *equi* outbreak, assessed by qPCR and culture and seM sequencing in silent carriers of strangles. *Vet Microbiol* **223**, 107-112, doi:10.1016/j.vetmic.2018.07.016 (2018).

33 Duffee, L. R., Stefanovski, D., Boston, R. C. & Boyle, A. G. Predictor variables for and complications associated with *Streptococcus equi* subsp *equi* infection in horses. *J Am Vet Med Assoc* **247**, 1161-1168, doi:10.2460/javma.247.10.1161 (2015).

34 Morris, E. R. A. *et al.* Comparison of whole genome sequences of *Streptococcus equi* subsp. *equi* from an outbreak in Texas with isolates from within the region, Kentucky, USA, and other countries. *Vet Microbiol* **243**, 108638, doi:10.1016/j.vetmic.2020.108638 (2020).

35 Pringle, J., Venner, M., Tscheschlok, L., Bachi, L. & Riihimaki, M. Long term silent carriers of *Streptococcus equi* ssp. *equi* following strangles; carrier detection related to sampling site of collection and culture versus qPCR. *Vet J* **246**, 66-70, doi:10.1016/j.tvjl.2019.02.003 (2019).

36 Pringle, J., Venner, M., Tscheschlok, L., Waller, A. S. & Riihimäki, M. Markers of long term silent carriers of *Streptococcus equi* ssp. *equi* in horses. *J Vet Intern Med* **34**, 2751-2757, doi:10.1111/jvim.15939 (2020).

37 Durham, A. E. & Kemp-Symonds, J. Failure of serological testing for antigens A and C of Streptococcus equi subspecies *equi* to identify guttural pouch carriers. *Equine Vet J* **53**, 38-43, doi:10.1111/evj.13276 (2021).

38 Chanter, N., Talbot, N. C., Newton, J. R., Hewson, D. & Verheyen, K. Streptococcus equi with truncated M-proteins isolated from outwardly healthy horses. *Microbiology (Reading)* **146 (Pt 6)**, 1361-1369, doi:10.1099/00221287-146-6-1361 (2000).

39 Heather, Z. *et al.* A novel streptococcal integrative conjugative element involved in iron acquisition. *Mol Microbiol* **70**, 1274-1292, doi:10.1111/j.1365-2958.2008.06481.x (2008).

40 Riley, L. W. & Blanton, R. E. Advances in molecular epidemiology of infectious diseases: Definitions, approaches, and scope of the field*. *Microbiol Spectr* **6**, doi:10.1128/microbiolspec.ame-0001-2018 (2018).

41 Maljkovic Berry, I. *et al.* Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: Approaches,

applications, and considerations for development of laboratory capacity. *J Infect Dis*, doi:10.1093/infdis/jiz286 (2019).

42 Jorm, L. R., Love, D. N., Bailey, G. D., McKay, G. M. & Briscoe, D. A. Genetic structure of populations of beta-haemolytic Lancefield group C streptococci from horses and their association with disease. *Res Vet Sci* **57**, 292-299 (1994).

43 Webb, K. *et al.* Development of an unambiguous and discriminatory multilocus sequence typing scheme for the *Streptococcus zooepidemicus* group. *Microbiology (Reading)* **154**, 3016-3024, doi:10.1099/mic.0.2008/018911-0 (2008).

44 Pelkonen, S. *et al.* Transmission of *Streptococcus equi* subspecies *zooepidemicus* infection from horses to humans. *Emerg Infect Dis* **19**, 1041-1048, doi:10.3201/eid1907.121365 (2013).

45 Chen, X. *et al.* Genetic characterization of *Streptococcus equi* subspecies *zooepidemicus* associated with high swine mortality in the United States. *Transbound Emerg Dis*, doi:10.1111/tbed.13645 (2020).

46 Byun, J. W., Yoon, S. S., Woo, G. H., Jung, B. Y. & Joo, Y. S. An outbreak of fatal hemorrhagic pneumonia caused by *Streptococcus equi* subsp. *zooepidemicus* in shelter dogs. *J Vet Sci* **10**, 269-271, doi:10.4142/jvs.2009.10.3.269 (2009).

47 Las Heras, A. *et al.* Unusual outbreak of clinical mastitis in dairy sheep caused by *Streptococcus equi* subsp. *zooepidemicus*. *J Clin Microbiol* **40**, 1106-1108, doi:10.1128/jcm.40.3.1106-1108.2002 (2002).

48 Blum, S. *et al.* Outbreak of *Streptococcus equi* subsp. *zooepidemicus* infections in cats. *Vet Microbiol* **144**, 236-239, doi:10.1016/j.vetmic.2009.12.040 (2010).

49 Bannister, M. F., Benson, C. E. & Sweeney, C. R. Rapid species identification of group C streptococci isolated from horses. *J Clin Microbiol* **21**, 524-526, doi:10.1128/JCM.21.4.524-526.1985 (1985).

50 Grant, S. T., Efstratiou, A. & Chanter, N. Laboratory diagnosis of strangles and the isolation of atypical *Streptococcus equi*. *Vet Rec* **133**, 215-216, doi:10.1136/vr.133.9.215 (1993).

51 Timoney, J. F. & Artiushin, S. C. Detection of *Streptococcus equi* in equine nasal swabs and washes by DNA amplification. *Vet Rec* **141**, 446-447, doi:10.1136/vr.141.17.446 (1997).

52 Webb, K. *et al.* Detection of *Streptococcus equi* subspecies *equi* using a triplex qPCR assay. *Vet J* **195**, 300-304, doi:10.1016/j.tvjl.2012.07.007 (2013).

53 Jolley, K. A. *et al.* Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology (Reading)* **158**, 1005-1015, doi:10.1099/mic.0.055459-0 (2012).

54 Kelly, C. *et al.* Sequence variation of the SeM gene of *Streptococcus equi* allows discrimination of the source of strangles outbreaks. *J Clin Microbiol* **44**, 480-486, doi:10.1128/JCM.44.2.480-486.2006 (2006).

55 Patty, O. & Cursons, R. The molecular identification of *Streptococcus equi* subsp. *equi* strains isolated within New Zealand. *N Z Vet J* **62**, 63-67, doi:10.1080/00480169.2013.841536 (2014).

56 Tartor, Y. H., El-Naenaeey, E. S. Y., Gharieb, N. M., Ali, W. S. & Ammar, A. M. Novel *Streptococcus equi* strains causing strangles outbreaks in Arabian horses in Egypt. *Transbound Emerg Dis* **67**, 2455-2466, doi:10.1111/tbed.13584 (2020).

57 Dong, J. *et al.* An outbreak of strangles associated with a novel genotype of *Streptococcus equi* subspecies equi in donkeys in China during 2018. *Equine Vet J* **51**, 743-748, doi:10.1111/evj.13114 (2019).

58 Libardoni, F. *et al.* Diversity of seM in *Streptococcus equi* subsp. *equi* isolated from strangles outbreaks. *Vet Microbiol* **162**, 663-669, doi:10.1016/j.vetmic.2012.09.010 (2013).

59 Corander, J., Marttinen, P., Sirén, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**, 539, doi:10.1186/1471-2105-9-539 (2008).

60 Newton, J. R., Laxton, R., Wood, J. L. N. & Chanter, N. Molecular epidemiology of *Streptococcus zooepidemicus* infection in naturally occurring equine respiratory disease. *Vet J* **175**, 338-345, doi:https://doi.org/10.1016/j.tvjl.2007.02.018 (2008).

61 Paillot, R. *et al.* Contribution of each of four superantigens to *Streptococcus equi*-induced mitogenicity, gamma interferon synthesis, and immunity. *Infect Immun* **78**, 1728-1739, doi:10.1128/iai.01079-09 (2010).

62 Paillot, R. *et al.* Identification of three novel superantigen-encoding genes in *Streptococcus equi* subsp. *zooepidemicus*, szeF, szeN, and szeP. *Infect Immun* **78**, 4817-4827, doi:10.1128/iai.00751-10 (2010).

63 Lindmark, H., Nilsson, M. & Guss, B. 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* **69**, 3159-3163, doi:10.1128/iai.69.5.3159-3163.2001 (2001).

64 Brüssow, H., Canchaya, C. & Hardt, W.-D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**, 560-602, doi:10.1128/mmbr.68.3.560-602.2004 (2004).

65 Miller, D. A., Luo, L., Hillson, N., Keating, T. A. & Walsh, C. T. Yersiniabactin synthetase. *Chem Biol* **9**, 333-344, doi:10.1016/s1074-5521(02)00115-1 (2002).

66 Anonymous. The economic impact of the horse industry on the United States. (Commissioned study by the American Horse Council Foundation, 2018).

67 Eybpoosh, S. *et al.* Molecular epidemiology of infectious diseases. *Electron Physician* **9**, 5149-5158, doi:10.19082/5149 (2017).

68 Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: A descriptive study. *Lancet Infect Dis* **13**, 130-136, doi:10.1016/s1473-3099(12)70268-2 (2013).

69 Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res* **45**, D535-d542, doi:10.1093/nar/gkw1017 (2017).

70 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

71 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

72 Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* **15**, 524, doi:10.1186/s13059-014-0524-x (2014).

73 Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, doi:https://doi.org/10.1099/mgen.0.000093 (2016).

74 R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2019).

75 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

76 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

77 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

78 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

79 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695 (2012).

80 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **44**, W3-w10, doi:10.1093/nar/gkw343 (2016).

81 Breakwell, D., MacDonald, B., Woolverton, C., Smith, K., Robison, R. Colony Morphology Protocol. (2007). <https://www.asmscience.org/content/education/protocol/protocol.3136>.

82 Boyle, A. G., Stefanovski, D. & Rankin, S. C. Comparison of nasopharyngeal and guttural pouch specimens to determine the optimal sampling site to detect *Streptococcus equi* subsp *equi* carriers by DNA amplification. *BMC Vet Res* **13**, 75, doi:10.1186/s12917-017-0989-4 (2017).

83 Lindahl, S., Baverud, V., Egenvall, A., Aspan, A. & Pringle, J. Comparison of sampling sites and laboratory diagnostic tests for *S. equi* subsp. *equi* in horses from confirmed strangles outbreaks. *J Vet Intern Med* **27**, 542-547, doi:10.1111/jvim.12063 (2013).

84 Maurer, P., Todorova, K., Sauerbier, J. & Hakenbeck, R. Mutations in *Streptococcus pneumoniae* penicillin-binding protein 2x: Importance of the C-terminal penicillin-binding protein and serine/threonine kinase-associated domains for beta-lactam binding. *Microb Drug Resist* **18**, 314-321, doi:10.1089/mdr.2012.0022 (2012).

85 Nichol, K. A., Zhanel, G. G. & Hoban, D. J. Penicillin-binding protein 1A, 2B, and 2X alterations in Canadian isolates of penicillin-resistant *Streptococcus pneumoniae*. *Antimicrob Agents Chemother* **46**, 3261-3264, doi:10.1128/aac.46.10.3261-3264.2002 (2002).

86 Charbonneau, A. R. L. *et al.* Defining the ABC of gene essentiality in streptococci. *BMC Genomics* **18**, 426, doi:10.1186/s12864-017-3794-3 (2017).

87 Meehan, M., Lynagh, Y., Woods, C. & Owen, P. The fibrinogen-binding protein (FgBP) of *Streptococcus equi* subsp. *equi* additionally binds IgG and contributes to virulence in a mouse model. *Microbiology (Reading)* **147**, 3311-3322, doi:10.1099/00221287-147-12-3311 (2001).

88 Cursons, R., Patty, O., Steward, K. F. & Waller, A. S. Strangles in horses can be caused by vaccination with Pinnacle I. N. *Vaccine* **33**, 3440-3443, doi:10.1016/j.vaccine.2015.05.009 (2015).

89 Paillot, R., Lopez-Alvarez, M. R., Newton, J. R. & Waller, A. S. Strangles: A modern clinical view from the 17th century. *Equine Vet J* **49**, 141-145, doi:10.1111/evj.12659 (2017).

90 Slater, J. D. Strangles, bastard strangles, vives and glanders: archaeological relics in a genomic age. *Equine Vet J* **35**, 118-120, doi:10.2746/042516403776114252 (2003).

91 Lindahl, S. B. *et al.* Outbreak of upper respiratory disease in horses caused by *Streptococcus equi* subsp. *zooepidemicus* ST-24. *Vet Microbiol* **166**, 281-285, doi:10.1016/j.vetmic.2013.05.006 (2013).

92 Björnsdóttir, S. *et al.* Genomic dissection of an Icelandic epidemic of respiratory disease in horses and associated zoonotic cases. *mBio* **8**, doi:10.1128/mbio.00826-17 (2017).

93 Preziuso, S., Moriconi, M. & Cuteri, V. Genetic diversity of *Streptococcus equi* subsp. *zooepidemicus* isolated from horses. *Comp Immun Microbiol Infect Dis* **65**, 7-13, doi:10.1016/j.cimid.2019.03.012 (2019).

94 Ozer, E. A., Allen, J. P. & Hauser, A. R. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* **15**, 737, doi:10.1186/1471-2164-15-737 (2014).

95 Gao, X.-Y., Zhi, X.-Y., Li, H.-W., Klenk, H.-P. & Li, W.-J. Comparative genomics of the bacterial genus *Streptococcus* illuminates evolutionary implications of species groups. *PLoS One* **9**, e101229, doi:10.1371/journal.pone.0101229 (2014).

96 Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461-465, doi:10.1038/nmeth.1459 (2010).

97 Blow, M. J. *et al.* The epigenomic landscape of prokaryotes. *PLoS Genet* **12**, e1005854, doi:10.1371/journal.pgen.1005854 (2016).

98 Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr Opin Microbiol* **25**, 9-16, doi:10.1016/j.mib.2015.03.004 (2015).

99 Nye, T. M. *et al.* DNA methylation from a Type I restriction modification system influences gene expression and virulence in *Streptococcus pyogenes*. *PLoS Pathog* **15**, e1007841, doi:10.1371/journal.ppat.1007841 (2019).

100 Gomez-Gonzalez, P. J. *et al.* An integrated whole genome analysis of Mycobacterium tuberculosis reveals insights into relationship between its genome, transcriptome and methylome. *Sci Rep* **9**, doi:10.1038/s41598-019-41692-2 (2019).

101 Casselli, T. *et al.* DNA methylation by restriction modification systems affects the global transcriptome profile in *Borrelia burgdorferi*. *J Bacteriol* **200**, doi:10.1128/jb.00395-18 (2018).

102 Furuta, Y. & Kobayashi, I. Mobility of DNA sequence recognition domains in DNA methyltransferases suggests epigenetics-driven adaptive evolution. *Mob Genet Elements* **2**, 292-296, doi:10.4161/mge.23371 (2012).

103 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).

104 Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* **5**, 8365, doi:10.1038/srep08365 (2015).

105 Ozer, E. A. ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. *BMC Bioinformatics* **19**, 150, doi:10.1186/s12859-018-2154-x (2018).

106 Shannon, P. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

107 Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093, doi:10.1093/bioinformatics/btp101 (2009).

108 Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608-1615, doi:10.1093/bioinformatics/btq249 (2010).

109 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

110 Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **43**, D298-299, doi:10.1093/nar/gku1046 (2015).

111 Poindexter, N. J. & Schlievert, P. M. Suppression of imunoglobulin-secreting cells from human peripheral blood by Toxic-Shock Syndrome Toxin-1. *J Infect Dis* **153**, 772-779, doi:10.1093/infdis/153.4.772 (1986).

112 Bohach, G. A., Fast, D. J., Nelson, R. D. & Schlievert, P. M. Staphylococcal and streptococcal pyrogenic toxins involved in Toxic Shock Syndrome and related illnesses. *Crit Rev Microbiol* **17**, 251-272, doi:10.3109/10408419009105728 (1990).

113 Brachmann, C. B. *et al.* The SIR2 gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression, and chromosome stability. *Genes Dev* **9**, 2888-2902, doi:10.1101/gad.9.23.2888 (1995).

114 Okumura, K. *et al.* Evolutionary paths of streptococcal and staphylococcal superantigens. *BMC Genomics* **13**, 404, doi:10.1186/1471-2164-13-404 (2012).

115 Fast, D. J., Schlievert, P. M. & Nelson, R. D. Nonpurulent response to toxic shock syndrome toxin 1-producing Staphylococcus aureus. Relationship to toxin-stimulated production of tumor necrosis factor. *J Immunol* **140**, 949-953 (1988).

116 Spaulding, A. R. *et al.* Staphylococcal and streptococcal Superantigen exotoxins. *Clin Microbiol Rev* **26**, 422-447, doi:10.1128/cmr.00104-12 (2013).

117 Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat Rev Genet* **20**, 157-172, doi:10.1038/s41576-018-0081-3 (2019).

118 Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N. & Uchiyama, I. Shaping the genome – restriction–modification systems as mobile genetic elements. *Curr Opin Genet Dev* **9**, 649-656, doi:10.1016/s0959-437x(99)00026-x (1999).

119 Conlan, S. *et al.* Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci Trans Med* **6**, 254ra126-254ra251, doi:10.1126/scitranslmed.3009845 (2014).

120 Sahu, P. & Pallaval Veera, B. 337-348 (Springer Singapore, 2018).

121 Xie, Z. *et al.* Identification of a quorum sensing system regulating capsule polysaccharide production and biofilm formation in *Streptococcus zooepidemicus*. *Front Cell Infect Microbiol* **9**, doi:10.3389/fcimb.2019.00121 (2019).

122 Segerman, B. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front Cell Infect Microbiol* **2**, doi:10.3389/fcimb.2012.00116 (2012).

123 Steward, K. F., Robinson, C. & Waller, A. S. Transcriptional changes are involved in phenotype switching in Streptococcus equi subspecies *equi*. *Mol Biosyst* **12**, 1194-1200, doi:10.1039/c5mb00780a (2016).

124 Roosaare, M. *et al.* StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. *PeerJ* **5**, e3353, doi:10.7717/peerj.3353 (2017).

125 Davis, J. J. *et al.* The PATRIC bioinformatics resource center: Expanding data and analysis capabilities. *Nucleic Acids Res*, doi:10.1093/nar/gkz943 (2019).

126 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).

127 Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* **38**, W695-W699, doi:10.1093/nar/gkq313 (2010).

128 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

129 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

130 Lun, A. T. L., Chen, Y. & Smyth, G. K. 391-416 (Springer New York, 2016).

131 Edgar, R. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210, doi:10.1093/nar/30.1.207 (2002).

132 Pringle, J., Venner, M., Tscheschlok, L., Bächi, L. & Riihimäki, M. Long term silent carriers of *Streptococcus equi* ssp. *equi* following strangles; carrier detection related to sampling site of collection and culture versus qPCR. V*et J* **246**, 66-70, doi:10.1016/j.tvjl.2019.02.003 (2019).

133 Mandlik, A. *et al.* RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe* **10**, 165-174, doi:10.1016/j.chom.2011.07.007 (2011).

134 Pusterla, N., Leutenegger, C. M., Barnum, S. M. & Byrne, B. A. Use of quantitative real-time PCR to determine viability of *Streptococcus equi* subspecies *equi* in respiratory secretions from horses with strangles. *Equine Vet J* **50**, 697-700, doi:10.1111/evj.12809 (2018).

135 Flock, M., Frykberg, L., Sköld, M., Guss, B. & Flock, J.-I. Antiphagocytic function of an IgG glycosyl hydrolase from *Streptococcus equi* subsp. *equi* and its use as a vaccine component. *Infect Immun* **80**, 2914-2919, doi:10.1128/iai.06083-11 (2012).

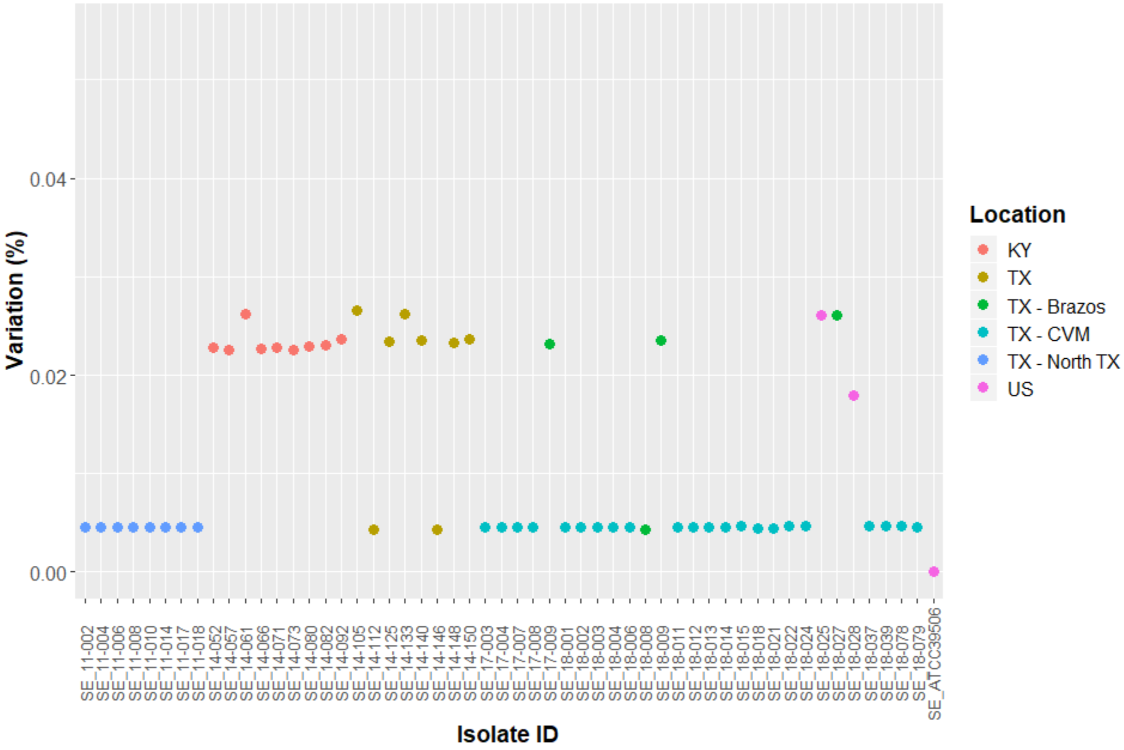
136 Jimenez, J. C. & Federle, M. J. Quorum sensing in group A *Streptococcus*. *Front Cellul Infect Microbiol* **4**, doi:10.3389/fcimb.2014.00127 (2014).

137 Charbonneau, A. R. L. *et al.* Identification of genes required for the fitness of *Streptococcus equi* subsp. *equi* in whole equine blood and hydrogen peroxide. *Microb Genom* **6**, doi:10.1099/mgen.0.000362 (2020).

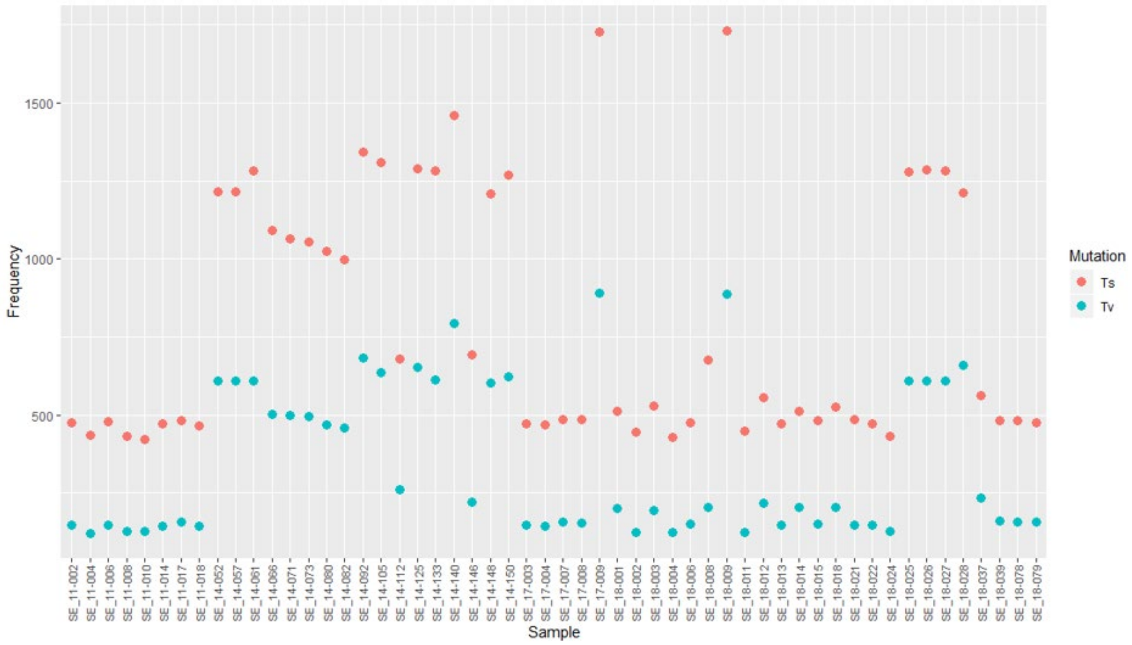
138 Avital, G. *et al.* scDual-Seq: mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing. *Genome Biol* **18**, doi:10.1186/s13059-017-1340-x (2017).

APPENDIX A

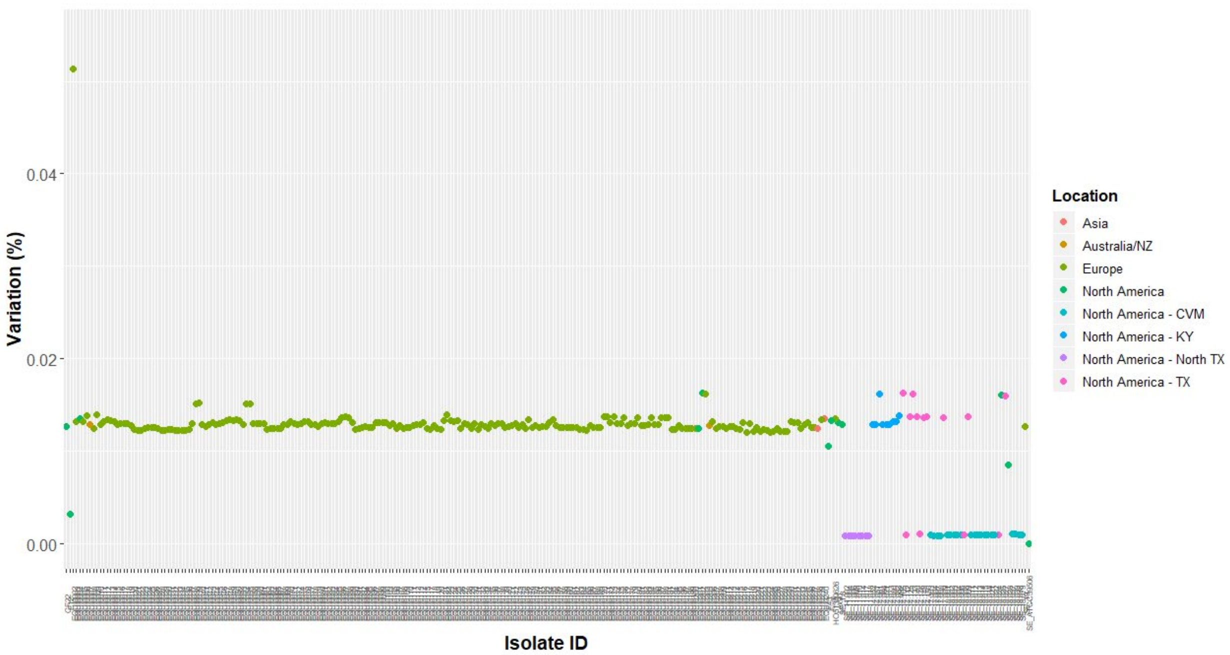
SUPPLEMENTARY MATERIAL: COMPARISON OF WHOLE GENOME SEQUENCES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM AN OUTBREAK IN TEXAS WITH ISOLATES FROM WITHIN THE REGION, KENTUCKY, USA, AND OTHER COUNTRIES



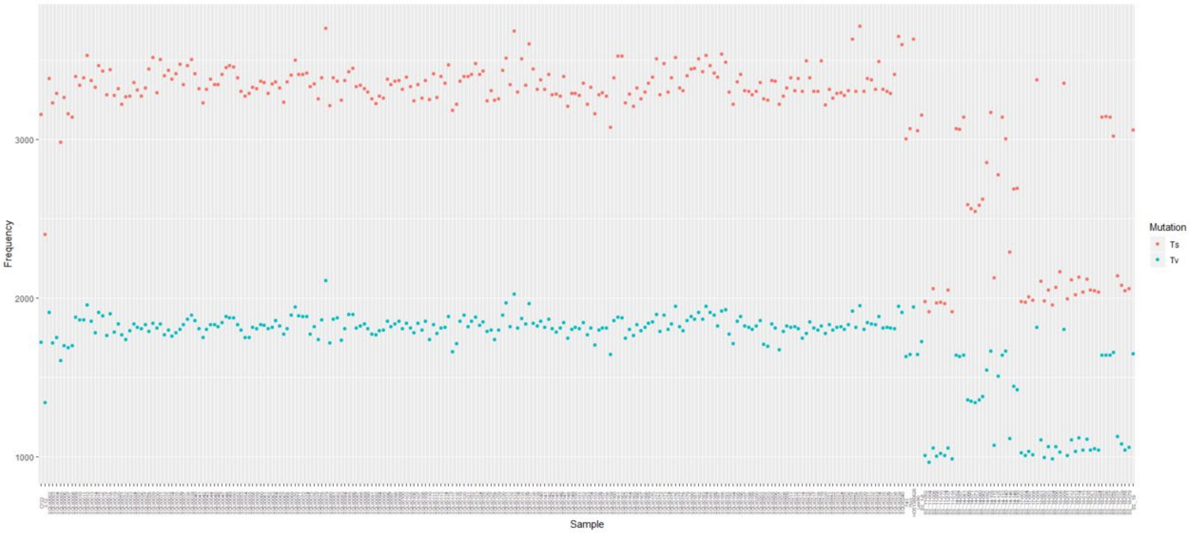
A-1 Fig. Percentage of variance for the core genome relative to the reference in 54 *S. equi* isolates from US. Variances among *S. equi* isolates from the same outbreak, such as the CVM (light blue) or north Texas outbreak (blue), approximately 0.0045%. *S. equi* isolates from unrelated incidences (red, yellow, green, and pink) exhibited a variance range of 0.0043% - 0.0265%.



A-2 Fig. Rate of transition (Ts; pink) and transversion (Tv; blue) mutation rates of entire genome for 54 Texas and Kentucky *S. equi* isolates relative to the reference. Overall, the 54 *S. equi* isolates had a lower frequency of Tv mutations than Ts, whereas, the *S. equi* isolates from a single outbreak had similar rates of Ts and Tv mutations.



A-3 Fig. Percentage of variance for the core genome relative to the reference in 54 United States and 230 publicly available *S. equi* isolates. *S. equi* isolates from various countries maintain an average variance of 0.0129%, whereas the isolates related to the College of Veterinary Medicine outbreak and the north Texas ranch outbreak both have an average of 0.0009% variance.



A-4 Fig. Rate of transition (Ts; pink) and transversion (Tv; blue) mutation rates of entire genome for 54 United States (Texas and Kentucky) and 230 publicly available *S. equi* isolates relative to the reference. Among all 284 *S. equi* isolates compared, there was a greater frequency of Ts mutations than the frequency of Tv mutations.

A-1 File. Detailed description of the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM) strangles outbreak and surrounding events.

The strangles outbreak at the College of Veterinary Medicine & Biomedical Sciences, Texas A&M University (CVM) occurred in late 2017 until early 2018. This was 5 months after the conclusion of a project in which 16 yearlings were challenged individually with SEE using the infection strain, 11-017 to test the efficacy of a candidate vaccine. This strain was from an outbreak of strangles in 2011 at a ranch in north Texas, for which we had multiple isolates in our laboratory's repository. The yearlings for this project were housed in isolation with biosecurity measures that included physical barriers, dedicate personal protective clothing (dedicated rubber boots, bleach foot baths and spray, disposable gloves, dedicated coveralls, and access restricted to trained personnel). Two weeks after resolution of clinical signs, all yearlings were required to have negative results of guttural pouch endoscopy and microbiologic culture results for SEE of guttural pouch lavage fluid on 3 consecutive examinations (Newton *et al.*, 2000) separated by 2 weeks prior to being reintroduced to the CVM teaching herd. The index case (strain 17-004) for the outbreak was noted to have clinical signs of bilateral purulent nasal discharge and depressed attitude on November 22, 2017. This horse had been housed for approximately 5 months with young horses that had been used in the aforementioned strangles vaccine project. The index case and her 5 pasture-mates were immediately isolated and biosecurity procedures were implemented to prevent dissemination of infection. Guttural pouch endoscopy and microbiologic culture and polymerase chain reaction (PCR) testing were performed on the 5 pasture-mates of the index case, and results were consistently negative for multiple samplings (at least 4 occasions for each horse). Despite the isolation of the index case, 8 more cases of strangles developed in the CVM herd. After the second case of strangles was identified (strain 17-003), all 15 horses in the CVM herd were isolated to their paddocks, and tested for SEE by PCR and microbiologic culture of nasopharyngeal lavage. Any horse identified as positive for SEE by culture or PCR was subsequently affected, these horses were housed in paddocks remote from the paddock of the index case, and had no direct contact with the horses in the paddock housing the index case. All 15 horses in the CVM herd were tested for SEE by PCR and microbiologic culture of nasopharyngeal lavage. Horses with clinical symptoms from the CVM outbreak were utilized as a direct contact challenge for a subsequent 2018 SEE vaccine study. This study was a continuation of the 2017 SEE candidate vaccine efficacy study.

lace text or figures/tables here.

A-1 Table. Description of the 230 publicly available SEE isolates from PATRIC.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.60	EQUI0238	Saudi Arabia	Asia
148942.29	EQUI0240	Saudi Arabia	Asia
148942.197	EQUI0007	Australia	Australia/NZ
148942.232	EQUI0205	New Zealand	Australia/NZ
148942.237	19	Ireland	Europe
148942.135	EQUI0002	United Kingdom	Europe
148942.151	EQUI0003	United Kingdom	Europe
148942.23	EQUI0005	United Kingdom	Europe
148942.183	EQUI0006	United States	Europe
148942.14	EQUI0008	United Kingdom	Europe
148942.57	EQUI0009	United Kingdom	Europe
148942.51	EQUI0010	United Kingdom	Europe
148942.94	EQUI0011	United Kingdom	Europe
148942.87	EQUI0012	United Kingdom	Europe
148942.127	EQUI0013	United Kingdom	Europe
148942.115	EQUI0014	United Kingdom	Europe
148942.104	EQUI0015	United Kingdom	Europe
148942.138	EQUI0016	United Kingdom	Europe
148942.150	EQUI0017	United Kingdom	Europe
148942.141	EQUI0018	United Kingdom	Europe
148942.225	EQUI0019	Sweden	Europe
148942.216	EQUI0020	United Kingdom	Europe
148942.196	EQUI0021	United Kingdom	Europe
148942.179	EQUI0022	United Kingdom	Europe
148942.67	EQUI0023	United Kingdom	Europe
148942.47	EQUI0024	United Kingdom	Europe
148942.44	EQUI0025	United Kingdom	Europe
148942.26	EQUI0026	United Kingdom	Europe
148942.12	EQUI0027	United Kingdom	Europe
148942.83	EQUI0028	United Kingdom	Europe
148942.73	EQUI0029	United Kingdom	Europe
148942.89	EQUI0030	United Kingdom	Europe
148942.156	EQUI0031	United Kingdom	Europe
148942.131	EQUI0032	United Kingdom	Europe
148942.167	EQUI0033	United Kingdom	Europe
148942.161	EQUI0034	United Kingdom	Europe
148942.193	EQUI0035	United Kingdom	Europe

Table A-1. Continued.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.204	EQUI0036	Ireland	Europe
148942.188	EQUI0037	United Kingdom	Europe
148942.24	EQUI0038	United Kingdom	Europe
148942.38	EQUI0039	United Kingdom	Europe
148942.27	EQUI0040	United Kingdom	Europe
148942.41	EQUI0041	United Kingdom	Europe
148942.61	EQUI0042	United Kingdom	Europe
148942.84	EQUI0043	United Kingdom	Europe
148942.100	EQUI0044	United Kingdom	Europe
148942.107	EQUI0045	United Kingdom	Europe
148942.121	EQUI0046	United Kingdom	Europe
148942.120	EQUI0047	United Kingdom	Europe
148942.190	EQUI0048	United Kingdom	Europe
148942.205	EQUI0049	United Kingdom	Europe
148942.164	EQUI0050	United Kingdom	Europe
148942.157	EQUI0052	United Kingdom	Europe
148942.136	EQUI0053	United Kingdom	Europe
148942.152	EQUI0054	United Kingdom	Europe
148942.103	EQUI0055	United Kingdom	Europe
148942.112	EQUI0058	United Kingdom	Europe
148942.18	EQUI0059	United Kingdom	Europe
148942.64	EQUI0060	United Kingdom	Europe
148942.42	EQUI0061	United Kingdom	Europe
148942.53	EQUI0062	United Kingdom	Europe
148942.180	EQUI0063	United Kingdom	Europe
148942.199	EQUI0064	United Kingdom	Europe
148942.215	EQUI0065	United Kingdom	Europe
148942.226	EQUI0067	United Kingdom	Europe
148942.85	EQUI0068	United Kingdom	Europe
148942.97	EQUI0069	United Kingdom	Europe
148942.163	EQUI0070	United Kingdom	Europe
148942.174	EQUI0071	United Kingdom	Europe
148942.139	EQUI0072	United Kingdom	Europe
148942.133	EQUI0074	United Kingdom	Europe
148942.212	EQUI0075	United Kingdom	Europe
148942.229	EQUI0076	United Kingdom	Europe
148942.187	EQUI0077	United Kingdom	Europe

Table A-1. Continued.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.21	EQUI0078	United Kingdom	Europe
148942.36	EQUI0079	United Kingdom	Europe
148942.55	EQUI0080	United Kingdom	Europe
148942.58	EQUI0081	United Kingdom	Europe
148942.75	EQUI0082	United Kingdom	Europe
148942.88	EQUI0083	United Kingdom	Europe
148942.98	EQUI0084	United Kingdom	Europe
148942.244	EQUI0085	United Kingdom	Europe
148942.123	EQUI0086	United Kingdom	Europe
148942.140	EQUI0087	United Kingdom	Europe
148942.23	EQUI0088	United Kingdom	Europe
148942.35	EQUI0089	United Kingdom	Europe
148942.49	EQUI0090	United Kingdom	Europe
148942.40	EQUI0091	United Kingdom	Europe
148942.62	EQUI0092	United Kingdom	Europe
148942.72	EQUI0094	United Kingdom	Europe
148942.106	EQUI0095	United Kingdom	Europe
148942.122	EQUI0096	United Kingdom	Europe
148942.119	EQUI0097	United Kingdom	Europe
148942.166	EQUI0098	United Kingdom	Europe
148942.159	EQUI0099	United Kingdom	Europe
148942.198	EQUI0100	United Kingdom	Europe
148942.177	EQUI0103	United Kingdom	Europe
148942.214	EQUI0105	United Kingdom	Europe
148942.178	EQUI0106	United Kingdom	Europe
148942.219	EQUI0107	United Kingdom	Europe
148942.145	EQUI0108	United Kingdom	Europe
148942.148	EQUI0109	United Kingdom	Europe
148942.158	EQUI0110	United Kingdom	Europe
148942.168	EQUI0111	United Kingdom	Europe
148942.80	EQUI0112	United Kingdom	Europe
148942.77	EQUI0113	United Kingdom	Europe
148942.91	EQUI0114	United Kingdom	Europe
148942.101	EQUI0115	United Kingdom	Europe
148942.240	EQUI0117	United Kingdom	Europe
148942.30	EQUI0118	United Kingdom	Europe
148942.13	EQUI0119	United Kingdom	Europe

Table A-1. Continued.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.210	EQUI0120	United Kingdom	Europe
148942.220	EQUI0121	United Kingdom	Europe
148942.182	EQUI0122	United Kingdom	Europe
148942.201	EQUI0123	United Kingdom	Europe
148942.200	EQUI0124	United Kingdom	Europe
148942.160	EQUI0125	United Kingdom	Europe
148942.170	EQUI0126	United Kingdom	Europe
148942.134	EQUI0127	United Kingdom	Europe
148942.33	EQUI0128	United Kingdom	Europe
148942.25	EQUI0129	United Kingdom	Europe
148942.66	EQUI0130	United Kingdom	Europe
148942.45	EQUI0131	United Kingdom	Europe
148942.92	EQUI0132	United Kingdom	Europe
148942.76	EQUI0133	United Kingdom	Europe
148942.116	EQUI0134	United Kingdom	Europe
148942.124	EQUI0135	United Kingdom	Europe
148942.111	EQUI0136	United Kingdom	Europe
148942.130	EQUI0137	United Kingdom	Europe
148942.142	EQUI0138	United Kingdom	Europe
148942.132	EQUI0139	United Kingdom	Europe
148942.224	EQUI0140	United Kingdom	Europe
148942.207	EQUI0141	United Kingdom	Europe
148942.206	EQUI0142	United Kingdom	Europe
148942.189	EQUI0143	United Kingdom	Europe
148942.59	EQUI0145	United Kingdom	Europe
148942.48	EQUI0146	United Kingdom	Europe
148942.37	EQUI0147	United Kingdom	Europe
148942.74	EQUI0149	United Kingdom	Europe
148942.86	EQUI0150	United Kingdom	Europe
148942.99	EQUI0151	United Kingdom	Europe
148942.81	EQUI0152	United Kingdom	Europe
148942.93	EQUI0153	United Kingdom	Europe
148942.102	EQUI0154	United Kingdom	Europe
148942.114	EQUI0155	United Kingdom	Europe
148942.19	EQUI0156	United Kingdom	Europe
148942.22	EQUI0157	United Kingdom	Europe
148942.39	EQUI0158	United Kingdom	Europe

Table A-1. Continued.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.203	EQUI0159	United Kingdom	Europe
148942.185	EQUI0160	United Kingdom	Europe
148942.184	EQUI0161	United Kingdom	Europe
148942.223	EQUI0162	United Kingdom	Europe
148942.209	EQUI0163	United Kingdom	Europe
148942.154	EQUI0164	United Kingdom	Europe
148942.137	EQUI0165	United Kingdom	Europe
148942.176	EQUI0166	United Kingdom	Europe
148942.165	EQUI0167	United Kingdom	Europe
148942.109	EQUI0168	United Kingdom	Europe
148942.90	EQUI0169	United Kingdom	Europe
148942.71	EQUI0170	United Kingdom	Europe
148942.56	EQUI0171	United Kingdom	Europe
148942.70	EQUI0172	United Kingdom	Europe
148942.50	EQUI0173	United Kingdom	Europe
148942.32	EQUI0174	United Kingdom	Europe
148942.15	EQUI0175	United Kingdom	Europe
148942.227	EQUI0176	United Kingdom	Europe
148942.217	EQUI0177	United Kingdom	Europe
148942.172	EQUI0178	United Kingdom	Europe
148942.192	EQUI0179	United Kingdom	Europe
148942.195	EQUI0180	United Kingdom	Europe
148942.208	EQUI0181	United Kingdom	Europe
148942.222	EQUI0182	United Kingdom	Europe
148942.228	EQUI0183	United Kingdom	Europe
148942.31	EQUI0184	United Kingdom	Europe
148942.11	EQUI0185	United Kingdom	Europe
148942.69	EQUI0186	United Kingdom	Europe
148942.54	EQUI0187	United Kingdom	Europe
148942.191	EQUI0188	United Kingdom	Europe
148942.173	EQUI0191	United Kingdom	Europe
148942.147	EQUI0192	United Kingdom	Europe
148942.144	EQUI0194	United Kingdom	Europe
148942.117	EQUI0195	United Kingdom	Europe
148942.125	EQUI0196	United Kingdom	Europe
148942.82	EQUI0197	United Kingdom	Europe
148942.10	EQUI0198	United Kingdom	Europe

Table A-1. Continued.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.63	EQUI0199	United Kingdom	Europe
148942.43	EQUI0200	United Kingdom	Europe
148942.181	EQUI0203	Netherlands	Europe
148942.155	EQUI0206	Belgium	Europe
148942.146	EQUI0207	Belgium	Europe
148942.171	EQUI0208	Belgium	Europe
148942.68	EQUI0209	Belgium	Europe
148942.52	EQUI0211	Belgium	Europe
148942.28	EQUI0212	Belgium	Europe
148942.17	EQUI0213	Belgium	Europe
148942.113	EQUI0214	Belgium	Europe
148942.108	EQUI0215	Ireland	Europe
148942.96	EQUI0216	Ireland	Europe
148942.79	EQUI0217	Ireland	Europe
148942.175	EQUI0218	Ireland	Europe
148942.129	EQUI0219	Ireland	Europe
148942.149	EQUI0220	Ireland	Europe
148942.211	EQUI0221	Ireland	Europe
148942.218	EQUI0222	Ireland	Europe
148942.231	EQUI0223	Ireland	Europe
148942.186	EQUI0224	Ireland	Europe
148942.202	EQUI0225	Ireland	Europe
148942.34	EQUI0226	Ireland	Europe
148942.20	EQUI0227	Ireland	Europe
148942.65	EQUI0228	Ireland	Europe
148942.78	EQUI0229	Sweden	Europe
148942.95	EQUI0230	Sweden	Europe
148942.105	EQUI0231	Sweden	Europe
148942.126	EQUI0232	Sweden	Europe
148942.128	EQUI0233	Sweden	Europe
148942.143	EQUI0234	Sweden	Europe
148942.153	EQUI0235	Sweden	Europe
148942.162	EQUI0236	Sweden	Europe
148942.169	EQUI0237	Sweden	Europe
148942.16	EQUI0239	Saudi Arabia	Europe
148942.221	HO51380626	United Kingdom	Europe
148942.236	1-8	United States	North America

Table A-1. Continued.

PATRIC Genome ID	Strain	Isolation Country	Continent
148942.235	CF22	United States	North America
148942.245	E12	United States	North America
148942.213	EQUI0004	Canada	North America
148942.247	F43	United States	North America
148942.246	Flint	United States	North America
148942.248	Lex90	United States	North America
148942.46	EQUI0201	United States	NorthAmerica
148942.194	EQUI0202	United States	NorthAmerica

A-2 Table. Colony morphology of the 54 Texas and Kentucky SEE isolates.

Isolate ID	Form	Elevation	Margin	Color
11-002	Circular	Umbonate	Entire	White
11-004	Circular	Umbonate	Entire	White
11-006	Circular	Umbonate	Entire	White
11-008	Circular	Umbonate	Entire	White
11-010	Circular	Umbonate	Entire	White
11-014	Circular	Umbonate	Entire	White
11-017	Circular	Umbonate	Entire	White
11-018	Circular	Umbonate	Entire	White
14-052	Circular	Umbonate	Entire	White
14-057	Circular	Umbonate	Entire	White
14-061	Circular	Umbonate	Entire	White
14-066	Circular	Convex	Entire	Salmon
14-071	Circular	Umbonate	Entire	White
14-073	Circular	Umbonate	Entire	White
14-080	Circular	Umbonate	Entire	White
14-082	Circular	Umbonate	Entire	White
14-092	Circular	Umbonate	Entire	White
14-105	Circular	Convex	Entire	Salmon
14-112	Circular	Umbonate	Entire	White
14-125	Circular	Umbonate	Entire	White
14-133	Circular	Convex	Entire	Salmon
14-140	Circular	Umbonate	Entire	White
14-146	Circular	Umbonate	Entire	White
14-148	Circular	Convex	Entire	Salmon

Table A-2. Continued.

Isolate ID	Form	Elevation	Margin	Color
14-150	Circular	Umbonate	Entire	White
17-003	Circular	Raised	Entire	White
17-004	Circular	Umbonate	Entire	White
17-007	Circular	Raised	Entire	White
17-008	Circular	Raised	Entire	White
17-009	Circular	Umbonate	Entire	White
18-001	Circular	Raised	Entire	White
18-002	Circular	Raised	Entire	White
18-003	Circular	Raised	Entire	White
18-004	Circular	Raised	Entire	White
18-006	Circular	Raised	Entire	White
18-008	Circular	Convex	Entire	Salmon
18-009	Circular	Umbonate	Entire	White
18-011	Circular	Raised	Entire	White
18-012	Circular	Raised	Entire	White
18-013	Circular	Raised	Entire	White
18-014	Circular	Raised	Entire	White
18-015	Circular	Raised	Entire	White
18-018	Circular	Raised	Entire	White
18-021	Circular	Raised	Entire	White
18-022	Circular	Raised	Entire	White
18-024	Circular	Raised	Entire	White
18-025	Circular	Umbonate	Entire	White
18-026	Circular	Umbonate	Entire	White
18-027	Circular	Umbonate	Entire	White
18-028	Circular	Umbonate	Entire	White
18-037	Circular	Raised	Entire	White
18-039	Circular	Raised	Entire	White
18-078	Circular	Raised	Entire	White
18-079	Circular	Raised	Entire	White

A-3 Table. Accession numbers for the submission of the 54 SEE isolates to the NCBI Sequence Read Archive (SRA) and Genbank.

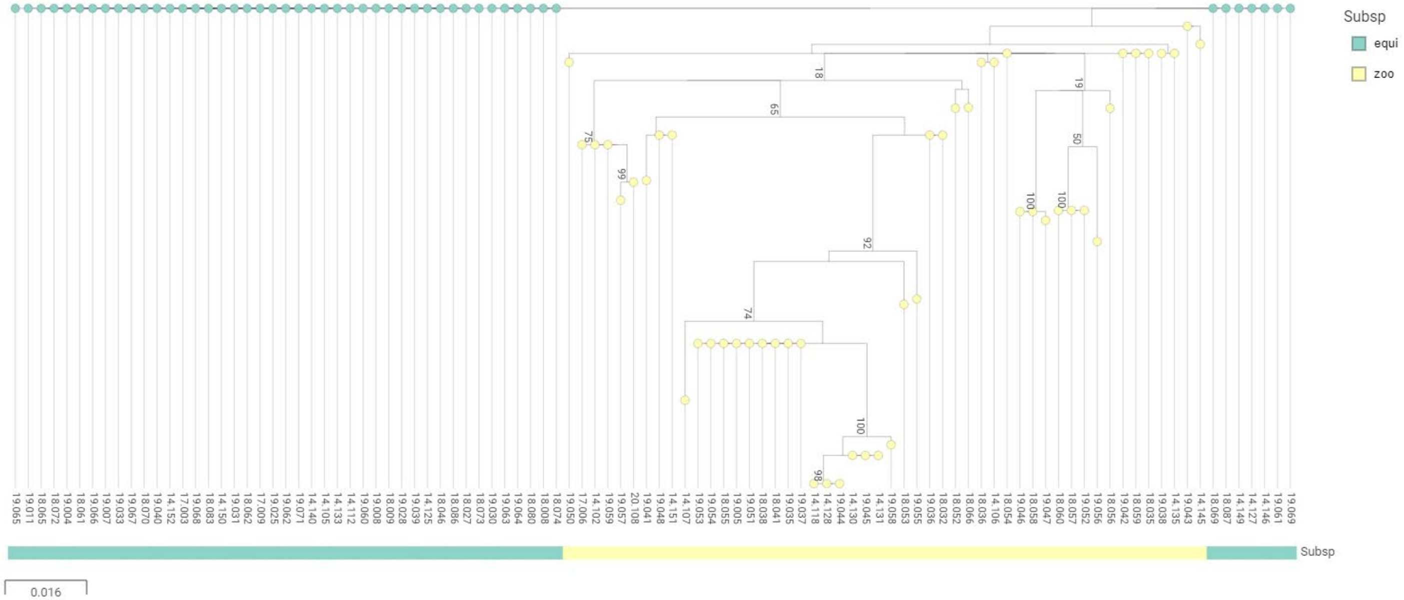
IsolateID	BioProject	SRA_BioSampleAccession	GenBank_BioSampleAccession
11-004	PRJNA575530	SAMN12898329	SAMN12908139
11-006	PRJNA575530	SAMN12898330	SAMN12908140
11-002	PRJNA575530	SAMN12898331	SAMN12908141
11-008	PRJNA575530	SAMN12898332	SAMN12908142
11-010	PRJNA575530	SAMN12898333	SAMN12908143
11-014	PRJNA575530	SAMN12898334	SAMN12908144
11-017	PRJNA575530	SAMN12898335	SAMN12908145
11-018	PRJNA575530	SAMN12898336	SAMN12908146
17-007	PRJNA575530	SAMN12898337	SAMN12908147
17-008	PRJNA575530	SAMN12898338	SAMN12908148
17-003	PRJNA575530	SAMN12898339	SAMN12908149
17-004	PRJNA575530	SAMN12898340	SAMN12908150
18-001	PRJNA575530	SAMN12898341	SAMN12908151
18-002	PRJNA575530	SAMN12898342	SAMN12908152
18-003	PRJNA575530	SAMN12898343	SAMN12908153
18-004	PRJNA575530	SAMN12898344	SAMN12908154
18-006	PRJNA575530	SAMN12898345	SAMN12908155
18-011	PRJNA575530	SAMN12898346	SAMN12908156
18-012	PRJNA575530	SAMN12898347	SAMN12908157
18-013	PRJNA575530	SAMN12898348	SAMN12908158
18-014	PRJNA575530	SAMN12898349	SAMN12908159
18-015	PRJNA575530	SAMN12898350	SAMN12908160
18-018	PRJNA575530	SAMN12898351	SAMN12908161
18-021	PRJNA575530	SAMN12898352	SAMN12908162
18-022	PRJNA575530	SAMN12898353	SAMN12908163
18-024	PRJNA575530	SAMN12898354	SAMN12908164
14-052	PRJNA575530	SAMN12898355	SAMN12908165
14-057	PRJNA575530	SAMN12898356	SAMN12908166
14-061	PRJNA575530	SAMN12898357	SAMN12908167
14-066	PRJNA575530	SAMN12898358	SAMN12908168
14-071	PRJNA575530	SAMN12898359	SAMN12908169
14-073	PRJNA575530	SAMN12898360	SAMN12908170
14-080	PRJNA575530	SAMN12898361	SAMN12908171
14-082	PRJNA575530	SAMN12898362	SAMN12908172
14-092	PRJNA575530	SAMN12898363	SAMN12908173
14-105	PRJNA575530	SAMN12898364	SAMN12908174

Table A-3. Continued.

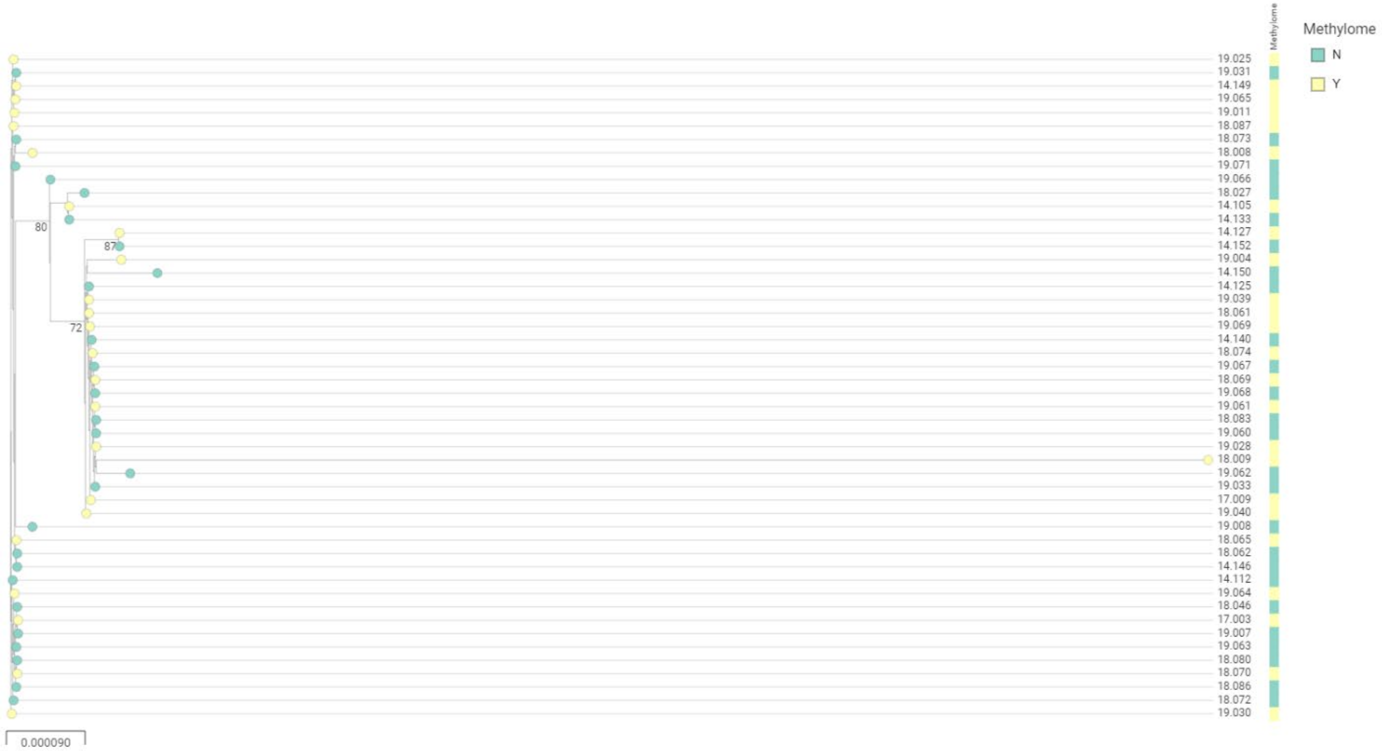
IsolateID	BioProject	SRA_BioSampleAccession	GenBank_BioSampleAccession
14-112	PRJNA575530	SAMN12898365	SAMN12908175
14-125	PRJNA575530	SAMN12898366	SAMN12908176
14-133	PRJNA575530	SAMN12898367	SAMN12908177
14-140	PRJNA575530	SAMN12898368	SAMN12908178
14-146	PRJNA575530	SAMN12898369	SAMN12908179
14-148	PRJNA575530	SAMN12898370	SAMN12908180
14-150	PRJNA575530	SAMN12898371	SAMN12908181
17-009	PRJNA575530	SAMN12898372	SAMN12908182
18-008	PRJNA575530	SAMN12898373	SAMN12908183
18-009	PRJNA575530	SAMN12898374	SAMN12908184
18-025	PRJNA575530	SAMN12898375	SAMN12908185
18-026	PRJNA575530	SAMN12898376	SAMN12908186
18-027	PRJNA575530	SAMN12898377	SAMN12908187
18-028	PRJNA575530	SAMN12898378	SAMN12908188
18-037	PRJNA575530	SAMN12898379	SAMN12908189
18-039	PRJNA575530	SAMN12898380	SAMN12908190
18-078	PRJNA575530	SAMN12898381	SAMN12908191
18-079	PRJNA575530	SAMN12898382	SAMN12908192

APPENDIX B

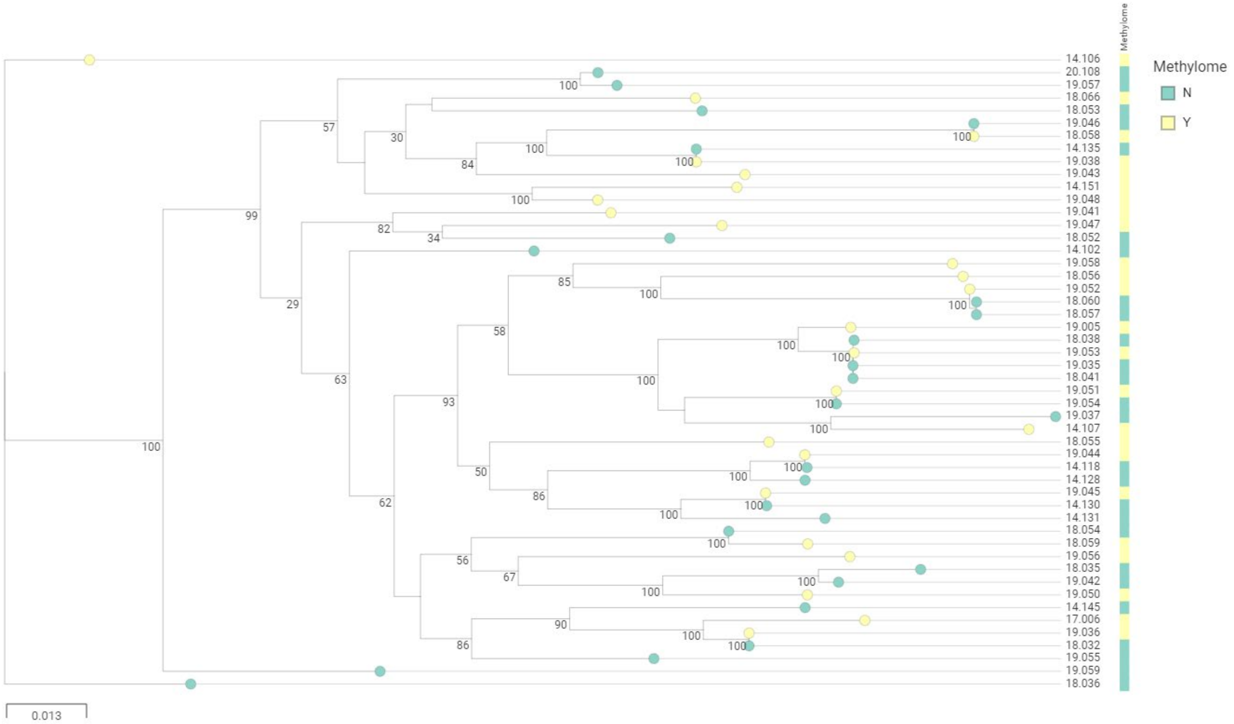
SUPPLEMENTARY MATERIAL: DIFFERENCES IN THE ACCESSORY GENOMES AND METHYLOMES OS STRAINS OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* AND OF *STREPTOCOCCUS EQUI* SUBSP. *ZOOEPIDEMICUS* OBTAINED FROM THE RESPIRATORY TRACT OF HORSES FROM TEXAS



B-1 Fig. Phylogenetic tree for SEE (n = 50) and respiratory SEZ (n = 50). Phylogenetic comparisons demonstrate the separation of the isolates by the respective subspecies. The blue squared denote the SEE isolates, and the yellow squares denote the SEZ isolates.



B-2 Fig. Phylogenetic tree of SEE (n = 50) isolates. A subset of SEE isolates (n = 24) selected for the methylome analysis based on relatedness in the phylogenetic tree. The blue squares denote SEE isolates that were not (N) used in the methylation analysis, whereas the yellow squares denote SEE isolates that were (Y) used in the methylation analysis.



B-3 Fig. Phylogenetic tree of SEZ (n = 50) isolates**.** A subset of SEZ isolates (n = 24) selected for the methylome analysis based on relatedness in the phylogenetic tree. The blue squares denote SEZ isolates that were not (N) used in the methylation analysis, whereas the yellow squares denote SEZ isolates that were (Y) used in the methylation analysis.

B-1 Table. Metadata for all 50 SEE and 50 SEZ genomes.

No	Isolate ID	Subsp	Year	Clinical Source	Location	Collection Source	Notes	Methylome
1	14-105	equi	2014	Abscess		TVMDL		Y
2	14-112	equi	2014	abscess		2014 TVMDL		N
3	14-125	equi	2014	abscess		TVMDL		N
4	14-127	equi	2014	abscess		TVMDL		Y
5	14-133	equi	2014	nasal		2014 TVMDL		N
6	14-140	equi	2014	guttural pouch		2014 TVMDL		N
7	14-146	equi	2014	guttural pouch		2014 TVMDL		N
8	14-149	equi	2014	gluttural pouch		TVMDL		Y
9	14-150	equi	2014	lymph node		2014 TVMDL		N
10	14-152	equi	2014	respiratory		TVMDL		N
11	17-003	equi	2017	nasal	College Station, TX	VetMed Outbreak		Y
12	17-009	equi	2017	lavage fluid	College Station, TX	ANSC	Dr. Leatherwood's herd	Y
13	18-008	equi	2018	abscess	Brazos Valley County	2018 TVMDL		Y
14	18-009	equi	2018	guttural pouch	College Station, TX	ANSC		Y
15	18-027	equi	2018	guttural pouch	College Station, TX	ANSC	Pinnacle Vx disease	N
16	18-046	equi	2018	nasal swab	Waller, TX	TVMDL		N
17	18-061	equi	2018	abscess	Waller, TX	TVMDL		Y
18	18-062	equi	2018	abcess	Victoria, TX	TVMDL		N
19	18-065	equi	2018	Guttural pouch	Bandera, TX	TVMDL		Y
20	18-069	equi	2018	Lymph Node	Kilgore, TX	TVMDL		Y
21	18-070	equi	2018	Wound	Weatherford, TX	TVMDL		Y
22	18-072	equi	2018	abscess	Bryan, TX	TVMDL		N
23	18-073	equi	2018	abscess	Weatherford, TX	TVMDL		N
24	18-074	equi	2018	abscess	Salado, TX	TVMDL		Y

Table B-1. Continued.

No	Isolate ID	Subsp	Year	Clinical Source	Location	Collection Source	Notes	Methylome
25	18-080	equi	2018	Nasal Swab	Needville TX	TVMDL		N
26	18-083	equi	2018	Gutteral Pouch	Aubrey, TX	TVMDL		N
27	18-086	equi	2018	submandibular lymph node aspirate	Lipan, TX	Clin Micro		N
28	18-087	equi	2018	Guttural Pouches		Clin Micro		Y
29	19-004	equi	2019	Guttural Pouch Lavage	Driftwood, TX	Clin Micro		Y
30	19-007	equi	2019	Guttural Pouch Lavage	Montgomery TX	Clin Micro		N
31	19-008	equi	2019	Retropharyngeal lymph node abscess	College Station, Tx	Clin Micro		N
32	19-011	equi	2019	abscessed lymph node	Anna, TX	TVMDL		Y
33	19-025	equi	2019	nasal wash	College Station, TX	VLCS	chronic carrier	Y
34	19-028	equi	2019	Guttural pouch	Weatherford, TX	TVMDL		Y
35	19-030	equi	2019	Draining abscess	Falfurrias, TX	TVMDL		Y
36	19-031	equi	2019	Abscess swab	Victoria, TX	TVMDL		N
37	19-033	equi	2019	Pus	Bryan, TX	TVMDL		N
38	19-039	equi	2019	Abscess swab	Aubrey, TX	TVMDL		Y
39	19-040	equi	2019	Abscess swab	Weatherford, TX	TVMDL		Y
40	19-060	equi	2014	abscess swab	College Station, Tx	Clin Micro		N
41	19-061	equi	2012	Nasal wash	Richmond, TX	Clin Micro		Y
42	19-062	equi	2014	abscess swab	Houston, TX	Clin Micro		N
43	19-063	equi	2017	abscess swab	Austin, TX	Clin Micro		N
44	19-064	equi	2015	nasal wash	Dripping Springs, TX	Clin Micro		Y
45	19-065	equi	2016	guttural pouch lavage	San Antonio, TX	Clin Micro		Y
46	19-066	equi	2016	guttural pouch lavage	Bellville, TX	Clin Micro		N
47	19-067	equi	2016	lymph node aspirate	College Station, TX	Clin Micro		N

Table B-1. Continued.

No	Isolate ID	Subsp	Year	Clinical Source	Location	Collection Source	Notes	Methylome
48	19-068	equi	2017	guttural pouch lavage	Crockett, TX	Clin Micro		N
49	19-069	equi	2018	Nasal wash	Aubrey, TX	Clin Micro		Y
50	19-071	equi	2019	abscess	Gonzales, TX	TVMDL		N
51	14-102	zoo	2014	Lung		TVMDL		N
52	14-106	zoo	2014	TTW		TVMDL		Y
53	14-107	zoo	2014	TTW		TVMDL		Y
54	14-118	zoo	2014	pharyngeal		TVMDL		N
55	14-128	zoo	2014	TTW		TVMDL		N
56	14-130	zoo	2014	nasal		TVMDL		N
57	14-131	zoo	2014	nasal		TVMDL		N
58	14-135	zoo	2014	TTW		TVMDL		N
59	14-145	zoo	2014	pleural effusion		TVMDL		N
60	14-151	zoo	2014	nasopharyngeal		TVMDL		Y
61	17-006	zoo	2017	nasal swab		VMP-8	Commensal	Y
62	18-032	zoo	2018	GP lavage		VETMED	Commensal	N
63	18-035	zoo	2018	abscess		TVMDL		N
64	18-036	zoo	2018	GP lavage		VETMED		N
65	18-038	zoo	2018	GP lavage		VETMED		N
66	18-041	zoo	2018	GP lavage		VETMED		N
67	18-052	zoo	2010	TTW		Clin Micro		N
68	18-053	zoo	2013	lung		Clin Micro		N
69	18-054	zoo	2013	Caudal maxillary sinus tissues		Clin Micro		N
70	18-055	zoo	2013	Guttural pouch		Clin Micro		Y

Table B-1. Continued.

No	Isolate ID	Subsp	Year	Clinical Source	Location	Collection Source	Notes	Methylome
71	18-056	zoo	2013	TTW		Clin Micro		Y
72	18-057	zoo	2013	Tracheal aspirate		Clin Micro		N
73	18-058	zoo	2014	nasal wash		Clin Micro		Y
74	18-059	zoo	2014	pharyngeal wash sample		Clin Micro		Y
75	18-060	zoo	2015	Guttural pouch		Clin Micro		N
76	18-066	zoo	2018	Guttural pouch		TVMDL		Y
77	19-005	zoo	2019	TTW,		Clin Micro		Y
78	19-035	zoo	2019	Nasal swab		TVMDL		N
79	19-036	zoo	2019	Nasal swab		TVMDL		Y
80	19-037	zoo	2019	Sinus fluid and swab		TVMDL		N
81	19-038	zoo	2019	TTW		TVMDL		Y
82	19-041	zoo	2010	TTW		Clin Micro	Commensal	Y
83	19-042	zoo	2010	pleural fluid		Clin Micro	pathogenic	N
84	19-043	zoo	2014	check abscess		Clin Micro	pathogenic	Y
85	19-044	zoo	2013	sinus swab		Clin Micro	pathogenic	Y
86	19-045	zoo	2013	sinus swab		Clin Micro	pathogenic	Y
87	19-046	zoo	2014	nasal wash		Clin Micro	pathogenic	N
88	19-047	zoo	2018	nasal wash		Clin Micro	Commensal	Y
89	19-048	zoo	2018	nasal wash		Clin Micro	Commensal	Y
90	19-050	zoo	2018	sinus fluid		Clin Micro	pathogenic	Y
91	19-051	zoo	2018	nasal wash		Clin Micro	commensal	Y
92	19-052	zoo	2018	guttural pouch lavage		Clin Micro	Commensal	Y
93	19-053	zoo	2018	nasal wash		Clin Micro	Commensal	Y

Table B-1. Continued.

No	Isolate ID	Subsp	Year	Clinical Source	Location	Collection Source	Notes	Methylome
94	19-054	zoo	2018	nasal wash		Clin Micro	Commensal	N
95	19-055	zoo	2018	guttural pouch lavage		Clin Micro	Commensal	N
96	19-056	zoo	2018	pleural fluid		Clin Micro	pathogenic	Y
97	19-057	zoo	2018	nasal wash		Clin Micro	commensal	N
98	19-058	zoo	2018	guttural pouch lavage		Clin Micro	commensal	Y
99	19-059	zoo	2019	Lung tissue sample		Clin Micro	pathogenic	N
100	20-108	zoo	2020	TTW		TVMDL	pathogenic	N

B-2 Table. ClueGO analysis summary of the AGE found SEE (n = 50) genomes.

GOID	GOTerm	Term PValue	Adjusted Term PValue	Group PValue	Ajusted Group PValue	Associated Genes Found
KEGG: 01053	Biosynthesis of siderophore group nonribosomal peptides	0.00	0.00	0.00	0.00	[SEQ_1243, SEQ_1242, SEQ_1240]
KEGG: 05150	Staphylococcus aureus infection	0.00	0.01	0.00	0.00	[SEQ_2036, SEQ_2037, SEQ_1728]
GO:0006304	DNA modification	0.00	0.03	0.11	0.34	[SEQ_0757, SEQ_0758, SEQ_1262]
GO:0003677	DNA binding	0.03	0.17	0.11	0.34	[SEQ_0756, SEQ_0757, SEQ_0758, SEQ_0787, SEQ_1102, SEQ_1231, SEQ_1252, SEQ_1262, SEQ_1762, SEQ_1246]
GO:0004519	endonuclease activity	0.05	0.25	0.12	0.23	[SEQ_0758, SEQ_0817, SEQ_0818]
GO:0016887	ATPase activity	0.35	0.35	0.35	0.35	[SEQ_1269, SEQ_1237, SEQ_1236, SEQ_1235]
GO:0006259	DNA metabolic process	0.20	0.40	0.11	0.34	[SEQ_0757, SEQ_0758, SEQ_0787, SEQ_1102, SEQ_1262]
GO:0090305	nucleic acid phosphodiester bond hydrolysis	0.14	0.41	0.12	0.23	[SEQ_0758, SEQ_0817, SEQ_0818]
GO:0004518	nuclease activity	0.14	0.41	0.12	0.23	[SEQ_0758, SEQ_0817, SEQ_0818]
GO:0016788	hydrolase activity, acting on ester bonds	0.12	0.47	0.12	0.23	[SEQ_0758, SEQ_0817, SEQ_0818, SEQ_1245]

B-3 Table. ClueGO analysis summary of the AGE found SEZ (n = 50) genomes.

GOID	GOTerm	Term PValue	Adjusted Term PValue	Group PValue	Adjusted Group PValue	Associated Genes Found
KEGG:00052	Galactose metabolism	0.00	0.00	0.00	0.00	[SZO_15230, SZO_15240, SZO_15220]

B-4 Table. Motif summary of global methylomes for SEE (n = 24) and SEZ (n = 24) genomes.

Genome ID	Subsp.	Motif Sequence	Center Pos	Modification Type	Fraction	Partner Motif Sequence	Mean Score	Mean IPD Ratio	Mean Coverage
14-105	equi	CATCC	2	m6A	0.97383		327.274	5.41924	215.545
14-127	equi	CATCC	2	m6A	0.98411		314.03	5.25248	212.204
17-009	equi	CATCC	2	m6A	0.98598		736.355	5.42285	568.592
18-009	equi	CATCC	2	m6A	0.95607		77.3578	5.24149	42.8583
18-061	equi	CATCC	2	m6A	0.9729		211.494	5.24969	135.282
18-069	equi	CATCC	2	m6A	0.98131		258.939	4.89464	184.764
18-074	equi	CATCC	2	m6A	0.97103		129.962	5.07654	79.1232
19-004	equi	CATCC	2	m6A	0.97664		306.448	5.47036	200.939
19-028	equi	CATCC	2	m6A	0.98037		341.501	5.23829	236.692
19-039	equi	CATCC	2	m6A	0.98411		492.208	5.36737	349.808
19-040	equi	CATCC	2	m6A	0.97383		426.21	5.41717	290.262
19-061	equi	CATCC	2	m6A	0.9785		322.648	5.35569	215.701
19-069	equi	CATCC	2	m6A	0.98224		352.395	5.32515	239.547
14-105	equi	CTGCAG	5	m6A	0.96585	CTGCAG	318.755	6.09583	221.216
14-127	equi	CTGCAG	5	m6A	0.97073	CTGCAG	305.487	5.96785	215.534
14-149	equi	CTGCAG	5	m6A	0.95366	CTGCAG	402.26	5.96997	293.111
17-003	equi	CTGCAG	5	m6A	0.94878	CTGCAG	373.341	6.30903	257.791
17-009	equi	CTGCAG	5	m6A	0.96585	CTGCAG	750.626	6.24939	573.891

Table B-4. Continued.

Genome ID	Subsp.	Motif Sequence	Center Pos	Modification Type	Fraction	Partner Motif Sequence	Mean Score	Mean IPD Ratio	Mean Coverage
18-008	equi	CTGCAG	5	m6A	0.92195	CTGCAG	72.1243	6.11037	42
18-009	equi	CTGCAG	5	m6A	0.94268	CTGCAG	68.0893	6.01758	39.5783
18-061	equi	CTGCAG	5	m6A	0.9561	CTGCAG	203.176	5.94216	137.767
18-065	equi	CTGCAG	5	m6A	0.94878	CTGCAG	416.959	5.63573	313.892
18-069	equi	CTGCAG	5	m6A	0.97073	CTGCAG	262.391	5.66261	186.411
18-070	equi	CTGCAG	5	m6A	0.94878	CTGCAG	263.461	5.6315	186.618
18-074	equi	CTGCAG	5	m6A	0.95488	CTGCAG	121.797	5.63307	78.3346
18-087	equi	CTGCAG	5	m6A	0.94878	CTGCAG	315.186	6.23387	217.712
19-004	equi	CTGCAG	5	m6A	0.96585	CTGCAG	287.578	6.30558	193.696
19-011	equi	CTGCAG	5	m6A	0.94878	CTGCAG	302.042	6.2246	206.72
19-025	equi	CTGCAG	5	m6A	0.95122	CTGCAG	349.31	6.35534	237.455
19-028	equi	CTGCAG	5	m6A	0.96829	CTGCAG	340.025	6.02206	238.73
19-030	equi	CTGCAG	5	m6A	0.94878	CTGCAG	307.107	6.04625	213.335
19-039	equi	CTGCAG	5	m6A	0.96585	CTGCAG	480.189	6.04991	351.163
19-040	equi	CTGCAG	5	m6A	0.9561	CTGCAG	398.844	6.07509	283.741
19-061	equi	CTGCAG	5	m6A	0.96585	CTGCAG	312.447	6.07676	216.271
19-064	equi	CTGCAG	5	m6A	0.94878	CTGCAG	373.959	6.0437	264.436
19-065	equi	CTGCAG	5	m6A	0.94878	CTGCAG	313.325	6.05207	217.887
19-069	equi	CTGCAG	5	m6A	0.96707	CTGCAG	344.494	6.0358	242.044
18-061	equi	GGATGH	3	m6A	0.28794		48.8243	1.95027	143.05
18-074	equi	GGATGNND	3	m6A	0.16293		42.5145	2.17029	85.3551
19-050	zoo	AAGANNNNNGGT	4	m6A	0.76995	ACCNNNNNTCTT	318.445	4.98134	236.104
14-006	zoo	ACAYNNNNNRGG	3	m6A	0.75887		415.657	5.41682	293.178

Table B-4. Continued.

Genome ID	Subsp.	Motif Sequence	Center Pos	Modification Type	Fraction	Partner Motif Sequence	Mean Score	Mean IPD Ratio	Mean Coverage
19-052	zoo	ACCCA	5	m6A	0.83396		222.417	5.17699	161.914
19-050	zoo	ACCNNNNNTCTT	1	m6A	0.76995	AAGANNNNNGGT	302.335	4.37482	236.128
19-044	zoo	AGTNNNNNNGTC	1	m6A	0.91812	GACNNNNNNACT	222.398	4.9719	166.269
19-050	zoo	AGTNNNNNNGTC	1	m6A	0.78223	GACNNNNNNACT	303.9	5.04107	236.022
14-006	zoo	CATCC	2	m6A	0.71152	GGATG	436.835	5.75398	281.248
17-006	zoo	CATCC	2	m6A	0.8073	GGATG	333.459	5.428	241.739
18-056	zoo	CATCC	2	m6A	0.626	GGATG	98.9016	5.34847	56.4408
18-058	zoo	CATCC	2	m6A	0.49487	GGATG	63.1567	5.54065	31.7442
18-066	zoo	CATCC	2	m6A	0.6317	GGATG	269.551	5.27038	174.278
19-036	zoo	CATCC	2	m6A	0.63968	GGATG	280.234	5.24904	182.578
19-038	zoo	CATCC	2	m6A	0.63968	GGATG	349.89	5.20144	238.504
19-041	zoo	CATCC	2	m6A	0.67959	GGATG	281.656	5.31602	181.126
19-044	zoo	CATCC	2	m6A	0.90764	GGATG	254.753	5.11192	165.367
19-045	zoo	CATCC	2	m6A	0.65336	GGATG	266.901	5.09318	177.394
19-047	zoo	CATCC	2	m6A	0.66819	GGATG	479.314	5.09189	353.186
19-050	zoo	CATCC	2	m6A	0.75941	GGATG	353.141	5.11831	243.027
14-006	zoo	GGATG	3	m6A	0.7138	CATCC	414.479	5.3238	282.064
17-006	zoo	GGATG	3	m6A	0.80844	CATCC	318.268	4.99244	241.183
18-056	zoo	GGATG	3	m6A	0.62486	CATCC	94.6058	5.02631	56.6679
18-058	zoo	GGATG	3	m6A	0.48119	CATCC	60.9526	5.16497	32.1043
18-066	zoo	GGATG	3	m6A	0.63284	CATCC	256.805	4.93153	174.137
19-036	zoo	GGATG	3	m6A	0.64424	CATCC	263.611	4.83671	181.685
19-038	zoo	GGATG	3	m6A	0.63398	CATCC	335.836	4.93146	238.038

Table B-4. Continued.

Genome ID	Subsp.	Motif Sequence	Center Pos	Modification Type	Fraction	Partner Motif Sequence	Mean Score	Mean IPD Ratio	Mean Coverage
19-041	zoo	GGATG	3	m6A	0.68187	CATCC	260.684	4.84707	181.159
19-044	zoo	GGATG	3	m6A	0.9065	CATCC	237.774	4.67949	164.931
19-045	zoo	GGATG	3	m6A	0.65564	CATCC	251.685	4.71115	177.489
19-047	zoo	GGATG	3	m6A	0.66705	CATCC	444.427	4.68894	353.456
19-050	zoo	GGATG	3	m6A	0.76397	CATCC	325.775	4.64767	244.1
18-066	zoo	CCANNNNNNNNNTAC	3	m6A	0.76693	GTANNNNNNNNNTGG	187.979	4.24659	176.191
14-151	zoo	CCANNNNNNTGA	3	m6A	0.80893	TCANNNNNNTGG	301.829	6.21469	207.565
19-048	zoo	CCANNNNNNTGA	3	m6A	0.79289	TCANNNNNNTGG	337.83	5.22095	256.522
18-059	zoo	CTCCAG	5	m6A	0.75802	CTGGAG	94.2134	6.61441	55.3029
19-043	zoo	CTCCAG	5	m6A	0.76667	CTGGAG	300.151	5.942	208.931
19-044	zoo	CTCCAG	5	m6A	0.95062	CTGGAG	239.403	5.85247	164.778
18-059	zoo	CTGGAG	5	m6A	0.75432	CTCCAG	90.9084	5.99157	55.1457
19-043	zoo	CTGGAG	5	m6A	0.7642	CTCCAG	281.389	5.30603	209.278
19-044	zoo	CTGGAG	5	m6A	0.94691	CTCCAG	226.021	5.19275	164.708
19-047	zoo	CYTANNNNNGTC	4	m6A	0.80211	GACNNNNNTARG	411.551	4.94677	363.336
19-038	zoo	GAANNNNNNNNTGC	3	m6A	0.80198	GCANNNNNNNNTTC	311.886	4.84451	241.441
19-044	zoo	GACNNNNNNACT	2	m6A	0.9216	AGTNNNNNNGTC	221.117	4.50038	167.227
19-050	zoo	GACNNNNNNACT	2	m6A	0.78223	AGTNNNNNNGTC	298.9	4.48508	238.056
19-041	zoo	GACNNNNNTARG	2	m6A	0.77053		230.014	4.4971	184.363
19-047	zoo	GACNNNNNTARG	2	m6A	0.80632	CYTANNNNNGTC	390.616	4.33172	364.961
19-058	zoo	GATC	2	m6A	0.83341	GATC	193.328	5.3576	117.045
19-056	zoo	GATGC	2	m6A	0.7968	GCATC	369.442	5.05178	259.527
19-038	zoo	GCANNNNNNNNTTC	3	m6A	0.80468	GAANNNNNNNNTGC	321.113	5.19038	241.47

Table B-4. Continued.

Genome ID	Subsp.	Motif Sequence	Center Pos	Modification Type	Fraction	Partner Motif Sequence	Mean Score	Mean IPD Ratio	Mean Coverage
19-056	zoo	GCATC	3	m6A	0.79497	GATGC	361.007	5.23803	260.06
19-045	zoo	GCTANAC	6	m6A	0.77072		234.027	4.41822	179.51
18-066	zoo	GTANNNNNNNNNTGG	3	m6A	0.76535	CCANNNNNNNNNTAC	184.87	4.09981	176.451
19-005	zoo	NA							
19-051	zoo	NA							
19-053	zoo	NA							
18-058	zoo	RAACNNNNNTGA	3	m6A	0.51783	TCANNNNNGTTY	48.6918	4.45082	30.5607
14-007	zoo	RGATCY	5	m4C	0.73629	RGATCY	328.635	3.78087	332.401
18-055	zoo	RGATCY	5	m4C	0.53916	RGATCY	49.8345	3.90952	32.9387
18-058	zoo	TCANNNNNGTTY	3	m6A	0.57895	RAACNNNNNTGA	52.6833	5.89709	29.6921
14-151	zoo	TCANNNNNNTGG	3	m6A	0.81311	CCANNNNNNTGA	306.968	6.37646	207.663
19-048	zoo	TCANNNNNNTGG	3	m6A	0.79568	CCANNNNNNTGA	342.672	5.31218	256.494
17-006	zoo	TCCAG	4	m6A	0.80622		159.074	3.50988	256.81
19-036	zoo	TCCAG	4	m6A	0.80958		266.225	5.97995	183.083
19-058	zoo	YACNNNNNGTR	2	m6A	0.8165	YACNNNNNGTR	155.258	4.10208	118.162

B-5 Table. ClueGO analysis summary of sites of methylation in SEE, but absence of methylation in SEZ.

GOID	GOTerm	Term PValue	Adjusted Term PValue	Group PValue	Adjusted Group PValue	Associated Genes Found
KEGG:00640	Propanoate metabolism	0.01	0.23	0.01	0.05	[SEQ_0045, SEQ_1625, SEQ_1627]
KEGG:02024	Quorum sensing	0.09	1.00	0.09	0.28	[SEQ_1918, SEQ_2009, SEQ_0435]
GO:1901575	organic substance catabolic process	0.03	0.75	0.03	0.13	[SEQ_0769, SEQ_1278, SEQ_0898, SEQ_0976]
GO:0044248	cellular catabolic process	0.04	1.00	0.03	0.13	[SEQ_0769, SEQ_1278, SEQ_0976]
GO:0046914	transition metal ion binding	0.13	0.91	0.07	0.29	[SEQ_0045, SEQ_0300, SEQ_0976]
GO:0008270	zinc ion binding	0.07	1.00	0.07	0.29	[SEQ_0045, SEQ_0300, SEQ_0976]
GO:0043169	cation binding	0.07	1.00	0.07	0.29	[SEQ_0045, SEQ_1278, SEQ_1597, SEQ_2210, SEQ_0898, SEQ_0300, SEQ_0976, SEQ_0435]
GO:0046872	metal ion binding	0.07	1.00	0.07	0.29	[SEQ_0045, SEQ_1278, SEQ_1597, SEQ_2210, SEQ_0898, SEQ_0300, SEQ_0976, SEQ_0435]
GO:0016817	hydrolase activity, acting on acid anhydrides	0.21	0.21	0.21	0.42	[SEQ_1277, SEQ_1407, SEQ_1410, SEQ_1129, SEQ_2152]
GO:0016818	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	0.21	0.21	0.21	0.42	[SEQ_1277, SEQ_1407, SEQ_1410, SEQ_1129, SEQ_2152]
GO:0016462	pyrophosphatase activity	0.21	0.42	0.21	0.42	[SEQ_1277, SEQ_1407, SEQ_1410, SEQ_1129, SEQ_2152]
GO:0017111	nucleoside-triphosphatase activity	0.19	0.77	0.21	0.42	[SEQ_1277, SEQ_1407, SEQ_1410, SEQ_1129, SEQ_2152]
GO:0016887	ATPase activity	0.09	1.00	0.21	0.42	[SEQ_1277, SEQ_1407, SEQ_1410, SEQ_1129, SEQ_2152]
GO:0006810	transport	0.01	0.31	0.01	0.10	[SEQ_0251, SEQ_0497, SEQ_1277, SEQ_1299, SEQ_1439, SEQ_1615, SEQ_1895, SEQ_1918, SEQ_1129, SEQ_0435]
GO:0055085	transmembrane transport	0.00	0.08	0.01	0.10	[SEQ_0251, SEQ_0497, SEQ_1277, SEQ_1299, SEQ_1439, SEQ_1615, SEQ_1895, SEQ_1918, SEQ_1129, SEQ_0435]
GO:0022857	transmembrane transporter activity	0.02	0.63	0.01	0.10	[SEQ_0251, SEQ_0497, SEQ_1277, SEQ_1299, SEQ_1615, SEQ_1895, SEQ_1129]
GO:0005886	plasma membrane	0.01	0.24	0.01	0.10	[SEQ_0251, SEQ_0497, SEQ_1299, SEQ_1318, SEQ_1439, SEQ_1615, SEQ_1918, SEQ_2009, SEQ_1129, SEQ_0435]
KEGG:02010	ABC transporters	0.01	0.30	0.01	0.10	[SEQ_0251, SEQ_1277, SEQ_1439, SEQ_1448, SEQ_1918, SEQ_1129]
GO:0071702	organic substance transport	0.07	1.00	0.01	0.10	[SEQ_0497, SEQ_1299, SEQ_1615, SEQ_1129, SEQ_0435]
GO:0022804	active transmembrane transporter activity	0.09	1.00	0.01	0.10	[SEQ_1277, SEQ_1299, SEQ_1615, SEQ_1129]

Table B-5. Continued.

GOID	GOTerm	Term PValue	Adjusted Term PValue	Group PValue	Adjusted Group PValue	Associated Genes Found
GO:0006810	transport	0.01	0.31	0.02	0.13	[SEQ_0251, SEQ_0497, SEQ_1277, SEQ_1299, SEQ_1439, SEQ_1615, SEQ_1895, SEQ_1918, SEQ_1129, SEQ_0435]
GO:0055085	transmembrane transport	0.00	0.08	0.02	0.13	[SEQ_0251, SEQ_0497, SEQ_1277, SEQ_1299, SEQ_1439, SEQ_1615, SEQ_1895, SEQ_1918, SEQ_1129, SEQ_0435]
GO:0022857	transmembrane transporter activity	0.02	0.63	0.02	0.13	[SEQ_0251, SEQ_0497, SEQ_1277, SEQ_1299, SEQ_1615, SEQ_1895, SEQ_1129]
GO:0005886	plasma membrane	0.01	0.24	0.02	0.13	[SEQ_0251, SEQ_0497, SEQ_1299, SEQ_1318, SEQ_1439, SEQ_1615, SEQ_1918, SEQ_2009, SEQ_1129, SEQ_0435]
GO:0071705	nitrogen compound transport	0.08	1.00	0.02	0.13	[SEQ_0497, SEQ_1129, SEQ_0435]
GO:0071702	organic substance transport	0.07	1.00	0.02	0.13	[SEQ_0497, SEQ_1299, SEQ_1615, SEQ_1129, SEQ_0435]
GO:0022804	active transmembrane transporter activity	0.09	1.00	0.02	0.13	[SEQ_1277, SEQ_1299, SEQ_1615, SEQ_1129]
GO:0006811	ion transport	0.18	1.00	0.02	0.13	[SEQ_1299, SEQ_1895, SEQ_1129]
GO:0034220	ion transmembrane transport	0.07	1.00	0.02	0.13	[SEQ_1299, SEQ_1895, SEQ_1129]
GO:0006812	cation transport	0.03	0.81	0.02	0.13	[SEQ_1299, SEQ_1895, SEQ_1129]
GO:0015075	ion transmembrane transporter activity	0.06	1.00	0.02	0.13	[SEQ_1299, SEQ_1895, SEQ_1129]
GO:0098655	cation transmembrane transport	0.01	0.29	0.02	0.13	[SEQ_1299, SEQ_1895, SEQ_1129]
GO:0008324	cation transmembrane transporter activity	0.01	0.29	0.02	0.13	[SEQ_1299, SEQ_1895, SEQ_1129]
KEGG:03010	Ribosome	0.05	1.00	0.67	0.67	[SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0043228	non-membrane-bounded organelle	0.09	1.00	0.67	0.67	[SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0043229	intracellular organelle	0.09	1.00	0.67	0.67	[SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0043232	intracellular non-membrane-bounded organelle	0.09	1.00	0.67	0.67	[SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0005840	ribosome	0.06	1.00	0.67	0.67	[SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0019843	rRNA binding	0.02	0.51	0.67	0.67	[SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0003723	RNA binding	0.19	0.94	0.67	0.67	[SEQ_0300, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]

Table B-5. Continued.

GOID	GOTerm	Term PValue	Adjusted Term PValue	Group PValue	Adjusted Group PValue	Associated Genes Found
GO:0043604	amide biosynthetic process	0.20	0.59	0.67	0.67	[SEQ_0300, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0043043	peptide biosynthetic process	0.10	1.00	0.67	0.67	[SEQ_0300, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0006412	translation	0.10	1.00	0.67	0.67	[SEQ_0300, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0000049	tRNA binding	0.02	0.49	0.67	0.67	[SEQ_0300, SEQ_0067, SEQ_0340]
GO:0043603	cellular amide metabolic process	0.12	1.00	0.67	0.67	[SEQ_0300, SEQ_0976, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0006518	peptide metabolic process	0.04	0.95	0.67	0.67	[SEQ_0300, SEQ_0976, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0019538	protein metabolic process	0.12	1.00	0.67	0.67	[SEQ_1597, SEQ_1920, SEQ_0300, SEQ_0976, SEQ_0067, SEQ_0070, SEQ_1651, SEQ_0340]
GO:0006508	proteolysis	0.13	0.91	0.67	0.67	[SEQ_1597, SEQ_1920, SEQ_0976]
GO:0008233	peptidase activity	0.12	1.00	0.67	0.67	[SEQ_1597, SEQ_1920, SEQ_0976]
GO:0008238	exopeptidase activity	0.01	0.29	0.67	0.67	[SEQ_1597, SEQ_1920, SEQ_0976]

B-6 Table. Sites of potential methylation present in SEZ, but absent in SEE.

CDS	Protein
SZO_00070	putative transcription-repair coupling factor
SZO_00940	DNA-directed RNA polymerase beta' chain
SZO_01440	leucyl-tRNA synthetase
SZO_01910	GTP pyrophosphokinase
SZO_06920	putative glutamine ABC transporter, glutamine-binding protein/permease protein
SZO_13500	probable potassium transport system protein
SZO_14700	ribonucleoside-diphosphate reductase alpha subunit
SZO_16270	putative membrane protein
SZO_16830	putative glutamine synthetase
SZO_18730	DNA mismatch repair protein MutS

APPENDIX C

LINUX AND R CODE: DIFFERENCES IN THE ACCESSORY GENOMES AND METHYLOMES OS STRAINS OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* AND OF *STREPTOCOCCUS EQUI* SUBSP. *ZOOEPIDEMICUS* OBTAINED FROM THE RESPIRATORY TRACT OF HORSES FROM TEXAS

C-1 Appendix. Linux and R code used for accessory genome, and methylome analysis of SEE and SEZ isolates.

```
### Streptococcus equi de novo genome assembly with CANU (v1.7) in Linux ###
module load Canu/1.7-intel-2017A-Perl-5.24.0
# command to run pipeline with -pacbio-raw option
canu useGrid=false -p SE_14.105 -d SE_14.105_CANU1.7_out genomeSize=2.1m \
-pacbio-raw ./Duke_Strep_PacBio/FastaFiles/SE_14.105.fasta \
corMhapSensitivity=high corMinCoverage=0 corOutCoverage=100

## Genomes assembled from CANU were annotated using RASTtk (https://rast.nmpdr.org/rast.cgi)

### SEE (n = 50) and SEZ (n = 50) - Spine, AGEnt, and ClustAGE in Linux ###
#Annotated genomes were reformated using Genbank Reformat (http://vfsmspineagent.fsm.northwestern.edu/cgi-
bin/gbk_reformat.cgi)

#Defining the core genome - Spine
module load Spine/0.3.2-GCCcore-7.3.0-Perl-5.28.0
spine.pl -f genome_files.txt

## Example of text in genome_files.txt below
./SZ_SE_AccessoryGenome/RastAnnotatedGenomes/SZ_14.102.gbk SZ_14.102     gbk
./SZ_SE_AccessoryGenome/RastAnnotatedGenomes/SZ_14.106.gbk SZ_14.106     gbk
./SZ_SE_AccessoryGenome/RastAnnotatedGenomes/SE_14.105.gbk SE_14.105     gbk
./SZ_SE_AccessoryGenome/RastAnnotatedGenomes/SE_14.112.gbk SE_14.112     gbk
./SZ_SE_AccessoryGenome/RastAnnotatedGenomes/SE_14.125.gbk SE_14.125     gbk

#Defining the accessory genome - AGEnt
module load AGEnt/0.3.1-GCCcore-7.3.0-Perl-5.28.0
AGEnt.pl -r output.backbone.fasta -q ./SZ_SE_AccessoryGenome/RastAnnotatedGenomes/SE_14.105.gbk -o
SE_14.105 ##each isolate is run individually

#Clustering and binning the accessory genome elements - ClustAGE
module load Magic-BLAST/1.3.0-x64-linux
module load ClustAGE/0.8-foss-2018b-Perl-5.28.0

ClustAGE.pl -f age_files.txt --annot annot_files.txt

## Example of text in age_files.txt below
SE_14.105.SE_14.105.accessory.fasta  SE_14.105     1
SE_14.112.SE_14.112.accessory.fasta  SE_14.112     1
SE_14.125.SE_14.125.accessory.fasta  SE_14.125     1
SZ_14.102.SZ_14.102.accessory.fasta  SZ_14.102     2
```

```
SZ_14.106.SZ_14.106.accessory.fasta  SZ_14.106           2
SZ_14.107.SZ_14.107.accessory.fasta  SZ_14.107           2

## Example of text in annot_files.txt below
SE_14.105.SE_14.105.accessory_loci.txt            SE_14.105
SE_14.112.SE_14.112.accessory_loci.txt            SE_14.112
SE_14.125.SE_14.125.accessory_loci.txt            SE_14.125
SZ_14.102.SZ_14.102.accessory_loci.txt            SZ_14.102
SZ_14.106.SZ_14.106.accessory_loci.txt            SZ_14.106
SZ_14.107.SZ_14.107.accessory_loci.txt            SZ_14.107

### R code for Accessory Genome Output - SEE and SEZ isolates ###
##R Version 4.0.3
subelem <- read.csv("./ClustAGEOutput_100Genomes/out_subelements.csv", header = T)
rownames(subelem) <- subelem[,1]
subelem.sums <- subelem[,3:ncol(subelem)]
sums <- colSums(subelem.sums)
#Adding up the bins of accessory genome elements [AGE](1 indicates presence of AGE, 0 indicates absence)

#Splitting isolates that are SEE
SE.subset <- subelem.sums[1:50,] ## numbers for SEE isolates
SE.sums <- colSums(SE.subset) # sums for only SEE isolates
SE.AGE <- SE.sums[SE.sums == 50]  # keeping bins that == 50
foo <- names(SE.AGE); SE.overall.subset <- subelem.sums[foo] #pulling out bins that == 50 from combined data
#Adding up the bins of accessory genome elements [AGE](1 indicates presence of AGE, 0 indicates absence)
SE.overall.sums <- colSums(SE.overall.subset)
SE.overall.sums <- SE.overall.sums[SE.overall.sums == 50] ## from combined data only keeping sites that == 50
names(SE.overall.sums) #Viewing if any sites fit criteria
write.csv(names(SE.overall.sums), "./ClustAGEOutput_100Genomes/50SE.AGE.csv") ##putting the finale output
into a CSV file

#Splitting isolates that are SEZ
SZ.subset <- subelem.sums[51:100,] ## numbers for SEZ isolates
SZ.sums <- colSums(SZ.subset) # sums for only SEE isolates
SZ.AGE <- SZ.sums[SZ.sums == 50] # keeping bins that == 50
foo <- names(SZ.AGE); SZ.overall.subset <- subelem.sums[foo]  #pulling out bins that == 50 from combined data
#Adding up the bins of accessory genome elements [AGE](1 indicates presence of AGE, 0 indicates absence)
SZ.overall.sums <- colSums(SZ.overall.subset)
SZ.overall.sums <- SZ.overall.sums[SZ.overall.sums == 50] ## from combined data only keeping sites that == 50
names(SZ.overall.sums) #Viewing if any sites fit criteria
write.csv(names(SZ.overall.sums), "./ClustAGEOutput_100Genomes/50SZ.AGE.csv") ##putting the finale output
into a CSV file

## Keeping the AGE that only have 95% of the protein identified in the AGE analysis
## Prior to being brought back into the R the CSV files were modified to put each individual protein into its own row
using Notepad++
SEE.prot <- read.csv("./ClustAGEOutput_100Genomes/AGEs_Proteins_SEE.csv", header = T)
SEZ.prot <- read.csv("./ClustAGEOutput_100Genomes/AGEs_Proteins_SEZ.csv", header = T)

library(dplyr); packageVersion("dplyr") ##1.0.2
#Filtering the protein list
SEE.prot.95 <- SEE.prot %>% filter(Percent >= 95.00)
SEZ.prot.95 <- SEZ.prot %>% filter(Percent >= 95.00)

#Writing the outputs to a comma separated file
write.csv(SEE.prot.95, "./ClustAGEOutput_100Genomes/AGEs_Proteins_SEE_95.csv", quote = F)
write.csv(SEZ.prot.95, "./ClustAGEOutput_100Genomes/AGEs_Proteins_SEZ_95.csv", quote = F)
```

```
#####################################################################################
### BaseMod Methylation pipeline for SEE (n = 24) & SEZ (n = 24) isolates using the SMRT-Link 8 command line
tools - Linux ###
## Example of pipeline for individual isolate

module load SMRT-Link/8.0.0.80529-cli-tools-only

#Aligning the raw BAM reads to the reference
pbmm2 align SE_4047.fasta SE_14.149.bam SE_14.149.aligned.bam
## pbmm2 align SZ_H70.fasta SZ_14.151.bam SZ_14.151.aligned.bam ## alignment for SEZ isolates
#Creating an index for the reference and the Streptococcus equi isolates
samtools faidx SE_4047.fasta ### Indexing the SEE 4047 reference gennome
#samtools faidx SZ_H70.fasta ### Indexing the SEZ H70 reference genome
pbindex SE_14.149.aligned.bam
#Analyzing the aligned sequences for base modifcations
ipdSummary SE_14.149.aligned.bam --reference SE_4047.fasta --gff SE_14.149.basemods.gff --csv
SE_14.149.basemods.csv --pvalue 0.001 --numWorkers 16 --identify m4C,m6A
#Identifying any consensus motifs
motifMaker find -f SE_4047.fasta -g SE_14.149.basemods.gff -o SE_14.149.motifs.csv ### requires more
computational sources than the ipdSummary command
#Creating a GFF file with all of the modification that are part of the motifs
motifMaker reprocess -f SE_4047.fasta -g SE_14.149.basemods.gff -m SE_14.149.motifs.csv -o SE_14.149.motifs.gff

### R code for to filter BaseMod GFF files prior to whole genome comparison set with BEDTools ###
##R Version 4.0.3

library(ape); packageVersion("ape") ## ape: 5.4.1
SE.14.127 <- read.gff("./Strep_equi/Motif_Gff/SE_14.127.motifs.gff", GFF3 = TRUE)
##Example of code for a single isolate

library(dplyr); packageVersion("dplyr") ##1.0.3
library(tidyr); packageVersion("tidyr") ##1.1.2

##### S. equi 14-127 #####
SE.14.127_filtered <- filter(SE.14.127, !grepl('modified_base', type)) #removing instances of modified base
SE.14.127_filt.motif <- filter(SE.14.127_filtered, grepl('motif', attributes)) #pulling out modification with motifs
out <- strsplit(as.character(SE.14.127_filt.motif$attributes), ";");
SE.14.127_filt.motif_attributes <- data.frame(t(sapply(out, '[')));
colnames(SE.14.127_filt.motif_attributes) <- c("context", "motif", "coverage", "IPDRatio", "id", "identifcationQv")
#splitting the attributes column into new columns by semi-colon
SE.14.127_filt.motif <- cbind(SE.14.127_filt.motif, SE.14.127_filt.motif_attributes)

SE.14.127_filt.nomotif <- filter(SE.14.127_filtered, !grepl('motif', attributes)) # pulling out modifications with out
motifs
out <- strsplit(as.character(SE.14.127_filt.nomotif$attributes), ";"); SE.14.127_filt.nomotif_attributes <-
data.frame(t(sapply(out, '['))); colnames(SE.14.127_filt.nomotif_attributes) <- c("coverage", "context", "IPDRatio",
"identifcationQv")
SE.14.127_filt.nomotif <- cbind(SE.14.127_filt.nomotif, SE.14.127_filt.nomotif_attributes) #splitting the attributes
column into new columns by semi-colon
na <- rep(NA, nrow(SE.14.127_filt.nomotif)); SE.14.127_filt.nomotif$motif <- na ; SE.14.127_filt.nomotif$id <- na
##creating columns of NAs to match columns seen in data with motifs

#Combining the data with and without motifs
SE.14.127_filtered <- rbind(SE.14.127_filt.motif, SE.14.127_filt.nomotif)
out <- strsplit(as.character(SE.14.127_filtered$identifcationQv), "="); SE.14.127_Qv <- data.frame(t(sapply(out, '[')));
colnames(SE.14.127_Qv) <- c("Qv", "QvScore"); SE.14.127_Qv$QvScore <- as.numeric(SE.14.127_Qv$QvScore)

SE.14.127_filtered <- cbind(SE.14.127_filtered, SE.14.127_Qv) #pulling out the QV score values
SE.14.127_QvScore30 <- filter(SE.14.127_filtered, QvScore >= 30) #Keeping only methylation with a QV score >=
30

#outputting the filtered data in text and gff file formats
write.table(SE.14.127_QvScore30, "./Strep_equi/SE.14.127_filtered.txt", sep = "\t", quote = F)
library(rtracklayer); packageVersion("rtracklayer") ##1.48.0
export(SE.14.127_QvScore30, "./Strep_equi/SE.14.127_filtered.gff", format = "gff3")

###Creating a annotated GFF file with the methylation events across all SEE & SEZ isolates by reference genome in
Linux ###
module load BEDTools/2.29.2-GCC-9.3.0

## Code for a SEE isolates
bedtools annotate -i SEE_4047.gff3 -files SE.14.105_filtered.gff SE.17.003_filtered.gff SE.19.025_filtered.gff
SE.14.127_filtered.gff \
SE.17.009_filtered.gff SE.18.008_filtered.gff SE.18.061_filtered.gff SE.18.074_filtered.gff SE.18.087_filtered.gff
SE.19.039_filtered.gff \
SE.19.040_filtered.gff SE.19.061_filtered.gff SE.19.065_filtered.gff SE_14.149_filtered.gff SE_18.009_filtered.gff
SE_18.065_filtered.gff \
SE_18.069_filtered.gff SE_18.070_filtered.gff SE_19.004_filtered.gff SE_19.011_filtered.gff SE_19.028_filtered.gff
SE_19.064_filtered.gff \
SE_19.069_filtered.gff SE_19.030_filtered.gff > All_SEE_Methylation_24.gff
## Code for SEZ isolates
bedtools annotate -i SZ_H70.gff3 -files SZ.17.006_filtered.gff SZ.19.005_filtered.gff SZ.14.151_filtered.gff
SZ.18.055_filtered.gff \
SZ.18.058_filtered.gff SZ.18.059_filtered.gff SZ.18.066_filtered.gff SZ.19.045_filtered.gff SZ.19.058_filtered.gff
SZ.19.038_filtered.gff \
SZ.19.043_filtered.gff SZ.19.052_filtered.gff SZ.19.036_filtered.gff SZ.14.106_filtered.gff SZ.14.107_filtered.gff
SZ.18.056_filtered.gff \
SZ.19.041_filtered.gff SZ.19.044_filtered.gff SZ.19.047_filtered.gff SZ.19.048_filtered.gff SZ.19.050_filtered.gff
SZ.19.051_filtered.gff \
SZ.19.053_filtered.gff SZ.19.056_filtered.gff > All_SEZ_Methylation_24.gff

### R code for identify site of methylation in SEE & SEZ isolates on homologous proteins (separately) ###
AllSEE_methy_anno <- read.delim("./All_SEE_Methylation_24_edited.txt", header=FALSE)
AllSEZ_methy_anno <- read.delim("./All_SEZ_Methylation_24_edited.txt", header=FALSE)

methy.localSEE <- AllSEE_methy_anno[,7:ncol(AllSEE_methy_anno)]
se <- c("SE.14.105", "SE.17.003", "SE.19.025", "SE.14.127", "SE.17.009", "SE.18.008", "SE.18.061", "SE.18.074",
       "SE.18.087", "SE.19.039", "SE.19.040", "SE.19.061", "SE.19.065", "SE_14.149", "SE_18.009", "SE_18.065",
"SE_18.069",
       "SE_18.070", "SE_19.004", "SE_19.011", "SE_19.028", "SE_19.064", "SE_19.069", "SE_19.030")
colnames(methy.localSEE) <- se

methy.localSEZ <- AllSEZ_methy_anno[,7:ncol(AllSEZ_methy_anno)]
sez <- c("SZ.17.006", "SZ.19.005", "SZ.14.151", "SZ.18.055", "SZ.18.058", "SZ.18.059", "SZ.18.066", "SZ.19.045",
"SZ.19.058",
        "SZ.19.038", "SZ.19.043", "SZ.19.052", "SZ.19.036", "SZ.14.106", "SZ.14.107", "SZ.18.056", "SZ.19.041",
"SZ.19.044",
        "SZ.19.047", "SZ.19.048", "SZ.19.050", "SZ.19.051", "SZ.19.053", "SZ.19.056")
colnames(methy.localSEZ) <- sez

### Keeping rows with only zeros in SEE isolates
NO.methy.localSEE <- methy.localSEE[apply(methy.localSEE[,-1], 1, function(x) all(x==0)),]
NO.methy.localSEE <- NO.methy.localSEE[rowSums(NO.methy.localSEE) == 0,]
B <- row.names(NO.methy.localSEE)
SEE_No_MethyAnnotate_Subset <- AllSEE_methy_anno[B, ]
dim(SEE_No_MethyAnnotate_Subset)
```

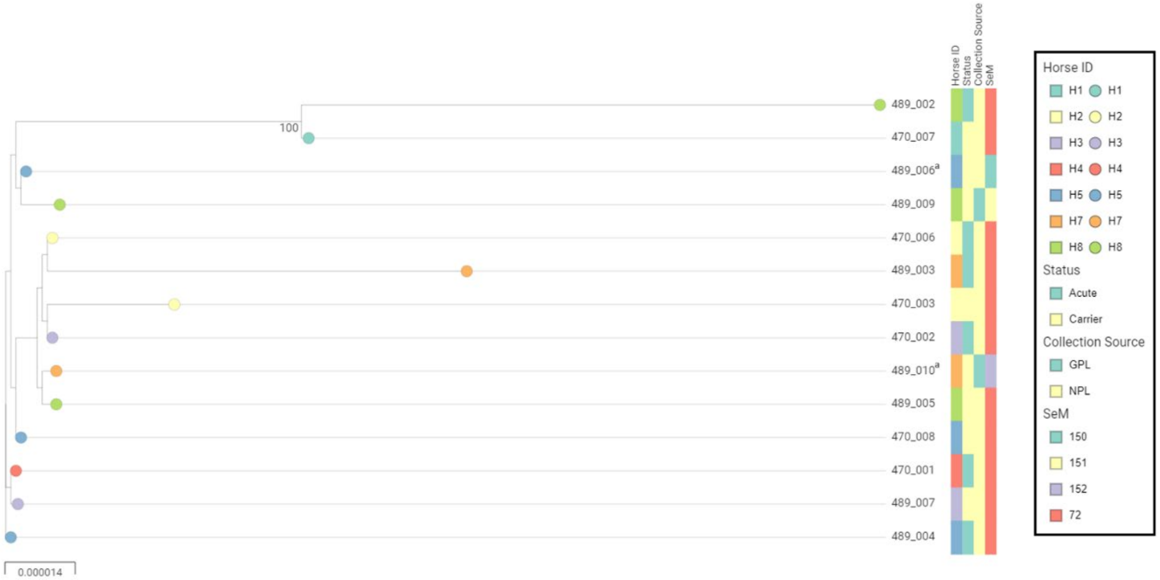
```
### Keeping rows with only zeros in SEZ isolates
NO.methy.localSEZ <- methy.localSEZ[apply(methy.localSEZ[,-1], 1, function(x) all(x==0)),] ### Keeping rows
with only zeros
NO.methy.localSEZ <- NO.methy.localSEZ[rowSums(NO.methy.localSEZ) == 0,]
B <- row.names(NO.methy.localSEZ)
SEZ_No_MethyAnnotate_Subset <- AllSEZ_methy_anno[B, ]
dim(SEZ_No_MethyAnnotate_Subset)

## Removing ANY rows that contain a zero value in SEE isolates
ALL.methy.localSEE <- methy.localSEE[apply(methy.localSEE, 1,function(x) !any(x==0)),]
B <- row.names(ALL.methy.localSEE)
SEE_ALL_MethyAnnotate_Subset <- AllSEE_methy_anno[B, ]
dim(SEE_ALL_MethyAnnotate_Subset)

## Removing ANY rows that contain a zero value in SEZ isolates
ALL.methy.localSEZ <- methy.localSEZ[apply(methy.localSEZ, 1,function(x) !any(x==0)),]
B <- row.names(ALL.methy.localSEZ)
SEZ_ALL_MethyAnnotate_Subset <- AllSEZ_methy_anno[B, ]
dim(SEZ_ALL_MethyAnnotate_Subset)

prot.list <- read.delim("SEEvsSEZ.txt", header = T)
names(prot.list)

library(dplyr)
############# Using PATRIC output with H70 as ref compared to 4047 & ATCC 39506 ##############
SEZ.prot.list <- read.delim("./SEEvsSEZ_ProteinList/genome_comparison_H70_Ref_edited.txt", header = T)
names(SEZ.prot.list)

SEZ.prot.list <- filter(SEZ.prot.list, SEE_4047_percent_identity >= 0.99 & SEE_4047_seq_coverage >= 0.99)
H70.filt_prot.list <- filter(SEZ.prot.list, SEE_ATCC39506_percent_identity >= 0.99 &
SEE_ATCC39506_seq_coverage >= 0.99)
nrow(H70.filt_prot.list)
## 623

# Merging the filtered genes that are >= 99% to entire genome comparison list from patric
SZH70.entirelist <- read.delim("./SEEvsSEZ_ProteinList/genome_comparison_H70_Ref.txt", header = T)
merged.H70 <- SZH70.entirelist[SZH70.entirelist$ref_SEZH70_genome_gene %in%
H70.filt_prot.list$H70_ref_genome_gene,]

locus.tag_4047 <- merged.H70$SEE_4047_locus_tag
locus.tag_H70 <- merged.H70$ref_SEZH70_genome_locus_tag
locus.tag <- data.frame(cbind(locus.tag_H70, locus.tag_4047))
write.table(locus.tag, "./SEEvsSEZ_ProteinList/Combined_LocusTags.txt", sep = "\t", quote = F)

## Reading into R the annotated presence and absence methylation data from SEE & SEZ
SEE_NoMeth_LocusTag <- read.delim("./SecondSet/SEE_No_MethyAnnotate_Subset.txt", header = T)
SEZ_NoMethy_LocusTag <- read.delim("./SecondSet/SEZ_No_MethyAnnotate_Subset.txt", header = T)
SEE_ALLMethy_LocusTag <- read.delim("./SecondSet/SEE_ALL_MethyAnnotate_Subset.txt", header =T)
SEZ_ALLMethy_LocusTag <- read.delim("./SecondSet/SEZ_ALL_MethyAnnotate_Subset.txt", header = T)

## Adding the protein IDs to the methylation presence and absence data
SEE_LocusTargets_NoMethy <- SEE_NoMeth_LocusTag[SEE_NoMeth_LocusTag$V32 %in%
locus.tag$locus.tag_4047,]
SEZ_LocusTargets_NoMethy <- SEZ_NoMethy_LocusTag[SEZ_NoMethy_LocusTag$V32 %in%
locus.tag$locus.tag_H70,]
SEE_LocusTargets_ALLMethy <- SEE_ALLMethy_LocusTag[SEE_ALLMethy_LocusTag$V32 %in%
locus.tag$locus.tag_4047,]
SEZ_LocusTargets_ALLMethy <- SEZ_ALLMethy_LocusTag[SEZ_ALLMethy_LocusTag$V32 %in%
locus.tag$locus.tag_H70,]
```

```
### Adding a column name to the last column of each dataframe so they can be merged
colnames(SEE_LocusTargets_NoMethy)[ncol(SEE_LocusTargets_NoMethy)] <- "locus.tag_4047"
colnames(SEZ_LocusTargets_NoMethy)[ncol(SEZ_LocusTargets_NoMethy)] <- "locus.tag_H70"
colnames(SEE_LocusTargets_ALLMethy)[ncol(SEE_LocusTargets_ALLMethy)] <- "locus.tag_4047"
colnames(SEZ_LocusTargets_ALLMethy)[ncol(SEZ_LocusTargets_ALLMethy)] <- "locus.tag_H70"

### Adding the homologous SEE/SEZ protien to the methylation profile.
library(dplyr); packageVersion("dplyr") ##1.0.3
SEE_LocusTargets_NoMethy <- full_join(SEE_LocusTargets_NoMethy, locus.tag, by = "locus.tag_4047")
SEE_LocusTargets_NoMethy <- na.omit(SEE_LocusTargets_NoMethy)
dim(SEE_LocusTargets_NoMethy)
# [1] 1376   34

SEZ_LocusTargets_NoMethy <- full_join(SEZ_LocusTargets_NoMethy, locus.tag, by = "locus.tag_H70")
SEZ_LocusTargets_NoMethy <- na.omit(SEZ_LocusTargets_NoMethy)
dim(SEZ_LocusTargets_NoMethy)
#[1] 484   34

SEE_LocusTargets_ALLMethy <- full_join(SEE_LocusTargets_ALLMethy, locus.tag, by = "locus.tag_4047")
SEE_LocusTargets_ALLMethy <-na.omit(SEE_LocusTargets_ALLMethy)
dim(SEE_LocusTargets_ALLMethy)
#[1] 251  34

SEZ_LocusTargets_ALLMethy <- full_join(SEZ_LocusTargets_ALLMethy, locus.tag, by = "locus.tag_H70")
SEZ_LocusTargets_ALLMethy <-na.omit(SEZ_LocusTargets_ALLMethy)
dim(SEZ_LocusTargets_ALLMethy)
#[1] 28  34

library(plyr); packageVersion("plyr") ###'1.8.6'
### Combining the absence of SEE methylation locations with the SEZ presence data at homologous proteins
SEE.No_vs_SEZ.All <- SEE_LocusTargets_NoMethy[SEE_LocusTargets_NoMethy$locus.tag_H70 %in%
SEZ_LocusTargets_ALLMethy$locus.tag_H70,]
dim(SEE.No_vs_SEZ.All)
## [1] 34 34
count(SEE.No_vs_SEZ.All$locus.tag_4047)
write.table(SEE.No_vs_SEZ.All, 'SEE.No_vs_SEZ.All_04Jan21_48isolates.txt', sep = "\t", quote = F)

### Combining the absence of SEZ methylation locations with the SEE presence data at homologous proteins
SEE.All_vs_SEZ.No <- SEE_LocusTargets_ALLMethy[SEE_LocusTargets_ALLMethy$locus.tag_H70 %in%
SEZ_LocusTargets_NoMethy$locus.tag_H70,]
dim(SEE.All_vs_SEZ.No)
## [1] 117  34
count(SEE.All_vs_SEZ.No$locus.tag_4047)
write.table(SEE.All_vs_SEZ.No, 'SEE.All_vs_SEZ.No_04Jan21_48isolates.txt', sep = "\t", quote = F)


#### Checking to be sure sites at which methylation occurred in all SEE isolates is homogenous in methylation type
and location.
library(dplyr); packageVersion("dplyr") ##1.0.2
SE.list <- list(SE.14.105,SE.14.127,SE.14.149,SE.17.003,SE.17.009,SE.18.008,SE.18.009,
            SE.18.061,SE.18.065,SE.18.069,SE.18.070,SE.18.074,SE.18.087,SE.19.004,
            SE.19.011,SE.19.025,SE.19.028,SE.19.030,SE.19.039,SE.19.040,SE.19.061,
            SE.19.064,SE.19.065,SE.19.069)
##Example of for a single homologous protein
SE.SEQ_0045 <- lapply(SE.list, function(x) subset(x, x$start >= 56643 & x$start <= 57695));
SE.SEQ_0045 <- bind_rows(SE.SEQ_0045) #selecting methylation that occurred on SEQ_0045

library(plyr); packageVersion("plyr") ##1.8.6
```

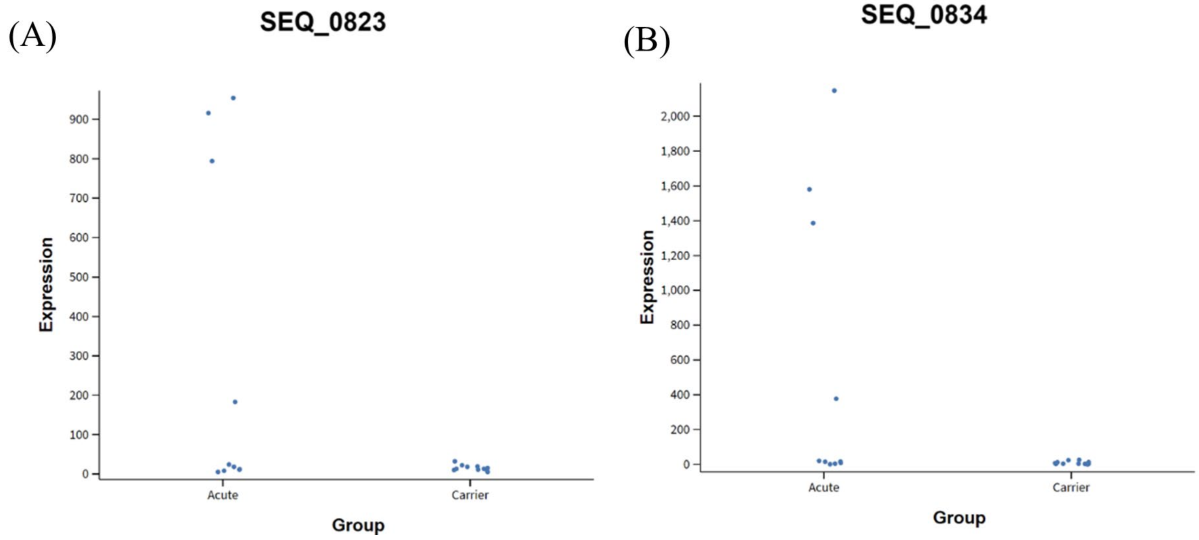
```
count(SE.SEQ_0045$start)

library(dplyr); packageVersion("dplyr") ##1.0.2
#Subsetting the dataframe by sites were all 24 SEE genomes have methylation present
All.SE.SEQ_0045 <- subset(SE.SEQ_0045, SE.SEQ_0045$start == 56855)

## Checking the time of methylation that occurs at those locations
count(All.SE.SEQ_0045$type)
```

APPENDIX D

SUPPLEMENTARY FIGURES AND TABLES: DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS



D-1 Fig. Phylogenetic tree of 14 SEE isolates from Sweden by horse. SEE isolates from the outbreak did not cluster by the individual horse from which the isolate was collected, but results demonstrate variation of isolates recovered from the same individual over time. [a]Denotes truncation in the SeM protein; GPL, Guttural pouch lavage; NL, Nasopharyngeal lavage; SeM, M-like protein.



D-2 Fig. RNA-Seq expression values for 2 SEE genes by disease status group. (A) Expression level (y-axis) of SEQ_0823 by disease presentation (x-axis). Only 3/10 of the acute SEE isolates had elevated expression levels. (B) Expression level (y-axis) of SEQ_0834 by disease presentation (x-axis). Only 3/10 of the acute SEE isolates had higher expression levels.

D-1 Table. Genome accession numbers for SEE isolates from Sweden and Pennsylvania**.**

Genome ID	Location	Status	BioProject	Genome Accession	BioSample Accession	GEO Accession
470_001	Sweden	Acute	PRJNA704656	CP071148	SAMN18051970	NA
470_002	Sweden	Acute	PRJNA704656	JAFKDV000000000	SAMN18051971	NA
470_003	Sweden	Carrier	PRJNA704656	CP071147	SAMN18051972	NA
470_006	Sweden	Acute	PRJNA704656	CP071146	SAMN18051973	NA
470_007	Sweden	Carrier	PRJNA704656	CP071145	SAMN18051974	NA
470_008	Sweden	Carrier	PRJNA704656	CP071144	SAMN18051975	NA
489_002	Sweden	Acute	PRJNA704656	CP071143	SAMN18051976	NA
489_003	Sweden	Acute	PRJNA704656	JAFJVL000000000	SAMN18051977	NA
489_004	Sweden	Acute	PRJNA704656	CP071142	SAMN18051978	NA
489_005	Sweden	Carrier	PRJNA704656	CP071141	SAMN18051979	NA
489_006	Sweden	Carrier	PRJNA704656	JAFKDW000000000	SAMN18051980	NA
489_007	Sweden	Carrier	PRJNA704656	JAFJVK000000000	SAMN18051981	NA
489_009	Sweden	Carrier	PRJNA704656	JAFKDX000000000	SAMN18051982	NA
489_010	Sweden	Carrier	PRJNA704656	CP071140	SAMN18051983	NA
20-080	Pennsylvania	Carrier	PRJNA704656	JAFJVJ000000000	SAMN18051984	GSM5114072
20-081	Pennsylvania	Carrier	PRJNA704656	JAFJVI000000000	SAMN18051985	GSM5114073
20-082	Pennsylvania	Carrier	PRJNA704656	JAFJVH000000000	SAMN18051986	GSM5114074
20-083	Pennsylvania	Carrier	PRJNA704656	JAFJVG000000000	SAMN18051987	GSM5114075
20-084	Pennsylvania	Carrier	PRJNA704656	JAFJVF000000000	SAMN18051988	GSM5114076
20-085	Pennsylvania	Carrier	PRJNA704656	JAFKDY000000000	SAMN18051989	GSM5114077
20-086	Pennsylvania	Carrier	PRJNA704656	JAFJVE000000000	SAMN18051990	GSM5114078
20-087	Pennsylvania	Carrier	PRJNA704656	JAFJVD000000000	SAMN18051991	GSM5114079
20-088	Pennsylvania	Carrier	PRJNA704656	JAFJVC000000000	SAMN18051992	GSM5114080
20-089	Pennsylvania	Carrier	PRJNA704656	JAFJVB000000000	SAMN18051993	GSM5114081
20-090	Pennsylvania	Carrier	PRJNA704656	JAFJVA000000000	SAMN18051994	GSM5114082

Table D-1. Continued.

Genome ID	Location	Status	BioProject	Genome Accession	BioSample Accession	GEO Accession
20-091	Pennsylvania	Acute	PRJNA704656	JAFKDZ000000000	SAMN18051995	GSM5114083
20-092	Pennsylvania	Acute	PRJNA704656	JAFJUZ000000000	SAMN18051996	GSM5114084
20-093	Pennsylvania	Acute	PRJNA704656	JAFJUY000000000	SAMN18051997	GSM5114085
20-094	Pennsylvania	Acute	PRJNA704656	JAFJUX000000000	SAMN18051998	GSM5114086
20-095	Pennsylvania	Acute	PRJNA704656	JAFKEA000000000	SAMN18051999	GSM5114087
20-096	Pennsylvania	Acute	PRJNA704656	JAFJUW000000000	SAMN18052000	GSM5114088
20-097	Pennsylvania	Acute	PRJNA704656	JAFJUV000000000	SAMN18052001	GSM5114089
20-098	Pennsylvania	Acute	PRJNA704656	JAFJUU000000000	SAMN18052002	GSM5114090
20-099	Pennsylvania	Acute	PRJNA704656	JAFJUT000000000	SAMN18052003	GSM5114091
20-100	Pennsylvania	Acute	PRJNA704656	JAFJUS000000000	SAMN18052004	GSM5114092

D-2 Table. Annotation and bin location for the accessory genome elements for the SEE isolates from Sweden.

Bin ID	Genome	Percentage	Annotation
bin1	470_001_01461	100.00%	Oxaloacetate decarboxylase alpha chain (EC 4.1.1.3)
bin1	470_001_01462	100.00%	Citrate lyase holo-[acyl-carrier-protein synthase (EC 2.7.7.61)
bin1	470_001_01463	100.00%	Citrate lyase alpha chain (EC 4.1.3.6)
bin1	470_001_01464	100.00%	Citrate lyase beta chain (EC 4.1.3.6)
bin1	470_001_01465	100.00%	Citrate lyase gamma chain, acyl carrier protein
bin1	470_001_01466	100.00%	FIG01114213: hypothetical protein
bin1	470_001_01467	100.00%	Oxaloacetate decarboxylase beta chain (EC 4.1.1.3)
bin1	470_001_01468	100.00%	Biotin carboxyl carrier protein of oxaloacetate decarboxylase; Biotin carboxyl carrier protein
bin1	470_001_01469	100.00%	FIG01114846: hypothetical protein
bin1	470_001_01470	100.00%	hypothetical protein

Table D-2. Continued.

Bin ID	Genome	Percentage	Annotation
bin1	470_001_01471	100.00%	Citrate/H+ symporter of CitMHS family
bin1	470_001_01472	100.00%	Transcriptional regulator, GntR family
bin1	470_001_01473	100.00%	Triphosphoribosyl-dephospho-CoA synthase (EC 2.4.2.52)
bin1	470_001_01474	100.00%	Putative membrane-spanning protein
bin1	470_001_01475	100.00%	Putative membrane-spanning protein
bin1	470_001_01476	100.00%	[Citrate [pro-3S-lyase ligase (EC 6.2.1.22)
bin2	489_007_00001	100.00%	photosystem I subunit II (PsaD)
bin2	489_007_00002	100.00%	hypothetical protein
bin2	489_007_00003	100.00%	hypothetical protein
bin2	489_007_00004	100.00%	hypothetical protein
bin2	489_007_00005	100.00%	hypothetical protein
bin2	489_007_00006	100.00%	hypothetical protein
bin2	489_007_00007	100.00%	hypothetical protein
bin2	489_007_00008	100.00%	hypothetical protein
bin2	489_007_00009	100.00%	hypothetical protein
bin2	489_007_00010	100.00%	hypothetical protein
bin2	489_007_00011	100.00%	hypothetical protein
bin2	489_007_00012	100.00%	hypothetical protein
bin2	489_007_00013	100.00%	hypothetical protein
bin2	489_007_00014	100.00%	hypothetical protein
bin2	489_007_00015	100.00%	hypothetical protein
bin2	489_007_00016	100.00%	hypothetical protein
bin2	489_007_00017	100.00%	hypothetical protein
bin2	489_007_00018	100.00%	hypothetical protein
bin2	489_007_00019	100.00%	hypothetical protein

Table D-2. Continued.

Bin ID	Genome	Percentage	Annotation
bin2	489_007_00020	100.00%	hypothetical protein
bin2	489_007_00021	100.00%	hypothetical protein
bin2	489_007_00022	100.00%	hypothetical protein
bin2	489_007_00023	100.00%	hypothetical protein
bin2	489_007_00024	100.00%	photosystem I subunit II (PsaD)
bin2	489_007_00025	100.00%	hypothetical protein
bin2	489_007_00026	100.00%	hypothetical protein
bin2	489_007_00027	100.00%	hypothetical protein
bin2	489_007_00028	100.00%	hypothetical protein
bin2	489_007_00029	100.00%	hypothetical protein
bin2	489_007_00030	100.00%	hypothetical protein
bin2	489_007_00031	100.00%	hypothetical protein
bin2	489_007_00032	100.00%	hypothetical protein
bin3	489_007_00056	100.00%	hypothetical protein
bin3	489_007_00057	100.00%	hypothetical protein
bin3	489_007_00058	100.00%	hypothetical protein
bin3	489_007_00059	100.00%	hypothetical protein
bin3	489_007_00060	100.00%	hypothetical protein
bin3	489_007_00061	100.00%	hypothetical protein
bin3	489_007_00062	100.00%	hypothetical protein

D-3 Table. Annotation and bin location for the accessory genome elements for the SEE isolates from Pennsylvania.

Bin ID	Genome	Percent	Annotation
bin1_se00001	SEE_20-080_01432	100.00	Iron-sulfur cluster assembly protein SufB
bin1_se00001	SEE_20-080_01433	100.00	Putative iron-sulfur cluster assembly scaffold protein for SUF system, SufE2
bin1_se00001	SEE_20-080_01434	100.00	Cysteine desulfurase (EC 2.8.1.7) => SufS
bin1_se00001	SEE_20-080_01435	100.00	Iron-sulfur cluster assembly protein SufD
bin1_se00001	SEE_20-080_01436	100.00	Iron-sulfur cluster assembly ATPase protein SufC
bin1_se00001	SEE_20-080_01437	100.00	Undecaprenyl-phosphate alpha-N-acetylglucosaminyl 1-phosphate transferase (EC 2.7.8.33)
bin1_se00001	SEE_20-080_01438	100.00	ClpCP protease substrate adapter protein MecA
bin1_se00001	SEE_20-080_01439	100.00	Undecaprenyl-diphosphatase (EC 3.6.1.27)
bin1_se00001	SEE_20-080_01440	100.00	ABC transporter, substrate-binding protein (cluster 3, basic aa/glutamine/opines) / ABC transporter, permease protein (cluster 3, basic aa/glutamine/opines)
bin1_se00001	SEE_20-080_01441	100.00	ABC transporter, ATP-binding protein (cluster 3, basic aa/glutamine/opines)
bin1_se00001	SEE_20-080_01442	100.00	Uncharacterized protein EF_3205
bin1_se00001	SEE_20-080_01443	100.00	Protein QmcA (possibly involved in integral membrane quality control)
bin1_se00001	SEE_20-080_01444	100.00	Dihydroxyacetone kinase-like protein, phosphatase domain / Dihydroxyacetone kinase-like protein, kinase domain
bin1_se00001	SEE_20-080_01445	100.00	FIG001802: Putative alkaline-shock protein
bin1_se00001	SEE_20-080_01446	100.00	LSU ribosomal protein L28p @ LSU ribosomal protein L28p, zinc-independent
bin1_se00001	SEE_20-080_01447	100.00	hypothetical protein
bin1_se00001	SEE_20-080_01448	100.00	FIG01119612: hypothetical protein
bin1_se00001	SEE_20-080_01449	100.00	hypothetical protein
bin1_se00001	SEE_20-080_01450	100.00	Fructose-bisphosphate aldolase class II (EC 4.1.2.13)
bin1_se00001	SEE_20-080_01452	100.00	Hydrolase, alpha/beta fold family
bin1_se00001	SEE_20-080_01453	100.00	hypothetical protein
bin1_se00001	SEE_20-080_01454	100.00	hypothetical protein
bin1_se00001	SEE_20-080_01455	100.00	CTP synthase (EC 6.3.4.2)
bin1_se00001	SEE_20-080_01456	100.00	DNA-directed RNA polymerase delta subunit (EC 2.7.7.6)
bin1_se00001	SEE_20-080_01457	100.00	Cell division trigger factor (EC 5.2.1.8)

D-3 Table. Continued.

Bin ID	Genome	Percent	Annotation
bin1_se00001	SEE_20-080_01458	100.00	Alpha-amylase (EC 3.2.1.1)
bin10_se00001	SEE_20-084_01853	100.00	Phage protein
bin10_se00001	SEE_20-084_01854	100.00	Phage protein
bin10_se00001	SEE_20-084_01855	100.00	hypothetical protein
bin10_se00001	SEE_20-084_01856	100.00	hypothetical phage protein
bin10_se00001	SEE_20-084_01857	100.00	FIG01114465: hypothetical protein
bin10_se00001	SEE_20-084_01858	100.00	FIG01115915: hypothetical protein
bin10_se00001	SEE_20-084_01859	100.00	FIG01117510: hypothetical protein
bin10_se00001	SEE_20-084_01860	100.00	Phage essential recombination function protein, Erf
bin11_se00001	SEE_20-090_00572	100.00	Streptococcal pyrogenic exotoxin C (SpeC); _Toximoron (Superantigen)
bin11_se00001	SEE_20-090_00573	100.00	hypothetical protein
bin11_se00001	SEE_20-090_00574	100.00	hypothetical protein
bin11_se00001	SEE_20-090_00575	100.00	Phage lysin, N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28)
bin12_se00001	SEE_20-095_00376	98.47	Phage protein
bin12_se00002	SEE_20-095_00377	98.66	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46), phage variant
bin12_se00004	SEE_20-095_00379	100.00	Paratox
bin13_se00001	SEE_20-080_02666	100.00	hypothetical protein
bin13_se00001	SEE_20-080_02667	100.00	Efflux ABC transporter, ATP-binding protein
bin13_se00001	SEE_20-080_02668	100.00	hypothetical protein
bin13_se00001	SEE_20-080_02665	96.80	metallo cofactor biosynthesis protein
bin14_se00001	SEE_20-084_01862	100.00	putative replication protein
bin14_se00001	SEE_20-084_01863	100.00	DNA primase, phage associated
bin15_se00001	SEE_20-084_01903	100.00	Phage protein
bin15_se00001	SEE_20-084_01904	100.00	Phage protein
bin15_se00001	SEE_20-084_01905	100.00	Streptococcal phospholipase A2; _Toximoron (Other)

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin15_se00001	SEE_20-084_01906	100.00	Phage protein
bin16_se00001	SEE_20-081_01490	100.00	Transposase
bin17_se00001	SEE_20-084_01612	100.00	hypothetical protein
bin17_se00001	SEE_20-084_01613	96.91	FIG01119143: hypothetical protein
bin18_se00003	SEE_20-080_02328	100.00	Phage protein
bin18_se00003	SEE_20-080_02329	100.00	Phage protein
bin19_se00001	SEE_20-089_02092	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02093	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02094	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02095	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02096	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02097	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02098	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02099	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02100	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02101	100.00	hypothetical protein
bin19_se00001	SEE_20-089_02102	100.00	hypothetical protein
bin2_se00001	SEE_20-096_00758	100.00	Positive transcriptional regulator, MutR family
bin2_se00001	SEE_20-096_00759	99.74	Phage integrase
bin2_se00003	SEE_20-096_00760	100.00	hypothetical protein
bin2_se00003	SEE_20-096_00761	100.00	hypothetical protein
bin2_se00003	SEE_20-096_00762	100.00	Phage transcriptional regulator
bin2_se00003	SEE_20-096_00763	100.00	hypothetical protein
bin2_se00003	SEE_20-096_00764	100.00	hypothetical protein
bin2_se00003	SEE_20-096_00765	100.00	Phage protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin2_se00003	SEE_20-096_00766	100.00	Phage excisionase
bin2_se00004	SEE_20-096_00768	100.00	Phage protein
bin2_se00004	SEE_20-096_00769	100.00	Phage protein
bin2_se00004	SEE_20-096_00770	100.00	Phage protein
bin2_se00004	SEE_20-096_00771	100.00	Phage protein
bin2_se00004	SEE_20-096_00772	100.00	Phage protein
bin2_se00004	SEE_20-096_00773	100.00	Phage protein
bin2_se00004	SEE_20-096_00774	100.00	Phage protein
bin2_se00004	SEE_20-096_00775	100.00	DNA polymerase, phage-associated
bin2_se00004	SEE_20-096_00776	100.00	DNA polymerase, phage-associated
bin2_se00004	SEE_20-096_00777	95.05	DNA primase, phage associated
bin2_se00007	SEE_20-096_00778	100.00	Phage protein
bin2_se00007	SEE_20-096_00779	100.00	DNA helicase (EC 3.6.4.12), phage-associated
bin2_se00009	SEE_20-096_00781	100.00	Phage protein
bin2_se00009	SEE_20-096_00782	100.00	hypothetical protein - phage associated
bin2_se00009	SEE_20-096_00783	100.00	hypothetical protein
bin2_se00009	SEE_20-096_00784	97.91	Phage protein
bin2_se00010	SEE_20-096_00785	100.00	Phage transcriptional activator
bin2_se00010	SEE_20-096_00786	100.00	Phage terminase, small subunit
bin20_se00001	SEE_20-080_00050	100.00	Phosphoglucomutase (EC 5.4.2.2)
bin21_se00002	SEE_20-100_02666	100.00	Phage protein
bin23_se00001	SEE_20-097_01642	100.00	Phage protein
bin23_se00003	SEE_20-097_01644	100.00	hypothetical protein
bin23_se00003	SEE_20-097_01645	100.00	hypothetical phage protein
bin23_se00003	SEE_20-097_01646	99.63	Phage protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin23_se00003	SEE_20-097_01643	96.03	hypothetical protein
bin24_se00001	SEE_20-098_00365	100.00	hypothetical protein
bin24_se00001	SEE_20-098_00366	100.00	Phage protein
bin24_se00001	SEE_20-098_00367	100.00	Helicase loader DnaI
bin25_se00001	SEE_20-080_02325	99.09	Phage protein
bin26_se00001	SEE_20-085_00131	100.00	Phage integrase
bin28_se00001	SEE_20-097_01620	100.00	Phage integrase
bin28_se00005	SEE_20-097_01621	98.04	hypothetical protein
bin29_se00001	SEE_20-080_00611	100.00	hypothetical protein
bin29_se00001	SEE_20-080_00610	96.87	Site-specific recombinase
bin3_se00001	SEE_20-090_00583	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00584	100.00	Phage tail length tape-measure protein T
bin3_se00001	SEE_20-090_00585	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00586	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00587	100.00	major tail protein b
bin3_se00001	SEE_20-090_00588	100.00	FIG00627453: hypothetical protein
bin3_se00001	SEE_20-090_00589	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00590	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00591	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00592	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00593	100.00	Phage major capsid protein
bin3_se00001	SEE_20-090_00594	100.00	Prophage Clp protease-like protein
bin3_se00001	SEE_20-090_00595	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00596	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00597	100.00	hypothetical protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin3_se00001	SEE_20-090_00598	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00599	100.00	Phage-associated homing endonuclease
bin3_se00001	SEE_20-090_00600	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00601	100.00	Phage integrase
bin3_se00001	SEE_20-090_00602	100.00	FIG01117886: hypothetical protein
bin3_se00001	SEE_20-090_00603	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00604	100.00	hypothetical protein
bin3_se00001	SEE_20-090_00605	100.00	Phage repressor
bin3_se00001	SEE_20-090_00606	100.00	hypothetical protein
bin30_se00001	SEE_20-084_01867	100.00	DNA-cytosine methyltransferase (EC 2.1.1.37)
bin30_se00001	SEE_20-084_01868	100.00	Type II, 5-methyl-cytosine DNA methyltransferase
bin33_se00001	SEE_20-080_01204	100.00	hypothetical protein
bin34_se00002	SEE_20-098_00378	97.77	Phage integrase
bin35_se00001	SEE_20-098_00371	100.00	hypothetical protein
bin35_se00001	SEE_20-098_00372	100.00	Phage antirepressor protein
bin35_se00001	SEE_20-098_00373	100.00	hypothetical protein
bin37_se00002	SEE_20-080_01138	100.00	Phage protein
bin38_se00001	SEE_20-084_01870	100.00	hypothetical protein
bin38_se00001	SEE_20-084_01871	100.00	hypothetical protein
bin39_se00001	SEE_20-087_02576	100.00	hypothetical phage protein
bin39_se00001	SEE_20-087_02577	100.00	Phage protein
bin4_se00001	SEE_20-096_00380	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00381	100.00	Phage tail length tape-measure protein T
bin4_se00001	SEE_20-096_00382	100.00	Phage protein (ACLAME 404)
bin4_se00001	SEE_20-096_00383	100.00	hypothetical protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin4_se00001	SEE_20-096_00384	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00385	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00386	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00387	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00388	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00389	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00390	100.00	Lactobacillus delbrueckii phage mv4 main capsid protein Gp34 homolog lin2390
bin4_se00001	SEE_20-096_00391	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00392	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00393	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00394	100.00	Phage protein
bin4_se00001	SEE_20-096_00395	100.00	hypothetical protein
bin4_se00001	SEE_20-096_00396	100.00	Portal protein [Bacteriophage A118]
bin4_se00001	SEE_20-096_00397	100.00	Phage terminase, large subunit
bin4_se00001	SEE_20-096_00398	100.00	Phage protein
bin4_se00004	SEE_20-096_00399	100.00	Phage protein
bin4_se00004	SEE_20-096_00400	100.00	hypothetical protein
bin42_se00001	SEE_20-098_00374	100.00	ORF070
bin43_se00001	SEE_20-080_02344	100.00	hypothetical protein
bin43_se00001	SEE_20-080_02345	100.00	Phage protein
bin43_se00001	SEE_20-080_02346	100.00	hypothetical protein
bin44_se00001	SEE_20-096_00425	100.00	Phage protein (ACLAME 1171)
bin45_se00001	SEE_20-089_02196	100.00	hypothetical protein
bin45_se00001	SEE_20-089_02195	97.62	hypothetical protein
bin47_se00001	SEE_20-087_01541	100.00	hypothetical protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin47_se00001	SEE_20-087_01540	98.91	hypothetical protein
bin5_se00001	SEE_20-081_01746	96.62	hypothetical phage protein
bin5_se00002	SEE_20-081_01747	100.00	hypothetical protein
bin5_se00004	SEE_20-081_01749	100.00	Phage protein
bin5_se00004	SEE_20-081_01750	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01751	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01752	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01753	100.00	Phage recombination protein Bet
bin5_se00005	SEE_20-081_01754	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01755	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01756	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01757	100.00	FIG01114174: hypothetical protein
bin5_se00005	SEE_20-081_01758	100.00	DNA replication protein DnaD
bin5_se00005	SEE_20-081_01759	100.00	Phage replication initiation protein
bin5_se00005	SEE_20-081_01760	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01761	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01762	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01763	100.00	hypothetical protein - phage associated
bin5_se00005	SEE_20-081_01764	100.00	Phage protein
bin5_se00005	SEE_20-081_01765	100.00	Phage antirepressor protein
bin5_se00005	SEE_20-081_01766	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01767	100.00	hypothetical protein
bin5_se00005	SEE_20-081_01768	100.00	hypothetical protein
bin5_se00006	SEE_20-081_01769	100.00	hypothetical protein within a prophage
bin5_se00009	SEE_20-081_01771	100.00	hypothetical protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin5_se00009	SEE_20-081_01772	100.00	mRNA interferase RelE
bin5_se00009	SEE_20-081_01773	100.00	hypothetical protein
bin5_se00009	SEE_20-081_01774	100.00	hypothetical protein
bin5_se00009	SEE_20-081_01775	100.00	hypothetical protein
bin52_se00001	SEE_20-087_02571	100.00	hypothetical protein
bin54_se00001	SEE_20-081_00760	100.00	hypothetical protein
bin55_se00001	SEE_20-096_00403	100.00	hypothetical protein
bin56_se00001	SEE_20-080_01999	100.00	hypothetical protein
bin57_se00001	SEE_20-080_00116	100.00	hypothetical protein
bin59_se00001	SEE_20-080_01171	100.00	hypothetical protein
bin59_se00001	SEE_20-080_01172	100.00	conserved hypothetical protein
bin6_se00001	SEE_20-080_01424	100.00	Oligopeptide ABC transporter, permease protein OppC (TC 3.A.1.5.1)
bin6_se00001	SEE_20-080_01425	100.00	Oligopeptide ABC transporter, permease protein OppB (TC 3.A.1.5.1)
bin6_se00001	SEE_20-080_01426	100.00	Oligopeptide ABC transporter, substrate-binding protein OppA (TC 3.A.1.5.1)
bin6_se00001	SEE_20-080_01427	100.00	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)
bin6_se00001	SEE_20-080_01428	100.00	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)
bin6_se00001	SEE_20-080_01429	100.00	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)
bin7_se00001	SEE_20-090_00610	100.00	Phage protein
bin7_se00001	SEE_20-090_00611	100.00	hypothetical phage protein
bin7_se00001	SEE_20-090_00612	100.00	Phage DNA replication protein O
bin7_se00001	SEE_20-090_00613	100.00	putative protein
bin7_se00001	SEE_20-090_00614	100.00	hypothetical protein
bin7_se00001	SEE_20-090_00615	100.00	hypothetical protein
bin7_se00001	SEE_20-090_00616	100.00	Phage antirepressor protein
bin7_se00001	SEE_20-090_00617	100.00	hypothetical protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin7_se00001	SEE_20-090_00618	100.00	hypothetical protein
bin7_se00001	SEE_20-090_00619	100.00	Phage protein
bin7_se00001	SEE_20-090_00620	100.00	Phage protein
bin7_se00001	SEE_20-090_00621	100.00	Phage protein
bin7_se00001	SEE_20-090_00622	100.00	Phage transcriptional regulator, Cro/CI family
bin7_se00001	SEE_20-090_00623	100.00	Putative cI repressor, metallo-proteinase motif (ACLAME 174)
bin7_se00001	SEE_20-090_00624	99.78	hypothetical protein
bin7_se00004	SEE_20-090_00625	98.02	Integrase
bin8_se00001	SEE_20-080_01181	98.87	Phage protein
bin8_se00003	SEE_20-080_01182	96.02	Phage protein
bin8_se00007	SEE_20-080_01185	100.00	hypothetical protein
bin8_se00007	SEE_20-080_01186	100.00	Phage protein
bin8_se00007	SEE_20-080_01187	100.00	Helicase loader DnaI
bin8_se00008	SEE_20-080_01189	100.00	hypothetical protein
bin8_se00008	SEE_20-080_01190	100.00	hypothetical protein
bin8_se00008	SEE_20-080_01191	100.00	hypothetical protein
bin8_se00008	SEE_20-080_01192	100.00	hypothetical protein
bin9_se00003	SEE_20-097_01623	100.00	hypothetical protein
bin9_se00003	SEE_20-097_01624	100.00	hypothetical protein
bin9_se00003	SEE_20-097_01625	99.82	hypothetical protein
bin9_se00004	SEE_20-097_01626	100.00	hypothetical protein
bin9_se00004	SEE_20-097_01627	100.00	putative cro protein
bin9_se00005	SEE_20-097_01628	97.39	Phage antirepressor protein
bin9_se00006	SEE_20-097_01629	100.00	hypothetical protein
bin9_se00006	SEE_20-097_01630	100.00	hypothetical protein

Table D-3. Continued.

Bin ID	Genome	Percent	Annotation
bin9_se00009	SEE_20-097_01631	96.56	hypothetical phage protein
bin9_se00010	SEE_20-097_01632	100.00	Phage protein

D-4 Table. Sites of methylation found in at least half (n ≥ 4) of either disease state in Swedish SEE isolates.

Isolate	Status	Location	Site	Methylation	Type	Motif
470_003	Carrier	SEQ_0106	NA	N	NA	NA
470_007	Carrier	SEQ_0106	118739	Y	m4C	Unknown
470_008	Carrier	SEQ_0106	118739	Y	m4C	Unknown
489_005	Carrier	SEQ_0106	NA	N	NA	NA
489_006	Carrier	SEQ_0106	NA	N	NA	NA
489_007	Carrier	SEQ_0106	118739	Y	m4C	Unknown
489_009	Carrier	SEQ_0106	118739	Y	m4C	Unknown
489_010	Carrier	SEQ_0106	NA	N	NA	NA
470_001	Acute	SEQ_0106	NA	N	NA	NA
470_002	Acute	SEQ_0106	NA	N	NA	NA
470_006	Acute	SEQ_0106	NA	N	NA	NA
489_002	Acute	SEQ_0106	NA	N	NA	NA
489_003	Acute	SEQ_0106	NA	N	NA	NA
489_004	Acute	SEQ_0106	NA	N	NA	NA
470_003	Carrier	SEQ_0128	NA	N	NA	NA
470_007	Carrier	SEQ_0128	NA	N	NA	NA
470_008	Carrier	SEQ_0128	NA	N	NA	NA
489_005	Carrier	SEQ_0128	NA	N	NA	NA

Table D-4. Continued.

Isolate	Status	Location	Site	Methylation	Type	Motif
489_006	Carrier	SEQ_0128	NA	N	NA	NA
489_007	Carrier	SEQ_0128	NA	N	NA	NA
489_009	Carrier	SEQ_0128	NA	N	NA	NA
489_010	Carrier	SEQ_0128	NA	N	NA	NA
470_001	Acute	SEQ_0128	136825	Y	m4C	Unknown
470_002	Acute	SEQ_0128	136825	Y	m4C	Unknown
470_006	Acute	SEQ_0128	136825	Y	m4C	Unknown
489_002	Acute	SEQ_0128	NA	N	NA	NA
489_003	Acute	SEQ_0128	136825	Y	m4C	Unknown
489_004	Acute	SEQ_0128	NA	N	NA	NA
470_003	Carrier	SEQ_0695	678862	Y	m6A	DNRTGCAGB
470_007	Carrier	SEQ_0695	NA	N	NA	NA
470_008	Carrier	SEQ_0695	NA	N	NA	NA
489_005	Carrier	SEQ_0695	678862	Y	m6A	DNRTGCAGB
489_006	Carrier	SEQ_0695	NA	N	NA	NA
489_007	Carrier	SEQ_0695	NA	N	NA	NA
489_009	Carrier	SEQ_0695	678862	Y	m6A	DNRTGCAGB
489_010	Carrier	SEQ_0695	678862	Y	m6A	DNRTGCAGB
470_001	Acute	SEQ_0695	NA	N	NA	NA
470_002	Acute	SEQ_0695	NA	N	NA	NA
470_006	Acute	SEQ_0695	NA	N	NA	NA
489_002	Acute	SEQ_0695	NA	N	NA	NA
489_003	Acute	SEQ_0695	NA	N	NA	NA
489_004	Acute	SEQ_0695	NA	N	NA	NA
470_003	Carrier	SEQ_0954	943185	Y	m6A	DNRTGCAGB

Table D-4. Continued.

Isolate	Status	Location	Site	Methylation	Type	Motif
470_007	Carrier	SEQ_0954	NA	N	NA	NA
470_008	Carrier	SEQ_0954	NA	N	NA	NA
489_005	Carrier	SEQ_0954	943185	Y	m6A	DNRTGCAGB
489_006	Carrier	SEQ_0954	NA	N	NA	NA
489_007	Carrier	SEQ_0954	NA	N	NA	NA
489_009	Carrier	SEQ_0954	943185	Y	m6A	DNRTGCAGB
489_010	Carrier	SEQ_0954	943185	Y	m6A	DNRTGCAGB
470_001	Acute	SEQ_0954	NA	N	NA	NA
470_002	Acute	SEQ_0954	NA	N	NA	NA
470_006	Acute	SEQ_0954	NA	N	NA	NA
489_002	Acute	SEQ_0954	NA	N	NA	NA
489_003	Acute	SEQ_0954	NA	N	NA	NA
489_004	Acute	SEQ_0954	NA	N	NA	NA
470_003	Carrier	SEQ_1931	1938942	Y	m6A	DNRTGCAGB
470_007	Carrier	SEQ_1931	NA	N	NA	NA
470_008	Carrier	SEQ_1931	1938942	Y	m6A	Unknown
489_005	Carrier	SEQ_1931	1938942	Y	m6A	DNRTGCAGB
489_006	Carrier	SEQ_1931	NA	N	NA	NA
489_007	Carrier	SEQ_1931	NA	N	NA	NA
489_009	Carrier	SEQ_1931	1938942	Y	m6A	DNRTGCAGB
489_010	Carrier	SEQ_1931	1938942	Y	m6A	DNRTGCAGB
470_001	Acute	SEQ_1931	NA	N	NA	NA
470_002	Acute	SEQ_1931	NA	N	NA	NA
470_006	Acute	SEQ_1931	NA	N	NA	NA
489_002	Acute	SEQ_1931	NA	N	NA	NA

Table D-4. Continued.

Isolate	Status	Location	Site	Methylation	Type	Motif
489_003	Acute	SEQ_1931	NA	N	NA	NA
489_004	Acute	SEQ_1931	NA	N	NA	NA
470_003	Carrier	SEQ_2001	NA	N	NA	NA
470_007	Carrier	SEQ_2001	NA	N	NA	NA
470_008	Carrier	SEQ_2001	NA	N	NA	NA
489_005	Carrier	SEQ_2001	NA	N	NA	NA
489_006	Carrier	SEQ_2001	NA	N	NA	NA
489_007	Carrier	SEQ_2001	NA	N	NA	NA
489_009	Carrier	SEQ_2001	NA	N	NA	NA
489_010	Carrier	SEQ_2001	NA	N	NA	NA
470_001	Acute	SEQ_2001	2021994	Y	m4C	Unknown
470_002	Acute	SEQ_2001	2021994	Y	m4C	Unknown
470_006	Acute	SEQ_2001	2021994	Y	m4C	Unknown
489_002	Acute	SEQ_2001	NA	N	NA	NA
489_003	Acute	SEQ_2001	NA	N	NA	NA
489_004	Acute	SEQ_2001	2021994	Y	m4C	Unknown
470_003	Carrier	SEQ_2169	2182631	Y	m6A	DNRTGCAGB
470_007	Carrier	SEQ_2169	NA	N	NA	NA
470_008	Carrier	SEQ_2169	NA	N	NA	NA
489_005	Carrier	SEQ_2169	2182631	Y	m6A	DNRTGCAGB
489_006	Carrier	SEQ_2169	NA	N	NA	NA
489_007	Carrier	SEQ_2169	NA	N	NA	NA
489_009	Carrier	SEQ_2169	2182631	Y	m6A	DNRTGCAGB
489_010	Carrier	SEQ_2169	2182631	Y	m6A	DNRTGCAGB
470_001	Acute	SEQ_2169	NA	N	NA	NA

Table D-4. Continued.

Isolate	Status	Location	Site	Methylation	Type	Motif
470_002	Acute	SEQ_2169	NA	N	NA	NA
470_006	Acute	SEQ_2169	NA	N	NA	NA
489_002	Acute	SEQ_2169	NA	N	NA	NA
489_003	Acute	SEQ_2169	NA	N	NA	NA
489_004	Acute	SEQ_2169	NA	N	NA	NA

D-5 Table. Sites of methylation found in at least half (n ≥ 6) of either disease state in Pennsylvania SEE isolates.

Genome	Status	Location	Site	Methylation	Type	Motif
20.080	Carrier	SEQ_0905	880423	Y	m4C	Unknown
20.081	Carrier	SEQ_0905	880423	Y	m4C	Unknown
20.083	Carrier	SEQ_0905	880423	Y	m4C	Unknown
20.084	Carrier	SEQ_0905	880423	Y	m4C	Unknown
20.085	Carrier	SEQ_0905	880423	Y	m4C	Unknown
20.088	Carrier	SEQ_0905	880423	Y	m4C	Unknown
20.082	Carrier	SEQ_0905	NA	N	NA	NA
20.086	Carrier	SEQ_0905	NA	N	NA	NA
20.087	Carrier	SEQ_0905	NA	N	NA	NA
20.089	Carrier	SEQ_0905	NA	N	NA	NA
20.090	Carrier	SEQ_0905	NA	N	NA	NA
20.091	Acute	SEQ_0905	NA	N	NA	NA
20.092	Acute	SEQ_0905	NA	N	NA	NA
20.093	Acute	SEQ_0905	NA	N	NA	NA
20.094	Acute	SEQ_0905	NA	N	NA	NA

Table D-5. Continued.

Genome	Status	Location	Site	Methylation	Type	Motif
20.095	Acute	SEQ_0905	NA	N	NA	NA
20.096	Acute	SEQ_0905	NA	N	NA	NA
20.097	Acute	SEQ_0905	NA	N	NA	NA
20.098	Acute	SEQ_0905	NA	N	NA	NA
20.099	Acute	SEQ_0905	NA	N	NA	NA
20.100	Acute	SEQ_0905	NA	N	NA	NA
20.080	Carrier	SEQ_1082	1073983	Y	m4C	Unknown
20.081	Carrier	SEQ_1082	1073983	Y	m4C	Unknown
20.083	Carrier	SEQ_1082	1073983	Y	m4C	Unknown
20.084	Carrier	SEQ_1082	1073983	Y	m4C	Unknown
20.085	Carrier	SEQ_1082	1073983	Y	m4C	Unknown
20.087	Carrier	SEQ_1082	1073983	Y	m4C	Unknown
20.082	Carrier	SEQ_1082	NA	N	NA	NA
20.086	Carrier	SEQ_1082	NA	N	NA	NA
20.088	Carrier	SEQ_1082	NA	N	NA	NA
20.089	Carrier	SEQ_1082	NA	N	NA	NA
20.090	Carrier	SEQ_1082	NA	N	NA	NA
20.091	Acute	SEQ_1082	NA	N	NA	NA
20.092	Acute	SEQ_1082	NA	N	NA	NA
20.093	Acute	SEQ_1082	NA	N	NA	NA
20.094	Acute	SEQ_1082	NA	N	NA	NA
20.095	Acute	SEQ_1082	NA	N	NA	NA
20.096	Acute	SEQ_1082	NA	N	NA	NA
20.097	Acute	SEQ_1082	NA	N	NA	NA
20.098	Acute	SEQ_1082	NA	N	NA	NA

Table D-5. Continued.

Genome	Status	Location	Site	Methylation	Type	Motif
20.099	Acute	SEQ_1082	NA	N	NA	NA
20.100	Acute	SEQ_1082	NA	N	NA	NA
20.080	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.081	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.083	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.085	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.086	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.088	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.090	Carrier	SEQ_2023	2052534	Y	m4C	Unknown
20.082	Carrier	SEQ_2023	NA	N	NA	NA
20.084	Carrier	SEQ_2023	NA	N	NA	NA
20.087	Carrier	SEQ_2023	NA	N	NA	NA
20.089	Carrier	SEQ_2023	NA	N	NA	NA
20.091	Acute	SEQ_2023	NA	N	NA	NA
20.092	Acute	SEQ_2023	NA	N	NA	NA
20.093	Acute	SEQ_2023	NA	N	NA	NA
20.094	Acute	SEQ_2023	NA	N	NA	NA
20.095	Acute	SEQ_2023	NA	N	NA	NA
20.096	Acute	SEQ_2023	NA	N	NA	NA
20.097	Acute	SEQ_2023	NA	N	NA	NA
20.098	Acute	SEQ_2023	NA	N	NA	NA
20.099	Acute	SEQ_2023	NA	N	NA	NA
20.100	Acute	SEQ_2023	NA	N	NA	NA

D-6 Table. Gene expression of transcripts with FDR < 0.1 identified by edgeR analysis.

Gene ID	logFC	logCPM	F-Statistic	PValue	FDR
SEQ_1976	-0.42209	9.498149	26.17902597	4.35E-05	0.05542
SEQ_0834	-6.0316	2.860015	23.89267467	7.61E-05	0.05542
SEQ_0823	-4.41396	2.015865	23.29317115	8.81E-05	0.05542
SEQ_0820	-3.72229	1.211648	20.17157841	0.000196	0.070974
SEQ_1667	-0.40178	6.853421	19.29441051	0.000246	0.070974
SEQ_1174	0.564532	6.846277	18.91771008	0.000273	0.070974
SEQ_1175	0.535149	4.716648	17.85523087	0.000368	0.070974
SEQ_0295	-0.29153	6.563539	17.57143263	0.000399	0.070974
SEQ_2060	-2.66168	2.661376	17.20741418	0.000448	0.070974
SEQ_2040	-1.55808	0.851829	17.16885295	0.000449	0.070974
SEQ_2143	0.410614	8.868365	17.0255004	0.000468	0.070974
SEQ_1500	-0.3559	8.385687	16.77730104	0.000503	0.070974
SEQ_1341	0.269749	7.530267	16.57862287	0.000533	0.070974
SEQ_1538	-0.37637	9.71697	16.48310889	0.000549	0.070974
SEQ_1977	-0.36133	8.36992	16.38863338	0.000564	0.070974
SEQ_0617	-0.51946	9.269869	15.71585908	0.00069	0.077508
SEQ_1323	0.388262	10.15832	15.55477294	0.000725	0.077508
SEQ_0400	-0.68981	10.55716	15.39836639	0.000761	0.077508
SEQ_0020	0.306409	8.712484	15.31434347	0.00078	0.077508
SEQ_2048	-2.25417	4.893223	14.98452053	0.000871	0.081197
SEQ_1773	-0.45146	7.552096	14.82844953	0.000907	0.081197
SEQ_0836	-1.04209	2.62963	14.58916005	0.000977	0.081197
SEQ_2046	-2.94757	2.693109	14.57394155	0.00099	0.081197
SEQ_0148	-5.1213	2.258311	14.15256674	0.00113	0.088874
SEQ_1668	-0.30614	7.568264	13.69343989	0.0013	0.093648

Table D-6. Continued.

Gene ID	logFC	logCPM	F-Statistic	PValue	FDR
SEQ_1577	-0.44246	8.540143	13.51460049	0.001377	0.093648
SEQ_1899	0.374589	7.674175	13.50655222	0.001381	0.093648
SEQ_0840	-1.60716	1.998133	13.45909848	0.001405	0.093648
SEQ_2213	0.277614	7.559426	13.37979035	0.001439	0.093648

APPENDIX E

LINUX AND R CODE: DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS

E-1 Appendix. Linux and R code used for accessory genome, methylome, and transcriptome analysis.

```
#E-1 Appendix. Linux and R code used for accessory genome, methylome and transcriptome analysis.
### Streptococcus equi de novo genome assembly with CANU (v1.7) in Linux ###
module load Canu/1.7-intel-2017A-Perl-5.24.0
# command to run pipeline with -pacbio-raw option
canu useGrid=false -p SEE_20-080 -d SEE_20-080_CANU1.7_out genomeSize=2.1m \
-pacbio-raw /scratch/user/ellenruth/Duke_Order6546/SEE_20-080.fasta \
corMhapSensitivity=high corMinCoverage=0 corOutCoverage=100

## Genomes assembled from CANU were annotated using RASTtk (https://rast.nmpdr.org/rast.cgi)

### Sweden and Pennsylvania Streptococcus isolates - Spine, AGEnt, and ClustAGE in Linux ###
#Annotated genomes were reformated using Genbank Reformat (http://vfsmspineagent.fsm.northwestern.edu/cgi-
bin/gbk_reformat.cgi)

#Defining the core genome - Spine
##Sweden and Pennsylvania Streptococcus isolates run separately
module load Spine/0.3.2-GCCcore-7.3.0-Perl-5.28.0
spine.pl -f genome_files.txt

## Example of text in genome_files.txt below
/PennSEE_AccessoryGenome/SEE_20-080.gbk   SEE_20-080        gbk
/PennSEE_AccessoryGenome/SEE_20-081.gbk   SEE_20-081        gbk
/PennSEE_AccessoryGenome/SEE_20-082.gbk   SEE_20-082        gbk
/PennSEE_AccessoryGenome/SEE_20-083.gbk   SEE_20-083        gbk
/PennSEE_AccessoryGenome/SEE_20-084.gbk   SEE_20-084        gbk
/PennSEE_AccessoryGenome/SEE_20-085.gbk   SEE_20-085        gbk

#Defining the accessory genome - AGEnt
##Sweden and Pennsylvania Streptococcus isolates run separately
module load AGEnt/0.3.1-GCCcore-7.3.0-Perl-5.28.0
AGEnt.pl -r output.backbone.fasta -q /PennSEE_AccessoryGenome/SEE_20-080.gbk -o SEE_20-080 ##each isolate
is run individually

#Clustering and binning the accessory genome elements - ClustAGE
##Sweden and Pennsylvania Streptococcus isolates run separately
module load Magic-BLAST/1.3.0-x64-linux
module load ClustAGE/0.8-foss-2018b-Perl-5.28.0

ClustAGE.pl -f age_files.txt --annot annot_files.txt
```

```
## Example of text in age_files.txt below
SEE_20-080.SEE_20-080.accessory.fasta         SEE_20-080        1
SEE_20-081.SEE_20-081.accessory.fasta         SEE_20-081        1
SEE_20-082.SEE_20-082.accessory.fasta         SEE_20-082        1
SEE_20-091.SEE_20-091.accessory.fasta         SEE_20-091        2
SEE_20-092.SEE_20-092.accessory.fasta         SEE_20-092        2
SEE_20-093.SEE_20-093.accessory.fasta         SEE_20-093        2

## Example of text in annot_files.txt below
SEE_20-080.SEE_20-080.accessory_loci.txt      SEE_20-080
SEE_20-081.SEE_20-081.accessory_loci.txt      SEE_20-081
SEE_20-082.SEE_20-082.accessory_loci.txt      SEE_20-082
SEE_20-083.SEE_20-083.accessory_loci.txt      SEE_20-083
SEE_20-084.SEE_20-084.accessory_loci.txt      SEE_20-084
SEE_20-085.SEE_20-085.accessory_loci.txt      SEE_20-085

### R code for Accessory Genome Output - Sweden and Pennsylvania Streptococcus equi isolates ###
##R Version 4.0.3
subelem <- read.csv("./ClustAGEOutput/out_subelements.csv", header = T)
rownames(subelem) <- subelem[,1]
subelem.sums <- subelem[,3:ncol(subelem)]
#Adding up the bins of accessory genome elements [AGE](1 indicates presence of AGE, 0 indicates absence)

#Splitting isolates that are carrier state
SE.carrier.subset <- subelem.sums[1:11,]
#SE.carrier.subset <- subelem.sums[1:8,] ## numbers for Sweden isolates
SE.carrier.sums <- colSums(SE.carrier.subset) # sums for only carrier isolates
SE.carrier.AGE <- SE.carrier.sums[SE.carrier.sums == 11] # keeping bins that == 11
#SE.carrier.AGE <- SE.carrier.sums[SE.carrier.sums == 8] ## numbers for Sweden isolates
foo <- names(SE.carrier.AGE); SEcar.overall.subset <- subelem.sums[foo] #pulling out bins that == 11 from
combined data
#Adding up the bins of accessory genome elements [AGE](1 indicates presence of AGE, 0 indicates absence)
SEcar.overall.sums <- colSums(SEcar.overall.subset)
SEcar.overall.sums <- SEcar.overall.sums[SEcar.overall.sums == 11] ## from combined data only keeping sites that
== 11
#SEcar.overall.sums <- SEcar.overall.sums[SEcar.overall.sums == 8] ## numbers for Sweden isolates
head(SEcar.overall.sums) #Viewing if any sites fit criteria

#Splitting isolates by clinical state
SE.clinical.subset <- subelem.sums[12:21,]
#SE.clinical.subset <- subelem.sums[9:14,] ## numbers for Sweden isolates
SE.clinical.sums <- colSums(SE.clinical.subset)  # sums for only clinical isolates
SE.clinical.AGE <- SE.clinical.sums[SE.clinical.sums == 10] # keeping bins that == 10
#SE.clinical.AGE <- SE.clinical.sums[SE.clinical.sums == 6] ## numbers for Sweden isolates
foo <- names(SE.clinical.AGE); SEclin.overall.subset <- subelem.sums[foo]
#Adding up the bins of accessory genome elements [AGE](1 indicates presence of AGE, 0 indicates absence)
SEclin.overall.sums <- colSums(SEclin.overall.subset)
SEclin.overall.sums <- SEclin.overall.sums[SEclin.overall.sums == 10] ## from combined data only keeping sites that
== 10
#SEclin.overall.sums <- SEclin.overall.sums[SEclin.overall.sums == 6] ## numbers for Sweden isolates
head(SEclin.overall.sums) #Viewing if any sites fit criteria

###########################################################################################
### BaseMod Methylation pipeline for Streptococcus equi isolates using the SMRT-Link 8 command line tools -
Linux ###
## Example of pipeline for individual isolate

module load SMRT-Link/8.0.0.80529-cli-tools-only
```

```
#Aligning the raw BAM reads to the reference
pbmm2 align SE_4047.fasta 470_003.subreads.bam 470_003.subreads.aligned.bam
#Creating an index for the reference and the Streptococcus equi isolates
samtools faidx SE_4047.fasta
pbindex 470_003.subreads.aligned.bam
#Analyzing the aligned sequences for base modifcations
ipdSummary 470_003.subreads.aligned.bam --reference SE_4047.fasta --gff 470_003.basemods.gff --csv
470_003.basemods.csv --pvalue 0.001 --numWorkers 16 --identify m4C,m6A
#Identifying any consensus motifs
motifMaker find -f SE_4047.fasta -g 470_003.basemods.gff -o 470_003.motifs.csv ### requires more computational
sources than the ipdSummary command
#Creating a GFF file with all of the modification that are part of the motifs
motifMaker reprocess -f SE_4047.fasta -g 470_003.basemods.gff -m 470_003.motifs.csv -o 470_003.motifs.gff

### R code for to filter BaseMod GFF files prior to whole genome comparison set with BEDTools ###
##R Version 4.0.3

library(ape); packageVersion("ape") ## ape: 5.4.1
SEE_20.080 <- read.gff("./SEE_20-080.motifs.gff", GFF3 = TRUE)
##Example of code for a single isolate

library(dplyr); packageVersion("dplyr") ##1.0.3
library(tidyr); packageVersion("tidyr") ##1.1.2

##### S. equi 20.080 - Carrier #####
SEE_20.080_filtered <- filter(SEE_20.080, !grepl('modified_base', type)) #removing instances of modified base
SEE_20.080_filt.motif <- filter(SEE_20.080_filtered, grepl('motif', attributes)) #pulling out modification with motifs
out <- strsplit(as.character(SEE_20.080_filt.motif$attributes), ";");
SEE_20.080_filt.motif_attributes <- data.frame(t(sapply(out, '[')));
colnames(SEE_20.080_filt.motif_attributes) <- c("context", "motif", "coverage", "IPDRatio", "id", "identifcationQv")
#splitting the attributes column into new columns by semi-colon
SEE_20.080_filt.motif <- cbind(SEE_20.080_filt.motif, SEE_20.080_filt.motif_attributes)

SEE_20.080_filt.nomotif <- filter(SEE_20.080_filtered, !grepl('motif', attributes)) # pulling out modifications with out
motifs
out <- strsplit(as.character(SEE_20.080_filt.nomotif$attributes), ";"); SEE_20.080_filt.nomotif_attributes <-
data.frame(t(sapply(out, '[')));
colnames(SEE_20.080_filt.nomotif_attributes) <- c("coverage", "context", "IPDRatio", "identifcationQv") #splitting
the attributes column into new columns by semi-colon
SEE_20.080_filt.nomotif <- cbind(SEE_20.080_filt.nomotif, SEE_20.080_filt.nomotif_attributes)
na <- rep(NA, nrow(SEE_20.080_filt.nomotif)) ##creating columns of NAs to match columns seen in data with motifs
SEE_20.080_filt.nomotif$motif <- na
SEE_20.080_filt.nomotif$id <- na

#Combining the data with and without motifs
SEE_20.080_filtered <- rbind(SEE_20.080_filt.motif, SEE_20.080_filt.nomotif)

out <- strsplit(as.character(SEE_20.080_filtered$identifcationQv), "=");
SEE_20.080_Qv <- data.frame(t(sapply(out, '[')));
colnames(SEE_20.080_Qv) <- c("Qv", "QvScore"); SEE_20.080_Qv$QvScore <-
as.numeric(SEE_20.080_Qv$QvScore)
#pulling out the QV score values
SEE_20.080_filtered <- cbind(SEE_20.080_filtered, SEE_20.080_Qv)

SEE_20.080_QvScore30 <- filter(SEE_20.080_filtered, QvScore >= 30) #Keeping only methylation with a QV score
>= 30

#outputting the filtered data in text and gff file formats
write.table(SEE_20.080_QvScore30, "./SEE_20.080_filtered.txt", sep = "\t", quote = F)
```

```
library(rtracklayer); packageVersion("rtracklayer") ##1.48.0
export(SEE_20.080_QvScore30, "./SEE_20.080_filtered.gff", format = "gff3")

#creating a annotated GFF file with the methylation events across all Streptococcus equi isolates (either from Sweden
or Pennsylvania)
module load BEDTools/2.29.2-GCC-9.3.0

bedtools annotate -i SEE_4047.gff3 -files Car.470_003_filtered.gff Car.470_007_filtered.gff Car.470_008_filtered.gff
Car.489_005_filtered.gff \
Car.489_006_filtered.gff Car.489_007_filtered.gff Car.489_009_filtered.gff Car.489_010_filtered.gff
Clin.470_001_filtered.gff Clin.470_002_filtered.gff \
Clin.470_006_filtered.gff Clin.489_001_filtered.gff Clin.489_002_filtered.gff Clin.489_003_filtered.gff
Clin.489_004_filtered.gff > Carrier_Clinical_Annotated.gff

### R code for identify site of methylation in carrier or clinical isolates from Sweden and Pennsylvania (separately)
###
All_methy_annotated <- read.delim("./Carrier_Clinical_Annotated_editted.txt", header=FALSE)

methy.local <- All_methy_annotated[,7:ncol(All_methy_annotated)]
see <- c("SEE_20.080", "SEE_20.081", "SEE_20.082", "SEE_20.083", "SEE_20.084", "SEE_20.085",
"SEE_20.086", "SEE_20.087", "SEE_20.088", "SEE_20.089", "SEE_20.090", "SEE_20.091",
        "SEE_20.092", "SEE_20.093", "SEE_20.094", "SEE_20.095", "SEE_20.096", "SEE_20.097", "SEE_20.098",
"SEE_20.099", "SEE_20.100")
#see <-
c("Car.470_003","Car.470_007","Car.470_008","Car.489_005","Car.489_006","Car.489_007","Car.489_009","Car.4
89_010","Clin.470_001","Clin.470_002","Clin.470_006",
#        "Clin.489_001","Clin.489_002","Clin.489_003","Clin.489_004") ## For the Sweden isolates
colnames(methy.local) <- see

methy.local <- methy.local[apply(methy.local[,-1], 1, function(x) !all(x==0)),] #removal of rows with all 0s

#Dividing the dataframe by disease state (carrier and clinical)
car.methy <- methy.local[,1:11]
#car.methy <- methy.local[,1:8] ## For the Sweden isolates
car.rows <- rowSums(car.methy) #Calculating row sums for carrier isolates
clin.methy <- methy.local[,12:21]
#clin.methy <- methy.local[,9:14] ## For the Sweden isolates
clin.rows <- rowSums(clin.methy )#Calculating row sums for clinical isolates
methy.rowsum <- cbind(car.rows, clin.rows) #Combining the row sums

colnames(methy.rowsum) <- c("CarrierRowSum", "ClinicalRowSum")
methy.rowsum <- as.data.frame(methy.rowsum)

library(dplyr); packageVersion("dplyr") ##1.0.3
foo <- methy.rowsum %>% filter_all(any_vars(. %in% 0.000000)) #Keeping only rows that have at least 1, zero
value.

B <- row.names(foo)
MethyAnnotate_Subset <- All_methy_annotated[B, ] #Subsetting annotated by the rows identified before, to keep
instances where only methylation occurs either in carrier or clinical isolates
#write.table(MethyAnnotate_Subset, "MethyAnnotate_Subset.txt", sep = "\t")

subset_methy.locat <- methy.local[B,  ]#Subsetting by the rows identified before, to keep instances where only
methylation occurs either in carrier or clinical isolates
write.table(subset_methy.locat, "subset_methy.locat.txt", sep = "\t")
#foo3 <- as.data.frame(rowSums(subset_methy.locat))

methylation <- subset_methy.locat[apply(subset_methy.locat, 1, function(x) sum(x != 0.000000)) >= 6,] ## getting rid
of sites without methylation
```

```
#methylation <- subset_methy.locat[apply(subset_methy.locat, 1, function(x) sum(x != 0.000000)) >= 4,] ## For the
Sweden isolates
C <- row.names(methylation)
methylation_sites <- All_methy_annotated[C, ]
write.table(methylation_sites, "methylation_sites.txt", sep = '\t')

##### Plotting of final sites #####
plot.data <- read.table("MethylationSites.txt", sep = "\t", header = T)
library(ggplot2); packageVersion("ggplot2") ##3.3.2
theme_set(theme_bw())
ggplot(plot.data, aes(x = ID, y = MethylationSum, color = Genome, shape = Status)) +
  geom_point(size = 4, position = position_dodge(width = 0.5)) +
  theme(axis.text.x = element_text(size = 10, angle = 90, vjust = 0.5), axis.text.y = element_text(size = 11))

final.plot <- read.delim("./FinalMethylationSites_Plot.txt", header = T)
ggplot(final.plot, aes(x = Location, y = Methylation, color = Genome, shape = Status)) +
  geom_point(size = 4, position = position_dodge(width = 0.5)) +
  theme(axis.text.x = element_text(size = 10, angle = 90, vjust = 0.5), axis.text.y = element_text(size = 11))

final.plot$TypeSite <- paste(final.plot$Type, final.plot$Site, sep = "_")
final.plot_na.rm <- na.omit(final.plot)
ggplot(final.plot_na.rm, aes(x = Location, y = Motif, color = TypeSite, shape = Status)) +
  geom_point(size = 4, position = position_dodge(width = 0.5)) +
  theme(axis.text.x = element_text(size = 10, angle = 90, vjust = 0.5), axis.text.y = element_text(size = 11))

##################################################################################
### Pennsylvania Streptococcus equi RNA-Seq Workflow - Linux ###
#Checking sequence quality - FastQC
module load FastQC/0.11.6-Java-1.8.0

fastqc -t 2 -o ./ /06_20_085/06-20-085_S21_L001_R1_001.fastq.gz /06_20_085/06-20-
085_S21_L001_R2_001.fastq.gz

#Performing RNA-Seq trimming based on quality output - Trimmomatic
module load Trimmomatic/0.39-Java-1.8.0

java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar PE -threads 8 01_20_080/01-20-
080_S1_L001_R1_001.fastq.gz 01_20_080/01-20-080_S1_L001_R2_001.fastq.gz \
TrimmedSequences/01-20-080_R1_trimmed.fastq.gz TrimmedSequences/01-20-080_R1_unpaired.fastq.gz
TrimmedSequences/01-20-080_R2_trimmed.fastq.gz \
TrimmedSequences/01-20-080_R2_unpaired.fastq.gz TRAILING:25 SLIDINGWINDOW:5:20 HEADCROP:10
LEADING:10 MINLEN:35

#Quantifying the expression of RNA transcripts - Salmon

#Creating the list decoys for the index step
grep "^>" Streptococcus_equi_subsp_equi_4047.ASM2658v1.dna.toplevel.fa | cut -d " " -f 1 > decoys.txt
sed -i.bak -e 's/>//g' decoys.txt
#Combining the cDNA and genome files into a single file for the index step
cat Streptococcus_equi_subsp_equi_4047.ASM2658v1.cdna.all_modified.fa
Streptococcus_equi_subsp_equi_4047.ASM2658v1.dna.toplevel.fa > SEE4047_gentrome.fa

module load Salmon/1.3.0-gompi-2020a
#Creating the reference genome index with 31-mers
salmon index -t SEE4047_gentrome.fa -i SEE4047_index31 --decoys decoys.txt -k 31 --gencode
#Quantifying the RNA transcripts for each of the Streptococcus equi isolates
salmon quant -i SEE4047_index31 -l A -1 01_20_080/01-20-080_S1_L001_R1_001.fastq.gz -2 01_20_080/01-20-
080_S1_L001_R2_001.fastq.gz \
-p 8 --validateMappings --gcBias -o quants_gcbias/20-080.
```

```
#Differential gene expression analysis using R - edgeR
##R Version 4.0.3
#BiocManager::install("tximport")
library(tximport); packageVersion("tximport") ##1.16.1
dir <- "."
samples <- read.csv("./Strep_RNA-seq/SampleData.csv") #importing metadata
rownames(samples) <- samples$run
samples$run <- as.factor(samples$run)
samples$Status <- as.factor(samples$Status)
samples$SeM <- as.factor(samples$SeM)
samples$Location <- as.factor(samples$Location)
files <- file.path(dir, samples$run, "quant.sf") #counts for each of the isolates
names(files) <- samples$run

tx2gene.maybe <- read.table("./Strep_RNA-seq/list.csv", header = T, sep = ',') #importing the list of gene names

txi <- tximport(files, type = "salmon", tx2gene = tx2gene.maybe)

library(edgeR); packageVersion("edgeR") ##3.30.3

cts <- txi$counts
normMat <- txi$length
Status <- samples$Status

y <- DGEList(counts=cts, group = Status)

#performing filtering and normalization
keep <- filterByExpr(y)
y <- y[keep,keep.lib.sizes=FALSE]
y <- calcNormFactors(y)
design <- model.matrix(~Status)
y <- estimateDisp(y,design)

#Running the GLM, quasi-mapping model
fit <- glmQLFit(y, design, robust = T)
qlf <- glmQLFTest(fit, coef = 2)
topTags(qlf) #Looking for any differentially expressed genes from model
tt.all <- topTags(qlf, n = nrow(qlf))

library(EnhancedVolcano); packageVersion("EnhancedVolcano") ##1.6.0
## Plotting a valcano plot, look ing for genes with a FDF <= 0.05; or logFC of < -1 or > 1.
EnhancedVolcano(logFC.qlf, lab = rownames(logFC.qlf), x = 'logFC', y = 'FDR', pCutoff = 0.05, FCcutoff = 1)

library(Glimma);packageVersion("Glimma") ##1.16.0
glMDPlot(qlf, counts=y$counts, groups=Status)

library(dplyr); packageVersion("dplyr") ##1.0.4
#Viewing genes by logFC value, regardless of FDR
nrow(subset(logFC.qlf, logFC < -1 | logFC > 1))
nrow(subset(logFC.qlf, logFC < -1))
nrow(subset(logFC.qlf, logFC > 1))
```

APPENDIX F

CLUSTAL OMEGA MULTIPLE SEQUENCE ALIGNMENT: DIFFERENCES IN THE GENOME, METHYLOME, AND TRANSCRIPTOME DO NOT DIFFERENTIATE ISOLATES OF *STREPTOCOCCUS EQUI* SUBSP. *EQUI* FROM HORSES WITH ACUTE CLINICAL SIGNS FROM ISOLATES OF INAPPARENT CARRIERS

F-1 Appendix. Clustal OMEGA multiple sequence alignment of the SeM DNA sequence (initial 360 base-pairs) from SEE isolates from Sweden (n = 14) and Pennsylania (n = 21).

```
CLUSTAL O(1.2.4) multiple sequence alignment


489_009      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
470_001      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
470_006      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
470_008      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_007      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_010      ------------------------------------------------------------    0
470_003      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
470_007      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_002      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_003      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_004      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
470_002      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_005      atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
489_006      ------------------------------------------------------------    0
20-090       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-081       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-082       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-083       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-080       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-093       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-087       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-095       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-086       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-089       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-096       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-092       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-094       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-100       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-098       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-091       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-084       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-097       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-099       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-088       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60
20-085       atgtttttgagaaataacaagcaaaaatttagcatcagaaaactaagtgccggtgcagca    60


489_009      tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt    120
470_001      tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt    120
470_006      tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt    120
470_008      tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt    120
```

```
489_007     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
489_010     ------------------------------------------------------------      0
470_003     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
470_007     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
489_002     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
489_003     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
489_004     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
470_002     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
489_005     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt     120
489_006     ------------------------------------------------------------      0
20-090     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgtggtt      120
20-081     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-082     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-083     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-080     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-093     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-087     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-095     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-086     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-089     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-096     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-092     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-094     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-100     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-098     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-091     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-084     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-097     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-099     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-088     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120
20-085     tcagtattagttgcaacaagtgtgttgggagggacaactgtaaaagcgaactctgaggtt      120


489_009     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
470_001     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
470_006     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
470_008     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_007     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_010     ------------------------------------------------------------      0
470_003     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
470_007     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_002     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_003     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_004     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
470_002     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_005     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc     180
489_006     ------------------------------------------------------------      0
20-090     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-081     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-082     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-083     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-080     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-093     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-087     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-095     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-086     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaaacgatatagcc      180
20-089     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgagatagcc      180
20-096     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-092     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgttatagcc      180
20-094     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-100     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-098     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-091     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-084     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-097     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-099     agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
```

```
20-088      agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180
20-085      agtcgtacggcgactccaagattatcgcgtgatttaaaaaatagattaagcgatatagcc      180


489_009      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
470_001      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
470_006      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
470_008      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_007      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_010      ------------------------------------------------------------      0
470_003      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
470_007      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_002      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_003      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_004      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
470_002      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_005      atagatagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
489_006      ------------------------------------------------------------      0
20-090      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-081      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-082      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-083      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-080      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-093      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-087      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-095      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-086      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-089      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-096      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-092      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-094      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-100      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-098      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-091      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-084      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-097      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-099      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-088      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240
20-085      ataagtagagatgcctcatcagcccaaaaagttcgaaatcttctaaaaggcgcctctgtt      240


489_009      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
470_001      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
470_006      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
470_008      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_007      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_010      ------------------------------------------------------------      0
470_003      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
470_007      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_002      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_003      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_004      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
470_002      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_005      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
489_006      ------------------------------------------------------------      0
20-090      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-081      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-082      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-083      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-080      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-093      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-087      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-095      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-086      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtaaagat      300
20-089      ggggatttacaggcattattgagaggtcttgattcagaaagggctgcgtatggtagagat      300
20-096      ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat      300
```

```
20-092    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-094    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-100    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-098    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-091    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-084    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-097    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-099    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-088    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300
20-085    ggggatttacaggcattattgagaggtcttgattcagcaagggctgcgtatggtagagat    300


489_009    gattattacaacttattgatacacctttcatcgatgttaaatgataaacctgatggggat    360
470_001    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
470_006    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
470_008    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_007    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_010    --------------------------------atgttaaatgataaacctgatggggat    27
470_003    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
470_007    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_002    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_003    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_004    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
470_002    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_005    gattattacaacttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
489_006    ------------------atgcacctttcatcgatgttaaatgataaacctgatggggat    42
20-090    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-081    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-082    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-083    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-080    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-093    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacttgatggggat    360
20-087    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacttgatggggat    360
20-095    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-086    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-089    gattattacaatttattgatgcacctttcatcgatgaaaaatgataaacctgatggggat    360
20-096    gattattacaatttattgatgcaactttcatcgatgttaaatgataaacctgatggggat    360
20-092    gattattacaatttattgatgcgcctttcatcgatgttaaatgataaacctgatggggat    360
20-094    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-100    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-098    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-091    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-084    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-097    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-099    gattattacaatttattgatgcacctttcatcgatgttaaatgataaacctgatggggat    360
20-088    gattattacaatttattgatgcgcctttcatcgatgttaaatgataaacctgatggggat    360
20-085    gattattacaatttattgatgcgcctttcatcgatgttaaatgataaacctgatggggat    360
                                    ***  *********** **********
```