The Use of Random Mutagenesis in the Functional Annotation of the *Streptococcus uberis* Genome

Adam Blanchard, BSc.

Thesis submitted to the
University of Nottingham
for the degree of
Doctor of Philosophy

July 2015



Try not to become a man of success, but rather try to become a man of value

- Albert Einstein

Abstract

Streptococcus uberis is a prolific and virulent mastitic pathogen, a disease that is a substantial welfare issue, costing the global dairy industry billions of pounds each year. Research into identifying control measures for this bacterium have so far been unsuccessful; partly due to it ubiquitous presence in the environment of dairy cattle and it's the multifactorial nature of the disease pathogenesis.

High-throughput, transposon insertions sequencing (TIS) protocols utilise libraries of bacteria, mutagenised using insertional transposon techniques, to screen for genomic regions involved in survival under particular environments. Significant advances in sequencing technology and the ease and availability of data generation have fuelled TIS experiments. However, the cost of these protocols coupled with the requirements to alter core sequencing-machine parameters, subsequently the experiments are implausible for those without free access to a sequencing platform.

Addressing these problems and wanting to provide an accessible, accurate, data analysis package, Pragmatic Insertion Mutant Mapping System (PIMMS) was developed. PIMMS builds on the use of the highly efficient, random mutagenic tool, pGhost9::IS*S1* and extends the principles of single mutant identification by inverse PCR. This enabled the enrichment of the transposon insertion/chromosomal DNA junction from a population of 100,000 mutant bacteria, generating a sample that could be used directly for high-throughput

sequencing.

Making use of multiple bioinformatic tools, a data analysis pipeline was created allowing for the raw sequence data to be analysed, isolating reads containing the DNA target, mapping them to the reference genome and creating a summary tabular output. This contains all of the insertion positions identified accompanied by an array of normalised data in order to identify genes of consequence to the selective condition under investigation. Using these methods allowed identification of a core set of 196 genes essential for growth of *Streptococcus uberis* on agar.

Further interrogation of this bacterium in host specific environments afforded the identification of four further subsets of genes important for growth in the associated conditions. During growth in liquid medium a total of 411 non-mutated genes were identified and a further 69 non-mutated genes were identified following the addition of lactoferrin (a divalent metal ion chelating protein) to the media. After growth in skimmed milk, 383 genes were identified as being conditionally essential. The metal binding protein, *mtuA,* was found to be devoid of insertions, in the presence of lactoferrin and in skimmed milk, complementing other published studies that show individual mutant strains of *S. uberis,* lacking MtuA are unable to grow in milk or in the presence of lactoferrin. Identification of the *mtuA* sequence using the PIMMS procedure provided a validation of the techniques ability to detect appropriate sequences. Following supplementation of skimmed milk bovine

serum (25%), a total of 656 CDS were deemed conditionally essential, incorporating most of those required for growth in skimmed milk alone, but also included additional sequences; seven of which were similar to those essential for growth of *Staphylococcus aureus* in a similar environment.

The precise protocol developed will enable the advancement of the initial stages of functional genomics in streptococci; due to the wide range of species in which pGhost9::IS*S1* will integrate. Also the bioinformatic pipeline can be coupled with any insertional mutagen to enable cost effective interrogation of any bacteria with an appropriate mutagenesis system.

Publications

Blanchard, A.M., Leigh, J.A., Egan, S.A., Emes, R.D. (2015) Transposon insertion mapping with PIMMS – Pragmatic Insertion Mutation Mapping System. *Frontiers in Genetics* 6: 139

Acknowledgments

Firstly I would like to thank the University of Nottingham, School of Veterinary Medicine and Science for allowing me to use their facilities, in order to undertake the research for this degree, and Zoetis for the generous funding of the project. I would like to give special thanks to my supervisors Professor Jamie Leigh for answering countless “quick questions” that turned out to be far from quick, Professor Richard Emes for dealing with the Friday afternoon data analysis panics, Dr Sharon Egan for sharing her knowledge and technical skills and for all of their support, advice and enthusiasm throughout my research. Thanks to the other SVMS post graduate students, especially Dr Amy Bowen for persuading me to talk to Jamie and Sharon in the first place, Maqsud Hossain and Dr Frank Wessely for discussions and help with bioinformatics, Ginny Sherwin for her shared dislike of bad radio stations and general lab chatter, Oli Maxwell and Pete Down for their time and efforts collecting milk on the farm (despite being in the middle of a snowy February). I would also like to thank Steve Brogden and John Remnant for the deep and meaningful discussions about cheese over morning coffee. I also appreciate the help from Dr Sabine Tötemeyer, Dr Mike Jones and Yvonne Allen with my demonstrating and Dr Sunir Malla for his technical ability and patience when discussing the finer points of NGS. On a personal level I would like to thank Clare and the rest of my family for all of their love, support and on-going encouragement through the last few years.

Table of Contents

Abstract i

Publications iv

Acknowledgments v

Table of Contents vi

Table of Figures xiii

Table of Tables xvi

List of Abbreviations xix

Chapter 1 Introduction 1

1.1. Streptococci 1

1.1.1. Taxonomy and nomenclature of the streptococcus genus 1

1.2.1. Disease caused by pyogenic streptococci 5

1.2.1.1. Disease caused by streptococci in humans 5

1.2.1.2. Disease caused by streptococci in animals 6

1.2.1.3. Disease caused by streptococci in cattle 8

1.2. Streptococcus uberis 9

1.2.1. The role of *S. uberis* in bovine mastitis 9

1.2.1.1. Epidemiology and strain typing of *S. uberis* 9

1.2.2. Host Pathogen Interactions 11

1.2.2.1. Physical defence mechanisms against mastitic pathogens 12

1.2.2.2. Soluble defence factors found within the bovine mammary gland 12

1.2.2.3. The activation of the immune response 15

1.2.2.4. Cellular defence mechanisms 16

1.2.2.5. The humoral defence system 17

1.2.2.6. Virulence factors of *S. uberis* 18

1.3. Essential genome 22

1.3.1. Transposon mutagenesis 23

1.3.1.1. Transposition using the Tn family of composite insertion elements 24

1.3.1.2. Mu bacteriophage transposition 24

1.3.1.3. Eukaryotic origin Mariner transposition 25

1.3.1.4. Insertional mutagenesis using IS insertional elements 25

1.3.2. High-throughput mutation mapping techniques 27

1.3.3. Sequencing of DNA 31

1.5. Aims and Objectives 35

Chapter 2 Materials and Methods 36

2.1. Generation of bacterial mutant Pools 36

2.2. Extraction of chromosomal DNA from *S. uberis* mutant pools 37

2.3. Quantification of DNA Concentration 38

2.4. Digestion of DNA using restriction endonuclease 38

2.5. Fragmentation of DNA using Adaptive-Focused-Acoustics 38

2.6. Preparation of DNA for inverse PCR reactions 39

2.7. Amplification of flanking insertion regions using Inverse PCR 40

2.8. Bioanalyser Quantification of DNA Fragment Distribution 40

2.9. Next Generation Sequencing 41

Chapter 3 Analysis of the enriched insertion and chromosomal DNA junction 42

3.1. Introduction 42

3.2. Results and discussion 48

3.2.1. Generation of mutant pools 48

3.2.2. Validation of inverse PCR on a population of mutants 49

3.2.3. Enrichment of insertion flanking regions for acoustically fragmented DNA 55

3.2.4. Sequencing of inverse PCR products 57

3.2.5. Analysis of data 57

3.2.6. Development of a data analysis pipeline 60

3.2.6.1. Identification of terminal insertion sequence and mapping to genome 62

3.2.6.2. Analysis of alignment methods 64

3.2.6.3. Processing insertion coordinates 70

3.4. Conclusions 75

3.4.1. Use of inverse PCR as a method of transposon site enrichment 75

3.4.2. Bioinformatic analysis of transposon insertion site sequence data 78

Chapter 4 Determination of essential genes in the *Streptococcus uberis* 0140J genome 80

4.1. Introduction 80

4.2. Results and discussion 85

4.2.1. Analysis of the *Streptococcus uberis* mutant insertion pools 85

4.2.3. Identification of *S. uberis* Essential genes 91

4.2.3.1. Functional categorisation of the essential *S. uberis* genes 91

4.2.3.2. Categorisation of the hypothetical and putative essential genes 96

4.2.3.3. Homology of the essential genes of *S. uberis* 99

4.2.4. Analysis of metabolic pathways associated with essential genes 102

4.3. Conclusions 111

Chapter 5 Identification of conditionally essential genes required for growth in liquid culture 114

5.1. Introduction 114

5.1.1. Examples of transposon insertion sequencing using phenotypic screening 115

5.1.2. Identification of additional essential genes associated with growth in liquid culture 118

5.1.3. Growth of *S. uberis* in the presence of lactoferrin 119

5.1.4. Analysis of conditionally essential genes in test vs. control experiments 120

5.2. Materials and Methods 123

5.2.1. Growth of bacteria in liquid cultures for phenotypic studies 123

5.2.2. Extraction of bacteria from liquid culture (THB) 123

5.3. Results 124

5.3.1. Growth of *S. uberis* in liquid culture 124

5.3.1.1. Growth parameters in liquid culture 124

5.3.1.2. Analysis of sequencing data 125

5.3.1.3. Identification of essential CDSs for growth in Todd Hewitt broth 126

5.3.2. Growth of *S. uberis* in the presence of lactoferrin 130

5.3.2.1. Growth parameters in the presence of lactoferrin 130

5.3.2.2. Analysis of the sequencing data 131

5.3.2.3. Identification of the CDSs essential for growth in THB containing lactoferrin 132

5.4. The identification of insertions that are selected or counter selected in the presence of lactoferrin 138

5.5. Discussion 140

Chapter 6 Identification of conditionally essential genes in biologically relevant environments 146

6.1 Introduction 146

6.1.1. Growth in skimmed milk 146

6.1.2. Growth of *S. uberis* in milk containing serum 148

6.1.3. Initial investigations required for robust data generation 149

6.1.4. Analysis of conditionally essential genes with multiple replicates 150

6.2. Materials and Methods 152

6.2.1. Preparation of milk samples for phenotypic growth studies 152

6.2.4. Extraction of bacterial colonies from complex liquid culture (Milk) 152

6.4. Results 153

6.4.1. Preliminary investigations 153

6.4.1.1. The use of fewer biological replicates 153

6.4.1.2. Effect of selective antibiotics 155

6.4.2. Growth of *S. uberis* in skimmed milk 156

6.4.2.1. Growth parameters 156

6.4.2.2. Analysis of the sequencing data from input and output pools 157

6.4.2.3. Identification of the essential genes for growth in skimmed milk 158

6.4.2.4. The use of differential insertion prevalence to identify phenotypically important genes 168

6.4.3. Growth of *S. uberis* in skimmed milk with serum 169

6.4.3.1. Growth parameters 169

6.4.3.2. Analysis of the sequencing data from input and output pools 170

6.4.3.3. Identification of the essential genes for growth in milk supplemented with serum 171

6.4.3.4. The use of differential insertion prevalence to identify phenotypically important genes 180

6.5. Discussion 182

6.5.1. Growth of *S. uberis* in skimmed milk 182

6.5.2. Growth of *S. uberis* in milk containing serum 185

Chapter 7 General discussion 190

7.1. Benchmarking the PIMMS process with other TIS experiments 190

7.2. Identification of truncated CDSs that are consistently found in these data 192

7.3. Conserved essential genes between different clinical and sub-clinical isolates of *S. uberis* 195

7.4. Future work 200

7.4.1. Linking more phenotypes to genotypes 200

7.4.2. Application to other bacteria 201

7.4.3. Genetic regulation by non-coding chromosomal features 203

References 205

Appendices 240

Appendix 1 241

Appendix 2 276

Table of Figures

Figure 1.1. Streptococci Phylogeny grouping based on core genome 4

Figure 1.2. Illustrative methodology of the four main transposon insertion sequencing protocols. 28

Figure 1.3. Representation of the stages involved in Reverse-terminator sequencing by synthesis 34

Figure 3.1. pGhost9::IS*S1* plasmid schematic representation 43

Figure 3.2. Validation of Oligonucleotide primers to be used in the inverse PCR on a population of mutants 51

Figure 3.3 Distribution of restriction endonuclease sites within the *S. uberis* genome 52

Figure 3.4. Electropherograms of acoustically fragmented DNA 54

Figure 3.5. Inverse PCR products using the acoustic fragmented DNA at different sizes 56

Figure 3.6a Association between Phred score and probability of error

Figure 3.6b Quality score graphs generated from each base Phred score 59

Figure 3.7. Schematic of the PIMMS data analysis process 61

Figure 3.8. Effect of read length on alignment confidence 63

Figure 3.9. Random sampling graphs generated by number of alignments 68

Figure 3.10. Venn diagrams demonstrating the overlap of alignments using bowte2 and BWA with the most stringent settings 69

Figure 3.11. Circular representation plots of unique insertion points identified by each sample preparation method 74

Figure 4.1. Flowchart demonstrating the laboratory procedure for amplification of unknown regions of DNA using inverse PCR 85

Figure 4.2. Circos representation of each insertion site identified within the *S. uberis* genome 90

Figure 4.3. Centile position of every insertion within all CDS 90

Figure 4.4. Proportional enrichment of essential genes associated with each RAST category 104

Figure 4.5. Aspects of the pathways associated with production of acetyl-CoA 109

Figure 4.6. Graphical representation of the fatty acid biosynthesis metabolic 110

Figure 5.1. Growth of *S. uberis* 0140J pGhost9::IS*S1* in Todd Hewitt Broth 124

Figure 5.2. Distribution of insertions identified within the *S. uberis* genome after growth in Todd Hewitt broth 125

Figure 5.3. A Venn diagram comparing the number of essential genes found after growth of *S. uberis* on solid medium and liquid media 126

Figure 5.4. Proportional enrichment of essential genes associated with each RAST category. 129

Figure 5.5. Growth curve of *S. uberis* 0140J pGhost9::ISS1 in THB with the addition of 1mg/ml lactoferrin 130

Figure 5.6. Distribution of insertions identified within the *S. uberis* genome after growth in Todd Hewitt broth with the addition of lactoferrin 131

Figure 5.7. A Venn diagram illustrating the overlap of essential genes between growth conditions 133

Figure 5.8. Proportional enrichment of essential genes associated with each RAST category 137

Figure 6.1. Saturation of disrupted CDSs using 10 pools of mutants 154

Figure 6.2. A Venn diagram showing the overlap of the number of essential CDSs from the five biological replicates 154

Figure 6.3. Growth curve assessing the effect of erythromycin on *S. uberis* numbers during culture in skimmed milk 155

Figure 6.4. Growth assessment of *S. uberis* 0140J pGhost9::IS*S1* in skimmed milk 156

Figure 6.5. Distribution of insertions identified within the *S. uberis* genome after growth in skimmed milk 158

Figure 6.6. Proportional enrichment of essential genes associated with each RAST category 166

Figure 6.7. *S. uberis* growth curve in skimmed milk with varying amounts of serum 169

Figure 6.8. Distribution of insertions identified within the *S. uberis* genome after growth in skimmed milk 171

Figure 6.9. Proportional enrichment of essential genes associated with each RAST category 178

Table of Tables

Table 1.1 – β-haemolytic streptococci with veterinary importance designated by Lancefield grouping 2

Table 3.1 – Bacterial counts from serial dilutions to determine numbers of insertions within a pool of mutants 49

Table 3.2 – The associated number of reads and average Phred scores for each sample 58

Table 3.3 – Associated number of reads that contain the transposon motif 64

Table 3.4 – Metrics of reads that aligned with BWA to the *S. uberis* genome 70

Table 3.5 – Metrics of insertion point discovery relative to CDS 72

Table 4.1 – Quantity of DNA at each stage of the enrichment process 87

Table 4.2 – Metrics of analysed sequencing data 88

Table 4.3 – Essential coding sequences and their associated RAST categories 93

Table 4.4 – Truncated coding sequences and their associated RAST categories 94

Table 4.5 – Non-homologous essential genes of *S. uberis* 101

Table 4.6 – Enrichment identification of essential gene RAST category 103

Table 5.1 – Essential coding sequences and their associated RAST categories after growth in Todd Hewitt broth 127

Table 5.2 – Truncated coding sequences and their associated RAST categories after growth in Todd Hewitt broth 128

Table 5.3 – Additional essential genes identified after growth in Todd Hewitt broth with the addition of lactoferrin 134

Table 5.4 – Truncated genes identified after growth in Todd Hewitt broth with the addition of lactoferrin 135

Table 5.5 – Additional genes identified only in growth of Todd Hewitt broth 136

Table 5.6 – Unique insertions that have statistically significant reduction in THB supplemented with lactoferrin 139

Table 6.1 – Additional essential CDSs identified in the milk input pool 160

Table 6.2 – Truncated CDSs identified in the milk input pool 161

Table 6.3 – Additional essential CDSs found after growth in milk 162

Table 6.4 – Truncated CDSs found after growth in milk 163

Table 6.5 – Essential CDSs found before growth in milk only 164

Table 6.6– Comparison between the numbers of CDSs lacking insertions in the skimmed milk phenotypic screen in association to RAST categories 165

Table 6.7 – Identification of RAST categories that show a level of enrichment after growth in skimmed milk 167

Table 6.8 – Genes with statistically significant fold change differences after screening in skimmed milk with serum 168

Table 6.9 – Essential CDSs identified before growth in milk with serum 172

Table 6.10 – Truncated CDSs identified before growth in milk with serum 173

Table 6.11 – Additional essential CDSs identified after growth in milk with serum 174

Table 6.12 – Truncated CDSs identified after growth in milk with serum 175

Table 6.13 – Additional essential CDSs identified before growth in milk with serum only 176

Table 6.14– Comparison between the numbers of CDSs lacking insertions in the skimmed milk supplemented with serum phenotypic screen in association with RAST categories. 177

Table 6.15 – Identification of RAST categories that show a level of enrichment after growth in milk with serum 179

Table 6.16 – Genes with statistically significant fold change differences after screening in skimmed milk with serum 180

Table 7.1 – Isolates used for homology comparison and the case status 194

Table 7.2 – Essential CDSs identified in skimmed milk and with the addition of serum that are not conserved between different isolates of *S. uberis* 196

List of Abbreviations

ABC	Adenosine Tri-phosphate Binding Transporter
ADP	Adenosine Di-Phosphate
AP-1	Activator Protein 1
ARTIST	Artificial Transposon Insertion Site Tracker
AS	Alignment Score
AT	Adenine Thymine
ATP	Adenosine Tri-Phosphate
Bayseq	Bayesian Sequencing Analysis
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BSA	Bovine Serum Albumin
°C	Degrees Celsius
CD	Cluster of Differentiation
CDD-BLAST	Conserved Domain Database BLAST
cDNA	Complementary Deoxyribonucleic acid
CFU	Colony Forming Units
CMN	Corynebacteria Mycobacteria Nocardia
ddNTP	Dideoxynucleoitde
DeADMAn	Designer Arrays for Defined Mutant Analysis
DEG	Database of Essential Genes
DESeq	Differential Expression Sequencing
DNA	Deoxyribonucleic Acid

dNTPs	Deoxynucleotides
EC	Enzyme Commission Number
edgeR	Empirical analysis of Digital Gene Expression in R
EDP	Entner-Doudoroff Pathway
EDTA	Ethylenediaminetetetraacetic Acid
EGGS	Essential Genes on a Genomic Scale
EMBL	European Molecular Biology Laboratory
EMP	Embden-Meyerhof Pathway
EOD	Early Onset Disease
Ery	Erythromycin
FAS	Fatty Acid Synthesis
FASTA	Fast file All
FASTN	Fast file Nucleotides
FASTP	Fast file Proteins
FASTQ	Fast file with Quality scores
FastQC	Fast file Quality Control
FASTX	Fast file toolkit
FTP	File Transfer Protocol
FU	Florescent Unit
g	Gravity
GC	Guanine Cytosine
gDNA	Genomic Deoxyribonucleic Acid
GFF	General Features Format File
IFN	Interferon

IL	Interleukin
In-Seq	Insertion Sequencing
IR	Inverted Repeat
IS*S1*	Bacterial Insertion Sequence 1
IV	Intra-venous
kb	kilobase pair
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LOD	Late Onset Disease
LTA	Lipoteichoic Acid
M	Molar
MATT	Microarray Tracking Transposon Mutants
MD	Mismatching bases
Mg	Magnesium
mg/L	Milligram per Litre
mg/ml	Milligram per Millilitre
MLST	Multi Locus Sequencing Type
mM	Millimolar
Mn	Manganese
MNF	Monomicrobial Necrotising Fasciitis
n	Number
NaCl	Sodium Chloride
NAPP	Nucleic Acid Phylogenetic Profiling
NFKB	Nuclear Factor Kappa B
ng/µl	Nanograms per microliter

NIM	Normalised Insertions Mapped
NMPDR	National Microbial Pathogen data Resource
NRM	Normalised Reads Mapped
OGEE	Online Gene Essentiality Database
p	Probability
PARTIC	Pathosystems Resource Integration Centre
PASS	Program to Align Short Sequences
PCR	Polymerase Chain Reaction
PDEs	Phosphodiesterases
PF	Pass Filter
PFGE	Pulse Filed Gel Electrophoresis
PhiX	Phage X 174
PIMMS	Pragmatic Insertion Mutant Mapping System
PRR	Pattern Recognition Receptor
PSI-BLAST	Position Specific Iterated BLAST
PTS	Phosphotransferase Transport System
Q	Quality Score
RAST	Rapid Annotation using Subsystem Technology
RNA	Ribonucleic Acid
RNASeq	Ribonucleic Acid Sequencing
rpm	Revolutions per minute
sam	Sequence Alignment/Map
SD	Standard Deviation
SDS	Sodium Dodecyl-Sulphate

ST	Sequence Type
STM	Signature Tagged Mutagenesis
SUAM	*Streptococcus uberis* Adhesion Molecule
TCS	Two-Component System
TE	Tris Ethylenediaminetetetraacetic Acid
THA	Todd Hewitt Agar
THA^e	Todd Hewitt Agar with Erythromycin
THB	Todd Hewitt Broth
TIS	Transposon Insertions Sequencing
TLR	Toll Like Receptor
Tn	Transposon
Tn-Seq	Transposon Sequencing
TNF	Tumour Necrosis Factor
TraDIS	Transposon Directed Insertions Sequencing
TraSH	Transposon Site Hybridisation
Tris-HCl	Tris Hydrochloride
tsv	Tab Separated Values
U	Units
μg	Micrograms
μg/ml	Micrograms per millilitre
μl	Microliter
UTR	Untranslated region
UniProt	Universal Protein Resource
UV	Ultra Violet

v/v	Volume per Volume
W	Watts
w/v	Weight per Volume

Chapter 1 Introduction

1.1. Streptococci

1.1.1. Taxonomy and nomenclature of the streptococcus genus

The genus *Streptococcus* was first observed and descriptively named in the early 1870s by the German surgeon Theodor Billroth. Streptococci are non-motile, non-sporeforming bacteria that appear in pairs or chains due to their cell division mechanism occurring on a single plane. There are approximately 40 species of streptococci, which are widely distributed throughout nature.

In a clinical setting, streptococci are often divided into three groups based on their haemolysis of blood agar: α-haemolytic streptococci cause a green colour change due to their ability to reduce the iron present in the haemoglobin, β-haemolytic streptococci rupture red blood cells creating a clear area around the colonies, whilst γ-streptococci have no impact on the agar colour or appearance. Early classification of β-haemolytic streptococci separated them into two groups: pyogenic and viridans (Sherman 1937) based on whether they were isolated from an animal or human source. Further characterisation was based on serological grouping deemed Lancefield testing. This is an anti-carbohydrate precipitin reaction test directed against specific cellular antigens present in the cell wall (Lancefield 1933). There are a total of 12 Lancefield groups (Table 1.1.), however a number of species do not react unambiguously with a specific typing serum and are designated non-typeable.

Table 1.1. β-haemolytic streptococci with veterinary importance designated by Lancefield grouping.

Lancefield group	Notable species
A	*Streptococcus pyogenes*
B	*Streptococcus agalactiae*
C	*Streptococcus equisimilis, equi, zooepidemicus and dysgalactiae*
D	Enterococcus spp. *and Streptococcus bovis*
E	*Streptococcus milleri and mutans*
F	*Streptococcus anginosus*
G	*Streptococcus canis and dysgalactiae*
H	*Streptococcus sanguis*
L	*Streptococcus dysgalactiae*
N	*Lactococcus lactis*
R&S	*Streptococcus suis*
Non-Typeable	*Streptococcus uberis, Streptococcus iniae*

Adapted from (Facklam 2002).

The Lancefield system of streptococcal classification is still often used in veterinary and human medicine as part of an identification procedure. However, it cannot be used in isolation to accurately identify specific species (Facklam 2002). The current means of identification of streptococci rely on phenotypic tests, which largely provide acceptable species differentiation (Täpp et al. 2003). However some stains may exhibit trait and biochemical variability leading to incorrect identification (Kikuchi et al. 1995). Streptococcal 16S DNA sequences have been shown to provide a robust method for determining species. Such sequences have also been used to

investigate genealogical relationships of the genus, extending the pyogenic cluster to include *Streptococcus agalactiae, Streptococcus uberis, Streptococcus parauberis* and *Streptococcus porcinus* (Bentley et al. 1991). A more recent and definitive method for grouping streptococci is by phylogenetic interrelationships. This is based on comparative genomics (Richards et al. 2014). Whole genome sequences for 44 streptococci were annotated for putative protein coding sequences (CDSs). Protein sequences were delineated into clusters based on shared homology using a Markov clustering method (Enright et al. 2002) and BLAST-P alignments (Altschul et al. 1990). This identified 159 gene clusters that were single gene copies, conserved across all of the species investigated. Evidence of recombination events led to the removal of 23 clusters, leaving 136 clusters. The use of the 136-cluster dataset enabled the generation of a high-resolution phylogeny tree identifying discrete variations. This approach led to the extension of the pyogenic cluster to include 17 different species, making it the largest phylogenetic group of streptococci (Figure 1.1.).

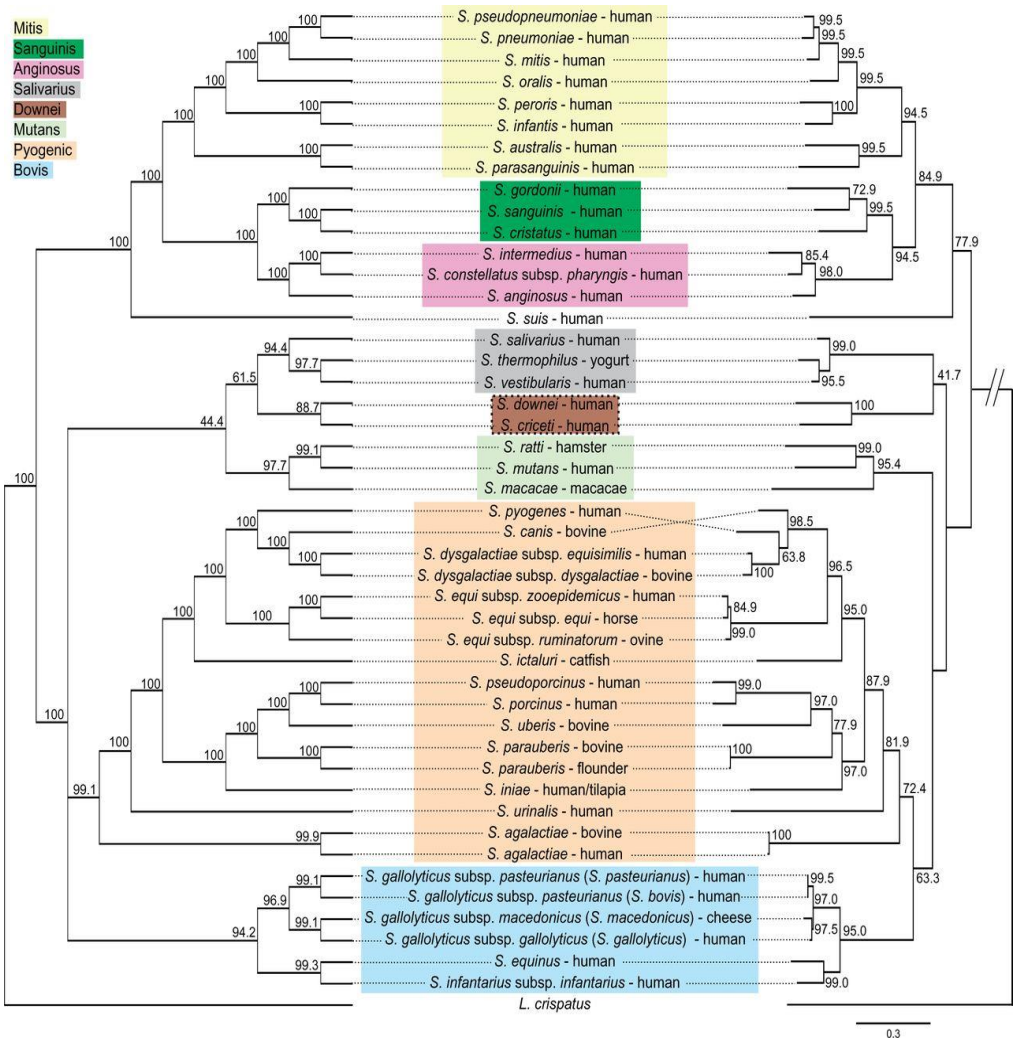

Figure 1.1. Streptococci Phylogeny grouping based on core genome. Each major group is colour shaded. Phylogenies derived from 136 core genes. Numbers on the left show proportions of genes that support the grouping. The right hand side shows maximum likelihood phylogeny derived from concatenation of the core genes. Divergent nomenclature from previous studies show in parentheses. Copied from (Richards et al. 2014).

1.2.1. Disease caused by pyogenic streptococci

Pyogenic streptococci are opportunistic pathogens capable of causing disease in distinct and different hosts. Many were known to cause infections in animals before they were considered agents of disease in humans and are often responsible for high morbidity and mortality outbreaks in fish, birds and various mammals (Chanter 1997)

1.2.1.1. Disease caused by streptococci in humans

Streptococcus pyogenes (*S. pyogenes*), *Streptococcus agalactiae* (*S. agalactiae*) and *Streptococcus dysgalactiae* subspecies *equisimilis* (*SDSE*) are major human pathogens. *S. pyogenes* is responsible for a range of diseases with variable severity. Most commonly infections originate in the throat with mild tonsillitis but certain strains, carrying phage T12, produce the erythrogenic toxin, streptococcal pyrogenic exotoxin A (*speA*), which can lead to scarlet fever (Stevens et al. 1989). Certain strains of *S. pyogenes* also produce streptococcal pyrogenic exotoxin C (*speC*), which can lead to streptococcal mediated toxic shock syndrome (Johansson et al. 2010). However, the most recent outbreak of streptococcal mediated toxic shock syndrome, was caused by *Streptococcus suis (S. suis).* The root of infection was deemed consumption of animals after culling for disease control (Tang et al. 2006; Sriskandan & Slater 2006) and led to 38 deaths in the Sichuan district of China.

S. agalactiae is an asymptomatic commensal bacteria of the human

gastrointestinal tract. In certain circumstances *S. agalactiae* can cause a severe invasive infection of placental tissues leading to neonatal septicaemia. Infants are infected *in utero* or during birth and the infections are separated into early onset (EOD) or late onset disease (LOD). EOD has a incidence of 0.56 in 1000 births in Europe and a neonate mortality rate of 9.6% (Edmond et al. 2012), whereas LOD which can cause meningitis has an incidence of 0.26 per 1000 births and a neonate mortality rate of 27% (Baker 2013).

Streptococcus dysgalactiae subspecies *equisimilis* (*SDSE)* was first divided from *Streptococcus dysgalactiae* subspecies *dysgalactiae* in 1996 to differentiate between human and animal pathogens (Vandamme et al. 1996). *SDSE* is often found in the upper respiratory, gastrointestinal and female genital tracts but is primarily found in soft tissue infections such as benign abscesses (Brandt & Spellerberg 2009) but SDSE can cause severe monomicrobial necrotising fasciitis (MNF) (Broyles et al. 2009). The mortality rate associated with MNF has not decreased over the past 30 years and remains at 25-30% (Anaya et al. 2013). Mortality is correlated with time to intervention and the highest incidence of MNF is in IV drug users. This could be related to diabetes control or narcotic abuse (Chen et al. 2001)

1.2.1.2. Disease caused by streptococci in animals

Many streptococci are responsible for causing a significant level of disease in animals. However, those that effect horses, pigs and cattle, have received the most attention because of the impact they pose on the economy of food

production, or animal welfare. The major pathogen involved in equine disease is *Streptococcus equi* subspecies *equi* (*S. equi*), which is capable of causing abcessation of the retropharyngeal and submandibular lymph nodes (Timoney & Kumar 2008). Despite low mortality rates (approximately 2%) (Robinson et al. 2013) the severity of the clinical signs means the enlarged lymph nodes can restrict the airways, which at worst, can lead to suffocation, or at least, make individuals reluctant to eat causing anorexia and weakness (Waller 2014). *S. zooepidemicus* despite sharing 97% genome identify with *S. equi*, is often regarded as a commensal bacterium; however, it has also been associated with respiratory disease (Wood et al. 2005) and uterine infections leading to abortion in horses (Smith, K. C. et al. 2003).

Aquaculture makes a major contribution to human nutrition, and diseases caused by *Streptococcus agalactiae*, *Streptococcus iniae*, and *Streptococcus parauberis* account for an annual loss of approximately $150 million to the fish industry (Shoemaker et al. 2000). Streptococcal infection has become a major problem, causing outbreaks of streptococcosis in wild-caught and farmed mariculture fish (Toranzo et al. 2005) and significantly reducing numbers of wild populations (Baeck et al. 2006).

In pigs, *S. suis* is of significant importance causing meningitis, pneumonia and septicaemia. *S. suis* is carried asymptomatically for a long period of time in tonsillar crypts, with onset of disease often being sudden and usually fatal if not treated (Chanter 1997).

1.2.1.3. Disease caused by streptococci in cattle

The principle streptococcal disease in cattle is bovine mastitis. Bovine mastitis is defined as the inflammation of the mammary gland and is typically in response to invasion of the udder and colonisation by a pathogenic microorganism (Taylor & Kerro Dego 2002). Mastitis is classified as either clinical or subclinical. Clinical mastitis is defined by the presence of overt inflammatory changes such as swelling and udder tenderness and/or dramatic alterations to milk composition. Subclinical infection by definition lacks any outward clinical signs meaning they are more challenging to diagnose and treat. Mastitis is one of the leading health and welfare concerns for the global dairy industry (Oviedo-Boyso et al. 2007). It was deemed the infectious disease that most impacted dairy cattle by the Farm Animal Welfare Council (Farm Animal Welfare Council 2009). The economic impact of mastitis alongside typical preventative measures account for an annual cost of approximately £200 million in the U.K. and around $35 billion worldwide (Biggs 2009). Infection can be caused by over 137 different organisms, with the majority of cases in the U.K. from 5 bacterial species; *Escherichia coli*, *Streptococcus uberis*, *Staphylococcus aureus*, *Streptococcus dysgalactiae* subsp. *dysgalactiae* and *Streptococcus agalactiae* (Bradley 2002). However, *S. uberis* accounts for 23.5% of clinical mastitis cases and 64% of sub-clinical cases (Bradley et al. 2007).

1.2. Streptococcus uberis

Streptococcus uberis (S. uberis), first described in the 1930s as a Gram-positive; catalyse negative cocci, which forms smooth, translucent, convex colonies when grown on solid media (Diernhofer 1932). *S. uberis* is non-typeable by the Lancefield scheme although occasional reaction to group E has been reported (Roguinsky 1977). Based on phylogenomics, *S. uberis* is part of the pyogenic cluster along with *Streptococcus equi, Streptococcus pyogenes, Streptococcus dysgalactiae* subsp. *dysgalactiae, Streptococcus agalactiae, Streptococcus porcinus, Streptococcus parauberis* and *Streptococcus iniae* (Richards et al. 2014).

1.2.1. The role of S. uberis in bovine mastitis

S. uberis can be routinely isolated from soil, pasture, and bedding meaning there is a continual source of infection for new individuals even without direct contact and contagious spread between cows (Kruze & Bramley 1982). Its ability to survive in the environment means it is difficult to completely eradicate from the population, and is a constant source of potential infection throughout herds. However, the significance and association between environmental contamination and infection of the udder remains unclear (Lopez-Benavides et al. 2007).

1.2.1.1. Epidemiology and strain typing of S. uberis

S. uberis can be isolated from multiple commensal sites of cattle such as skin,

tonsils, gut, coat, teats and genital tract (Zadoks et al. 2003) but is a significant cause of bovine mastitis. *S. uberis* was identified as one of the most common bacterial causes of mastitis from isolates as early as the 1900s and was first acknowledged as a major barrier to reducing the incidences of mastitis by the 1920s (Bramley 1984).

Despite its high prevalence, there is a poor understanding of the epidemiology of *S. uberis* mastitis. It is well documented that persistent and chronic infections are normally attributed to a predominant strain (Zadoks et al. 2003) and after experimental challenge with multiple strains, one tends to dominate within the mammary gland (Pryor et al. 2009). Protocols for ribotyping *S. uberis* strains were first used to sufficiently discriminate between several U.K. strains using *Hin*dIII endonuclease digestion (Hill & Leigh 1989). The first global strain typing of 45 strains (from the U.S.A., Germany and the U.K.) suggested *S. uberis* showed an overall similarity of fragment patterns (Groschup et al. 1991). However these methods are not suitable for analysis of closely related strains and this limitation could account for the conclusion of a homogenous population.

Further technological developments in strain typing by pulse-field gel electrophoresis (PFGE) appeared to produce highly discriminatory results (Baseggio et al. 1997) and was used to identify 330 strains from 343 isolates (Douglas et al. 2000). However PGFE suffers from poor reproducibility (Cookson et al. 1996; Van Belkum et al. 1998; Murchan et al. 2003) and is

unable to quantitate genetic relationships between isolates (Maiden et al. 1998). To overcome these limitations, the use of Multi Locus Sequence Type (MLST) was investigated. MLST relies on comparison of DNA sequence from several different housekeeping genes that are highly conserved due to their requirement for cellular function (Maiden et al. 1998). MLST carried out on *S. uberis* was able to differentiate between strains with the same ribotype (Zadoks 2005; Coffey & Pullinger 2006). This typing method identified 3 major clonal complexes (ST-5, ST-143 and ST-86) from nearly 1000 isolates within the U.K., New Zealand, Australia and Portugal (Pullinger et al. 2006; Rato et al. 2008). All of these clonal complexes have been shown capable of establishing intramammary infections along with the ability to establish commensal and/or environmental colonisation (Pullinger et al. 2006). No particular clonal complex was associated with persistent infections, indicating the ability to infect the bovine mammary gland is intrinsic within the population as a whole. A view that has been further substantiated by the observation that all known virulence determinants appear to reside in the core genome (Hossain et al. 2015)

1.2.2. Host Pathogen Interactions

For successful colonisation of the mammary gland *S. uberis* must gain entry and either adhere to the host tissue or multiply at a sufficient rate to prevent elimination during milking. *S. uberis* has to overcome a number of host defence mechanisms to do so.

1.2.2.1. Physical defence mechanisms against mastitic pathogens

The teat canal has a thick layer of keratin, which provides a physical barrier to bacterial influx; low levels of keratin have been shown to correlate with an increased rate of infection (Lacy-Hulbert & Hillerton 1995). The mesh-like structure of keratin is thought to impede progress of bacteria, any that become attached are flushed out during milking (Williams & Mein 1985). Keratin also has bacteriostatic properties which has been attributed in part to its lipid content (Hogan et al. 1986) and fatty acid $C_{18:3}$ which inhibits growth of *S. uberis* (Hogan et al. 1988).

1.2.2.2. Soluble defence factors found within the bovine mammary gland

Cationic antimicrobial proteins are part of the bactericidal effectors in phagocytic cells, but are expressed by many cell types including epithelial cells (Schonwetter et al. 1995). Their mechanism is only partly understood, however, they have a high membrane affinity, leading to disruption of the membrane phospholipids and resulting in loss of membrane integrity and ultimately lysis of the bacteria cell (Shai 1995). These proteins consist of small peptides with a positive charge and work synergistically with soluble proteins such as lactoferrin and lysozyme (Gudmundsson et al. 1999). Lactoferrin, lactoperoxidase and lysozyme are whey proteins present in the non-casein fraction of milk and non-lactating secretions. These proteins are produced by the cells of the glandular epithelium (Molenaar et al. 1996). Lactoferrin concentrations rise during infection (Harmon et al. 1976) and is, as a single

entity, deemed bacteriostatic due to its ability to chelate iron. However because of their low requirement for iron, streptococci are less sensitive to this defence mechanism (Chaneton et al. 2008). However, lactoferrin also has the ability to bind zinc (Blakeborough et al. 1983), copper (Eaton et al. 1989) and manganese ions (Davidson & Lönnerdal 1989). Of which, manganese has been shown to be an essential trace metal for growth of *S. uberis* (Smith, A. J. et al. 2003).

The role of lactoferrin in the control of *S. uberis* infection has not been investigated thoroughly, although the ability of *S. uberis* to bind lactoferrin to the cell surface (Fang & Oliver 1999; Fang et al. 2000) by SUB0145 (*lpb*), the *S. uberis* specific lactoferrin binding protein, has been described as a putative virulence factor (Almeida et al. 2006). There is some evidence that suggests lactoferrin aids internalisation of *S. uberis* into mammary epithelial cells (Almeida et al. 2006). *S. uberis* has been shown to posses a lactoferrin adhesion molecule (SUAM) which may help *S. uberis* create a bridge between the bacterium, the lactoferrin and a putative lactoferrin receptor on the epithelial cell surface (Patel et al. 2009). By doing so it is postulated *S. uberis* increases its presence and persistence of over multiple lactations (Oliver et al. 1998). However as lactoferrin is a highly charged protein, binding could be independent of a specific binding protein.

Lactoperoxidase is present in high concentration in colostrum and is an effective bactericidal for Gram-negative bacteria and bacteriostatic for Gram-

positive species (Marshall et al. 1986). It is less effective during lactation because lactoperoxidase relies on hydrogen peroxide and thiocynate to produce its antimicrobial compound oxidising hypothiocynate (Fang et al. 1998). These cofactors are not present in high enough concentrations in milk potentially due to catalase activity (Mickelson et al. 1985).

Lysozyme is a glycoside hydrolase, which is able to lyse cells by catalysing hydrolysis of β-linkages between the peptidoglycans in the cell wall (Mckenzie & White Jr. 1991). Lysozyme has been shown to work in conjunction with lactoferrin to lyse some Gram-positive bacteria such as *Staphylococcus aureus.* This is achieved by binding to the lipoteichoic acid surface component (Rainard et al. 2006). The effect of lysozyme on streptococci is limited because certain strains contain peptidoglycan modifications. N-acetylglucosamine deacetylase (*pgdA*) and peptidoglycan O-acetyltransferase (*oatA*) change the structure of the peptidoglycan to infer lysozyme resistance (Vollmer & Tomasz 2000; Davis et al. 2008; Crisóstomo et al. 2006). Whilst Mur ligase creates additional cross linkages of peptidoglycan increasing resistance further (Filipe et al. 2002). Streptococcal species lacking these peptidoglycan modification genes are susceptible to high concentrations of lysozyme (250μg/ml) (Schreur et al. 2012), however, due to its low concentration of 0.12μg/ml in normal milk and 1.7μg/ml during intramammary infection (Carlsson et al. 1989) it is unlikely to have a strong antimicrobial role in mastitis prevention.

1.2.2.3. The activation of the immune response

The stimulation of the immune system relies on the appropriate activation of pathogen recognition receptors (PRR) stimulating the release of effector molecules. The major family of PRRs, found in mammals, are Toll-like receptors (TLRs), which are particularly important for the identification of a distinct set of molecular patterns that are not present in healthy cells (Goldammer et al. 2004). In the case of an invading Gram-positive pathogen, such as *S. uberis*, lipoteichoic acid (LTA) is the microbial ligand present that is recognised by TLR-2 and TLR-4, which enable the initiation of the antimicrobial response (Yoshimura et al. 1999). The recognition of LTA by the TLR ectodomain activates the production of transcriptional factors (O'Neill & Bowie 2007) activator protein (AP-1) and nuclear factor kappa B (NF-κB) (Kawai & Akira 2007). The release of NF-κB induces the expression of cytokines IL-1β, IL-10, IL-12, IFN-γ, TNF-α, CD14, CXCL8 and complement component C5a. The sustained elevation of IL-1β, IL-10, IL-12, IFN-γ and C5a is particular to *S. uberis* infection (Bannerman et al. 2004). These cytokines are responsible for different aspects of the immune response. IL-1β is responsible for activating lymphocytes and initiating the production of IL-6, which alongside TNF-α stimulates energy metabolism that leads to an increase in body temperature. CXCL8 and C5a are responsible for the recruitment of neutrophils. CXCL8 induces the migration of the neutrophils from the blood and the C5a acts as a chemoattractant to the infected cell.

1.2.2.4. Cellular defence mechanisms

As the mammary gland is not a holocrine secretory organ, there are very few immune cells present when healthy. Epithelial cells and macrophages account for the majority of cells present, with an increase of leukocytes in early lactation (Rainard & Iollet 2006). During intra-mammary infection the proportions of each cell type differs with type and severity of the intramammary infection (Kehrli & Shuster 1994), however more than 90% of these cells are neutrophils (Paape et al. 1981). Neutrophils are the first immune cells to arrive at the site of infection in large numbers and their main role is to phagocytose the bacteria. Their part in controlling some mastitic bacteria such as *S. aureus* and *E. coli* is generally accepted (Schalm et al. 1976; Hill et al. 1978). Their role when *S. uberis* is the invading pathogen is questioned (Leigh 1999) as it seems able to resist phagocytosis from neutrophils, potentially due to as yet uncharacterised 'interfering factors' produced by the bacteria (Leigh & Field 1994). Macrophage, activated by IFN-γ, are also phagocytic cells which, alongside the neutrophils, ingest common mastitis pathogens and display antigens on their cell membrane for recognition by lymphocytes (Politis et al. 1992). There are two subsets of lymphocytes, T and B, where T-lymphocytes can be further divided into T-helper and T-suppressor cells. In recognition of the macrophage antigen complex, IL-12 induces differentiation of T-cells into T-helper cells, which produce cytokine CD4, activating the B-lymphocytes. This stimulates the maturation of the B-lymphocytes into antibody secreting plasma cells (Sordillo & Streicher 2002).

1.2.2.5. The humoral defence system

Antibodies, also referred to as Immunoglobulins (Ig), are one of the main constituents of humoral immunity and prevent the spread of intracellular infection by pathogen destruction. There are several isotopes of Ig present in the bovine mammary gland (IgA, IgM, IgG_1, IgG_2) and immunoglobulin G (IgG_1 and IgG_2) is the most abundant (Lascelles & McDowell 1974). IgG_1 and IgG_2 are transported into the mammary gland during neutrophil diapedesis where they bind to the antigen epitope on the bacterial cell surface. The function of this is two fold. Firstly their presence on the cell surface opsonises bacteria to enhance phagocytosis. Secondly IgG blocks the binding of the bacteria to host cell surface receptors, reducing the ability of the bacteria to colonise on the cell surface or inhibiting internalisation into the mammary epithelial cells (Sordillo et al. 1997). The isotype IgA is responsible for binding to bacterial secreted toxins, neutralising their effect on the host (Sordillo & Streicher 2002) and IgM is responsible for the activation of the complement system (Sordillo & Streicher 2002). Complement is a collection of serum-derived proteins which activities are mediated through receptors located on a variety of cells (Rainard 2003). Functions of complement are related to lysis of bacteria, opsonisation and acting as attractants for phagocytes (Korhonen et al. 2000), however, the role of the complement system in the defence of the mammary gland is yet to be fully defined (Sordillo & Streicher 2002).

1.2.2.6. Virulence factors of S. uberis

Virulence factors facilitate bacterial colonisation through various mechanisms, such as invasion of the mammary gland, adherence to epithelial lining and protection against antibacterial factions of the immune system. Research into common streptococcal virulence factors such as capsule formation, plasminogen activators, sortase anchored proteins and adhesion molecules has identified determinants essential for pathogenesis. During infection of the lactating gland, *S. uberis* is found predominantly in the lumen of the secretory alveoli and ductular tissue, indicating a large percentage of growth occurs in newly synthesised milk (Thomas et al. 1994). This environment lacks free amino acids and as *S. uberis* is known to be highly auxotrophic for particular amino acids, it was thought that it must be able to acquire essential nutrients via an alternative method (Leigh et al. 1999).

The caseinolytic enzyme, plasmin, is thought to facilitate the colonisation of bacteria in the bovine mammary gland (Leigh 1993). Plasminogen activators play a major role in the generation of plasmin from plasminogen (Leigh & Lincoln 1997). Some streptococci have the ability to produce plasminogen activators called streptokinase and *S. uberis* produces the alternative plasminogen activator, PauA (Rosey et al. 1999). It is known *S. uberis* can bind plasmin to its surface and it is postulated that this may allow for the production of nutritionally beneficial peptides in close proximity to the cell (Ward & Leigh 2004). A vaccination study conducted using a cell culture filtrate containing PauA antigen, inferred some protection against

intramammary infection (Leigh et al. 1999). However, a challenge study later showed that *S .uberis* unable to produce PauA, due to an inactivation of the gene, was able to colonise a mammary gland, in a manner similar to the wild-type. Indicating PauA has little or no effect on growth in milk or intramammary infection (Ward et al. 2003). An additional plasminogen activator was identified in *S. uberis* strain SK880 (Johnsen et al. 1999). Analysis of the DNA sequence in this locus found that this strain lacked the *PauA* open reading frame and encoded an alternative plasminogen activator, PauB (Ward & Leigh 2002). A genetic screen of a panel of isolates from U.K. and Danish farms revealed that PauB was only found in the SK880 strain and represents a novel but rare plasminogen activator (Ward & Leigh 2004).

Hyaluronic acid capsule has been shown to play an important role in virulence of group A streptococci during severe infections (Wessels et al. 1994). It was postulated that a capsular strain of *S. uberis* would therefore have a higher resistance to opsonisation and phagocytosis (Leigh et al. 1994). *In vitro*, bacteria were grown in skimmed milk in the presence of peripheral blood neutrophils. The ability of *S. uberis* to resist phagocytosis was considerably reduced when the capsule was removed by hyaluronidase treatment (Leigh & Field 1994). Disruption of the capsule genes *hasABC* showed that *S. uberis* lost its ability to resist neutrophils *in vitro* (Ward et al. 2001). An experimental challenge study investigated the capsules importance within the host using non-capsular mutant strains. However, levels of infection were shown to be indistinguishable between capsular and non-capsular strains indicating

virulence was independent of the capsule (Field et al. 2003). The occurrence of strains possessing a hyaluronic acid capsule is at a higher frequency in isolates associated with mastitis when compared to those that are not (Field et al. 2003). This suggests that the capsule may have more of an effect on the ability of the bacteria to penetrate the mammary gland rather than influencing its ability to cause infection (Field et al. 2003)

Hyaluronidase, a protein secreted by some strains of *S. uberis*, has been suggested to facilitate the bacterial dissemination into mucosal barriers aiding spread of infection (Oliver et al. 1998). It is capable of degrading the hyaluronic component of connective tissue facilitating the spread of the bacteria through tissues (Hynes & Walton 2000). Together with hyaluronic acid capsule, hyaluronidase influences the proliferation of mammary epithelial cells (Matthews et al. 1994). Which could have important consequences during the preparturient period, when mammary tissue undergoes rapid differentiation, enabling the spread of the bacteria from the initial site of infection. Hyaluronidase is encoded by either the *hylA* or phage assisted *hylP* coding sequence, neither of which are present in the reference strain (0140J, Accession number AM946015) (Ward et al. 2009), suggesting this alone is not responsible for its increased virulence.

MtuA is a lipoprotein receptor antigen, responsible for the acquisition of manganese, has been shown to be essential for growth in milk and infection of the bovine mammary gland (Smith, A. J. et al. 2003). Its position within the

bacterial cell membrane, on the outside surface of the cytoplasmic membrane but beneath the cell wall, means it is not a plausible target for growth inhibiting antibodies (Jones et al. 2004). A further two lipoproteins were predicted within the *S. uberis* genome, OppA1 and OppA2, which are able to import amino acids produced by the hydrolysis of casein by PauA (Smith et al. 2002). The expression of these oligopeptide binding proteins is shown to be up-regulated when *S. uberis* is grown in milk and mutant strains lacking *oppA1* or *oppA2* were shown to display a lower growth rate in milk (Taylor et al. 2003).

Surface proteins are frequently required for virulence and sortase (SrtA) is responsible for binding many such proteins to the cell wall. Investigation into the presence of SrtA in *S. uberis* and whether it, or associated proteins, played a role in the ability to cause mastitis, led to the identification of ten potential cell wall anchored proteins (Egan et al. 2010). A mammary challenge study showed that the presence of functional SrtA and individual SrtA anchored proteins; lactoferrin binding protein (Lbp), collagen-like protein and C5a peptidase precursor played a significant role in virulence (Leigh et al. 2010). Recently, Egan et al. 2012 also demonstrated that a mutation in *vru*; encoding a regulatory protein, showed reduced ability to colonise the bovine mammary gland and induce clinical signs of mastitis. Analysis of bacterial gene expression indicated that Vru regulates a number of known or suspected virulence determinants, including *pauA* (Ward et al. 2003), *lbp* (Almeida et al. 2006), Collagen-like protein (Leigh et al. 2010) and the capsule forming genes

hasA and *hasB* (Wessels et al. 1994). Therefore there is a need to widen the investigative approach to look for numerous genes or proteins that work in a synergistic manner to cause disease rather than on an individual basis.

1.3. Essential genome

The concept of an "essential genome" assumes there is a minimum number of genes that bacteria require for survival and that the remainder can be disrupted or removed without obvious deleterious effects (Mushegian & Koonin 1996). Therefore, the essential genes must encode proteins that maintain central metabolism and cellular structure, replication, translation and the uptake of necessary nutrients and export of any potential toxic metabolites.

Creating an essential genome map is often achieved operationally, relying heavily on disruptive mutagenesis and the ability to randomly create a genetic lesion within the bacterial genome. The method of targeted single gene deletions was deemed the "gold standard" for determining gene essentiality. Any intragenic insertion would disrupt gene function; if this gene were non-essential the bacteria would be able to proliferate, however if the insertion was within an essential gene the bacteria would not survive.

The system of transposition, for disruptive mutagenesis, has been a useful tool in bacterial genetic evaluation, helping to ascertain gene function and

regulatory factors (Maguin et al. 1996). Transposons are mobile genetic elements that translocate within bacterial DNA using a form of recombination (Mizuuchi 1992). They are found in all organisms and create an important source of genetic variation, such as nucleotide substitutions that may result in improved protein function (Bogarad & Deem 1999) or acquired DNA repair systems increasing genetic stability (Arber 2000). All transposition relies on a common reaction that breaks the target DNA and joins it to the donor strands, which is catalysed by transposases (Haapa-Paananen et al. 2002).

1.3.1. Transposon mutagenesis

There are many transposon systems used in bacteriology, but the four most frequently used are; the transposable resistance elements (Tn), Bacteriophage (Mu), *mariner* transposon derivatives and Insertional Sequences (IS) (Shapiro 1979). There are differences in the behaviour of theses elements, such as frequency of insertion, specificity of insertion site and utilisation of different delivery systems but they are all capable of insertion of foreign DNA segments into the host DNA of various species.

In general, transposons all have a similar basic structure; they are flanked by the presence of short, identical DNA sequences, termed inverted repeats (IR). The function of the IR regions is to form a binding site for the transposase enzyme that activates the transposition (Reznikoff & Winterberg 2008).

1.3.1.1. Transposition using the Tn family of composite insertion elements

Tn transposition relies on the non-replicative cut and paste mechanism. The transposase protein binds to the donor DNA and cuts at the site of integration and host DNA repair proteins integrate the two strands (Bender & Kleckner 1986). There are multiple variations of Tn, which are identified by a numeral suffix. The Tn family of transposons have limited use as a random mutagen in some bacterial species due to insertion site specificity, for example the Tn5 transposon always inserts at 5'-NGCTANAGCN-3' (Matsutani 1991). It can also suffer from low transposition frequency in many Gram-positive strains (Chopin & Chopin 1989), meaning the transposon does not manifest in a sufficiently high frequency to generate saturated mutagenesis.

1.3.1.2. Mu bacteriophage transposition

An alternative method of insertional mutagenesis that is more applicable for Gram-positive bacteria is the bacteriophage Mu, which is the most extensively studied replicative transpositional mutagen (Haapa-Paananen et al. 2002); its use has been documented in *S. aureus, S. pyogenes* and *S. suis* (Pajunen et al. 2005). The insertional element is replicated during integration through an RNA intermediate which rapidly increase the copy number present allowing for higher transposition frequency (Mizuuchi 1992). The Mu bacteriophage has a strong insertional bias to a central CGG sequence in a five base region (5'-NCGGN-3') (Manna et al. 2005; Green et al. 2012). As discussed previously insertional recognition sites are not desirable for mutagenesis systems when

investigating gene essentiality. This is because some coding sequences may be missed due to their complexity or lack of a target sequence (Green et al. 2012).

1.3.1.3. Eukaryotic origin Mariner transposition

Originally identified in drosophila (Jacobson et al. 1986), *mariner* transposons have been more extensively used in prokaryotes (Lampe 2010). *Mariner* Transposons have also been used in Gram-positive bacteria to assay essential genes (Zhang et al. 2012). As with the Tn and Mu, the mariner system suffers from pseudo-random insertion bias. It has a TA recognition site with a higher preference to palindromic TA repeats (5'-ATATATAT-3'). Both of which are superseded by bendable DNA configurations (Vigdal et al. 2002), which are morphological DNAse binding structures for optimal strand transfer (Brukner et al. 1995). Their method of integration is a cut and paste mechanism (Hartl 2001), relying on a double stranded scission, allowing the entire element to integrate into a staggered cut within a TA nucleotide recognition site (Muñoz-López et al. 2010).

1.3.1.4. Insertional mutagenesis using IS insertional elements

Bacterial Insertional sequences (*IS*) represent the simplest class of transposable elements. They have the ability to rearrange genetic organisation and effect expression and regulation of genes with the presence of translation and transcription termination signals (Romero & Klaenhammer 1993). *IS*

elements are formed by small segments of bacterial DNA flanked by inverted repeats and encode their own transposition functions (Griffiths et al. 2000). They were first identified in *Escherichia coli* within a galactose metabolism operon (Irani et al. 1983) and have been used in numerous Lactococcus species (Renault & Heslot 1987; Polzin & Shimizu-Kadota 1987; Haandrikman et al. 1990; Cluzel et al. 1991; Gasson & Fitzgerald 1994). *IS* elements have been shown to integrate randomly within coccus chromosomes (Romero & Klaenhammer 1993) but are limited to strains with high transformation frequencies (Maguin et al. 1996).

To broaden the application of *IS* elements to other coccus species, Maguin et al. (1996) used a pG$^+$host plasmid, which carries the *IS* insertion element and acts as a temperature sensitive delivery vector. The pGhost9::IS*S1* mutagenesis system has been used extensively in determining many factors of virulence in numerous coccus bacteria (Spellerberg et al. 1999; Maguin et al. 1996; Biswas & Biswas 2011; Baureder & Hederstedt 2012). Insertion of IS*S1* and its flanking DNA (pGhost9) undergoes replicative transposition leading to an integrated plasmid sequence between duplicated IS elements (Maguin et al. 1996). The benefit of using the pG$^+$host:IS*S1* system is that if a desired mutant is identified, the bacteria can be cured (plasmid sequence removed) leaving the disruptive IS element within the gene thus removing the antibiotic resistance marker (Maguin et al. 1996). There has been some concerns about its application for some streptococcal species because of potential presence of endogenous IS*S1* elements that can recombine and create chromosomal

deletions (Thibessard et al. 2002). However, the pGhost9::IS*S1* vector system has already been found to provide efficient, high frequency random mutagenesis in a wide range of streptococci including *S. uberis* (Ward et al. 2001) and has been used successfully to assess individual mutants identifying virulence factors (Egan et al. 2010; Leigh et al. 2010).

1.3.2. High-throughput mutation mapping techniques

Deletion and analysis of individual genes is a time consuming process. Identification of numerous insertions simultaneously, termed high-throughput mutation mapping, initially used DNA hybridisation to track mutants through a mouse infection model (Hensel et al. 1995). However, this method was soon superseded by the use of a microarray. These methods relied on signal intensity before and after a defined selective process. There were three methods developed almost simultaneously; TraSH (Sassetti et al. 2001), MATT (Salama et al. 2004) and DeADMAn (Lamichhane et al. 2005). There are some issues related to the use of microarrays for this purpose, as probes can give false readings which can lead to inaccurate or missed readings (Wheelan et al. 2006). There are also issues with spatial artefacts (bad resolution of neighbour fragments), mis-hybridisation, difficulty in detecting low signal and probe saturation (Chaudhuri, Allen, et al. 2009).

With technological advances in DNA sequencing, a more accurate method to identify insertion positions was achieved compared to microarray analysis. The use of random insertional mutagenesis coupled with high throughput

sequencing has enabled identification of conditionally essential genes for many pathogenic bacteria. Interestingly the first method developed predated the microarray approach (Hutchison III et al. 1999) but the limitation of sequencing technologies, at that time, hampered its widespread adoption. Various approaches have since been developed to achieve transposon-insertion sequencing (TIS); Tn-Seq (van Opijnen et al. 2009), INSeq (Goodman et al. 2009), HITS (Gawronski et al. 2009) and TraDIS (Langridge et al. 2009), each of which follow a series of steps; transposon mutagenesis, DNA extraction and fragmentation and enrichment of transposon-chromosome junctions (Figure 1.2.).

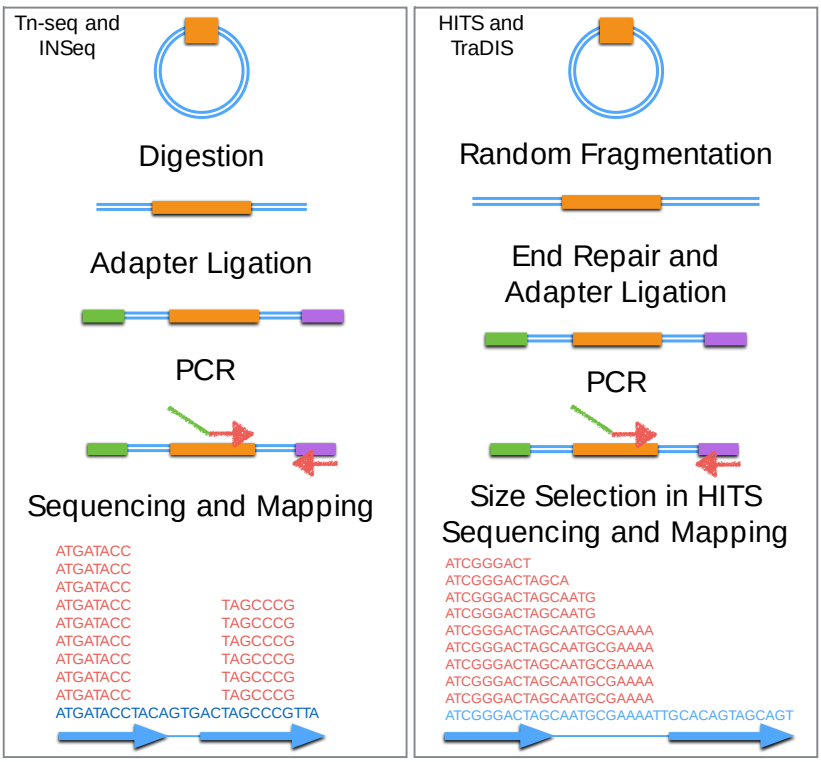

Figure 1.2. Illustrative methodology of the four main transposon insertion sequencing protocols. Each of the protocols starts with pooled genomic DNA. a) Tn-Seq and INSeq have very similar method with the differences based on purification. Tn-Seq uses PAGE gel separation and INSeq uses agarose. b) TraDIS and HITS are more similar with the deviation occurring when the PCR products in HITS are purified by streptavidin capture whereas TraDIS samples are sequenced directly.

Extracted DNA is fragmented either by endonuclease digestion (TN-Seq and INSeq) or mechanical shearing (HITS and TraDIS). The fragmented DNA is then enriched to amplify the presence of the desired fragments, containing the transposon insertion site. The enriched fragments are then prepared for sequencing by the ligation of sequencing adaptors. Apart from the variable transposon used in each experimental technique, the differences mainly occur during the transposon-chromosomal junction enrichment. These differences mean each protocol is unique and have their own benefits for certain applications.

Some protocols add an additional purification step. The Tn-Seq Circle method follows the same library preparation as Tn-Seq and INSeq up to the point of sequencing. Here samples are digested with a restriction endonuclease directed at one end of the transposon. The single stranded DNA is then circularised by template-dependant ligation. This method joins two ends of a DNA fragment at a specific DNA sequence, directed at the digested transposon region (Gallagher et al. 2011). Fragments which have not been circularised are digested with an exonuclease; removing those which do not contain the transposon junctions (Gallagher et al. 2011). Although this makes the protocol more laborious the modifications increase the sensitivity of the technique. High throughput insertion tracking (HITS) uses an approach by which a biotinylated primer, that is complementary to the 3’ end of the insertion, is incorporated during the PCR enrichment stage (Gawronski et al. 2009). A methodology modification was also made to the IN-Seq protocol, with the

introduction of a PCR with biotinylated primers (Goodman et al. 2011). This step reduces the amount of enzymes used during the protocol, but the subsequent magnetic streptavidin purification requires a large amount of DNA to be used prior to the affinity purification stage to allow for sufficient DNA to be available for sequencing.

There are some limitations associated with TIS techniques. Tn-Seq and INSeq are limited to pathogens that are mutable using the mariner transposon system. The protocol is based around the use of an *Mme*I endonuclease site, which is introduced into the insertion sequence. This allows for the digestion of the chromosomal DNA 20bp downstream of the transposon site. As previously discussed, there is an issue relating to how the mariner system is biased during insertion. Also the recognition site is 4bp before the end of the insertion terminal sequence meaning only 16bp of chromosomal DNA is left to align to the reference genome, which in some cases may not be sufficient to align to a reference genome unambiguously.

The most recent method, TraDIS (Transposon Directed Insertion-site Sequencing), has taken a new approach to the development of the sample library generation. Rather than relying on the sequencer to generate the desired reads within a large pool of data, modifications to the sequencing parameters increase the production of desired fragments (Langridge et al. 2009). A PCR using custom primers, chimeras of the Illumina sequence adaptors and a transposon specific sequence, enabled production of desired fragments that incorporated a target sequence. Transposon specific primers

were used in a custom-sequencing protocol so only the target sequence is generated (Langridge et al. 2009). In this case the innovation of TraDIS is also its greatest limitation. The requirement for the custom-sequencing protocol means core parameters of the sequencer need to be altered in order for the protocol to function correctly and the use of custom primers mean costly reagents are often wasted in optimisation. This may not be available to those without direct access to the sequencing machine, as commercial sequencing services may be unwilling to alter settings due to quality assurance restrictions, limiting the scope of the TraDIS protocol for most researchers.

1.3.3. Sequencing of DNA

DNA sequencing was dominated for over 20 years by the methods developed by Fredrick Sanger in the late 1970s. Dideoxynucleoitde or Sanger sequencing uses a complement of the four deoxynucleotides (dNTPs), target primers and a population of amplified DNA molecules along with a single dideoxynucleotide (ddTNP). When a polymerase is added, the reaction mixture undergoes DNA extension before being terminated by the ddNTP (Sanger & Coulson 1975; Sanger 1977; Maxam & Gilbert 1977). The reaction is replicated with each ddNTP species and the resulting products are resolved on an acrylamide gel. Each band then corresponds to a different DNA base. Progressive advancements lead to fluorescence-labelled ddNTPs (Smith et al. 1986; Smith & Fung 1985) and capillary electrophoresis (Jorgenson & DeArman Lukacs 1981; Kasper et al. 1988), allowing for single sequencing reactions, reducing the cost and improving the throughput (Karger & Guttman

2009). Large-scale sequencing projects initiated the demand for a technology that would reduce the cost and time spent generating genomic data. In 2005 Solexa (later being incorporated with Illumina) introduced the Genome Analyser, which being similar in principle to Sanger sequencing, used dye-terminator technology. This enabled hundreds of thousands of simultaneous sequencing reactions (Fedurco et al. 2006) generating more data, reducing costs even further (Bentley et al. 2008) and provided the opportunity to combine essential genome experiments with affordable high-throughput sequencing (Loman et al. 2012).

The Illumina MiSeq's chemistry uses reverse-terminator sequencing by synthesis (Figure 1.3). Illumina sequencing relies on "clusters" and bridge amplification and during library preparation; "adaptors" are added to the end of the fragmented DNA template. The 5' adaptor allows the fragments to attach to the flowcell using surface bound oligonucleotides. A complementary strand of DNA is generated using dNTPs and the original strand is removed by denaturation. The adapter of the 3' end of the fragment is then annealing to another surface bound oligonucleotide creating a "bridge", which is the new site for strand synthesis. This reaction is repeated through multiple cycles of annealing, extension and denaturation, termed cluster generation. The clusters are then cleaved at a single site in one of the adapters as to ensure all strands are identically orientated. The strand for sequencing is synthesised using reversible terminators with removable fluorescent moieties and a special DNA polymerase capable of incorporating them into growing chains of

oligonucleotides (Morozova & Marra 2008). After each cycle of incorporation the identity of the inserted base is determined by laser-induced excitation and imaged (Bentley et al. 2008). Compared to traditional Sanger sequencing the automated base calling on the MiSeq allows for high accuracy reads (0.0064% error) and high quality scores (overall 70% >Q30) using the V3 reagents at 300bp reads (Schrimer et al. 2015). The Illumina chemistry does not have the limitations associated with other sequencers when homopolymeric regions are present. This is because nucleotides are incorporated in clusters of four bases, increasing resolution (Bentley et al. 2008). The main problem highlighted with the MiSeq technology relates to low complexity libraries. During cluster generation, the MiSeq conducts an initial quality control phase to ensure no errors have occurred which is achieved by the camera focusing on the DNA strands. If the camera cannot focus, the clusters are deemed too similar and the run will be aborted. Within a bacterial genome and specifically TIS experiments, such as TraDIS, many of the strands will be identical due to the presence of the same primer sequence. To overcome this technical issue, it is recommended that a “spike” control of irrelevant DNA (normally phage PhiX) be used to offset this issue. This variation allows the camera to focus on individual clusters and therefore complete the sequencing run. The addition of the control DNA however, then accounts for 5-10% of the overall sequencing reads, resulting in a proportion of less useful data.

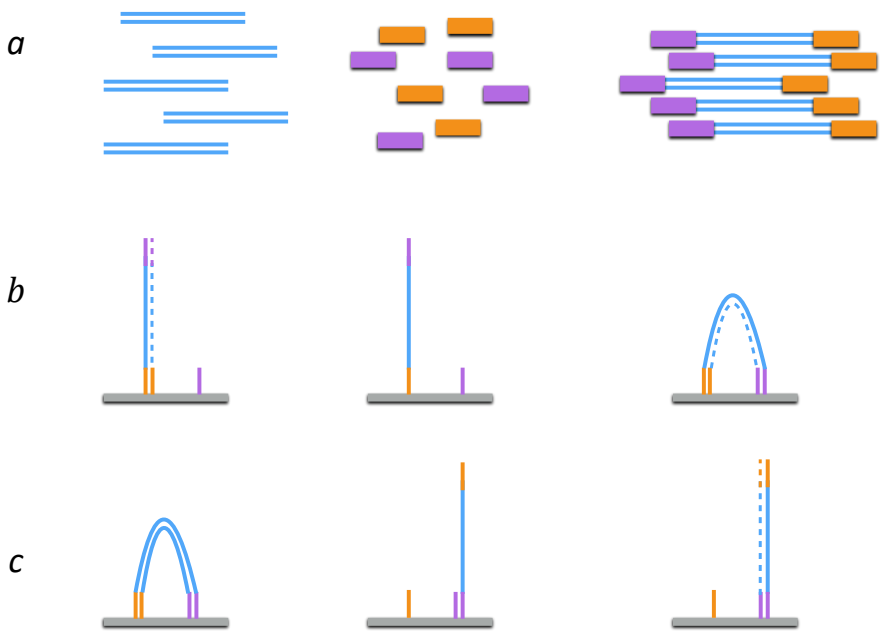

Figure 1.3. Representation of the stages involved in Reverse-terminator sequencing by synthesis. *a*, DNA is fragmented and ligated to oligonucleotides resulting in double stranded, blunt ended fragments. *b*, The denatured double stranded DNA is annealed to complementary oligonucleotides attached to the flowcell. A complementary strand is synthesised using dNTPs, until completion of the strand. The complementary strand is then annealed to another surface bound oligonucleotide forming a bridge. *c*, The original strand is removed by denaturation and the process is repeated for the second strand. Multiple cycles of annealing, extension, and denaturation in varying isothermal conditions generate clusters of identical surface bound strands of DNA. The strand of DNA for sequencing is formed using fluorescent dNTPs in groups. Laser excitation and image capture decipher the florescent discharge from the dNTPs into base calls, building a sequence read for each cluster. (Bentley et al. 2008).

1.5. Aims and Objectives

The aim of this project was to use an existing bank of insertional mutants of *S. uberis* 0140J (AM946015) (Ward et al. 2001) each containing a single copy of pGhost9::IS*S1* integrated into the genome to:

- Develop a protocol capable of targeted enrichment of chromosomal DNA flanking the point of an insertion

- Develop a bioinformatic pipeline to analyse high-throughput sequence data, enabling the identification of the chromosomal region flanking the insertion and map the read to the reference DNA sequence

- Conduct various *in vitro* analyses related to bacterial growth in host specific environments relating to the bovine mammary gland

- Use the afore mentioned bioinformatics pipeline to compare the coding sequences lacking insertions, to determine those with a defined role in various environments

Chapter 2 Materials and Methods

2.1. Generation of bacterial mutant Pools

A culture of *S. uberis* 0140J that had been mutagenised with the thermosensitive plasmid containing the insertion sequence element *S1* (pGhost9:IS*S1*) and stored at -80°C was used throughout this study. The viability and frequency of pGhost9:IS*S1* insertions within the culture were assessed by serial dilution plate counts on Todd-Hewitt agar (THA) in the presence and absence of a selective antibiotic erythromycin (Ery; 1µg/ml; Sigma Aldrich UK). Total counts in the presence/absence of Ery were used to calculate the total number and the proportion of mutant bacteria within the culture. Subsequently, a sample of the mutagenised culture, diluted appropriately, was plated and grown to single colonies on THA containing Ery (1µg/ml). Ten pools each containing approximately 10^4 colonies were scraped from plates into phosphate buffered saline (PBS) and resulting bacterial suspension collected by centrifugation (8000 x *g*, for 10 minutes) washed (3 times) in PBS, and finally suspended in pyrogen free saline (Sigma Aldrich U.K.) containing 50% (v/v) glycerol. These pools were stored in aliquots and frozen at -80°C.

2.2. Extraction of chromosomal DNA from *S. uberis* mutant pools

Samples containing approximately $8x10^8$ CFU (range $3.9x10^8$ to $1.6x10^9$) were centrifuged (12 000*g*, 5 minutes) and the supernatant the cell pellet was resuspended in 100µl TE-buffer (1M Tris-HCl, 0.5M EDTA). After centrifugation (12 000*g*, 5 minutes) the cells were then resuspended in 375µl of cell wall disruption buffer (1M Tris-HCl, 0.5M EDTA, 30U/ml Mutanolysin and 10mg/ml Lysozyme) and incubated for 30 minutes at 37°C. To lyse cells, 20µl of SDS (20% w/v in 50mM Tris and mM EDTA) was added along with 20mg/ml Proteinase K and the sample incubated for a further hour at 37°C. To precipitate the cell debris, 200µl Saturated NaCl (approximately 6M) was added and the sample agitated before being centrifuged (12 500*g*, 10 minutes). The supernatant was removed to a fresh tube (400µl) and an equal volume of the lower phase of Tris-equilibrated phenol:chloroform:isoamyl alcohol was added, inverted and centrifuged (12 000*g*, 5 minutes). The upper aqueous phase was retained (300µl) and DNA was precipitated by the addition of two volumes of absolute ethanol and incubated at 4°C for an hour. After centrifugation (12 000*g*, 5 minutes) the absolute ethanol was replaced with 70% ethanol. The final DNA sample was obtained by centrifugation (12 000*g*, 5 minutes), and following removal of the 70% ethanol was allowed to air dry at ambient temperature before being suspended in TE buffer containing 20µg/ml RNAse A. Typically samples were stored at 4°C overnight prior to being incubated at 37°C for 30 minutes, and left to cool back to room temperature.

2.3. Quantification of DNA Concentration

DNA was quantified using the Qubit dsDNA Broad Range Fluorometric Assay kit (Life Sciences U.K.), by the addition of 2µl of DNA to 198µl of Qubit buffer. The DNA was incubated at room temperature for 5 minutes before being analysed.

2.4. Digestion of DNA using restriction endonuclease

Restriction digests of DNA using *Hin*dIII and *Eco*R1 were performed by the addition of 10 units of restriction enzyme to a total reaction volume of 50µl using 1µg of DNA and incubated for 1 hour at 37°C. The reaction was then heat inactivated at 80°C for 20 minutes. The digested DNA was purified using PCR cleanup kit (Machery and Nagel, U.S.A.) and eluted using 30µl of pre-heated (70°C) elution buffer.

2.5. Fragmentation of DNA using Adaptive-Focused-Acoustics

Approximately 2µg of genomic DNA was suspended in 200µl TE buffer and fragmented to an average size of 3kb using Covaris Adaptive Focused Acoustics (Covaris Inc., USA) using a Peak incident Power of 5W, 20% Duty Factor, 1000 cycles per burst for 600 seconds at 20°C.

2.6. Preparation of DNA for inverse PCR reactions

The fragmented DNA was purified using Agencourt SPRI beads (Beckman and Coulter, UK) by adding 1.8x volume of beads to the DNA, mixed by pipetting and allowed to incubate at room temperature for 5 minutes. The beads containing DNA were separated from the supernatant using a magnetic stand and the supernatant aspirated. Beads were washed twice with high purity ethanol (80% v/v), and the DNA was eluted into 95µl of molecular biology grade water (Fisher Scientific, UK). Fragmented DNA was blunt end repaired using the NEBNext End Repair module (New England Biolabs Inc., U.S.A.), using 10µl buffer and 5µl enzyme mix and being incubated at 20°C for 30 minutes. End repaired DNA was purified with 1.8x SPRI beads (as previously described) and resuspended in 50µl molecular biology grade water before being quantified using Qubit BR Broad Range Fluorometric Assay kit (Life Sciences U.K.). The end-repaired DNA (1 µg) was suspended in 750µl ligase buffer in the presence of 1000U T4 ligase (New England Biolabs Inc. U.S.A.); and incubated at 22°C overnight. The DNA was purified and concentrated using a PCR cleanup kit (Machery and Nagel, U.S.A.) and eluted using 30µl of pre-heated (70°C) elution buffer.

2.7. Amplification of flanking insertion regions using Inverse PCR

An inverse PCR was conducted to enrich the sequence flanking the IS*S1* element. In a 50µl reaction volume, 100ng of re-circularised DNA was used as template with 10µl Phusion HF buffer (New England Biolabs Inc., USA), 200µM dNTPs and 0.5µM of each primer (P064 5'-CCAACAGCGACAATAATCACATC-3' and P082 5'-AGAACCGAAGAATTCGAACGCTC-3'). The reaction was incubated for 5 minutes at 98°C before the addition of 1U of Phusion High fidelity DNA polymerase (New England Biolabs Inc., USA) to initiate the reaction followed by a further 2 minutes at 98°C then 35 cycles of 98°C for 10 seconds, 63°C for 30 seconds, 72°C for 1 minute with a final extension of 8 minutes at 72°C) on a Applied Bioscience GeneAmp 2700 thermocycler. The amplified PCR product was purified using 1.8x Agencourt SPRI beads (Beckman and Coulter, UK) as previously described and suspended in 50µl of Molecular biology water (Fisher Scientific, UK).

2.8. Bioanalyser Quantification of DNA Fragment Distribution

Agilent reagents were equilibrated to room temperature for 1 hour and thoroughly resuspended using a vortex mixer. Dye concentrate and gel matrix were mixed and centrifuged for 10 minutes at 15,000*g.* The gel mix was added to the microfluidics chip and distributed using the priming station. DNA sample, marker and ladder DNA sample were added to the associated wells and vortexed for 60 seconds at 2400*rpm.*

2.9. Next Generation Sequencing

The purified PCR products were fragmented to 600-700bp using Covaris Adaptive Focused Acoustics (Covaris Inc. USA) using a Peak incident Power of 50W, 20% Duty Factor, 200 cycles per burst for 45 seconds at 20°C. The samples were then prepared for sequencing using the Illumina TruSeq library preparation kit (Illumina Inc., U.S.A.) by DeepSeq (University of Nottingham) and sequenced on the Illumina MiSeq platform.

Chapter 3 Analysis of the enriched insertion and chromosomal DNA junction

This chapter is an expansion of a previously published article "Transposon insertion mapping with PIMMS – Pragmatic Insertional Mutation Mapping System. Blanchard, A.M. et al (2015) Frontiers in Genetics." (Appendix 2)

3.1. Introduction

The combination of random mutagenesis and transposon insertion sequencing (TIS) is a powerful tool in determining an essential genome. Its application to *S. uberis* was achieved by transformation with the pGhost9::IS*S1* transposition system (Maguin et al. 1996). The initial protocol allowed for individual identification of a single mutant using enzymatically digested DNA fragments. The choice of endonuclease was based on the presence of restriction sites within the plasmid and their comparable frequency within the *S. uberis* genome (Ward et al. 2001) (Figure 3.1.). These fragments were then transformed into an *E. coli* host strain for amplification. The DNA was extracted from the host strain and digested again using a *Hin*dIII or *Eco*R1 endonucleases, which have restriction sites within the pGhost9 replicon (Figure 3.1.). Digested fragments were then circularised by ligation and amplified using restriction site and insertion sequence primers (Maguin et al. 1996). The PCR product was then sequenced using Sanger sequencing and

homologs detected using BLAST (Altschul et al. 1990). This method has investigated, through the exploitation of single isogenic mutants, roles of genes thought to be important for the pathogenesis of *S. uberis,* including three sortase (SrtA) surface anchored proteins; Sub1154, Sub1095 and Sub0145 (Leigh et al. 2010), the manganese ABC-transporter metal binding protein, MtuA (Smith, A. J. et al. 2003) and the stand alone regulator, Vru (Egan et al. 2012). Conversely, the investigation of individual mutants lacking the ability to produce hyaluronate synthase, HasA (responsible for capsule biosynthesis) or the plasminogen activator, PauA (Ward et al. 2001; Ward et al. 2003) did not show any role for these molecules in pathogenesis. This method of individually identifying insertion positions within a bank of mutants is a laborious task. Also to assay mutants within all genes of a pathogen is impractical, limiting the scope of this methodology (Hamer et al. 2001).

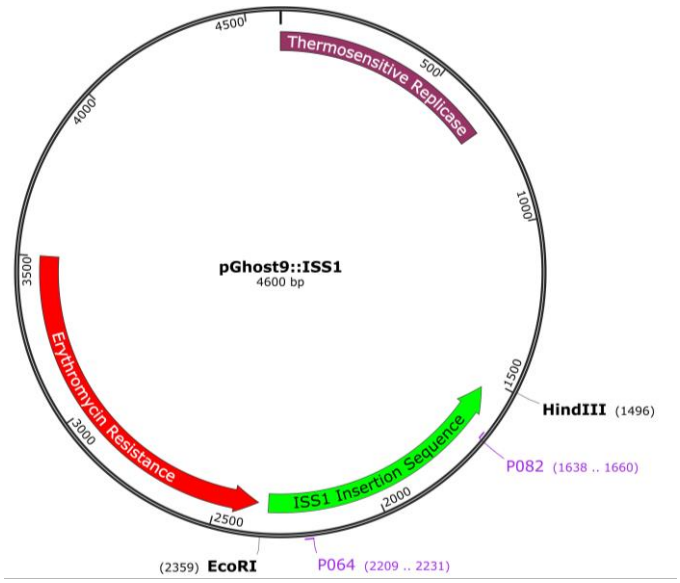

Figure 3.1. pGhost9::IS*S1* plasmid schematic representation. The circular sequence map shows the locations of thermosensitive replicase, antibiotic resistance cassette, endonuclease restriction sites, the insertion element and corresponding primer binding regions for the inverse PCR reaction.

Advances in sequencing technology have allowed routine bacterial genome sequencing and annotation. This in turn led to the development of protocols, enabling many transposon mutants to be assayed simultaneously. These high-throughput methods (discussed in Chapter 1) require a series of complex molecular biological steps and some specialised sequencing (or sequence detection) strategies in conjunction with complex bioinformatic analysis to manage the vast amounts of data produced from a single experiment.

Data analysis is an important part of TIS experiments. The current methods; TN-Seq (van Opijnen et al. 2009), INSeq (Goodman et al. 2009), HITS (Gawronski et al. 2009) and TraDIS (Langridge et al. 2009) each offer specific protocols for data analysis, however all are 'counting' methods, in which the number of the sequence reads aligned within a coding sequence (CDS) correlates to its contributed fitness. The greater the number of insertions or reads within the CDS, the lesser the essentiality to the bacterium, conversely if a CDS is void of insertions, this can be assumed to be of importance for growth and proliferation under the tested conditions. Most published TIS experiments do not identify, in depth, how the analysis is achieved and protocols are often developed in-house and not made publically available. In many cases the sequence reads of interest produced by TIS experiments consist of chromosomal and plasmid or transposon DNA, standard mapping protocols to align reads directly to a reference genome will not be successful. To overcome this limitation published methods use an array of software packages to separate the read into its component parts, which is mainly

achieved using custom computational tools only suited to an experienced bioinformatician. The issues related to this are two fold. To enable the analysis of the data, the user requires prior knowledge of available software. Also due to the vast variability of this type of software, an understanding of the capabilities and limitations are also needed to enable justification for their use. Some software packages have been developed with the ability to analyse insertional sequence reads. The first, specifically for TIS experiments, was ESSENTIALS. This allows for automated filtering of raw sequence reads, identifying transposon regions and separating the chromosomal DNA. This is achieved using the barcode removal tool of FASTX toolkit (Hannon 2009), with the transposon sequence as its target. The filtered reads are then aligned using PASS (Campagna et al. 2009), which uses a method similar to BLAST (Altschul et al. 1990) allowing large databases to be searched quickly. Utilising a genome index of 12 base pair segments, a separate index is created from the reads, the two are compared, and insertion counts are determined. The reads are filtered further to remove reads with multiple alignments within a genome (Zomer et al. 2012) and any statistical analysis is completed using the freely available R package, edgeR (Robinson et al. 2010).

As a purely web-based application, ESSENTIALS does not require any performance computer hardware. However this style of interface is limited by network upload speeds and more importantly, security of sensitive data sets. Raw data has to be hosted on a server, which is then uploaded to the ESSENTIALS server by file transfer protocol (FTP) to complete the analysis. This

situation may not be suitable for all users, especially those analysing commercially sensitive data. There are security implications involved with FTP, as it is not a secure method for data transfer, but more importantly the data is stored on unknown servers. The sites disclaimer states data is secure and only held for 3 weeks, but is not accompanied by any confidentiality agreement. Sponsors may not allow for this dissemination of data, which then limits the scope of the packages used. There is a standalone package available for local installations, at the discretion of the developer, however it requires modifications for specific computer specifications, which may be beyond the abilities of a novice user. There is also the need to generate a tab separated value (tsv) configuration file containing the network address of the input files, transposon sequence and barcode sequence. Sequence data for each end of the transposon and each barcoded sample also needs to be included, which can be time consuming to generate.

More recently, a statistical toolkit has been released for analysis of essential genes called ARTIST (Pritchard et al. 2014), which is based on the use of custom scripts for MATLAB. The aim of ARTIST was to overcome the redundancy of normalisation in TIS experiments, which is achieved by simulation based resampling. A control dataset is generated using the non-parametric Mann-Whitney U test, providing an estimate of how the dataset can be biased during the laboratory-based experiments. This data is then used to train a Hidden Markov Model that identifies regions of the reference genome that have a different number of insertions to the simulated control.

ARTIST has a software requirement for MATLAB 2012b with specific bioinformatics and statistics toolboxes. The costs involved for buying a twelve-month license are currently £100 for a student or £675 for academic staff. Some institutions provide MATLAB under their software libraries, however several additional toolboxes need to be purchased separately, which are £150 each. The ARTIST pipeline is purely for statistical analysis and the cost implication can be often be discouraging when the output can be achieved using free tools such as R (Ihaka & Gentleman 1996) and edgeR (Robinson et al. 2010). Also the custom scripts are not automated and require a great deal of user input at each step meaning analysis of the data is a time consuming process.

Currently, there is no complete commercial software package to deal with TIS reads from raw data to statistical analysis. Individual software is available to achieve certain aspects, but their accessibility is limited by bioinformatic knowledge. There is a need to develop a complete bioinformatic pipeline for TIS experiments, allowing a relative bioinformatic novice to take the raw data received from sequencing centres and directly achieve interpretable results.

3.2. Results and discussion

3.2.1. Generation of mutant pools

An aliquot from an existing *S. uberis* 0140J pGhost9::IS*S1* bank of mutants was amplified by growing as individual colonies to generate 10 pools, each containing approximately $1x10^4$ mutations. This allowed approximately $1x10^5$ individual mutations to be assessed, and the reproducibility of the enrichment method to be evaluated. If distributed randomly, this number of mutations would equate to a single mutation every 18bp throughout the *S. uberis* genome. The number of bacteria with plasmid integration was determined by enumeration of viable bacteria in the presence and absence of erythromycin. As pGhost9 incorporates an erythromycin resistance cassette, using THA^E (Todd Hewitt Agar) plates, with erythromycin as a selective antibiotic (1µg/ml), enabled positive identification of the number of plasmid integrants. This was compared with growth on THA plates to identify the total number of viable bacteria. The requirement for the overall population of bacteria in the pools to be mutants was achieved by comparison of the total viable bacteria and the number with antibiotic resistance (Table 3.1.). Pools in which the total number of viable bacteria markedly exceeded the number of erythromycin resistant bacteria were discarded (Pools 2, 4, 7, 9). Overall each mutant pool contained approximately 10^{11} CFU (approximately 10^7 copies of each mutation).

Table 3.1. Bacterial counts from serial dilutions to determine numbers of insertions within a pool of mutants.

Sample	Number of Mutants	Solid media type counts		THA^E cfu/ml	THA cfu/ml	Total Mutants cfu/pool
		THA^E	THA			
1	1.60E+04	3.20E+01	5.50E+01	1.60E+10	2.75E+10	1.95E+11
2	1.03E+04	1.30E+01	1.03E+02	6.50E+07	5.15E+08	7.93E+08
3	1.60E+04	2.20E+01	3.40E+01	1.10E+09	1.70E+09	1.34E+10
4	1.60E+04	6.00E+01	5.40E+01	3.00E+08	2.70E+09	3.66E+09
5	1.07E+04	4.60E+01	4.60E+01	2.30E+10	2.30E+10	2.81E+11
6	1.60E+04	3.10E+01	4.90E+01	1.55E+10	2.45E+10	1.89E+11
7	1.10E+04	1.80E+01	4.40E+01	9.00E+08	2.20E+9	1.10E+10
8	1.60E+04	3.70E+01	9.30E+01	1.85E+09	4.65E+09	2.26E+10
9	1.60E+04	7.00E+00	5.50E+01	3.50E+08	2.75E+09	4.27E+09
10	1.03E+04	3.20E+01	4.40E+01	1.60E+10	2.20E+10	1.95E+11
11	6.00E+03	9.10E+01	7.80E+01	4.55E+09	3.90E+09	5.55E+10
12	6.90E+03	7.80E+01	8.10E+01	3.90E+09	4.05E+09	4.76E+10
13	7.80E+03	7.20E+01	6.80E+01	3.60E+09	3.40E+09	4.39E+10
14	7.60E+03	8.10E+01	8.60E+01	4.05E+09	4.30E+09	4.94E+10

THA^E = Todd Hewitt Agar with Erythromycin (1µg/ml), pools consist of 12x1ml aliquots.

3.2.2. Validation of inverse PCR on a population of mutants

The use of inverse PCR to amplify fragments flanking insertions for the identification of individual mutants is well published in multiple bacteria including *Pseudomonas abietaniphila* (Martin & Mohn 1999), *Mycoplasma genitalium* (Hutchison III et al. 1999), *Xanthomonas albilineans* (Huang et al. 2000), *Streptococcus uberis* (Ward et al. 2001) and *Helicobacter pylori* (Salama et al. 2004). However no methods describe the simultaneous application of

this protocol to a population of different mutants. To assess the use of the inverse PCR method for chromosome/transposon region amplification, on a larger population of random mutants, DNA was extracted from one of the mutant pools and inverse PCR conducted following the method outlined for individual mutant identification. Chromosomal DNA was digested using either *Hin*dIII or *Eco*R1 restriction endonucleases; completion of digestion was assessed using 1% agarose gel electrophoresis. The chromosomal DNA was then re-circularised by ligation, and used as the template in an inverse PCR reaction, using primers P064 and P082 (Ward et al. 2001). The PCR product was separated on a 1% agarose gel by electrophoresis and visualised after ethidium bromide staining using a UV transilluminator. The separation pattern of DNA taken from the population of mutant strains is notably different from that of the individual strains, appearing as a smear of DNA from 10kb to 400bp. This is likely to be due to the different locations of endonuclease restriction sites in relation to the plasmid integration site. Therefore the PCR template will be of varying sizes generating an array of many different sized products, making distinction of individual bands impossible (Figure 3.2). This however suggested the plausibility of inverse PCR as a method of enrichment for a population of mutants.

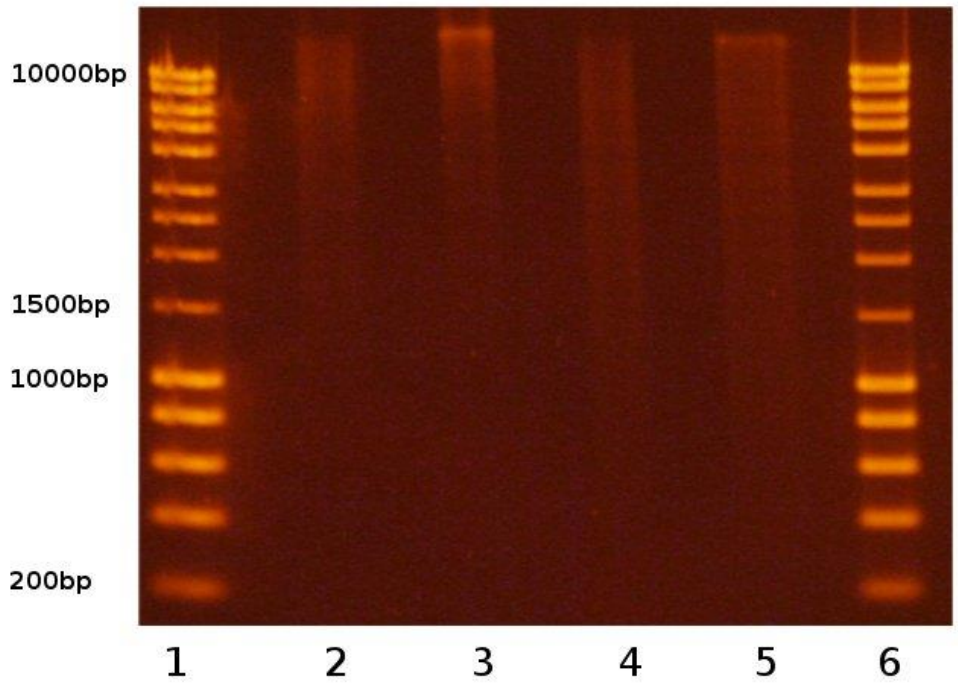

Figure 3.2. Validation of Oligonucleotide primers to be used in the inverse PCR on a population of mutants. The PCR products are visualised as a smear, rather than distinct bands due to the large number of individual mutants within the pool chosen, creating numerous bands of similar sizes. Lanes 1 and 6 contain Hyperladder I molecular weight marker. Lanes 2 and 4 contain a HindIII digested inverse PCR products and lanes 3 and 5 contain EcoR1 digested inverse PCR products.

The use of restriction endonucleases can introduce fragment size bias when they are used to generate DNA fragments. This is due to the non-random distribution of their specific recognition sites within the *S. uberis* 0140J genome (Accession number AM946015) (Ward et al. 2001) (Figure 3.3). There are alternative digestion methods, to overcome restriction site bias, using a mixture of 'rare-cutter' enzymes called Fragmentase. However these can result in a broad fragment distribution and introduce artificial DNA modifications (Head et al. 2014).

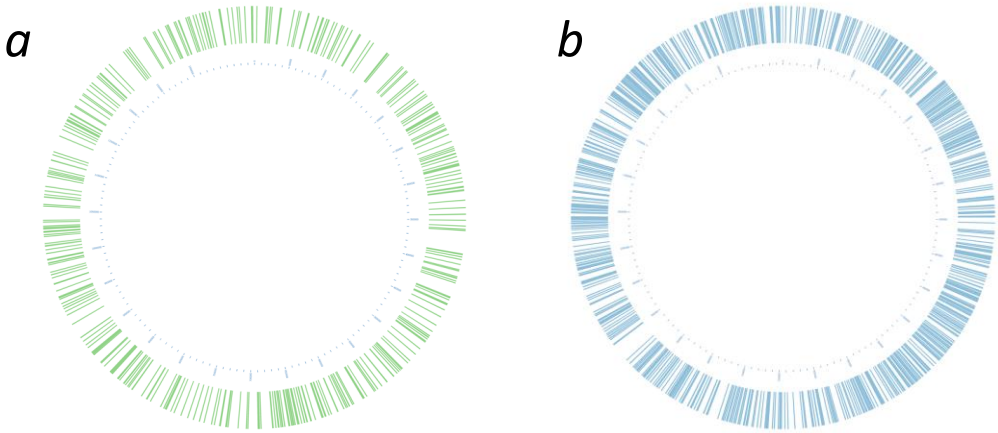

Figure 3.3 Distribution of restriction endonuclease sites within the *S. uberis* genome. Genomic coordinates of endonuclease restriction sites were plotted using Circos (Krzywinski et al. 2009) a) Shows the 491 *Eco*R1 restriction sites b) Shows the 1086 *Hind*III restriction sites.

To assess the impact of endonuclease digestion on the DNA products, used for ligation and subsequent inverse PCR amplification, a comparison was performed using a method of random DNA fragmentation. To be a true substitute for enzymatic digestion the generated fragments need to be of relatively equal length and truly random. This overcomes the issue of PCR amplification bias, where smaller ligated fragments amplify more efficiently than larger fragments, soon becoming dominant in the sample and the pseudo-random identification of mutations relating to the presence of endonuclease restriction sites. Alternative methods routinely used are acoustic fragmentation, sonication and nebulisation. Nebulisation is inexpensive, however it requires a significant amount of input DNA, approximately 10μg in 2ml of appropriate buffer (Quail 2010). The nebulised DNA has a broad fragment distribution and the desired size range only accounts for 5% of the final volume of DNA, therefore is uneconomical in

terms of sample retention (Quail et al. 2009). Sonication is often a source of DNA contamination and also has a wide fragment distribution, resulting in wasted DNA (Quail et al. 2009). Acoustic fragmentation is achieved by focusing acoustic energy into a small-localised zone. This creates cavitation bubbles which collapse, breaking the double stranded DNA. This method has several advantages over those previously discussed. Only a small amount of DNA is required, and the closed independent vessel means the contamination risk is low (Fisher et al. 2011). As fragment size plays an important role in inverse PCR, a range of size fragments was generated based on previously documented results by Ochman et al. 1988. A fragment size of 2kb greater than the insertion core region has been shown to amplify the most efficiently using standard PCR reagents. As the IS*S1* insertion is 800bp, a range of sizes was produced (1.5kb, 2kb, 2.5kb and 3kb), to enable an average size, optimal for inverse PCR, to be identified. DNA was fragmented using Covaris Adaptive-Focused-Acoustics and the distribution of fragmented DNA was quantified using fluorescent microfluidic electrophoresis. This separates the fluorescently labelled DNA in a microchannel and measures the relative fluorescent units (RFU). This was compared to a reference ladder for fragment size measurement. The ranges of size distributions were a) 1594bp (expected 1.5kb), b) 2182bp (expected 2kb), c) 2587bp (expected 2.5kb) and d) 2824bp (expected 3kb) in length (Figure 3.4).

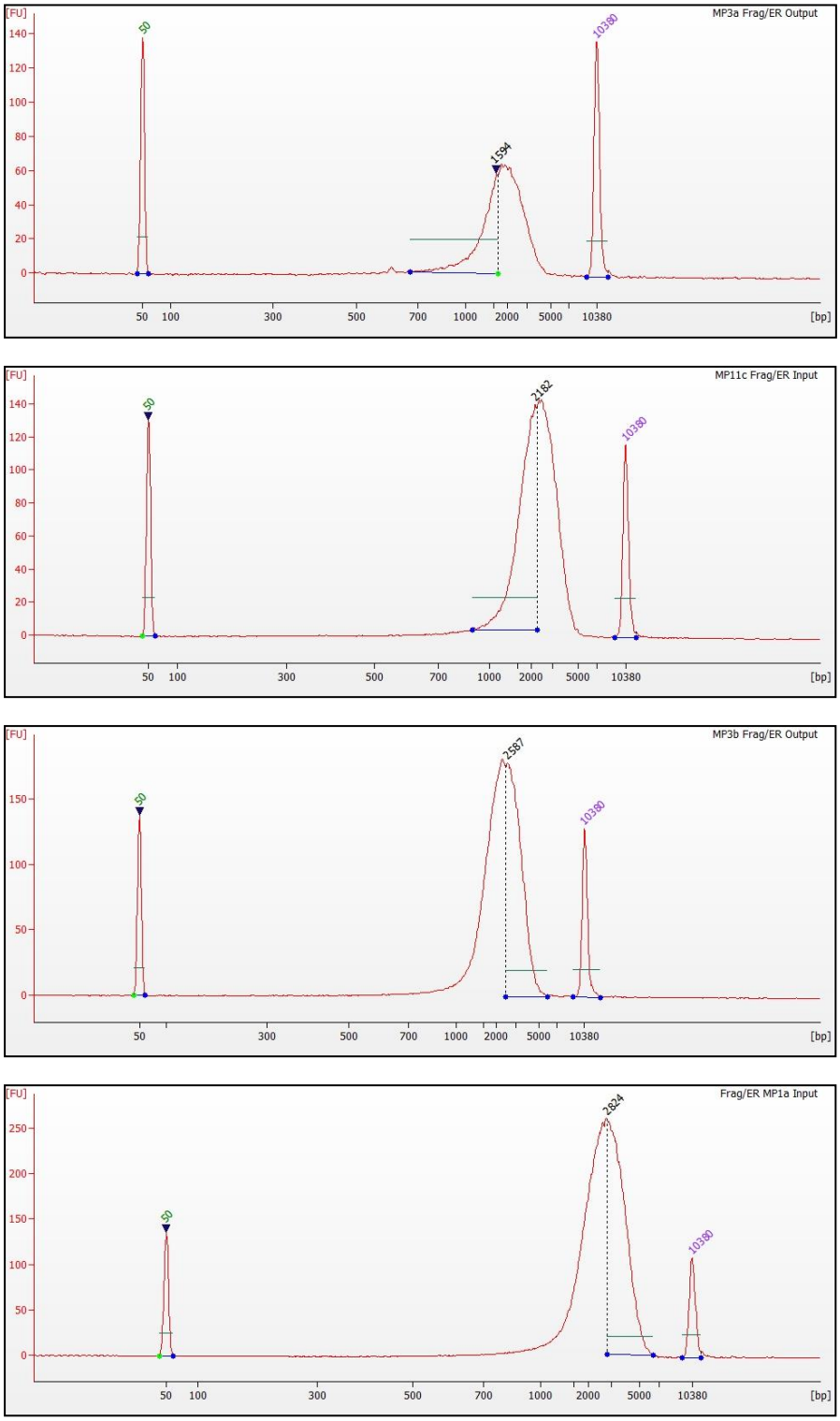

Figure 3.4. Electropherograms of acoustically fragmented DNA. DNA was fragmented to a range of approximate lengths using Adaptive-Focused-Acoustics. A sample of the fragmented DNA was diluted and the fragment distribution assessed using fluorescent microfluidic electrophoresis. The average fragment sizes were determined to be approximately a) 1594bp, b) 2182bp, c) 2587bp and d) 2824bp in length.

3.2.3. Enrichment of insertion flanking regions for acoustically fragmented DNA

The physical characteristics of fragmented DNA generated by acoustic methods differ from endonuclease-digested samples. An enzyme will create either blunt ended fragments, where both strands of DNA are the same length, or sticky ended fragments, where each strand will have an complementary overhang of equal length. When acoustic fragmentation is used, DNA is stressed producing staggered ends, where there are longer overhangs of either the positive or negative strand. This inhibits the circularisation of fragments, as the ends are of unequal lengths. To enable efficient ligation of acoustic fragmented DNA, the strands must be enzymatically end-repaired in order to create blunt ended fragments. However, blunt end repaired fragments have an increased chance of concatenation over recircularisation during ligation, and this is thought to be due to concentration of reactants and hence reaction volume. Large reaction volumes have been shown to reduce the incidence of concatenation in similar studies (Baureder & Hederstedt 2012). To assess the potential for this occurring, a comparison of ligation reaction volumes was performed. A 50µl reaction volume, as used for evaluation of inverse PCR on individual mutants (Ward et al. 2001), was compared to a greater 750µl ligation volume. The acoustically fragmented DNA samples generated (1.5kb, 2kb, 2.5kb and 3kb) were ligated in either the 50µl or 750µl volumes, before purification, and use the template in the inverse PCR. The resulting amplified products were separated on a 1% agarose gel by electrophoresis and visualised after

ethidium bromide staining using a UV transilluminator (Figure 3.5). Each of the different fragmented size ranges produced a DNA smear from approximately 10kb to 100bp in the 50µl reaction volume, regardless of the size of DNA fragment used for recircularisation and inverse PCR. The samples ligated in the larger 750µl reaction volume only produced a product when the average fragment size used for re-circularisation was 3kb (Figure 3.5, Lane 12). This suggests the larger 3kb fragments are the optimal size to generate an inverse PCR product but also that the 750µl ligation reaction volume is best suited to reduce concatenation and create circularised DNA fragments.

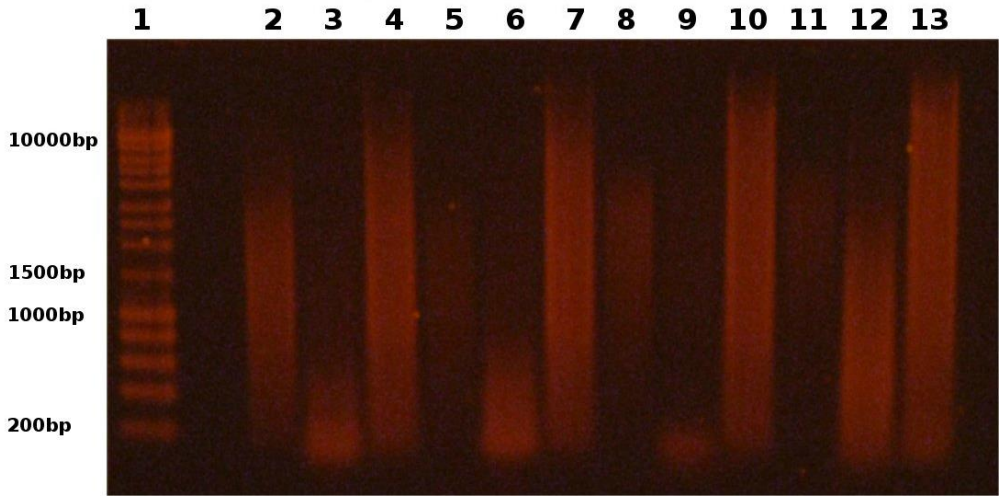

Figure 3.5. Inverse PCR products using the acoustic fragmented DNA at different sizes. DNA was fragmented using Adaptive-Focused-Acoustics to 1.5kb, 2kb, 2.5kb and 3kb then end repaired. 1µg of the fragmented DNA was ligated in either a 50µl or 750µl reaction volume and each subjected to inverse PCR using primers P064 and P082 (Ward et al. 2001). **Lane 1** – Hyperladder I, **Lane 2** – DNA fragmented to 1.5kb (50ng), **Lane 3** – DNA fragmented to 1.5kb inverse PCR product 750µl ligation, **Lane 4** – DNA fragmented to 1.5kb inverse PCR product 50µl ligation, **Lane 5** – DNA fragmented to 2kb (50ng), **Lane 6** – DNA fragmented to 2kb inverse PCR product 750µl ligation, **Lane 7** – DNA fragmented to 2kb inverse PCR product 50µl ligation, **Lane 8** – DNA fragmented to 2.5kb (50ng), **Lane 9** – DNA fragmented to 2.5kb inverse PCR product 750µl ligation, **Lane 10** – DNA fragmented to 2.5kb inverse PCR product 50µl ligation, **Lane 11** – DNA fragmented to 3kb (50ng), **Lane 12** – DNA fragmented to 3kb inverse PCR product 750µl ligation, **Lane 13** – DNA fragmented to 3kb inverse PCR product 50µl ligation

3.2.4. Sequencing of inverse PCR products

Successful PCR amplification of samples suggested inverse PCR was a plausible method for enrichment of DNA flanking transposon insertion sites within a population of mutants. The acoustically fragmented, inverse PCR enriched sample (referred to as Covaris), was compared to a PCR template generated by a restriction endonuclease digest using either *Hin*dIII or *Eco*R1. Also, in order to assess the benefits of inverse PCR, a fourth sample of DNA was fragmented acoustically and was sequenced without any inverse PCR enrichment (referred to as gDNA).

3.2.5. Analysis of data

The DNA samples were sequenced using an Illumina MiSeq. The sequencing equipment generates a primary quality filter, called Pass Filter (PF) reads, which determines the reliability of the sequence reads. The score, of which PF reads are determined, is calculated by the ratio of the two brightest clusters. The thresholds for a PF read is determined by no more than one base call having a score of <0.6 for the first 25 cycles. Total yield of sequence data was 15,405,584 reads with 14,242,067 reads passing filter. From the PF reads, 13,498,417 reads (94.8%) were allocated to the indexed samples and were distributed accordingly (Table 3.2). After successful sequencing, quality control is an important stage that is often overlooked but is essential for meaningful downstream analysis. Assessments of sequence duplication levels, over represented sequences, sequence length distribution and quality scores need

to be implemented to allow for erroneous data to be identified and rectified (Patel & Jain 2012). The quality of the data was assessed by Phred scores (Ewing et al. 2005a; Ewing et al. 2005b) using FastQC (version 0.10.1 http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Phred scores are calculated for each base of every read and are widely used for characterising the quality of the data and the efficiency of the sequence run. A base Phred score of 30 (99.9% accuracy probability) (Figure 3.6a) is the suggested quality standard cut-off (Ewing et al. 2005b). The Phred approach uses base call parameters to look up a corresponding quality score in a look-up table. These tables are generated from correct base calls and contain a number of acceptable variations. Phred scores are a method of handling low error probabilities and are calculated by log transformation of error probability (*p*) using the equation:

$$Phred\ score = -10\log_{10}(p)$$

All samples were shown to have an average Phred score of >30 and no single base had a Phred score of lower than 31 (Table 3.2, Figure 3.6b). As the quality of the sequence data generated by the inverse PCR samples (Covaris, Table 3.2) is comparable to the non-enriched whole genomic DNA sample (gDNA,

Table 3.2. The associated number of reads and average Phred scores for each sample.

Method	Total PF Reads	Average Phred Score	Minimum base Phred Score	Maximum base Phred Score
*Eco*R1	3848211	37.26	33.33	38.48
*Hin*dIII	3294706	37.31	33.6	38.48
gDNA	2977706	37.48	32.23	38.35
Covaris	3377794	36.75	31.59	38.62

Table 3.2), it indicates that inverse PCR enrichment could be used as a sequence library input and generates sequence data that is of high quality.

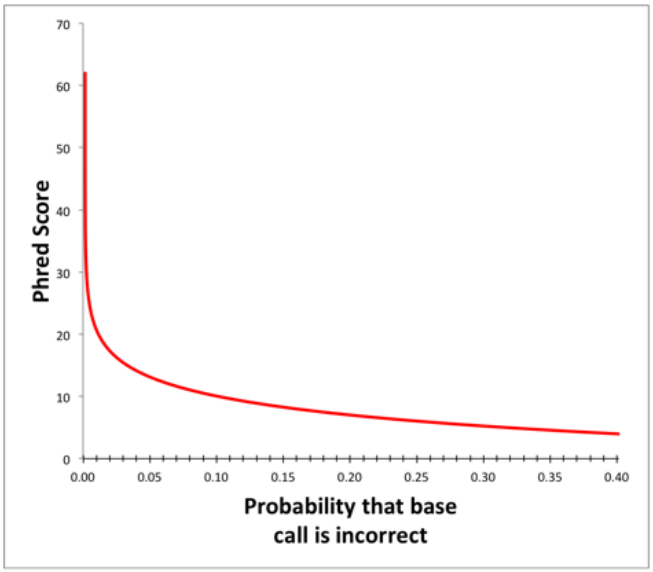

Figure 3.6a Association between Phred score and probability of error For example a Phred score of 30 represents a 1 in 1000 chance of being incorrect, or 99.9% accurate

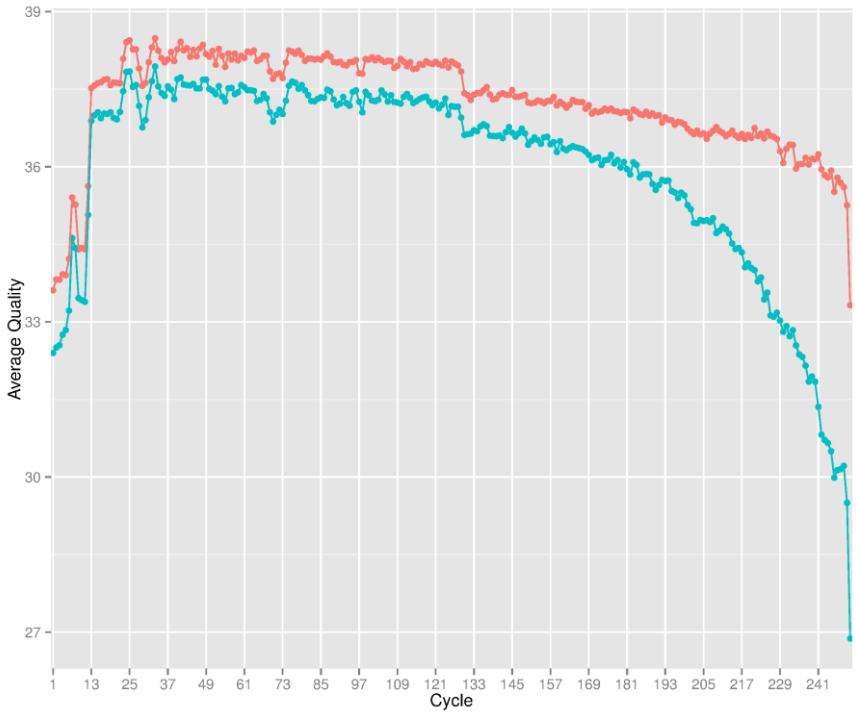

Figure 3.6b Quality score graphs generated from each base Phred score. Bar plot showing the proportion of quality calls per cycle. Colours are presented in red and blue, where red identifies read one and blue identifies read two. Images generated using RQC 3.1 (https://github.com/labbcb/Rqc).

3.2.6. Development of a data analysis pipeline

The software tool Pragmatic Insertion Mutagenesis Mapping System (PIMMS) was developed to analyse these data (https://github.com/ADAC-UoN/PIMMS) (Blanchard et al. 2015). The bioinformatic pipeline is broken down into component modules (mapping, process.sam and counts) (Figure 3.7) that are written in Perl, a general Unix scripting language. Each module can be called through the command line using

```
PIMMS.pl –m [module]
```

This initiates a prompt for the remaining parameters and inputs for the module to run, which will be discussed in each corresponding section of this Chapter.

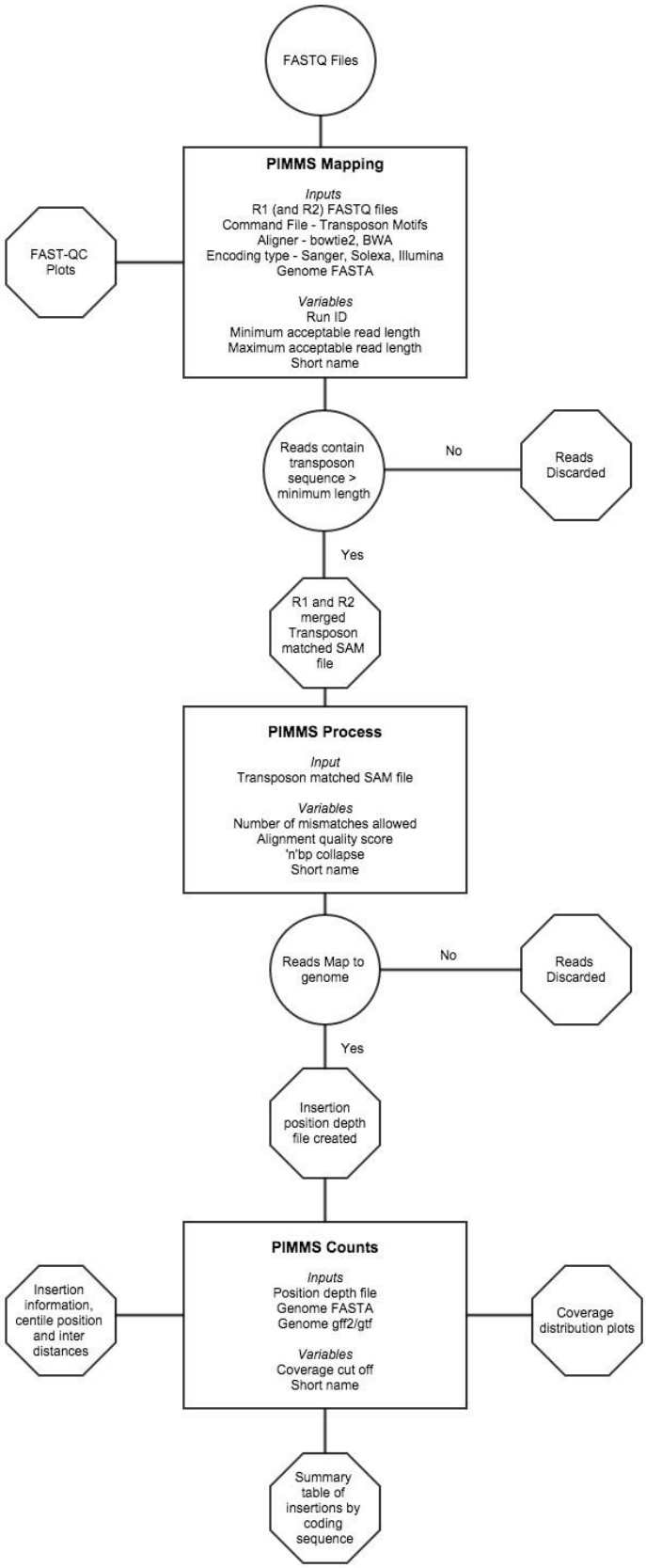

Figure 3.7. Schematic of the PIMMS data analysis process Inputs are represented by circles, outputs by hexagons and variables by squares.

3.2.6.1. Identification of terminal insertion sequence and mapping to genome

The first aspect of the PIMMS pipeline, mapping, uses the raw FASTQ files generated from the sequencer along with a reference genome file and a command file containing the transposon motifs. This was be initiated by the following command:

```
PIMMS.pl –m [mapping] –c [command file]
–i [Fastq file 1] –j [Fastq file 2 or FALSE]
–g [Reference genome file] –e [Sequencing
encoding] –r [Sequencer ID] –min [Minimum length
of sequence to be retained] –max [Maximum length
of sequence to be retained] –n [name for output]
```

The mapping module identifies the transposon terminal sequence. This 23bp region is recognised in both orientations from 5' and 3' ends of the insertion sequence. Once identified the sequence adjacent to the transposon motif is copied to an additional FASTQ file along with the corresponding Phred quality score. The length of this fragment is dependant on the command input but the default is a minimum of 20bp and a maximum of 50bp. This aims to allow for an unequivocal alignment to the reference genome but avoids using reads where the fragment re-enters the transposon when generated from a circular PCR product (Figure 3.8).

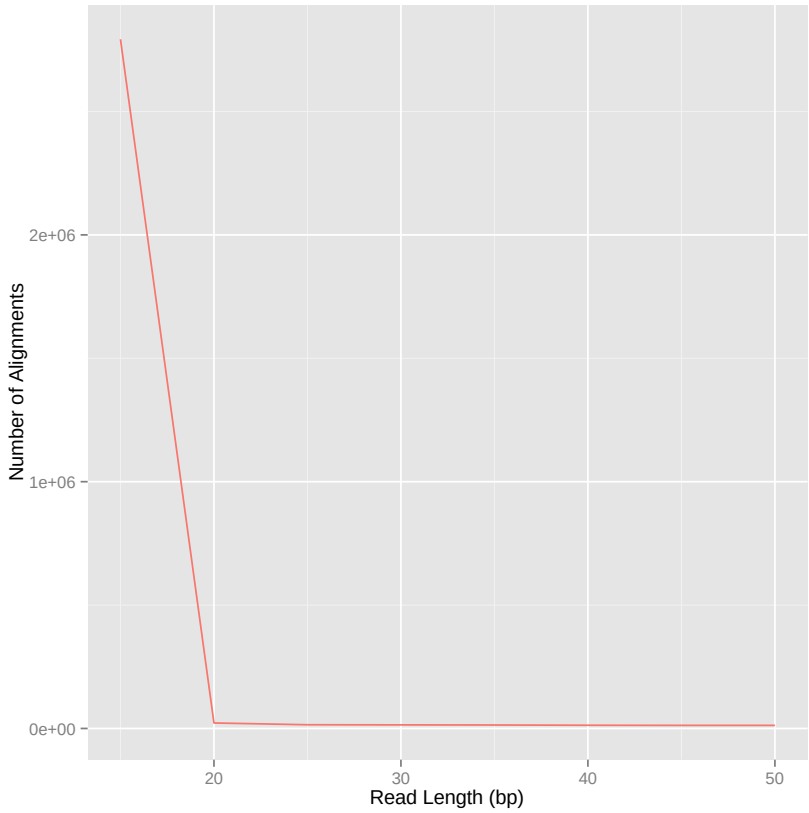

Figure 3.8. Effect of read length on alignment confidence. The number of times a single read, at varying lengths could be aligned to the reference *S. uberis* genome (Accession number AM946015) (Ward et al. 2001) was recorded.

Analysis of the sequence data generated from the *Eco*R1, *Hin*dIII, Covaris, gDNA samples, was carried out using the PIMMS pipeline. The number of reads that contained the transposon terminal sequence (Matched Reads) was similar for the endonuclease-digested samples with 62.3% and 62.9% of the total PF reads including the target sequence. The Covaris sample showed fewer matched reads with only 40.8% containing the terminal insertion sequence and unenriched genomic DNA sample only produced 0.2% (Table 3.3). The efficiency of the pipeline to identify matched reads is in line with expectations. The Covaris samples are produced by an approximate 2kb PCR product, when this is fragmented further (approximately 500bp) for library

preparation this would generate 2-4 fragments for sequencing equating to approximately 25-40% matched reads. The occurrence of an increased number of reads that contain the terminal sequence with an endonuclease-digested sample is expected. This is due to a higher percentage of fragments containing the transposon junction because of the recognition site and non-random digestion.

Table 3.3. Associated number of reads that contain the transposon motif.

Method	Total PF Reads*	Matched Reads	Percentage (Matched/Total)
*Eco*R1	3848211	2400817	62.39
*Hin*dIII	3294706	2075014	62.98
Covaris	3377794	1380699	40.88
gDNA	2977706	5817	0.20

*Reads were generated using 2x250bp Illumina MiSeq chemistry reagents

3.2.6.2. Analysis of alignment methods

Alignment of DNA sequences is an integral part of many bioinformatic pipelines and there are numerous sequence aligners available. The decision about which to use is an important consideration and often eludes novice users and the misuse can lead to incorrectly interpreted data. Aligning a short sequence to a much longer genome or database of genomes has been approached and reassessed many times over. Dynamic programming using a matrix of all possible alignment positions, was first proposed as an optimal solution to overcome this problem in the early 1970s (Needleman & Wunsch 1970) and later refined to take into account a sequence similarity scoring system in the early 1980s (Smith & Waterman 1981). This method becomes

impractical with the volume of NGS data as the matrix required soon exceeds computational memory restraints. Heuristic solutions such as BLAST (Altschul et al. 1990) overcome this by building indices of the sequence, but are very time consuming to run. Therefore more powerful algorithms have been developed with additional refinements to deal with repetitive genomic regions. To empirically determine the affect of the chosen aligner, the efficiency of mapping was investigated using two commonly used alignment packages; bowtie2 (Langmead & Salzberg 2012) and BWA-MEM (Li 2013). The decision to compare bowtie2 and BWA was due to their active use in the community, high level of user support and also that each method provides per read mapping quality scores allowing subsequent filtering, which makes them ideally suited for this application.

Random subsets of reads were taken from each sample (*Eco*R1, *Hin*dIII, Covaris, gDNA), containing 1,000,000, 2,000,000, 3,000,000 and total reads. Each subset was generated in triplicate to assess variation of mapping. The reads were then aligned to the *S. uberis* 0140J reference genome (Accession number AM946015) (Ward et al. 2001) with varying stringency settings for the two different sequence read aligners. The number of alignments was then plotted in a scatterplot and a LOWESS (local weighted scatterplot smoothing) curve added to represent a smoothed value for each replicate. Knowing there were a finite number of mutations present in the samples sequenced, it was postulated that with increasing number of reads there would be a plateau of newly discovered insertion points as insertions were repeatedly identified.

Alignments of 50bp reads were carried out using bowtie2 with differential stringencies set at 0, -10 and -30. Using 0, the alignment must be a perfect match the entire length of the read. The use of -10 and -30 are increasingly lenient penalty scores for mismatches. These relate to gaps of alignment within the read, with a greater penalty assigned in relation to a mismatch of a base with higher quality scores (Figure 3.9a). With the mismatch allowance set at -30 the subset replicates show a high degree of variation when compared to the most stringent setting of zero. This is evident by the discrimination of all three points of the scatterplot at each subset number of reads for the -10 and -30 settings. Using BWA-MEM, the alignments scores are different with 50 being a perfect alignment for the whole length of the read; alignment scores of 10 and 30 being a percentage of mismatches relative to the read length and 0 being the least stringent allowing for all suboptimal alignments. There is less variation between the replicates even with the more relaxed alignment settings, when compared with the bowtie2 alignments (Figure 3.9b). This shows that stringency has a dominant effect on both aligners. When both aligners are set to zero mismatches there is little variability between the subsets with either aligner. However based on bacterial colony counts performed when the mutant pools were generated, approximately 16,000 unique insertions were expected. BWA shows more evidence of the plateau expected, suggesting that the higher number of insertions found by bowtie2 may be due to erroneous alignments (Figure 3.9b). The number of insertions found with the endonuclease prepared sample would be limited with the number of restriction sites so could be expected to be lower than the

randomly fragmented sample.

To assess the degree of variation of the unique insertions made by both aligners, the insertion coordinates from bowtie2 and BWA were compared (Figure 3.10). Alignments generated by BWA are almost entirely contained within the bowtie2 alignments for each sample. With further investigation the removal of insertions with only a single read coverage dramatically reduced the number of insertions identified by bowtie2 alone. Suggesting that low read depth (as might be expected for erroneous alignments) is largely responsible for the disparities. BWA seems to be more reproducible when aligning reads from a randomly selected number of reads and the positions identified are within the ranges expected from colony counts. Therefore BWA was used as the default aligner for further experiments.

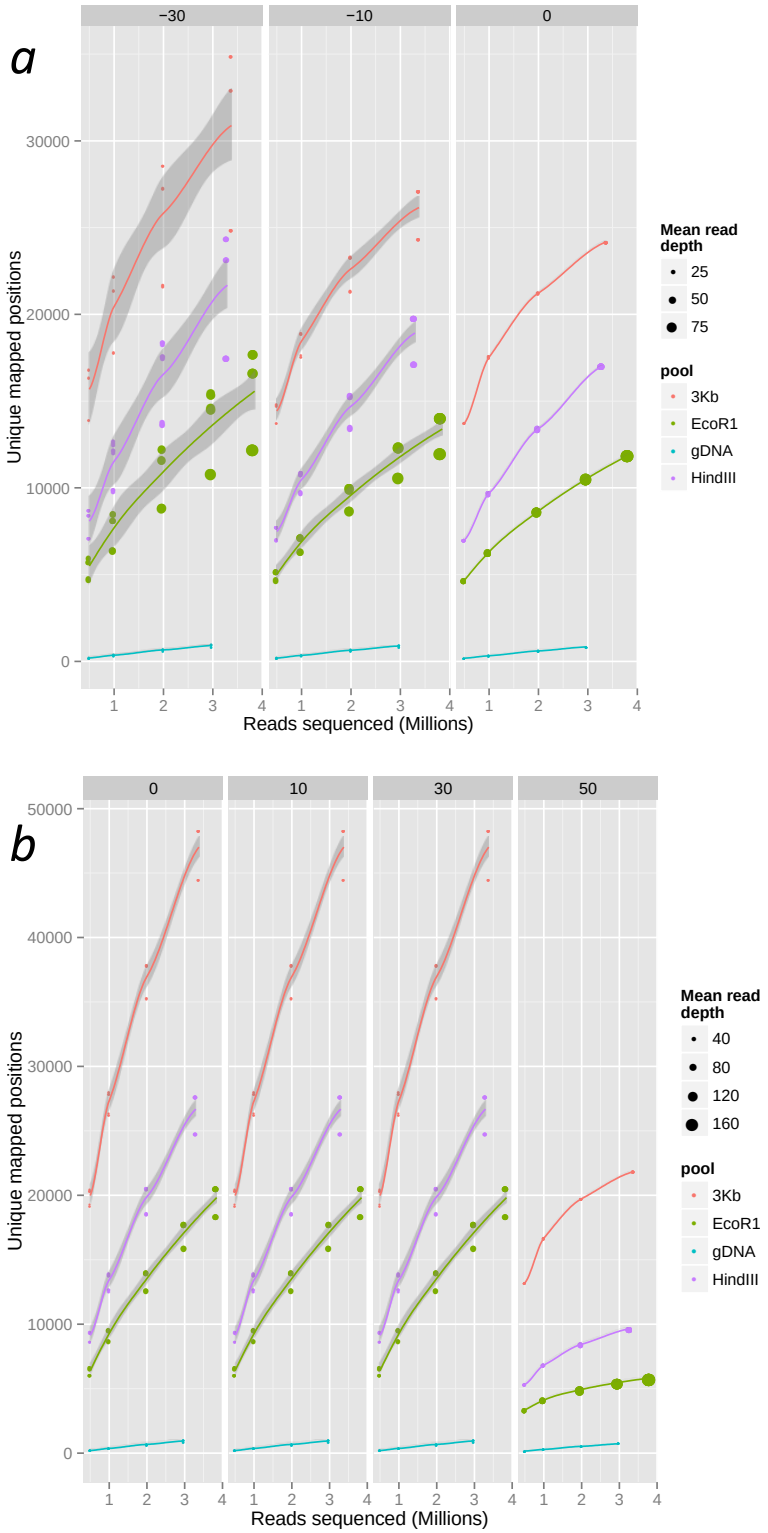

Figure 3.9. Random sampling graphs generated by number of alignments. a) Alignments using bowtie2 b) using BWA. Columns of the plot show changes with mismatches, and rows changes in alignment scores. Size of circles represents the mean number of reads at each unique position.

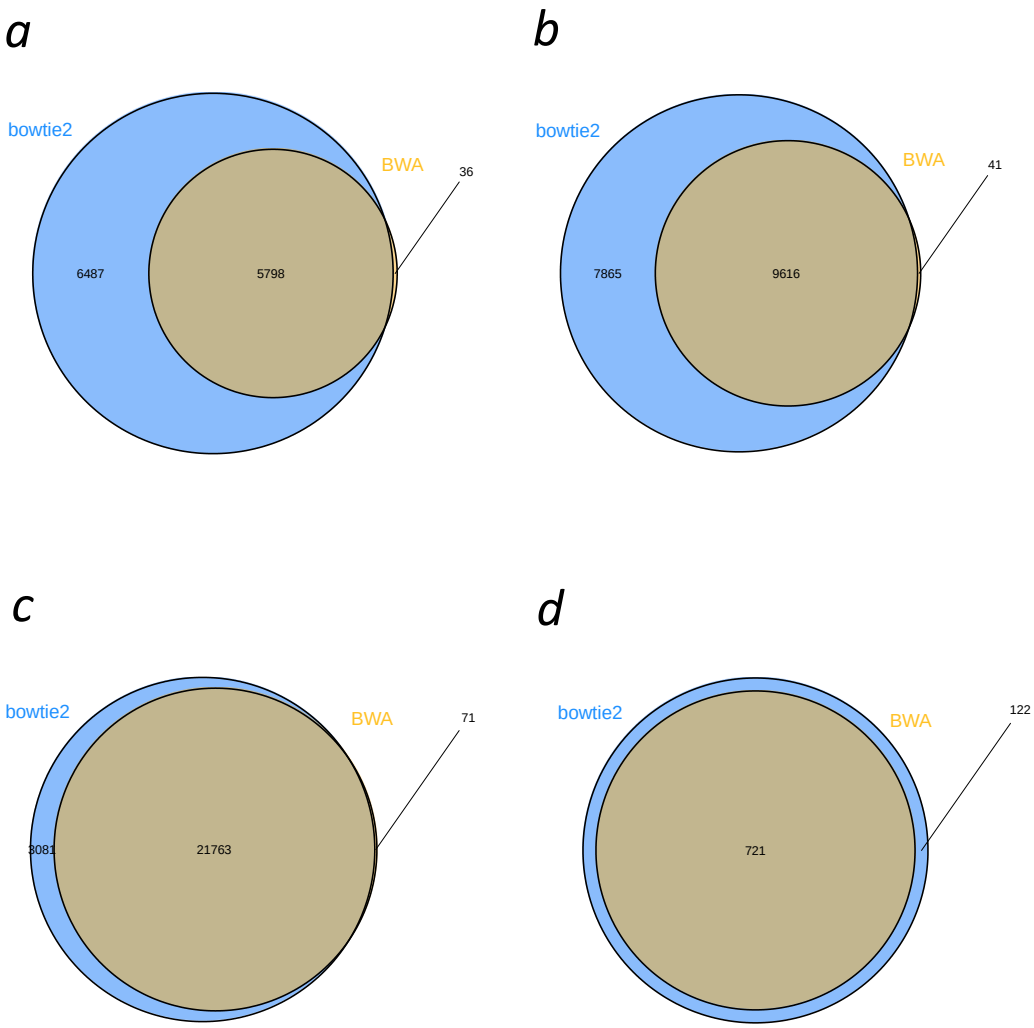

Figure 3.10. Venn diagrams demonstrating the overlap of alignments using bowte2 and BWA with the most stringent settings. a) are alignments using the *Eco*R1 digested sample. b) are alignments using the *Hind*III digested sample. c) are alignments using the Covaris fragmented sample. d) are alignments using the unenriched genomic DNA sample.

Based on the results gained from the alignment accuracy assessment, the new FASTQ files, which contain the filtered reads, were then aligned to the reference *S. uberis* 0140J genome (AM946015) (Ward et al. 2001) using BWA-MEM (Li 2013) as the default PIMMS aligner. Using the highest stringency settings, BWA aligned 24% of the Covaris fragmented samples and unenriched samples. The *Eco*R1 endonuclease digestion sample consisted of 39% reads that mapped to the reference genome and the *Hin*dIII contained 29.5% (Table 3.4).

Table 3.4. Metrics of reads that aligned with BWA to the *S. uberis* genome.

Method	Matched Reads	Mapped Reads	Percentage (Mapped/Matched)
*Eco*R1	2400817	937081	39.03
*Hin*dIII	2075014	613441	29.56
Covaris	1380699	341562	24.74
gDNA	5817	1439	24.74

3.2.6.3. Processing insertion coordinates

The reads that map to the reference genome, generate a Sequence Alignment Map (SAM) file; which includes for each mapped read, the associated number of mismatched bases (MD) and overall alignment score (AS). The SAM file is processed further to take into account the quality of the alignment and the occurrence of the inverted 8bp repeat that is generated by the insertion using the pGhost9::IS*S1* plasmid vector. The mapping module is used by:

```
PIMMS.pl –m [process.sam] –s [.sam file] –n
[name] –a [stringency of alignment] –m [number
of allowed mismatches in read]
```

For analysis of the data the highest stringency setting for BWA was used (50) and no mismatches were tolerated. Using these settings the process.sam module will only use the reads that have MD and AS tags of 50 and 0 respectively. This means only reads that align for the whole length of the read with zero mismatches are used. It will then look for the read start and stop coordinates. If two insertion positions are separated by exactly 8bp they are classed as the same insertion event and the read depths at these positions are collapsed. By default the lower number insertion position is retained.

These filtered reads are used to generate a processed.sam file, which contains a count of the number of reads at each genomic coordinate. The final module is incited by:

```
PIMMS.pl -m [counts] -d [processed SAM file]
-gtf [general feature format file] -g [reference
genome file] -n [short name for output file]
```

Insertion number per coding sequence (CDS) is determined by comparison to the genome annotation contained in the corresponding general feature format (GFF) file. The insertion coordinates are parsed against this file, if an insertion falls within the start and end positions they are added to a count file related to each CDS. From this a summary table identifying the number of reads associated to each CDS and the number of unique insertion positions within the CDS is generated. From this essential CDS can be inferred. Using the random fragmentation method 21834 unique insertions sites were identified and 720 were identified in the unenriched gDNA. For the two samples

prepared for inverse PCR using endonuclease digestion, the number of unique insertions identified in the *Hin*dIII samples is almost double to that of the *Eco*R1 (Table 3.5). Which may be expected, as there are 1086 *Hin*dIII and only 495 *Eco*R1 restriction sites within this particular *S. uberis* genome.

Table 3.5. Metrics of insertion point discovery relative to CDS

Method	Unique Insertion Points	Average Read Depth	Average Inter-insertion Distance
*Eco*R1	5835	160.6	89
*Hin*dIII	9657	63.5	72
Covaris	21834	15.6	41
gDNA	721	1.9	1054

The coordinates were used to generate a circular representation of the relative position of insertions within the *S. uberis* genome (Figure 3.11). The four plots show the unique insertion points for the mutations found with the associated read depth for each of the sample preparation methods. The endonuclease-digested samples (A and B) show a noticeable 'patchy' distribution when compared to the plot generated by the randomly fragmented Covaris sample (D). This is likely to be related to the number of recognition sites for either one of the endonuclease enzymes. The random nature of the Covaris fragmentation has generated a more even distribution of insertion points. There are also regions where no insertions have been identified. These regions also correspond to non-mutated regions on the *Eco*R1 and *Hin*dIII prepared samples suggesting the positions of essential genes.

As expected the gDNA sample (E) shows very few insertion sites, the number was approximated to be 803, this was achieved by using the formula:

$$\mathrm{RS}/g = gn$$

where 'R' is the number of PF reads for the sample (2977706), 'S' is the number of bases sequenced per read (500), 'g' is the length of the genome (1852352) and 'gn' equals the number of genomes sequenced (803). This is slightly higher than the amount of unique insertions discovered, however some insertions were found more than once (Figure 3.11 E) potentially accounting for the slight overestimation. This shows that the Covaris sample gave 30-fold enrichment over the gDNA sample.

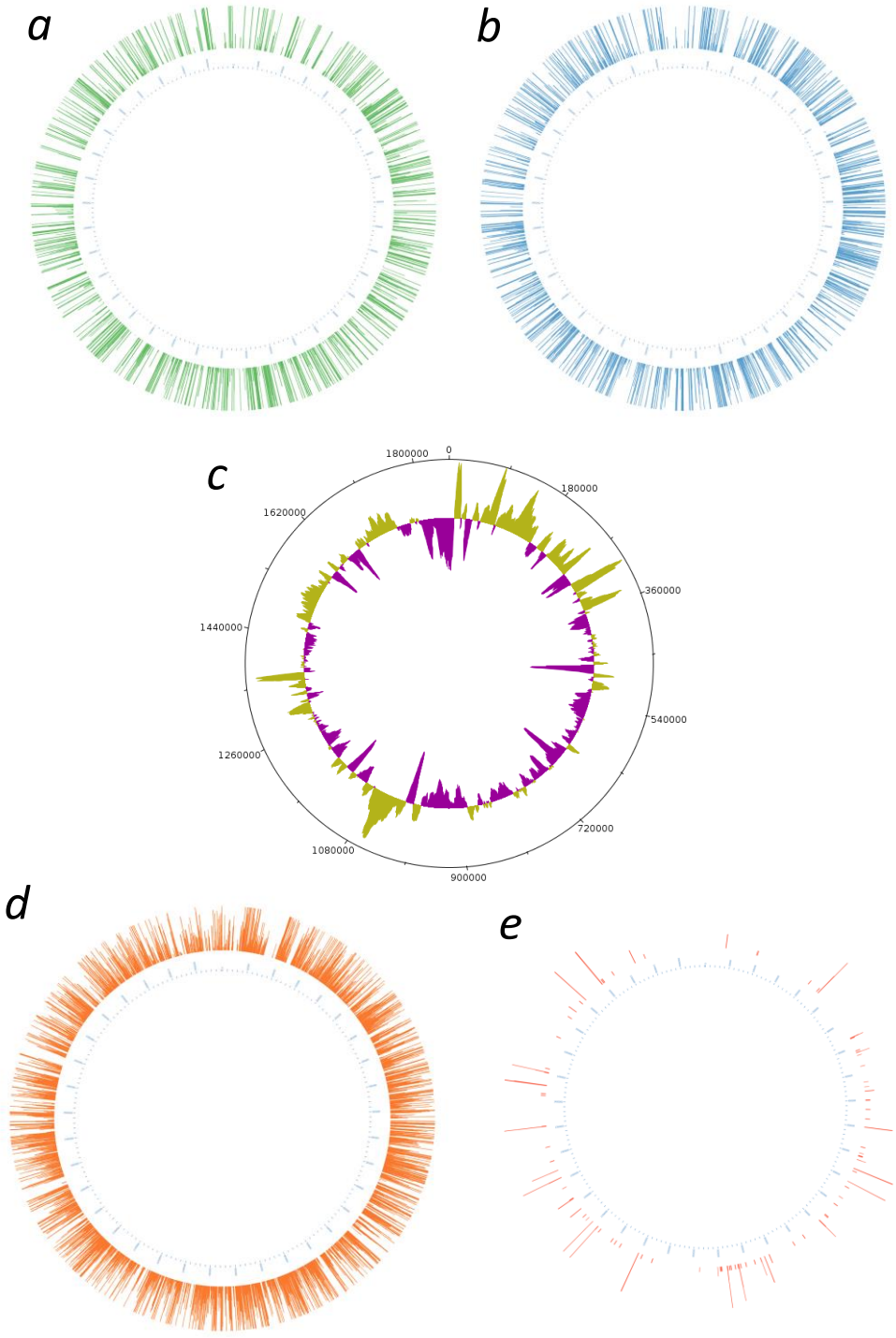

Figure 3.11. Circular representation plots of unique insertion points identified by each sample preparation method. Each of the coordinate files were used to generate a Circos plot (Krzywinski et al. 2009) of insertion points identified in the *S. uberis* genome using each fragmentation/enrichment method and the genomic DNA control. a) *Eco*R1 digested DNA – inverse PCR enriched, b) *Hin*dIII digested DNA – inverse PCR enriched, c) GC bias throughout the genome, d) Covaris fragmented – inverse PCR enriched and e) Unenriched control sample.

3.4. Conclusions

3.4.1. Use of inverse PCR as a method of transposon site enrichment

Inverse PCR, as a means of enrichment, is a simple, effective and widely used laboratory protocol. It has been successfully utilised in the identification of genetic rearrangement (van Bakel et al. 1995), the location of viral DNA incorporated into the human genome (Neves et al. 1998; Tönjes et al. 1999; Ohshima et al. 1997) and the elucidation of cDNA sequences (Chowers et al. 1995). Its application of identifying transposon insertion flanking regions has been extensively used in several bacteria (Huang et al. 2000; Martin & Mohn 1999; Hutchison III et al. 1999; Salama et al. 2004), and more importantly its function has been thoroughly validated in *S. uberis* (Ward et al. 2001; Kerro Dego et al. 2011; Johnsen et al. 1999; Wirawan & Klesse 2006). Until now inverse PCR has only been used to identify single transposon insertion sites and its application to find flanking regions within a population had not been reported. By combining inverse PCR enrichment with massively parallel DNA sequencing, it is possible to comprehensively identify multiple mutated coding sequences. The ability to utilise populations mutated with pGhost9::IS*S1* is a significant advance in the tools available for functional genomics of *Streptococcus* and other Gram-positive bacteria.

The transfer of this methodology into a high-throughput system had to address key issues, such as the use of endonuclease digestion within a population, alternative fragmentation methods and the handling of large

volumes of data with subsequent analysis.

The comparison of endonuclease digests with mechanical fragmentation methods have been previously reported (Knierim et al. 2011), however there is little information in the literature regarding the relative advantages or disadvantages of this approach. Other than the well-documented issue of finite recognition sites within genomes, endonucleases have also been found to introduce nucleotide insertions and deletions (Quail et al. 2012). These are introduced during digestion where incomplete cuts of the DNA and potential cut repairs form a loss of nucleotides (Knierim et al. 2011). This can lead to alignment errors especially with the stringency needed, which would subsequently lead to these reads being ignored.

Due to its consistency, accuracy and low sample requirements the preferred alternative for endonuclease digestion is acoustic fragmentation. Acoustically fragmented DNA has been shown to produce 17% of the sample to the desired length whilst being accurate and reproducible. However the nature of acoustically fragmented DNA does introduce other aspects for consideration. For successful amplification of acoustically fragmented DNA by inverse PCR, the fragments need to be end repaired to allow for circularisation and are required to be of sufficient length for the ligase enzyme to generate a circular product. The optimal DNA fragment length was found to be approximately 3kb (actual length in this case was 2824bp), which is in line with other studies that have shown the optimal size for an inverse PCR template must be 2kb larger

than the separation of the primers, in this case 800bp core region of the IS*S1* element (Ochman et al. 1988).

To avoid the occurrence of concatenated blunt ended DNA fragments meant the difference in reaction volumes for the ligation was an important parameter to assess. The larger 750µl reaction introduced a volume-excluding effect, meaning the DNA fragments are less likely to be in close proximity to each other so it is less likely that the ligase can join separate fragments together, reducing the chance of concatenation. This step however does introduce a potential limitation; the need to reduce the volume from the ligation reaction before the use of the ligated DNA in the PCR can lead to DNA losses. This enforces the need to have sufficient copies of each mutation in the initial DNA preparation to ensure their presence in the to PCR reaction. Similarly there are no current methods for determining the efficacy of the circularisation, therefore the quantity of PCR template can only be based on standard DNA quantification methods and the ratio of ligated to linear DNA cannot be validated.

The use of endonuclease-digested samples, in the application of TIS experiments, has been shown to provide robust data. However, acoustic fragmentation should be considered a necessary replacement due to the limitations associated with endonuclease digestion. Downstream analysis could be skewed, when determining essential genes, because of the lack of restriction sites. This ambiguity reduces confidence, as a CDS identified as

"essential" as it has no insertions may just lack a restriction site and not be truly essential.

3.4.2. Bioinformatic analysis of transposon insertion site sequence data

The difficulty associated with data analysis of transposon insertion site sequencing is mainly related to the composition of the reads. Current alignment software is unable to deal with this chimeric arrangement so it is necessary to separate the two parts. Achieving this is challenging without previous bioinformatic knowledge and the only current described method is to utilise the barcode removal tool of FASTX, previously discussed, with the transposon sequence as the target for removal. The method omits identification of the remaining size of fragment, where a small read could have multiple alignment positions potentially resulting in incorrect or overestimated placements, skewing results. To best approach this, it is important to also consider the location of the inverse PCR primers used in the in relation to their position within the transposon sequence. Due to the low complexity of the terminal transposon sequence PCR primers are positioned further into the sequence. When identifying sequence to remove, the terminal sequence is needed rather than the primer sequence. This ensures the correct region is identified and the remainder of the sequence to be copied is the desired bacterial chromosome. The minimum and maximum remaining read length is important so alignments are unequivocal. This was achieved by a random selection of different length reads (15-30bp) being aligned until ambiguous reads were no longer present, increasing confidence

in the data.

The insertional mutagen pGhost9::IS*S1* introduces an 8bp repeat when replicative transposition occurs (Polzin & Shimizu-Kadota 1987). The importance to take into account these 8bp regions cannot be overlooked as this could potentially falsely inflate the number of insertions deemed unique. In reality, achieving this collapse does introduce a limitation of the bioinformatics. Every insertion 8bp apart might not be the same transposition event and even if they are the collapsed might not be in the right direction. There is no method to distinguish this correctly, so a consistent approach was used to collapse all insertions 8bp apart to the lowest genome coordinate. This reduces unique insertion numbers but increases the confidence in the final data.

Chapter 4 Determination of essential genes in the *Streptococcus uberis* 0140J genome

4.1. Introduction

Essential genes are defined as those in which the absence of the gene product would be lethal to the organism (Mushegian & Koonin 1996). Even the simplest cells have a complex array of functional systems that are necessary to support life (Koonin 2003). The notion of an essential genome has its origins in synthetic biology, where the elucidation of a common set of genes would allow for the production of a minimal autonomous cell (Gibson et al. 2010). The phrase “essential genome” is often used as a substitution for “essential genes” however the two are not interchangeable. Essential genomes consist of a set of essential genes, but are those necessary for growth in a defined environment, such as laboratory medium or within a host. This may differ to the smallest complement of genes to enable independent growth (Gibson et al. 2010).

The movement to high-throughput transposon insertion sequencing (TIS) experiments has led to additional factors being considered when defining an essential gene, making the definition of a truly essential gene both a complex and subjective process. As numerous mutations are assessed simultaneously the abundance of any given mutation is represented by sequence read depth. Assuming the rate of sequencing is equal across every strand of DNA, this can

be directly associated with the frequency a unique insertion occurs within the population. However, if the base calls of the sequenced DNA are incorrect, then sequence reads could align to an alternative location within the reference genome, producing the identification of a false insertion point. To avoid this, the implementation of an insertion read depth cut-off can be used for data analysis. The level of which to apply this cut off is speculative, but identifying a single insertion $\geq$3 times suggests it is not a spurious sequencing artefact that has been amplified by chance (Klein et al. 2012).

Other determined cut-offs for TIS experiments have been formulated to enhance the confidence of the generated data. For example any CDS completely lacking insertion elements within the given population are the first to be categorised as essential (van Opijnen et al. 2009). However, it has been postulated that many genes can tolerate insertions in terminal proportions of their sequence (Goodman et al. 2009; Griffin et al. 2011; Zomer et al. 2012) or outside of certain regions of functional domains (Zhang et al. 2012). Another positional effect can be based around alternative start codons that are able to produce shorter but still functional gene products, or alternatively, the loss of amino acids from the C-terminus of the protein having no deleterious effect on the protein function (Jacobs et al. 2003). This truncation has led to mutations in the initial and terminal 10% of the CDS being disregarded (Gallagher et al. 2011) and also CDS that have less than 3 unique insertions within the central 80% of its coding sequence are also omitted so they are categorised as essential (Goodman et al. 2009).

Analysis of an essential gene in isolation does not provide any greater understanding on how a bacterium can overcome host immune systems in order to colonise and cause infection or disease. A more comprehensive analysis of essential genes can be gained by investigating interactions within the context of metabolic pathways. A metabolic pathway is a cascade of linked biochemical reactions. The product of an initial reaction is then the reactant or catalyst of the subsequent reaction (Karp 2001). There have been numerous metabolic pathway databases created to hold information on multiple bacterial species for functional comparison. Some are no longer maintained like PUMA2, which generated a pathway based on homology to known protein sequences (Maltsev et al. 2006). Others are organism specific, like the EcoCyc database and are based on the functions of individual gene products of *E.coli,* identifying their regulation and organisation into operons (Keseler et al. 2011). The largest and the most frequently updated database that contains information from multiple species is the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Ogata et al. 1999). The KEGG database was developed for systematic analysis of gene functions in terms of networks of genes and molecules. The KEGG database contains graphical representations of known metabolic pathways and regulatory pathways. The information within the KEGG database is generated based on enzyme commission (EC) numbers from orthologous and paralogous gene groups (Kanehisa & Goto 2000). A combined approach called rapid annotation of microbial genomes using subsystems technology (RAST), mines data from various online repositories, such as searching literature databases and gathering annotations

from KEGG (Kanehisa & Goto 2000), GenBank (Benson et al. 2013) and Swiss-Prot (UniProt 2013). This provides a method of grouping genes with their functionally related protein families (Overbeek et al. 2014), allowing for a broader identification of the function of the gene.

As an extension to the metabolic pathway databases, online repositories have been developed, focusing on data generated by various TIS experiments. The Essential Genes on a Genomic Scale (EGGS) (McNeil et al. 2007), the Online Gene Essentiality database (OGEE) (Chen et al. 2012) and the Database of Essential Genes (DEG 1.0) (Zhang et al. 2004), all hold records of genes identified as essential or non-essential from various disruptive mutagenesis experiments. The National Microbial Pathogen Data Resource Centre (NMPDR) discontinued the EGGS database in 2009, and is now maintained by the Pathosystems Resource Integration Centre (PATRIC). This allowed for integration of the TIS data into other online comparative analysis pipelines, however the data is not displayed in a particularly user-friendly manner for searching and data comparison and some tools are only available with paid subscriptions. The OGEE database is maintained by the European Molecular Biology Laboratory (EMBL) and has additional datasets that are generated by data mining the National Centre for Biotechnology Information (NCBI) for gene specific knockout experiments. Finally, the DEG database is the most comprehensive database available, and includes data on yeast, archaea, prokaryotes and eukaryotes. It is regularly maintained and updated by the Centre of Bioinformatics at the Tianjin University to incorporate additional

data (DEG 5.0) (Zhang & Lin 2009) and comparative tools (DEG 10) (Luo et al. 2014) to allow for specific, environmentally conditional, essential genes to be identified and the ability to directly perform a BLAST search against the database. This allows for custom searches to identify homologous essential genes and compare the conservation across multiple organisms using amino acid or nucleotide sequences.

The data in these repositories is limited when investigating Gram-positive bacteria, especially streptococci. *Streptococcus sanguinis (S. sanguinis)* (Xu et al. 2011) and *Streptococcus pneumoniae* (*S. pneumoniae*) (van Opijnen et al. 2009; Thanassi et al. 2002; Song et al. 2005) are the only streptococci present within the essential gene databases with *Bacillus subtilis* (*B. subtilis*) (Kobayashi et al. 2003), and *Staphylococcus aureus* (*S. aureus*) (Christiansen et al. 2014; Bae et al. 2004; Forsyth et al. 2002) being the only other Gram-positive bacteria investigated.

4.2. Results and discussion

4.2.1. Analysis of the Streptococcus uberis mutant insertion pools

The pools of mutant *Streptococcus uberis* (*S. uberis*) were prepared for sequencing following the acoustic fragmentation and inverse PCR enrichment. To assess the reproducibility of the method, the quantities of DNA were recorded at each stage of the protocol (Figure 4.1) (Table 4.1).

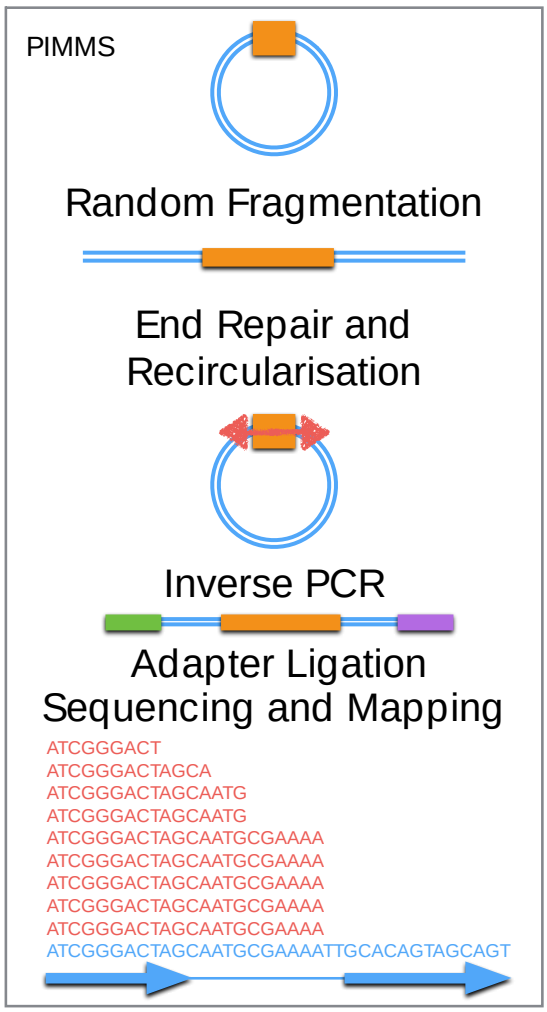

Figure 4.1. Flowchart demonstrating the laboratory procedure for amplification of unknown regions of DNA using inverse PCR. The extracted DNA is randomly fragmented using Covaris adaptive focused acoustics. These fragments are treated to repair the ends and recircularised to create a PCR template. The primer pair used in the reaction amplifies from the two terminal ends of the IS*S1* element in an outwards direction. The PCR product is then prepared for sequencing using Illumina library preparation reagents.

Each pool contained approximately 10,000 individual *S. uberis* mutants (range 7800-16000), estimated by bacterial colony counts. Subsequent sequence analysis and mapping identified an average of 11,549 unique insertions per pool (range 6004-19561), with an average read depth of 98 (range 73-204).

The sequence data from all 10 pools was concatenated and analysed allowing for approximately 100,000 mutants of *S. uberis* to be analysed. Using an implemented cut-off where only insertions with ≥3 occurrences are considered as 'true' insertional events (Klein et al. 2012) a total of 80,617 unique insertions throughout the genome were identified, within this population (Table 4.2). This equates to a mutation every 15bp with an average read depth of 182 reads and an average of 30 insertions per CDS, meaning any regions of the genome greater than 15bp, devoid of mutant insertion points, can be considered as important for the growth of *S. uberis* on nutrient rich solid media.

Table 4.1. Quantity of DNA at each stage of the enrichment process. All samples were quantified using Qubit dsDNA broad range reagents.

Pool Number	Extraction (ng)	Fragmentation (ng)[1]	Following End Repair (ng)[2]	Circularisation (ng)[3]	PCR Amplification (ng)[4]
1	4505	1836	1220	459	1740
3	4385	2150	1485	549	1865
5	3550	1742	1395	609	1995
6	3915	1640	1410	711	2065
8	4665	1844	1640	477	2235
10	4965	2490	1185	534	1815
11	4805	1946	1255	597	3020
12	3910	1810	1350	573	2340
13	4770	1836	1180	567	2150
14	3690	1895	1470	558	2410
Average	4316	1918	1359	563	2163
SD	535	253	150	63	363

DNA quantities used in each step; [1] 2000ng, [2] 1500ng, [3] 1000ng, [4] 100ng PCR template

Table 4.2. Metrics of analysed sequencing data Values represent the number of sequence reads that are associated with every stage of the bioinformatics pipeline for each pool and the corresponding percentage.

Pool Number	Total number of Reads	Matched Reads (%)	Mapped Reads (%)	Unique insertions with ≥3 occurrences	Average Read Depth	Number of CDS lacking insertions*
1	8953812	4437287 (49.6)	2178162 (49)	10647	204	574
3	7861879	2426829 (30.9)	1071066 (44)	13669	78	473
5	8277653	4183217 (50.5)	1335140 (32)	15287	87	450
6	8115818	3003635 (37)	1153172 (38)	15611	73	446
8	8268603	3672825 (44.4)	176272 (4.8)	6004	28	648
10	6761822	1942516 (28.7)	797029 (41)	7356	108	695
11	8636714	2446844 (28.3)	836874 (34)	8856	94	596
12	6478983	1795475 (27.7)	881896 (49)	9719	90	618
13	7266178	2187573 (30.1)	961217 (43.9)	8780	109	650
14	8526696	4220350 (49.5)	2143901 (50.7)	19561	109	375
Pooled	79148158	30316551 (38.3)	11534729 (38)	80617	182	199

* Total coding sequences in *S. uberis* 0140J genome (Accession number AM946015) (Ward et al. 2001) of which, 1826 code for proteins, 15 code for Ribosomal RNA and 20 code for tRNA.

As transposon site bias has been deemed a limiting factor in TIS experiments (Langridge et al. 2009), base composition was assessed at each insertion site, and was compared to the base composition within the AT rich (63.4%) *S. uberis* genome. The base composition of unique insertions was calculated at 57.7% AT, indicating a slight preference towards insertion within GC rich regions of the genome. However, nucleotide sequences surrounding the insertion sites found no overrepresentation of a specific insertion motif. To ensure the overall frequency was evenly distributed, the coordinates of each transposon insertion site were used to generate a graphical representation of their position with the *S. uberis* genome (Figure 4.2.). This indicated that there are distinct regions within the genome in which the transposon did not integrate, possibly relating to coding sequences of importance to cellular function.

The distribution of the centile positions for every unique insertion point was used to assess any positional insertional bias within individual CDS (Figure 4.3.). This indicated a slight preference of pGhost9:IS*S1* insertion towards each end of the CDS, and this could be related to the positional effect of insertion, as previously discussed (van Opijnen et al. 2009). For example, if an individual CDS can tolerate insertions within the last 10% and still produce a functional product, then the identification of an increased number of insertions within such a region across essential and non-essential CDSs would be expected.

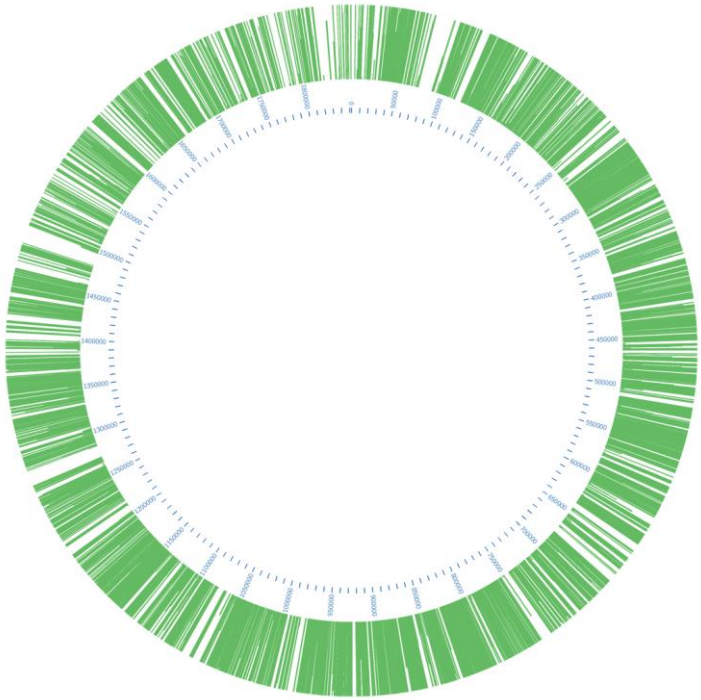

Figure 4.2. Circos representation of each insertion site identified within the *S. uberis* genome. The coordinate file was used to generate a Circos plot (Krzywinski et al. 2009) of 63,187 unique insertion points identified in the *S. uberis* genome.

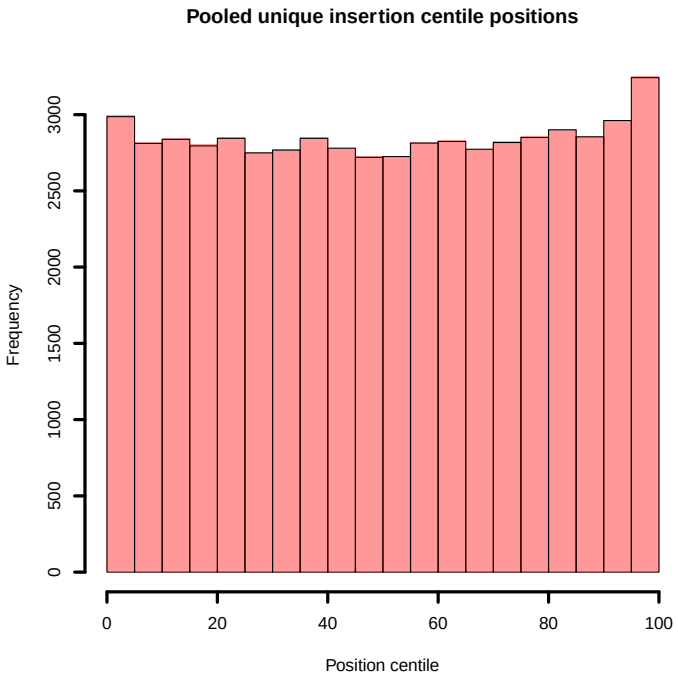

Figure 4.3. Centile position of every insertion within all CDS. The centile location for every insertion was calculated and the frequency of this position was assessed within each coding sequence.

4.2.3. Identification of S. uberis Essential genes

For simplicity in this chapter, essential genes will be defined as a CDS completely devoid of insertional events throughout its entire length. Any CDSs with insertions present in the final 10%, with a read depth of ≥3, will be referred to as truncated.

Whole genome analysis using sequence data generated by all 10 pools of mutants identified a total of 196 essential genes (Table 4.3). An additional 67 truncated genes were identified having insertions only within the C-terminal 10% of the CDS (Table 4.4).

When analysing the data generated from the *S. uberis* TIS experiment, it was noted that a specific CDS (M42 glutamyl aminopeptidase), a gene that is required for growth of *Lactococcus lactis* in milk, but is not deemed essential (L'Anson et al. 1995), had an abnormally high read depth in *S. uberis.* The occurrence (534811) was twice the amount when compared to the next most frequently hit CDS, SUB0885 (201192). The incidence of an overrepresented insertion could be associated with a insertion recognition sequence or a molecular preference site resulting in a ratio of insertions at this point, however after investigation the gene is mutated every 8bp amassing 137 unique mutations throughout the 1052bp CDS.

4.2.3.1. Functional categorisation of the essential S. uberis genes

Characterisation of the essential CDS was performed using RAST (Rapid Annotation using Subsystem Technology) (Overbeek et al. 2014). This allocates

genes into categories based on protein families and related functions. Unsurprisingly, the majority of essential CDSs were associated with transcription and translation with the remainder related to other basic cellular functions including those involved in catabolism and cell wall and capsule formation (Table 4.3 and 4.4).

A number of essential genes were associated with a role in virulence in other bacterial species. SUB0552 (*vicR*), part of the *vic* operon; consisting of *vicR*, a regulator protein, *vicK* a histidine kinase and *vicX* a putative mismatch repair exonuclease, was the only CDSs within the operon that was found to be essential. The *vic* operon has been well characterised in *Streptococcus pneumoniae* (*S. pneumoniae*) (Wagner & Saizieu 2002) and is part of a two-component regulatory system that regulates the cell cycle, allowing the bacterium to detect a favourable environment for growth (Winkler & Hoch 2008). This operon also has a promoter region prior to the *vicR* CDS. This 82bp sequence also had 5 unique insertions represented 135 times. The data generated for *S. uberis* supported the observation in *S. pneumoniae* where single gene deletions of *vicK* and *vicX,* were found to be non-essential in vivo. Interestingly *vic* operon mutants, with an overexpressed *vicR* gene, have been shown to have a significant decrease in virulence in mouse models (Wagner & Saizieu 2002).

Table 4.3. Essential coding sequences and their associated RAST categories Each protein coding sequence that was deemed essential for growth in *S. uberis*. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

RAST Category	Count	Gene
Amino Acids and Derivatives	2	*aspS, alr*
Carbohydrates	12	*fba, pgk, plr, gpmA, acoL, pfk, pdhC, pdhB, ptsH, gapN, pgi, gpsA*
Cell Division and Cell Cycle	8	*ftsA recU, ftsL, ftsZ*, SUB1127, 1285, 1404, 1092
Cell Wall and Capsule	14	*glr, dltC, murF, ddlA, murG, rmlB, murE, pbpX, glmS, rmlA, rlmC*, SUB0010, 0696, 0697
Cofactors, Vitamins, Prosthetic Groups	4	*ppnK, dpfB, birA*, SUB0641
DNA Metabolism	11	*mecA, dnaI, dnaG, hlpA, ssb, dnaC, parE, dnaH, plsX, xseB*, SUB1777
Fatty Acids, Lipids, and Isoprenoids	18	*aacA, aacD, aacC, fabZ, uppS, dgk, mvaD, fni, fabD, fabK, acpP, fabH, acpS, mvaS, fabE, fabF,* SUB1015, 1501
Hypothetical Proteins	9	*veg,* SUB0149, 0332, 0388, 0726, 0930, 1286, 1547, 1619
Iron acquisition and metabolism	0	-
Membrane Transport	6	SUB0581, 1004, 1158, 1413, 1799, 1853
Miscellaneous	11	*prsA1, secG, ftsE*, SUB0223, 0382, 0399, 1093, 1472, 1620, 1732, 1775
Nucleosides and Nucleotides	6	*prsA2, tmk, nrdH, adk, ybeY* SUB1225
Phosphorus Metabolism	0	-
Protein Metabolism	58	*alaS, engA, engC, fus, gatA, gatB, gltX, groES, infA, infB, infC, pheS, prfB, rimM, rplB, rplD, rplD, rplE, rplF, rplJ, rplK, rplK, rplN, rplO, rplP, rplR, rplS, rplU, rplV, rplW, rplX, rpmA, rpmB, rpmC, rpmD, rpme, rpmF, rpml, rpmJ, rpsB, rspC, rpsD, rpsE, rpsF, rpsG, rpsH, rpsI, rpsJ, rpsK, rpsL, rpsM, rpsN, rpsO, rpsQ, rpsS, rpsU, tufA*, SUB1008, 1732A
Regulation and Cell signalling	1	*hisS*
Respiration	7	*atpB, atpD, atpE, atpF, atpA, atpG, atpH*
RNA Metabolism	13	*leuS, asnS, glyS, proS, rnpA, era, trmD,* SUB0849, 1470, 1467, 1616, 1618, 1847
Stress Response	3	*dnaJ, grpE*, SUB0009
Virulence, Disease and Defence	4	*vicR*, SUB0502, 0505, 0506
Total	196	

Table 4.4 Truncated coding sequences and their associated RAST category Each protein coding sequence that was deemed truncated be insertions within the terminal 10% of the encoded protein. Each coding sequence was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

RAST Category	Count	Gene
Amino Acids and Derivatives	2	*cysE, mtnN*
Carbohydrates	2	*ptsK, tpi*
Cell Division and Cell Cycle	2	*ftsW, ftsX*
Cell Wall and Capsule	10	*bacA, gcaD, hasC, mraY, murB, murC, murM, rgpG*, SUB426, 0700
Cofactors, Vitamins, Prosthetic Groups	7	*coaA, coaD, coaC, dyr, metK, nadE*, SUB0356
DNA Metabolism	5	*dnaA, dnaN, holB, parC, gyrA*
Fatty Acids, Lipids, and Isoprenoids	5	*mvaA, mvaK2*, SUB253, 0333, 1246
Hypothetical Proteins	2	SUB1434, 1834
Iron acquisition and metabolism	0	-
Membrane Transport	5	*secA,* SUB0511, 1005, 1852, 1854
Miscellaneous	4	*tyrS,* SUB0019, 0393, 0704
Nucleosides and Nucleotides	5	*gmk, pgmA, thiD*, SUB0745, 1227
Phosphorus Metabolism	1	*ppaC*
Protein Metabolism	11	*argS, efp, prfA, pth, rplL, rplM, rplQ, serS, thrS, trsA*, SUB0345
Regulation and Cell signalling	0	-
Respiration	0	-
RNA Metabolism	6	*ileS, nusA, rpoD*, SUB0013, 0873, 1479
Stress Response	0	-
Virulence, Disease and Defence	0	-
Total	67	

There are three other essential genes associated with virulence, disease and defence, which encode for proteinaceous antimicrobial substances (bacteriocins). Some strains of *S. uberis* are prolific producers of bacteriocins, including ubericin, uberolysin and nisin U (Heng et al. 2007). Bacteriocins are novel inhibitory agents, which are active against other bacterial species. These are considered to offer a growth advantage for *S. uberis* within the host environment, by reducing competition for nutrients. The *ubl* operon, which is associated with uberolysin production, contains five genes. In these data the *ubl* operon contains a large number of insertions indicating this is not essential for growth under these conditions. The nisin U operon, present in some strains of *S. uberis* is absent in the 0140J experimental strain (Wirawan et al. 2006). The genomic region incorporating the CDS SUB0500 – SUB0516 encodes multiple bacteriocins and associated immunity proteins (Ward et al. 2009). This contains the three deemed essential in the current investigation (SUB0502, 505 and 506) and SUB0512 (ubericin) CDSs. These CDSs are relatively small, at 129, 267, 201 and 228bp respectively.

It is also interesting to note that there is a Phage protein deemed essential from this data. The coding sequence SUB0223, has no mutations detected even though it is approximately 200bp in length. This CDS is also conserved within numerous streptococci suggesting some importance greater than redundant regions of genome, potentially inferring some immunity against phage infection.

Those genes previously described in literature as being putative *S. uberis* virulence associated genes (*mtuA, pauA, OppA1, hasABC, srtA and vru*) were all found to have numerous insertions present. Only *hasC*, which has been identified as a capsule formation protein was categorised as truncated. However strains with mutations mapped within the *hasC* CDS have been used in other studies to identify capsule deficient mutants (Ward et al. 2001). In the previous study esculin blood agar was used as the growth medium, possibly indicating *hasC* is not essential during growth on this medium.

4.2.3.2. Categorisation of the hypothetical and putative essential genes

Some essential genes required further investigation, especially those with the miscellaneous RAST categorisation. The absence of classification may reflect their association with multiple metabolic pathways, or as the *S. uberis* genome has received relatively little attention with respect to characterisation of gene function, there is a possibility some annotations were missing.

Coding sequences that were categorised as miscellaneous by RAST were assessed from sequences in the NCBI gene database (Maglott 2004). Of these CDS, three were integral to multiple metabolic pathways and therefore correctly assigned to the miscellaneous pathway category. SUB0019 (amidase) is a hydrolase that acts on acrylamides, SUB0113 (thioredoxin) is an antioxidant that is associated with multiple cellular mechanisms and SUB0223 a phage protein, which could provide some immunity to bacteriophage

infection.

Most of the remaining miscellaneous CDSs were actually found to have a defined role. The *prsA1* CDS is associated with membrane transport as its folds proteins (Ma et al. 2006), however as it is shorter than some orthologous genes, the *S. uberis* specific *prsA1* may not be assigned during automated category grouping techniques. Other CDSs, such as *secA, secG*, SUB1732 (*secE*) and SUB0399 are responsible for secretion (Driessen et al. 1998), with *ftsE* and SUB0282 associated with cellular division (Corbin et al. 2007). SUB1093, was identified as a mur-ligase responsible for peptidoglycan synthesis and therefore associated with structural integrity of the cell wall (Longenecker et al. 2005). Mur ligase have recently been identified as a potential antimicrobial target, because they are functionally essential and highly conserved across bacteria without a eukaryotic counterpart (Kouidmi et al. 2014). SUB1620 works in conjunction with SUB0490 (*ybeY*) and was previously thought to be part of the phosphorus metabolism pathway, but more recently has been found to function as an endoribonuclease (Grinwald & Ron 2013), and both CDSs are responsible for the degradation of RNA. SUB1775 is a holiday junction resolvase, which could be associated with cell cycle as it coils DNA. SUB1472 is a phosphotransferase, which if grouped with other CDSs in the PTS system is most likely associated with central carbohydrate metabolism.

To try to identify the function of the putative or hypothetical proteins, the amino acid sequences were used in a position-specific iterated (PSI) BLAST

search (Altschul et al. 1997). This method generates a matrix of sequence motifs, which allows for multiple alignments within a sequence. This is a more sensitive approach than the traditional BLAST searches and identifies potential ancestral homology. A series of 5 essential CDSs, SUB0223, SUB0581, SUB1158, SUB1413 and SUB1799, were found to be uncharacterised membrane proteins, whilst SUB0332 and SUB1547 were stress response proteins and SUB1286 was identified as a putative cell division protein. SUB0641 is a flavodoxin protein, which is responsible for electron transfer and plays a role in multiple metabolic pathways, however, its function as a cofactor in oxidative stress protection has been recently suggested (Lindhoud et al. 2012). SUB1470 is a ribosomal maturation protein, which is part of the *nusA* operon (Shazand et al. 1993). The *nusA* operon also includes essential genes SUB1466 and SUB1467, which interact with the S10 ribosomal proteins to pause RNA transcription termination (Shunsuke et al. 1984). SUB1501 was identified as a transcriptional regulator protein. SUB0697 was a putative ABC transporter permease which could play a part in importing essential nutrients into the cell (Davidson et al. 2008). The *veg* gene (SUB1842) encodes a small 86kDa cytoplasmic protein; highly conserved across many bacteria, but with an unknown function (Fukushima et al. 2003). SUB0388 was a cystathionine γ-synthase which is a catalyst for the production of cystathionine from homoserine and cysteine (Ravanel et al. 1998) and SUB0930 was found to share 98% homology with a DNA binding protein found in *Bacillus subtilis* (Marciniak et al. 2012). There were four CDSs that did not return a result from the PSI-BLAST, SUB0502, SUB0505 and SUB0506 are the *S. uberis* specific

bacteriocins, as previously described and SUB0149, a 139bp protein, which may well be a *S. uberis* specific gene.

4.2.3.3. Homology of the essential genes of S. uberis

The list of essential genes was used to query the DEG 10 database (Luo et al. 2014) to ascertain the homology of the identified essential genes within other bacterial species. The amino acid sequences were parsed against the existing collection of essential genes from 34 other bacteria. Of the 187 genes essential in *S. uberis*, 158 (85%) were homologous to other genes essential in other bacterial species. The remaining 28 CDSs did not return any homologous matches (Table 4.5). From the 28 CDSs there were 10 genes that did not show sequence homology but have a gene name (Table 4.5). Searching the DEG 10 database by gene name identified 6 more genes with the same function but different sequence. The genes *ftsL*, *dgk*, *pheS, infB*, *atpH* and *grpE* were all found to have an essential homologue within other bacterial species. This occurrence could be potentially due to errors in automated annotation or the unintentional duplication of names being used to describe CDSs with different sequences. Examination of these CDSs showed that *dgk*, *pheS, infB*, *atpH* and *grpE,* are multi-domain proteins. These CDSs consist of small functional regions of DNA, separated by highly variable sequences in different bacteria, meaning each species has a unique CDS with the same function (Vogel et al. 2004). This explains the lack of homology at an amino acid level, and after parsing the DEG 10 database by gene identifier 27 different bacteria were

found to have the same CDS. The final gene, *ftsL*, encodes a small septum formation protein that is an essential part of the cell division machinery (Villanelo et al. 2011). Its amino acid sequence is low in complexity but highly divergent, which accounts for the lack of a homologous match to other bacteria in the database. Similarly when the *ftsL* identifier was used in the DEG 10 database another 13 species of bacteria have *ftsL* listed as an essential gene. Genes *dltC, pbpX, mecA* and *veg* were identified in the non-essential gene database, suggesting their essential role in bacterial growth is unique to *S. uberis* or the laboratory environment in which the pools of mutants where grown.

Table 4.5 Non-homologous essential genes of *S. uberis* Essential genes that did not have a homologous match within DEG 10 (http://www.essentialgene.org). Any gene that was identified as either hypothetical or putative was parsed against the NCBI database to identify any similarities between other bacteria.

Identifier	Name	Product	Subsystem
SUB0149	NA	hypothetical protein	Hypothetical Protein
SUB0223	NA	putative membrane protein	Membrane Transport
SUB0332	NA	putative alkaline shock protein	Stress Response
SUB0340	*mecA*	adapter protein MecA	DNA Metabolism
SUB0388	NA	putative cystathionine gamma synthase	Protein Metabolism
SUB0491	*dgk*	diacylglycerol kinase	Fatty Acids, Lipids, and Isoprenoids
SUB0502	NA	putative bacteriocin	Virulence, Disease and Defence
SUB0505	NA	putative bacteriocin	Virulence, Disease and Defence
SUB0506	NA	putative bacteriocin	Virulence, Disease and Defence
SUB0581	NA	putative membrane protein	Membrane Transport
SUB0641	NA	putative flavodoxin	Cofactors, Vitamins, Prosthetic Groups
SUB0669	*atpH*	putative ATP synthase delta chain	Respiration
SUB0678	*pheS*	phenylalanyl-tRNA synthetase	Protein Metabolism
SUB0697	NA	putative ABC transporter permease protein	Cell Wall and Capsule
SUB0930	NA	putative DNA binding protein	DNA Metabolism
SUB1147	*dltC*	D-alanyl carrier protein	Cell Wall and Capsule
SUB1158	NA	putative membrane protein	Membrane Transport
SUB1286	NA	putative cell division protein	Cell Division and Cell Cycle
SUB1413	NA	putative membrane protein	Membrane Transport
SUB1419	*pbpX*	putative penicillin binding protein	Cell Wall and Capsule
SUB1420	*ftsL*	putative cell division protein	Cell Division and Cell Cycle
SUB1466	*infB*	translation initiation factor IF-2	Protein Metabolism
SUB1470	*ylxS*	putative ribosomal maturation protein	RNA Metabolism
SUB1501	NA	MarR family regulatory protein	DNA Metabolism
SUB1505	*grpE*	GrpE protein (HSP-70 cofactor)	Stress Response
SUB1547	NA	putative alkaline shock protein	Stress Response
SUB1799	NA	putative membrane protein	Membrane Transport
SUB1842	veg	conserved hypothetical protein	Miscellaneous

4.2.4. Analysis of metabolic pathways associated with essential genes

The RAST categories (Table 4.3) were used to link into the KEGG database to allow for identification of the metabolic pathway associated with the specific essential genes. Unsurprisingly, most of the essential genes for *S. uberis* are responsible for core cellular processes such as protein, DNA and RNA metabolism, cell division and cellular integrity. Using the RAST categories, the proportion of the genes throughout the genome was used to compare the composition of the essential genes (Figure 4.4). The percentage of CDSs associated with cell division, cell wall, fatty acid biosynthesis, DNA metabolism and protein metabolism account for a greater proportion of essential genes when compared to the relative proportion throughout the genome as a whole. To assess if there was any enrichment for a specific biological RAST subsystem a chi-square analysis was performed (Table 4.6). After incorporating a p-value cut-off of ≤ 0.05, it was found that there was a positive enrichment of essential genes for those associated with protein metabolism, fatty acid biosynthesis and respiration (production of ATP). Indicating these subsystems of functionally related proteins represents a greater proportion of essential genes in comparison to their abundance in the whole genome. However membrane transport, carbohydrate utilisation and amino acid biosynthesis are less likely to be essential.

Table 4.6. Enrichment identification of essential gene RAST categories Chisq.pval = the associated p-value showing the significance of the enrichment for the given RAST category.

Category	chisq.pval
Protein Metabolism	4.11506E-34
Fatty Acids, Lipids, and Isoprenoids	8.11145E-12
Respiration	0.00086022
Membrane Transport	0.001349588
Carbohydrates	0.009775391
Amino Acids and Derivatives	0.047460902
Cell Division and Cell Cycle	0.095842391
DNA Metabolism	0.178353703
Hypothetical Phage Proteins	0.183204814
Cell Wall and Capsule	0.275706545
Regulation and Cell signalling	0.444117255
Cofactors, Vitamins, Prosthetic Groups	0.660403506
Phages, Transposable elements, Plasmids	0.760398489
Stress Response	0.764243738
Potassium metabolism	0.851059548
RNA Metabolism	0.97791306
Virulence, Disease and Defence	0.986106494
Hypothetical Proteins	1
Miscellaneous	1
Iron acquisition and metabolism	1
Phosphorus Metabolism	1
Nucleosides and Nucleotides	1
Sulphur Metabolism	1

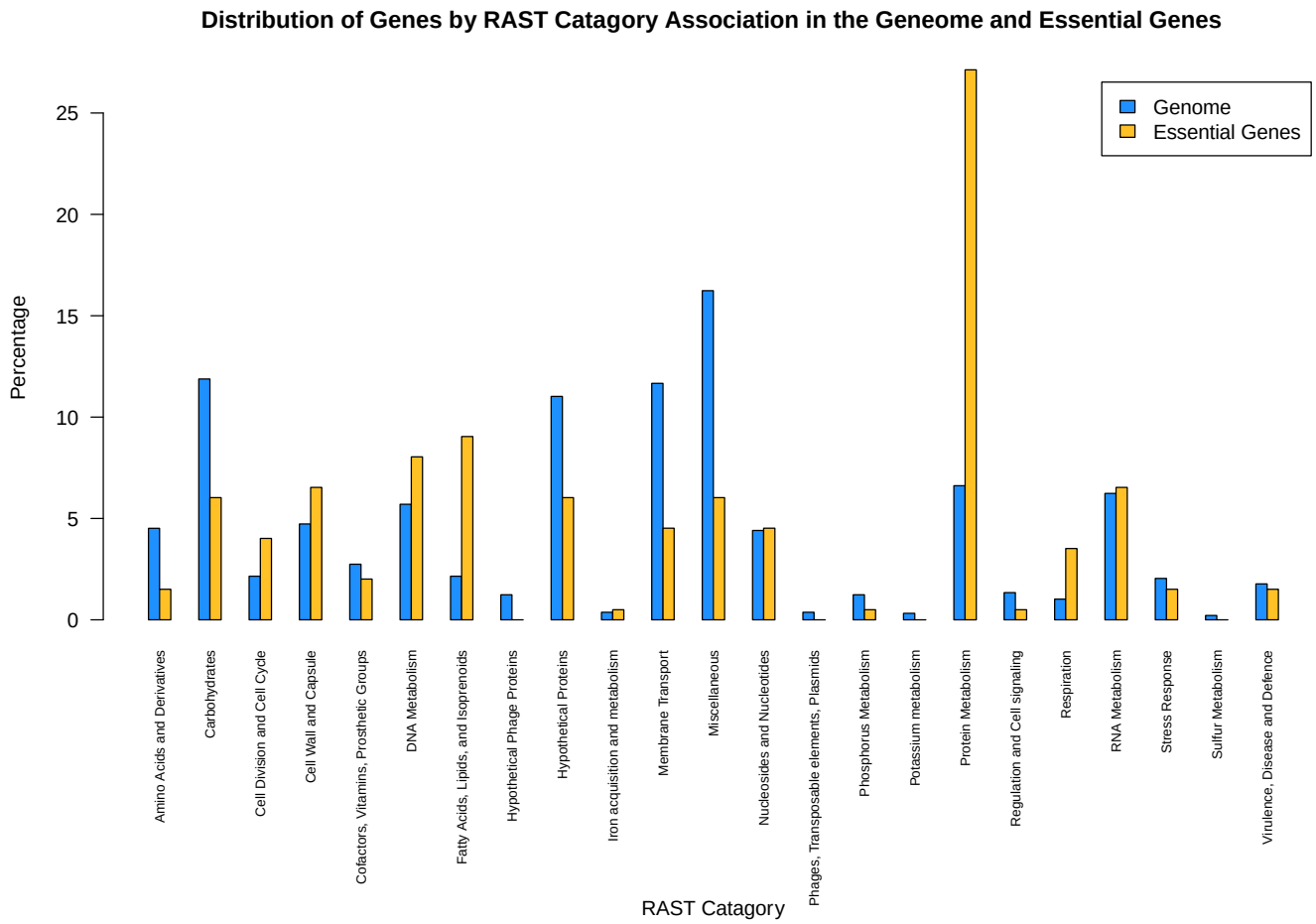

Figure 4.4. Proportional enrichment of essential genes associated with each RAST category The percentage of essential genes associated with each RAST category was calculated for the whole genome and those found to be essential for growth on THA.

The chi-squared analysis indicated a highly significant enrichment for CDS associated with fatty acid biosynthesis, with a p-value of ≤0.0001, highlighting its importance for *S. uberis* proliferation. Interestingly, genes associated with carbohydrate utilisation are less likely to be essential, however those that are appeared to be associated with the fatty acids biosynthesis category. For example the essential genes involved in carbohydrate metabolism are all part of the Embden-Meyerhof-Parnas pathway (EMP). The EMP pathway comprises of glycolysis, pentose phosphate and pyruvate metabolism pathways that feed directly into fatty acid biosynthesis through the production of acetyl-CoA. The EMP is the most common sequence of reactions in all domains of life, generating energy and metabolites for many cellular processes. This pathway plays a major role in the generation of acetyl-CoA, which is used in many biochemical reactions, such as the activation of bacterial toxins (Issartel et al. 1991) but is primarily required used in the synthesis of fatty acids.

Fatty acid synthesis is a central conserved process in both eukaryotic and prokaryotic cellular function and can be divided into two principle classes. Fatty acid synthase I (FASI), found predominantly in eukaryotes and the corynebacterium-mycobacterium-nocarid (CMN) group of bacteria, uses multifunctional polypeptides to produce palmitic acid, the precursor to longer chain fatty acids (Jenke-Kodama et al. 2005). FASII is present in all other bacterial species and covalently binds fatty acid intermediates via an acyl coenzyme A (acyl-CoA) primer and malonyl-CoA chain extender (Kaneda 1991). The essential nature of this pathway is well documented and used as a

target for control of Gram-positive bacteria resistant to some antibiotics (Wang et al. 2007). It has been show that numerous streptococci, enterococci, pneumococci and staphylococci with inactivated *fabF* or *fabH* genes are unable to grow on standard Todd Hewitt agar, however with a supplementation of a source of exogenous fatty acids, normal growth is restored, confirming the essential nature of the pathway (Brinster et al. 2009).

The gluconeogenesis pathway is one of the main mechanisms used to source glucose, which is achieved by conversion of non-carbohydrate carbon substrates. From these data, it appears that gluconeogenesis is not essential for growth of *S. uberis in vitro* (Figure 4.5 A). This is potentially due to the presence of extracellular glucose as the main carbohydrate source within Todd Hewitt agar. Glucose can be imported into the cell directly by one of three transporter systems; PTS-mannose (SUB0045), PTS-cellobiose (SUB0195) or by a non-PTS permease *glcU*, which is a unique transporter identified only in low GC Gram-positive bacteria (Lorca et al. 2007; Castro et al. 2009). The original genome annotation of *S. uberis* did not indicate the presence of *glcU*, however analysis of the amino acid sequence from the (somewhat related) Gram positive bacteria, *Staphylococcus xylosus*, using BLAST (Altschul et al. 1990), identified a homologous gene (SUB1861) that had been annotated as a putative sugar uptake protein. As there are three methods for glucose uptake, these systems are classified as mutually redundant. Once imported, the glucose is then converted into glyceraldehyde-3P, with the concomitant release of 2 ATP, by a succession of reactions performed by the essential genes *pgi, pfk* and *fba* (Deutscher et al. 2006).

The next stage of the EMP pathway is the generation of pyruvate. Pyruvate is a product of multiple processes outlined in Figure 4.5 C, none of which were identified as essential. The main mechanism is conversion of phosphoenolpyruvate to pyruvate by *pyk* (SUB1000), however there are alternative methods for its synthesis by either conversion of malate by SUB0825 to pyruvate or reduction of L-lactate by SUB0879 (*ldh*). This suggests as there are multiple methods for its production, and that if one was disrupted, another can provide the necessary components for generation of acetyl-CoA. Alternatively, Todd Hewitt broth, which was the media component for the solid growth agar used for the isolation of the original *S. uberis* mutant strains, contains a protein component that was derived from meat. This protein source has been shown to contain pyruvic acid (Hidalgo et al. 2013), which may provide a sufficient resource, avoiding the need for *S. uberis* to synthesise the metabolite. Interestingly, an alternative pathway has been identified in *Enterococcus faecalis (E. faecalis)*, the Entner-Doudoroff pathway (EDP), which uses 6-phosphogluconate dehydrase, 2-keto-deoxyphosphogluconate aldolase and 2-keto-3-deoxy-6-phosphogluconaye to convert glucose into pyruvate (Goddard & Sokatch 1964). The genes responsible for the EDP (*kgdAT, idnO*) are present in *S. uberis* (SUB1601, 1504 and 1605), suggesting the potential for an alternative pathway for pyruvate production.

Pyruvic acid is then converted into acetyl-CoA, for production of fatty acids (Figure 4.6). Acetyl-CoA is used to generate malonyl-CoA and malonyl-[acp] by

the addition of a HCO_3^- group by acetyl-CoA carboxylase (*accABCD*). The malonyl-CoA is converted into malonyl-[acp] by malonyl-CoA-[acp] transacylase (*fabD*), which in the first condensation reaction, catalysed by *fabH*, combines with an aceytl-CoA group forming β-oxobutyryl-[acp]. These set of reactions are repeated in a cyclic manner to extend the chain by two carbon units each time by *fabF* until the full length of the fatty acid is reached. The β-oxoacyl reductase (*fabG*) catalyses the next reaction to form β-hydroxyacyl-[acp] before the removal of a hydroxyl group by *fabZ* and conversion of the chain into a trans-2-enoyl group. The metabolism of an acetyl-[acp] into an hydoxyacetyl-[acp] by non-essential *fabG* is achievable due to the presence of paralogous genes (SUB0465 and SUB1496). There are two genomic locations for *fabG* within the *S. uberis* genome, and therefore a mutation in only one would allow the bacterium to still produce fatty acids. The final step reduces the double bond by the enoyl-reductase *fabk* producing the essential *S. uberis* fatty acids Octanoic, decanoic, dodecanoic acid and n-7 unsaturated acyl-[acp].

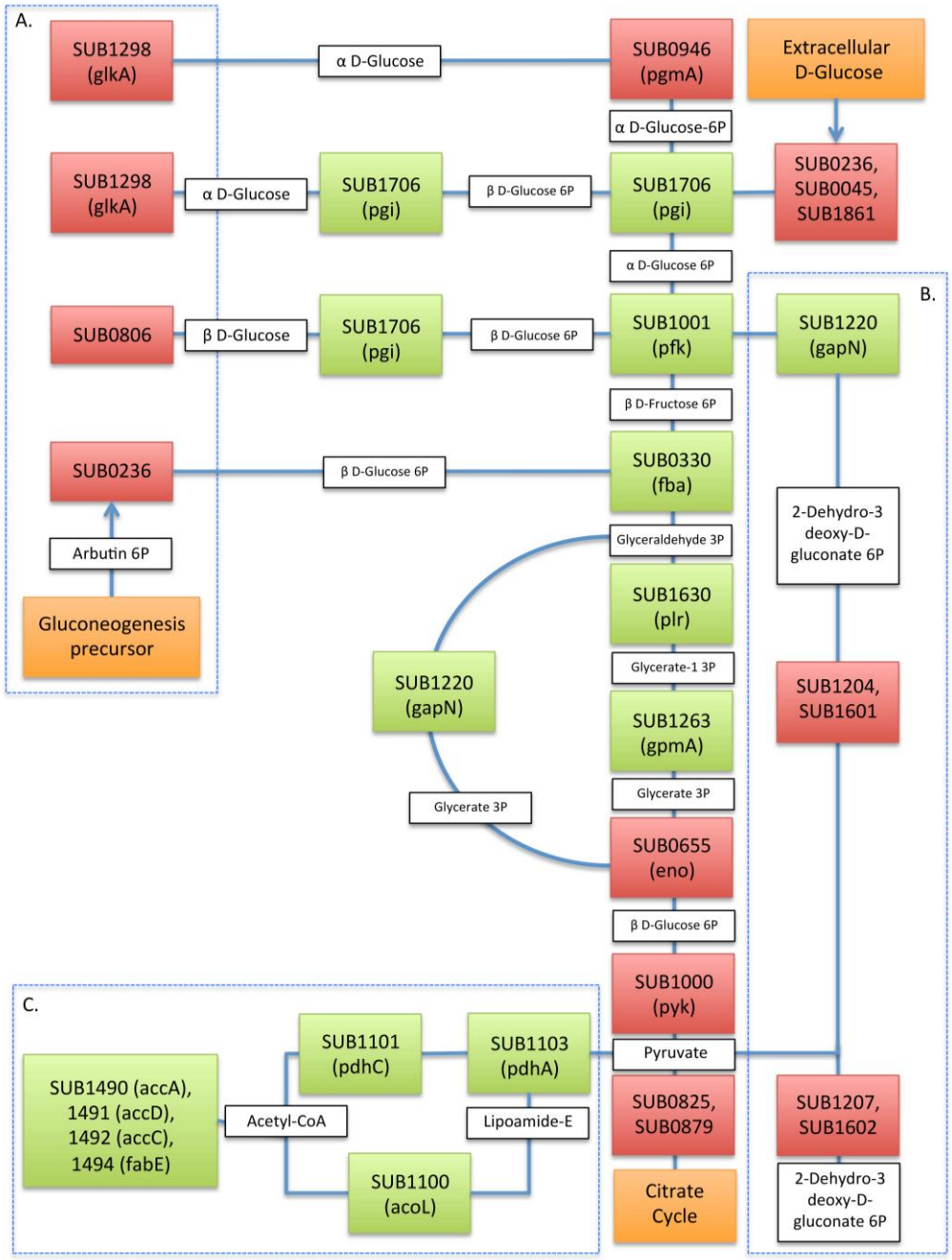

Figure 4.5. Aspects of the pathways associated with production of acetyl-CoA. Subsection A. Gluconeogenesis, B. Pentose phosphate pathway, C. Pyruvate metabolism. Red boxes represent non-essential genes, green represent essential genes, white boxes represent metabolites and orange boxes are metabolite imports from other pathways.

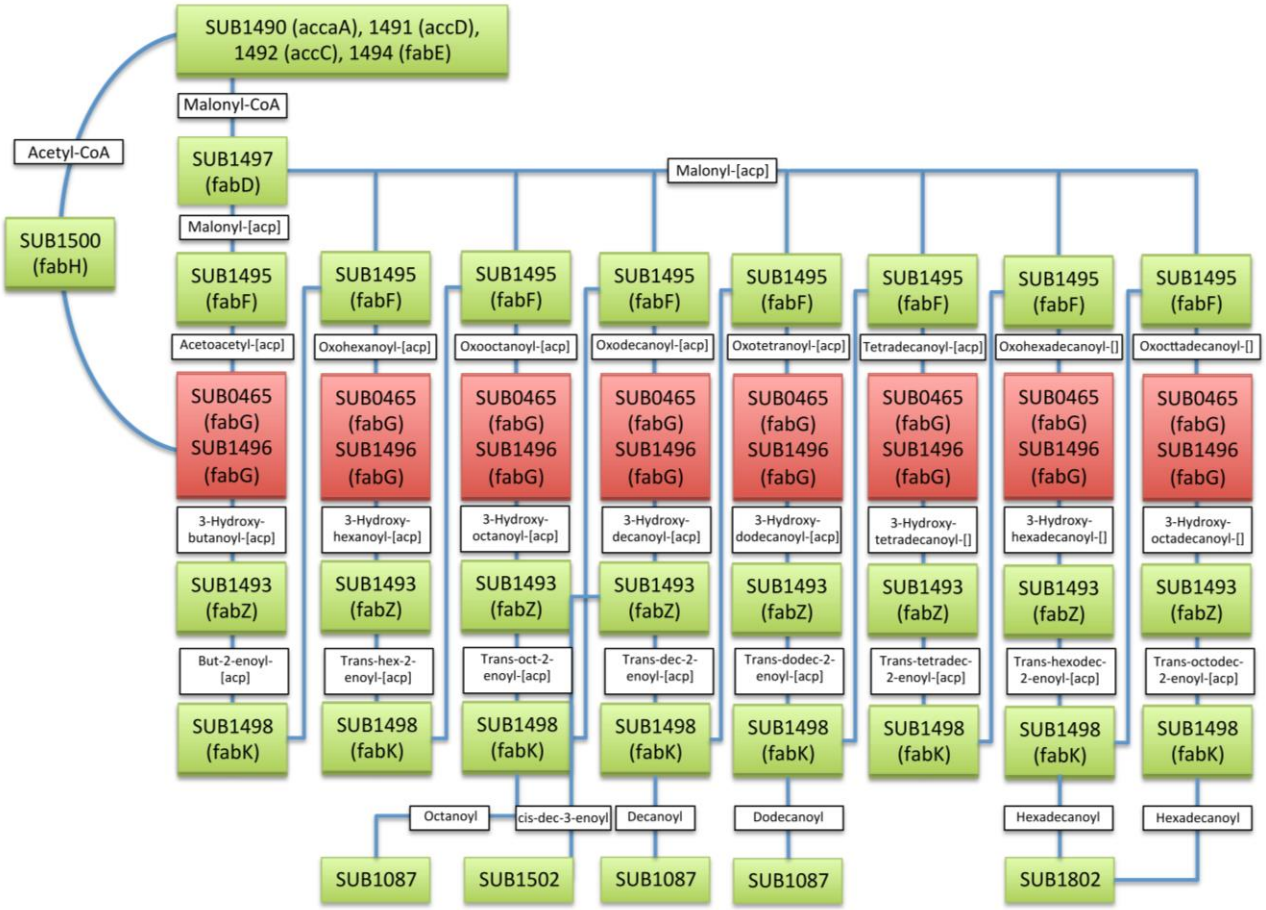

Figure 4.6 Graphical representation of the fatty acid biosynthesis metabolic pathway Red boxes represent non-essential genes, green boxes represent essential genes and white boxes represent metabolites.

4.3. Conclusions

Essential gene identification and TIS experiments have been limited in Gram-positive bacteria to studies on *Bacillus subtilis* (*B. subtilis*) (Kobayashi et al. 2003), *Streptococcus sanguinis (S. sanguinis)* (Xu et al. 2011), *Streptococcus pneumoniae* (*S. pneumoniae*) (van Opijnen et al. 2009; Thanassi et al. 2002; Song et al. 2005) and *Staphylococcus aureus* (*S. aureus*) (Christiansen et al. 2014; Bae et al. 2004; Forsyth et al. 2002). This provides a good benchmark to assess the application of the PIMMS inverse PCR protocol to identify essential genes. In order to try and evaluate the confidence in the data generated from the essential gene list, a comparison was made using the Database of Essential Genes (DEG) (Zhang et al. 2004; Zhang & Lin 2009; Luo et al. 2014). This holds a record of all currently available essential genomic elements of numerous Gram-negative bacteria and the data from the Gram-positive *B. subtilis, S. sanguinis, S. pneumoniae* and *S. aureus* studies. From the 187 essential protein-coding sequences of *S. uberis*, 152 were also found to be essential in *B. subtilis, S. sanguinis, S. pneumoniae* or *S. aureus.*

Chi-squared analysis was used to identify any enrichment of biological processes. Using a p-value score of $\leq$0.0001, four metabolic subsystems, protein metabolism, fatty acid biosynthesis and respiration, were found to be highly significant for growth of *S. uberis* in vitro. The essential genes corresponding to genetic information processing pathways were conserved in

all four Gram-positive species bacteria that have been investigated with TIS experiments.

The EDP pathway has been implicated to play a role in virulence for *Xanthomonas campestris* as a simple, energy efficient method for generating metabolic precursors for glycolysis (Lu et al. 2009). Comparison of the alternative pyruvate synthesis EDP genes, SUB1601 and 1605, against the *S. sanguinis* genome revealed two homologous hypothetical genes SSA1586 and SSA1544. The identification of these genes in both *S. uberis* and *S. sanguinis* could be an interesting avenue of investigation, as the presence of this pathway has only been shown in very few Gram-negative bacteria and one Gram-positive bacterium, *Enterococcus faecalis* (Conway 1992).

Due to the importance of the fatty acid biosynthesis and associated carbohydrate metabolism pathways in *S. uberis,* these pathways were of most interest to investigate in other Gram-positive bacteria. In *S. sanguinis and S. pneumoniae*, both the glycolysis and fatty acid synthesis pathways are also recognised as essential for bacterial growth. Alternatively, the glycolysis pathway is essential in *B. subtilis*, and *S. aureus* however the fatty acid biosynthesis pathway is not, with disruptive insertions in genes *fabEFHKZ* (Thanassi et al. 2002) having no deleterious effect on bacterial growth. Importantly, *S. aureus* and *B. subtilis* lack the acyl-[acp]-thioesterse and enoyl-CoA-hydratase, and have therefore lost the ability to covert the 2-enoyl-[acp] metabolities into the anoyl-[acp] precursors of fatty acids. These proteins are

both present in *S. uberis*, *S. sanguins* and *S. pneumoniae* showing that the requirement for octanoic, decanoic and dedecanoic acids are likely to be limited to Streptococci and not Gram-positive bacteria in general.

The major enzyme complexes present in bacterial membranes contribute to energy generation through a proton electrochemical gradient. Like *S. uberis*, *S. sanguinis* lacks a respiratory chain and uses the F-ATPase and ATP hydrolysis pathway to provide energy for nutrients and waste transport. This is an indispensable pathway, crucial to maintain cellular homeostasis, which results in the genes encoding F-ATPase and ATP-binding proteins being essential in both *S. uberis* and *S. sanguinis*. In contrast, the essential energy generation mechanism in *S. aureus* and *B. subtilis* is through the electron transport chain NADH:quinone oxidoreductase, which directly couples electron transport to the generation of a sodium gradient (Juárez & Barquera 2012). Both *B. subtilis* and *S. aureus* are likely to maintain cellular homeostasis via a Na1 / H1 transporter, and subsequent analysis shows that genes involved in this process are essential in *B. subtilis* and paralogous in *S. aureus* (Xu et al. 2011). This transporter is not present in *S. sanguinis, S. pneumoniae* or *S. uberis*.

Chapter 5 Identification of conditionally essential genes required for growth in liquid culture

5.1. Introduction

Predicting genes that are required for bacterial survival in standard laboratory media may reveal interesting aspects of the biological processes of *Streptococcus uberis* (*S. uberis*). However, directly linking these genes to a defining role in mammary gland colonisation and subsequent infection is more difficult. Genome sequencing and annotation can suggest functions of essential genes but other methods must be utilised to enable a better understanding of their role in pathogenesis. A negative selection, or phenotypic pressure, whereby an individual mutant is lost through a functional screen representing conditionally essential genes, can be used to interrogate environmental niches directly.

Until recently, phenotypic screening has been limited to the investigation of a small subset of genes based on prior knowledge of specific virulence determinants often by comparison of genes identified as important for growth in other bacterial species. Recent advances in transposon insertion sequencing (TIS) methodology provides the ability to directly assess a large number of mutant bacteria in a single experiment, enabling the identification of those strains unable to survive in a given host specific environment or biologically

relevant condition. To accurately assess the effect of the environmental screen, the experiment needs to ideally compare an original mutant population, termed 'input pool', to the same mutant population that is subjected to growth within a biological relevant condition, termed 'output pool'. Both must follow the same laboratory procedures apart from the environmental screen ensuring any essential genes identified in the output pool are a result of the environmental pressure rather than spurious absences due to the laboratory procedure.

5.1.1. Examples of transposon insertion sequencing using phenotypic screening

Signature Tagged Mutagenesis (STM) was the first high-throughput protocol to investigate the effect of a specific environmental condition on a number of transposon mutants collectively, in a single experiment (Hensel et al. 1995). STM has been used in numerous murine infection models, to identify genes associated with growth within a host. Originally used to identify genes associated with typhoid fever like symptoms, a bank of 1152 mutant *Salmonella typhimurium* were created and injected into mice. Spleen homogenates were collected three days later and plated onto laboratory media. A total of 40 input mutants were absent from the output pool; thirteen of which had previously been identified as putative virulence factors in this species. Similarly to STM, high throughput insertion tracking by deep sequencing (HITS) has also been used to successfully identify genes associated

with persistence of *Haemophilus influenzae* during pulmonary system infection in mice (Gawronski et al. 2009). Initial screens identified 358 genes that were classified as essential for viability on laboratory medium. Direct comparison of genes essential for murine lung infection identified 30 additional genes, which were implicated in bacterial survival *in vivo*. These genes were functionally diverse but mainly associated with amino acid transport and metabolism and other central biological processes (Gawronski et al. 2009).

Goodman et al. (2009) used INSeq to identify genes required by *Bacteroides thetaiotaomicron (B. thetaiotaomicron)* to establish infection in the gut of gnotobiotic mice (Goodman et al. 2009). *B. thetaiotaomicron* is a major gut symbiont, however their proportion in relation to other gut microbiota varies depending on the host. This biological screen identified which genetic mechanisms were shaped by the murine gut to enable colonisation. The input pool identified insertions within 72% (3435/4779) of the predicted coding sequences (CDSs) implying that the remaining 1344 genes were essential for growth in defined media. The output pool, after the *in vivo* screen, identified 280 additional genes essential for host colonisation. These were mostly associated with assembling polysaccharides, protein based cellular structures and acquisition and utilisation of vitamin B_{12} (Goodman et al. 2009).

Tn-Seq has also been used to identify genes involved with murine nasopharynx and lung infection by *Streptococcus pneumoniae (S.*

pneumoniae) (van Opijnen & Camilli 2012), identifying genes responsible for the bacterial response to pH stress, and the relationship between a two component system and core biosynthetic pathways (van Opijnen & Camilli 2012).

TraDIS has be used to identify genes conditionally essential for growth and survival of *Bacillus thuringiensis* in soil (Bishop et al. 2014), *E. coli* and its response to ionizing radiation (Byrne et al. 2014) and the ability of *Salmonella enterica* Serovar Typhi (*S.* Typhi) to tolerate bile and resistance to bacteriophage infection (Langridge et al. 2009; Pickard et al. 2013). However the only Gram-positive bacterium that has been investigated using this method was *Staphylococcus aureus* (*S. aureus*). This study identified 23 genes directly associated with growth in porcine blood, most of which were related to purine biosynthesis and shock response mechanisms. Seven were hypothetical genes and a further two were without a known function; highlighting an interesting subset of genes to investigate further (Christiansen et al. 2014).

Most other TIS experiments in Gram-positive bacteria have not investigated the effect of environmental pressures on essential gene subsets (Kobayashi et al. 2003; Bae et al. 2004; Forsyth et al. 2002; Xu et al. 2011), so existing comparative data is limited to *Streptococcus pyogenes*, a cause of persistent upper respiratory tract infections (Le Breton et al. 2013), *Streptococcus pneumoniae,* a causative agent of pneumonia and meningitis (van Opijnen &

Camilli 2012) and *Staphylococcus aureus* which is capable of causing infections in different sites in many different host species (Christiansen et al. 2014).

TraSH was used to identify essential genes associated with *Streptococcus pyogenes* growth in human blood (Le Breton et al. 2013). A library was generated consisting of 38,000 mutants. These were grown in human blood for 12 hours, and sampled every 4 hours. A serial analysis of all time points identified a total of 139 essential genes, largely responsible for protein metabolism, carbohydrate metabolism and replication (Le Breton et al. 2013). Tn-Seq has been used for the investigation of *S. pneumoniae* in 17 different nutrient environments including differing carbon sources or with defined stressors; reduced divalent cation concentration, DNA damage induced by methyl methanesulfonate, exposure to hydrogen peroxide, acidic pH(6), antibiotic exposure (norfloxacin) or reduced temperature (30°C) (van Opijnen & Camilli 2012). Around 6% of genes were found to be uniquely associated with increased fitness in certain situations. This poses some interesting evolutionary questions as it suggests genes are retained even though they potentially have limited function in other conditions.

5.1.2. Identification of additional essential genes associated with growth in liquid culture

The growth of bacteria on solid media may induce a selective pressure on certain individual mutant strains. During growth on solid media, bacteria in

single colonies are present at every stage of their growth, from exponential replication to cellular death (Abuaita & Withey 2011). This may impact on the identification of phase specific gene expression of particular proteins, or transcription and regulation of fitness associated antimicrobial compounds like bacteriocins (Lee et al. 2011). There are also global growth dependent factors such as nutrient abundance and inhibitory concentrations of metabolites during various growth phases, which can result in pH changes in the surrounding environment, promoting expression of stress response proteins (Klumpp et al. 2009). Therefore, growth in liquid culture may better mimic the interactions between single bacteria with the potential of identifying different signalling pathways or genes responsible for competitive fitness (Qingfang et al. 2004). Considering the impact of growth in liquid culture can also help develop a more comprehensive understanding of the essential genes identified in other, more restrictive phenotypic screens.

5.1.3. Growth of S. uberis in the presence of lactoferrin

Lactoferrin is considered a factor in the defence mechanism of the mammary gland. It is considered bacteriostatic for iron dependent bacteria such as staphylococci and coliforms (Sordillo et al. 1997). This is through the depletion of metal ions, thus removing essential nutrients for bacterial growth (Farnaud & Evans 2003), however this is not applicable to streptococci due to their low requirement for metal ions (Todhunter et al. 1995). Iron depletion could also be involved in the increased bacterial degradation caused by phagocytes due

to the ability of lactoferrin to halt production of dismutase. This enzyme is produced by some bacteria, inactivating superoxide radicals, halting production of peroxide (Sordillo et al. 1997). Peroxide is used by phagocytes in conjunction with lactoferrin to produce hydroxyl radicals which are potent bactericides (Clifford & Repine 1982). Other postulated iron-independent immune roles of lactoferrin are associated with the disruption of the cell wall in Gram-negative bacteria (Ellison et al. 1990) and the blockade of carbohydrate metabolism (Arnold et al. 1982).

Lactoferrin is found at its highest level in bovine milk, but concentrations are still not clearly established varying from 115mg/L to 189mg/L (Sánchez et al. 1992; Hagiwara et al. 2003; Soyeurt et al. 2007; Cheng et al. 2008). Lactoferrin concentrations also alter at different stages of lactation (Cheng et al. 2008) and a substantial increase has been reported during mammary gland infection; up to a concentration of 2000mg/L (Harmon et al. 1976; Hagiwara et al. 2003; Komine et al. 2005).

5.1.4. Analysis of conditionally essential genes in test vs. control experiments

The comparison of input and output pools can be achieved in two different ways. Firstly the difference in the number of essential genes can initially identify genes that are required for growth in the specific environment. This binary fashion of looking for 'new' essential genes in an output pool may not by entirely appropriate for identifying changes in mutant prevalence with a

'test vs. control' experimental design. Due to the difference in the size of the CDSs present within bacterial genomes direct comparison cannot be performed, as larger CDS are likely to contain more insertions. To take this into account when comparing different CDSs, a normalisation of the insertions present is necessary. A basic method to overcome the size variability would be to create an insertion frequency for a known length, for example a region of 1kb. This allows for a simple method of direct comparison. However additional methods are also needed to take into account the amount of sequence reads gained and how representative this number is within the population. Normalised values (NRM and NIM scores) were developed to account for this (Chapter 3). NRM – Normalised Reads Mapped (total number of reads / length of gene in Kb) / (total mapped read count / 10^6) and NIM – Normalised Insertions Mapped (total unique insertions mapped / Length of gene in Kb) / (total insertions mapped / 10^6) provide a indication of the disruption of a given gene in comparison to others within the population and also takes into account the variability of the number of mapped sequence reads for each experiment (Blanchard et al. 2015).

Secondly an assessment of fitness can be generated for those CDSs that have a reduced amount of mutant insertions present in the output when compared to their relative frequency in the input pool. Using statistical analysis would help identify finer differences in prevalence, however standard non-parametric statistical tests such as T-test and Mann-Whitney U tests are not applicable due to the low number of replicates used in TIS experiments,

normally just two. The comparison of the input and output pools with multiple replicates can be achieved using differential count software packages, such as DESeq2 (Love et al. 2014), edgeR (Robinson et al. 2010), cutdiff (Trapnell et al. 2013), bayseq (Hardcastle & Kelly 2010) or NOISeq (Tarazona et al. 2010). However with the in house method PIMMS.compare (Blanchard et al. 2015) was developed to enable the analysis of experiments with fewer replicates. PIMMS.compare identifies common and unique mutant insertions between experimental conditions. Using files generated by PIMMS.mapping (Blanchard et al. 2015), the comparison module can be used by:

```
PIMMS.pl –m [compare] –i [input position depth file] -o [output position depth file] –d [collapse by number of base pairs] –n [short name of output file]
```

This generates three tables of data (input only, output only and shared positions). The tables consist of insertion positions, normalised, observed and expected number of reads with associated gene information. For the shared positions an indication of the magnitude of deviation from an expected norm is determined. Within an experiment, the natural logarithm (base e) transformed proportion ratio (the share of reads mapped at a location) approximates a normal distribution. Using the mean (shared mean) and standard deviation (shared sd) of this population, for each insertion the input/output proportion ratio (Proportion ratio), Zscore is calculated (Zscore = ((log(Proportion ratio)) - (shared mean)) / (shared sd)) (Blanchard et al. 2015).

5.2. Materials and Methods

5.2.1. Growth of bacteria in liquid cultures for phenotypic studies

A vial of the mutagenised *S. uberis* glycerol stock generated in Chapter 4, was centrifuged (8 000*g*, 5 minutes) and supernatant aspirated and the cell pellet was resuspended in 100µl TE-buffer (10mM Tris-HCl, 1mM EDTA). Samples were centrifuged (8 000*g*, 5 minutes) and the cell pellet resuspended in 1ml pyrogen free saline. Samples were diluted appropriately and approximately $5.5x10^5$ CFU was inoculated into 10ml THB and incubated at 37°C.

5.2.2. Extraction of bacteria from liquid culture (THB)

After the predetermined time of growth the entire culture was centrifuged (8 000*g*, 5 minutes) and the supernatant aspirated. The cell pellet was suspended in 1ml TE-buffer (10mM Tris-HCl, 1mM EDTA). The sample was centrifuged (8 000*g*, 5 minutes) and the final bacterial pellet was resuspended in 5ml pyrogen free saline containing 50% (v/v) glycerol, aliquoted and frozen at -80°C.

5.3. Results

5.3.1. Growth of S. uberis in liquid culture

5.3.1.1. Growth parameters in liquid culture

Growth of *S. uberis* 0140J pGhost9::IS*S1* was assessed for 8 hours at 37°C to understand the amplification of the bacteria in liquid culture. This allowed parameters to be defined that would allow for sufficient mutant representation in the final population (Figure 5.1). It was concluded that culturing for 5 hours would provide 100-fold amplification of each mutant strain inoculated. Harvested bacteria were subjected to DNA extraction and enrichment of chromosomal/insertion DNA junction as described in Chapter 3.

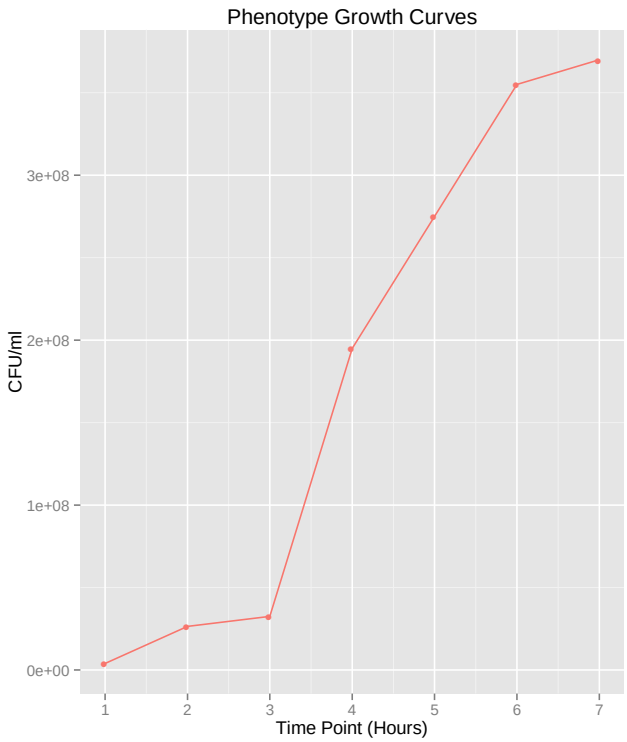

Figure 5.1 Growth of *S. uberis* 0140J pGhost9::IS*S1* in Todd Hewitt Broth

5.3.1.2. Analysis of sequencing data

The sequencing data was parsed through the PIMMS pipeline (Chapter 3) to identify CDSs essential for growth in liquid medium. There was a total of 8,054,130 reads, of which 4,102,362 (51%) contained the terminal transposon sequence. From those 1,814,730 (44%) mapped to the *S. uberis* 0140J genome (Accession number AM946015) (Ward et al. 2001) with 16,560 unique insertion sites identified (Figure 5.2). This achieved a mean coverage of 129 reads per insertion with a median inter-insertion distance of 42bp. There was a mean incidence of 6 insertions per disrupted CDS. These data identified 402 CDSs without insertions and 60 CDSs were truncated (i.e. contained an insertion only in the terminal 10% of the sequence) (Table 5.4).

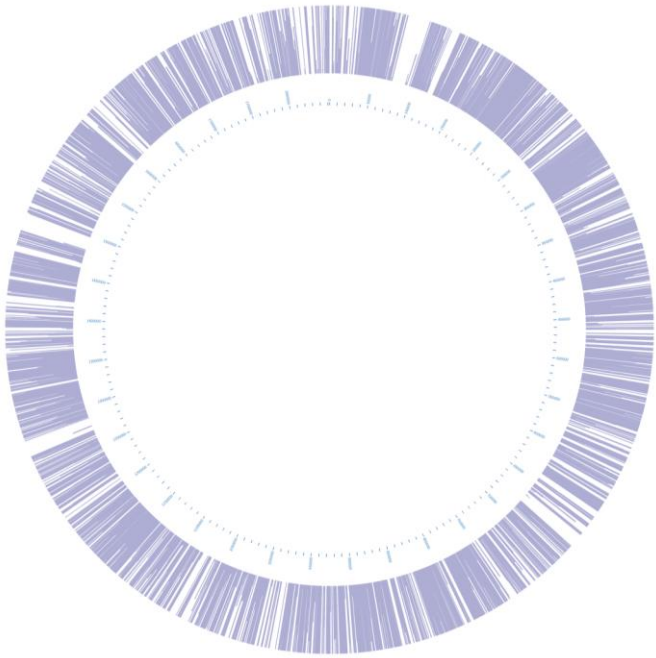

Figure 5.2. Distribution of insertions identified within the *S. uberis* genome after growth in Todd Hewitt broth. The coordinate file generated by the PIMMS bioinformatics pipeline was used to create a Circos plot (Krzywinski et al. 2009).

5.3.1.3. Identification of essential CDSs for growth in Todd Hewitt broth

Comparisons were made with the essential CDS data obtained from the same mutant pool in Chapter 4. In liquid medium, an additional 251 CDSs were identified that were devoid of insertions (Table 5.3). The genes identified as essential for growth on solid media were also essential for growth in liquid medium with the exception of one gene, SUB1706 (*pgi*, glucose-6-phosphate isomerase) (Figure 5.3).

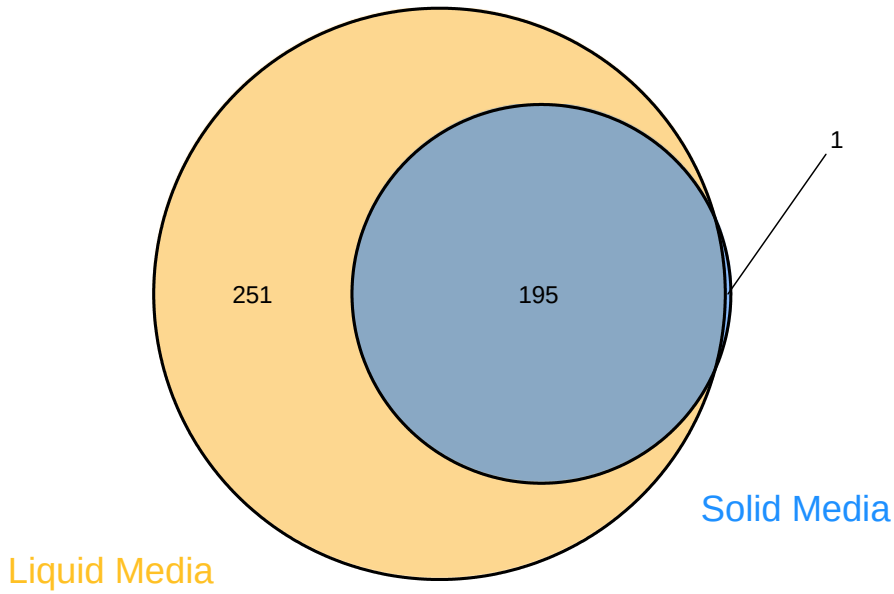

Figure 5.3 A Venn diagram comparing the number of essential genes found after growth of *S. uberis* on solid medium and liquid media.

The CDSs devoid of insertions were aligned with other proteins of a similar function using RAST (Overbeek et al. 2014). When comparing the RAST category allocations from bacteria grown on solid media to bacteria grown in liquid media, hypothetical and miscellaneous proteins were highly represented (Figure 5.4).

Table 5.1. Essential CDSs and their associated RAST categories after growth in Todd Hewitt broth Each protein coding sequence that was deemed essential for growth of *S. uberis* in Todd Hewitt Broth. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

THB RAST Category	Count	Genes
Amino Acids and Derivatives	11	*aroD, asd, dapB, hom, prpC, sdhB*, SUB0316, 0608, 1009, 1156, 1529
Carbohydrates	13	*kduD, ldh, pdhA, pstI, pstK, rbsD, rpiA, treA*, SUB0529, 0679, 0748, 1347
Cell Division and Cell Cycle	11	*ftsJ, ftsY, ftsX, pbp1A, sepF*, SUB0479, 0894, 0923 1090, 1292, 1576
Cell Wall and Capsule	15	*galE, gcaD, mraY, murC, murD, murM, murN, rgpG, rmlD*, SUB0043, 0538, 0695, 0698, 0699, 0700,
Cofactors, Vitamins, Prosthetic Groups	9	*coaA, coaC, coaD, folC1, foldD, mraW, nadE*, SUB1098, 1261
DNA Metabolism	16	*dnaA, dnaN, gyrA, gyrB, holB, lig, recA, recR, ruvA, spxA, tag*, SUB0353, 0498, 0499, 0648, 1710
Fatty Acids, Lipids, and Isoprenoids	11	*fps, fabG, mvaA, mvaK1, mvaK2, pgsA*, SUB0024, 0755, 1107, 1246, 1502
Hypothetical Phage Proteins	8	SUB0221, 0224, 0225, 0227, 1819, 1825, 1827, 1834
Hypothetical Proteins	24	SUB0064, 0148, 0154, 0193, 0216, 0268, 0274, 0283, 0335, 0362, 0377, 0624, 0754, 0847, 0987, 1026, 1215, 1280, 1434, 1571, 1598, 1649, 1787, 1817
Iron Acquisition	1	*vicX*
Membrane Transport	29	*mtuB, mtuC, gla,* SUB0058, 0100, 0272, 0480, 0500, 0504, 0511, 0597, 0611, 0709, 1033, 1064, 1150, 1287, 1415, 1476, 1552, 1573, 1578, 1635, 1814, 1852, 1854, 1856
Miscellaneous	23	*cysS, citD, gidA, pnpA, tyrS*, SUB0003, 0019, 0113, 0361, 0704, 0913, 0931, 1135, 1178, 1185, 1401, 1438, 1514, 1681, 1798, 1805, 1811, 1832A
Nucleosides and Nucleotides	13	*ccd, gmk, pyrH*, SUB0133, 0141, 0240, 0381, 0401, 0951, 1412, 1559, 1749, 1807
Phages, Transposable elements, Plasmids	4	SUB0222, 0229, 0805, 1839
Phosphorus Metabolism	1	*phoU*
Potassium metabolism	1	SUB0371
Protein Metabolism	27	*cipX, efp, engB, prfA, pth, rplA, rplC, rplI, rplL, rplM, rplQ, rplT, rpmG, rpmH, rpsN, rpsP, rpsT, rrf, secY, serS, thrS, trsA*, SUB0929, 1157, 1194, 1296, 1562
Regulation and Cell Signalling	1	*clpP*
Respiration	3	*atpC*, SUB0614, 0708
RNA Metabolism	22	*acdR, ctsR, glyQ, greA, ileS, lysS, nusA, nusG, rpoA, rpoB, rpoD, valS,* SUB0013, 0355, 0730, 0872, 0873, 1417, 1468, 1479, 1617, 1707
Stress Response	6	*csp, cspC, dnaK, hrcA, sodA,* SUB0656
Virulence, Disease and Defence	2	*pedA*, SUB0512
Total	251	

Table 5.2. Truncated CDSs and their associated RAST categories after growth in Todd Hewitt broth Each protein coding sequence only contained insertions within the last 10%. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

THB RAST Category	Count	Genes
Amino Acids and Derivatives	3	*mtnN, recO*, SUB0065
Carbohydrates	2	*pgi, tpi*
Cell Division and Cell Cycle	2	*ftsH, ftsW*
Cell Wall and Capsule	6	*bacA, hasC, murB*, SUB0714, 0426, 1031
Cofactors, Vitamins, Prosthetic Groups	9	*coaE, dyr, folQ, metK, ribC*, SUB0356, 0945, 1410, 1709
DNA Metabolism	2	*parC, ruvB*
Fatty Acids, Lipids, and Isoprenoids	3	SUB0253, 0333, 0653
Hypothetical Phage Proteins	-	
Hypothetical Proteins	3	SUB0011, 0742, 1125
Iron Acquisition	-	
Membrane Transport	5	*secA,* SUB0112, 0151, 0515, 1005
Miscellaneous	4	SUB0728, 0772A, 1769, 1832A
Nucleosides and Nucleotides	5	*pgmA, pyrE, thyA*, SUB0745, 1227
Phages, Transposable elements, Plasmids	1	SUB1832
Phosphorus Metabolism	1	*ppaC*
Potassium metabolism	-	
Protein Metabolism	5	*argS, fmt, groEL, tsf*, SUB0345
Regulation and Cell Signalling	1	*ccpA*
Respiration	1	*oxaA*
RNA Metabolism	5	*rbfA, rnc, rnh*, SUB1809
Stress Response	2	*rex,* SUB1094
Virulence, Disease and Defence	-	
Total	60	

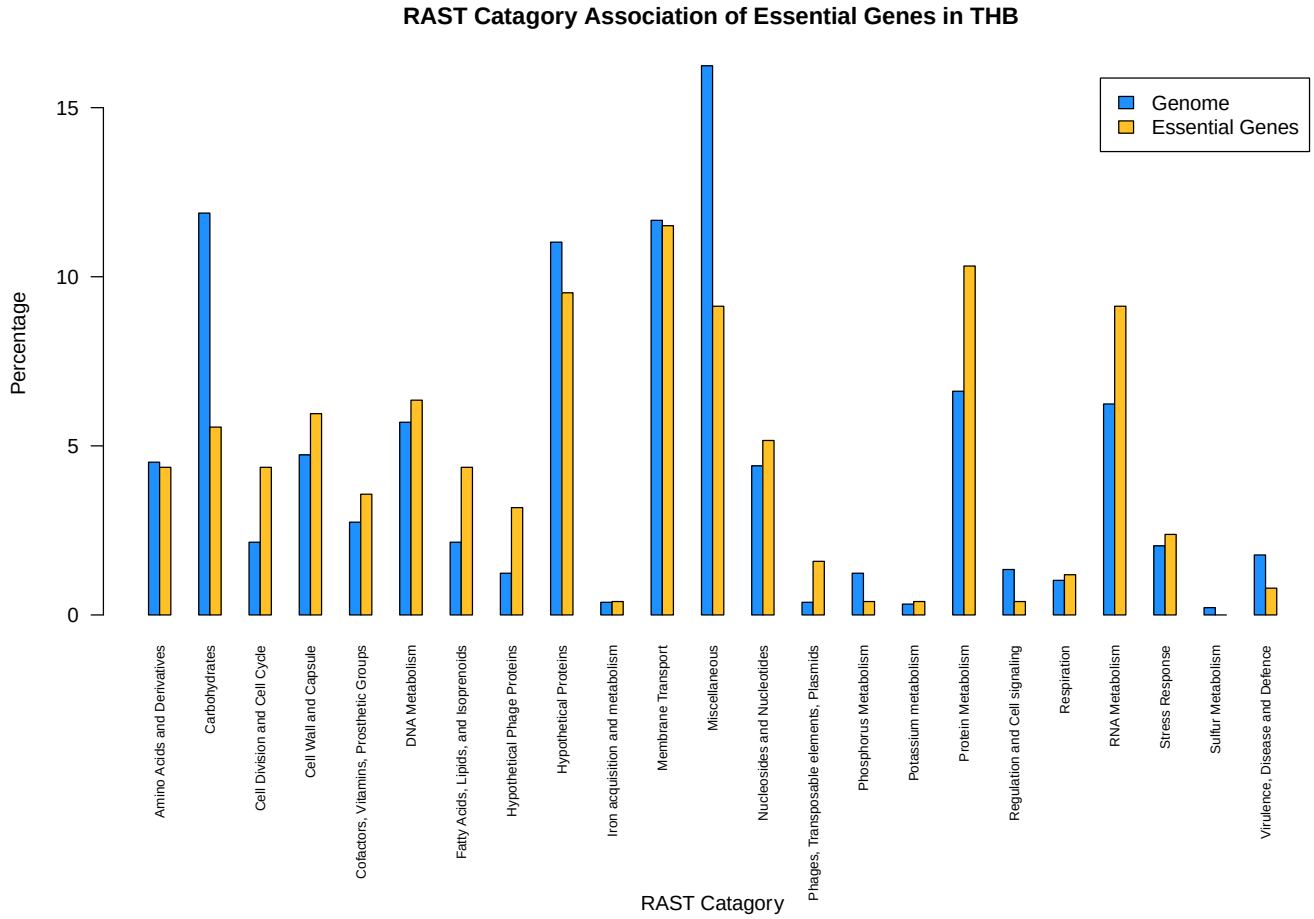

Figure 5.4. Proportional enrichment of essential genes associated with each RAST category. Their percentage of essential genes associated with each RAST category was calculated for the whole genome and those found to be essential for growth in THB.

5.3.2. Growth of S. uberis in the presence of lactoferrin

5.3.2.1. Growth parameters in the presence of lactoferrin

Growth of *S. uberis* 0140J pGhost9::IS*S1* was reassessed to account for any potential effect of lactoferrin on the bacteria. A second growth study was conducted in order to ascertain the amount of time needed to allow for sufficient mutant library complexity (Figure 5.5). The bacteria were cultured for 6 hours, which would still allow for 100-fold growth, thus making the data comparable to the mutants identified during growth in liquid media alone. Harvested bacteria were subjected to DNA extraction and enrichment of chromosomal/insertion DNA junction as described in Chapter 3.

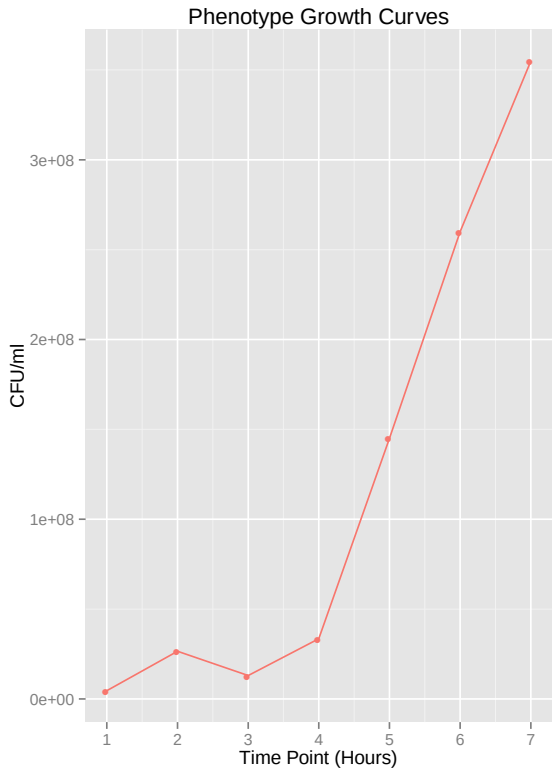

Figure 5.5 Growth curve of *S. uberis* 0140J pGhost9::ISS1 in THB with the addition of 1mg/ml lactoferrin

5.3.2.2. Analysis of the sequencing data

There was a total yield of 8,441,171 reads, of which 4,795,431 (57%) contained the terminal transposon site. A total of 1,769,355 (37%) mapped to the *S. uberis* genome (Accession number AM946015) (Ward et al. 2001) with 15,612 unique insertions (Figure 5.6). There was a mean read depth of 136 and a median inter-insertion distance of 42bp. There were 469 CDSs devoid of insertions (Table 5.3) and 62 truncated CDSs (i.e. contained an insertion only in the terminal 10% of the sequence) (Table 5.4) and a mean incidence of 6 insertions per disrupted CDS.

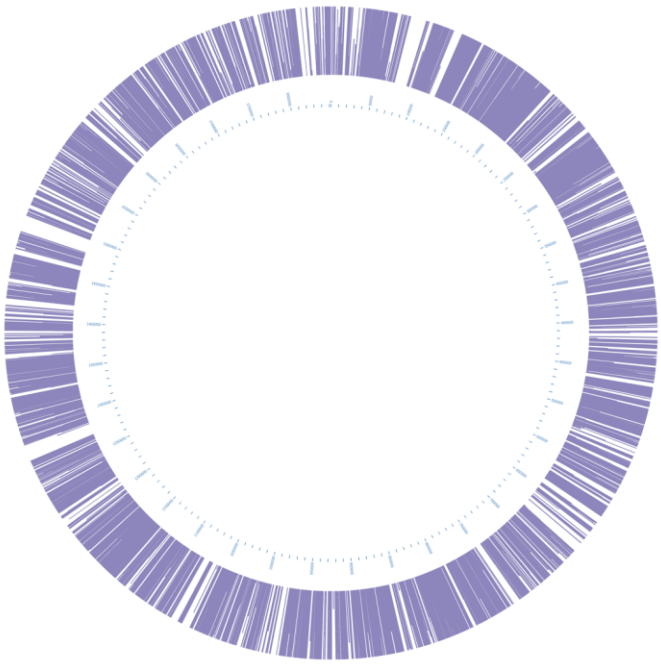

Figure 5.6. Distribution of insertions identified within the *S. uberis* genome after growth in Todd Hewitt broth with the addition of lactoferrin. The coordinate file generated by the PIMMS bioinformatics pipeline was used to create a Circos plot (Krzywinski et al. 2009).

5.3.2.3. Identification of the CDSs essential for growth in THB containing lactoferrin

There were 215 genes identified as essential for growth in THB containing lactoferrin that were also essential for growth in THB alone and 185 genes that were also essential for growth in THB and on THA. There was another unique gene for growth in THB containing lactoferrin, SUB0255 (*proS*), which is a proyl-tRNA synthase (Figure 5.7). There were also 36 genes that were identified as essential for growth in THB but non-essential for growth in THB containing lactoferrin (Table 5.5).

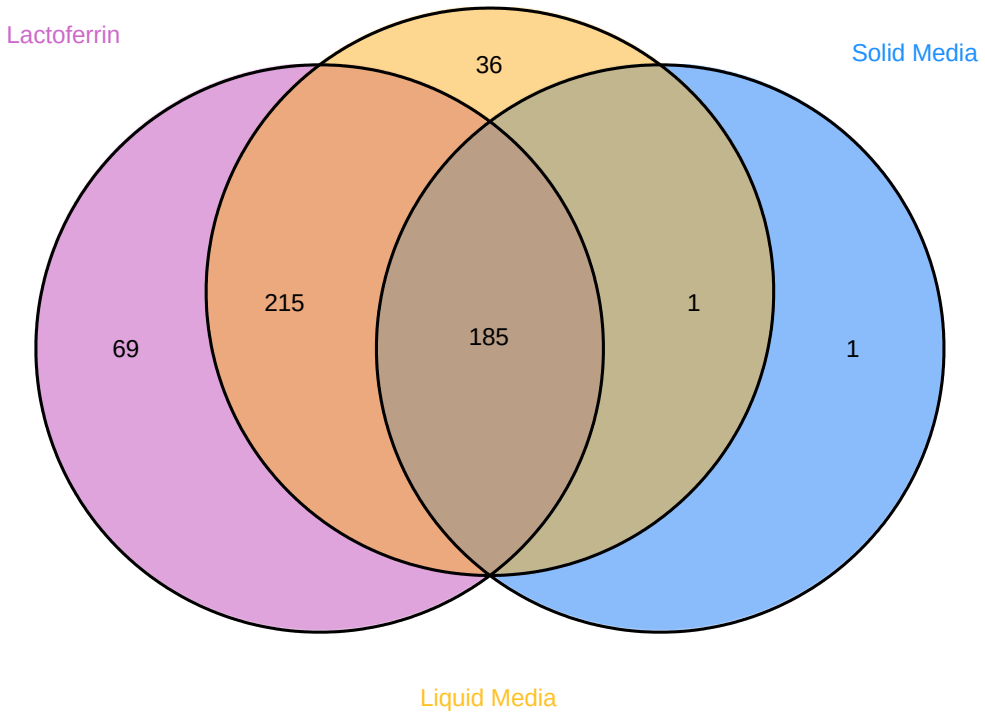

Figure 5.7 A Venn diagram illustrating the overlap of essential genes between growth conditions Essential genes that were identified after growth in THB supplemented with lactoferrin (Lactoferrin), compared the when grown in THB (liquid media) alone and on THA (solid media).

There were 69 additional CDSs devoid of insertions identified after growth in THB supplemented with lactoferrin (at 1mg/ml), when compared to those identified after growth in THB alone (Table 5.3). The majority of these additional essential genes are associated with core metabolic processes and include RNA metabolism, protein metabolism and cell division and cell cycle. Interestingly, hypothetical proteins account for the largest group of the additional essential genes (Figure 5.8).

Table 5.3. Additional essential CDSs identified after growth in Todd Hewitt broth with the addition of lactoferrin Each protein coding sequence that was deemed essential for growth of *S. uberis* in Todd Hewitt broth with the addition of lactoferrin. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

THB plus Lactoferrin RAST Category	Count	Genes
Amino Acids and Derivatives	3	*cysE, thrB,* SUB1289
Carbohydrates	4	*lacF2, tpi,* lSUB0978, 1764
Cell Division and Cell Cycle	5	*scpB,* SUB0476, 1091, 1405, 1577
Cell Wall and Capsule	5	*ditB, hasC*, SUB0539, 0714, 1031
Cofactors, Vitamins, Prosthetic Groups	6	*cinA, tdk, thiN,* SUB0356, 1410, 1709
DNA Metabolism	4	*parC, polC,* SUB1549, 1845
Fatty Acids, Lipids, and Isoprenoids	3	SUB0253, 0333, 0653
Hypothetical Phage Proteins	-	
Hypothetical Proteins	10	SUB0011, 0292, 0510A, 0786, 0967, 1184, 1299, 1346, 1526, 1750
Iron Acquisition	-	
Membrane Transport	7	*mtuA,* SUB0299, 0410, 0508, 0707, 0876, 1005
Miscellaneous	3	*mur1.1, stk1,* SUB1099
Nucleosides and Nucleotides	4	*pyrR,* SUB0031, 0292A, 1392
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	-	
Potassium metabolism	1	*mscL*
Protein Metabolism	6	*codY, gatC, rpmGB, trmE,* SUB0354, 1647
Regulation and Cell Signalling	-	
Respiration	-	
RNA Metabolism	7	*rpoE, rbfA, rnh, rpoZ, truB, yrzB,* SUB0636
Stress Response	2	*rex,* SUB0441
Virulence, Disease and Defence	-	
Total	69	

Table 5.4. Truncated CDSs identified after growth in Todd Hewitt broth with the addition of lactoferrin Each protein coding sequence only contained insertions within the last 10%. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005)

THB plus Lactoferrin RAST Category	Count	Genes
Amino Acids and Derivatives	3	*mtnN, recO*, SUB0065
Carbohydrates	3	*nox, pgi, ptsK*
Cell Division and Cell Cycle	3	*ezrA, ftsH, ftsW*
Cell Wall and Capsule	3	*bacA, murB*, SUB0426
Cofactors, Vitamins, Prosthetic Groups	7	*coaE, dyr, folQ, metK, ribC,* SUB0945, 0950,
DNA Metabolism	2	*dnaA, ruvB*
Fatty Acids, Lipids, and Isoprenoids	1	SUB0755
Hypothetical Phage Proteins	-	
Hypothetical Proteins	5	SUB0148, 0286, 0377, 0847, 1125
Iron Acquisition	-	
Membrane Transport	4	*secA,* SUB0112, 0151, 0647
Miscellaneous	9	*topA,* SUB0179, 0273, 0704, 0728, 0772A, 0913, 1769, 1832A
Nucleosides and Nucleotides	6	*gmk, pgmA, purR, thiD*, SUB0745, 1227
Phages, Transposable elements, Plasmids	1	SUB1832
Phosphorus Metabolism	1	SUB0418
Potassium metabolism	-	
Protein Metabolism	7	*argS, groEL, pheT, rplL, serS, spi, tsf*
Regulation and Cell Signalling	1	*ccpA*
Respiration	1	*oxaA*
RNA Metabolism	6	*cca, proS, rnc, rpoC,* SUB0703, 1809
Stress Response	-	
Virulence, Disease and Defence	-	
Total	63	

Table 5.5. Additional essential CDSs identified only in growth of Todd Hewitt broth Each protein coding sequence that was deemed essential for growth of *S. uberis* in Todd Hewitt broth. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

THB plus Lactoferrin RAST Category	Count	Genes
Amino Acids and Derivatives	2	*aroD,* SUB1156
Carbohydrates	2	*kduD, ptsK*
Cell Division and Cell Cycle	2	*pbp1A,* SUB0923
Cell Wall and Capsule	1	SUB0043
Cofactors, Vitamins, Prosthetic Groups	-	
DNA Metabolism	2	*dnaA, recR*
Fatty Acids, Lipids, and Isoprenoids	1	SUB0755
Hypothetical Phage Proteins	-	
Hypothetical Proteins	7	SUB0148, 0216, 0268, 0283, 0335, 0377, 0847
Iron Acquisition	1	*vicX*
Membrane Transport	3	SUB0500, 0597, 1150
Miscellaneous	3	SUB0704, 0913, 1178
Nucleosides and Nucleotides	2	*gmk,* SUB0401
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	-	
Potassium metabolism	1	SUB0371
Protein Metabolism	3	*rpsL, rpsP, serS*
Regulation and Cell Signalling	1	*clpP*
Respiration	-	
RNA Metabolism	3	*nusG,* SUB0730, 0872
Stress Response	1	SUB0656
Virulence, Disease and Defence	1	SUB0143
Total	36	

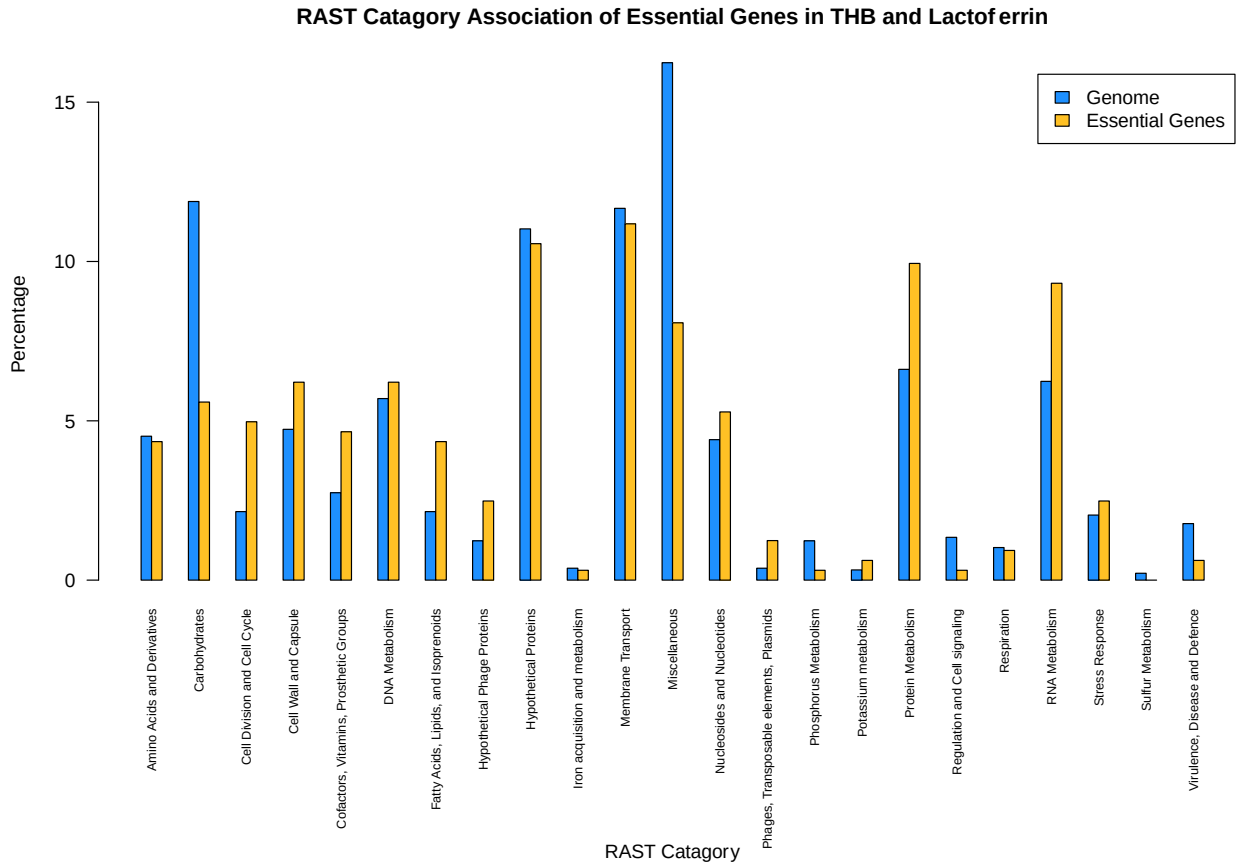

Figure 5.8 Proportional enrichment of essential genes associated with each RAST category. Their percentage of essential genes associated with each RAST category was calculated for the whole genome and those found to be essential for growth in THB with the addition of lactoferrin.

5.4. The identification of insertions that are selected or counter selected in the presence of lactoferrin

The comparison of the THB and THB with supplemented lactoferrin growth conditions, identified seven CDSs with a Z score of >3 (an associated p-value of <0.001). This identifies insertions with the greatest difference from the mean, accompanied with their statistical significance.

There were seven insertions present in seven different CDS that showed a statistically significant decrease in presence of lactoferrin (Table 5.6). These seven CDSs also showed a 3-fold change from the mean, indicating a high magnitude reduction. Investigation of the CDS revealed that even though there is a significant reduction of a particular insertion, there still are a number of insertions present elsewhere in the CDS.

The position of the insertion within the CDSs (insertion %) was investigated to identify any potential positional effect (Table 5.6). In SUB0345, which had a NRM of 1047 after growth in THB but had a NRM of 77 after growth in THB supplemented with lactoferrin. The majority of insertions were present between 78 and 97% of the CDS with one at 1.9%. However after growth in THB there were also insertions present in 8, 15, 19 and 77% of the CDS. This could indicate the truncation effect discussed previously that was found in other TIS experiment data (Gallagher et al. 2011). The other six CDSs all had a

distribution of other insertions throughout the entire length, showing no other indication of positional effect.

Table 5.6. Unique insertions that have a statistically significant reduction in the THB supplemented with lactoferrin

Locus	Product	Insertion (%)*	Input[1]	Output[2]	NRM	Z score
SUB0048	putative amidase	25.5	1162	3	425	-3.554
SUB0192	response regulator protein	14.9	1696	6	842	-3.376
SUB0345	putative cysteine desulfurase	97.2	2060	3	77	-3.878
SUB0826	putative surface-anchored subtilase family protein	38	1502	4	693	-3.535
SUB0932	conserved hypothetical protein	57.9	1409	5	841	-3.375
SUB1398	putative permease	11.5	1002	3	432	-3.47
SUB1583	conserved hypothetical protein	9.1	745	3	520	-3.302

*Insertion (%) represents the percentile position within the CDS where the insertions are present. [1] Number of reads found in association with that unique insertion in the data generated after growth in THB. [2] Number of reads associated with that unique insertion in the data generated after growth in THB supplemented with lactoferrin.

5.5. Discussion

The concentration of lactoferrin used in this study was chosen to replicate levels recorded within a mammary gland infected with *S. uberis* during a bovine challenge model (0.3mg/ml) and a known sub-inhibitory concentration (2mg/ml) (Chaneton et al. 2008). This was expected to mimic resource restrictions imposed by lactoferrin during an intra-mammary infection, allowing the bacteria to grow, whilst imposing a selective pressure. The ability of lactoferrin to sequester iron is probably not directly involved in the control of *S. uberis* due to its low requirement for iron. However, its ability to bind other divalent metal ions such as manganese (Mn) (Steijns & van Hooijdonk 2000) may be critical as concentrations of Mn in milk are very low (10 to 50µg/L) (Thomas 1970) and its acquisition is essential for *S. uberis* to grow in milk (Smith, A. J. et al. 2003). The gene involved in transport of Mn, at low concentrations, is MtuA, a metal binding protein, which is part of a high affinity Mn transport system. After growth of *S. uberis* in THB containing lactoferrin, *mtuA* was found to be completely lacking insertions. The data generated in Chapter 4, from growth on THA, identified the *mtuA* CDS as having an NRM of 53 and an NRM of 39 when grown in THB. The other components of the *mtu* operon, *mtuB* and *mtuC,* were also found to be essential for growth in THB and THB containing lactoferrin. In environments with high levels of Mn, such as nutrient rich laboratory media, *S. uberis* appears to have passive mechanisms for Mn ion import, as insertion mutants with inactivating lesions in the *mtuA*, m*tuB* or *mtuC* genes, that form the Mn

transport operon are still able to proliferate (Smith, A. J. et al. 2003). However at low levels of Mn, such as that typically found in milk (0.8mg/L) or in this case media containing lactoferrin and thus depleted Mn ions, *mtuA* is essential. Therefore the identification of *mtuA* as an essential gene under these conditions acts as a positive internal control for the PIMMS process in identifying conditionally essential genes under different selective pressures or environments.

There was a single gene identified as being essential for growth on solid media when compared to essential genes after growth in THB or THB containing lactoferrin. The gene SUB1792 (*pgi*) is part of the Embden-Meyerhof pathway, discussed in Chapter 4, which is responsible for glucose catabolism (Smith & Doolittle 1992) and had a single insertion after growth in THB and THB with lactoferrin. Coordinates of the insertion revealed a truncation in both incidences at 99.8% of the CDS, resulting in the loss of two amino acids. This truncation is outside the isomerase domain and polypeptide-binding interface so is therefore unlikely to have a negative effect on the function of the enzyme. Another gene SUB0255 (*proS*) a prolyl-tRNA synthase, acts as a ligase between the amino acid and its cognate tRNA (Eriani et al. 1990). The gene was found to be essential for growth in THB and on THA but not in THB containing lactoferrin. Similarly to the insertional occurrence in found in *pgi* after growth in THB*, proS* only has a single insertion in the terminal 10% of the CDS. However, as the associated read depth for this insertion was only 4, and taking into account the method of bacterial extraction from the culture, the

bacterium containing this insertion might have been present but non-viable for replication.

There were 37 CDSs that were found to be essential for growth in THB, which were found to contain insertions after the addition of lactoferrin (Table 5.5). The incidence of these disrupted CDSs was investigated and revealed that each CDS had either one or two insertions identified numerous times. Their absence from the list of essential genes, after growth in THB containing lactoferrin, could be a limitation of the method related to the subsampling of the population of mutants for inoculation of liquid media. As two different aliquots, from the pool of mutants, were used to obtain the desired number of mutant bacteria, a different selection may have been chosen in each instance.

There were a high proportion of hypothetical proteins found in the essential gene lists. To investigate any putative function previously allocated to the amino acid sequences these were parsed through PSI-BLAST (Altschul et al. 1997). SUB0193 and SUB0624 were only found to be present in *S. uberis,* whist SUB0216 was also found in the closely related *Streptococcus parauberis*, however there were no further functions or domains allocated to these hypothetical genes. SUB0268 was found to have an SseB N-terminal domain, which was initially found in the S*almonella* pathogenicity island (PAI). PAIs are mobile genomic elements that play a pivotal role in virulence (Hacker et al. 1990). This domain has been linked to bacterial replication inside macrophages (Ruiz-Albert et al. 2003) and has also been found to regulate

serine sensitivity in *Escherichia coli* (Hama et al. 1994). The presence of PAIs in streptococci is limited when compared to other bacterial species. *S. pyogenes* has numerous virulence factors that have been acquired through phage mediated gene transfer but lacks the structure of a PAI (Ferretti et al. 2001), whereas *S. pneumoniae* has a defined chromosomal region of approximately 27kb which contains transposase and recombinase genes related to PAIs (Brown et al. 2001). However this PAI is absent from other streptococci (Schmidt & Hensel 2004).

SUB1215 was found to have a GAF domain, which are typically present in cyclic nucleotide phosphodiesterases (PDEs) and have been shown to interact with transcriptional factors, predominantly associated with gene regulation in bacteria (Ho et al. 2000; Zoraghi et al. 2004). The remaining hypothetical protein SUB0847 was identified as a ribose-phosphate pyrophosphatekinase, which converts ribose-5-phosphate into 5-phosphate-α-1-pyrophosphate as part of the pentose phosphate pathway and is deemed essential for growth in *Bacillus subtilis* (Eriksen et al. 2000).

The role of resource competition, could also have an impact on the CDSs required for proliferation in a liquid media (Case & Gilpin 1974). When bacteria are grown on solid medium, there is little or no nutritional competition. However, in liquid media more bacteria are using resources, quickly depleting freely available nutrients (Nadell et al. 2009). Nutritional limitation has widespread effects on gene regulation and is controlled through

the use of alarmones and quorum sensing (Cornforth & Foster 2013). The intracellular alarmones Guanosine pentaphosphate, guanosine tetrephosphate ((p)ppGpp) and associated ribosome enzymes; guanosine hydrolase and synthase respond to low levels of amino acids, sugars, phosphates and fatty acids (Bouvert & Battesti 2011). The production of (p)ppGpp initially reduces the biosynthesis of macromolecules whilst attempting to restart metabolism but can arrest cell growth and activate stress survival genes (Cornforth & Foster 2013).

The primary mechanism for quorum sensing in Gram-positive bacteria is through short ribosomally produced, post transitionally modified signal peptides (Lyon & Novick 2004). The quorum sensing peptides in *S. uberis* are encoded by the genes *luxS* and *agrA.* The gene *luxS* was initially discovered in *Salmonella typhimurium* and later characterised in 30 other bacteria (Surette & Bassler 1998), whilst *agrA* was first discovered in *S. aureus* and was characterised by its ability to allow *S. aureus* to respond to changes in bacterial density (Peng et al. 1988). The stimulation of these receptor molecules by an secreted inducer molecule from another bacterium, leads to the regulation of several physical properties such as competence, biofilm formation and virulence (Cvitkovitch et al. 2003). The induction of the quorum sensing peptides can also lead to the production of antimicrobial peptides such as bacteriocins. This is especially important in mutagenesis studies, as a bacterium without a functioning resistance protein will be susceptible to the cognate bacteriocin. However, it has also been shown that some streptococci

(*S. mutans* and *S. gordonii*) are able to degrade the signal peptide, which stops the production of bacteriocins, nullifying the competitive advantage held by those bacteria with the bacteriocin resistance gene (Bing-yan Wang & Kuramitsu 2005).

There are also cooperative strategies that can be induced by population-density. The most well documented incidence of this in bacteria is regarding siderophores, which are involved in iron acquisition (Clancy et al. 2006). Siderophores are iron-chelating compounds secreted by some bacteria which enable the release of iron from high-affinity iron binding glycoproteins such as lactoferrin and transferrin (Weinberg 1986). If these are released in sufficient concentrations the environment becomes rich in soluble iron, removing this nutrient limitation and allowing bacterial proliferation. The utilisation of siderophores is well characterised in Gram-negative bacterial species, however, with the exception of *Streptococcus equi* subspecies *equi* (Heather et al. 2008), no evidence has been shown relating to siderophore biosynthesis in streptococci (Eichenbaum et al. 1996). Despite having a low iron dependency, *S. uberis* still has the *fhu* operon that encodes for a siderophore mediated Mn/Fe transport system, suggesting the potential presence of an uncharacterised metal ion acquisition system. Analysis of the data generated for the genes associated with alarmones, quorum sensing and siderophore released ions, has shown they do contain insertions and may not play an important role in growth in liquid media.

Chapter 6 Identification of conditionally essential genes in biologically relevant environments

6.1 Introduction

6.1.1. Growth in skimmed milk

As *S. uberis* is primarily a mastitic pathogen, bacterial growth in milk would closely mimic the nutritional aspects available in the host environment. This would aid the discovery of specific genes required for the bacterium to grow. Skimmed milk has been used previously to successfully replicate the nutritional aspects of the mammary gland, leading to the characterisation of the lipoprotein *mtuA* and its function within *S. uberis* for manganese acquisition (Smith, A. J., et al. 2003).

Due to the compositional complexity of milk, it is not possible to extract bacteria directly from culture without enrichment. Milk also contains PCR inhibitors that can be carried through during bacterial DNA isolation (Wilson 1997), which may be problematic for the inverse PCR step of the PIMMS protocol. This is a well-documented problem in the food industry, particularly for detection of bacterial contamination of dairy products during quality control and food safety testing (Pirondini et al. 2010). Because of this, there are numerous methods that have been developed to circumvent such problems including the use of multiple modified DNA extraction methods (Quigley et al. 2012). However these approaches do not produce sufficient

amounts of bacterial DNA with the purity appropriate for the PIMMS inverse PCR amplification.

There are methods to remove host cells, fats, proteins and calcium that have been designed or adapted to facilitate the extraction of bacteria directly from milk. Buoyant floatation separation, involves a density gradient centrifugation of the milk sample with three different colloidal solutions and is used to separate biological particles and microorganisms by sedimentation and flotation (Wolffs et al. 2004). This method suffers from some limitations; including the carryover of materials that could potentially inhibit the amplification reaction (Wilson 1997). Enzymatic milk clearing using savinase was developed specifically for milk samples (Gunasekera et al. 2000). Savinase catalyses the hydrolysis of the peptide bonds in the milk proteins (Meisel et al. 2011). However, it has been shown that savinase is bactericidal, which could have an impact on the isolation of intact bacteria from milk culture (Smith, E. M. et al. 2003). Finally microbial plating onto solid media with selective antibiotics, although time consuming, would result in a robust sampling of the population of mutant *S. uberis*. This method has been used in other TIS experiments in order to isolate the output bacterial population for the environmental screen (Hensel et al. 1995; Christiansen et al. 2014). This would ensure all experimental factors were identical and allows the impact of the independent variable to be accurately assessed. However as bacteria needed to be plated out onto agar after a phenotypic screen, a control sample also needed to be created. The use of plating out bacteria onto solid media after

the phenotypic screen may infer additional selective pressures, but these would only be linked to the overall fitness of the bacterial mutation. Some studies have used repeated passaging of the bacteria to remove the lower incidence mutant insertions (Gawronski et al. 2009; Langridge et al. 2009; Christiansen et al. 2014), which would lead to a similar selection of bacteria with increased fitness. To ensure the additional step of growth on solid media did not inflate the number of CDSs found lacking mutant insertions, comparative analysis should be performed between the corresponding input and output pools. The use of microbial plating means that only a subset of the cultured population would be collected for sequencing. This could also introduce a potential limitation on transposon insertion discovery, as there would potentially be a higher incidence of bacteria containing the same mutation.

6.1.2. Growth of S. uberis in milk containing serum

During the course of mammary gland infection there is an increased presence of serum derived compounds present in the mammary gland secretion (Korhonen et al. 2000). Their occurrence is due to the disruption of the integrity to the mammary epithelial lining (Batavani et al. 2007). There are several important functions related to these secretions, firstly the immunoglobulin present is thought to play a role in the inhibition of bacterial growth and prevent adherence to the epithelial membrane (Nickerson 1989). There are also specific proteins such as α-lactalbumin and β-lactoglobulin, the levels of which increase during *S. uberis* infection (Hoeben et al. 1999) that

have perpetrated immunoenhancing effects on the production of IL-1β (Wong et al. 1997). The supplementation of milk with serum would, to some extent, be considered to mimic the nutritional environment within the lactating mammary gland during infection.

6.1.3. Initial investigations required for robust data generation

When using "test vs. control" sequencing experiments, there are a few important factors to consider. Each sequencing experiment will generate a finite number of reads, which has implications on read depth and target discovery, such as mutant insertions (Nakamura et al. 2011). To accurately assess the impact of an independent variable, like a phenotypic screen, other experimental factors need to be identical, which can be problematic with high-throughput sequencing (Schrimer et al. 2015). Technical variation as a result of changing reagents or the use of a different flowcell during a sequencing run, can result in amplification preferences for certain cluster populations, skewing the data generated (Aird et al. 2011). The proposed benefits of running input and output pools on the same sequencing run, equalising the potential effects of sequencing bias, is an important factor to consider. Therefore the effect of the reduction of the number of biological replicates used in the comparative analysis should be investigated.

The use of selective antibiotics during growth, used to ensure the population of bacteria are those containing the insertions, could also have an impact on what genes are deemed essential in the analysis. It has been suggested that

incubation of bacteria in the presence of erythromycin can enrich for insertions in certain regulatory genes, thus having implications on the function of their cognate genes (Sun et al. 2010). It is therefore necessary to either account for this in the data analyses or include measures to reduce the effect of the proposed mechanism.

6.1.4. Analysis of conditionally essential genes with multiple replicates

Genomic analysis tools have been developed to address the low number of replicate samples used for differential analysis of transcription levels in RNA-Seq experiments, which can be also be applied to any sequence based count data. Their application in TIS phenotypic screens is also documented (Christiansen et al. 2014) meaning their use could provide a widely applicable method to compare the differential insertion frequency. These algorithms normalise the data to account for variation in total numbers of reads within the sample. The ratios of reads are then used to calculate a $\log_2$ fold change. A difference reflects an attenuated or hypercompetitive mutant strain, with differences labelled as infinite or negative-infinite reflecting a result where zero insertions are found in one or the other pool.

There are two well-known packages currently used in modern RNA-Seq analysis; edgeR (Robinson et al. 2010) and DESeq2 (Love et al. 2014). Both assume the data has a negative binomial distribution, which is a natural occurrence of repeated sampling from random variables. This distribution has been shown to be accurate by analysis of repeated sequencing of a set of

specific oligonucleotides (Marioni et al. 2008) and to be a robust assumption by repeated analysis of sequencing results from the same sample (Lu et al. 2005). The two packages have been compared numerous times using different sample data sets (Soneson & Delorenzi 2013; Nookaew et al. 2012; Rapaport et al. 2013), however neither seems to dominate the other across every aspect of the data analysis (Anders et al. 2013). To identify differential prevalence of insertion mutants, edgeR is favoured because of its comprehensive data output (Robles et al. 2012). The p-value gives an indication of the confidence in the difference of prevalence between the two conditions, given the negative binomial distribution of the count data, and the $\log_2$ fold change (log-FC) affords an estimation of the magnitude of the difference. The log-FC is calculated by

$$logFC_g = \log(n_{g,b}) - \log(n_{g,a})$$

where g = the genomic feature (CDS), $n_{g,b}$ is the normalised average of the read count in the test condition, $n_{g,a}$ is the normalised average read counts in the control condition. The subtraction calculates the log of the ratio, which acts as a filter for negligible changes. The log-FC p-value is also controlled using Benjamini and Hochberg's procedure (Benjamini & Hochberg 1995) to reduce the false discovery rate (FDR).

6.2. Materials and Methods

6.2.1. Preparation of milk samples for phenotypic growth studies

Milk was collected using standard aseptic hand milking technique from each quarter of six Holstein/Freisan dairy cows and stored at 4°C. Each sample was assessed for sterility by inoculating 1ml of milk into 9ml of THB and incubated overnight at 37°C. A sample of the culture was plated onto THA and incubating for a further 16 hours at 37°C. Samples that did not show any microbial growth were skimmed by centrifugation (2000*g*, 20 minutes). The fat layer was removed and 1ml of the skimmed milk was incubated for a further 6 hours at 37°C, plated onto THA plates and incubated overnight at 37°C. Samples devoid of microbial growth were pooled, aliquoted and stored at -20°C.

6.2.4. Extraction of bacterial colonies from complex liquid culture (Milk)

A volume of culture containing approximately 1x10^{6} CFU was diluted into 10ml of pyrogen free saline and was divided onto two 225ml bioassay dishes of THA containing erythromycin (1µg/ml) and grown at 37°C overnight to produce single colonies. Colonies were collected using 5ml PBS for each plate and the bacterial suspension centrifuged (8 000*g*, 10 minutes) and re-suspended in TE buffer three times. The final bacterial pellet was resuspended in 10ml pyrogen free saline containing 50% (v/v) glycerol, aliquoted and frozen at -80°C.

6.4. Results

6.4.1. Preliminary investigations

6.4.1.1. The use of fewer biological replicates

Data collected from the non-competitive agar growth study, generated in Chapter 4, was used to compare which CDSs were disrupted and how well these disruptions were conserved across each pool of mutants. A random permutation saturation plot was generated in order to identify how many biological replicates would be required to gain similar results from fewer samples (Figure 6.1). Pools of mutants were chosen at random and assessed for the amount of essential CDSs. The number of essential CDSs in any one given pool was then used to identify a mean and standard deviation for the use of *n+1* pools. This was repeated using *n+2* pools up to and including *n+9* pools of mutants. By using this method of analysis, it was found that the use of five biological replicates would allow the approximate identification of 85% of the CDSs that have been disrupted in the *S. uberis* 0140J pGhost9::IS*S1* bank of mutants (Figure 6.1). The essential CDSs, from each pool, were compared to assess their distribution. The data obtained in Chapter 4 identified 199 CDSs lacking transposon insertions when the bacteria are grown on THA. Choosing the five pools of mutants, which have the highest level of variation (Figure 6.2), there are 289 CDSs identified without transposon insertions. The 90 additional CDSs mostly contained a single insertion throughout the entire sequence and were often only identified less the ten times. These CDSs were only identified as essential in two specific pools of

mutants (10 and 13) so were not representative of the overall population.

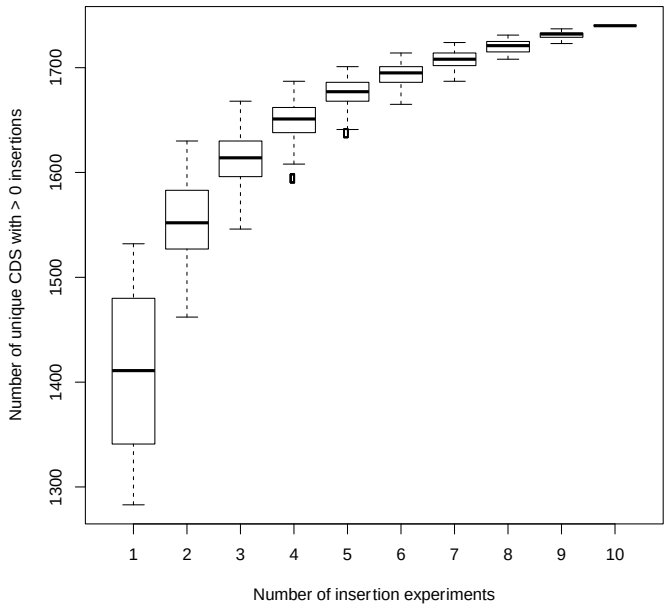

Figure 6.1. Saturation of disrupted CDSs using 10 pools of mutants. The number of disrupted CDSs was calculated from each pool of mutants, to assess the reduction of biological replicates on the data gained

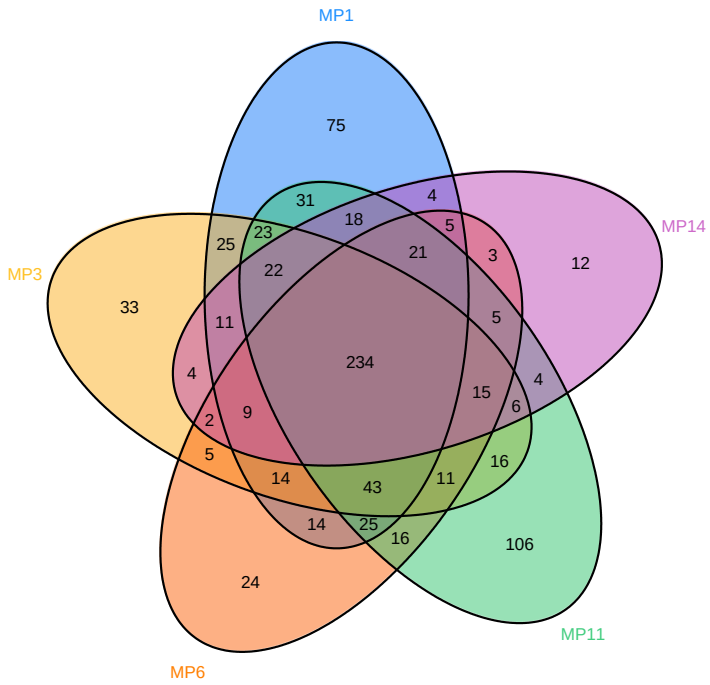

Figure 6.2. A Venn diagram showing the overlap of the number of essential CDSs from the five biological replicates. The five pools chosen allow for 85% of the CDSs, which have been shown to contain transposon insertions, whilst reducing the number of samples needing to be processed.

6.4.1.2. Effect of selective antibiotics

Growth of *S. uberis* 0140J pGhost9::IS*S1* was assessed in skimmed milk for 8 hours at 37°C, with and without the addition of erythromycin. The number of bacteria counted was almost identical at each time point except for the transition between growth and stationary phase (Figure 6.3). The bacteria without the selective erythromycin were actively growing for longer. The final number of bacteria present after 8 hours was almost identical as the sample containing erythromycin grew to approximately $1.9x10^8$ and without erythromycin bacterial numbers reached $1.95x10^8$.

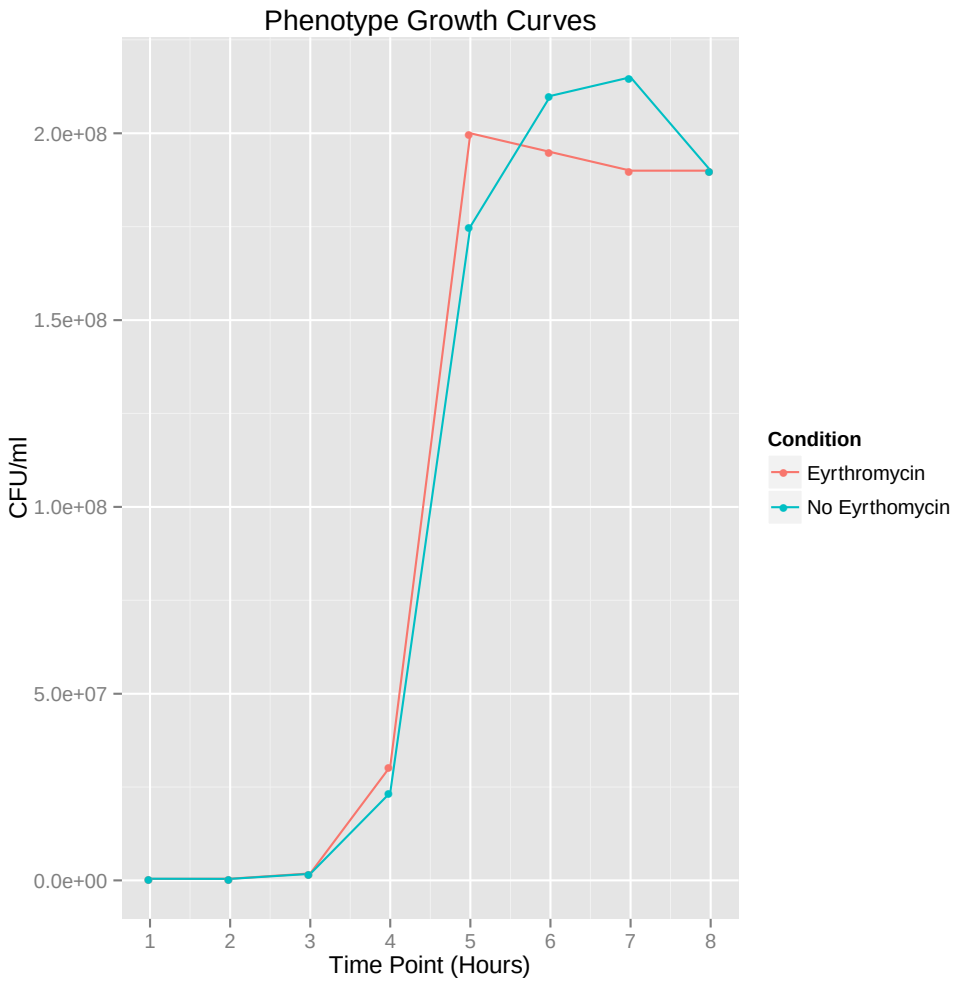

Figure 6.3 Growth curve assessing the effect of erythromycin on *S. uberis* numbers during culture in skimmed milk.

6.4.2. Growth of S. uberis in skimmed milk

6.4.2.1. Growth parameters

Raw milk was collected from six healthy adult Friesian/Holstein dairy cows. The milk was tested for microbial contamination before being skimmed by centrifugation and pooled, nullifying individual cattle variation, which could potentially skew results. Growth of *S. uberis* 0140J pGhost9::IS*S1* was assessed for 8 hours at 37°C (Figure 6.4).

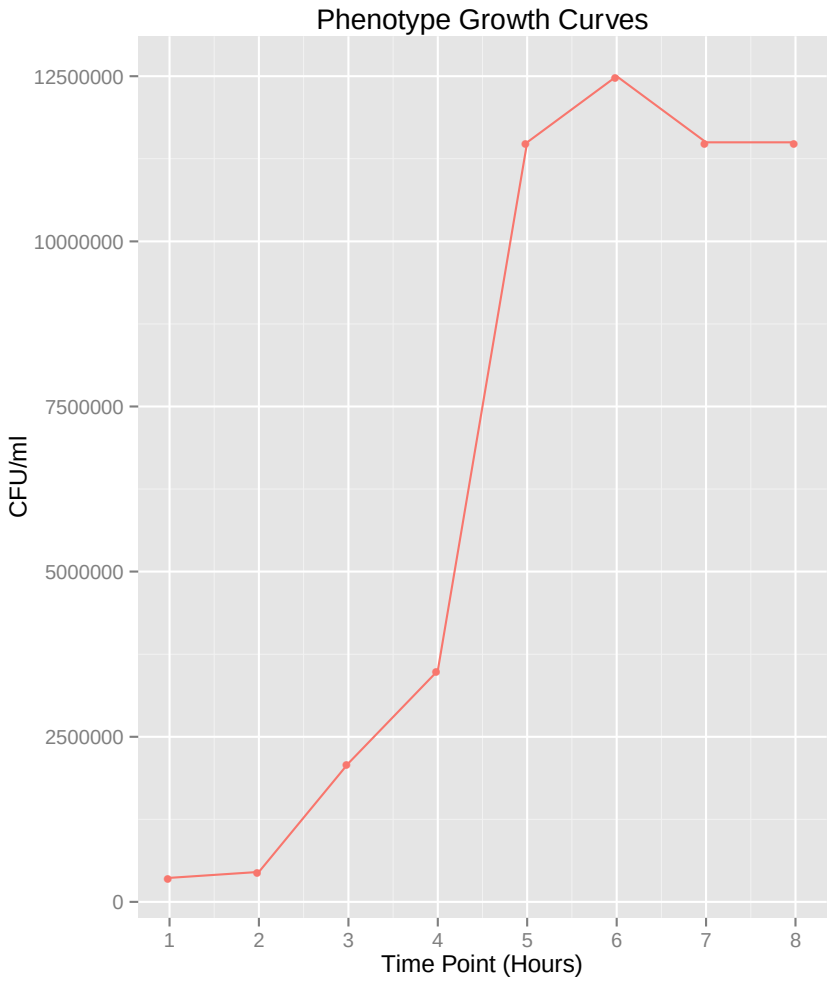

Figure 6.4 Growth assessment of *S. uberis* 0140J pGhost9::IS*S1* in skimmed milk

To allow for sufficient mutant library complexity, bacteria were inoculated at $5x10^5$ CFU and cultured to allow for 100-fold growth (approximately 5-6 generations). This allows for mutant strains with a lower fitness or those that have an insertion with a deleterious effect to be lost. Two subsamples of the culture, each containing approximately $5x10^5$ CFU, were plated onto solid agar. Thus enabling the collection of approximately 1,000,000 mutants.

6.4.2.2. Analysis of the sequencing data from input and output pools

The samples generated without growth in skimmed milk yielded 23,931,932 reads and 14,107,428 (59%) contained the terminal transposon site. A total of 3,355,909 (24%) mapped to the *S. uberis* genome (Accession number AM946015) (Ward et al. 2001) with 41,949 unique insertions (Figure 6.5a). These insertions had a mean coverage of 108 reads, with a median inter-insertion distance of 28bp. The disrupted CDSs contained a mean of 6 insertions.

The samples created after growth in skimmed milk generated 23,338,346 reads. A total of 12,783,608 (55%) contained the terminal transposon site, and 3,569,262 (23%) mapped to the *S. uberis* genome (Accession number AM946015) (Ward et al. 2001). There were 37,262 unique insertions (Figure 6.5b) with a mean coverage of 134 reads and a median inter-insertion distance of 21bp. Each disrupted CDSs contained a mean of 6 insertions.

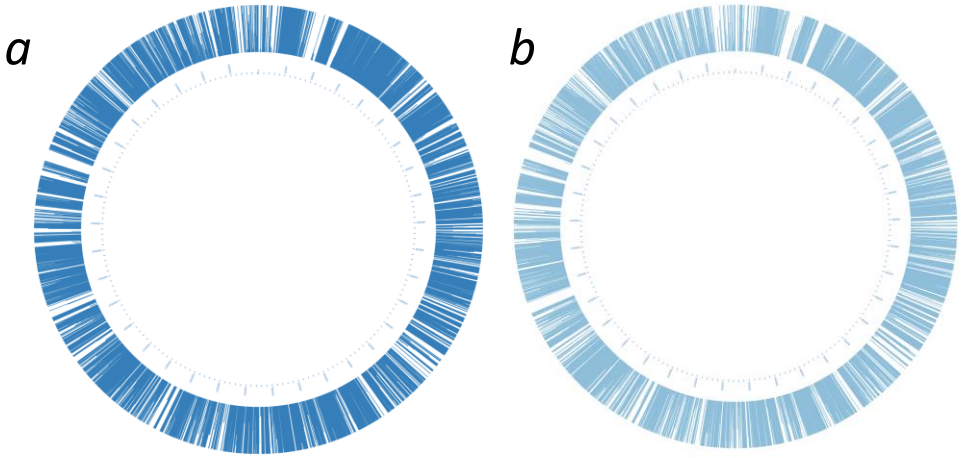

Figure 6.5. Distribution of insertions identified within the *S. uberis* genome after growth in skimmed milk. The coordinate files generated by the PIMMS bioinformatics pipeline a) milk input pool and b) milk output pool were used to create a Circos plot (Krzywinski et al. 2009).

6.4.2.3. Identification of the essential genes for growth in skimmed milk

There were 313 essential genes identified in the input pool. These were compared to the essential genes already identified as essential for growth on solid media, using the reduced number of biological replicates (Section 6.3.1), revealing an additional 24 essential genes (Table 6.1). The analysis also identified 71 truncated genes (i.e. contained an insertion only in the terminal 10% of the sequence) (Table 6.2).

The output pool contained an additional 70 essential genes (Table 6.3) and 59 truncated genes (Table 6.4) when compared to the input pool. There were also 39 genes identified as essential in the input that contained insertions in the output pool (Table 6.5).

The essential genes were grouped into corresponding RAST categories based

on their functionally related protein families (Table 6.6). Analysis of potential RAST category enrichment was performed using chi-squared, to assess any preferential use of pathways associated with growth in skimmed milk (Table 6.7). Using a p-value cut-off of <0.05 the categories associated with protein metabolism, fatty acid biosynthesis, cell division, cell wall, cofactors, vitamins, prosthetic groups were identified as being statistically enriched pathways (Figure 6.6).

Table 6.1. Additional essential CDSs identified in the milk input pool Each protein coding sequence that was deemed essential for growth of *S. uberis* on agar. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed milk input RAST category	Count	Genes
Amino Acids and Derivatives	1	SUB0316
Carbohydrates	1	*tpi*
Cell Division and Cell Cycle	-	
Cell Wall and Capsule	1	SUB1037
Cofactors, Vitamins, Prosthetic Groups	3	*coaA, coaD,* SUB0469
DNA Metabolism	2	*dnaG,* SUB0353
Fatty Acids, Lipids, and Isoprenoids	2	*mvaK2,* SUB0253
Hypothetical Phage Proteins	-	
Hypothetical Proteins	1	SUB1089
Membrane Transport	2	*gla,* SUB1852
Miscellaneous	3	SUB0019, 0361
Nucleosides and Nucleotides	1	SUB0240
Phages, Transposable elements, Plasmids	1	SUB1832
Phosphorus Metabolism	-	
Potassium metabolism	-	
Protein Metabolism	4	*rpmG, serS*, SUB0005, 0929
Respiration	-	
RNA Metabolism	1	*era*
Stress Response	1	SUB0656
Virulence, Disease and Defence	-	
Total	24	

Table 6.2. Truncated CDSs identified in the milk input pool Each protein coding sequence only contained insertions within the last 10%. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed milk input RAST category	Count	Genes
Amino Acids and Derivatives	5	*cysE, mtnN, recO, thrB*, SUB0065
Carbohydrates	4	*nox, pgk, pgi, ptsK*
Cell Division and Cell Cycle	5	*ftsh, ftsW, ftsX*, SUB0714, 0923
Cell Wall and Capsule	7	*bacA, hasC, murB, murC*, SUB0426, 0698, 0714
Cofactors, Vitamins, Prosthetic Groups	8	*coaC, coaE, metK, ribC*, SUB0356, 0641, 1410, 1709,
DNA Metabolism	3	*dnaN, holB, ruvB*
Fatty Acids, Lipids, and Isoprenoids	2	SUB0333, 0653
Hypothetical Phage Proteins	2	SUB1819, 1834
Hypothetical Proteins	5	SUB0377, 0624, 1125, 1434, 1537
Membrane Transport	4	*secA, mtuC*, SUB0674, 1854
Miscellaneous	5	*cysS, tyrS*, SUB0728, 0931, 1438A
Nucleosides and Nucleotides	5	*pgmA, thiD*, SUB0745, 1227, 1412
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	-	
Potassium metabolism	-	
Protein Metabolism	8	*fmt, groEL, prfA, rplA, rplL, rplT, trsA, tsf*
Respiration	1	*oxaA*
RNA Metabolism	7	*nusA, rpoC, rpoD, rnc*, SUB0013, 0730, 0873
Stress Response	-	
Virulence, Disease and Defence	-	
Total	71	

Table 6.3. Additional essential CDSs found after growth in milk Each protein coding sequence that was deemed essential for growth of *S. uberis* in skimmed milk. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed milk output RAST category	Count	Genes
Amino Acids and Derivatives	2	*cysE, prpC*
Carbohydrates	8	*eno, glkA, pgk, ptsK, pyk, treA*, SUB0679, 0978
Cell Division and Cell Cycle	5	*ftsZ,* SUB0467, 1091, 1292, 1613
Cell Wall and Capsule	4	*murB, murC, srtA*, SUB0714
Cofactors, Vitamins, Prosthetic Groups	7	*folD, folQ, metK*, SUB0356, 0641, 1098, 1410
DNA Metabolism	3	*dnaN, rpe*, SUB0499
Fatty Acids, Lipids, and Isoprenoids	3	*mvaA*, SUB0653, 1246
Hypothetical Phage Proteins	1	SUB1819
Hypothetical Proteins	4	SUB0613, 0987, 1061, 1434
Membrane Transport	4	*mtuA, secA*, SUB0211, 1854
Miscellaneous	-	
Nucleosides and Nucleotides	3	*prsA2, thyA*, SUB1559
Phages, Transposable elements, Plasmids	2	SUB0228, 0229
Phosphorus Metabolism	3	*csrR, ppaC, phoU*
Potassium metabolism	1	*mscL*
Protein Metabolism	7	*rplB, rplD, rplI, rpmGB, rrf*, SUB1118, 1562
Respiration	2	*oxaA*, SUB0614
RNA Metabolism	8	*asnS, greA, nusA, rbfA, yrzB*, SUB0355, 1315, 1468
Stress Response	2	*hrcA*, SUB0441
Virulence, Disease and Defence	1	SUB1561
Total	70	

Table 6.4. Truncated CDSs found after growth in milk Each protein coding sequence only contained insertions within the last 10%. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed milk output RAST category	Count	Genes
Amino Acids and Derivatives	2	*aspC, mtnN*
Carbohydrates	1	*pgi*
Cell Division and Cell Cycle	6	*ffh, ftsH, ftsW, ftsX,* SUB0479, 0923
Cell Wall and Capsule	5	*gcaD, hasC,* SUB0426, 0698, 0700
Cofactors, Vitamins, Prosthetic Groups	6	*coaC, coaD, coaE, folC1, ribC*, SUB1709
DNA Metabolism	3	*holB, recA, ruvB*
Fatty Acids, Lipids, and Isoprenoids	2	*mvaK2,* SUB0024
Hypothetical Phage Proteins	1	SUB1834
Hypothetical Proteins	2	SUB1132, 1787
Membrane Transport	4	*mtuC,* SUB1005, 1216, 1852
Miscellaneous	5	*cysS, tyrS,* SUB0019, 0728, 0913
Nucleosides and Nucleotides	5	*pgmA, thiD,* SUB0318, 0745, 1227
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	-	
Potassium metabolism	-	
Protein Metabolism	9	*codY, groEL, prfA, rplA, rplL, rplT, serS, trsA, tsf*
Respiration	-	
RNA Metabolism	7	*rpoC, rpoD,* SUB0013, 0555, 0873, 1479, 1618
Stress Response	1	SUB656
Virulence, Disease and Defence	-	
Total	59	

Table 6.5. Essential CDSs found before growth in milk only Each protein coding sequence that was deemed essential for growth of *S. uberis* in skimmed milk. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed milk output RAST category	Count	Genes
Amino Acids and Derivatives	5	*aspS, dapB, hom,* SUB0316 1529
Carbohydrates	-	
Cell Division and Cell Cycle	-	
Cell Wall and Capsule	4	*gcaD, murE*, SUB0695, 0700
Cofactors, Vitamins, Prosthetic Groups	2	*coaD,* SUB0469
DNA Metabolism	2	*gyrB, recA*
Fatty Acids, Lipids, and Isoprenoids	3	*mvaK1, mvaK2,* SUB0024
Hypothetical Phage Proteins	-	
Hypothetical Proteins	4	SUB0011, 0378, 1026, 1089
Membrane Transport	5	*mtuB,* SUB0470, 1004, 1578, 1852
Miscellaneous	4	*pnpA,* SUB0019, 0361, 0869
Nucleosides and Nucleotides	-	
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	1	SUB0490
Potassium metabolism	-	
Protein Metabolism	2	*serS,* SUB1008
Respiration	-	
RNA Metabolism	3	*ileS,* SUB1479, 1618
Stress Response	3	*dnaK, sodA*, SUB0656
Virulence, Disease and Defence	-	
Total	39	

Table 6.6. Comparison between the numbers of CDS lacking insertions in the skimmed milk phenotypic screen in the association to RAST categories.

Category	Mutant Pool	Milk Input	Milk Output
Amino Acids and Derivatives	9	10	12
Carbohydrates	15	16	24
Cell Division and Cell Cycle	13	13	18
Cell Wall and Capsule	26	27	31
Cofactors, Vitamins, Prosthetic Groups, Pigments	8	11	18
DNA Metabolism	20	22	25
Fatty Acids, Lipids, and Isoprenoids	26	28	31
Hypothetical Phage Proteins	1	1	2
Hypothetical Proteins	6	7	11
Iron acquisition and metabolism	1	1	1
Membrane Transport	22	24	28
Miscellaneous	8	11	11
Nucleosides and Nucleotides	8	9	12
Phages, Transposable elements, Plasmids	2	3	5
Phosphorus Metabolism	1	1	4
Potassium metabolism	0	0	1
Protein Metabolism	72	76	83
Regulation and Cell signalling	1	1	1
Respiration	8	8	10
RNA Metabolism	28	29	37
Stress Response	10	11	13
Sulphur Metabolism	0	0	0
Virulence, Disease and Defence	4	4	5
Total	289	313	383

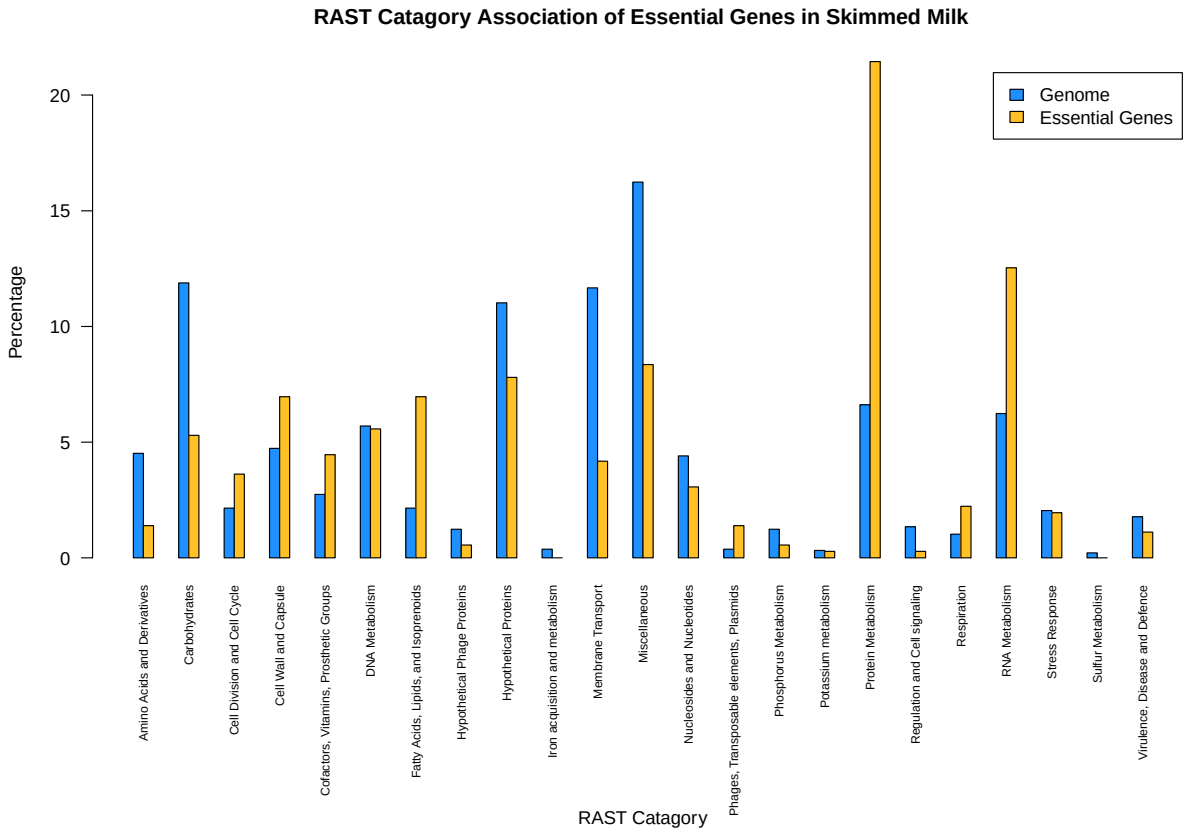

Figure 6.6 Proportional enrichment of essential genes associated with each RAST category. Their percentage of essential genes associated with each RAST category was calculated for the whole genome and those found to be essential for growth in skimmed milk.

Table 6.7. Identification of RAST categories that show a level of enrichment after growth in skimmed milk. Chisq.pval = the associated p-value showing the significance of the enrichment for the given RAST category.

Category	chisq.pval
Protein Metabolism	1.12E-39
Fatty Acids, Lipids, and Isoprenoids	6.07608E-14
Miscellaneous	2.19053E-12
Hypothetical Proteins	5.37605E-08
Carbohydrates	8.73033E-05
Membrane Transport	0.000239133
Cell Division and Cell Cycle	0.001108244
Respiration	0.002455266
Amino Acids and Derivatives	0.009727231
Cell Wall and Capsule	0.012437767
Phages, Transposable elements, Plasmids	0.019323497
RNA Metabolism	0.027838022
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.041847508
Regulation and Cell signalling	0.089778884
Nucleosides and Nucleotides	0.293501301
Hypothetical Phage Proteins	0.30255431
Iron acquisition and metabolism	0.414153066
Phosphorus Metabolism	0.55557006
Stress Response	0.58156347
Virulence, Disease and Defence	0.641269199
Sulphur Metabolism	0.730075652
DNA Metabolism	0.827407273
Potassium metabolism	1

6.4.2.4. The use of differential insertion prevalence to identify phenotypically important genes

The data was used in edgeR statistical analysis and the output sorted in a descending order of false discovery rate (FDR) incorporating a threshold of <0.05. This identified five genes in the skimmed milk phenotypic screen that were found to have differential gene essentiality (Table 6.8). The genes *purC* and *purA* are involved with fatty acid biosynthesis, the gene *folE* is associated by RAST category with cofactors and vitamins as part of the folate biosynthesis pathway. SUB0856 is associated with nucleotide synthesis and SUB1662 is involved with cell division and cell cycle. From those identified, two genes (SUB0152 and SUB1662) had a fold difference from the mean (Log-FC) of greater than two, suggesting a significant loss of insertions after being subjected to the phenotypic screen. However when looking at the number of insertions present in this CDS for the output pool, it is apparent there is only a small decrease in the mutant insertions compare to the input pool.

Table 6.8. Genes with statistically significant fold change differences after screening in skimmed milk

Loci	Gene	Function	Log-FC	FDR*
SUB0025	purC	phosphoribosylaminoimidazole-succinocarboxamide synthase	-1.303941052	5.82976E-05
SUB0865	NA	GMP reductase	-1.679583778	0.000796235
SUB0152	purA	conserved hypothetical protein	-3.226200355	0.031519095
SUB1662	NA	putative DNA-binding protein	-2.909731639	0.031519095
SUB0813	folE	GTP cyclohydrolase I	-1.631423812	0.047970513

* p-value adjusted Benjamini and Hochberg's procedure (Benjamini & Hochberg 1995).

6.4.3. Growth of S. uberis in skimmed milk with serum

6.4.3.1. Growth parameters

Growth of *S. uberis* 0140J pGhost9::IS*S1* was reassessed in order to account for any potential effect of serum on the bacteria. To allow for sufficient mutant library complexity, bacteria were inoculated at 5x10^5 CFU and grown for four hours (approximately 5-6 generations) equivalent to that used previously during the investigation of genes required for growth in milk (Figure 6.7).

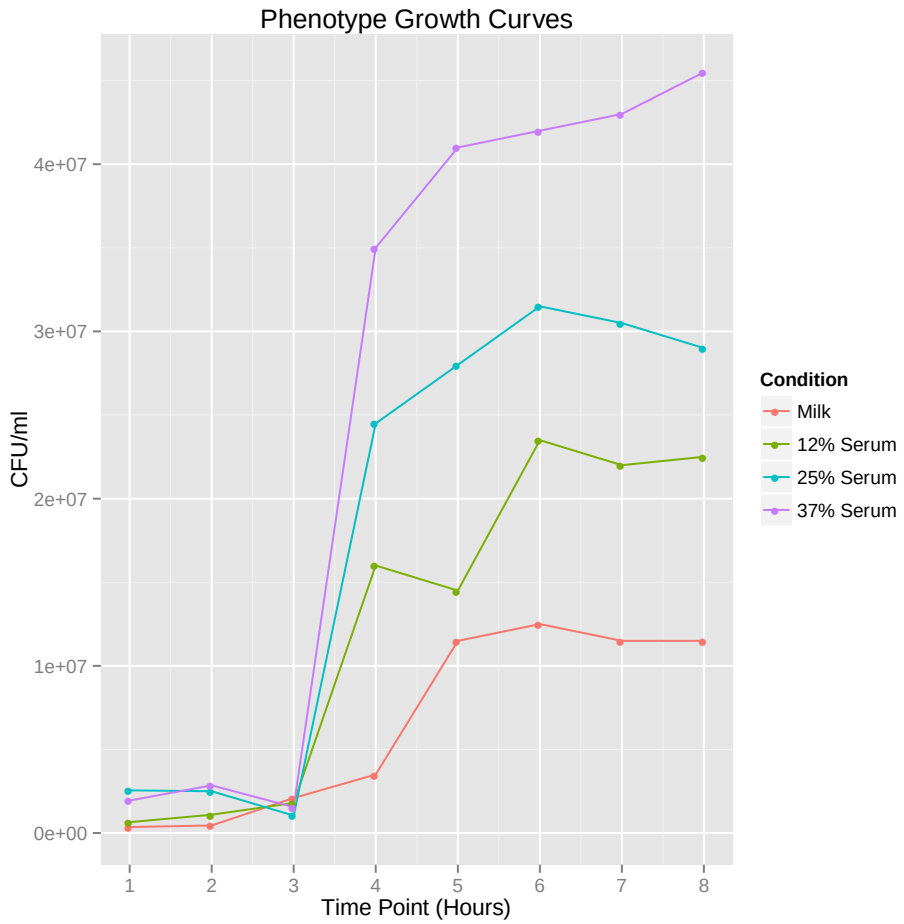

Figure 6.7. *S. uberis* growth curve in skimmed milk with varying amounts of serum. Each medium was inoculated with 5.5x10^5 CFU of wild type *S. uberis* and grown for 8 hours. Samples were taken every hour to attain colony counts by plating onto THA using 8 fold serial dilutions.

6.4.3.2. Analysis of the sequencing data from input and output pools

The samples generated without growth in milk supplemented with 25% serum produced 22,993,308 reads and 6,812,214 (30%) contained the terminal transposon site. From those containing the transposon motif 986,723 (14%) mapped to the *S. uberis* genome (Accession number AM946015) (Ward et al. 2001). There were 16,420 unique insertions (Figure 6.8a) with a mean coverage of 102 and a median inter-insertion distance of 55bp. The disrupted CDSs contained a mean of 6 insertions.

After growth in milk supplemented with serum the samples yielded 23,297,193 reads, of which 8,108,527 (35%) contained the terminal transposon site. A total of 1,357,728 (17%) mapped to the *S. uberis* genome (Accession number AM946015) (Ward et al. 2001) with 22,030 unique insertions (Figure 6.8b). The mean coverage was 98 reads and the median-inter insertion distance was 42bp and a mean of 6 insertions per disrupted CDS.

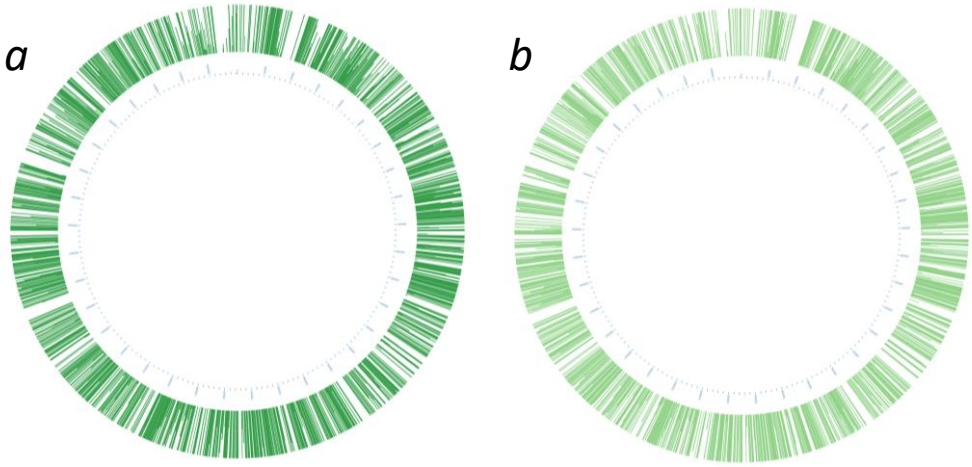

Figure 6.8 Distribution of insertions identified within the *S. uberis* genome after growth in skimmed milk. The coordinate files generated by the PIMMS bioinformatics pipeline a) milk with serum input pool and b) milk with serum output pool were used to create a Circos plot (Krzywinski et al. 2009).

6.4.3.3. Identification of the essential genes for growth in milk supplemented with serum

There were a total of 401 essential genes identified in the input pool. When compared to data generated in Chapter 4, using the five pools of mutants chosen (Section 6.3.1), an additional 179 essential genes were identified (Table 6.9). The analysis also identified 69 truncated genes (i.e. contained an insertion only in the terminal 10% of the sequence) (Table 6.10). The output pool contained an additional 255 essential genes (Table 6.11) and 66 truncated genes (Table 6.12) when compared to the input pool. There were 134 genes only identified as essential in the input that contained insertions in the output pool (Table 6.13).

The essential genes from the output were grouped into RAST categories based

on their functionally related protein families (Table 6.14). Analysis of potential RAST category enrichment was performed using chi-squared, to assess any preferential use of pathways associated with growth in skimmed milk with the addition of serum (Table 6.15). Using a p-value cut-off of <0.005 the category for phages, transposable elements and plasmids was identified as an additional statistically enriched pathway.

Table 6.9. Essential CDSs identified before growth in milk with serum Each protein coding sequence was deemed essential for growth of *S. uberis* on agar. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed Milk plus Serum RAST Category	Count	Genes
Amino Acids and Derivatives	4	*asd, aspC, cysE, recO*
Carbohydrates	12	*ldh, nox, pgk, ptsK, rbsD, rpiA, treA,* SUB0288, 0293, 0529, 0679, 0748
Cell Division and Cell Cycle	8	*ftsH, ftsX ftsY, ftsZ,* SUB0479, 1091, 1577, 1613
Cell Wall and Capsule	8	*bacA, dltD, galE, hasC, murC,* SUB0170, 0426, 0714
Cofactors, Vitamins, Prosthetic Groups	12	*coaC, folC1, folD, folQ, metK, ribC, tdk*, SUB0356, 0641, 1098, 1410, 1709
DNA Metabolism	14	*dnaN, holB, parC, priA, recF, recG, recR, rpe, tag, xseA*, SUB1243, 1638, 1845
Fatty Acids, Lipids, and Isoprenoids	4	*mvaA,* SUB0769, 0333, 0653, 1246
Hypothetical Phage Proteins	8	SUB0218, 0225, 0227, 1820, 1822, 1827, 1836, 1837
Hypothetical Proteins	17	SUB0148, 0153, 0210, 0268, 0292, 0613, 0754, 0801, 0987, 1061, 1434, 1571, 1680, 1735, 1817, 1858
Iron acquisition and metabolism	-	
Membrane Transport	15	*secA, mtuA,* SUB0061, 0186, 0204, 0211, 0238, 0410, 0480, 0500, 0737, 1476, 1566, 1635, 1854
Miscellaneous	16	*gidA, lmrP, stk1,* SUB0251, 0275, 0728, 0736, 0772A, 0913, 0931, 1123, 1303, 1510, 1769, 1810, 1832A
Nucleosides and Nucleotides	10	*ccd, hpt, prsA2, thyA,* SUB0031, 0292A, 0745, 1392, 1559, 1683
Phages, Transposable elements, Plasmids	4	SUB0228, 0229, 0805, 1824
Phosphorus Metabolism	2	*csrR, ppaC*
Potassium metabolism	2	*mscL,* SUB0317
Protein Metabolism	19	*codY, def, fmt, groEL, pheT, prfA, rplA, rplB, rplD, rpll, rplK, rplL, rplL, rpsL, rpsN, rrf, trmE, trsA*, SUB1194, 1562
Regulation and Cell signalling	3	*apt, ccpA, clpP*
Respiration	3	*atpC, oxaA,* SUB0614
RNA Metabolism	16	*apbE, asnS, cca, lysS, nusA, rbfA, rnh, truB, valS, yrzB,* SUB0828, 0873, 1315, 1468, 1809
Stress Response	1	SUB0441
Virulence, Disease and Defence	1	SUB1826
Total	179	

Table 6.10. Truncated CDSs identified before growth in milk with serum Each protein coding sequence only contained insertions within the last 10%. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed Milk plus Serum RAST Category	Count	Genes
Amino Acids and Derivatives	4	*asnB, mtnN, recO,* SUB0066
Carbohydrates	2	*nox,* SUB1459
Cell Division and Cell Cycle	6	*ezrA, ffh, ftsH, ftsX,* SUB0923, 1577
Cell Wall and Capsule	4	*bacA, murB,* SUB0426, 1658
Cofactors, Vitamins, Prosthetic Groups	4	*cinA, coaE, folQ*, SUB1410
DNA Metabolism	6	*addA, addB, holB, ruvB,* SUB0498, 1565
Fatty Acids, Lipids, and Isoprenoids	1	SUB1522
Hypothetical Phage Proteins	3	SUB0221, 1819, 1829
Hypothetical Proteins	9	SUB0377, 0632, 0659, 1125, 1132, 1162, 1511, 1767, 1787
Iron acquisition and metabolism	-	
Membrane Transport	4	*mtuC,* SUB0112, 1005, 1085
Miscellaneous	4	*cyrS, topA, tyrS*, SUB0728
Nucleosides and Nucleotides	3	*pgmA, pyrS, thiD*
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	2	*ppaC,* SUB0452
Potassium metabolism	-	
Protein Metabolism	6	*rplA, rpsP, serS, trsA, tsf,* SUB0929
Regulation and Cell signalling	1	SUB1190
Respiration	-	
RNA Metabolism	6	*metG, rbfA, rpoC, rnc*, SUB0730, 1380
Stress Response	3	*hrcA, rex,* SUB0409
Virulence, Disease and Defence	1	SUB0509
Total	69	

Table 6.11. Additional essential CDSs identified after growth in skimmed milk and serum Each protein coding sequence that was deemed essential for growth of *S. uberis* in skimmed milk with the addition of normal bovine serum. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed Milk plus Serum RAST Category	Count	Genes
Amino Acids and Derivatives	13	*argR, asnB, glnA, mtnN, potA, potC, prpC, sdhA, thrB*, SUB0588, 1009, 1130, 1272
Carbohydrates	21	*ackA, atoA2, deoC, glkA, kduD, lacB2, lacR, pgi, ppc, pta, srlE, tpi,* SUB0045, 0303, 0545, 0978, 1003, 1117, 1725, 1764, 1850
Cell Division and Cell Cycle	8	*exrA, ftsW, sepF,* SUB0057, 0467, 0647, 1292, 1405
Cell Wall and Capsule	11	*dlyB, dltD, pbp1B, srtA, wzg,* SUB0539, 0698, 0904, 1037, 1041, 1416
Cofactors, Vitamins, Prosthetic Groups	5	*coaA, folP,* SUB0469, 0747, 1241
DNA Metabolism	13	*addB, comFC, dnaE, exoA, fpg, mutL*, SUB0122, 0498, 0499, 1232, 1648, 1710
Fatty Acids, Lipids, and Isoprenoids	2	*mvaK2,* SUB0940
Hypothetical Phage Proteins	8	SUB0219, 1059, 1819, 1823, 1825, 1828, 1834, 1835
Hypothetical Proteins	40	SUB0064, 0154, 0167, 0274, 0283, 0362, 0431, 0510A, 0514, 0624, 0632, 0693, 0788, 0790A, 0852, 0858, 0875, 0934, 0941, 0967, 0973, 1032, 1083, 1125, 1138, 1162, 1170A, 1184, 1215, 1280, 1305, 1340, 1400, 1445, 1512, 1575, 1597, 1817
Iron acquisition and metabolism	3	*vicK, vicX,* SUB1684
Membrane Transport	35	*sagG,* SUB0018, 0049, 0058, 0111, 0151, 0189, 0295, 0299, 0461, 0515, 0587, 0599, 0984, 0991, 0992, 1005, 1033, 1964, 1085, 1109, 1256, 1268, 1356, 1473, 1553, 1590, 1661, 1689, 1790, 1852, 1856, 1864, 1865
Miscellaneous	33	*citE, gidB, htrA, mur1.1, pepV, qacH, tyrS,* SUB0012, 0062, 0157, 0166, 0212, 0230, 0327A, 0361, 0427, 0787, 0803, 0861, 0980, 1018, 1052, 1130, 1163, 1269, 1349, 1368, 1348A, 1594, 1600, 1798, 1857, 1861
Nucleosides and Nucleotides	18	*ndk, purA, purB, purE, purR, pyrD, pryF, thiD,* SUB0031, 0240, 0738, 1227, 1306, 1412, 1475, 1751, 1797, 1807
Phages, Transposable elements, Plasmids	1	SUB1831
Phosphorus Metabolism	3	*phoU,* SUB0263, 0963
Potassium metabolism	-	
Protein Metabolism	13	*ksgA, lepA, msrA, pepA, prfC, rpmGB, rpsP, sipC*, SUB1118, 1196, 1296, 1375, 1771
Regulation and Cell signalling	1	SUB1477
Respiration	1	SUB1048
RNA Metabolism	11	*miaA, rnc, rnr, rpoD,* SUB0013, 0142, 0636, 0872, 12181250, 1765
Stress Response	7	*eep, gshA, hrcA,* SUB0440, 0656, 1530, 1714
Virulence, Disease and Defence	5	*copY, oedA, ublA,* SUB0408, 1561
Sulphur Metabolism	3	*ahpC, ndh,* SUB1654
Total	255	

Table 6.12. Truncated CDSs identified after growth in skimmed milk with serum Each coding sequence only contained insertions within the last 10%. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed Milk plus Serum RAST Category	Count	Genes
Amino Acids and Derivatives	5	*aspC, asnC, recO,* SUB0065, 1086
Carbohydrates	3	*nox,* SUB0196, 1764
Cell Division and Cell Cycle	4	*ezrA, ftsH, ftsW,* SUB0479
Cell Wall and Capsule	3	*hasC, murB,* SUB0426
Cofactors, Vitamins, Prosthetic Groups	4	*coaD, coaE, folC1, ribC*
DNA Metabolism	4	*pcrA, ruvB,* SUB0136, 0498
Fatty Acids, Lipids, and Isoprenoids	2	*mvaK2*, SUB0653
Hypothetical Phage Proteins	1	SUB1829
Hypothetical Proteins	6	SUB0011, 0377, 0613, 0847, 1125, 1511
Iron acquisition and metabolism	1	SUB1684
Membrane Transport	3	*mtuC,* SUB0597, 1852
Miscellaneous	6	*cysS, tryS, topA,* SUB0019, 0931, 1798
Nucleosides and Nucleotides	6	*pgmA, pyrR, purh, purK, thiD,* SUB1807
Phages, Transposable elements, Plasmids	-	
Phosphorus Metabolism	-	
Potassium metabolism	-	
Protein Metabolism	8	*rplA, rplT, rpsP, rpsT, serS, smpB, tsf*
Regulation and Cell signalling	1	SUB0851
Respiration	1	*atpC*
RNA Metabolism	5	*cca, rpoC, rpoD,* SUB0730, 1218
Stress Response	2	*rex,* SUB0656
Virulence, Disease and Defence	1	SUB1826
Sulphur Metabolism	-	
Total	66	

Table 6.13. Additional essential CDSs identified before growth in skimmed milk with serum only Each coding sequence that was deemed essential for growth of *S. uberis* in skimmed milk with the addition of normal bovine serum. Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005).

Skimmed Milk plus Serum RAST Category	Count	Genes
Amino Acids and Derivatives	4	*aspC, potC,* SUB0310
Carbohydrates	9	*rbsD*, SUB0156, 0288, 0529, 0531, 0748, 1570, 1724, 1764
Cell Division and Cell Cycle	3	*ftsK, ftsW,* SUB1292
Cell Wall and Capsule	2	*dltD, srtA*
Cofactors, Vitamins, Prosthetic Groups	6	*birA, coaD, folC1, ribC*, SUB0469, 1241
DNA Metabolism	5	*pcrA, recF,* SUB0237, 0499, 1638
Fatty Acids, Lipids, and Isoprenoids	5	*mvaK2, fps*, SUB0333, 0653, 0755
Hypothetical Phage Proteins	7	SUB0218, 0225, 1060, 1820, 1822, 1825, 1835
Hypothetical Proteins	22	SUB0011, 0148, 0153, 0210, 0215, 0216, 0378, 0646, 0754, 0782, 0801, 0847, 0852, 1089, 1170A, 1305, 1400, 1445, 1571, 1607A, 1649, 1858
Iron acquisition and metabolism	-	
Membrane Transport	21	SUB0058, 0151, 0186, 0189, 0204, 0213, 0313, 0410, 0500, 0599, 0611, 0990, 1216, 1265, 1476, 1580, 1635, 1664, 1814, 1852, 1856
Miscellaneous	21	*gidA, qacH, stk1*, SUB0019, 0180, 0230, 0251, , 0275, 0327A, 0399, 0736, 0869, 0931, 1438A, 1478, 1514, 1594, 1600, 1810, 1813, 1851
Nucleosides and Nucleotides	5	*hpt, purR, pyrF,* SUB0031, 0133
Phages, Transposable elements, Plasmids	1	SUB0222
Phosphorus Metabolism	-	
Potassium metabolism	-	
Protein Metabolism	5	*ksgA, rplT, rpsT,* SUB0354, 1118
Regulation and Cell signalling	-	
Respiration	4	*atpC,* SUB0449, 0708, 1048
RNA Metabolism	6	*cca, lysS, miaA, rpoD*, SUB1218, 1617
Stress Response	4	*eep,* SUB0402, 0441, 0656
Virulence, Disease and Defence	2	SUB0513, 1826
Sulphur Metabolism	2	*ahpC, tpx*
Total	134	

Table 6.14. Comparison between the numbers of CDS lacking insertions in the skimmed milk supplemented with serum phenotypic screen in the association to RAST categories.

Category	Mutant Pool	Serum Input	Serum Output
Amino Acids and Derivatives	3	7	20
Carbohydrates	14	26	47
Cell Division and Cell Cycle	12	20	28
Cell Wall and Capsule	20	28	39
Cofactors, Vitamins, Prosthetic Groups, Pigments	4	16	21
DNA Metabolism	15	29	42
Fatty Acids, Lipids, and Isoprenoids	20	24	26
Hypothetical Phage Proteins	0	8	16
Hypothetical Proteins	2	19	59
Iron acquisition and metabolism	1	1	4
Membrane Transport	11	26	61
Miscellaneous	3	19	52
Nucleosides and Nucleotides	7	17	35
Phages, Transposable elements, Plasmids	2	6	7
Phosphorus Metabolism	1	3	6
Potassium metabolism	0	2	2
Protein Metabolism	68	87	100
Regulation and Cell signalling	1	4	5
Respiration	7	10	11
RNA Metabolism	21	37	48
Stress Response	6	7	14
Sulphur Metabolism	0	0	5
Virulence, Disease and Defence	4	5	8
Total	222	401	656

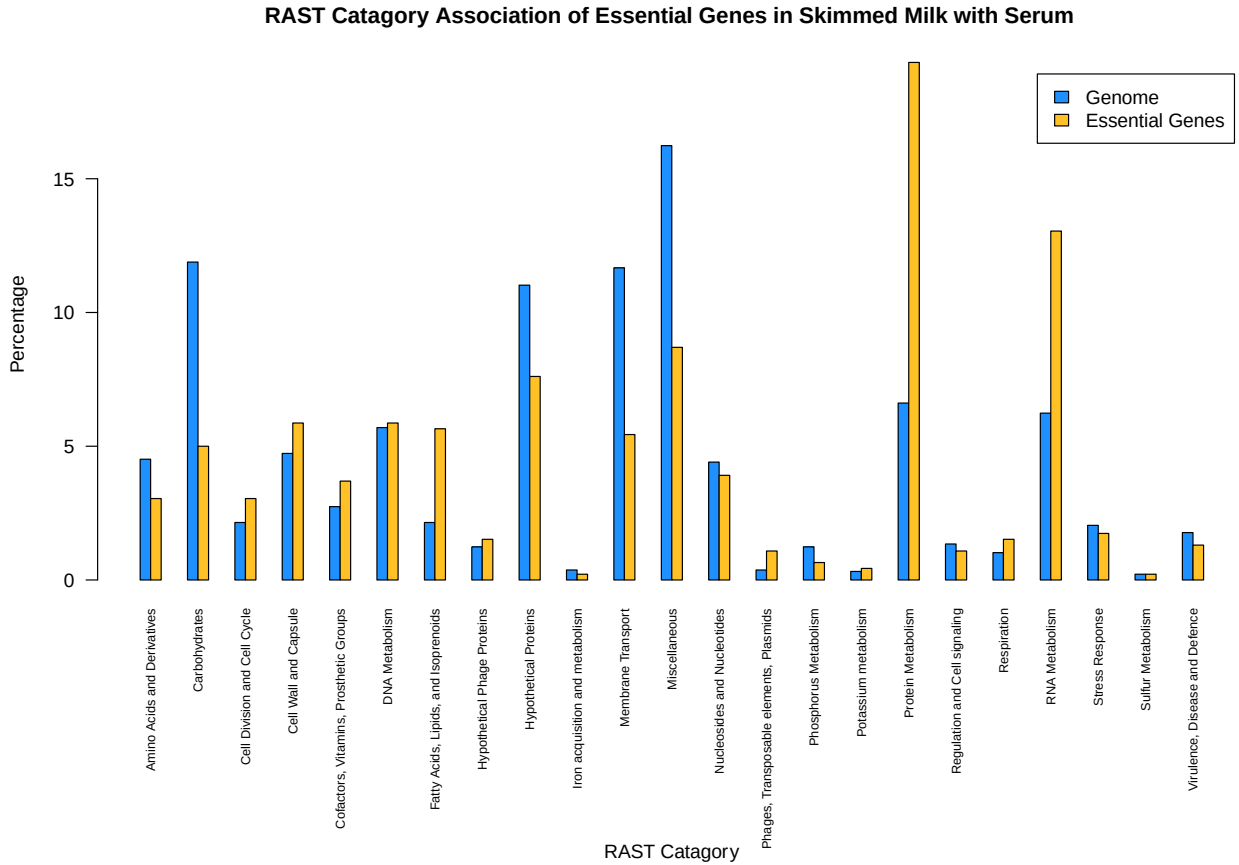

Figure 6.9. Proportional enrichment of essential genes associated with each RAST category. Their percentage of essential genes associated with each RAST category was calculated for the whole genome and those found to be essential for growth in skimmed milk with serum.

Table 6.15. Identification of RAST categories that show a level of enrichment after growth in milk with serum. Chisq.pval = the associated p-value showing the significance of the enrichment for the given RAST category.

Category	chisq.pval
Protein Metabolism	2.23344E-39
Miscellaneous	2.84166E-16
RNA Metabolism	4.06124E-15
Fatty Acids, Lipids, and Isoprenoids	1.39696E-08
Hypothetical Proteins	4.07498E-08
Carbohydrates	1.82645E-06
Membrane Transport	0.000176479
Phages, Transposable elements, Plasmids	0.000932243
Cell Division and Cell Cycle	0.004936832
Respiration	0.092597267
Cell Wall and Capsule	0.117985821
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.200817289
DNA Metabolism	0.278081513
Amino Acids and Derivatives	0.306703771
Phosphorus Metabolism	0.494246887
Hypothetical Phage Proteins	0.692993953
Virulence, Disease and Defence	0.715476612
Regulation and Cell signalling	0.749983421
Iron acquisition and metabolism	0.839208368
Nucleosides and Nucleotides	0.943428422
Stress Response	0.967421538
Potassium metabolism	0.987803718
Sulphur Metabolism	1

6.4.3.4. The use of differential insertion prevalence to identify phenotypically important genes

The data was used in edgeR statistical analysis and the output sorted in a descending order of false discovery rate (FDR) incorporating a threshold of <0.05. This identified seven genes in the skimmed milk with serum that were

found to have differential gene essentiality (Table 6.16). The genes *gauA, purH, purF* and SUB0026 were involved in nucleotide synthesis, whilst SUB0541 and SUB1416 were involved in carbohydrate metabolism. These genes were involved in metabolic pathways previously shown to be statistically important for growth in this environment, and therefore this result is not surprising. There were three CDSs with a fold difference from the mean (Log-FC) of greater than two (SUB0541, 0891 and 1416), of which one has substantially lower insertions in samples from the output pool compared to the input pool (SUB0891). This suggests *gauA*, a catalyst for purine metabolism, has an important role during growth in milk containing serum albeit one that is not essential.

Table 6.16. Genes with statistically significant fold change differences after screening in skimmed milk with serum

Loci	Gene	Function	log-FC	FDR*
SUB0891	guaA	GMP synthase [glutamine-hydrolyzing]	-3.21032501	2.61584E-14
SUB0026	NA	phosphoribosylformylglycinamidine synthase protein	-1.138423381	1.63161E-05
SUB0988	NA	sodium:alanine symporter family protein	-1.50966916	0.00415481
SUB0030	purH	bifunctional purine biosynthesis protein	-1.888193049	0.00982641
SUB0541	NA	dihydroxyacetone kinase family protein	-2.191361677	0.01005817
SUB0027	purF	amidophosphoribosyltransferase precursor	-1.234635486	0.01354875
SUB1416	NA	extracellular solute-binding protein	-2.733981639	0.01616898

* p-value adjusted Benjamini and Hochberg's procedure (Benjamini & Hochberg 1995).

6.5. Discussion

Selective antibiotics are used as part of the mutagenesis system to ensure the population of bacteria contain the mutant insertion. The antibiotic used in this study, erythromycin, inhibits protein synthesis by binding to the 50S subunit of the bacterial ribosome (Moazed & Noller 1987). This undoubtedly has an effect on bacterial replication, even with the resistance gene introduced by the pGhost9 vector. The use of antibiotics has also been suggested to enrich for mutations found within certain regulatory genes, therefore affecting the function of their cognate genes (Sun et al. 2010). Instead of having to account for this in the analyses, growth parameters were investigated to understand the effect on the number of mutant bacteria in the absence of the selective antibiotic (Figure 6.3). The number of bacteria counted was almost identical at each time point except for the transition between growth and stationary phase. The bacteria without the selective erythromycin were actively growing for longer, confirming the selective pressure on the bacteria. However the number of bacteria present during stationary phase was almost identical, meaning growth in the biological relevant environments could be achieved minus erythromycin without deleterious effects on the overall number of mutant bacteria.

6.5.1. Growth of S. uberis in skimmed milk

A study on gene expression related to growth in milk of *Staphylococcus aureus*

(*S. aureus*) highlighted 28 genes that were essential for growth on milk agar (Lammers et al. 2000). From the genes identified, eight were involved in peptidoglycan biosynthesis, five were related to DNA metabolism, two were transcriptional regulators and the remaining 13 were hypothetical genes. Comparison of these genes with the data generated in the current study revealed seven homologous genes that were identified as essential for growth of *S. uberis,* interestingly in milk containing serum rather than milk alone. The *S. uberis* essential genes *ptsK* and *phoU* are part of a two-component system, which is linked to environmental adaption and cellular invasion of mammary epithelial cells in *E. coli* (Sinai & Bavoil 1993). The genes *purH, purD and pyrD,* are responsible for nucleic acid biosynthesis, which are linked to milk specific replication (Lammers et al. 2000) were identified as part of the edgeR analyses for growth in milk containing serum. Finally, *murB* and *asd* are responsible for peptidoglycan biosynthesis a major component of the cell wall. The cell wall plays an important role in the resistance to lysozyme, but the reason why this would be essential *in vitro*, in the absence of lysozyme, is unclear. However, *asd* and *murB* are responsible for purine biosynthesis and their essential nature could be linked to replication rather than peptidoglycan biosynthesis.

Even though relatively few additional genes were identified for growth in skimmed milk, the gene *mtuA*, which was previously determined to be unable to grow in milk (Smith, A. J. et al. 2003) was identified as essential in these data. The role of *mtuA* in *S. uberis* pathogenesis was determined to be due to its involvement in acquisition of environmental manganese (Smith, A. J. et al.

2003). This was also found essential in the Todd Hewitt with lactoferrin data, which is unsurprising given the role of this protein in metal chelation. Both discoveries add confidence in the data gained from the phenotypic screening method.

Other genes essential for growth in milk, that are of interest, are associated with the Fatty acid biosynthesis pathway. In the data generated from the growth on solid media in Chapter 4, genes *eno* (input NRM 16.8) and *pyk* (input NRM 1.7) were non-essential. Both genes are part of the glycolysis pathway and convert glycerate-2P into pyruvate. However, in the milk output pool both genes were found to be essential. The difference in essential gene status for *pyk* may be explained by the alternative mechanisms *S. uberis* has for energy production through the generation of pyruvic acid, however in a favourable environment this may be the most efficient method for its production. Interestingly pyruvic kinase is regulated by the presence of glucose and has been shown to play a role in glucose catabolite repression (Cunningham et al. 2009). Catabolite repression allows bacteria to quickly adapt to use a preferred carbon source. In the case of the original solid agar, essential genes were directly required for metabolism of the main carbohydrate source glucose. While the main carbohydrate source in milk is lactose, thus requiring an alternative set of genes for metabolism.

A few of the CDSs that were identified as lacking transposon insertions have already been highlighted in previously published data. Sortase A (*srtA*)

(SUB0881) encodes a transamidase which is capable of covalently anchoring a specific subset of proteins to the peptidoglycan cell wall (Egan et al. 2010). The sortase-anchored proteins (SUB0135, 0145, 0207, 0241, 0826, 0888, 1095, 1154, 1370 and 1730) contain a conserved amino acid motif (LPxxG or LPxxxD/E) with a secretory leader peptide. These SrtA-anchored proteins, SUB0145, 1095 and 1154, have been shown to be of importance for infection. Strains in which these genes were inactivated were shown to be attenuated in an experimental infection model (Leigh et al. 2010). In the current data, *srtA* was found to be essential for proliferation in milk, however the individual sortase-anchored proteins were not. This may reflect a cumulative role for all SrtA-anchored proteins with regard for growth. Consequently, loss of any one of the sortase-anchored proteins does not have a major deleterious effect on bacterial growth in milk.

6.5.2. Growth of S. uberis in milk containing serum

The serum used in this study would have been devoid of any immune cells but would still have contained a variety of host proteins and other components that would enhance or inhibit bacterial growth. This would enable identification of genes utilised by *S. uberis* for acquisition of nutrients available in the mammary gland without interference from factors associated with the mammary gland immune response. The investigation into the amount of sera used to supplement the milk was calculated to be equivalent to that present in milk from late in the lactation cycle, sub-clinical and a clinically mastitic

mammary gland based on known bovine serum albumin concentration (0.3, 0.6 and 10mg/ml respectively) (Rainard 2003). Analysis of growth characteristics of wild type *S. uberis* 0140J in milk with variable concentrations of serum has shown an accelerated replication rate that correlates to the increase in serum concentration (Figure 6.5). This could have implications on the results generated from this study as a differential in fitness, could have resulted in a proportion of bacteria replicating more frequently, skewing the results. This competitive advantage could be reflected in the number of essential genes identified in certain RAST categories related to replication. The proportion of essential genes identified in the milk with serum output pool, that are associated with Protein and RNA metabolism, are shown to be statistically significant in the chi-squared analysis (Table 6.14) as 60 of the 116 (52%) RNA metabolism genes and 89 of 123 (77%) genes associated with protein metabolism are deemed essential. The genes deemed essential are all part of the transcription and translation process suggesting an altered level of protein expression, which could be an adaptive response to a new environment.

Studies that have investigated genes required for growth in the presence of serum have been mainly based on serum alone, not as an additional component of another environment. However the results they have generated are quite similar, identifying purine and pyrimidine biosynthesis pathways as essential for growth of *E. coli, Bacillus anthracis, Salmonella typhimurium* (Samant et al. 2008), and *S. aureus* (Christiansen et al. 2014). *Streptococcus*

pyogenes was used to identify genes required for growth of group A streptococci in whole human blood, which identified a more comprehensive list of essential genes (Le Breton et al. 2013). The gene lists identified carbohydrate metabolism, amino acids and derivatives and nucleotides and nucleosides as the functional categories most highly represented. These results are not wholly reflected in the current study with the exception of carbohydrate metabolism. However, this disparity could be related to the differences between genes required for growth in whole blood compared to milk containing serum. The transcription regulator *cpsY* was found to be essential for growth in human blood and has also been independently confirmed in be essential for growth of *Streptococcus iniae* (Allen & Neely 2011). The amino acid sequence for *cpsY* was parsed against the *S. uberis* protein CDSs and a homologous match was identified as SUB1004. This CDS was found to be essential in all nutritional environments and is therefore not limited to survival in serum, for *S. uberis.*

As the number of essential genes identified for the milk with serum input pool was greater than those identified for the milk input pool (Tables 6.1 and 6.8) bacterial DNA was re-extracted from the frozen glycerol stocks and the whole PIMMS process repeated, to ensure the results from the milk with serum were not erroneous. Sequencing of these additional samples resulted in a similar set of results, with any new mutational insertions being found within the same CDSs rather than disrupting any of the CDSs previously identified as essential. The additional CDSs identified by the milk with serum input pool mostly

contained a single insertion within each CDS. The absence or presence of such mutants may be explained by the fact that the input pools were generated from a sub-set of the entire mutant pool; therefore it is possible some mutants of low abundance were lost. The incidence of stochastic loss is a limitation of the plating method used in the test vs. control design. Comparison of the sequence data for both input pools showed little variation in the frequency of the unique insertions (every 42bp compared to every 32bp in the milk). This means that even with poor experimental efficiency, it is still possible to identify phenotypically important CDSs using this comparative method.

Similarly to the THB and THB supplemented with lactoferrin data, there were 39 CDSs, devoid of insertions only found in the milk input pool (Table 6.5) and 134 found only in the milk/serum input pool (Table 6.13). This could be due to the occurrence of hypercompetitive mutants, where a disruption of a CDS enables the bacterium to perform better in a certain environment compared to another. This has been shown in other TIS experiments where mutant strains have been found to out-compete other strains. Examples include *fnr* a regulatory gene involved in cellular migration in *Salmonella enterica* Serovar Typhimurium (Chaudhuri, et al. 2009), *srpA*, a host platelet aggregation protein in *Streptococcus sanguinis* (Turner et al. 2009) and pANL51, a protein of unknown function found in *Staphylococcus aureus* (Christiansen et al. 2014).

As previously discussed, the genes present in the *vic* operon are responsible for determining environmental suitability for proliferation as part of a two component signal transduction system (TCS) (Wagner & Saizieu 2002). In contrast to the milk output essential genes, *vicK* and *vicX* were found to be essential for growth in milk containing serum, with an input NRM score of 7 and 37, respectively. The *vic* operon has been found to regulate both pathogenicity in *Bacillus subtilis* (Appleby et al. 1996) and cellular viability in *Caulobacter crescentus* (Quon et al. 1996). Overexpression of the *vic* operon has been detected in *Streptococcus pneumoniae* in media supplemented with 5% fetal calf serum (Wagner & Saizieu 2002), indicating a potential importance for growth in the presence of serum. Potential drug inhibitors of TCSs are currently being investigated for use against multiple human pathogenic streptococci (Gotoh et al. 2010; Chauhan & Calderone 2008) and may prove useful in the future control strategies for *S. uberis* infection.

Chapter 7 General discussion

7.1. Benchmarking the PIMMS process with other TIS experiments

The PIMMS methodology described was an adaption and extension of techniques used previously to map individual mutations generated by the insertional mutagen pGhost9:IS*S1* (Maguin et al. 1996). This process has inferred functions of some of genetic features that have been investigated by the exploitation of single isogenic mutants generated using the random insertional mutagen, pGhost9:IS*S1*. Such studies have indicated roles in infection for three sortase (SrtA) surface anchored proteins; Sub1154, Sub1095 and Sub0145 (Leigh et al. 2010), the manganese ABC-transporter metal binding protein, MtuA (Smith, A. J. et al. 2003) and the stand alone regulator, Vru (Egan et al. 2012). Conversely, individual mutants lacking the ability to produce hyaluronate synthase, HasA (responsible for capsule biosynthesis) or the plasminogen activator, PauA (Ward et al. 2001; Ward et al. 2003) have been shown not to play a role in pathogenesis in isolation.

The movement towards high-throughput transposon insertion sequencing (TIS) has so far focused mainly on Gram-negative bacteria, with the exceptions of studies on *Streptococcus pneumonia* (van Opijnen et al. 2009) and *Staphylococcus aureus* (Christiansen et al. 2014). A comparison of these

methodologies provides a good benchmark for the efficiency of the PIMMS protocol. The number of unique mutations discovered compared to the number of mutants in these studies is relatively low. The investigation of *S. pneumoniae* using Tn-Seq found 23,875 unique insertions in a population of 150,000 mutants and *S. aureus* using TraDIS identified 140,330 unique mutations from 1,000,000 mutants. When compared to the *S. uberis* PIMMS data 80,617 unique mutations were identified from 100,000 mutants. This could be attributed to the efficiency of the pGhost^{+}9::IS*S1* mutagen compared to the *mariner* transposon used in Tn-Seq and Tn5 used in TraDIS. The pGhost9::IS*S1* mutagen does not have a transposition bias to a specific insertion motif (Green et al. 2012) or hold an insertional preference to specific structural features of DNA (Lampe et al. 1998). Both of which can lead to pseudo-random insertions and thus limiting the range and variability of mutations that can be created within a population.

In comparison to previous Gram-positive studies, one aim of the PIMMS based analysis was to increase the confidence of the precise location of every insertion position identified. To be captured for PIMMS analysis, firstly the sequence read must contain the terminal 23bp of the IS*S1* sequence, which allows for accurate identification of the insertion chromosome junction within each matched and subsequently mapped sequence. In comparison with other existing methods, the insertion identification in TraDIS is 10bp (Christiansen et al. 2014) and in Tn-seq (van Opijnen & Camilli 2010), none of the original transposon sequence is left within the sequence read. The QC of sequence

reads are not discussed in the Tn-seq protocol (van Opijnen & Camilli 2010) but TraDIS (Christiansen et al. 2014) uses a phred score threshold of >15, which therefore allows incorporation of sequences that have a probability of greater than 1 in 50 of containing an error. PIMMS applies a more stringent Phred score threshold of >30 which equates to a probable error rate of 1 in 1000. In PIMMS, all reads that satisfy these quality parameters must also contain a minimum of 20bp adjacent to the IS*S1* element to enable unambiguous genome mapping. In most cases (98.8%) these fragments reach the maximum permitted read length of 50bp. This compares favourably with other mapping systems; Tn-Seq uses 20bp and TraDIS uses 33bp for mapping. The stringency of the alignments were also not discussed for Tn-Seq and TraDIS, however in PIMMS, BWA-MEM is used for alignment (Li 2013), due to its accuracy with short un-gapped reads without any mismatches.

7.2. Identification of truncated CDSs that are consistently found in these data

The incidence of truncated genes during insertional mutagenesis experiments is not always identified. However, in these data, there are a few small groups of truncated CDSs that consistently appear in certain environments. The first group of truncated CDSs are only found when the bacteria are grown in liquid medium. The ATP-dependant metalloprotease *ftsH*, associated with cell division (Nilsson et al. 1994), only contains insertions in the terminal 4% of the

CDS when the bacteria is grown in liquid media. This region encodes for one of three zinc-binding sites. The role of *ftsH* has been implicated with the degradation of *ftsZ* in a Zn^{2+} dependant manner (Srinivasan et al. 2008), suggesting the reduction in one binding site may lead to decreased efficiency, whilst maintaining overall function.

The other three identified have not been subjected to as much research and therefore have a lesser-known impact on bacterial growth. The RuvA-RuvB complex promotes strand exchange reactions in recombination (Majewski et al. 2000). The truncated *ruvB* found in these data, only contains insertions within the terminal 2% of the coding sequence equating to the loss of 6 amino acids. Co-enzyme E (*coaE*) is an integral part of the fatty acid biosynthesis pathway. Its role is the production of co-enzyme A (*coaA*) a precursor to acyl-CoA, which is used in fatty acid elongation. CoaE was found to be essential in *Streptococcus mutans* (Bing Wang & Kuramitsu 2005). SUB0426 is an uncharacterised membrane protein involved in polysaccharide biosynthesis. It contains an RfbX domain, which are found in proteins responsible for export of teichoic acid to the cell wall. The CDS only contains insertions within the terminal 5%, which may involve the loss of the C-terminal membrane region (Miyafusa et al. 2012).

The second group comprise of two CDSs that are found only in complex liquid media. The cell division protein, *ftsW,* only contains insertions at 95%, loosing 21 amino acids. It has been suggested that *ftsW* may play a role in the

establishment of a stable *ftsZ* based structure, at septum formation, during bacterial division (Khattar et al. 1997). Whereas *ribC,* a riboflavin biosynthesis protein, only has insertions at 96% of its CDS, loosing 12 amino acids. This region is an ATP binding site, which has been shown to control the flavin mononucleotide synthesase activity in *Bacillus subtilis*, a biomolecule that functions as a prosthetic group in various oxidoreductases (Karelov et al. 2011).

Finally those that are always found in growth on solid media consist of *mtnN, bacA* and *pgm*. The CDS *mtnN* is part of the amino acid biosynthesis pathway. The insertions found within the CDS are in the terminal 5%, which encodes a proton donor (Lee et al. 2005). Similarly *bacA* only contains insertions in the terminal 10% of the CDS. This encodes for a bacitracin resistance protein. Bacitracin is a peptide antibiotic produced by certain species of *Bacillus* (Storm 1974). A strain of *S. pneumoniae* with a disrupted *bacA* gene was shown to be highly attenuated in a murine infection model (Chalker et al. 2000). However, *in vivo* it plays a role in the regulation of cell biosynthesis during late exponential growth phase in *Enterococcus faecalis* (Umbreit et al. 1972). The terminal sequence of *bacA* encodes for one of four transmembrane helices, which may account for the ability of this CDS to still function, if the protein is still bound to the membrane by three other helices. The CDS *pgmA,* a phosphomannomutase/phosphoglucomutase, only has insertions within the terminal 3%, which equates to 17 amino acids. It plays an important role in cell wall morphology and capsule expression (Hardy et al. 2000) and has been

found to play a role in virulence in *S. pneumoniae* in a murine infection model (Hardy et al. 2001). It has also been found to infer resistance of *Streptococcus iniae* against the innate immune response of fish (Buchanan et al. 2005).

7.3. Conserved essential genes between different clinical and sub-clinical isolates of *S. uberis*

A comparison of genetic composition between 13 isolates of *S. uberis* with varying virulence status, was recently completed (Hossain et al. 2015). The isolates represented were the highly virulent strain 0140J (reference strain), an avirulent strain EF20 and 11 isolates from various clinical and sub-clinical mammary gland infections of lactating dairy cattle (Table 7.1).

Table 7.1. Isolates used for homology comparison and the case status

Isolate	Strains	Number of CDS
0140J	Virulent*	1858
C6344	Clinical	1999
C5072	Clinical	2004
S6261	Clinical	1970
EF20	Avirulent*	1995
6736	Sub-clinical	1980
C8329	Sub-clinical	2020
Ab71	Clinical	1886
C9359	Clinical	1918
C5388	Sub-clinical	1918
B190	Sub-clinical	1967
6780	Clinical	2011
B362	Sub-clinical	1935

(*Hill 1988; other data from Hossain et al. 2015)

This led to the identification of a *S. uberis* core genome, which are genes conserved between all 13 isolates, and accessory genes, which are those

unique to each isolate. The core genome created from the comparative genomics study was used in conjunction with the data generated using the PIMMS process to identify the 0140J essential accessory genes. This could identify unique genes present in 0140J that could explain its increased virulence when compared to other strains. There are 54 genes found only in the 0140J strain that have been found to be essential in this study (Table 7.2), of which, any could attribute to the higher rate of virulence of this strain when compared to other isolates. The smaller genome size could also offer a fitness advantage with a more optimised process for nutrient acquisition and energy production or a lower energy requirement during replication.

The essential accessory gene list has a high number of putative membrane proteins and all of the miscellaneous genes were also found to be associated membrane proteins. However there is no distinction as to whether they are responsible for transport, cell adhesion, enzymatic activity or membrane receptor relays. Investigation of some individual genes highlighted a few of interest. Secretion of and resistance to bacteriocins has been shown to play an important role in competitive fitness in virulent *Escherichia coli* (*E. coli*) strains (Durso et al. 2004; Lenski et al. 1998). The reference *S. uberis* (0140J) has three bacteriocin CDSs (SUB0502, 0504 and 0505) that are not found in EF20.

Table 7.2. Essential genes identified in skimmed milk and with serum that are not conserved between different isolates of *S. uberis*
Each gene was associated to its RAST category based on the subsystem of function (Overbeek et al. 2005). Putative virulence proteins are highlighted in red, as determined by VirulentPred (Garg & Gupta 2008).

RAST Category	Count	Genes
Carbohydrates	3	*glkA,* SUB0293, 1205
Cell Division and Cell Cycle	1	SUB0476
Cell Wall and Capsule	3	SUB0699, 0700, 1031
DNA Metabolism	2	*dnaI,* SUB1710
Fatty Acids, Lipids, and Isoprenoids	1	*acpS*
Hypothetical Phage Proteins	7	SUB0219, 0224, 0227, 1819, 1827, 1834, 1837
Hypothetical Proteins	8	SUB0064, 0149, 0193, 0292, 0624, 1061, 1597, 1817
Membrane Transport	8	SUB0211, 0233, 0504, 0511, 1033, 1150, 1578, 1590
Miscellaneous	3	SUB0772A, 1135, 1832A
Nucleosides and Nucleotides	1	SUB0292A
Phages, Transposable elements, Plasmids	7	SUB228, 0229, 1824, 1831, 1832, 1838, 1839
Protein Metabolism	4	*rplD, rpmGB, rpmJ,* SUB1732A
Respiration	1	SUB1592
RNA Metabolism	1	SUB0726
Virulence, Disease and Defence	4	*pedA,* SUB0502, 0504, 0505
Total	54	

Other than the genes associated with DNA and RNA metabolism, none of the other 0140J accessory genes have a homologous match with a known gene in other bacterial species apart from the exopolysaccharide biosynthesis protein (SUB1031). This has been found to be important for growth of certain strains of *Streptococcus thermophillus* in skimmed milk (De Vuyst et al. 1998).

There are a high number of hypothetical proteins, which are surprisingly deemed essential and are unique to the 0140J genome. Some hypothetical proteins have been shown to affect virulence through the mimicry of host immune signalling pathways, subverting cell processes (Newman et al. 2006), or inferring resistance to bacteriocins (Nascimento et al. 2012) and antibiotics (Holden et al. 2004). To investigate the role of the essential unique hypothetical proteins in *S. uberis 0140J*, the amino acid sequences were parsed through CDD-BLAST to identify any conserved domains to determine protein function (Marchler-Bauer et al. 2015). There were two domain similarities found within SUB1817 and SUB1837. The first, SUB1817, has a nudix-hydrolase domain at the N terminus of the protein. They require a divalent cation either magnesium (Mg) or manganese (Mn) and their main role is hydrolyse nucleoside triphosphates (McLennan 2006; Bessman et al. 1996) preventing AT:CG transversions during replication (Bhatnagar et al. 1991). This could be linked to the requirement for Mn for growth of *S. uberis* previously discussed (Smith, A. J. et al. 2003). The second protein with an identified domain was SUB1837 phage protein, which might infer some protection against phage infection (Henthorn & Friedman 1995). However it

has been suggested that this gene is an integration host factor, which regulates the expression of flagellin, in *E. coli* (Craig & Nash 1984). As streptococci do not have flagella the regulation of such a protein is unlikely, however the prospect of the phage protein regulating another CDS that aids virulence is possible.

VirulencePred (Garg & Gupta 2008) is an analysis package that offers an alternative method to identify putative virulence factors by using amino acid, dipeptide and high order dipeptide sequences to compare composition to known virulent factors using PSI-BLAST (Altschul et al. 1997). The amino acid sequences from the hypothetical genes were parsed thought VirulencePred using an implemented cut-off of >1, ensuring an 80% likeliness to amino acid similarity over the length of the sequence. From the fifteen sequences submitted, eleven were found to have some sequence similarity to known virulence factors in other bacterial species (Table 7.2). Unfortunately this only identifies potential matching of motifs and does not offer any further information of what virulence determinants they match. Interestingly the phage resistance protein (SUB1837) was one of the identified virulence determinants.

With the information presented by the phenotypic screening process and the comparative genomics, there are certain connections to be made. These can help formulate a putative rationale behind the increased virulence of the 0140J strain. The possibility of enhanced expression levels by the phage

protein SUB1837 of virulence factors. The benefits of a smaller genome could be accompanied by a more efficient transcription method by *dnaI* (SUB0383). The 0140J strain of *S. uberis* has a high number of genes that are essential for growth in milk supplemented with serum. This could be because it is highly optimised for this environment, which is supported by increased growth of wild type *S. uberis* 0140J in the presence of high levels of serum in milk when compared to growth in milk alone (Figure 6.5). This could be linked with the wide array of essential membrane transport proteins, which could help acquire nutrients from the environment. The array of 0140J accessory hypothetical proteins that are essential for growth in a cell-density dependent environment could be linked to the postulated neutrophil toxin (Leigh & Field 1994). Also due to 0140J having these additional virulence factors the capsule formation *has* operon, found not essential for 0140J growth (Field et al. 2003), might be more important for less virulent strains.

7.4. Future work

7.4.1. Linking more phenotypes to genotypes

The data generated in this study has highlighted, that in the presence of certain nutritional factors, *S. uberis* requires a diverse complement of genetic factors in order to proliferate. The ability of *S. uberis* to persist in various niches, suggests it has multiple survival systems at its disposal. In order to elucidate these phenotypic differences PIMMS could investigate the impact of disruptions in certain coding sequences throughout a range of relevant

environments. The ability of *S. uberis* to survive outside of the bovine host and within milking machinery would provide useful data. However, there are also numerous other host specific conditions that could be investigated such as, colostrum, dry cow secretions and growth in the presence of antimicrobial peptides.

These *in vitro* environments can mimic nutritional factors faced by the bacteria, however during the infection of the mammary gland the environment alters. There is an increase in the permeability of the blood-milk barrier, leading to an elevated level of host derived antimicrobial peptides and the influx of immune cells. This could lead to the succession from an initial subset of genes being required to initiate infection and a further subset needed to survive the host response, which alters as infection progresses. The movement to *in vivo* testing will provide a list of genes that are directly involved in colonisation, persistence and infection in the face of a mounting host response.

7.4.2. Application to other bacteria

There are a number of veterinary diseases and zoonoses, which could benefit from the implementation of technologies to identify all key genes involved in virulence during a given environmental niche. The pGhost9::IS*S1* temperature sensitive insertional mutagen has been successfully used to generate random mutants in *Streptococcus*, *Lactococcus* and *Enterococcus* (Maguin et al. 1996;

Haaber et al. 2008; Ward & Leigh 2004; Asam et al. 2013); genera that are typically not amenable to insertional mutagenesis. By combining high-density random mutagenesis and massively parallel DNA sequencing, PIMMS exploits the ability to generate large quantities of individual sequence reads and is able to comprehensively and simultaneously identify multiple mutated coding sequences. The ability to utilise populations mutagenised with pGhost9::IS*S1* in this manner is a significant advance in the tools available for functional genomics of streptococci; and related bacterial genera. The PIMMS technology can be applied to identify key determinants from bacteria of specific interest with veterinary importance such as other mastitic pathogens *Streptococcus dysgalactiae* subsp. *dysgalactiae* and *Streptococcus agalactiae, Streptococcus suis* in pigs, *Streptococcus iniae* in the farmed fisheries or *Streptococcus equi* subspecies *equi* in horses that all have a large financial and welfare impact. The methods can also be used to study bacteria with human medical importance such as *Streptococcus pneumoniae* and *Streptococcus pyogenes,* which contribute to mortality and morbidity ranging from pneumonia to meningitis and rheumatic heart disease but also Enterococcus species that contribute to endocarditis, meningitis and dental disease amongst others. The identification of essential genes is not limited to those which are involved in pathogenesis and can also be applied to improve food production, for example identification of key genes required for Lactococcus species, used for cheese and yogurt production.

7.4.3. Genetic regulation by non-coding chromosomal features

Bacterial genomes are considered to mostly consist of genes for protein synthesis and structural RNAs, with small intragenic regions. The presence of non-coding DNA seems to be determined by selective pressures to minimise non-functional DNA, while maintaining advantageous regulatory signals (Rogozin et al. 2002). The initial theory of non-coding genomic regions was proposed in the early 70s (Spiegelman et al. 1972), however their presence was not confirmed until 1980, with the identification of these regions producing non-coding RNAs (Itoh & Tomizawa 1980). There are several classes of non-coding RNAs, which are often transcribed in the opposite direction to the DNA template (Wagner & Simons 1994). The two main classes of non-coding RNAs which either act as post transcription regulators or riboswitches (Westhof 2010). Riboswitches are regions of messenger RNA (mRNA), which regulates transcription and translation of the cognate gene found in the 5'-untranslated region of the mRNA (5'-UTR) and often regulate biosynthesis, membrane transport or catabolism (Serganov & Patel 2009). However, in *Vibrio cholera* a riboswitch has been implicated in the regulation of virulence associated genes (Kulshina et al. 2009). A region of non-coding genome affecting the transcription of a CDS has previously been discovered in *S. uberis*. An insertion discovered in the upstream non-coding region of SUB0538 was found to have a direct impact on protein expression for SUB0538 and SUB0539, which are both glycosyltransferases (Czabańska 2013). This information could prove a useful insight into the role of non-coding genetic motifs in the *S. uberis* genome. Further research could lead to the

identification of regions that are important for bacterial virulence but are often overlooked, as they are not within a protein coding sequence.

References

Abuaita, B.H. & Withey, J.H., 2011. Genetic Screening for Bacterial Mutants in Liquid Growth Media By Flouorescence-Activated Cell Sorting. *Journal of Microbiological Methods*, 84(1), pp.109–113.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., et al., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology*, 12(2), p.R18.

Allen, J.P. & Neely, M.N., 2011. The *Streptococcus iniae* transcriptional regulator *cpsY* is required for protection from neutrophil-mediated killing and proper growth in vitro. *Infection and Immunity*, 79(11), pp.4638–4648.

Almeida, R.A., Luther, D.A., Park, H.-M. & Oliver, S.P., 2006. Identification, isolation, and partial characterization of a novel *Streptococcus uberis* adhesion molecule (SUAM). *Veterinary microbiology*, 115(1-3), pp.183–91.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–10.

Altschul, S.F., Madden, T.L., Schäffer, A. a., Zhang, J., et al., 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17), pp.3389–3402.

Anaya, D.A., McMahon, K., Nathens, A.B., Sullivan, S.R., et al., 2013. Predictors of mortality and limb loss in necrotizing soft tissue infections. *American Surgeon*, 140(10), pp.1102–1105.

Anders, S., McCarthy, D., Chen, Y., Okoniewski, M., et al., 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, 8(9), pp.1765–1786.

Appleby, J.L., Parkinson, J.S. & Bourret, R.B., 1996. Signal Transduction via the Minireview Multi-Step Phosphorelay: Not Necessarily a Road Less Traveled. *Cell*, 86, pp.1–4.

Arber, W., 2000. Genetic variation: Molecular mechanisms and impact on microbial evolution. *FEMS Microbiology Reviews*, 24(1), pp.1–7.

Arnold, R.R., Russell, J.E. & Champion, W.J., 1982. Bactericidal activity of human lactoferrin: Differentiation from the stasis of iron deprivation. *Infection and Immunity*, 35(3), pp.792–799.

Asam, D., Mauerer, S., Walheim, E. & Spellerberg, B., 2013. Identification of β-haemolysin-encoding genes in *Streptococcus anginosus*. *Molecular oral microbiology*, 28(4), pp.302–15.

Bae, T., Banger, A.K., Wallace, A., Glass, E.M., et al., 2004. *Staphylococcus aureus* virulence genes identified by bursa aurealis mutagenesis and nematode killing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33), pp.12312–12317.

Baeck, G.W., Kim, J.H., Gomez, D.K. & Park, S.C., 2006. Isolation and characterization of *Streptococcus sp*. from diseased flounder in Jeju Island. *Journal of veterinary science*, 7, pp.53–58.

Van Bakel, I., Holt, S., Craig, I. & Boyd, Y., 1995. Sequence analysis of the breakpoint regions of an X;5 translocation in a female with Duchenne muscular dystrophy. *American journal of human genetics*, 57(Cockburn 1991), pp.329–336.

Baker, C.J., 2013. The spectrum of perinatal group B streptococcal disease. *Vaccine*, 31(S4), pp.D3–D6.

Bannerman, D., Paape, M., Goff, J., Kimura, K., et al., 2004. Innate immune response to intramammary infection with Serratia marcescens and *Streptococcus uberis*. *Veterinary Reserach*, 35, pp.681–700.

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

Batavani, R.A., Asri, S., Naebzadeh, H., R.a, B., et al., 2007. The Effect of Subclinical mastitis on milk composition in dairy cows. *Iranian Journal of Veterinay Reserch*, 8(3), pp.205–211.

Baureder, M. & Hederstedt, L., 2012. Genes important for catalase activity in *Enterococcus faecalis*. *PloS one*, 7(5), p.e36725.

Van Belkum, A., Van Leeuwen, W., Kaufmann, M.E., Cookson, B., et al., 1998. Assessment of resolution and intercenter reproducibility of results of genotyping *Staphylococcus aureus* by pulsed-field gel electrophoresis of SmaI macrorestriction fragments: A multicenter study. *Journal of Clinical Microbiology*, 36(6), pp.1653–1659.

Bender, J. & Kleckner, N., 1986. Genetic evidence that Tn10 transposes by a nonreplicative mechanism. *Cell*, 45(6), pp.801–15.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp.289 – 300.

Benson, D. a., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., et al., 2013. GenBank. *Nucleic Acids Research*, 41(D1), pp.36–42.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Geoffrey, P., et al., 2008. Accurate Whole Genome Sequencing using Reversible Terminator Chemistry. *Nature*, 456(7218), pp.53–59.

Bentley, R.W., Leigh, J.A. & Collins, M.D., 1991. Intrageneric structure of *Streptococcus* based on comparative analysis of small-subunit rRNA sequences. *International journal of systematic bacteriology*, 41(4), pp.487–94.

Bessman, M.J., Frick, D.N. & O'Handley, S.F., 1996. The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *Journal of Biological Chemistry*, 271(41), pp.25059–25062.

Bhatnagar, S.K., Bullions, L.C. & Bessman, M.J., 1991. Characterization of the *mutT* nucleoside triphosphatase of *Escherichia coli*. *Journal of Biological Chemistry*, 266(14), pp.9050–9054.

Biggs, A., 2009. *Mastitis in Cattle*, The Crowood Press Ltd U.K.

Bishop, A.H., Rachwal, P. a & Vaid, A., 2014. Identification of genes required by *Bacillus thuringiensis* for survival in soil by transposon-directed insertion site sequencing. *Current microbiology*, 68(4), pp.477–85.

Biswas, S. & Biswas, I., 2011. Role of *vltAB*, an ABC transporter complex, in viologen tolerance in *Streptococcus mutans*. *Antimicrobial agents and chemotherapy*, 55(4), pp.1460–9.

Blakeborough, P., Salter, D.N. & Gurr, M.I., 1983. Zinc binding in cow's milk and human milk. *The Biochemical journal*, 209(2), pp.505–512.

Blanchard, A.M., Leigh, J.A., Egan, S.A. & Emes, R.D., 2015. Transposon insertion mapping with PIMMS – Pragmatic Insertional Mutation Mapping System. *Frontiers in Genetics*, 6(April), pp.1–8.

Bogarad, L.D. & Deem, M.W., 1999. A hierarchical approach to protein molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6), pp.2591–2595.

Bouvert, E. & Battesti, A., 2011. General stress responses. In G. Storz & R. Hengge, eds. *Bacterial Stress Responses*. Washington D.C.: American Society for Microbiology (ASM), pp. 231–250.

Bradley, A.J., 2002. Bovine Mastitis: An Evolving Disease. *The Veterinary Journal*, 164(2), pp.116–128.

Bradley, A.J., Leach, K.A., Breen, J.E., Green, L.E., et al., 2007. Survey of the incidence and aetiology of mastitis on dairy farms in England and Wales. *Veterinary Record*, (February 2007).

Bramley, A.J., 1984. *Streptococcus uberis* udder infection - a major barrier to reducing mastitis incidence. *The British veterinary journal*. 140(4), pp.328-335.

Bramley, A.J. & Kruze, J., 1982. Sources of *Streptococcus uberis* in the dairy herd. *Journal of Dairy Research*, (49), pp.369–373.

Brandt, C.M. & Spellerberg, B., 2009. Human infections due to S*treptococcus dysgalactiae* subspecies *equisimilis*. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 49(5), pp.766–772.

Le Breton, Y., Mistry, P., Valdes, K.M., Quigley, J., et al., 2013. Genome-wide identification of genes required for fitness of group a *Streptococcus* in human blood. *Infection and Immunity*, 81(3), pp.862–875.

Brinster, S., Lamberet, G., Staels, B., Trieu-Cuot, P., et al., 2009. Type II fatty acid synthesis is not a suitable antibiotic target for Gram-positive pathogens. *Nature*, 458(7234), pp.83–86.

Brown, J.S., Gilliland, S.M. & Holden, D.W., 2001. A *Streptococcus pneumoniae* pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Molecular Microbiology*, 40(3), pp.572–585.

Broyles, L.N., Van Beneden, C., Beall, B., Facklam, R., et al., 2009. Population-based study of invasive disease due to beta-hemolytic streptococci of groups other than A and B. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 48(6), pp.706–712.

Brukner, I., Sánchez, R., Suck, D. & Pongor, S., 1995. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *The EMBO journal*, 14(8), pp.1812–1818.

Buchanan, J.T., Stannard, J. a, Lauth, X., Ostland, V.E., et al., 2005. Streptococcus iniae Phosphoglucomutase Is a Virulence Factor and a Target for Vaccine Development *Streptococcus iniae* Phosphoglucomutase Is a Virulence Factor and a Target for Vaccine Development. *American Society for Microbiology*, 73(10), pp.6935–6944.

Byrne, R.T., Chen, S.H., Wood, E. a, Cabot, E.L., et al., 2014. Surviving extreme exposure to ionizing radiation: *Escherichia coli* genes and pathways. *Journal of bacteriology*, (July).

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., et al., 2009. PASS: A program to align short sequences. *Bioinformatics*, 25(7), pp.967–968.

Carlsson, a, Björck, L. & Persson, K., 1989. Lactoferrin and lysozyme in milk during acute mastitis and their inhibitory effect in Delvotest P. *Journal of dairy science*, 72(12), pp.3166–75.

Case, T.J. & Gilpin, M.E., 1974. Interference Competition and Niche Theory. *Proceedings of the National Academy of Sciences*, 71(8), pp.3073–3077.

Castro, R., Neves, A.R., Fonseca, L.L., Pool, W. a., et al., 2009. Characterization of the individual glucose uptake systems of *Lactococcus lactis*: Mannose-PTS, cellobiose-PTS and the novel *glcU* permease. *Molecular Microbiology*, 71(3), pp.795–806.

Chalker, A.F., Ingraham, K. a., Lunsford, R.D., Bryant, A.P., et al., 2000. The bacA gene, which determines bacitracin susceptibility in *Streptococcus pneumoniae* and *Staphylococcus aureus*, is also required for virulence. *Microbiology*, 146(7), pp.1547–1553.

Chaneton, L., Tirante, L., Maito, J., Chaves, J., et al., 2008. Relationship between milk lactoferrin and etiological agent in the mastitic bovine mammary gland. *Journal of dairy science*, 91(5), pp.1865–73.

Chanter, N., 1997. Streptococci and enterococci as animal pathogens. *Society for Applied Bacteriology symposium series*, 26, p.100S–109S.

Chaudhuri, R.R., Allen, A.G., Owen, P.J., Shalom, G., et al., 2009. Comprehensive identification of essential *Staphylococcus aureus* genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC genomics*, 10, p.291.

Chaudhuri, R.R., Peters, S.E., Pleasance, S.J., Northen, H., et al., 2009. Comprehensive identification of *Salmonella enterica* serovar typhimurium genes required for infection of BALB/c mice. *PLoS Pathogens*, 5(7).

Chauhan, N. & Calderone, R., 2008. Two-component signal transduction proteins as potential drug targets in medically important fungi. *Infection and Immunity*, 76(11), pp.4795–4803.

Chen, J.L., Fullerton, K.E. & Flynn, N.M., 2001. Necrotizing fasciitis associated with injection drug use. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 33(1), pp.6–15.

Chen, W.H., Minguez, P., Lercher, M.J. & Bork, P., 2012. OGEE: An online gene essentiality database. *Nucleic Acids Research*, 40(November 2011), pp.901–906.

Cheng, J.B., Wang, J.Q., Bu, D.P., Liu, G.L., et al., 2008. Factors affecting the lactoferrin concentration in bovine milk. *Journal of dairy science*, 91(3), pp.970–976.

Chopin, M. & Chopin, A., 1989. Insertion and amplification of foreign genes in the *Lactococcus lactis* subsp. *lactis* chromosome. *Applied and environmental microbiology*, 55(7), pp.1769–1774.

Chowers, Y., Holtmeier, W., Morzycka-Wroblewska, E. & Kagnoff, M.F., 1995. Inverse PCR amplification of rare T cell receptor messages from mucosal biopsy specimens. *Journal of Immunological Methods*, 179(94), pp.261–263.

Christiansen, M.T., Kaas, R.S., Chaudhuri, R.R., Holmes, M.A., et al., 2014. Genome-Wide High-Throughput Screening to Investigate Essential Genes Involved in Methicillin-Resistant *Staphylococcus aureus* Sequence Type 398 Survival. *PloS one*, 9(2), p.e89018.

Clancy, A., Loar, J.W., Speziali, C.D., Oberg, M., et al., 2006. Evidence for siderophore-dependent iron acquisition in group B *Streptococcus*. *Molecular Microbiology*, 59(2), pp.707–721.

Clifford, D.P. & Repine, J.E., 1982. Hydrogen peroxide mediated killing of bacteria. *Molecular and Cellular Biochemistry*, 49(3), pp.143–149.

Cluzel, P.J., Chopin, a., Ehrlich, S.D. & Chopin, M.C., 1991. Phage abortive infection mechanism from *Lactococcus lactis* subsp. *lactis*, expression of which is mediated by an Iso-ISS1 element. *Applied and Environmental Microbiology*, 57(12), pp.3547–3551.

Coffey, T. & Pullinger, G., 2006. First insights into the evolution of *Streptococcus uberis*: a multilocus sequence typing scheme that enables investigation of its population biology. *Applied and environmental microbiology*, 72(2), pp.1420–1428.

Conway, T., 1992. The Entner-Doudoroff pathway: history, physiology and molecular biology. *FEMS microbiology reviews*, 9(1), pp.1–27.

Cookson, B.D., Aparicio, P., Deplano, A., Struelens, M., et al., 1996. Inter-centre comparison of pulsed-field gel electrophoresis for the typing of methicillin-resistant *Staphylococcus aureus*. *Journal of Medical Microbiology*, 44(3), pp.179–184.

Corbin, B.D., Wang, Y., Beuria, T.K. & Margolin, W., 2007. Interaction between cell division proteins FtsE and FtsZ. *Journal of Bacteriology*, 189(8), pp.3026–3035.

Cornforth, D.M. & Foster, K.R., 2013. Competition sensing: the social side of bacterial stress responses. *Nature reviews. Microbiology*, 11(4), pp.285–93.

Craig, N.L. & Nash, H.A., 1984. E. coli integration host factor binds to specific sites in DNA. *Cell*, 39(3), pp.707–716.

Crisóstomo, M.I., Vollmer, W., Kharat, A.S., Inhülsen, S., et al., 2006. Attenuation of penicillin resistance in a peptidoglycan O-acetyl transferase mutant of *Streptococcus pneumoniae*. *Molecular Microbiology*, 61(6), pp.1497–1509.

Cunningham, D.S., Liu, Z., Domagalski, N., Koepsel, R.R., et al., 2009. Pyruvate kinase-deficient *Escherichia coli* exhibits increased plasmid copy number and cyclic AMP levels. *Journal of Bacteriology*, 191(9), pp.3041–3049.

Cvitkovitch, D.G., Li, Y.H. & Ellen, R.P., 2003. Quorum sensing and biofilm formation in streptococcal infections. *Journal of Clinical Investigation*, 112(11), pp.1626–1632.

Czabańska, A., 2013. Immunochemical investigations of the cell envelope components isolated from *Streptococcus uberis*. MSc Thesis, University of Lubeck.

Davidson, A.L., Dassa, E., Orelle, C. & Chen, J., 2008. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiology and molecular biology reviews : MMBR*, 72(2), pp.317–364, table of contents.

Davidson, L. a & Lönnerdal, B., 1989. Fe-saturation and proteolysis of human lactoferrin: effect on brush-border receptor-mediated uptake of Fe and Mn. *The American journal of physiology*, 257(6 Pt 1), pp.G930–G934.

Davis, K.M., Akinbi, H.T., Standish, A.J. & Weiser, J.N., 2008. Resistance to mucosal lysozyme compensates for the fitness deficit of peptidoglycan modifications by *Streptococcus pneumoniae*. *PLoS Pathogens*, 4(12).

Deutscher, J., Francke, C. & Postma, P.W., 2006. How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiology and molecular biology reviews : MMBR*, 70(4), pp.939–1031.

Diernhofer, K., 1932. Aesculinbouillon als Hilfsmittel für die Differenzierung von Euter-und Milchstreptokokken bei Massenuntersuchungen. *Milchwirtschaftl. Forsch*, 13, pp.368–378.

Douglas, V.L., Fenwick, S.G., Pfeiffer, D.U., Williamson, N.B., et al., 2000. Genomic typing of *Streptococcus uberis* isolates from cases of mastitis, in New Zealand dairy cows, using pulsed-field gel electrophoresis. *Veterinary Microbiology*, 75(1), pp.27–41.

Driessen, a J., Fekkes, P. & van der Wolk, J.P., 1998. The Sec system. *Current opinion in microbiology*, 1(2), pp.216–222.

Durso, L.M., Smith, D. & Hutkins, R.W., 2004. Measurements of Fitness and Competition in Commensal *Escherichia coli* and *E . coli* O157 : H7 Strains Measurements of Fitness. *Society*, 70(11), pp.6466–6472.

Eaton, S.S., Dubach, J., More, K.M., Eaton, G.R., et al., 1989. Comparison of the electron spin echo envelope modulation (ESEEM) for human lactoferrin and transferrin complexes of copper(II) and vanadyl ion. *Journal of Biological Chemistry*, 264(9), pp.4776–4781.

Edmond, K.M., Kortsalioudaki, C., Scott, S., Schrag, S.J., et al., 2012. Group B streptococcal disease in infants aged younger than 3 months: Systematic review and meta-analysis. *The Lancet*, 379(9815), pp.547–556.

Egan, S. a, Ward, P.N., Watson, M., Field, T.R., et al., 2012. Vru (Sub0144) controls expression of proven and putative virulence determinants and alters the ability of *Streptococcus uberis* to cause disease in dairy cattle. *Microbiology (Reading, England)*, 158(Pt 6), pp.1581–92.

Egan, S.A., Kurian, D. & Ward, P.N., 2010. Identification of sortase A (SrtA) substrates in *Streptococcus uberis*: evidence for an additional hexapeptide (LPXXXD) sorting motif. *Journal of Proteome Research*, pp.1088–1095.

Eichenbaum, Z., Muller, E., Morse, S. a. & Scott, J.R., 1996. Acquisition of iron from host proteins by the group A *Streptococcus*. *Infection and Immunity*, 64(12), pp.5428–5429.

Ellison, R.T., LaForce, F.M., Giehl, T.J., Boose, D.S., et al., 1990. Lactoferrin and transferrin damage of the gram-negative outer membrane is modulated by Ca2+ and Mg2+. *Journal of general microbiology*, 136(7), pp.1437–1446.

Enright, A.J., Dongen, S. V & Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), pp.1575–1584.

Eriani, G., Delarue, M., Poch, O., Jean, G., et al., 1990. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Letters To Nature*, 346, pp.183–187.

Eriksen, T.A., Kadziola, A., Bentsen, A.K., Harlow, K.W., et al., 2000. Structural basis for the function of *Bacillus subtilis* phosphoribosyl-pyrophosphate synthetase. *Nature structural biology*, 7(4), pp.303–308.

Ewing, B., Ewing, B., Hillier, L., Hillier, L., et al., 2005a. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, pp.175–185.

Ewing, B., Ewing, B., Hillier, L., Hillier, L., et al., 2005b. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, (206), pp.175–185.

Facklam, R., 2002. What happened to the streptococci: Overview of taxonomic and nomenclature changes. *Clinical Microbiology Reviews*, 15(4), pp.613–630.

Fang, W., Almeida, R. a & Oliver, S.P., 2000. Effects of lactoferrin and milk on adherence of *Streptococcus uberis* to bovine mammary epithelial cells. *American journal of veterinary research*, 61(3), pp.275–9.

Fang, W., Luther, D. a, Almeida, R. a & Oliver, S.P., 1998. Decreased growth of *Streptococcus uberis* in milk from mammary glands of cows challenged with the same mastitis pathogen. *Zentralblatt für Veterinärmedizin. Reihe B. Journal of veterinary medicine. Series B*, 45(9), pp.539–49.

Fang, W. & Oliver, S.P., 1999. Identification of lactoferrin-binding proteins in bovine mastitis-causing *Streptococcus uberis*. *FEMS microbiology letters*, 176, pp.91–96.

Farm Animal Welfare Council, 2009. *Opinion on the Welfare of the Dairy Cow*,

Farnaud, S. & Evans, R.W., 2003. Lactoferrin - A multifunctional protein with antimicrobial properties. *Molecular Immunology*, 40(7), pp.395–405.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., et al., 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, 34(3), p.e22.

Ferretti, J.J., McShan, W.M., Ajdic, D., Savic, D.J., et al., 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8), pp.4658–4663.

Field, T.R., Ward, P.N., Pedersen, L.H. & Leigh, J.A., 2003. The Hyaluronic Acid Capsule of *Streptococcus uberis* Is Not Required for the Development of Infection and Clinical Mastitis. *American Society for Microbiology*, 71(1), pp.132–139.

Filipe, S.R., Severina, E. & Tomasz, A., 2002. The *murMN* operon: a functional link between antibiotic resistance and antibiotic tolerance in *Streptococcus pneumoniae*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3), pp.1550–1555.

Fisher, S., Barry, A., Abreu, J., Minie, B., et al., 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology*, 12(1), p.R1.

Forsyth, R.A., Haselbeck, R.J., Ohlsen, K.L., Yamamoto, R.T., et al., 2002. A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Molecular microbiology*, 43(6), pp.1387–400.

Fukushima, T., Ishikawa, S., Yamamoto, H., Ogasawara, N., et al., 2003. Transcriptional, functional and cytochemical analyses of the *veg* gene in *Bacillus subtilis*. *Journal of Biochemistry*, 133(4), pp.475–483.

Gallagher, L., Shendure, J. & Manoil, C., 2011. Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. *MBio*, 2(1), pp.1–8.

Garg, A. & Gupta, D., 2008. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC bioinformatics*, 9, p.62.

Gasson, M.J. & Fitzgerald, G.F., 1994. Gene transfer systems and transposition. In M. Gasson & W. De Vos, eds. *Genetics and Biotechnology of Lactic Acid Bacteria SE - 1*. Springer Netherlands, pp. 1–51.

Gawronski, J.D., Wong, S.M.S., Giannoukos, G., Ward, D. V, et al., 2009. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38), pp.16422–7.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., et al., 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329(5987), pp.52–56.

Goddard, J.L. & Sokatch, J.R., 1964. 2-Ketogluconate Fermentation By *Streptococcus Faecalis*. *Journal of bacteriology*, 87(4), pp.844–851.

Goldammer, T., Zerbe, H., Molenaar, A., Schuberth, H., et al., 2004. Mastitis Increases Mammary mRNA Abundance of beta-Defensin 5, but Not TLR9 in Cattle. *American Society for Microbiology*, 11(1), pp.174–185.

Goodman, A., Wu, M. & Gordon, J., 2011. Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. *Nature Protocols*, 6(12), pp.1969–1980.

Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., et al., 2009. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*, 6(3), pp.279–289.

Goryshin, I.Y., Miller, J. a, Kil, Y. V, Lanzov, V. a, et al., 1998. Tn5/IS50 target recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 95(18), pp.10716–21.

Gotoh, Y., Eguchi, Y., Watanabe, T., Okamoto, S., et al., 2010. Two-component signal transduction as potential drug targets in pathogenic bacteria. *Current Opinion in Microbiology*, 13(2), pp.232–239.

Green, B., Bouchier, C., Fairhead, C., Craig, N.L., et al., 2012. Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mobile DNA*, 3(1), p.3.

Griffin, J.E., Gawronski, J.D., Dejesus, M. a, Ioerger, T.R., et al., 2011. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS pathogens*, 7(9), p.e1002251.

Griffiths, J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., et al., 2000. *An Introduction to Genetic Analysis* 7th Editio.,

Grinwald, M. & Ron, E.Z., 2013. The *Escherichia coli* Translation-Associated Heat Shock Protein YbeY Is Involved in rRNA Transcription Antitermination. *PLoS ONE*, 8(4).

Groschup, M.H., Hahn, G. & Timoney, J.F., 1991. Antigenic and genetic homogeneity fo *Streptococcus uberis* strains from the bovine udder. *Epidemiology Infections*, 107, pp.297–310.

Gudmundsson, G.H. & Agerberth, B., 1999. Neutrophil antibacterial peptides, multifunctional effector molecules in the mammalian immune system. *Journal of immunological methods*, 232(1-2), pp.45–54.

Gunasekera, T.S.T., Attfield, P. V & Veal, D.A., 2000. A flow cytometry method for rapid detection and enumeration of total bacteria in milk. *Applied and environmental Microbiology*, 66(3), pp.1228–1232.

Haaber, J., Moineau, S., Fortier, L.-C. & Hammer, K., 2008. AbiV, a novel antiphage abortive infection mechanism on the chromosome of *Lactococcus lactis* subsp. *cremoris* MG1363. *Applied and environmental microbiology*, 74(21), pp.6528–37.

Haandrikman, A.J., Van Leeuwen, C., Kok, J., Vos, P., et al., 1990. Insertion elements of lactococcal proteinase plasmids. *Applied and Environmental Microbiology*, 56(6), pp.1890–1896.

Haapa-Paananen, S., Rita, H. & Savilahti, H., 2002. DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection in vitro. *The Journal of biological chemistry*, 277(4), pp.2843–51.

Hacker, J., Bender, L., Ott, M., Wingender, J., et al., 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. *Microbial pathogenesis*, 8(3), pp.213–225.

Hagiwara, S., Kawai, K., Anri, A. & Nagahata, H., 2003. Lactoferrin concentrations in milk from normal and subclinical mastitic cows. *The Journal of veterinary medical science / the Japanese Society of Veterinary Science*, 65(3), pp.319–323.

Hama, H., Kayahara, T., Ogawa, W., Tsuda, M., et al., 1994. Enhancement of serine-sensitivity by a gene encoding rhodanese-like protein in Escherichia coli. *Journal of biochemistry*, 115(6), pp.1135–1140.

Hamer, L., DeZwaan, T., Montenegro-Chamorro, M., Frank, S., et al., 2001. Recent advances in large-scale transposon mutagenesis. *Current opinion in chemical biology*, 5, pp.67–73.

Hannon, G., 2009. FASTX-Toolkit. Available at: http://hannonlab.cshl.edu/fastx_toolkit/ (accessed July 2015).

Hardcastle, T.J. & Kelly, K.A., 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11, p.422.

Hardy, G.G., Caimano, M.J. & Yother, J., 2000. Capsule biosynthesis and basic metabolism in *Streptococcus pneumoniae* are linked through the cellular phosphoglucomutase. *Journal of Bacteriology*, 182(7), pp.1854–1863.

Hardy, G.G., Magee, A.D., Ventura, C.L., Caimano, M.J., et al., 2001. Essential role for cellular phosphoglucomutase in virulence of type 3 *Streptococcus pneumoniae*. *Infection and Immunity*, 69(4), pp.2309–2317.

Harmon, R.J., Schanbacher, F.L., Ferguson, L.C. & Smith, K.L., 1976. Changes in lactoferrin, immunoglobulin G, bovine serum albumin, and alpha-lactalbumin during acute experimental and natural coliform mastitis in cows. *Infection and immunity*, 13(2), pp.533–42.

Hartl, D., 2001. Discovery of the Transposable Element Mariner. *Genetics*, (157), pp.471–467.

Head, S.R., Kiyomi Komori, H., LaMere, S.A., Whisenant, T., et al., 2014. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56, pp.61–77.

Heather, Z., Holden, M.T.G., Steward, K.F., Parkhill, J., et al., 2008. A novel streptococcal integrative conjugative element involved in iron acquisition. *Molecular Microbiology*, 70(5), pp.1274–1292.

Heng, N.C.K., Burtenshaw, G.A., Jack, R.W. & Tagg, J.R., 2007. Ubericin A, a class IIa bacteriocin produced by *Streptococcus uberis*. *Applied and Environmental Microbiology*, 73(23), pp.7763–7766.

Hensel, M., Shea, J.E., Gleeson, C., Jones, M.D., et al., 1995. Simultaneous identification of bacterial virulence genes by negative selection. *Science (New York, N.Y.)*, 269(5222), pp.400–3.

Henthorn, K.S. & Friedman, D.I., 1995. Identification of related genes in phages ϕ80 and P22 whose products are inhibitory for phage growth in Escherichia coli IHF mutants. *Journal of Bacteriology*, 177(11), pp.3185–3190.

Hidalgo, F.J., Navarro, J.L., Delgado, R.M. & Zamora, R., 2013. Determination of α-keto acids in pork meat and Iberian ham via tandem mass spectrometry. *Food Chemistry*, 140(1-2), pp.183–188.

Hill, A. & Leigh, J., 1989. DNA fingerprinting of *Streptococcus uberis*: a useful tool for epidemiology of bovine mastitis. *Epidemiology and infection*, 103, pp.165–171.

Hill, A.W., 1988. Pathogenicity of two strains of *Streptococcus uberis* infused into lactating and non-lactating bovine mammary glands. *Research in veterinary science*, 45(3), pp.400–404.

Hill, A.W., Shears, A.L. & Hibbitt, K.G., 1978. The elimination of serum-resistant *Escherichia coli* from experimentally infected single mammary glands of healthy cows. *Research in veterinary science*, 25(1), pp.89–93.

Ho, Y.S., Burden, L.M. & Hurley, J.H., 2000. Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. *The EMBO journal*, 19(20), pp.5288–5299.

Hoeben, D., Burvenich, C., Eppard, P.J., Byatt, J.C., et al., 1999. Effect of bovine somatotropin on neutrophil functions and clinical symptoms during *Streptococcus uberis* mastitis. *Journal of dairy science*, 82(7), pp.1465–1481.

Hogan, J.S., Duthie, A.H. & Pankey, J.W., 1986. Fatty acid composition of bovine teat canal keratin. *Journal of dairy science*, 69(9), pp.2424–7.

Hogan, J.S., Smith, K.L., Todhunter, D. a & Schoenberger, P.S., 1988. Growth responses of environmental mastitis pathogens to long-chain fatty acids. *Journal of dairy science*, 71(1), pp.245–9.

Holden, M.T.G., Feil, E.J., Lindsay, J. a, Peacock, S.J., et al., 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26), pp.9786–9791.

Hossain, M., Egan, S.A., Coffey, T., Ward, P.N., et al., 2015. Virulence related sequences; insights provided by comparative genomics of *Streptococcus uberis* of differing virulence. *BMC Genomics*, 16(1), pp.1–13.

Huang, G., Zhang, L. & Birch, R.G., 2000. Rapid amplification and cloning of Tn5 flanking fragments by inverse PCR. *Letters in applied microbiology*, 31(2), pp.149–53.

Hutchison III, C.A., Peterson, S.N.N., Gill, S.R.R., Cline, R.T.T., et al., 1999. Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science*, 286(5447), pp.2165–2169.

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

Ihaka, R. & Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and graphical Statistics*, 5(3).

Irani, M.H., Orosz, L. & Adhya, S., 1983. A control element within a structural gene: the *gal* operon of *Escherichia coli*. *Cell*, 32(3), pp.783–788.

Issartel, J.P., Koronakis, V. & Hughes, C., 1991. Activation of *Escherichia coli* prohaemolysin to the mature toxin by acyl carrier protein-dependent fatty acylation. *Nature*, 351(6329), pp.759–761.

Itoh, T. & Tomizawa, J., 1980. Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proceedings of the National Academy of Sciences of the United States of America*, 77(5), pp.2450–2454.

Jacobs, M.A., Alwood, A., Thaipisuttikul, I., Spencer, D., et al., 2003. Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24), pp.14339–44.

Jacobson, J.W., Medhora, M.M. & Hartl, D.L., 1986. Molecular structure of a somatically unstable transposable element in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 83(November), pp.8684–8688.

Jenke-Kodama, H., Sandmann, A., Müller, R. & Dittmann, E., 2005. Evolutionary implications of bacterial polyketide synthases. *Molecular Biology and Evolution*, 22(10), pp.2027–2039.

Johansson, L., Thulin, P., Low, D.E. & Norrby-Teglund, A., 2010. Getting under the skin: the immunopathogenesis of *Streptococcus pyogenes* deep tissue infections. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 51(1), pp.58–65.

Johnsen, L.B., Poulsen, K., Kilian, M. & Petersen, T.E., 1999. Purification and cloning of a streptokinase from *Streptococcus uberis*. *Infection and immunity*, 67(3), pp.1072–8.

Jones, C.L., Monaghan, P., Field, T.R., Smith, A.J., et al., 2004. Localization of *mtuA*, an *lraI* homologue in *Streptococcus uberis*. *Journal of applied microbiology*, 97(1), pp.149–57.

Jorgenson, J.W. & DeArman Lukacs, K., 1981. Zone electrophoresis in open-tubular glass capillaries: Preliminary data on performance. *Journal of High Resolution Chromatography*, 4(5), pp.230–231.

Juárez, O. & Barquera, B., 2012. Insights into the mechanism of electron transfer and sodium translocation of the Na+-pumping NADH:quinone oxidoreductase. *Biochimica et Biophysica Acta - Bioenergetics*, 1817(10), pp.1823–1832.

Kaneda, T., 1991. Iso- and anteiso-fatty acids in bacteria: biosynthesis, function, and taxonomic significance. *Microbiological reviews*, 55(2), pp.288–302.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27–30.

Karelov, D. V., Kreneva, R.A., Errais Lopes, L., Perumov, D.A., et al., 2011. Mutational analysis of the *ribC* gene of *Bacillus subtilis*. *Russian Journal of Genetics*, 47(6), pp.757–761.

Karger, B.L. & Guttman, A., 2009. DNA Sequecing by Capillary Electrophoresis. *Electrophoresis*, 30(Suppl 1), pp.1–11.

Karp, P., 2001. Pathway databases: a case study in computational symbolic theories. *Science*, 293(September), pp.2040–2045.

Kasper, T.J., Melera, M., Gozel, P. & Brownlee, R.G., 1988. Separation and detection of DNA by capillary electrophoresis. *Journal of chromatography*, 458, pp.303–312.

Kawai, T. & Akira, S., 2007. Signaling to NF-κB by Toll-like receptors. *Trends in Molecular Medicine*, 13(11), pp.460–469.

Kehrli, M.E. & Shuster, D.E., 1994. Factors affecting milk somatic cells and their role in health of the bovine mammary gland. *Journal of dairy science*, 77(2), pp.619–627.

Kerro Dego, O., Prado, M.E., Chen, X., Luther, D.A., et al., 2011. pGh9:ISS1 transpositional mutations in *Streptococcus uberis* UT888 causes reduced bacterial adherence to and internalization into bovine mammary epithelial cells. *Veterinary microbiology*, 151(3-4), pp.379–85.

Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., et al., 2011. EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic acids research*, 39(Database issue), pp.D583–90.

Khattar, M.M., Addinall, S.G., Stedul, K.H., Boyle, D.S., et al., 1997. Two polypeptide products of the *Escherichia coli* cell division gene ftsW and a possible role for ftsW in ftsZ function. *Journal of Bacteriology*, 179(3), pp.784–793.

Kikuchi, K., Enari, T., Totsuka, K.I. & Shimizu, K., 1995. Comparison of phenotypic characteristics, DNA-DNA hybridization results, and results with a commercial rapid biochemical and enzymatic reaction system for identification of viridans group streptococci. *Journal of Clinical Microbiology*, 33(5), pp.1215–1222.

Klein, B. a, Tenorio, E.L., Lazinski, D.W., Camilli, A., et al., 2012. Identification of essential genes of the periodontal pathogen *Porphyromonas gingivalis*. *BMC genomics*, 13(1), p.578.

Klumpp, S., Zhang, Z. & Hwa, T., 2009. Growth-rate dependent global effects on gene expression in bacteria. *Cell*, 139(7), pp.1366–1375.

Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M., et al., 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PloS one*, 6(11), p.e28240.

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., et al., 2003. Essential *Bacillus subtilis* genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), pp.4678–83.

Komine, K., Komine, Y., Kuroishi, T., Kobayashi, J., et al., 2005. Small molecule lactoferrin with an inflammatory effect but no apparent antibacterial activity in mastitic mammary gland secretion. *The Journal of veterinary medical science*, 67(7), pp.667–77.

Koonin, E. V, 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature reviews. Microbiology*, 1(November), pp.127–136.

Korhonen, H., Marnila, P. & Gill, H.S., 2000. Milk immunoglobulins and complement factors. *The British journal of nutrition*, 84 Suppl 1, pp.S75–S80.

Kouidmi, I., Levesque, R.C. & Paradis-Bleau, C., 2014. The biology of Mur ligases as an antibacterial target. *Molecular Microbiology*, 94(2), pp.242–253.

Krzywinski, M., Schein, J., Birol, I., Connors, J., et al., 2009. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9), pp.1639–45.

Kulshina, N., Baird, N.J. & Ferré-D'Amaré, A.R., 2009. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nature structural & molecular biology*, 16(12), pp.1212–1217.

L'Anson, K.J., Movahedi, S., Griffin, H.G., Gasson, M.J., et al., 1995. A non-essential glutamyl aminopeptidase is required for optimal growth of Lactococcus lactis MG1363 in milk. *Microbiology (Reading, England)*, 141 (Pt 1, pp.2873–81.

Lacy-Hulbert, S.J. & Hillerton, J.E., 1995. Physical characteristics of the bovine teat canal and their influence on susceptibility to streptococcal infection. *The Journal of dairy research*, 62(3), pp.395–404.

Lamichhane, G., Tyagi, S. & Bishai, W.R., 2005. Designer Arrays for De ned Mutant Analysis To Detect Genes Essential for Survival of. *Society*, 73(4), pp.2533–2540.

Lammers, A., Kruijt, E., Van De Kuijt, C., Nuijten, P.J.M., et al., 2000. Identification of *Staphylococcus aureus* genes expressed during growth in milk: A useful model for selection of genes important in bovine mastitis? *Microbiology*, 146(4), pp.981–987.

Lampe, D.J., 2010. Bacterial genetic methods to explore the biology of mariner transposons. *Genetica*, 138, pp.499–508.

Lampe, D.J., Grant, T.E. & Robertson, H.M., 1998. Factors affecting transposition of the Himar1 mariner transposon in vitro. *Genetics*, 149(1), pp.179–87.

Lancefield, R., 1933. A serological differentiation of human and other groups of hemolytic streptococci. *The Journal of experimental medicine*, 1919(1).

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

Langridge, G.C., Phan, M., Turner, D.J., Perkins, T.T., et al., 2009. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome research*, 19, pp.2308–2316.

Lascelles, A.K. & McDowell, G.H., 1974. Localized humoral immunity with particular reference to ruminants. *Transplantation reviews*, 19, pp.170–208.

Lee, J.E., Smith, G.D., Horvatin, C., Huang, D.J.T., et al., 2005. Structural snapshots of MTA/AdoHcy nucleosidase along the reaction coordinate provide insights into enzyme and nucleoside flexibility during catalysis. *Journal of Molecular Biology*, 352(3), pp.559–574.

Lee, J.H., Li, X. & O'Sullivan, D.J., 2011. Transcription analysis of a lantibiotic gene cluster from *Bifidobacterium longum* DJO10A. *Applied and Environmental Microbiology*, 77(17), pp.5879–5887.

Leigh, J.A., 1993. Activation of bovine plasminogen by *Streptococcus uberis*. *FEMS Microbiology Letters*, 114(1), pp.67–71.

Leigh, J.A., 1999. *Streptococcus uberis*: A Permanent Barrier to the Control of Bovine Mastitis? *The Veterinary Journal*, 157(3), pp.225-238.

Leigh, J.A., Egan, S.A., Ward, P.N., Field, T.R., et al., 2010. Sortase anchored proteins of *Streptococcus uberis* play major roles in the pathogenesis of bovine mastitis in dairy cattle. *Veterinary research*, 41(5), p.63.

Leigh, J.A. & Field, T.R., 1994. *Streptococcus uberis* resists the bactericidal action of bovine neutrophils despite the presence of bound immunoglobulin. *Infection and immunity*, 62(5), pp.1854–9.

Leigh, J.A., Finch, J.M., Field, T.R., Real, N.C., et al., 1999. Vaccination with the plasminogen activator from *Streptococcus uberis* induces an inhibitory response and protects against experimental infection in the dairy cow. *Vaccine*, 17(7-8), pp.851–7.

Leigh, J.A. & Lincoln, R.A., 1997. *Streptococcus uberis* acquires plasmin activity following growth in the presence of bovine plasminogen through the action of its specific plasminogen activator. *FEMS Microbiology Letters*, 154(1), pp.123–129.

Lenski, R.E., Mongold, J.A., Sniegowski, P.D., Travisano, M., et al., 1998. Evolution of competitive fitness in experimental populations of *E. coli*: What makes one genotype a better competitor than another? *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 73(1), pp.35–47.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Oxford University Press*, 00(00), p.3.

Lindhoud, S., van den Berg, W.A.M., van den Heuvel, R.H.H., Heck, A.J.R., et al., 2012. Cofactor binding protects flavodoxin against oxidative stress. *PLoS ONE*, 7(7), pp.2–9.

Loman, N.J., Misra, R. V, Dallman, T.J., Constantinidou, C., et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), pp.434–9.

Longenecker, K.L., Stamper, G.F., Hajduk, P.J., Fry, E.H., et al., 2005. Structure of MurF from *Streptococcus pneumoniae* co-crystallized with a small molecule inhibitor exhibits interdomain closure. *Protein science : a publication of the Protein Society*, 14(12), pp.3039–3047.

Lopez-Benavides, M.G., Williamson, J.H., Pullinger, G.D., Lacy-Hulbert, S.J., et al., 2007. Field observations on the variation of *Streptococcus uberis* populations in a pasture-based dairy farm. *Journal of dairy science*, 90(12), pp.5558–66.

Lorca, G.L., Barabote, R.D., Zlotopolski, V., Tran, C., et al., 2007. Transport capabilities of eleven Gram-positive bacteria: Comparative genomic analyses. *Biochimica et Biophysica Acta - Biomembranes*, 1768(6), pp.1342–1366.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, pp.1–21.

Lu, G.T., Xie, J.R., Chen, L., Hu, J.R., et al., 2009. Glyceraldehyde-3-phosphate dehydrogenase of *Xanthomonas campestris* pv. *campestris* is required for extracellular polysaccharide production and full virulence. *Microbiology*, 155(5), pp.1602–1612.

Lu, J., Tomfohr, J.K. & Kepler, T.B., 2005. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC bioinformatics*, 6, p.165.

Luo, H., Lin, Y., Gao, F., Zhang, C.-T., et al., 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic acids research*, 42(1), pp.D574–80.

Lyon, G.J. & Novick, R.P., 2004. Peptide signaling in *Staphylococcus aureus* and other Gram-positive bacteria. *Peptides*, 25(9), pp.1389–1403.

Ma, Y., Bryant, A.E., Salmi, D.B., Hayes-Schroer, S.M., et al., 2006. Identification and characterization of bicistronic *speB* and *prsA* gene expression in the group A *Streptococcus*. *Journal of Bacteriology*, 188(21), pp.7626–7634.

Maglott, D., 2004. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33, pp.D54–D58.

Maguin, E., Prévost, H., Ehrlich, S.D., Gruss, A., et al., 1996. Efficient insertional mutagenesis in lactococci and other Gram-positive bacteria. *Journal of bacteriology*, 178(3), pp 931-935.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., et al., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(March), pp.3140–3145.

Majewski, J., Zawadzki, P., Pickerill, P., Frederick, M., et al., 2000. Barriers to Genetic Exchange between Bacterial Species : *Streptococcus pneumoniae* Transformation Barriers to Genetic Exchange between Bacterial Species : *Streptococcus pneumoniae* Transformation. *Journal of bacteriology*, 182(4), pp.1016–1023.

Maltsev, N., Glass, E., Sulakhe, D., Rodriguez, A., et al., 2006. PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic acids research*, 34(Database issue), pp.D369–72.

Manna, D., Deng, S., Breier, A.M. & Higgins, N.P., 2005. Bacteriophage Mu targets the trinucleotide sequence CGG. *Journal of Bacteriology*, 187(10), pp.3586–3588.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., et al., 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1), pp.D222–D226.

Marciniak, B.C., Trip, H., Fusetti, F. & Kuipers, O.P., 2012. Regulation of *ykrL* (*htpX)* by *rok* and *ykrK*, a novel type of regulator in bacillus subtilis. *Journal of Bacteriology*, 194(11), pp.2837–2845.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., et al., 2008. comparison with gene expression arrays RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, pp.1509–1517.

Marshall, V.M., Cole, W.M. & Bramley, a J., 1986. Influence of the lactoperoxidase system on susceptibility of the udder to *Streptococcus uberis* infection. *The Journal of dairy research*, 53(4), pp.507–14.

Martin, V. & Mohn, W., 1999. An alternative inverse PCR (IPCR) method to amplify DNA sequences flanking Tn5 transposon insertions. *Journal of microbiological methods*, 35, pp.163–166.

Matsutani, S., 1991. Multiple copies of IS10 in the *Enterobacter cloacae* MD36 chromosome. *Journal of bacteriology*, 173(24), pp.7802–9.

Matthews, K., Almeida, R. & Oliver, S., 1994. Bovine mammary epithelial cell invasion by *Streptococcus uberis. Infection and immunity*, 62(12), pp.5641-5646.

Maxam, A.M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2), pp.560–564.

Mckenzie, H.A. & White Jr., F.H., 1991. Lysozyme and α-Lactalbumin: Structure, Function, and Interrelationships. In J. T. E. C.B. Anfinsen Frederic M. Richards & D. S. Eisenberg, eds. Advances in Protein Chemistry. Academic Press, pp. 173–315.

McLennan, A.G., 2006. The Nudix hydrolase superfamily. *Cellular and Molecular Life Sciences*, 63(2), pp.123–143.

McNeil, L.K., Reich, C., Aziz, R.K., Bartels, D., et al., 2007. The National Microbial Pathogen Database Resource (NMPDR): A genomics platform based on subsystem annotation. *Nucleic Acids Research*, 35, pp.347–353.

Meisel, S., Stöckel, S., Elschner, M., Rösch, P., et al., 2011. Assessment of two isolation techniques for bacteria in milk towards their compatibility with Raman spectroscopy. *The Analyst*, 136(23), pp.4997–5005.

Mickelson, M.N. & Brown, R.W., 1985. Physiological characteristics of *Streptococcus dysgalactiae* and *Streptococcus uberis* and the effect of the lactoperoxidase complex on their growth in a chemically-defined medium and milk. *Journal of dairy science*, 68(5), pp.1095–102.

Miyafusa, T., Caaveiro, J.M.M., Tanaka, Y. & Tsumoto, K., 2012. Reveals a Unique Architecture Composed of Two Functional Domains. *Biochemical Journal*, 443(3), pp.671–680.

Mizuuchi, K., 1992. Transpositional recombination: Mechanistic Insights from Studies of Mu. *Annual Review of Biochemistry*, 61, pp.1011–1051.

Moazed, D. & Noller, H.F., 1987. Chloramphenicol, erythromycin, carbomycin and vernamycin B protect overlapping sites in the peptidyl transferase region of 23S ribosomal RNA. *Biochimie*, 69(8), pp.879–884.

Molenaar, a J., Kuys, Y.M., Davis, S.R., Wilkins, R.J., et al., 1996. Elevation of lactoferrin gene expression in developing, ductal, resting, and regressing parenchymal epithelium of the ruminant mammary gland. *Journal of dairy science*, 79(7), pp.1198–208.

Morozova, O. & Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5), pp.255–64.

Muñoz-López, M. & García-Pérez, J.L., 2010. DNA transposons: nature and applications in genomics. *Current genomics*, 11(2), pp.115–28.

Murchan, S., Kaufmann, M.E., Deplano, A., Struelens, M., et al., 2003. Harmonization of Pulsed-Field Gel Electrophoresis Protocols for Epidemiological Typing of Strains of Methicillin-Resistant *Staphylococcus aureus*: a Single Approach Developed by Consensus in 10 European Laboratories and Its Application for Tracing the Spread of Related Strains. *Journal of Clinical Microbiology*, 41(4), pp.1574–1585.

Mushegian, A.R. & Koonin, E. V, 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), pp.10268–73.

Nadell, C.D., Xavier, J.B. & Foster, K.R., 2009. The sociobiology of biofilms. *FEMS Microbiology Reviews*, 33(1), pp.206–224.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., et al., 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13), pp.1–13.

Nascimento, J.D.S., Coelho, M.L.V., Ceotto, H., Potter, A., et al., 2012. Genes involved in immunity to and secretion of aureocin A53, an atypical class II bacteriocin produced by *Staphylococcus aureus* A53. *Journal of Bacteriology*, 194(4), pp.875–883.

Needleman, S.B. & Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48, pp.443–453.

Neves, M., Périès, J. & Saïb, a., 1998. Study of human foamy virus proviral integration in chronically infected murine cells. *Research in Virology*, 149, pp.393–401.

Newman, R., Salunkhe, P., Godzik, A. & Reed, J.C., 2006. Identification and characterization of a novel bacterial virulence factor that shares homology with mammalian Toll/interleukin-1 receptor family proteins. *Infection and immunity*, 74(1), pp.594–601.

Nickerson, S.C., 1989. Immunological aspects of mammary involution. *Journal of dairy science*, 72(6), pp.1665–1678.

Nilsson, D., Lauridsen, a. a., Tomoyasu, T. & Ogura, T., 1994. A *Lactococcus lactis* gene encodes a membrane protein with putative ATPase activity that is homologous to the essential *Escherichia coli ftsH* gene product. *Microbiology*, 140(10), pp.2601–2610.

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., et al., 2012. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 40(20), pp.10084–10097.

O'Neill, L.A.J. & Bowie, A.G., 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nature reviews. Immunology*, 7(5), pp.353–364.

Ochman, H., Gerber, a S. & Hartl, D.L., 1988. Genetic applications of an inverse polymerase chain reaction. *Genetics*, 120(3), pp.621–3.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., et al., 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1), pp.29–34.

Ohshima, K., Mukai, Y., Shiraki, H., Suzumiya, J., et al., 1997. Clonal integration and expression of human T-cell lymphotropic virus type I in carriers detected by polymerase chain reaction and inverse PCR. *American Journal of Hematology*, 54(November 1996), pp.306–312.

Oliver, S.P., Almeida, R. a & Calvinho, L.F., 1998. Virulence factors of *Streptococcus uberis* isolated from cows with mastitis. *Zentralblatt für Veterinärmedizin. Reihe B. Journal of veterinary medicine. Series B*, 45(8), pp.461–71.

Van Opijnen, T., Bodi, K.L. & Camilli, A., 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganiams. *Nature Methods*, 6(10), pp.767–772.

Van Opijnen, T. & Camilli, A., 2012. A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome research*, 22(12), pp.2541–51.

Van Opijnen, T. & Camilli, A., 2010. Genome-wide fitness and genetic interactions determined by Tn-seq, a high-throughput massively parallel sequencing method for microorganisms. *Current protocols in microbiology*, 7-16.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J. V., et al., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), pp.5691–5702.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., et al., 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, 42, pp.D206–214.

Oviedo-Boyso, J., Valdez-Alarcón, J.J., Cajero-Juárez, M., Ochoa-Zarzosa, A., et al., 2007. Innate immune response of bovine mammary gland to pathogenic bacteria responsible for mastitis. *The Journal of infection*, 54(4), pp.399–409.

Paape, M.J., Wergin, W.P., Guidry, A.J. & Schultze, W.D., 1981. Phagocytic defense of the ruminant mammary gland. *Advances in experimental medicine and biology*, 137, pp.555–578.

Pajunen, M.I., Pulliainen, A.T., Finne, J. & Savilahti, H., 2005. Generation of transposon insertion mutant libraries for Gram-positive bacteria by electroporation of phage Mu DNA transposition complexes. *Microbiology (Reading, England)*, 151(Pt 4), pp.1209–18.

Patel, D., Almeida, R.A., Dunlap, J.R.J. & Oliver, S.S.P., 2009. Bovine lactoferrin serves as a molecular bridge for internalization of *Streptococcus uberis* into bovine mammary epithelial cells. *Veterinary microbiology*, 137(3-4), pp.297–301.

Patel, R.K. & Jain, M., 2012. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2), e30619

Peng, H.L., Novick, R.P., Kreiswirth, B., Kornblum, J., et al., 1988. Cloning, characterization and sequencing of an accessory gene regulator (agr) in *Staphylococcus aureus*. *Journal of Bacteriology*, 170(9), pp.4365–4372.

Pickard, D., Kingsley, R.A., Hale, C., Turner, K., et al., 2013. A genome wide mutagenesis screen identifies multiple genes contributing to Vi capsular expression in *Salmonella enterica* serovar Typhi. *Journal of bacteriology*, 195(6), pp.1320–6.

Pirondini, A., Bonas, U., Maestri, E., Visioli, G., et al., 2010. Yield and amplificability of different DNA extraction procedures for traceability in the dairy food chain. *Food Control*, 21(5), pp.663–668.

Politis, I., Zhao, X., McBride, B.W. & Burton, J.H., 1992. Function of bovine mammary macrophages as antigen-presenting cells. *Veterinary immunology and immunopathology*, 30, pp.399–410.

Polzin, K.M. & Shimizu-Kadota, M., 1987. Identification of a new insertion element, similar to gram-negative IS26, on the lactose plasmid of *Streptococcus lactis* ML3. *Journal of Bacteriology*, 169(12), pp.5481–5488.

Pritchard, J.R., Chao, M.C., Abel, S., Davis, B.M., et al., 2014. ARTIST: High-Resolution Genome-Wide Assessment of Fitness Using Transposon-Insertion Sequencing P. H. Viollier, ed. *PLoS Genetics*, 10(11), p.e1004782.

Pryor, S.M., Cursons, R.T., Williamson, J.H. & Lacy-Hulbert, S.J., 2009. Experimentally induced intramammary infection with multiple strains of *Streptococcus uberis*. *Journal of dairy science*, 92(11), pp.5467–5475.

Pullinger, G.D., Lo, M., Coffey, T.J., Williamson, J.H., et al., 2006. Application of *Streptococcus uberis* Multilocus Sequence Typing: Analysis of the Population Structure Detected among Environmental and Bovine Isolates from New Zealand and the United Kingdom. *Applied and environmental microbiology*, 72(2), pp.1429–1436.

Qingfang, W., Frye, J.G., Mcclelland, M. & Harshey, R.M., 2004. Gene expression patterns during swarming in Salmonella tyhpimurium: genes specific to surface growth and putative new motility and pathogenicity genes. *Molecular Microbiology*, 52, pp.169–187.

Quail, M.A., 2010. DNA : Mechanical Breakage. *Encyclopedia of Life Sciences*, pp.1–5.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., et al., 2009. A large genome center's improvements to the Illumina sequencing system. *Nature methods*, 1010(June), pp.1005–1010.

Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., et al., 2012. Optimal enzymes for amplifying sequencing libraries. *Nature methods*, 9(1), pp.10–1.

Quail, M.A., Swerdlow, H. & Turner, D.J., 2009. Improved protocols for the illumina genome analyzer sequencing system. *Current protocols in human genetics*, Chapter 18(July), 18.2.

Quigley, L., O'Sullivan, O., Beresford, T.P., Paul Ross, R., et al., 2012. A comparison of methods used to extract bacterial DNA from raw milk and raw milk cheese. *Journal of applied microbiology*, 113(1), pp.96–105.

Quon, K.C., Marczynski, G.T. & Shapiro, L., 1996. Cell cycle control by an essential bacterial two-component signal transduction protein. *Cell*, 84(1), pp.83–93.

Rainard, P., 2003. The complement in milk and defense of the bovine mammary gland against infections. *Veterinary research*, 34, pp.647–670.

Rainard, P. & Iollet, C.R., 2006. Innate immunity of the bovine mammary gland. *Veterinary Reserach*, 37, pp.369–400.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., et al., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9), p.R95.

Rato, M.G., Bexiga, R., Nunes, S.F., Cavaco, L.M., et al., 2008. Molecular epidemiology and population structure of bovine *Streptococcus uberis*. *Journal of dairy science*, 91(12), pp.4542–51.

Ravanel, S., Gakière, B., Job, D. & Douce, R., 1998. Cystathionine c-synthase from *Arabidopsis thaliana*: purification and biochemical characterization of the recombinant enzyme overexpressed in *Escherichia coli*. *Biochemical Journal*, 331, pp.639–648.

Renault, P.P. & Heslot, H., 1987. Selection of *Streptococcus lactis* Mutants Defective in Malolactic Fermentation. *American Society for Microbiology*, 53(2), pp.320–324.

Reznikoff, W. & Winterberg, K., 2008. Transposon-Based Strategies for the Identification of Essential Bacterial Genes. *Methods in Molecular Biology*, 416, pp.107–113.

Richards, V.P., Palmer, S.R., Pavinski Bitar, P.D., Qin, X., et al., 2014. Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome biology and evolution*, 6(4), pp.741–53.

Robinson, C., Steward, K.F., Potts, N., Barker, C., et al., 2013. Combining two serological assays optimises sensitivity and specificity for the identification of *Streptococcus equi* subsp. *equi* exposure. *Veterinary Journal*, 197(2), pp.188–191.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), pp.139–40.

Robles, J.A., Qureshi, S.E., Stephen, S.J., Wilson, S.R., et al., 2012. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, 13(1), p.484.

Rogozin, I.B., Makarova, K.S., Natale, D. a, Spiridonov, A.N., et al., 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic acids research*, 30(19), pp.4264–4271.

Roguinsky, M., 1977. Comparison of *Streptococcus uberis* and S. infrequens. Pathogenicity for cow udder. *Ann. Rech.* vol. 8, no. 2, pp. 153-157. 1977.

Romero, D.A. & Klaenhammer, T.R., 1993. Transposable elements in Lactococci: a review. *Journal of dairy science*, 76(1), pp.1–19.

Rosey, E.L., Lincoln, R.A., Ward, P.N., Yancey, R.J., et al., 1999. PauA: a novel plasminogen activator from *Streptococcus uberis*. *FEMS microbiology letters*, 178(1), pp.27–33.

Ruiz-Albert, J., Mundy, R., Yu, X.J., Beuzón, C.R., et al., 2003. SseA is a chaperone for the SseB and SseD translocon components of the Salmonella pathogenicity-island-2-encoded type III secretion system. *Microbiology*, 149(5), pp.1103–1111.

Salama, N.R., Shepherd, B. & Falkow, S., 2004. Global Transposon Mutagenesis and Essential Gene Analysis of *Helicobacter pylori*. *Journal of bacteriology*, 186(23), pp.7926–7935.

Samant, S., Lee, H., Ghassemi, M., Chen, J., et al., 2008. Nucleotide biosynthesis is critical for growth of bacteria in human blood. *PLoS Pathogens*, 4(2), e37.

Sánchez, L., Calvo, M. & Brock, J.H., 1992. Biological role of lactoferrin. *Archives of disease in childhood*, 67(5), pp.657–61.

Sanger, F., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3), pp.441–448.

Sassetti, C.M., Boyd, D.H. & Rubin, E.J., 2001. Comprehensive identification of conditionally essential genes in mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22), pp.12712–7.

Schalm, O.W., Lasmanis, J. & Jain, N.C., 1976. Conversion of chronic staphylococcal mastitis to acute gangrenous mastitis after neutropenia in blood and bone marrow produced by an equine anti-bovine leukocyte serum. *American journal of veterinary research*, 37(8), pp.885–890.

Schmidt, H. & Hensel, M., 2004. Pathogenicity Islands in Bacterial Pathogenesis. *Clinical Microbiology Reviews*, 17(1), pp.14–56.

Schonwetter, B.S., Stolzenberg, E.D. & Zasloff, M. a, 1995. Epithelial antibiotics induced at sites of inflammation. *Science (New York, N.Y.)*, 267(5204), pp.1645–8.

Schreur, P.J., van Weeghel, C., Rebel, J.M.J., Smits, M.A., et al., 2012. Lysozyme resistance in *Streptococcus suis* is highly variable and multifactorial. *PLoS ONE*, 7(4).

Schrimer, M., Ijaz, U.Z., D'Amore, R., Hall, N., et al., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research*, gku1341.

Serganov, A. & Patel, D.J., 2009. Amino acid recognition and gene regulation by riboswitches. *Biochimica et biophysica acta*, 29(6), pp.997–1003.

Shai, Y., 1995. Molecular recognition between membrane-spanning polypeptides. *Trends in biochemical sciences*, 20.11 (1995): 460-464.

Shapiro, J.A., 1979. Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, 76(4), pp.1933–7.

Shazand, K., Tucker, J., Grunberg-Manago, M., Rabinowitz, J.C., et al., 1993. Similar organization of the nusA-infB operon in *Bacillus subtilis* and *Escherichia coli*. *Journal of Bacteriology*, 175(10), pp.2880–2887.

Sherman, J.M., 1937. The *Streptococci*. *Bacteriological reviews*, 1, pp.3–97.

Shoemaker, C.A., Evans, J.J. & Klesius, P.H., 2000. Density and dose: Factors affecting mortality of *Streptococcus iniae* infected tilapia (Oreochromis niloticus). *Aquaculture*, 188, pp.229–235.

Shunsuke, I., Kuroki, K. & Imamoto, F., 1984. tRNA gene in the leader region of the *nusA* operon in *Escherichia coli*. *Biochemistry*, 81(January), pp.409–413.

Sinai, A.P. & Bavoil, P.M., 1993. Hyper-invasive mutants define a novel Pho-regulated invasion pathway in *Escherichia coli. Molecular microbiology*, 10(5), pp.1125–1137.

Smith, A.J., Kitt, A.J., Ward, P.N. & Leigh, J.A., 2002. Isolation and characterization of a mutant strain of *Streptococcus uberis*, which fails to utilize a plasmin derived beta-casein peptide for the acquisition of methionine. *Journal of applied microbiology*, 93(4), pp.631–9.

Smith, A.J., Ward, P.N., Field, T.R., Jones, C.L., et al., 2003. MtuA, a lipoprotein receptor antigen from *Streptococcus uberis*, is responsible for acquisition of manganese during growth in milk and is essential for infection of the lactating. *Infection and Immunity*, 71(9), pp.4842–4849.

Smith, E.M., Green, L.E. & Mason, D., 2003. Savinase Is a Bactericidal Enzyme. *Applied and environmental microbiology*, 69(1), pp.719–721.

Smith, K.C., Blunden, A.S., Whitwell, K.E., Dunn, K.A., et al., 2003. A survey of equine abortion, stillbirth and neonatal death in the UK from 1988 to 1997. *Equine veterinary journal*, 35(5), pp.496–501.

Smith, L. & Fung, S., 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic acids …*, 13(7), pp.2399–2412.

Smith, L., Sanders, J., Kaiser, R. & Hughes, P., 1986. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321, pp.674–679.

Smith, M.W. & Doolittle, R.F., 1992. Anomalous phylogeny involving the enzyme glucose-6-phosphate isomerase. *Journal of Molecular Evolution*, 34(6), pp.544–545.

Smith, T.. & Waterman, M.., 1981. Identification of common molecular subsequences. *Journal of molecular biology*, (147), pp.195–197.

Soneson, C. & Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1), p.91.

Song, J., Ko, K., Lee, J., Baek, J., et al., 2005. Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Molecules and Cells*, 19(3), pp.365–374.

Sordillo, L.M., Shafer-Weaver, K. & DeRosa, D., 1997. Immunobiology of the mammary gland. *Journal of dairy science*, 80(8), pp.1851–65.

Sordillo, L.M. & Streicher, K.L., 2002. Mammary gland immunity and mastitis susceptibility. *Journal of mammary gland biology and neoplasia*, 7(2), pp.135–46.

Soyeurt, H., Colinet, F.G., Arnould, V.M.R., Dardenne, P., et al., 2007. Genetic variability of lactoferrin content estimated by mid-infrared spectrometry in bovine milk. *Journal of dairy science*, 90(9), pp.4443–4450.

Spellerberg, B., Pohl, B., Haase, G., Martin, S., et al., 1999. Identification of genetic determinants for the hemolytic activity of *Streptococcus agalactiae* by ISS1 transposition. *Journal of bacteriology*, 181(10), pp.3212–9.

Spiegelman, W.G., Reichardt, L.F., Yaniv, M., Heinemann, S.F., et al., 1972. Bidirectional transcription and the regulation of Phage lambda repressor synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 69(11), pp.3156–3160.

Srinivasan, R., Rajeswari, H. & Ajitkumar, P., 2008. Analysis of degradation of bacterial cell division protein FtsZ by the ATP-dependent zinc-metalloprotease FtsH in vitro. *Microbiological Research*, 163(1), pp.21–30.

Sriskandan, S. & Slater, J.D., 2006. Invasive disease and toxic shock due to zoonotic *Streptococcus suis*: an emerging infection in the East? *PLoS medicine*, 3(5), p.e187.

Steijns, J.M. & van Hooijdonk, A.C., 2000. Occurrence, structure, biochemical properties and technological characteristics of lactoferrin. *The British journal of nutrition*, 84 Suppl 1, pp.S11–S17.

Stevens, D.L., Tanner, M.H., Winship, J., Swarts, R., et al., 1989. Severe Group A Streptococcal Infections Associated with a Toxic Shock-like Syndrome and Scarlet Fever Toxin A. *New England Journal of Medicine*, 321(1), pp.1–7.

Storm, D.R., 1974. Mechinism of bacitracin action: A specific lipid-peptide interaction. *Annals of the New York Academy of Sciences*, 235(1), pp.387–398.

Sun, F., Cho, H., Jeong, D.W., Li, C., et al., 2010. Aureusimines in *Staphylococcus aureus* Are Not Involved in Virulence. *PLoS ONE*, 5(12).

Surette, M.G. & Bassler, B.L., 1998. Quorum sensing in *Escherichia coli* and *Salmonella typhimurium*. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12), pp.7046–7050.

Tang, J., Wang, C., Feng, Y., Yang, W., et al., 2006. Streptococcal toxic shock syndrome caused by *Streptococcus suis* serotype 2. *PLoS medicine*, 3(5), p.e151.

Täpp, J., Thollesson, M. & Herrmann, B., 2003. Phylogenetic relationships and genotyping of the genus *Streptococcus* by sequence determination of the RNase P RNA gene, rnpB. *International Journal of Systematic and Evolutionary Microbiology*, 53, pp.1861–1871.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., et al., 2010. Differential expression in RNA-seq: a matter of depth. *Genome Research*, 21.12 (2011): 2213-2223

Taylor, D.L., Ward, P.N., Rapier, C.D., Leigh, J.A., et al., 2003. Identification of a Differentially Expressed Oligopeptide Binding Protein (OppA2) in *Streptococcus uberis* by Representational Difference Analysis of cDNA. *Journal of bacteriology*, 185(17), pp.5210–5219.

Taylor, P. & Dego, O.K., 2002. Factors involved in the early pathogenesis of bovine *Staphylococcus aureus* mastitis with emphasis on bacterial adhesion and invasion. *Veterinary Quarterly*, (July 2012), pp.37–41.

Thanassi, J., Hartman-Neumann, S., Dougherty, T., Dougherty, B., et al., 2002. Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic acids research*, 30(14), pp.3152–3162.

Thibessard, A., Fernandez, A., Gintz, B., Decaris, B., et al., 2002. Transposition of pGh9: IS S1 is random and efficient in *Streptococcus thermophilus* CNRZ368. *Canadian journal of microbiology*, 48(5), pp.473–478.

Thomas, J.W., 1970. Metabolism of iron and manganese. *Journal of dairy science*, 53(8), pp.1107–1123.

Thomas, L. & Haider, W., 1994. Pathologic findings of experimentally induced *Streptococcus uberis* infection in the mammary gland of cows. *American journal of Veterinary research*, 55.12 (1994): 1723-1728.

Timoney, J.F. & Kumar, P., 2008. Early pathogenesis of equine *Streptococcus equi* infection (strangles). *Equine veterinary journal*, 40(7), pp.637–642.

Todhunter, D.A., Smith, K.L. & Hogan, J.S., 1995. Environmental *Streptococcal* intramammary infections of the bovine mammary gland. *Journal of dairy science*, 78(11), pp.2366–74.

Tönjes, R.R., Czauderna, F. & Kurth, R., 1999. Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K. *Journal of virology*, 73(11), pp.9187–9195.

Toranzo, A.E., Magariños, B. & Romalde, J.L., 2005. A review of the main bacterial fish diseases in mariculture systems. *Aquaculture*, 246, pp.37–61.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., et al., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1), pp.46–53.

Turner, L.S., Kanamoto, T., Unoki, T., Munro, C.L., et al., 2009. Comprehensive evaluation of *Streptococcus sanguinis* cell wall-anchored proteins in early infective endocarditis. *Infection and Immunity*, 77(11), pp.4966–4975.

Umbreit, J.N., Stone, K.J. & Strominger, J.L., 1972. Isolation of Polyisoprenyl Alcohols from *Streptococcus faecalis*. *Journal of bacteriology*, 112(3), pp.1302–1305.

UniProt, 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41(1), pp.43–47.

Vandamme, P., Pot, B., Falsen, E., Kersters, K., et al., 1996. Taxonomic study of lancefield streptococcal groups C, G, and L (*Streptococcus dysgalactiae*) and proposal of *S. dysgalactiae* subsp. *equisimilis*. nov. *International journal of systematic bacteriology*, 46(3), pp.774–781.

Vigdal, T.J., Kaufman, C.D., Izsvák, Z., Voytas, D.F., et al., 2002. Common Physical Properties of DNA Affecting Target Site Selection of Sleeping Beauty and other Tc1/mariner Transposable Elements. *Journal of Molecular Biology*, 323(3), pp.441–452.

Villanelo, F., Ordenes, A., Brunet, J., Lagos, R., et al., 2011. A model for the *Escherichia coli* FtsB/FtsL/FtsQ cell division complex. *BMC structural biology*, 11(1), p.28.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., et al., 2004. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, 14, pp.208–216.

Vollmer, W. & Tomasz, A., 2000. The *pgdA* gene encodes for a peptidoglycan N-acetylglucosamine deacetylase in *Streptococcus pneumoniae*. *Journal of Biological Chemistry*, 275(27), pp.20496–20501.

De Vuyst, L., Vanderveken, F., Van De Ven, S. & Degeest, B., 1998. Production by and isolation of exopolysaccharides from Streptococcus thermophilus grown in a milk medium and evidence for their growth-associated biosynthesis. *Journal of Applied Microbiology*, 84(6), pp.1059–1068.

Wagner, C. & Saizieu, A. De, 2002. Genetic analysis and functional characterization of the *Streptococcus pneumoniae vic* operon. *Infection and immunity*, 70(11), pp.6121–6128.

Wagner, G. & Simons, R.W., 1994. Antisense Rna Control in Bacteria, Phages and PLasmids. *Annu. Rev. Microbial.*, 48, pp.713–42.

Waller, A.S., 2014. New Perspectives for the Diagnosis, Control, Treatment, and Prevention of Strangles in Horses. *Veterinary Clinics of North America: Equine Practice*, 30(3), pp.591–607.

Wang, B. & Kuramitsu, H.K., 2005. Inducible Antisense RNA Expression in the Characterization of Gene Functions in Streptococcus mutans. *American Society for Microbiology*, 73(6), pp.3568–3576.

Wang, B. & Kuramitsu, H.K., 2005. Interactions between Oral Bacteria: Inhibition of *Streptococcus mutans* bacteriocin production by Streptococcus gordonii. *Society*, 71(1), pp.354–362.

Wang, J., Kodali, S., Lee, S.H., Galgoci, A., et al., 2007. Discovery of platencin, a dual FabF and FabH inhibitor with in vivo antibiotic properties. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18), pp.7612–7616.

Ward, P.N., Field, T.R., Ditcham, W.G.F., Maguin, E., et al., 2001. Identification and Disruption of Two Discrete Loci Encoding Hyaluronic Acid Capsule Biosynthesis Genes *hasA , hasB*, and *hasC* in *Streptococcus uberis*. *American Journal for Microbiology*, 69(1), pp.392–399.

Ward, P.N., Field, T.R., Rapier, C.D. & Leigh, J.A., 2003. The Activation of Bovine Plasminogen by PauA Is Not Required for Virulence of *Streptococcus uberis*. *American Society for Microbiology*, 71(12), pp.7193–7196.

Ward, P.N., Holden, M.T.G., Leigh, J.A., Lennard, N., et al., 2009. Evidence for niche adaptation in the genome of the bovine pathogen *Streptococcus uberis*. *BMC genomics*, 10, p.54.

Ward, P.N. & Leigh, J.A., 2004. Genetic analysis of *Streptococcus uberis* plasminogen activators. *The Indian journal of medical research*, 119(May), pp.136–40.

Ward, P.N.P. & Leigh, J.A., 2002. Characterization of PauB, a novel broad-spectrum plasminogen activator from *Streptococcus uberis*. *Journal of bacteriology*, 184(1), pp.119–125.

Weinberg, E.D., 1986. Iron and Infection. *American Society for Microbiology*, 5(1), p.132.

Wessels, M.R., Goldberg, J.B., Moses, A.E. & DiCesare, T.J., 1994. Effects on virulence of mutations in a locus essential for hyaluronic acid capsule expression in group A streptococci. *Infection and immunity*, 62(2), pp.433–41.

Westhof, E., 2010. The amazing world of bacterial structured RNAs. *Genome biology*, 11(3), p.108.

Wheelan, S.J., Scheifele, L.Z., Martínez-Murillo, F., Irizarry, R. a, et al., 2006. Transposon insertion site profiling chip (TIP-chip). *Proceedings of the National Academy of Sciences of the United States of America*, 103(47), pp.17632–7.

Williams, D.M. & Mein, G.A., 1985. The role of machine milking in the invasion of mastitis organisms and implications for maintaining low infection-rates. *Kieler Milchwirtschaftliche Forschungsberichte*, 37(4), pp.415–425.

Wilson, I.G., 1997. Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology*, 63(10), pp.3741–3751.

Winkler, M.E. & Hoch, J.A., 2008. Essentiality, bypass, and targeting of the YycFG (VicRK) two-component regulatory system in gram-positive bacteria. *Journal of Bacteriology*, 190(8), pp.2645–2648.

Wirawan, R. & Klesse, N., 2006. Molecular and genetic characterization of a novel nisin variant produced by *Streptococcus uberis*. *Applied and environmental microbiology*, 72(2), pp.1148–1156.

Wirawan, R.E., Jack, R.W. & Tagg, J.R., 2006. A novel combination of bacteriocins produced by Streptococcus uberis strain 42. *International Congress Series*, 1289, pp.355–358.

Wolffs, P., Knutsson, R., Norling, B. & Radstrom, P., 2004. Rapid Quantification of Yersinia enterocolitica in Pork Samples by a Novel Sample Preparation Method, Flotation, Prior to Real-Time PCR. *Journal of Clinical Microbiology*, 42(3), pp.1042–47.

Wong, C.W., Seow, H.F., Husband, A.J., Regester, G.O., et al., 1997. Effects of purified bovine whey factors on cellular immune functions in ruminants. *Veterinary Immunology and Immunopathology*, 56(1-2), pp.85–96.

Wood, J.L.N., Newton, J.R., Chanter, N. & Mumford, J.A., 2005. Association between Respiratory Disease and Bacterial and Viral Infections in British Racehorses Association between Respiratory Disease and Bacterial and Viral Infections in British Racehorses. *Journal of Clinical Microbiology*, 43(1), pp.120–126.

Xu, P., Ge, X., Chen, L., Wang, X., et al., 2011. Genome-wide essential gene identification in *Streptococcus sanguinis*. *Scientific reports*, 1, p.125.

Yoshimura, A., Lien, E., Ingalls, R.R., Tuomanen, E., et al., 1999. Cutting edge: recognition of Gram-positive bacterial cell wall components by the innate immune system occurs via Toll-like receptor 2. *Journal of immunology (Baltimore, Md. : 1950)*, 163(1), pp.1–5.

Zadoks, R., 2005. Multilocus Sequence Typing of Streptococcus uberis Provides Sensitive and Epidemiologically Relevant Subtype Information and Reveals Positive Selection in the Virulence Gene *pauA*. *Journal of clinical microbiology*, 43(5), pp.2407–2417.

Zadoks, R.N., Gillespie, B.E., Barkema, H.W., Sampimon, O.C., et al., 2003. Clinical, epidemiological and molecular characteristics of *Streptococcus uberis* infections in dairy herds. *Epidemiology and infection*, 130(2), pp.335–49.

Zhang, R. & Lin, Y., 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*, 37(Database issue), pp.D455–8.

Zhang, R., Ou, H.-Y. & Zhang, C.-T., 2004. DEG: a database of essential genes. *Nucleic acids research*, 32(Database issue), pp.D271–2.

Zhang, Y.J., Ioerger, T.R., Huttenhower, C., Long, J.E., et al., 2012. Global assessment of genomic regions required for growth in *Mycobacterium tuberculosis*. *PLoS pathogens*, 8(9), p.e1002946.

Zomer, A., Burghout, P., Bootsma, H.J., Hermans, P.W.M., et al., 2012. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PloS one*, 7(8), p.e43012.

Zoraghi, R., Corbin, J.D. & Francis, S.H., 2004. Properties and functions of GAF domains in cyclic nucleotide phosphodiesterases and other proteins. *Molecular pharmacology*, 65(2), pp.267–278.

Appendices

Appendix 1

Table documenting annotations and RAST category of each coding sequence for the *S. uberis* 0140J genome (Accession number AM946015) used during this study.

Amino Acids and Derivatives

Locus	Gene	Product
SUB0022	recO	DNA repair protein RecO
SUB0065	NA	putative threonine synthase
SUB0130	proC	putative pyrroline-5-carboxylate reductase
SUB0269	cysE	serine acetyltransferase
SUB0310	NA	putative peptidase
SUB0315	NA	haloacid dehalogenase-like hydrolase
SUB0316	NA	putative aspartokinase
SUB0324	NA	haloacid dehalogenase-like hydrolase
SUB0345	csdB	putative cysteine desulfurase
SUB0366	NA	putative lipoprotein
SUB0367	NA	putative peptidase
SUB0368	NA	putative D-methionine transport ATP-binding protein
SUB0369	NA	putative D-methionine transport system permease protein
SUB0370	NA	putative branched-chain amino acid transport system carrier protein
SUB0398	NA	pyridoxal-dependent decarboxylase
SUB0462	brnQ	putative branched-chain amino acid transport system carrier protein
SUB0471	mtnN	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase
SUB0558	NA	haloacid dehalogenase-like hydrolase
SUB0582	NA	putative peptidase
SUB0583	NA	putative peptidase
SUB0593	NA	haloacid dehalogenase-like hydrolase
SUB0608	NA	haloacid dehalogenase-like hydrolase
SUB0618	NA	haloacid dehalogenase-like hydrolase
SUB0628	aspC	aspartate aminotransferase
SUB0642	NA	chorismate mutase type II protein
SUB0651	NA	haloacid dehalogenase-like hydrolase
SUB0687	NA	putative aminotransferase
SUB0716	aroD	3-dehydroquinate dehydratase
SUB0717	aroF	chorismate synthase
SUB0756	dapB	putative dihydrodipicolinate reductase
SUB0762	gdh	NADP-specific glutamate dehydrogenase
SUB0810	hom	homoserine dehydrogenase
SUB0811	thrB	homoserine kinase
SUB0818	potA	putative spermidine/putrescine ABC transporter ATP-binding protein
SUB0819	potB	putative spermidine/putrescine ABC transporter permease protein
SUB0820	potC	spermidine/putrescine ABC transporter permease protein
SUB0821	potD	spermidine/putrescine extracellular binding protein
SUB0836	NA	haloacid dehalogenase-like hydrolase
SUB0844	NA	putative peptidase
SUB0859	NA	haloacid dehalogenase-like hydrolase
SUB0874	glyA	serine hydroxymethyltransferase

SUB0909	NA	putative carboxypeptidase
SUB0938	dapA	dihydrodipicolinate synthase
SUB0939	asd	aspartate-semialdehyde dehydrogenase
SUB1009	NA	putative branched-chain amino acid aminotransferase
SUB1010	NA	putative branched-chain-amino-acid aminotransferase
SUB1062	NA	haloacid dehalogenase-like hydrolase
SUB1063	NA	haloacid dehalogenase-like hydrolase
SUB1086	NA	haloacid dehalogenase-like hydrolase
SUB1156	NA	putative UDP-N-acetylglucosamine 1-carboxyvinyltransferase 2
SUB1183	NA	Crp family regulatory protein
SUB1192	aroA	3-phosphoshikimate 1-carboxyvinyltransferase
SUB1199	NA	haloacid dehalogenase-like hydrolase
SUB1234	NA	haloacid dehalogenase-like hydrolase
SUB1272	NA	putative repressor protein
SUB1289	NA	conserved hypothetical protein
SUB1316	NA	carbon-nitrogen hydrolase family protein
SUB1317	NA	Porphyromonas-type peptidyl-arginine deiminase
SUB1318	NA	putative carboxynorspermidine decarboxylase
SUB1319	NA	saccharopine dehydrogenase family protein
SUB1320	speE	putative spermidine synthase
SUB1321	speA	putative arginine decarboxylase
SUB1323	asnA	aspartate--ammonia ligase
SUB1324	arcC	carbamate kinase
SUB1326	NA	C4-dicarboxylate anaerobic carrier protein
SUB1327	arcB	ornithine carbamoyltransferase
SUB1328	arcA	arginine deiminase
SUB1329	NA	Crp family regulatory protein
SUB1330	argR	arginine repressor
SUB1361	aroD	3-dehydroquinate dehydratase
SUB1362	aroE	shikimate 5-dehydrogenase
SUB1379	cysM	putative cysteine synthase
SUB1386	prpC	serine/threonine phosphatase
SUB1399	luxS	S-ribosylhomocysteinase
SUB1422	proA	gamma-glutamyl phosphate reductase
SUB1423	proB	glutamate 5-kinase
SUB1447	NA	haloacid dehalogenase-like hydrolase
SUB1529	NA	putative aminotransferase
SUB1531	NA	haloacid dehalogenase-like hydrolase
SUB1532	asnB	putative L-asparaginase
SUB1535	alr	alanine racemase
SUB1626	glnA	putative glutamine synthetase
SUB1698	NA	putative peptidase
SUB1788	argR	putative arginine repressor
SUB1793	aspS	aspartyl-tRNA synthetase
SUB1848	sdhB	putative L-serine dehydratase, beta chain
SUB1849	sdhA	putative L-serine dehydratase, alpha chain

Carbohydrates

SUB0038	NA	GntR family regulatory protein
SUB0040	lacD3	tagatose 1,6-diphosphate aldolase 3
SUB0041	NA	GntR family regulatory protein
SUB0042	NA	putative beta-galactosidase precursor
SUB0045	NA	Sugar (PTS), mannose/fructose/sorbose family, IID component
SUB0126	ackA	acetate kinase
SUB0135	fruA	putative fructan beta-fructosidase precursor

SUB0156	NA	putative pyruvate formate-lyase activating enzyme
SUB0158	NA	putative sugar-binding regulatory protein
SUB0160	pflD	putative formate acetyltransferase
SUB0161	fsaA	putative fructose-6-phosphate aldolase 1
SUB0195	NA	Sugar (PTS), lactose/cellobiose-specific family, IIB component
SUB0196	NA	Sugar (PTS), lactose/cellobiose-specific family, IIA component
SUB0197	NA	Sugar (PTS), IIC component
SUB0198	NA	putative beta-glucosidase
SUB0199	NA	6-phosphogluconate dehydrogenase family protein
SUB0200	NA	6-phospho-beta-glucosidase 1
SUB0201	NA	putative gluconokinase
SUB0235	NA	LacI family regulatory protein
SUB0236	NA	putative glucose-specific (PTS), IIABC component
SUB0245	msmK	multiple sugar-binding transport ATP-binding protein
SUB0259	NA	putative NADPH-dependent FMN reductase
SUB0285	RPE	putative ribulose-phosphate 3-epimerase
SUB0286	NA	putative 6-phospho-3-hexuloisomerase
SUB0287	NA	putative mannitol-1-phosphate 5-dehydrogenase
SUB0288	NA	sugar phosphotransferase system (PTS), IIA component
SUB0289	NA	putative mannitol-specific (PTS), IIBC component
SUB0290	tal2	probable transaldolase 2
SUB0291	tkt2	putative transketolase 2
SUB0293	NA	acetoin(diacetyl) reductase (pseudogene)
SUB0303	NA	LacI family regulatory protein
SUB0307	NA	beta-fructofuranosidase
SUB0308	scrK2	fructokinase 2
SUB0309	NA	6-phospho-beta-glucosidase 2
SUB0330	fba	fructose-bisphosphate aldolase
SUB0417	pflC	pyruvate formate-lyase activating enzyme
SUB0489	NA	uracil DNA glycosylase superfamily protein
SUB0526	gloA	putative lactoylglutathione lyase
SUB0529	NA	Sugar (PTS), lactose/cellobiose-specific family IIB component
SUB0531	NA	Sugar (PTS), lactose/cellobiose-specific family, IIB component
SUB0533	NA	sugar phosphotransferase system (PTS), IIC component
SUB0534	NA	conserved hypothetical protein
SUB0537	NA	putative alpha-amylase
SUB0541	NA	dihydroxyacetone kinase family protein
SUB0543	NA	dihydroxyacetone kinase family protein
SUB0544	NA	dihydroxyacetone kinase family protein
SUB0545	NA	sugar phosphotransferase system (PTS), IIA component
SUB0546	glpF2	glycerol uptake facilitator protein 2
SUB0577	ptsK	HPr kinase/phosphorylase
SUB0594	NA	phosphoglycerate mutase family protein
SUB0598	NA	glycosyl hydrolases family protein
SUB0602	ppc	putative phosphoenolpyruvate carboxylase
SUB0605	tpi	triosephosphate isomerase
SUB0616	NA	glycerophosphoryl diester phosphodiesterase
SUB0619	NA	putative fructose-1,6-bisphosphatase
SUB0625	NA	metallo-beta-lactamase superfamily protein
SUB0655	eno	enolase
SUB0660	NA	putative pullulanase
SUB0661	glgB	1,4-alpha-glucan branching enzyme
SUB0662	glgC	glucose-1-phosphate adenylyltransferase
SUB0663	NA	putative glycogen biosynthesis protein
SUB0664	glgA	glycogen synthase
SUB0665	glgP	glycogen phosphorylase
SUB0679	NA	acetyltransferase (GNAT) family protein

SUB0703	NA	glycosyl transferase
SUB0740	NA	putative glycerophosphodiester phosphodiesterase
SUB0748	NA	DeoR family regulatory protein
SUB0749	fruK	putative 1-phosphofructokinase
SUB0750	fruA	putative fructose-specific (PTS), IIABC component
SUB0759	NA	acetyltransferase (GNAT) family protein
SUB0792	lacR	lactose phosphotransferase system repressor
SUB0793	lacA2	galactose-6-phosphate isomerase LacA subunit 2
SUB0794	lacB2	galactose-6-phosphate isomerase LacB subunit 2
SUB0795	lacC2	tagatose-6-phosphate kinase 2
SUB0796	lacD2	tagatose 1,6-diphosphate aldolase 2
SUB0798	lacF2	putative lactose-specific (PTS), IIA component 2
SUB0799	lacE2	putative lactose-specific (PTS), IIBC component 2
SUB0800	lacG2	6-phospho-beta-galactosidase 2
SUB0806	NA	putative 6-phospho-beta-glucosidase
SUB0822	NA	response regulator protein
SUB0823	NA	putative sensor kinase
SUB0824	malP	sodium-dependent malate transporter
SUB0825	NA	NAD-dependent malic enzyme
SUB0830	NA	Beta-glucoside-specific (PTS), IIABC component
SUB0834	NA	6-phospho-beta-glucosidase 3
SUB0837	NA	6-phospho-beta-glucosidase 4
SUB0838	NA	phosphoglycerate mutase family protein
SUB0839	NA	phosphoglycerate mutase family protein
SUB0841	NA	6-phospho-beta-glucosidase 5
SUB0855	pta	putative phosphate acetyltransferase
SUB0856	NA	putative NADP-dependent L-serine/L-allo-threonine dehydrogenase
SUB0864	NA	putative 6-phospho-beta-glucosidase
SUB0877	nox	NADH oxidase
SUB0879	ldh	L-lactate dehydrogenase
SUB0893	NA	GntR family regulatory protein
SUB0905	NA	6-phospho-beta-glucosidase 6
SUB0927	maa	maltose O-acetyltransferase
SUB0952	deoC	deoxyribose-phosphate aldolase
SUB0978	NA	GntR family regulatory protein
SUB0994	mtlD	mannitol-1-phosphate 5-dehydrogenase
SUB0995	mtlF	putative mannitol-specific (PTS), IIA component
SUB0997	mtlA	putative mannitol-specific (PTS), IIBC component
SUB1000	pyk	putative pyruvate kinase
SUB1001	pfk	6-phosphofructokinase
SUB1003	NA	GntR family regulatory protein
SUB1049	deoB	phosphopentomutase
SUB1050	rpiA	ribose 5-phosphate isomerase A
SUB1066	NA	aldo/keto reductase family protein
SUB1072	NA	aldo/keto reductase family protein
SUB1100	acoL	dihydrolipoamide dehydrogenase
SUB1101	pdhC	dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase
SUB1102	pdhB	putative pyruvate dehydrogenase E1 component, beta subunit
SUB1103	pdhA	putative pyruvate dehydrogenase E1 component, alpha subunit
SUB1112	aldB	alpha-acetolactate decarboxylase
SUB1113	alsS	acetolactate synthase
SUB1117	NA	conserved hypothetical protein
SUB1140	glgP	putative glycogen phosphorylase
SUB1141	malM	putative 4-alpha-glucanotransferase
SUB1142	NA	LacI family transcriptional repressor
SUB1143	malX	putative maltose/maltodextrin-binding protein precursor
SUB1144	malC	putative maltose/maltodextrin ABC transport system permease protein

SUB1145	malD	putative maltose/maltodextrin ABC transport system permease protein
SUB1197	NA	glycosyl hydrolase family protein
SUB1198	NA	glycosyl hydrolase family protein
SUB1200	NA	putative beta-glucuronidase
SUB1201	kduD	putative 2-deoxy-D-gluconate 3-dehydrogenase
SUB1202	uxuA	putative mannonate dehydratase
SUB1203	uxaC	uronate isomerase
SUB1204	dgoA	putative 2-dehydro-3-deoxy-6-phosphogalactonate aldolase
SUB1205	NA	GntR family regulatory protein
SUB1206	uidA	putative beta-glucuronidase
SUB1207	NA	PfkB family carbohydrate kinase
SUB1220	gapN	putative NADP-dependent glyceraldehyde-3-phosphate dehydrogenase
SUB1222	pstI	phosphoenolpyruvate-protein phosphotransferase
SUB1223	ptsH	histidine-containing phosphocarrier protein (HPr)
SUB1263	gpmA	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase
SUB1298	glkA	glucokinase
SUB1307	rbsB	D-ribose-binding protein precursor
SUB1308	rbsC	ribose transport system permease protein RbsC
SUB1309	rbsA	ribose transport ATP-binding protein RbsA
SUB1310	rbsD	high affinity ribose transport protein RbsD
SUB1311	rbsK	ribokinase
SUB1312	NA	putative ribose transport operon repressor
SUB1345	NA	putative acetyltransferase
SUB1347	NA	putative aldose 1-epimerase
SUB1365	NA	short chain dehydrogenase
SUB1394	atoB2	acetyl-CoA acetyltransferase 2
SUB1395	atoD2	acetate CoA-transferase alpha subunit 2
SUB1396	atoA2	acetate CoA-transferase beta subunit 2
SUB1427	tkt1	putative transketolase 1
SUB1428	tal1	probable transaldolase 1
SUB1430	NA	pyridine nucleotide-disulphide oxidoreductase
SUB1431	glpF1	glycerol uptake facilitator protein 1
SUB1432	glpO	alpha-glycerophosphate oxidase
SUB1433	glpK	glycerol kinase
SUB1440	NA	sugar phosphotransferase system (PTS), IIC component (fragment)
SUB1443	NA	putative N-acetylglucosamine-6-phosphate deacetylase
SUB1448	lacG1	6-phospho-beta-galactosidase 1
SUB1451	lacF1	putative lactose-specific (PTS), IIA component 1
SUB1454	lacD1	tagatose 1,6-diphosphate aldolase 1
SUB1455	lacC1	tagatose-6-phosphate kinase 1
SUB1456	lacB1	galactose-6-phosphate isomerase LacB subunit 1
SUB1457	lacA1	galactose-6-phosphate isomerase LacA subunit 1
SUB1458	NA	sugar (PTS), galactitol-specific family, IIC component
SUB1459	NA	Sugar (PTS), IIB component
SUB1460	NA	Sugar (PTS), IIA component
SUB1461	lacR1	lactose phosphotransferase system repressor 1
SUB1482	manL	putative mannose-specific (PTS), IIAB component
SUB1485	manN	putative mannose-specific (PTS), IID component
SUB1486	NA	conserved hypothetical protein
SUB1509	NA	phosphoglycerate mutase family protein
SUB1519	ppdK	pyruvate, phosphate dikinase
SUB1533	NA	aldo/keto reductase family protein
SUB1538	pmi	mannose-6-phosphate isomerase
SUB1539	NA	putative beta-glucosidase
SUB1540	NA	sugar phosphotransferase system (PTS), IIC component
SUB1542	scrK1	fructokinase 1
SUB1543	scrA	putative sucrose-specific (PTS), IIABC component

SUB1544	scrB	putative sucrose-6-phosphate hydrolase
SUB1545	scrR	putative sucrose operon repressor
SUB1568	pfl	formate acetyltransferase
SUB1570	NA	putative beta-lactamase
SUB1579	NA	6-phospho-beta-glucosidase 7
SUB1583	NA	conserved hypothetical protein
SUB1584	NA	putative carboxylase subunit
SUB1601	kgdA	putative KHG/KDPG aldolase
SUB1602	NA	putative sugar kinase
SUB1604	idnO	putative gluconate 5-dehydrogenase
SUB1605	kdgT	2-keto-3-deoxygluconate permease
SUB1606	kduI	4-deoxy-l-threo-5-hexosulose-uronate ketol-isomerase
SUB1629	pgk	phosphoglycerate kinase
SUB1630	plr	glyceraldehyde-3-phosphate dehydrogenase
SUB1651	NA	acetyl xylan esterase (AXE1) family protein
SUB1671	NA	putative sensor kinase
SUB1672	NA	putative sensor kinase (pseudogene)
SUB1675	NA	putative sensor kinase
SUB1685	NA	putative NADPH-dependent FMN reductase
SUB1686	NA	putative NADPH-dependent FMN reductase
SUB1691	gpsA	glycerol-3-phosphate dehydrogenase [NAD(P)+]
SUB1694	NA	5-formyltetrahydrofolate cyclo-ligase family protein
SUB1700	srlB	putative glucitol/sorbitol-specific (PTS), IIA component
SUB1701	srlE	putative glucitol/sorbitol-specific (PTS), IIBC component
SUB1702	srlA	putative glucitol/sorbitol-specific (PTS), IIC2 component
SUB1703	gutM	glucitol operon activator protein
SUB1705	sorD	sorbitol-6-phosphate 2-dehydrogenase
SUB1706	pgi	glucose-6-phosphate isomerase
SUB1720	NA	conserved hypothetical protein
SUB1722	NA	L-ribulose 5-phosphate 4-epimerase
SUB1723	NA	putative hexulose-6-phosphate isomerase
SUB1724	NA	putative hexulose-6-phosphate synthase
SUB1725	NA	Sugar (PTS), IIA component
SUB1726	NA	sugar (PTS), lactose/cellobiose-specific family, IIB component
SUB1727	NA	putative sugar-specific permease, SgaT/UlaA family
SUB1728	NA	putative glycerate kinase
SUB1757	NA	putative transketolase subunit
SUB1758	NA	putative transketolase subunit
SUB1762	treA	putative trehalose-6-phosphate hydrolase
SUB1763	NA	sugar phosphotransferase system (PTS), IIABC component
SUB1764	NA	GntR family regulatory protein
SUB1801	NA	putative coenzyme A transferase
SUB1803	NA	putative fusion protein
SUB1804	thlA	acetyl-CoA acetyltransferase
SUB1806	hbd	3-hydroxybutyryl-CoA dehydrogenase
SUB1850	NA	putative 5'-nucleotidase

Cell Division and Cell Cycle

SUB0015	ftsH	putative cell division protease FtsH
SUB0017	NA	rod shape-determining protein MreC
SUB0057	NA	putative DNA-binding protein
SUB0127	NA	putative DNA-binding protein
SUB0382	NA	replication initiation and membrane attachment protein
SUB0405	scpA	segregation and condensation protein A

SUB0406	scpB	segregation and condensation protein B
SUB0407	rluB	ribosomal large subunit pseudouridine synthase B
SUB0476	NA	putative transposase
SUB0479	NA	DNA translocase FtsK
SUB0556	smc	putative chromosome partition protein
SUB0559	ftsY	putative cell division protein FtsY
SUB0603	ftsW	putative cell division protein
SUB0622	ftsE	putative cell division ATP-binding protein
SUB0623	ftsX	putative cell division protein
SUB0647	NA	putative peptidoglycan biosynthesis protein
SUB0650	ezrA	septation ring formation regulator
SUB0894	NA	putative DNA-binding protein
SUB0895	ffh	signal recognition particle protein
SUB0923	NA	putative DNA-binding protein
SUB1067	NA	putative chromate transport protein (pseudogene)
SUB1090	NA	putative phosphoglucosamine mutase
SUB1091	NA	putative exported protein
SUB1092	NA	putative membrane protein
SUB1127	NA	putative DNA replication protein DnaD
SUB1195	NA	putative DNA-binding protein
SUB1273	ftsJ	cell division protein
SUB1285	NA	putative cell-division protein DivIVA
SUB1288	sepF	putative cell division protein
SUB1290	ftsZ	cell division protein FtsZ
SUB1291	ftsA	putative cell division protein
SUB1292	NA	putative cell division protein
SUB1404	NA	DivIVA protein
SUB1405	NA	conserved hypothetical protein
SUB1406	recU	putative recombination protein U homologue
SUB1407	pbp1A	putative penicillin-binding protein 1A
SUB1420	ftsL	putative cell division protein
SUB1429	NA	putative DNA-binding protein
SUB1521	NA	putative DNA-binding protein
SUB1576	NA	putative DNA-binding protein
SUB1577	NA	putative DNA-binding protein
SUB1609	NA	putative DNA-binding protein (pseudogene)
SUB1613	NA	putative DNA-binding protein
SUB1619	NA	putative transcriptional regulator
SUB1662	NA	putative DNA-binding protein
SUB1869	NA	putative chromosome partitioning protein

Cell Wall and Capsule

SUB0008	NA	putative polysaccharide biosynthesis protein
SUB0010	NA	putative septum formation initiator protein
SUB0043	NA	Sugar (PTS), sorbose subfamily, IIB component
SUB0044	NA	Sugar (PTS), sorbose-specific family, IIC component
SUB0046	NA	Sugar (PTS), fructose family, IIA component
SUB0114	pbp1B	putative penicillin-binding protein 1B
SUB0170	NA	RpiR family regulatory protein
SUB0175	NA	putative microcin immunity protein
SUB0207	NA	putative surface-anchored protein
SUB0249	galE	UDP-glucose 4-epimerase (pseudogene)
SUB0306	NA	extracellular solute-binding protein
SUB0339	bacA	undecaprenol kinase (bacitracin resistance protein)

SUB0341	rgpG	putative glycosyl transferase
SUB0348	NA	putative surface-anchored protein (pseudogene)
SUB0349	NA	putative D-alanyl-D-alanine carboxypeptidase
SUB0350	NA	putative D-alanyl-D-alanine carboxypeptidase
SUB0351	NA	putative D-alanyl-D-alanine carboxypeptidase
SUB0365	NA	extracellular solute-binding protein
SUB0389	murC	UDP-N-acetylmuramate--alanine ligase
SUB0400	glr	glutamate racemase
SUB0425	murE	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-dia minopimelate ligase
SUB0426	NA	putative polysaccharide biosynthesis protein
SUB0468	gcaD	bifunctional GlmU protein
SUB0538	NA	putative glycosyltransferase
SUB0539	NA	putative glycosyltransferase
SUB0606	murN	putative peptidoglycan branched peptide synthesis protein
SUB0607	murM	putative peptidoglycan pentaglycine interpeptide biosynthesis protein
SUB0675	murA1	UDP-N-acetylglucosamine 1-carboxyvinyltransferase 1
SUB0688	NA	extracellular solute-binding protein
SUB0694	rmlD	dTDP-4-dehydrorhamnose reductase
SUB0695	NA	putative rhamnosyltransferase
SUB0696	NA	putative glycosyltransferase
SUB0697	NA	putative polysaccharide export ABC transporter permease protein
SUB0698	NA	putative polysaccharide export system ATP-binding protein
SUB0699	NA	putative glycosyl transferase
SUB0700	NA	rhamnan synthesis protein F family protein
SUB0705	NA	N-acetylmuramoyl-L-alanine amidase
SUB0714	NA	putative sulfatase
SUB0808	NA	pyridine nucleotide-disulphide oxidoreductase family protein
SUB0809	NA	putative polysaccharide deacetylase
SUB0817	murB	UDP-N-acetylenolpyruvoylglucosamine reductase
SUB0881	srtA	sortase A
SUB0888	NA	putative surface-anchored protein
SUB0904	NA	rpiR-family regulatory protein
SUB0998	glmS	glucosamine--fructose-6-phosphate aminotransferase [isomerizing]
SUB1019	NA	extracellular solute-binding protein
SUB1027	hasB2	UDP-glucose 6-dehydrogenase 2
SUB1028	NA	putative nucleotide sugar epimerase
SUB1029	mnaA	UDP-N-acetylglucosamine 2-epimerase
SUB1031	NA	putative exopolysaccharide biosynthesis protein
SUB1034	NA	putative glycosyl transferase
SUB1036	NA	putative glycosyl transferase
SUB1037	NA	putative UDP-glucose 4-epimerase
SUB1038	NA	putative glycosyl-1-phosphate-transferase
SUB1040	wze	tyrosine-protein kinase Wze
SUB1041	NA	putative chain length determinant protein
SUB1042	wzh	protein-tyrosine phosphatase Wzh
SUB1043	wzg	integral membrane regulatory protein Wzg
SUB1093	NA	putative UDP-N-acetylmuramyl peptide synthase
SUB1119	rmlB	putative dTDP-glucose-4,6-dehydratase
SUB1120	rmlC	putative dTDP-4-keto-6-deoxyglucose-3,5-epimerase
SUB1121	rmlA	glucose-1-phosphate thymidyl transferase
SUB1146	dltD	putative D-alanyl-lipoteichoic acid biosynthesis protein
SUB1147	dltC	D-alanyl carrier protein
SUB1148	dltB	putative activated D-alanine transport protein
SUB1149	dltA	D-alanine--poly(phosphoribitol) ligase subunit 1
SUB1210	murZ	UDP-N-acetylglucosamine 1-carboxyvinyltransferase 2
SUB1219	NA	polysaccharide deacetylase family protein
SUB1239	nagB	glucosamine-6-phosphate isomerase

SUB1255	murF	putative UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase
SUB1257	ddlA	D-alanine--D-alanine ligase
SUB1293	murG	putative UDP-N-acetylglucosamine transferase
SUB1294	murD	UDP-N-acetylmuramoylalanine--D-glutamate ligase
SUB1416	NA	extracellular solute-binding protein
SUB1418	mraY	phospho-N-acetylmuramoyl-pentapeptide-transfera se
SUB1419	pbpX	putative penicillin binding protein 2x
SUB1508	NA	putative D-alanyl-D-alanine carboxypeptidase
SUB1541	NA	RpiR family regulatory protein
SUB1582	NA	RpiR family regulatory protein
SUB1652	NA	ROK family protein
SUB1653	nanH	putative N-acetylneuraminate lyase
SUB1655	NA	extracellular solute-binding lipoprotein
SUB1657	NA	putative transport system permease
SUB1658	NA	putative N-acetylmannosamine-6-phosphate 2-epimerase
SUB1692	hasC	UTP--glucose-1-phosphate uridylyltransferase
SUB1696	hasB1	UDP-glucose 6-dehydrogenase 1
SUB1697	hasA	hyaluronan synthase
SUB1730	NA	putative surface-anchored protein
SUB1733	pbp2A	penicillin-binding protein 2a
SUB1739	NA	putative surface-anchored protein (pseudogene)

Cofactors, Vitamins and Prosthetic Groups

SUB0098	NA	putative lipoprotein
SUB0137	NA	putative lipoprotein
SUB0138	NA	putative lipoprotein
SUB0248	NA	putative lipoprotein
SUB0264	NA	Crp family regulatory protein
SUB0356	NA	putative nicotinate-nucleotide adenylyltransferase
SUB0445	serC	putative phosphoserine aminotransferase
SUB0447	NA	putative D-isomer specific 2-hydroxyacid dehydrogenase
SUB0469	NA	putative ADP-ribose pyrophosphatase
SUB0518	coaE	dephospho-CoA kinase
SUB0600	NA	putative lipoprotein
SUB0641	NA	putative flavodoxin
SUB0747	NA	putative ketopantoate reductase
SUB0772	dyr	dihydrofolate reductase
SUB0812	folC1	putative folypolyglutamate synthase 1
SUB0813	folE	GTP cyclohydrolase I
SUB0814	folP	dihydropteroate synthase
SUB0815	folQ	dihydroneopterin aldolase
SUB0816	folK	2-amino-4-hydroxy-6-hydroxymethyldihydropteridi ne pyrophosphokinase
SUB0853	ppnK	putative inorganic polyphosphate/ATP-NAD kinase
SUB0870	tdk	putative thymidine kinase
SUB0878	NA	conserved hypothetical protein
SUB0921	NA	protein Gid homolog
SUB0936	NA	GntR family regulatory protein
SUB0942	fhs1	formate--tetrahydrofolate ligase 1
SUB0943	dpfB	putative pantothenate metabolism flavoprotein
SUB0944	coaC	putative phosphopantothenoylcysteine decarboxylase
SUB0945	NA	putative membrane protein
SUB0950	NA	putative lipoprotein
SUB0955	coaA	pantothenate kinase
SUB0969	ribC	putative riboflavin biosynthesis protein

SUB1007	NA	putative lipoprotein
SUB1047	punA	putative purine nucleoside phosphorylase
SUB1098	NA	putative lipoate-protein ligase A
SUB1155	NA	nicotinamide mononucleotide transporter
SUB1211	metK	S-adenosylmethionine synthetase
SUB1213	birA	BirA bifunctional protein
SUB1241	NA	putative lipoprotein
SUB1261	NA	putative membrane protein
SUB1262	NA	conserved hypothetical protein
SUB1277	folD	FolD bifunctional protein
SUB1314	coaD	phosphopantetheine adenylyltransferase
SUB1409	nadE	NH(3)-dependent NAD(+) synthetase
SUB1410	NA	nicotinate phosphoribosyltransferase (NAPRTase) family protein
SUB1421	mraW	S-adenosyl-methyltransferase MraW
SUB1444	NA	putative lipoprotein
SUB1527	NA	isochorismatase family protein
SUB1639	thiN	putative thiamin pyrophosphokinase
SUB1709	NA	BioY family protein
SUB1779	cinA	CinA-like protein
SUB1802	NA	putative fatty-acid CoA ligase

DNA Metabolism

SUB0001	dnaA	chromosomal replication initiator protein
SUB0002	dnaN	DNA polymerase III, beta chain
SUB0004	NA	putative DNA-binding protein
SUB0007	trcF	putative transcription-repair coupling factor
SUB0023	plsX	fatty acid/phospholipid synthesis protein
SUB0059	ruvB	Holliday junction DNA helicase, subunit B
SUB0118	NA	putative competence protein
SUB0119	NA	putative competence protein
SUB0120	comYC	putative competence protein
SUB0121	NA	putative competence protein
SUB0122	NA	putative exported protein
SUB0123	NA	putative membrane protein
SUB0124	NA	putative exported protein
SUB0136	NA	putative single stranded DNA-binding protein
SUB0214	NA	putative ATPase
SUB0237	NA	endonuclease/exonuclease/phosphatase family protein
SUB0257	polC	DNA polymerase III PolC-type
SUB0279	NA	type I restriction-modification system M protein
SUB0282	NA	type I restriction-modification system R protein
SUB0340	mecA	adapter protein MecA
SUB0352	comX	putative competence-specific global transcription modulator
SUB0353	NA	HAD family hydrolase
SUB0383	dnaI	putative primosomal protein
SUB0385	NA	putative helicase
SUB0439	holB	putative DNA polymerase III, delta' subunit
SUB0448	ogt	putative methylated-DNA--protein-cysteine methyltransferase
SUB0454	exoA	exodeoxyribonuclease
SUB0498	NA	response regulator protein
SUB0499	NA	sensor histidine kinase
SUB0501	NA	bacteriocin transport/processing ATP-binding protein
SUB0517	fpg	formamidopyrimidine-DNA glycosylase
SUB0574	NA	putative transcription accessory protein

SUB0601	NA	putative oligopeptidase
SUB0627	NA	putative ATP-dependent DNA helicase
SUB0648	NA	putative haloacid dehalogenase-like hydrolase
SUB0649	gyrB	DNA gyrase subunit B
SUB0658	lig	DNA ligase
SUB0676	epuA	putative competence associated protein
SUB0677	endA	putative competence associated endonuclease
SUB0685	addB	putative ATP-dependent exonuclease subunit B
SUB0686	addA	putative ATP-dependent exonuclease subunit A
SUB0691	dnaG	DNA primase
SUB0776	clpL	putative ATP-dependent protease ATP-binding subunit ClpL
SUB0807	NA	putative transcriptional regulator
SUB0843	radC	DNA repair protein RadC
SUB0880	gyrA	DNA gyrase subunit A
SUB0900	NA	putative oxidoreductase
SUB0930	NA	putative DNA_binding protein
SUB0968	spxA	regulatory protein Spx
SUB0976	NA	TetR family regulatory protein
SUB0986	pcrA	putative ATP-dependent DNA helicase
SUB0996	mtlR	putative transcriptional regulator
SUB1002	dnaE	DNA polymerase III alpha subunit
SUB1011	parC	topoisomerase IV subunit A
SUB1014	parE	topoisomerase IV subunit B
SUB1017	ung	uracil-DNA glycosylase
SUB1056	uvrC	UvrABC system protein C
SUB1126	nth	putative endonuclease III
SUB1129	NA	putative single-stranded-DNA-specific exonuclease
SUB1151	uvrB	UvrABC system protein B (UvrB protein)
SUB1186	NA	putative DNA recombinase (fragment)
SUB1187	NA	putative modification methylase
SUB1214	dnaH	DNA polymerase III subunit gamma/tau
SUB1232	NA	putative O-methyltransferase
SUB1235	pepB	putative oligopeptidase
SUB1236	NA	competence protein CoiA-like family protein
SUB1243	NA	putative DNA polymerase III delta subunit
SUB1244	comEC	putative competence protein
SUB1245	comEA	putative competence protein
SUB1258	recR	recombination protein RecR
SUB1267	hlpA	DNA-binding protein HU
SUB1271	recN	putative DNA repair protein
SUB1275	xseB	putative exodeoxyribonuclease VII small subunit
SUB1276	xseA	putative exodeoxyribonuclease VII large subunit
SUB1300	NA	type IV leader peptidase family protein
SUB1334	NA	putative oxidoreductase
SUB1336	NA	Putative helicase
SUB1376	comFC	putative late competence protein
SUB1377	comFA	putative late competence protein
SUB1389	priA	putative primosomal protein N'
SUB1446	NA	TetR family regulatory protein
SUB1534	recG	ATP-dependent DNA helicase
SUB1549	NA	putative cytidine and deoxycytidylate deaminase
SUB1551	uvrA	UvrABC system protein A (UvrA protein)
SUB1556	ssb	single strand binding protein (SSB)
SUB1558	NA	putative A/G-specific adenine glycosylase
SUB1560	mutS2	putative DNA mismatch repair protein
SUB1565	NA	conserved hypothetical protein
SUB1567	dinB	DNA polymerase IV

SUB1611	NA	conserved hypothetical protein
SUB1612	NA	putative exonuclease
SUB1638	NA	RmuC family protein
SUB1640	rpe	putative ribulose-phosphate 3-epimerase
SUB1648	NA	TatD related DNase
SUB1650	NA	putative transcription regulator
SUB1670	fasA	response regulator protein
SUB1682	sms	putative DNA repair protein
SUB1699	NA	putative transferase
SUB1710	NA	CHY zinc finger protein
SUB1716	polA	DNA polymerase I
SUB1721	NA	putative transcriptional antiterminator
SUB1770	NA	putative oxidoreductase
SUB1777	NA	regulatory protein Spx
SUB1778	recA	RecA recombinase (recombinase A)
SUB1780	tag	DNA-3-methyladenine glycosylase I
SUB1781	ruvA	Holliday junction DNA helicase RuvA
SUB1783	mutL	DNA mismatch repair protein MutL
SUB1786	mutS	DNA mismatch repair protein MutS
SUB1816	NA	TetR family regulatory protein
SUB1843	dnaC	replicative DNA helicase
SUB1845	NA	DHH family phosphoesterase
SUB1860	recF	DNA replication and repair protein RecF

Fatty Acids and Isoprenoids

SUB0024	NA	putative acyl carrier protein
SUB0025	purC	phosphoribosylaminoimidazole-succinocarboxamide synthase
SUB0063	adhE	aldehyde-alcohol dehydrogenase 2
SUB0101	NA	zinc-binding dehydrogenase
SUB0108	NA	polyprenyl synthetase
SUB0109	ispE	4-diphosphocytidyl-2-c-methyl-D-erythritol kinase
SUB0252	uppS	undecaprenyl pyrophosphate synthetase
SUB0253	NA	putative phosphatidate cytidylyltransferase
SUB0333	NA	conserved hypothetical protein
SUB0465	NA	putative short chain dehydrogenase
SUB0491	dgk	diacylglycerol kinase
SUB0653	NA	putative phosphopentomutase
SUB0755	NA	DegV family protein – fatty acid binding protein
SUB0765	mvaK1	mevalonate kinase
SUB0766	mvaD	mevalonate diphosphate decarboxylase
SUB0767	mvaK2	phosphomevalonate kinase
SUB0768	fni	isopentenyl-diphosphate delta-isomerase
SUB0769	mvaA	3-hydroxy-3-methylglutaryl-coenzyme A reductase
SUB0770	mvaS	3-hydroxy-3-methylglutaryl coenzyme A synthase
SUB0940	NA	putative cardiolipin synthetase
SUB1015	NA	putative membrane protein
SUB1076	NA	putative zinc-binding alcohol dehydrogenase
SUB1107	NA	Metallo-beta-lactamase (conversion of glutathione to d-lactic acid)
SUB1246	NA	putative acyltransferase
SUB1274	fps	putative geranyltranstransferase
SUB1397	NA	putative D-beta-hydroxybutyrate dehydrogenase
SUB1398	NA	putative permease
SUB1490	accA	acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha
SUB1491	accD	acetyl-coenzyme A carboxylase carboxyl transferase subunit beta

SUB1492	accC	biotin carboxylase subunit of acetyl-CoA carboxylase
SUB1493	fabZ	(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase
SUB1494	fabE	biotin carboxyl carrier protein of acetyl-CoA carboxylase
SUB1495	fabF	3-oxoacyl-[acyl-carrier-protein] synthase II
SUB1496	fabG	3-oxoacyl-[acyl-carrier protein] reductase
SUB1497	fabD	putative malonyl CoA-acyl carrier protein transacylase
SUB1498	fabK	enoyl-ACP reductase
SUB1499	acpP	acyl carrier protein (ACP)
SUB1500	fabH	3-oxoacyl-[acyl-carrier-protein] synthase III
SUB1501	NA	MarR family regulatory protein
SUB1502	NA	enoyl-CoA hydratase/isomerase family protein
SUB1522	NA	3-hydroxybutyryl-CoA dehydrogenase (fragment)
SUB1536	acpS	holo-[acyl-carrier protein] synthase
SUB1855	pgsA	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase

Hypothetical Phage Proteins

SUB0217	NA	hypothetical phage protein
SUB0218	NA	hypothetical phage protein
SUB0219	NA	hypothetical phage protein
SUB0221	NA	hypothetical phage protein
SUB0224	NA	hypothetical phage protein
SUB0225	NA	hypothetical phage protein
SUB0227	NA	hypothetical phage protein
SUB1059	NA	hypothetical phage protein
SUB1060	NA	hypothetical phage protein
SUB1819	NA	hypothetical phage protein
SUB1820	NA	hypothetical phage protein
SUB1821	NA	hypothetical phage protein
SUB1822	NA	hypothetical phage protein
SUB1823	NA	hypothetical phage protein
SUB1825	NA	hypothetical phage protein
SUB1827	NA	hypothetical phage protein
SUB1828	NA	hypothetical phage protein
SUB1829	NA	hypothetical phage protein
SUB1830	NA	hypothetical phage protein
SUB1834	NA	hypothetical phage protein
SUB1835	NA	hypothetical phage protein
SUB1836	NA	hypothetical phage protein
SUB1837	NA	hypothetical phage protein

Hypothetical Proteins

SUB0011	NA	conserved hypothetical protein
SUB0051	NA	hypothetical protein
SUB0064	NA	hypothetical protein
SUB0117	NA	conserved hypothetical protein
SUB0125	NA	conserved hypothetical protein
SUB0148	NA	conserved hypothetical protein
SUB0149	NA	hypothetical protein
SUB0150	NA	conserved hypothetical protein
SUB0153	NA	conserved hypothetical protein
SUB0154	NA	conserved hypothetical protein

SUB0159	NA	conserved hypothetical protein
SUB0165	NA	conserved hypothetical protein
SUB0167	NA	conserved hypothetical protein
SUB0193	NA	hypothetical protein
SUB0209	NA	hypothetical protein
SUB0210	NA	conserved hypothetical protein
SUB0215	NA	conserved hypothetical protein
SUB0216	NA	hypothetical protein
SUB0232	NA	conserved hypothetical protein
SUB0250	NA	conserved hypothetical protein
SUB0256	NA	conserved hypothetical protein
SUB0268	NA	conserved hypothetical protein
SUB0271	NA	conserved hypothetical protein
SUB0274	NA	conserved hypothetical protein
SUB0278	NA	conserved hypothetical protein
SUB0283	NA	conserved hypothetical protein
SUB0292	NA	hypothetical protein
SUB0300	NA	conserved hypothetical protein
SUB0302	NA	conserved hypothetical protein
SUB0314	NA	conserved hypothetical protein
SUB0320	NA	conserved hypothetical protein
SUB0335	NA	conserved hypothetical protein
SUB0344	sufD	conserved hypothetical protein
SUB0347	sufB	conserved hypothetical protein
SUB0357	NA	conserved hypothetical protein
SUB0358	NA	conserved hypothetical protein
SUB0359	NA	conserved hypothetical protein
SUB0360	NA	conserved hypothetical protein
SUB0362	NA	conserved hypothetical protein
SUB0363	NA	conserved hypothetical protein
SUB0364	NA	conserved hypothetical protein
SUB0377	NA	conserved hypothetical protein
SUB0378	NA	hypothetical protein
SUB0396	NA	conserved hypothetical protein
SUB0403	NA	conserved hypothetical protein
SUB0419	NA	conserved hypothetical protein
SUB0420	NA	conserved hypothetical protein
SUB0430	NA	conserved hypothetical protein
SUB0431	NA	conserved hypothetical protein
SUB0437	NA	conserved hypothetical protein
SUB0453	NA	conserved hypothetical protein
SUB0464	NA	conserved hypothetical protein
SUB0466	NA	conserved hypothetical protein
SUB0485	NA	conserved hypothetical protein
SUB0487	NA	conserved hypothetical protein
SUB0510A	NA	conserved hypothetical protein (fragment)
SUB0514	NA	hypothetical protein
SUB0520	NA	conserved hypothetical protein
SUB0542	NA	conserved hypothetical protein
SUB0547	NA	conserved hypothetical protein
SUB0560	NA	conserved hypothetical protein
SUB0565	NA	conserved hypothetical protein
SUB0584	NA	conserved hypothetical protein
SUB0585	NA	conserved hypothetical protein
SUB0591	NA	conserved hypothetical protein
SUB0596	NA	conserved hypothetical protein
SUB0610	NA	conserved hypothetical protein

SUB0613	NA	hypothetical protein
SUB0624	NA	hypothetical protein
SUB0631	NA	conserved hypothetical protein
SUB0632	NA	conserved hypothetical protein
SUB0646	NA	conserved hypothetical protein
SUB0654	NA	conserved hypothetical protein
SUB0659	NA	conserved hypothetical protein
SUB0681	NA	conserved hypothetical protein
SUB0693	NA	conserved hypothetical protein
SUB0718	NA	conserved hypothetical protein
SUB0742	NA	conserved hypothetical protein
SUB0753	NA	conserved hypothetical protein
SUB0754	NA	conserved hypothetical protein
SUB0779	NA	conserved hypothetical protein
SUB0780	NA	conserved hypothetical protein
SUB0781	NA	conserved hypothetical protein
SUB0782	NA	conserved hypothetical protein
SUB0786	NA	conserved hypothetical protein
SUB0788	NA	conserved hypothetical protein
SUB0790A	NA	conserved hypothetical protein (fragment)
SUB0801	NA	conserved hypothetical protein
SUB0804	NA	conserved hypothetical protein (pseudogene)
SUB0847	NA	conserved hypothetical protein
SUB0848	NA	conserved hypothetical protein
SUB0852	NA	conserved hypothetical protein
SUB0858	NA	conserved hypothetical protein
SUB0875	NA	conserved hypothetical protein
SUB0887	NA	conserved hypothetical protein (fragment)
SUB0896	NA	hypothetical protein
SUB0897	NA	hypothetical protein
SUB0901	NA	conserved hypothetical protein
SUB0906	NA	conserved hypothetical protein
SUB0915	NA	hypothetical protein
SUB0920	NA	conserved hypothetical protein
SUB0932	NA	conserved hypothetical protein
SUB0934	NA	hypothetical protein
SUB0941	NA	conserved hypothetical protein
SUB0967	NA	conserved hypothetical protein
SUB0971	NA	conserved hypothetical protein
SUB0973	NA	conserved hypothetical protein
SUB0979	NA	conserved hypothetical protein
SUB0981	NA	conserved hypothetical protein
SUB0987	NA	hypothetical protein
SUB1022	NA	conserved hypothetical protein
SUB1026	NA	conserved hypothetical protein
SUB1032	NA	hypothetical protein
SUB1045	NA	conserved hypothetical protein
SUB1057	NA	conserved hypothetical protein
SUB1061	NA	hypothetical protein
SUB1074	NA	conserved hypothetical protein
SUB1079	NA	conserved hypothetical protein
SUB1083	NA	conserved hypothetical protein
SUB1089	NA	conserved hypothetical protein (pseudogene)
SUB1114	NA	conserved hypothetical protein
SUB1125	NA	conserved hypothetical protein
SUB1132	NA	conserved hypothetical protein
SUB1138	NA	conserved hypothetical protein

SUB1139	NA	conserved hypothetical protein
SUB1162	NA	conserved hypothetical protein
SUB1165	NA	conserved hypothetical protein
SUB1170	NA	conserved hypothetical protein
SUB1170A	NA	conserved hypothetical protein (pseudogene)
SUB1182	NA	hypothetical protein
SUB1184	NA	hypothetical protein
SUB1215	NA	conserved hypothetical protein
SUB1248	NA	conserved hypothetical protein
SUB1252	NA	hypothetical protein
SUB1270	NA	conserved hypothetical protein
SUB1280	NA	conserved hypothetical protein
SUB1283	NA	conserved hypothetical protein
SUB1299	NA	conserved hypothetical protein
SUB1305	NA	conserved hypothetical protein
SUB1331	NA	conserved hypothetical protein
SUB1332	NA	conserved hypothetical protein
SUB1340	NA	conserved hypothetical protein (fragment)
SUB1343	NA	conserved hypothetical protein
SUB1344	NA	conserved hypothetical protein
SUB1346	NA	conserved hypothetical protein (fragment)
SUB1351	NA	conserved hypothetical protein
SUB1354	NA	conserved hypothetical protein
SUB1355	NA	conserved hypothetical protein
SUB1366	NA	conserved hypothetical protein
SUB1373	NA	conserved hypothetical protein
SUB1374	NA	conserved hypothetical protein
SUB1378	NA	conserved hypothetical protein
SUB1400	NA	hypothetical protein
SUB1426	NA	conserved hypothetical protein
SUB1434	NA	conserved hypothetical protein
SUB1445	NA	conserved hypothetical protein
SUB1511	NA	conserved hypothetical protein
SUB1512	NA	conserved hypothetical protein
SUB1520	NA	conserved hypothetical protein
SUB1523	NA	conserved hypothetical protein
SUB1526	NA	conserved hypothetical protein
SUB1571	NA	conserved hypothetical protein
SUB1575	NA	conserved hypothetical protein
SUB1593	NA	conserved hypothetical protein
SUB1597	NA	hypothetical protein
SUB1598	NA	hypothetical protein
SUB1603	NA	conserved hypothetical protein
SUB1607A	NA	hypothetical protein
SUB1649	NA	hypothetical protein
SUB1680	NA	conserved hypothetical protein
SUB1719	NA	conserved hypothetical protein
SUB1734	NA	conserved hypothetical protein
SUB1735	NA	conserved hypothetical protein
SUB1746	NA	conserved hypothetical protein
SUB1750	NA	conserved hypothetical protein
SUB1767	NA	conserved hypothetical protein
SUB1787	NA	conserved hypothetical protein
SUB1800	NA	conserved hypothetical protein
SUB1817	NA	conserved hypothetical protein (fragment)
SUB1842	veg	conserved hypothetical protein
SUB1858	NA	conserved hypothetical protein

Iron Acquisition and Metabolism

SUB0202	NA	sensor histidine kinase
SUB0203	NA	response regulator protein
SUB0205	NA	permease protein
SUB0552	vicR	response regulator protein
SUB0553	vicK	sensor histidine kinase
SUB0554	vicX	metallo-beta-lactamase superfamily protein
SUB1684	NA	putative membrane protein

Membrane Transport

SUB0018	NA	putative membrane protein
SUB0033	ublB	putative membrane protein
SUB0034	ublC	putative membrane protein
SUB0035	ublD	ABC transporter ATP-binding protein
SUB0036	ublE	putative membrane protein
SUB0049	NA	putative membrane protein
SUB0052	NA	putative membrane protein
SUB0058	NA	putative membrane protein
SUB0061	NA	putative membrane protein
SUB0100	NA	putative membrane protein
SUB0106	cydC	ABC transporter ATP-binding membrane protein
SUB0107	cydD	ABC transporter ATP-binding membrane protein
SUB0111	NA	ABC transporter ATP-binding protein
SUB0112	NA	ABC transporter permease protein
SUB0128	NA	putative membrane protein
SUB0140	NA	ABC transporter ATP-binding protein
SUB0151	NA	putative membrane protein
SUB0163	NA	binding-protein-dependent transport system membrane protein
SUB0181	NA	putative extracellular sugar-binding protein
SUB0182	NA	sensor histidine kinase
SUB0183	NA	response regulator protein
SUB0184	NA	extracellular sugar-binding protein
SUB0185	NA	Sugar (PTS), fructose-specific family, IIA component
SUB0186	NA	Sugar (PTS), sorbose subfamily, IIB component
SUB0187	NA	Sugar (PTS), mannose/fructose family, IIC component
SUB0188	manZ	Sugar (PTS), mannose/fructose/sorbose family, IID component
SUB0189	NA	putative membrane protein
SUB0204	NA	ABC transporter ATP-binding protein
SUB0208	NA	putative membrane protein
SUB0211	NA	putative membrane protein
SUB0213	NA	putative membrane protein
SUB0223	NA	putative phage membrane protein
SUB0238	NA	putative membrane protein
SUB0272	NA	putative membrane protein (fragment)
SUB0295	NA	putative membrane protein
SUB0296	NA	ABC transporter, ATP-binding protein
SUB0297	NA	putative membrane protein
SUB0298	NA	putative membrane protein
SUB0299	NA	putative membrane protein
SUB0301	NA	putative membrane protein

SUB0304	NA	binding-protein-dependent transport system membrane protein
SUB0305	NA	binding-protein-dependent transport system membrane protein
SUB0311	NA	putative transport protein
SUB0312	NA	ABC transporter ATP-binding protein
SUB0313	NA	putative membrane protein
SUB0319	NA	putative membrane protein
SUB0334	NA	putative membrane protein
SUB0338	NA	putative membrane protein
SUB0343	sufC	putative ABC transporter, ATP-binding protein
SUB0386	NA	putative membrane protein
SUB0387	NA	putative membrane protein
SUB0397	NA	putative membrane protein
SUB0410	NA	putative membrane protein
SUB0413	NA	ABC transporter, ATP-binding protein
SUB0414	NA	putative membrane protein
SUB0432	NA	branched-chain amino acid ABC transporter, amino acid-binding protein
SUB0433	NA	branched-chain amino acid transport system permease protein
SUB0434	NA	branched-chain amino acid transport system permease protein
SUB0435	NA	branched-chain amino acid transport ATP-binding protein
SUB0436	NA	branched-chain amino acid transport ATP-binding protein
SUB0443	NA	putative membrane protein
SUB0456	NA	putative membrane protein
SUB0457	NA	putative membrane protein
SUB0458	NA	putative membrane protein
SUB0460	NA	putative membrane protein
SUB0461	NA	putative membrane protein (fragment)
SUB0470	NA	putative membrane protein
SUB0472	scaR	metal-dependent transcriptional regulator
SUB0473	mtuA	extracellular manganese-binding protein
SUB0474	mtuB	metal cation ABC transporter ATP-binding protein
SUB0475	mtuC	metal cation ABC transporter membrane protein
SUB0478	NA	putative membrane protein
SUB0480	NA	putative membrane protein
SUB0495	NA	putative membrane protein
SUB0496	NA	putative membrane protein
SUB0497	NA	ABC transporter, ATP-binding protein
SUB0500	NA	putative membrane protein
SUB0504	NA	putative membrane protein (pseudogene)
SUB0508	NA	putative membrane protein
SUB0511	NA	putative membrane protein
SUB0515	NA	putative membrane protein
SUB0519	NA	putative membrane protein
SUB0522	secG	putative protein-export membrane protein
SUB0532	NA	putative membrane protein
SUB0563	NA	putative membrane protein
SUB0570	NA	ABC transporter permease protein
SUB0571	NA	ABC transporter permease protein
SUB0576	NA	putative membrane protein
SUB0581	NA	putative membrane protein
SUB0586	NA	putative extracellular oligopeptide-binding protein
SUB0587	NA	putative membrane protein
SUB0595	NA	putative membrane protein
SUB0597	NA	putative membrane protein
SUB0599	NA	putative membrane protein
SUB0611	NA	putative membrane protein (fragment)
SUB0612	NA	putative membrane protein (fragment)
SUB0615	NA	putative membrane protein

SUB0657	NA	putative membrane protein
SUB0674	NA	putative membrane protein
SUB0683	NA	ABC transporter ATP-binding protein
SUB0701	NA	putative membrane protein
SUB0702	NA	putative membrane protein
SUB0707	NA	putative membrane protein
SUB0709	NA	putative membrane protein
SUB0720	NA	putative membrane protein
SUB0723	NA	putative membrane protein
SUB0737	NA	ABC transporter ATP-binding protein
SUB0739	NA	putative membrane protein
SUB0746	NA	putative membrane protein
SUB0758	NA	ABC transporter ATP-binding protein
SUB0760	NA	ABC transporter, ATP-binding/permease protein
SUB0761	NA	ABC transporter, ATP-binding/permease protein
SUB0775	NA	putative membrane protein
SUB0784	NA	putative membrane protein
SUB0842	NA	putative membrane protein
SUB0876	NA	putative membrane protein
SUB0886	NA	putative membrane protein
SUB0889	NA	putative di-/tripeptide transporter
SUB0910	NA	putative membrane protein
SUB0911	NA	putative membrane protein
SUB0919	NA	ABC transporter ATP-binding protein
SUB0922	NA	putative membrane protein
SUB0935	NA	ABC transporter ATP-binding protein
SUB0947	NA	ABC transporter permease protein
SUB0948	NA	ABC transporter permease protein
SUB0949	NA	ABC transporter ATP-binding protein
SUB0972	NA	putative membrane protein
SUB0974	NA	ABC transporter permease protein
SUB0975	NA	ABC transporter ATP-binding protein
SUB0982	NA	putative membrane protein
SUB0983	NA	putative membrane protein
SUB0984	NA	putative membrane protein
SUB0985	NA	putative membrane protein
SUB0990	NA	putative amino acid ABC transporter, binding protein
SUB0991	NA	putative amino acid ABC transporter ATP-binding protein
SUB0992	NA	putative amino acid ABC transporter permease protein
SUB1004	NA	ABC transporter ATP-binding protein
SUB1005	NA	putative membrane protein
SUB1006	NA	putative membrane protein
SUB1013	NA	putative membrane protein
SUB1033	NA	putative membrane protein
SUB1064	NA	putative membrane protein (pseudogene)
SUB1069	NA	putative membrane protein
SUB1071	NA	putative membrane protein
SUB1085	NA	putative membrane protein
SUB1104	NA	ABC transporter ATP-binding protein
SUB1108	NA	ABC transporter ATP-binding protein
SUB1109	NA	ABC transporter permease protein
SUB1115	NA	putative membrane protein
SUB1137	NA	putative membrane protein
SUB1150	NA	putative membrane protein
SUB1158	NA	putative membrane protein
SUB1167	NA	ABC transporter ATP-binding protein
SUB1171	NA	putative ABC transport/processing ATP-binding protein (pseudogene)

SUB1173	NA	putative membrane protein
SUB1174	NA	ABC transporter, ATP-binding protein
SUB1177	NA	putative membrane protein
SUB1216	NA	putative membrane protein
SUB1228	NA	putative membrane protein
SUB1254	NA	putative membrane protein
SUB1256	NA	putative membrane protein
SUB1265	NA	putative membrane protein
SUB1268	NA	putative membrane protein
SUB1279	NA	putative amino acid ABC transporter permease protein
SUB1287	NA	putative membrane protein
SUB1295	NA	putative membrane protein
SUB1301	NA	ABC transporter, ATP-binding/permease protein
SUB1302	NA	ABC transporter, ATP-binding/permease protein
SUB1337	NA	putative membrane protein
SUB1339	NA	putative membrane protein
SUB1356	NA	ABC transporter, ATP-binding protein
SUB1357	NA	putative membrane protein
SUB1384	NA	putative membrane protein
SUB1402	NA	putative membrane protein
SUB1413	NA	putative membrane protein
SUB1414	NA	ABC transporter ATP-binding protein
SUB1415	NA	ABC transporter permease protein
SUB1424	NA	putative membrane protein
SUB1425	NA	ABC transporter ATP-binding protein
SUB1473	NA	ABC transporter protein EcsB
SUB1474	NA	ABC transporter ATP-binding protein
SUB1476	NA	conserved hypothetical protein
SUB1487	NA	putative membrane protein
SUB1537	secA	putative preprotein SecA subunit
SUB1552	NA	putative magnesium transporter
SUB1553	NA	putative membrane protein
SUB1566	NA	putative membrane protein
SUB1572	NA	putative membrane protein
SUB1573	gla	glycerol uptake permease
SUB1578	NA	putative membrane protein
SUB1580	NA	putative membrane protein
SUB1581	NA	putative membrane protein
SUB1590	NA	putative membrane protein
SUB1596	NA	putative membrane protein
SUB1614	NA	putative membrane protein
SUB1621	dppF	di-tripeptide transporter ATP-binding protein
SUB1622	dppD	di-tripeptide transporter ATP-binding protein
SUB1623	dppC	di-tripeptide transporter permease protein
SUB1624	dppB	di-tripeptide transport system permease protein
SUB1625	dppA	putative peptide binding protein
SUB1628	NA	putative membrane protein
SUB1635	NA	putative membrane anchored protein
SUB1643	NA	putative membrane protein
SUB1646	NA	putative membrane protein
SUB1661	NA	putative membrane protein
SUB1663	NA	ABC-2 type transporter membrane protein
SUB1664	sagG	ABC transporter ATP-binding protein
SUB1688	NA	ABC transporter ATP-binding membrane protein
SUB1689	NA	ABC transporter ATP-binding membrane protein
SUB1713	NA	putative membrane protein
SUB1732	NA	putative preprotein translocase SecE subunit

SUB1747	NA	putative membrane protein
SUB1773	NA	putative membrane protein
SUB1790	NA	putative membrane protein
SUB1791	NA	putative membrane protein
SUB1792	NA	putative membrane protein
SUB1799	NA	putative membrane protein
SUB1812	NA	putative membrane protein
SUB1814	NA	putative membrane protein
SUB1815	NA	putative membrane protein
SUB1852	NA	putative cobalt transport protein
SUB1853	NA	ABC transporter ATP-binding protein
SUB1854	NA	ABC transporter ATP-binding protein
SUB1856	NA	putative membrane protein
SUB1859	yaaA	S4 domain-containing protein (reduces intracellular iron levels)
SUB1864	NA	putative membrane protein
SUB1865	NA	ABC transporter, ATP-binding protein
SUB1866	NA	putative membrane protein

Miscellaneous

SUB0003	NA	putative GTPase
SUB0012	NA	putative exported protein
SUB0016	NA	putative amino acid permease
SUB0019	NA	putative amidase
SUB0037A	NA	bifunctional purine biosynthesis protein PurH
SUB0039	agaS	putative tagatose-6-phosphate aldose/ketose isomerase
SUB0048	NA	putative amidase
SUB0060	NA	putative phosphotyrosine protein phosphatase
SUB0062	NA	acyltransferase family protein
SUB0066	NA	multi antimicrobial extrusion (MATE) family transporter
SUB0096	NA	AraC family regulatory protein (fragment)
SUB0099	NA	pyridoxamine 5'-phosphate oxidase protein
SUB0102	NA	UbiA prenyltransferase family protein
SUB0113	tyrS	putative tyrosyl-tRNA synthetase
SUB0129	NA	CAAX amino terminal protease family membrane protein
SUB0132	NA	putative exported protein
SUB0139	NA	transport system permease protein
SUB0147	NA	prophage maintenance system killer protein (fragment)
SUB0155	NA	basic membrane protein
SUB0157	NA	DeoR family regulatory protein
SUB0162	qacH	putative quaternary ammonium compound-resistance protein
SUB0164	NA	putative fusion protein (pseudogene)
SUB0166	NA	DNA-binding protein
SUB0172	NA	oxidoreductase family protein
SUB0173	NA	NADH:flavin oxidoreductase / NADH oxidase family protein
SUB0174	NA	putative ribosomal RNA large subunit methyltransferase A
SUB0176	NA	putative exported protein
SUB0178	NA	CAAX amino terminal protease family protein
SUB0179	NA	Acetyltransferase (GNAT) family protein
SUB0180	NA	putative haloacid dehalogenase-like hydrolase
SUB0190	NA	putative exported protein
SUB0191	NA	sensor histidine kinase
SUB0206	NA	putative exported protein
SUB0212	NA	putative exported protein
SUB0220	NA	putative phage replication protein

SUB0230	NA	putative phage membrane protein
SUB0231	NA	putative phage integrase
SUB0244	lrp	leucine-rich protein
SUB0246	dexB	glucan 1,6-alpha-glucosidase
SUB0247	NA	putative exported protein
SUB0251	NA	putative preprotein translocase subunit
SUB0258	NA	MarR family regulatory protein
SUB0261	NA	PAP2 superfamily
SUB0265	NA	Major Facilitator Superfamily protein
SUB0267	pnpA	polyribonucleotide nucleotidyltransferase
SUB0270	cysS	cysteinyl-tRNA synthetase
SUB0273	NA	SpoU rRNA methylase family protein
SUB0275	NA	DegV family protein
SUB0280	NA	type I restriction-modification system S protein (pseudogene)
SUB0284	NA	DeoR family regulatory protein
SUB0323	tig	trigger factor (prolyl isomerase)
SUB0327A	NA	antiholin-like protein (fragment)
SUB0329	NA	putative exported protein
SUB0336	glnQ2	putative glutamine ABC transporter, ATP-binding protein 2
SUB0337	NA	glutamine ABC transporter, glutamine-binding protein/permease protein
SUB0342	NA	putative exported protein
SUB0346	iscU	NifU-like protein
SUB0361	NA	putative regulatory protein
SUB0373	gidB	methyltransferase GidB
SUB0375	NA	LemA family protein
SUB0376	htpX	probable protease HtpX homolog
SUB0380	csrS	sensor histidine kinase
SUB0390	NA	acetyltransferase (GNAT) family protein
SUB0391	NA	putative aminodeoxychorismate lyase
SUB0394	NA	putative acylphosphatase
SUB0395	NA	SpoU rRNA Methylase family protein
SUB0399	NA	putative exported protein
SUB0404	xerD	tyrosine recombinase XerD
SUB0411	NA	PAP2 superfamily protein
SUB0412	NA	radical SAM superfamily protein
SUB0416	NA	putative transport protein
SUB0423	fhuD	putative ferrichrome-binding protein precursor
SUB0427	NA	putative aminotransferase
SUB0446	NA	acetyltransferase (GNAT) family protein
SUB0463	NA	putative exported protein
SUB0494	NA	putative exported protein
SUB0525	csbB	putative glucosyl transferase
SUB0527	NA	nitroreductase family protein
SUB0528	NA	putative exported protein
SUB0535	pepQ	putative Xaa-Pro dipeptidase
SUB0557	NA	putative haloacid dehalogenase-like hydrolase
SUB0561	NA	putative N-acetylmuramoyl-L-alanine amidase
SUB0579	NA	putative exported protein
SUB0580	NA	putative exported protein
SUB0589	NA	ammonium transporter family protein (pseudogene)
SUB0626	NA	putative esterase
SUB0630	NA	P-loop ATPase protein family protein
SUB0634	adcA	zinc-binding protein AdcA precursor
SUB0638	NA	putative extracellular amino acid-binding protein
SUB0639	hisC	putative histidinol-phosphate aminotransferase
SUB0643	NA	voltage gated chloride channel protein
SUB0684	NA	putative neutral zinc metallopeptidase

SUB0704	NA	putative sugar translocase
SUB0728	NA	LysR family regulatory protein
SUB0736	NA	putative exported protein
SUB0751	mur1.1	mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase family protein
SUB0752	NA	mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase family protein
SUB0764	NA	putative surface-anchored 5'-nucleotidase (pseudogene)
SUB0772A	NA	putative exported protein
SUB0783	NA	putative FtsK/SpoIIIE family protein
SUB0785	NA	putative plasmid replication protein
SUB0787	NA	integrase
SUB0789	NA	putative replication initiator protein
SUB0790	NA	C-5 cytosine-specific DNA methylase (pseudogene)
SUB0802	lacX2	aldose 1-epimerase 2
SUB0803	NA	putative DNA_binding protein
SUB0827	NA	voltage gated chloride channel family protein
SUB0831	NA	AraC family regulatory protein
SUB0835	NA	GDSL-like Lipase/acylhydrolase family protein
SUB0840	NA	putative transporter protein
SUB0860	NA	zinc-binding dehydrogenase
SUB0861	NA	MarR family regulatory protein
SUB0869	NA	putative 4-oxalocrotonate tautomerase
SUB0883	NA	putative exported protein
SUB0884	lmb	laminin binding protein
SUB0885	NA	Streptococcal histidine triad protein
SUB0890	NA	putative Na(+)/H(+) antiporter
SUB0892	NA	putative exported protein
SUB0898	NA	impB/mucB/samB family
SUB0899	NA	putative phage repressor-like protein
SUB0903	NA	putative 2,5-diketo-D-gluconic acid reductase
SUB0907	NA	NADPH-dependent FMN reductase
SUB0913	NA	tyrosine recombinase XerC
SUB0917	NA	response regulator protein
SUB0918	NA	putative transporter permease protein
SUB0924	NA	LysR family regulatory protein
SUB0925	topA	DNA topoisomerase I
SUB0931	NA	haemolysin-III related membrane protein
SUB0933	NA	putative exported protein
SUB0954	NA	putative methyltransferase
SUB0958	ciaR	response regulator protein
SUB0959	pepN	putative lysyl-aminopeptidase
SUB0966	NA	putative myo-inositol-1(or 4)-monophosphatase
SUB0980	NA	CsbD-like protein
SUB0988	NA	sodium:alanine symporter family protein
SUB1018	NA	transport system membrane protein
SUB1020	NA	putative amidase
SUB1021	NA	Phosphatidylethanolamine-binding protein
SUB1023	NA	acetyltransferase (GNAT) family protein
SUB1030	NA	putative acetyltransferase
SUB1035	NA	putative galacturonic acid acetylase (fragment)
SUB1044	NA	LysR family regulatory protein
SUB1052	NA	CAAX amino terminal protease family protein (pseudogene)
SUB1054	pepV	putative Xaa-His dipeptidase
SUB1055	NA	nitroreductase family protein
SUB1058	NA	putative succinate-semialdehyde dehydrogenase
SUB1070	NA	putative lipase
SUB1073	NA	carboxymuconolactone decarboxylase family protein
SUB1075	NA	MerR family regulatory protein

SUB1077	NA	putative short chain dehydrogenase
SUB1078	NA	Aldo/keto reductase family
SUB1087	NA	putative acyl-ACP thioesterase
SUB1095	sclB	collagen-like surface-anchored protein
SUB1099	NA	cation transporting ATPase (fragment)
SUB1105	NA	putative protease
SUB1106	NA	putative esterase
SUB1110	NA	putative exported protein
SUB1111	fbpS	fibronectin/fibrinogen-binding protein
SUB1122	NA	FAD dependent oxidoreductase
SUB1123	NA	ZIP zinc transporter family protein
SUB1124	NA	NIF3 (NGG1p interacting factor 3) family protein
SUB1130	NA	short chain dehydrogenase
SUB1135	NA	putative exported protein
SUB1136	NA	C4-dicarboxylate transporter/malic acid transport protein
SUB1160	NA	CorA-like3 Mg2+ transporter protein
SUB1161	NA	putative ribosomal small subunit pseudouridine synthase B
SUB1163	NA	thioesterase superfamily protein
SUB1164	NA	major facilitator superfamily protein
SUB1166	NA	putative membrane protein
SUB1168	NA	thioesterase superfamily protein
SUB1169	NA	acetyltransferase (GNAT) family protein
SUB1176	NA	MutT/NUDIX family protein
SUB1178	NA	relaxase (fragment)
SUB1178A	NA	putative transposase (fragment)
SUB1185	NA	puative DNA-binding protein
SUB1188	NA	putative type III site-specific deoxyribonuclease
SUB1208	NA	putative sugar transporter (pseudogene)
SUB1229	NA	voltage gated chloride channel
SUB1231	prsA1	foldase protein PrsA 1 precursor
SUB1233	NA	major facilitator superfamily protein
SUB1237	NA	major facilitator superfamily protein
SUB1238	NA	putative ribosomal small subunit pseudouridine synthase A
SUB1259	penA	penicillin-binding protein 2b (pseudogene)
SUB1260	NA	putative formate/nitrite transporter family protein
SUB1269	NA	GDSL-like lipase/acylhydrolase protein
SUB1281	clpE	putative ATP-dependent Clp protease ATP-binding subunit
SUB1303	NA	VanZ like family protein
SUB1304	NA	radical SAM superfamily protein
SUB1313	NA	putative exported protein
SUB1325	NA	petidase family M20/M25/M40 protein
SUB1333	NA	CorA-like Mg2+ transporter protein
SUB1349	NA	MarR family regulatory protein
SUB1352	NA	major facilitator superfamily protein
SUB1353	NA	putative amidase
SUB1359	NA	putative transporter protein
SUB1363	NA	major facilitator superfamily protein
SUB1364	NA	major facilitator superfamily protein
SUB1367	NA	LysR family regulatory protein
SUB1368	NA	putative esterase
SUB1369	NA	acetyltransferase (GNAT) family protein
SUB1370	NA	putative zinc carboxypeptidase
SUB1372	NA	RecX family protein
SUB1382	NA	response regulator protein
SUB1385	stk1	serine/threonine-protein kinase
SUB1387	NA	putative ribosomal RNA small subunit methyltransferase
SUB1393	NA	LysR family regulatory protein

SUB1401	NA	hypothetical protein with signal peptide
SUB1411	NA	amino acid permease
SUB1438	NA	isoprenylcysteine carboxyl methyltransferase (ICMT) family protein
SUB1438A	NA	transposase (pseudogene)
SUB1450	lacE1	putative lactose-specific (PTS), IIBC component 1
SUB1453	lacX1	aldose 1-epimerase 1
SUB1478	NA	acetyltransferase (GNAT) family protein
SUB1481	NA	putative haloacid dehalogenase-like hydrolase
SUB1483	manM	putative mannose-specific (PTS), IIC component (pseudogene)
SUB1507	NA	putative N-acetyl-muramidase
SUB1510	NA	CsbD-like protein
SUB1514	NA	transglycosylase associated protein
SUB1524	NA	putative pyridine nucleotide-disulphide oxidoreductase
SUB1569	NA	acetyltransferase (GNAT) family protein
SUB1574	pepXP	Xaa-Pro dipeptidyl-peptidase
SUB1585	citX	puatative apo-citrate lyase phosphoribosyl-dephospho-CoA transferase
SUB1586	citF	putative citrate lyase alpha subunit
SUB1587	citE	putative citrate lyase beta subunit (pseudogene)
SUB1589	citD	putative citrate lyase acyl carrier protein
SUB1594	NA	putative Mg2+/citrate transporter (fragment)
SUB1600	NA	putative transposase (fragment)
SUB1607	NA	LacI family regulatory protein
SUB1608	NA	putative plasmid replication protein (pseudogene)
SUB1610	NA	putative integrase
SUB1627	glnR	MerR family regulatory protein
SUB1637	NA	putative 3'-5' exoribonuclease
SUB1642	NA	major facilitator superfamily protein
SUB1645	NA	putative exported protein
SUB1656	NA	putative transport system permease
SUB1660	NA	putative exported protein
SUB1668	argH	argininosuccinate lyase
SUB1669	argG	argininosuccinate synthase
SUB1681	NA	putative carbonic anhydrase
SUB1687	NA	putative exported protein
SUB1690	NA	MarR family regulatory protein
SUB1693	NA	rhomboid family membrane protein
SUB1708	NA	metallo-beta-lactamase superfamily protein
SUB1715	NA	putative CoA binding protein
SUB1737	NA	putative exported protein
SUB1756	pepO	putative endopeptidase O
SUB1759	NA	putative sugar-specific permease, SgaT/UlaA family
SUB1760	NA	sugar (PTS), lactose/cellobiose-specific family, IIB component
SUB1761	NA	putative PTS multi-domain regulator
SUB1769	NA	acetyltransferase (GNAT) family protein
SUB1776	NA	regulatory protein
SUB1782	lmrP	putative transporter protein
SUB1784	NA	putative exported protein
SUB1798	NA	putative DNA-binding membrane protein (pseudogene)
SUB1805	NA	putative acyl-CoA dehydrogenase
SUB1810	NA	NUDIX hydrolase
SUB1811	NA	PadR family regulatory protein
SUB1813	NA	putative exported protein
SUB1818	NA	Fic protein family phage protein
SUB1832A	NA	putative phage membrane protein
SUB1840	NA	phage integrase
SUB1846	gidA	glucose inhibited division protein A
SUB1851	NA	putative transglycosylase protein

SUB1857	NA	putative protease
SUB1861	NA	putative sugar uptake protein
SUB1868	htrA	serine protease

Nucleotides and Nucleosides

SUB0014	hpt	hypoxanthine-guanine phosphoribosyltransferase
SUB0020	prsA1	ribose-phosphate pyrophosphokinase 1
SUB0021	NA	putative aminotransferase
SUB0026	NA	putative phosphoribosylformylglycinamidine synthase protein
SUB0027	purF	putative amidophosphoribosyltransferase precursor
SUB0028	purM	phosphoribosylformylglycinamidine cyclo-ligase
SUB0029	purN	phosphoribosylglycinamide formyltransferase
SUB0030	purH	bifunctional purine biosynthesis protein PurH
SUB0031	NA	response regulator protein (pseudogene)
SUB0053	purD	phosphoribosylamine-glycine ligase
SUB0054	purE	putative phosphoribosylaminoimidazole carboxylase catalytic subunit
SUB0055	purK	putative phosphoribosylaminoimidazole carboxylase ATPase subunit
SUB0056	purB	adenylosuccinate lyase
SUB0089	adk	adenylate kinase
SUB0133	NA	putative thioredoxin
SUB0141	NA	putative deoxyadenosine kinase protein
SUB0152	purA	adenylosuccinate synthetase
SUB0240	NA	putative NrdI-like protein
SUB0241	NA	putative surface-anchored 2',3'-cyclic-nucleotide 2'-phosphodiesterase
SUB0292A	NA	putative dihydrolipoamide dehydrogenase (fragment)
SUB0318	thiD	putative phosphomethylpyrimidine kinase
SUB0326	pyrG	putative CTP synthase
SUB0381	NA	ATP-binding protein
SUB0401	NA	HAM1 protein homolog
SUB0428	upp	uracil phosphoribosyltransferase
SUB0438	tmk	thymidylate kinase
SUB0483	pyrH	uridylate kinase
SUB0564	NA	putative permease
SUB0588	NA	conserved hypothetical protein
SUB0609	NA	putative phosphohydrolase
SUB0617	NA	putative phosphoesterase
SUB0633	pepD	putative dipeptidase
SUB0682	NA	putative permease
SUB0710	cmk	cytidylate kinase
SUB0731	pyrR	PyrR bifunctional protein
SUB0732	pyrP	uracil permease
SUB0733	pyrB	aspartate carbamoyltransferase
SUB0734	carA	carbamoyl-phosphate synthase small chain
SUB0735	carB	carbamoyl-phosphate synthase large chain
SUB0738	NA	putative permease
SUB0745	NA	pyridine nucleotide-disulphide oxidoreductase family protein
SUB0771	thyA	thymidylate synthase
SUB0850	prsA2	putative ribose-phosphate pyrophosphokinase 2
SUB0865	NA	GMP reductase
SUB0866	xpt	putative xanthine phosphoribosyltransferase
SUB0867	pbuX	putative xanthine permease
SUB0891	guaA	GMP synthase [glutamine-hydrolyzing]
SUB0914	NA	putative NUDIX hydrolase
SUB0937	NA	ATP-binding protein
SUB0946	pgmA	putative phosphomannomutase
SUB0951	cdd	cytidine deaminase

SUB0953	deoA	thymidine phosphorylase
SUB1016	pyrC	putative dihydroorotase
SUB1024	pyrE	orotate phosphoribosyltransferase
SUB1025	pyrF	orotidine 5'-phosphate decarboxylase
SUB1046	deoD	putative purine nucleoside phosphorylase
SUB1084	ndk	nucleoside diphosphate kinase
SUB1116	mutX	mutator MutX protein (7,8-dihydro-8-oxoguanine-triphosphatase)
SUB1217	udk	uridine kinase
SUB1224	nrdH	glutaredoxin-like protein
SUB1225	NA	putative ribonucleoside-diphosphate reductase alpha chain
SUB1227	NA	putative ribonucleoside-diphosphate reductase beta chain
SUB1264	pyrD	putative dihydroorotate dehydrogenase
SUB1282	NA	Mut/NUDIX family protein
SUB1306	NA	putative phosphoesterase
SUB1335	nudG	putative CTP pyrophosphohydrolase
SUB1391	gmk	guanylate kinase
SUB1392	NA	putative phosphohydrolase
SUB1412	NA	putative thioredoxin reductase
SUB1475	NA	putative histidine triad protein
SUB1480	NA	putative permease
SUB1559	NA	putative thioredoxin
SUB1636	purR	putative pur operon repressor
SUB1683	NA	deoxyuridine 5'-triphosphate nucleotidohydrolase
SUB1738	NA	putative dipeptidase
SUB1748	NA	putative dioxygenase
SUB1749	NA	putative dioxygenase
SUB1751	NA	putative phospholipase/carboxylesterase
SUB1768	nrdG	putative anaerobic ribonucleoside-triphosphate reductase activating
SUB1772	nrdD	anaerobic ribonucleoside-triphosphate reductase
SUB1797	NA	NUDIX hydrolase
SUB1807	NA	putative thioredoxin
SUB1808	NA	putative permease
SUB1862	guaB	inosine-5'-monophosphate dehydrogenase

Phages, Transposable elements and Plasmids

SUB0222	NA	putative DNA-binding phage protein
SUB0228	NA	putative DNA-binding phage protein
SUB0229	NA	putative DNA-binding phage protein
SUB0805	NA	putative integrase (pseudogene)
SUB1824	NA	phage membrane protein
SUB1831	NA	putative DNA-binding phage protein
SUB1832	NA	putative DNA-binding phage protein
SUB1838	NA	putative DNA-binding phage protein
SUB1839	NA	putative DNA-binding phage protein

Phosphorus Metabolism

SUB0262	ppk	polyphosphate kinase
SUB0263	NA	Ppx/GppA phosphatase family protein
SUB0379	csrR	transcriptional regulator protein
SUB0418	ppaC	manganese-dependent inorganic pyrophosphatase
SUB0451	NA	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein

SUB0452	NA	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein
SUB0488	NA	PhoH-like protein
SUB0490	NA	conserved hypothetical protein
SUB0566	NA	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein
SUB0567	phoP	alkaline phosphatase synthesis transcriptional regulatory protein
SUB0568	phoR	putative alkaline phosphatase synthesis sensor protein
SUB0569	NA	putative lipoprotein
SUB0572	pstB3	phosphate import ATP-binding protein 3
SUB0573	NA	phosphate transport system protein
SUB0652	NA	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein
SUB0882	NA	glyoxalase/bleomycin resistance protein
SUB0960	phoU	putative phosphate transport system protein
SUB0961	pstB1	phosphate import ATP-binding protein 1
SUB0962	pstB2	phosphate import ATP-binding protein 2
SUB0963	NA	putative phosphate ABC transporter permease protein
SUB0964	pstC	putative phosphate ABC transporter permease protein
SUB0965	pstS	putative phosphate ABC transporter binding protein
SUB0993	NA	PhnA protein
SUB1766	NA	glyoxalase/bleomycin resistance protein/dioxygenase superfamily protein

Potassium Metabolism

SUB0321	NA	mechanosensitive ion channel protein
SUB0371	NA	TrkA family transport protein
SUB0372	NA	cation transport protein
SUB0690	mscL	large-conductance mechanosensitive channel
SUB1249	NA	probable potassium transport system protein
SUB1297	NA	putative membrane protein

Protein Metabolism

SUB0005	NA	putative GTP-binding protein
SUB0006	pth	peptidyl-tRNA hydrolase
SUB0067	rpsJ	30S ribosomal protein S10
SUB0068	rplC	50S ribosomal protein L3
SUB0069	rplD	50S ribosomal protein L4
SUB0070	rplW	50S ribosomal protein L23
SUB0071	rplB	50S ribosomal protein L2
SUB0072	rpsS	30S ribosomal protein S19
SUB0073	rplV	50S ribosomal protein L22
SUB0074	rpsC	30S ribosomal protein S3
SUB0075	rplP	50S ribosomal protein L16
SUB0076	rpmC	50S ribosomal protein L29
SUB0077	rpsQ	30S ribosomal protein S17
SUB0078	rplN	50S ribosomal protein L14
SUB0079	rplX	50S ribosomal protein L24
SUB0080	rplE	50S ribosomal protein L5
SUB0081	rpsN	30S ribosomal protein S14
SUB0082	rpsH	30S ribosomal protein S8
SUB0083	rplF	50S ribosomal protein L6
SUB0084	rplR	50S ribosomal protein L18
SUB0085	rpsE	30S ribosomal protein S5
SUB0086	rpmD	50S ribosomal protein L30

SUB0087	rplO	50S ribosomal protein L15
SUB0088	secY	preprotein translocase SecY subunit
SUB0090	infA	translation initiation factor IF-1
SUB0091	rpmJ	50S ribosomal protein L36
SUB0092	rpsM	30S ribosomal protein S13
SUB0093	rpsK	30S ribosomal protein S11
SUB0095	rplQ	50S ribosomal protein L17
SUB0131	pepA	putative glutamyl-aminopeptidase
SUB0233	NA	ribosomal protein L11 methyltransferase
SUB0239	NA	M42 gutamyl aminopeptidase
SUB0260	def	polypeptide deformylase
SUB0266	rpsO	30S ribosomal protein S15
SUB0276	rplM	50S ribosomal protein L13
SUB0277	rpsI	30S ribosomal protein S9
SUB0331	rpmB	50S ribosomal protein L28
SUB0354	NA	GTP-binding protein
SUB0384	engA	GTP-binding protein EngA
SUB0388	NA	putative cystathionine gamma synthase
SUB0481	rplK	50S ribosomal protein L11
SUB0482	rplA	50S ribosomal protein L1
SUB0484	rrf	ribosome recycling factor
SUB0486	msrA	peptide methionine sulfoxide reductase
SUB0521A	rpmGB	50S ribosomal protein L33 2
SUB0524	smpB	SsrA-binding protein
SUB0540	thrS	threonyl-tRNA synthetase
SUB0578	lgt	prolipoprotein diacylglyceryl transferase
SUB0604	tufA	elongation factor Tu (EF-Tu)
SUB0621	prfB	peptide chain release factor 2
SUB0635	rpmE	50S ribosomal protein L31
SUB0644	rplS	50S ribosomal protein L19
SUB0678	pheS	phenylalanyl-tRNA synthetase alpha chain
SUB0680	pheT	phenylalanyl-tRNA synthetase beta chain
SUB0689	rpsU	30S ribosomal protein S21
SUB0711	infC	translation initiation factor IF-3
SUB0712	rpmI	50S ribosomal protein L35
SUB0713	rplT	50S ribosomal protein L20
SUB0715	NA	conserved hypothetical protein
SUB0725	rplU	50S ribosomal protein L21
SUB0727	rpmA	50S ribosomal protein L27
SUB0729	lspA	lipoprotein signal peptidase
SUB0741	rpsP	30S ribosomal protein S16
SUB0743	rimM	16S rRNA processing protein
SUB0763	fms	peptide deformylase
SUB0773	clpX	ATP-dependent Clp protease ATP-binding subunit ClpX
SUB0774	engB	putative GTP-binding protein
SUB0777	rplJ	50S ribosomal protein L10
SUB0778	rplL	50S ribosomal protein L7/L12
SUB0871	prfA	peptide chain release factor 1
SUB0929	NA	putative GTPase protein
SUB0956	rpsT	30S ribosomal protein S20
SUB0999	sipC	putative signal peptidase I
SUB1008	NA	putative 30S ribosomal protein S1
SUB1039	NA	polysaccharide biosynthesis protein
SUB1051	trmE	tRNA modification GTPase TrmE
SUB1080	NA	putative metal-dependent phosphohydrolase
SUB1081	msrB	peptide methionine sulfoxide reductase
SUB1082	lepA	GTP-binding protein LepA

SUB1118	NA	putative isoleucyl-tRNA synthetas
SUB1133	NA	putative GTPase
SUB1157	NA	putative GTP-binding protein
SUB1159	pepS	aminopeptidase PepS
SUB1194	NA	putative metallopeptidase
SUB1196	NA	acetyltransferase (GNAT) family protein
SUB1230	alaS	alanyl-tRNA synthetase
SUB1253	prfC	peptide chain release factor 3
SUB1296	NA	BipA family GTPase
SUB1375	NA	sigma 54 modulation protein / S30EA ribosomal protein
SUB1388	fmt	methionyl-tRNA formyltransferase
SUB1408	pepC	aminopeptidase C
SUB1466	infB	translation initiation factor IF-2
SUB1489	serS	seryl-tRNA synthetase
SUB1516	gatB	glutamyl-tRNA amidotransferase subunit B
SUB1517	gatA	glutamyl-tRNA amidotransferase subunit A
SUB1518	gatC	glutamyl-tRNA amidotransferase subunit C
SUB1528	codY	GTP-sensing transcriptional pleiotropic repressor
SUB1548	efp	elongation factor P (EF-P)
SUB1550	NA	putative metallopeptidase
SUB1555	rpsR	30S ribosomal protein S18
SUB1557	rpsF	30S ribosomal protein S6
SUB1562	NA	putative seryl-tRNA synthetase
SUB1564	spi	putative signal peptidase I
SUB1615	rpsN	30S ribosomal protein S14
SUB1631	fus	elongation factor G (EF-G)
SUB1632	rpsG	30S ribosomal protein S7
SUB1633	rpsL	30S ribosomal protein S12
SUB1641	engC	probable GTPase EngC
SUB1644	ksgA	dimethyladenosine transferase
SUB1647	NA	conserved hypothetical protein
SUB1659A	rpmH	50S ribosomal protein L34
SUB1679	gltX	glutamyl-tRNA synthetase
SUB1732A	NA	50S ribosomal protein L33
SUB1736	NA	RNA pseudouridylate synthase
SUB1741	groEL	60 kDa chaperonin
SUB1742	groES	10 kDa chaperonin
SUB1743	clpC	putative stress response-related Clp ATPase
SUB1754	rpsB	30S ribosomal protein S2
SUB1755	tsf	elongation factor Ts
SUB1771	NA	30S ribosomal protein S10
SUB1775	NA	putative Holliday junction resolvase
SUB1789	argS	arginyl-tRNA synthetase
SUB1795	rpmF	50S ribosomal protein L32
SUB1796	rpmG	50S ribosomal protein L33
SUB1841	rpsD	30S ribosomal protein S4
SUB1844	rplI	50S ribosomal protein L9
SUB1863	trsA	tryptophanyl-tRNA synthetase
SUB1867	NA	conserved hypothetical protein

Regulation and Cell Signalling

SUB0144	NA	putative Mga-like regulatory protein
SUB0145	lbp	lactoferrin binding protein
SUB0192	NA	response regulator protein

SUB0242	relA	GTP pyrophosphokinase
SUB0243	NA	D-tyrosyl-tRNA(Tyr) deacylase
SUB0327	NA	sensor histidine kinase (fragment)
SUB0429	clpP	putative ATP-dependent Clp protease proteolytic subunit
SUB0455	NA	ArsR family regulatory protein
SUB0536	ccpA	catabolite control protein A
SUB0637	NA	acetyltransferase (GNAT) family protein
SUB0826	NA	putative surface-anchored subtilase family protein
SUB0851	NA	adenylate cyclase family protein
SUB0916	NA	sensor histidine kinase
SUB0957	ciaH	sensor histidine kinase
SUB1128	apt	adenine phosphoribosyltransferase
SUB1154	scpA	C5A peptidase precursor
SUB1190	NA	cell envelope-related transcriptional attenuator domain protein
SUB1191	aroK	putative shikimate kinase
SUB1383	NA	sensor histidine kinase
SUB1477	NA	LytR family regulatory protein
SUB1599	NA	sensor histidine kinase (fragment)
SUB1674	NA	sensor histidine kinase
SUB1676	NA	sensor histidine kinase
SUB1677	NA	sensor histidine kinase (pseudogene)
SUB1794	hisS	histidyl-tRNA synthetase

Respiration (ATP Synthesis)

SUB0103	NA	pyridine nucleotide-disulphide oxidoreductase
SUB0104	cydA	cytochrome d ubiquinol oxidase subunit I
SUB0105	cydB	cytochrome d ubiquinol oxidase subunit II
SUB0393	oxaA	cytochrome oxidase biogenesis protein
SUB0449	NA	ArsC family protein
SUB0562	NA	conserved hypothetical protein
SUB0614	NA	cation transporting ATPase
SUB0666	atpE	ATP synthase C chain
SUB0667	atpB	ATP synthase A chain
SUB0668	atpF	ATP synthase B chain
SUB0669	atpH	putative ATP synthase delta chain
SUB0670	atpA	ATP synthase alpha chain
SUB0671	atpG	ATP synthase gamma chain
SUB0672	atpD	ATP synthase beta chain
SUB0673	atpC	ATP synthase epsilon chain
SUB0708	NA	putative ferredoxin
SUB0908	NA	NADPH-dependent FMN reductase
SUB1048	NA	ArsC family protein
SUB1360	NA	NADH:flavin oxidoreductase / NADH oxidase family protein
SUB1591	gcdB	putative Na+-transporting glutaconyl-CoA decarboxylase beta subunit
SUB1592	NA	putative decarboxylase gamma chain

RNA Metabolism

SUB0013	NA	PP-loop family protein
SUB0094	rpoA	DNA-directed RNA polymerase alpha chain
SUB0110	acdR	putative transcriptional repressor
SUB0115	rpoB	DNA-directed RNA polymerase beta chain

SUB0116	rpoC	DNA-directed RNA polymerase beta' chain
SUB0134	NA	putative tRNA-binding protein
SUB0142	NA	putative tRNA-dihydrouridine synthase
SUB0177	NA	RNA polymerase sigma factor protein
SUB0194	NA	transcriptional regulator (antiterminator)
SUB0255	proS	prolyl-tRNA synthetase
SUB0317	truA	tRNA pseudouridine synthase A
SUB0325	rpoE	putative DNA-directed RNA polymerase, delta subunit
SUB0355	NA	RNA-binding protein
SUB0392	greA	transcription elongation factor
SUB0415	NA	putative rRNA methylase
SUB0459	metG	methionyl-tRNA synthetase
SUB0477	NA	putative cyclophilin type peptidyl-prolyl cis-trans isomerase
SUB0492	era	GTP-binding protein Era homolog
SUB0523	rnr	putative exoribonuclease R
SUB0530	NA	transcriptional regulator (antiterminator)
SUB0555	rnc	ribonuclease III
SUB0575	NA	SprT-like protein
SUB0592	lysS	lysyl-tRNA synthetase
SUB0620	NA	putative iron-sulphur protein
SUB0629	asnS	asparaginyl-tRNA synthetase
SUB0636	NA	conserved hypothetical protein
SUB0692	rpoD	RNA polymerase sigma factor RpoD
SUB0721	NA	putative cysteine desulfhydrase
SUB0722	thiI	putative thiamine biosynthesis protein
SUB0724	NA	putative inosine-uridine preferring nucleoside hydrolase
SUB0726	NA	ribosome-associated protein
SUB0730	NA	putative RNA pseudouridylate synthase
SUB0744	trmD	tRNA (Guanine-N(1)-)-methyltransferase
SUB0757	cca	CCA-adding enzyme
SUB0797	lacT2	transcription antiterminator LacT 2
SUB0828	NA	transcription antiterminator (pseudogene)
SUB0849	NA	putative cysteine desulfurase
SUB0854	NA	putative RNA pseudouridylate synthase
SUB0868	apbE	putative thiamine biosynthesis protein
SUB0872	NA	putative methyltransferase
SUB0873	NA	conserved hypothetical protein
SUB0926	NA	SMF family protein
SUB0928	rnh	ribonuclease HII
SUB0970	truB	tRNA pseudouridine synthase B
SUB1131	rnz	ribonuclease Z
SUB1134	miaA	tRNA delta(2)-isopentenylpyrophosphate transferase
SUB1189	NA	putative RNA methyltransferase
SUB1193	NA	ribonuclease BN-like family protein
SUB1218	NA	putative RNA helicase
SUB1240	queA	S-adenosylmethionine:tRNA ribosyltransferase-isomerase
SUB1250	NA	DEAD box helicase family protein
SUB1251	NA	conserved hypothetical protein
SUB1284	ileS	isoleucyl-tRNA synthetase
SUB1286	NA	putative RNA-binding protein
SUB1315	NA	putative rRNA methyltransferase
SUB1342	valS	valyl-tRNA synthetase
SUB1371	NA	putative RNA methyltransferase
SUB1380	NA	putative S1 RNA binding domain protein
SUB1381	NA	cyclophilin type peptidyl-prolyl cis-trans isomerase protein
SUB1390	rpoZ	DNA-directed RNA polymerase omega chain
SUB1403	NA	putative RNA methylase family protein

SUB1417	NA	putative helicase
SUB1435	glyS	glycyl-tRNA synthetase beta chain
SUB1436	glyQ	glycyl-tRNA synthetase alpha chain
SUB1452	lacT1	transcription antiterminator LacT 1
SUB1465	rbfA	ribosome-binding factor A
SUB1467	NA	ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
SUB1468	NA	conserved hypothetical protein
SUB1469	nusA	putative N utilization substance protein A
SUB1470	ylxS	putative ribosomal maturation protein
SUB1471	NA	tRNA (guanine-N(7)-)-methyltransferase
SUB1479	NA	putative P-loop hydrolase
SUB1525	NA	conserved hypothetical protein
SUB1546	nusB	N utilization substance protein B homolog (NusB protein)
SUB1563	NA	ribonuclease HIII
SUB1616	NA	putative glycoprotease
SUB1617	NA	acetyltransferase (GNAT) family protein
SUB1618	NA	glycoprotease family protein
SUB1620	NA	ribonuclease J
SUB1667	rnpA	ribonuclease P protein component
SUB1704	NA	transcription antiterminator
SUB1707	NA	putative deaminase
SUB1711	tgt	queuine tRNA-ribosyltransferase
SUB1729	leuS	leucyl-tRNA synthetase
SUB1731	nusG	transcription antitermination factor
SUB1744	ctsR	transcriptional regulator
SUB1765	NA	transcriptional regulator protein
SUB1774	yrzB	putative anti-sigma factor
SUB1809	NA	putative transcription regulator
SUB1847	NA	putative tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase

Stress Response

SUB0009	NA	S4 domain containing protein
SUB0146	gshA	putative glutamate--cysteine ligase
SUB0234	NA	conserved hypothetical protein
SUB0254	eep	putative pheromone-processing membrane metalloprotease
SUB0332	NA	Putative alkaline shock protein
SUB0402	NA	putative phosphoesterase
SUB0409	NA	SpoU rRNA methylase family protein
SUB0421	fhuG	ferrichrome transport permease
SUB0422	fhuB	ferrichrome transport permease
SUB0424	fhuC	ferrichrome transport ATP-binding protein
SUB0440	NA	conserved hypothetical protein
SUB0441	NA	initiation-control protein YabA
SUB0442	NA	tetrapyrrole (corrin/porphyrin) methylase family protein
SUB0548	NA	putative amino acid ABC transporter, permease protein
SUB0549	NA	putative amino acid ABC transporter, permease protein
SUB0550	NA	putative glutamine-binding protein precursor
SUB0551	glnQ4	putative glutamine transporter, ATP-binding protein 4
SUB0640	NA	putative glutathione S-transferase
SUB0656	NA	cold shock protein
SUB0706	pepT	peptidase T
SUB0719	gor	glutathione reductase
SUB0845	rex	putative redox-sensing transcriptional repressor
SUB0862	NA	ABC-type glycine betaine transport system ATP-binding protein

SUB0863	NA	ABC-type glycine betaine transport system protein
SUB1088	hemN	putative oxygen-independent coproporphyrinogen III oxidase
SUB1094	NA	putative glutamine amidotransferase
SUB1152	NA	putative glutamine ABC transporter, glutamine-binding protein protein
SUB1153	glnQ1	putative glutamine ABC transporter, ATP-binding protein 1
SUB1242	sodA	superoxide dismutase [Mn]
SUB1247	NA	conserved hypothetical protein
SUB1278	glnQ3	putative glutamine transporter, ATP-binding protein 3
SUB1503	dnaJ	chaperone protein DnaJ
SUB1504	dnaK	chaperone protein DnaK (heat shock protein 70)
SUB1505	grpE	GrpE protein (HSP-70 cofactor)
SUB1506	hrcA	heat-inducible transcription repressor HrcA
SUB1513	cspC	cold shock protein CspC
SUB1530	NA	putative universal stress protein
SUB1547	NA	putative alkaline shock protein
SUB1714	NA	ferric uptake regulator family protein
SUB1717	NA	putative glycine-betaine binding permease protein
SUB1718	NA	putative glycine betaine transport ATP-binding protein
SUB1745	csp	cold shock protein

Sulphur Metabolism

SUB0977	tpx	putative thiol peroxidase
SUB1654	NA	conserved hypothetical protein
SUB1752	ahpC	alkyl hydroperoxide reductase subunit C
SUB1753	ndh	NADH dehydrogenase

Virulence, Disease and Defence

SUB0032	ublA	putative bacteriocin
SUB0143	NA	33 kDa chaperonin (heat shock protein 33 homolog)
SUB0169	NA	putative membrane protein
SUB0408	NA	conserved hypothetical protein
SUB0444	NA	CutC family protein
SUB0450	NA	cation efflux family protein
SUB0502	NA	putative bacteriocin (pseudogene)
SUB0505	NA	putative bacteriocin
SUB0506	NA	putative bacteriocin
SUB0509	NA	putative bacteriocin
SUB0510	NA	putative bacteriocin immunity protein
SUB0512	pedA	putative bacteriocin
SUB0513	NA	putative bacteriocin immunity protein
SUB0516	NA	putative bacteriocin immunity protein
SUB0521	pmrA	multi-drug resistance efflux pump
SUB0989	NA	cation efflux family protein
SUB1097	NA	AraC family regulatory protein
SUB1180	copZ	putive copper chaperone CopZ
SUB1181	NA	putative copper importing ATPase
SUB1212	NA	putative lipoprotein
SUB1266	NA	putative cation-transporting ATPase
SUB1350	NA	MATE family efflux protein
SUB1358	NA	DedA family protein
SUB1462	copZ	putative copper chaperone

SUB1463	copA	copper-transporting ATPase
SUB1464	copY	putative negative regulator of copper transport operon
SUB1472	NA	putative aminoglycoside phosphotransferase
SUB1515	NA	putative membrane protein
SUB1561	NA	putative membrane protein
SUB1665	NA	putative ssDNA-binding protein
SUB1666	NA	membrane protein OxaA 1 precursor
SUB1695	NA	MATE family efflux protein
SUB1785	NA	PauA protein precursor (streptokinase precursor)
SUB1826	NA	putative autolysin regulatory protein

Appendix 2