# Bioinformatic analysis of *Streptococcus uberis* genes and genomes

MUHAMMAD MAQSUD HOSSAIN




School of Veterinary Medicine and Science
UNIVERSITY OF NOTTINGHAM

A dissertation submitted to
the **University of Nottingham**
in accordance with the requirements of the degree of
DOCTOR OF PHILOSOPHY in the Faculty of Health Sciences.

SEPTEMBER 2015

# ABSTRACT

*Streptococcus uberis* is a Gram-positive, catalase-negative member of the family Streptococcaceae and is an important environmental pathogen primarily responsible for a significant amount of bovine intramammary infections. This thesis describes the sequencing and comparison of multiple strains from clinical and sub-clinical infections. Following *de novo* assembly, these are compared to the single reference strain (0140J).

The assemblies of strains sequenced with two technologies (Illumina and SOLiD) were compared. From these assemblies, annotation allowed the comparison of gene content, the pan and core genomes and gene gain/loss of coding sequences associated with clustered regularly interspaced short palindromic repeats (CRISPRs), prophage and bacteriocin production. Identification of sequence variants allowed identification of highly conserved and highly variant genes.

Inferred intraspecies and interspecies (host-*S. uberis*) protein-protein interaction networks revealed pathways of bovine proteins enriched with potentially interacting pathogen proteins. These identified known and predicted pathways and also novel interaction partners.

This was the first "whole-genome" comparison of multiple *S. uberis* strains isolated from clinical vs non-clinical intramammary infections including the type virulent vs non-virulent strains. These data allowed the first insight into potential evolutionary forces behind virulence differences.

# List of publications

M. HOSSAIN, S. A. EGAN, T. COFFEY, P. N. WARD, R. WILSON, J. A. LEIGH, AND R. D. EMES, *Virulence related sequences; insights provided by comparative genomics of Streptococcus uberis of differing virulence*, BMC genomics, 16 (2015), p. 1.

## ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor, Professor Richard Emes for introducing me to microbial pathogenomics and providing me admirable support and guidance over the past four years. He extended all possible academic and non academic help to ensure safety to all my endeavors. I am grateful to him from the bottom of my heart. It is difficult to express in words how Prof. Emes cared about my life and difficulties with great priority. I express my sincere gratitude to my second supervisor Professor James Leigh for his expert guidance, teachings and affection. Both of them have always been greatly approachable and accessible from the beginning till the end of my doctoral research. Special thanks goes to Dr Sharon Egan who helped me to conduct wet lab experiments and provided me with useful experimental data to validate my computational predictions. I would like to acknowledge all those who contributed to the studies presented here: Tracey Coffey, Philip Ward, Ray Wilson and Mike Jones. Special thanks are due to Tom Giles, Frank, Andrew Warry , Jo Moreton and everyone from the Advanced Data Analysis Centre (ADAC) for their help with the server and helpful bioinformatics support when needed. Nick Loman (University of Birmingham) helped many times with some useful scripts and constructive suggestions and helped networking with some great Bioinformaticians by arranging the 'Balti and Bioinformatics' conference. Society of General Microbiology and UK NGS conference arranged by Deep Seq, QMC helped me with the conference and travel funds to present my research works. Dr George Russell (Moredun Research Institute, Edinburgh, UK)

made tremendous contribution in many aspects of my career and personal life. I am indebted to Helena Percival , Susan, Ali and other officials at the School of Veterinary Medicine and Science for helping me on official needs. I cannot forget my best friend Sujith Sebastian who has always been there for me whenever I used to stress out and needed a break from my research and thanks to Sudhakar for accompanying us most of the time and helping with his expert advice in the coffee area. I am privileged to have some wonderful moments with my colleagues and friends who supported me near or off my desk in the corner in one way or the other and with whom I have shared many memorable moments during my time in B31, B32, SB, and Nottingham: Adeelah, Katy, Adam, Mansi, , Pengxiang, James, Jasmine, Donna, Ibrahim, Basim, Osama, Muhammad, Azad and many others. Special thanks to Dr Debarka Sengupta for helping me during the final compilation of the thesis. I am beholden to my family members , my three sisters Sumi, Sonia and Tania who stood by me through thick and thin I feel the blessings of my father (Late MD Anowar Hossain), who taught me to be an independent thinker and urged me to become a good human being with high ambition. My mother (Late Ambia Hossain) always believed in my abilities and reminded me like a constant trainer and it's only her faith and sacrifices that could bring this success to my life. I owe a lot to my wife Fariza Shams for all her effortless support during this four year journey. Thanks to in laws for constant nagging about finishing my PhD. Thanks to Nahwan, my son for being the greatest inspiration to finish my thesis within the quickest possible time. Finally, I would like to dedicate this thesis to my beloved parents, wife and son for their uninterrupted support, encouragement and affection.

# Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

# TABLE OF CONTENTS

**Page**

**List of Tables** **xi**

**List of Figures** **xiii**

**Abbreviations** **xvii**

**1 Introduction** **1**

1.1 Bovine mastitis . . . . . 1

1.2 Mastitis-causing bacteria . . . . . 2

1.3 *Streptococcus uberis* . . . . . 4

1.4 Host response to mastitis . . . . . 6

1.5 Virulence factors of *Streptococcus uberis* . . . . . 10

1.6 Strain-typing of *S. uberis* . . . . . 13

1.7 *Streptococcus uberis* genome overview . . . . . 16

1.8 Next-generation sequencing technologies . . . . . 19

1.9 Aims and objectives . . . . . 23

**2 Materials and Methods** **25**

2.1 Bacterial isolates and genome sequencing . . . . . 25

2.2 *De novo* assembly of *Streptococcus uberis* genomes from SOLiD short-read data sets . . . . . 26

2.2.1 SOLiD Accuracy Enhancement Tool . . . . . 27

2.2.2 Preprocessor . . . . . 28

2.2.3 Choice of short read assemblers . . . . . 28

2.3 *De novo* assembly of *Streptococcus uberis* genomes from Illumina MiSeq reads . . . . . 31

2.3.1 Quality control and *de novo* assembly . . . . . 31

2.3.2 Database submission of Illumina assembly . . . . . 31

2.4 Assembly evaluation . . . . . 32

2.5 Genome annotation . . . . . 33

2.6 Pan and core-genome analysis . . . . . 34

2.7 Multi locus sequence typing . . . . . 34

2.8 Read mapping and variant calling . . . . . 35

2.9 Whole genome phylogeny . . . . . 36

2.10 Heatmap clustering and statistical analysis . . . . . 36

2.11 Analysis of genes undergoing positive selection . . . . . 37

2.12 Network visualization and analysis . . . . . 37

2.13 Prediction of *Streptococcus uberis* protein-protein interactions . . . . . 38

2.14 Homology based host-pathogen protein-protein interaction prediction . . . 38

2.15 Scripts . . . . . 39

2.16 Additional files . . . . . 39

**3 *De novo* assembly and comparative analysis of SOLiD and Illumina sequencing datasets 41**

3.1 Introduction . . . . . 41

3.1.1 Technical aspects of NGS data . . . . . 43

3.2 Results . . . . . 48

3.2.1 Genome sequencing and assembly using SOLiD . . . . . 48

3.2.2 Genome sequencing and assembly using Illumina MiSeq . . . . . 54

3.2.3 Multi locus sequence typing . . . . . 63

3.3 Discussion . . . . . 66

**4 Comparative genomics of *Streptococcus uberis* isolates of differing virulence 69**

4.1 Introduction . . . . . 69

4.2 Results . . . . . 71

4.2.1 General features of *Streptococcus uberis* genomes . . . . . 71

4.2.2 Pan and core-genome analysis . . . . . 84

4.2.3 Comparison of virulence factors . . . . . 88

4.2.4 Analysis of the has operon . . . . . 90

4.2.5 Analysis of CRISPR-Cas proteins . . . . . 91

4.2.6 Analysis of prophage regions . . . . . 94

4.2.7 Analysis of bacteriocin production . . . . . 96

4.3 Discussion . . . . . 97

**5 Genomic variation analysis of *Streptococcus uberis* 102**

5.1 Background . . . . . 102

5.2 Results . . . . . 105

5.2.1 Whole-genome sequencing to identify the relationship among *Streptococcus uberis* isolates . . . . . 105

5.2.2 Gene-wide variation analysis . . . . . 105

5.2.3 Variation of virulence genes . . . . . 124

5.2.4 Genes under positive selection . . . . . 129

5.3 Discussion . . . . . 133

**6 Computational analysis of inter- and intraspecies protein-protein interaction networks 138**

6.1 Introduction . . . . . 138

6.2 Results and Discussion . . . . . 140

6.2.1 *Streptococcus uberis* protein-protein interaction . . . . . 140

6.2.2 *Streptococcus uberis*-bovine protein-protein interaction . . . . . 146

6.3 Conclusion . . . . . 161

**7 Conclusions 163**

7.1 Summary . . . . . 163

7.2 Outlook . . . . . 166

7.2.1 Gap closure . . . . . 166

7.2.2 Identification of virulence genes . . . . . 167

7.2.3 Analysis of host-pathogen interaction. . . . . 170

**A Appendix A** **172**

A.1 Colorspace reads . . . . . 172

A.2 Command lines . . . . . 175

**Bibliography** **176**

# List of Tables


**Table** **Page**

1.1 Prevalence of mastitis pathogens (clinical mastitis) in dairy herds in UK, Quarter . . . . . 4

2.1 Isolates used in the study for whole genome sequencing. All twenty two isolates were used for SOLiD sequencing and the isolates with an asterisk were chosen for MiSeq Illumina sequencing . . . . . 26

2.2 Major scripts used in the study are available at https://github.com/mdmaqsud/Cov.R 40

3.1 Q-scores, error and accuracy. . . . . 45

3.2 Quality scores of Illumina reads and corresponding ASCII codes . . . . . 47

3.3 Effect of K-mer on N50 values of twenty isolates using SOLiD reads . . . . . 51

3.4 Optimum coverage distribution for different isolates used during Velvet assembly . . . . . 53

3.5 *De novo* assembly statistics of *S. uberis* genomes using SOLiD reads. . . . . 61

3.6 *De novo* assembly statistics of *S. uberis* genomes using Illumina MiSeq reads 62

4.1 Genes present in 0140J and absent from EF20 . . . . . 72

4.2 Genes present in EF20 and absent from 0140J . . . . . 76

4.3 Distribution of prophage regions among 13 isolates . . . . . 96

5.1 SNP differences among different isolates using Illumina reads . . . . . 108

5.2 List of potential genes affecting the grouping of strains into two major categories 123

5.3 SNP distribution of (a) known virulence genes (b) sortase-anchored proteins (c) has genes. *denotes genes previously associated with virulence which also contain a predicted sortase-anchor. . . . . 125

5.4 Genes under positive selection . . . . . 132

6.1 Canonical pathways interacting with EF20 and 0140J genes predicted by IPA 148

A.1 SOLiD error filtering . . . . . 173

A.2 Sortase prediction by CW-PRED . . . . . 174

# List of Figures

**Figure** **Page**


1.1 Bacterial isolation, somatic cell count and clinical response following challenge with *S. uberis* 0140J and EF20 in dairy cattle. . . . . 7

1.2 BURST analysis of UK isolates . . . . . 15

1.3 Frequency of different clonal complexes in New Zealand and UK. . . . . 15

1.4 Bayesian phylogeny based on combined 16srRNA and *rnpB*. . . . . 18

1.5 Template generation via bridge amplification . . . . . 21

2.1 Schematic diagram of *de novo* assembly pipelines for SOLiD reads . . . . . 27

3.1 Di-nucleotide to colorspace encoding . . . . . 44

3.2 Q-score as the function of base calling probability . . . . . 45

3.3 Example of a HiSeq FASTQ header. . . . . 46

3.4 Plot showing the relationship between position in sequence read and quality value (isolate 0140J) . . . . . 50

3.5 Coverage distribution of filtered SOLiD short reads of the genome of 0140J. . 52

3.6 Per base quality value of Illumina reads . . . . . 55

3.7 The per sequence quality score. This shows the subset of EF20 sequences have universally high quality values with most of the reads with QV > 30. . . 56
3.8 Scatterplot showing the distribution N50 as a function of different K-mers on MiSeq reads. . . . . 56
3.9 Coverage distribution of illumina reads of the genome EF20. . . . . 57
3.10 Distribution of cumulative lengths of assembled genomes . . . . . 59
3.11 Comparison of N50 distribution of SOLiD and Illumia reads . . . . . 60
3.12 Phylogenetic tree based on concatenated MLST genes determined by PhyML 64
3.13 MLST distribution of 13 isolates compared to the other UK strains available in the PubMLST database. . . . . 65

4.1 Relationship between gene count and genome size . . . . . 86
4.2 Distribution of accessory genes not in 0140J. . . . . 87
4.3 Pan-genome clustering . . . . . 88
4.4 Comparison of has operon between 0140J and B362 . . . . . 91
4.5 Comparative analyses of Type II CRISPRs regions . . . . . 93
4.6 Comparative analyses of Type III CRISPRs regions in B362 and 0140J . . . . 94
4.7 Comparison of the CDS of the intact prophage . . . . . 95

5.1 Whole genome phylogeny of *S. uberis* isolates using SOLiD reads . . . . . 106
5.2 Whole genome phylogeny of *S. uberis* isolates using Illumina reads. . . . . 107
5.3 Cumulative frequency plot showing differnce of mean SNP/bp of core and accessory genes . . . . . 108

5.4 Violin plot showing the distribution of mean SNP/bp between accessory and core genes. . . . . 109

5.5 Cumulative frequency plot of SNP/bp of each genome as a function of genome position. . . . . 110

5.6 Distribution of different functional groups based on (a) GO molecular function (b) GO Cellular location . . . . . 111

5.7 Violin plot showing the distribution of mean SNP/bp in core genomes . . . . . 113

5.8 Heatmap clustering of shared SNPs within genes annotated as GO molecular function . . . . . 116

5.9 SNP density (SNP/bp) of individual genes associated with enriched GO cellular location categories. Heatmap visualization of clustering of genes with similar SNP density between *S. uberis* strains. Colours show SNP density, yellow = higher mean SNP/bp. (a) cell wall (b) cytoplasm (c) extracellular (d) intracellular (e) membrane (f) ribosome. . . . . 122

5.10 Heatmap showing the distribution and clustering of SNP/bp of putative virulence genes among different isolates . . . . . 126

5.11 Variation in the *vru* upstream regions. . . . . 128

5.12 GO Cellular location of genes under positive selection . . . . . 130

5.13 Cumulative frequency (dn/ds) of (a) different functional groups of genes under positive selection . . . . . 131

6.1 Visualtization of PPIN data in cytoscape . . . . . 142

6.2 Simple statistical network parameters . . . . . 143

6.3 Betweenness centrality . . . . . 144

6.4 PPI network of *vru* . . . . . 145

6.5 IL-10 signaling pathway in *S uberis*. . . . . 154

6.6 IL-6 signaling pathway (a) 0140J (b) EF20 . . . . . 156

6.7 Comparative analysis of 0140J and EF20 genes interacting with IL-6 signalling pathway. . . . . 157

# ABBREVIATIONS

| | |
|---|---|
| SCC | Somatic cell counts |
| MEC | Mammary epithelial cells |
| TLR | Toll-like receptors |
| PRR | Pathogen recognition receptors |
| LTA | Lipoteichoic acid |
| PMN | Polymorphonuclear neutrophils |
| TNF$\alpha$ | Tumor necrosis factor alpha |
| DNA | Deoxyribonucleic acid |
| IMI | Intramammary infection |
| MLST | Multi locus sequence typing |
| PFGE | Pulsed field gel electrophoresis |
| MLEE | Multi locus enzyme electrophoresis |
| ST | Sequence type |
| MGE | Mobile genetic elements |
| MP | Maximum parsimony |
| ME | Minimum evolution |

| | |
|---|---|
| WGS | Whole genome sequencing |
| NGS | Next-generation sequencing |
| SBS | Sequencing by synthesis |
| SBL | Sequencing by ligation |
| PPI | Protein-protein interaction |
| SAET | SOLiD accuracy enhancement tool |
| HTQC | High throughput quality control |
| CRISPR | Clustered regularly interspaced short palindromic sequence |
| ACT | Artemis comparison tool |
| CDS | Coding sequence |
| PGAP | Pan genome analysis pipeline |
| SNP | Single nucleotide polymorphisms |
| VCF | Variant calling format |
| Indel | Insertion/deletion |
| CSFASTA | Colorspace fasta |
| QV | Quality value |
| PTS | Phosphotransferase system |
| RATT | Rapid annotation transfer tool |
| RAST | Rapid annotation server |
| RNA | Ribonucleic acid |
| HMM | Hidden markov model |
| My | Million year |

| | |
|---|---|
| PCR | Polymerase chain reaction |
| GWAS | Genome-wide association study |
| vs | versus |
| bp | base pair |
| Mbp | Mega base pair |
| Gb | Giga base |
| ATP | Adenosie tri-phosphate |
| GTP | Guanosine tri-phosphate |
| GO | Gene ontology |
| Y2H | Yeast two-hybrid |
| MS | Mass spectrometry |
| TAP | Tandem affinity purification |
| IL | Interleukin |
| APRIL | A proliferation-inducing ligand |
| BCMA | B-cell maturation antigen |
| MAPK | Mitogen activated protein kinase |
| PI3K/Akt | Phosphatidylinositol 3-kinase |
| PAK | p21-activated kinase |

CHAPTER 1

# INTRODUCTION

## 1.1 Bovine mastitis

The bovine mammary gland is typically a sterile environment and entry of any foreign body invokes a localized immune response leading to infection of the gland which is commonly known as mastitis [20]. Mastitis is mostly caused by bacteria, with more than 100 microbial species isolated from the mammary gland of the cow [189] and is one of the most common infectious diseases in dairy cattle. The economic impact of both sub-clinical and clinical mastitis in the UK dairy industry is in excess of £200 million/annum with worldwide economic loss estimated at US$35 billion [22].

Mastitis can be classified into two major categories: clinical and sub-clinical mastitis. However, only a small subset of species is responsible for most cases of clinical mastitis. Clinical mastitis is a more severe form and is characterized by abnormal milk, protein

aggregates and clotting, swelling of the udder sometimes resulting in elevated rectal temperature, lethargic condition or anorexia. In severe cases it can be fatal due to bacteraemia or septicaemia [105].

It is difficult for farmers to recognize sub-clinical mastitis as there is no obvious clinical signs and this can cause significant reduction in milk production. In sub-clinical mastitis there is an influx of somatic cells to the mammary gland primarily polymorphonuclear neutrophils (PMNs) [117] which release reactive oxygen intermediates and proteolytic enzymes and can damage mastitis mammary tissues [230].

In milk from uninfected glands, leukocyte counts, including macrophages, neutrophils and lymphocytes are typically below 150,000 cells/mL. Subclinically infected quarters usually have a cell count in excess of 250 000 cells/mL, but this figure may vary in clinical mastitis [48, 91]. Somatic cell count (SCC) is often used as an indirect estimation and there has been a fall in the SCC of the bulk milk supply in the UK from around 500,000 cells/mL in 1971 to an estimated 180 000 cells/mL in 1996 [19], suggesting that sub-clinical infection has also been reduced by the improvements to dairy hygiene and husbandry.

## 1.2 Mastitis-causing bacteria

Mastitis can be categorized as 'contagious' and 'environmental' based on the mode of transmission of the bacterial pathogens. The major contagious pathogens *Staphylococcus aureus* and *Streptococcus agalactiae* can survive within the mammary gland and can

be transmitted from cow-to-cow or during the time of milking [20, 202]. During the 1960s the implementation of five point control measures including the implementation of antibiotic dry cow therapy, post-milking teat disinfection, and routine maintenance of milking machines, reduced the incidence of mastitis caused by these bacteria [151]. However, whilst such control measures were effective in reducing the incidences associated with the contagious pathogens, they were not as effective against environmental pathogens [179]. Environmental pathogens are considered to be opportunistic and they are thought to be able to adapt to survive within the mammary gland [20]. It is also hypothesized that a particular niche can be occupied by environmental pathogens due to the reduction of contagious pathogens [95]. *Escherichia coli*, *Klebsiella pneumoniae*, *Streptococcus dysgalactiae* and *Streptococcus uberis* [159, 207] are the major groups of environmental pathogens and among them *S. uberis* has been reported [78]. Teat end contamination is one of the major predisposing factors in the development of environmental mastitis and this can be influenced by housing materials as environmental pathogens are able to survive and multiply in organic bedding materials. Since the 1980s, environmental pathogens such as *S. uberis* and *E. coli* are known to be most prevalent mastitis causing pathogens in UK [20] and different epidemiological studies have showed that the prevalence of mastitis causing pathogens are changing (Table 1.1). Analysis of 2,257 samples in a study from clinical mastitis in UK showed that the most prevalent bacteria was *S. uberis* (37%) followed by enterobacteriaceae (23%) and coagulase negative *Staphylococcus* spp. (10%) [142]. Another study carried out in England and Wales found the prevalence of *S. uberis* and *E. coli* were 23.5% and 19.8%, respectively in clinical

mastitis cases. Reports from other countries including US, Canada, Australia and the Netherlands showed that in clinical mastitis the prevalence of *S. uberis* ranged between 14 to 26% of [95, 104]. The dynamics of prevalence of mastitis pathogens suggests that *S. uberis* is emerging as a major mastitis pathogen worldwide.

Table 1.1: Prevalence of mastitis pathogens (clinical mastitis) in dairy herds in UK, Quarter: cases per 100 cases per year [20]

| Pathogens | 1967 | 1982 | 1998 |
|---|---|---|---|
| *S. aureus* | 67(43.8%) | 7(17.5%) | 2.2(5.3%) |
| *S. agalactiae* | 6(4%) | 1(2.5%) | |
| *S. dysgalactiae* | 16(10.5%) | 4(10%) | 2.0(4.8%) |
| *S. uberis* | 7(4.6%) | 9(22.5%) | 5.3(12.7%) |
| *E. coli* | 7(4.6%) | 10(25%) | 14.4(34.6%) |
| Others | 50(32.7%) | 9(22.5%) | 17.7(42.6%) |
| | 153(100%) | 40(100%) | 41.6(100%) |

## 1.3 *Streptococcus uberis*

*S. uberis* is a Gram-positive, catalase-negative cocci and key biochemical diagnostic characteristics include hydrolysis of hippurate, esculin [218] and acid production from inulin [139]. *S. uberis* has been isolated from multiple areas of the the body of cows including the skin, genital tract, rumen, tonsils, lips, udder, teat skin and faeces [43, 43, 226]. The infection caused by *S. uberis* can remain sub-clinical during long periods of time and, in the absence of treatment, this may cause a reduction in milk production and alter milk composition. These changes in turn adversely affect the quality and quantity of dairy products [104]. As a limited number of pathogens have been found to be associated with bovine mastitis, a significant achievement to control the incidence of disease may

be made by procedures applied against an individual bacterial species such as *S. uberis*. Vaccination might be the ultimate integral and essential part of more comprehensive mastitis control plan however, the development of such vaccines is hampered by a lack of information on the interaction between pathogen and host that form the basis of infection and disease [119].

Different strains of *S. uberis* have been isolated and are predicted to cause differing types of infection. Very few direct comparisons of virulence have been conducted. Prof Leigh University of Nottingham provided access to data relating to one such experiment. The representative avirulent strain EF20 was compared in an experimental infection model, to the clinical virulent 0140J strain. All challenged quarters became infected and shed bacteria at detectable levels from the first milking (12 hours post challenge). Those challenged with strain 0140J shed bacteria at $10^4$ cfu/ml of milk 12h post-challenge) rising to $10^6$-$10^7$ cfu/ml of milk by 48h post challenge (Fig 1.1 a). In contrast, those challenged with strain EF20 shed considerably fewer bacteria $\leq 10^3$ cfu/ml and typically declined to around $10^2$ cfu/ml of milk by 48h post challenge. The speed of cellular infiltration into the mammary gland in response to infection with either strain was similar, however, the magnitude of the infiltration was 10 fold less following challenged with strain EF20. The somatic cell count detected following challenge with strain 0140J was similar to those reported previously for this strain [61, 120, 188] (Fig 1.1 b). The levels of cellular infiltration and bacterial colonization for each challenged quarter showed a significant positive correlation over the time-course of the experiment ($R^2$ = 0.404, P <0.001). Acute clinical signs of mastitis (change in milk composition, swollen

and inflamed udder quarter) occurred in all animals challenged with strain 0140J (Fig 1.1 c). In contrast, those challenged with strain EF20 showed few, if any, changes in milk composition and/or quarter appearance, thus substantiating and adding detail to the previous data [88] regarding the nature of virulence of these two strains and confirming their suitability for direct comparison in the elucidation of virulence related features. Vaccination might be the ultimate integral and essential part of more comprehensive mastitis control plan however, the development of such vaccines is hampered by a lack of information on the interaction between pathogen and host that form the basis of infection and disease [119].

## 1.4 Host response to mastitis

In comparison to the host response to mastitis caused by *Staphylococcus aureus* and *E. coli*, *S. uberis* mastitis is not well understood [179, 220]. A number of resident, recruited, and inducible immune factors are involved in the immune response associated with mastitis [171]. During the early stages of infection, innate immune response is induced rapidly at the site of infection [191] and mastitis causing Gram-positive and Gram-negative organisms exert a distinctive innate immune response in the mammary gland [10].

Delayed onset of inflammation is a characteristic feature of *S. uberis* infection [9, 120]. Failure of *S. uberis* to induce pro-inflammatory cytokine and chemokine-encoding gene exprssion, is revealed by comparison of global transcriptome profiling of *S. uberis* and *E.*

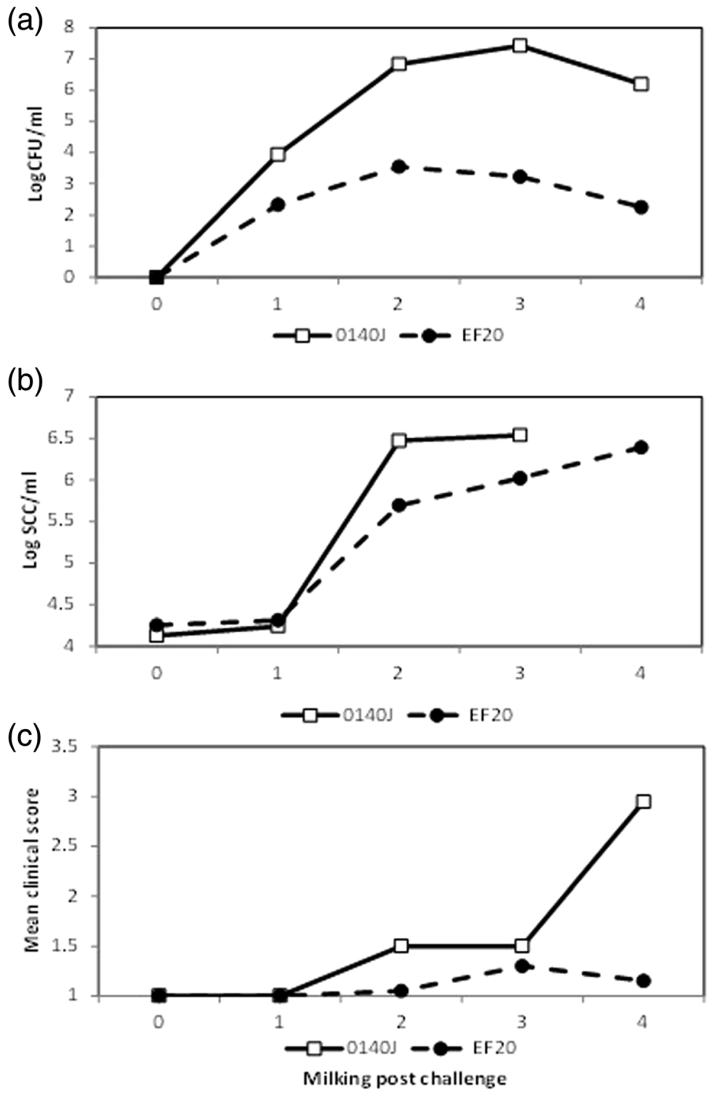



Figure 1.1: Bacterial isolation, somatic cell count and clinical response following challenge with *S. uberis* 0140J and EF20 in dairy cattle. Comparison of challenge of animals with either *S. uberis* strain 0140J (n = 10) or the EF20 (n = 10). (a) Bacterial recovery of *S. uberis* at each milking following challenge, measured by cfu/ml of milk (b) Cellular influx at each milking following *S. uberis* challenge, measured by somatic cell count/ml of milk (c) Combined clinical scores from clinical manifestations following *S. uberis* challenge. Data is represented as the arithmetic mean of clinical scores given for the appearance of the quarter and appearance of the milk as previously described [120].

*coli* infection trials in the udder [145, 149, 162, 196]. IL-10 and IL-6 directed inflammation reducing pathways, were upregulated by *S. uberis* infection [149].

The udder cells are predominantly mammary epithelial cells (MEC) which come into contact with invading pathogens soon after infection [29, 30]. Toll-like receptors (TLRs) are among the pathogen recognition receptors (PRR) used by the epithelial cells to recognise pathogens [199]. There are thirteen described mammalian TLRs among which TLR2 provides efficient immune defence against Gram-positive bacteria [50, 55]. Lipoproteins of these organisms are the natural ligands of TLR2 [97, 209]. Previously Farhat et al. reported inactivation of bovine TLR2 receptor in HEK293 cells by mitomycin-C inactivated *S. uberis* preparations [59]. *S. uberis* strains including virulent strain 0140J and avirulent strain, EF20, caused insignificant activation of immune genes TNF, CXCL8, IL-1A, IL-6, CCL5, SAA3, $\beta$-defensin and LAP in pbMEC [82].

Together these results provide insights into the slow and weak immune response against *S. uberis*. Although induction of a strong immune gene response and activation of NF-$\kappa$B has been attributed to lipoteichoic acid (LTA) from two strains of *S. uberis* (233 and T1-18), the arrangement of LTA in the cell envelope prevents the recognition of *S. uberis* by the pbMEC [82]. Due to inadequate immune response generated by MEC, *S. uberis* isolated from both clinical and sub-clinical mastitis, escape immune surveillance resulting in weak and delayed inflammatory response during *S. uberis* mastitis [82].

Being the major first line defense, macrophages phagocytise pathogens in the lactating mammary gland [171]. As demonstrated by challenge with *S. uberis* strain O140J, bacteria were observed in the cytoplasm of macrophages indicating a possible of role of

macrophage in clearing the mastitis causing streptococci. It should also be noted that in vitro experiments were unable to show actual killing of *S. uberis* [45, 206]. In a recent report Tassi et al. demostrated killing of non-virulent *S. uberis* strain FSL Z1-124 by bovine macrophages at the same time the inability of the macrophages to kill a virulent strain of *S. uberis* FSLZ1-048 [201].

Surprisingly, heat-killed particles from different *S. uberis* strains strongly induced CXCL2 ,TNF, IL-6 and CCL5 genes in parallel to significant activation of NF-$\kappa$B factors, in murine macrophage model cell line RAW264.7. Therefore, *S. uberis* is recognised as a threatening pathogen by professional immune cells like macrophages [83].

PMNs play a crucial role in intramammary infection (IMI) caused by *E.coli* and *S.aureus* by killing the ingested bacteria at infection site [180]. In contrast, killing of *S. uberis* has been linked with strain-specific virulence as shown by Hill [88]. According to Tassi etal, cows challenged with a putative host adapted strain of *S. uberis* FSL Z1-048 presented mastitis accompanied with clinical signs, influx of PMNs and lymphocyte, and increase of multiple cytokine responses including IL-1$\beta$, IL-6, CXCL8, IL-10, IL-12p40, and tumor necrosis factor-$\alpha$ (TNF- $\alpha$) whereas a putative host non-adapted strain FSL Z1-124 was unable to exert any clinical, cytological, or immunological response [201]. In another study described in Section 1.3, it was found that *S. uberis* O140J which is a virulent strain, was able to resist the killing by PMNs whereas *S. uberis* EF20, a non-virulent strain was more easily killed. Similar results were observed in a recent report and it has been stated that a host adapted strain of *S. uberis* FSL Z1-048 and a non-adapted strain of *S. uberis* FSL Z1-124 were both killed by bovine PMN [201].

## 1.5 Virulence factors of *Streptococcus uberis*

Several virulence factors have been identified in *S. uberis* and most are similar to those found in other streptococci. Strains associated with clinical disease have been associated with the presence of sortase-anchored proteins [120]. Sortase is a transpeptidase which covalently links important cell surface proteins usually containing a LPXTGX amino acid motif to bacterial cell walls [41, 138] and cleaves the bond between thr (T) and gly (G) residues in the motif [138] to anchor these proteins to the cell wall of Gram-positive pathogens including *S. uberis* [156]. The reference strain 0140J contains a single copy of gene (*SUB0881*) homologous to sortase A (*srtA*). Mutation of the *srtA* has been shown to reduce *S. uberis* infective potential [120]. This suggests that the sortase-anchored proteins contain one or more virulence factors important for establishment of infection. It is likely that inhibition of function of neutrophils is of vital importance to the outcome of *S. uberis* infection as they constitute one of the major defensive cells of the lactating gland. Molecules secreted by *S. uberis* that are capable of inhibiting neutrophil bactericidal function have not been identified [61] and this is an on-going area of research. The role of potential iron-binding proteins, lactoferrin binding proteins (*Lbp*) (*SUB0145*) and *S. uberis* adhesion molecule (*SUAM*) (*SUB1635*) have been investigated in the pathogenesis of *S. uberis* [3, 148]. The *SUAM* protein is thought to be important during the initial stages of the infection, colonization and adherence to the bovine mammary epithelial cell line in vitro [3]. This protein has been found to have an affinity for lactoferrin, however, using *lbp* mutants in adherence experiments [216] has suggested that this is not required for attachment of *S. uberis* to host epithelial cells or for growth in iron-limiting conditions

[216].

Activation of plasminogen to plasmin is another mechanism which is likely to facilitate a variety of streptococcal infections. Plasmin is a known serine protease and binds to the surface of a number of bacterial pathogens. Streptococci express specific plasminogen activators which are associated with the respective host's plasminogen to form an activation complex. Most *S. uberis* isolates encode *pauA* (*SUB1785*) which is located between the *hexA* and *hexB* genes which are homologous to the *mutS* and *mutL* DNA mismatch repair genes [173]. A number of strains have been found lacking the gene *pauA* containing an alternative novel plasminogen activator (*pauB*) and this is also found at the hexAB locus [217].

Hyaluronidases have been found to break down the hyaluronic acid component of connective tissues facilitating the spread of the pathogens. Hyaluronidase accelerates the spread of high molecular weight compounds rather than bacterial cells itself [194] and facilitates the effects of other streptococcal virulence factors such as pneumolysin [60].

Hyaluronic acid (HA) capsule is a well-studied virulence factor in *S. uberis*. The arrangement of the operon containing hyaluronic acid biosynthetic genes in *S. uberis* differs from the hasABC arrangement of Group A Streptococci (GAS) [215]. The genes *hasA* (*SUB1697*) encoding hyaluronan synthase and *hasB* (*SUB1696*) encoding UDP-glucose dehydrogenase, is arranged in a similar fashion as in GAS. The *hasC* homologue (*SUB1691*), encoding UDP-glucose pyrophosphorylase, was found downstream of *hasB* in the reverse orientation and separated from the *hasA*/*B* genes by approximately 3 kb coding regions, however, this is not involved in capsule biosynthesis [216]. Studies have

shown conflicting results of the protection provided by the capsule for *S. uberis* invasion of the bovine mammary gland [88, 118, 215]. Hill et al. found that an encapsulated strain showed increased infectivity and virulence with experimental challenge of the lactating mammary gland compared with the acapsular isolate [88]. In support of this, Ward et al. investigated the resistance of capsular and acapsular strains to phagocytosis by bovine neutrophils in vitro and found that the capsular strains were more resistant to killing by neutrophils [215]. This finding was complemented by in vitro studies by Field et al. however, this was not evident with the experimentally challenged lactating bovine mammary gland with either an encapsulated, or genetically altered acapsular mutant as both strains caused similar levels of infection [61]. Although these studies suggest that HA capsule may provide minimal protection against bovine neutrophils in the invasion of the bovine mammary gland in vivo [167] the carriage of hasABC genotype has been found at a higher frequency in isolates associated with disease, suggesting that the capsule is required for some aspects of pathogenesis [61]. It has also been suggested that *hasA* gene could either have a direct influence on virulence or may affect the presence and activity of other genes involved in virulence [168].

Substantial efforts have been invested towards designing a vaccine against *S. uberis*. Hill et al. infected cattle with *S. uberis* via the teat canal and found that following cure of the initial infection a significant proportion of animals were resistant to subsequent infection and it was suggested that the observed T-cell increase might be considered responsible for the protection seen [88]. Subsequent studies using crude vaccines have shown that heat-killed *S. uberis* administered locally to the mammary gland can increase protection

against the same strain as the vaccine [63]. Attempts at protection via a humoral response against *S. uberis* have consisted of vaccinating cows with a cloned *GapC* molecule [67], plasminogen activator, *PauA* [117] and with a *SUAM* molecule [3]. Vaccination typically induced high levels of antibodies in mammary gland secretions but the effects of these antibodies on protection against IMI and clinical mastitis has not been convincing. In the case of *GapC*, vaccination did result in a less severe inflammatory response [67] and *PauA* vaccination protected against IMI with *S. uberis*. However, a number of strains have been found lacking the gene encoding *PauA* and *S. uberis* isolates with alternative pathways using *PauB* have been shown to exist on commercial dairies [217]. The major challenge in the design of an effective vaccine against *S. uberis* is the high level of genetic variation as such variability will make the selection of vaccine antigen candidates more difficult. The differences between antibody-based or cellular-based memory may be of particular importance in designing vaccines for *S. uberis* IMI.

## 1.6 Strain-typing of *S. uberis*

There are a number of molecular and non-molecular typing methods available for *S. uberis*. However, the non-molecular typing methods such as phage typing [198] serotyping [221] or bacteriocin-like inhibitor typing schemes [89] are unsuitable for the analysis of closely related strains. Molecular methods including DNA restriction analysis by pulsed-field gel electrophoresis (PFGE), multilocus enzyme electrophoresis (MLEE) [183] and multiple locus sequencing typing (MLST) [225] are more reliable and have been found

to be highly discriminatory [49, 95, 164] as well as more suitable for understanding the long-term epidemiology of *S. uberis*. MLST involves characterization of isolates of microbial species on the basis of DNA sequences of internal fragments of multiple housekeeping genes. Seven loci have been used for MLST profiling of UK isolates [38] including carbamate kinase (*arcC*), d-alanine-d-alanine ligase (*ddl*), glucose kinase (*gki*), transketolase (*recP*), thymidine kinase (*tdk*), triosephosphate isomerase (*tpi*), and acetyl-coenzyme A acetyltransferase (*yqiL*) and each unique allelic profile is assigned as a sequence type (ST).

Analysis of 160 UK isolates from milk samples was resolved into 57 STs and ST-5 was predicted as the founder of the lineage consisting of the highest number of single locus variants and isolated most frequently (112 isolates (70%) and 24 different STs) from UK population (Fig. 1.2). The ST-5 complex includes the well characterized virulent isolate 0140J which was assigned to ST-1. The non-virulent isolate EF20 was assigned to ST-55 which was not assigned to any clonal complex . There was no clear association between disease status and STs. A similar outcome was revealed from the analysis of 253 New Zealand isolates from milk, other bovine sites, and the environment in regards to the relationship with the disease status. New Zealand isolates were categorized into 131 different sequence types (STs)and a few of the New Zealand isolates (16 isolates and 7 STs) were assigned to the clonal complex ST-5. Identification of two new clonal complexes i.e. ST-143 and ST-86 also revealed that the ST-143 complex being the most predominant containing 106 isolates and 44 STs (Fig 1.3). The ST-86 complex consisting of 27 New Zealand isolates resolved into 20 STs [168]. Using the MLST technique, carriage of

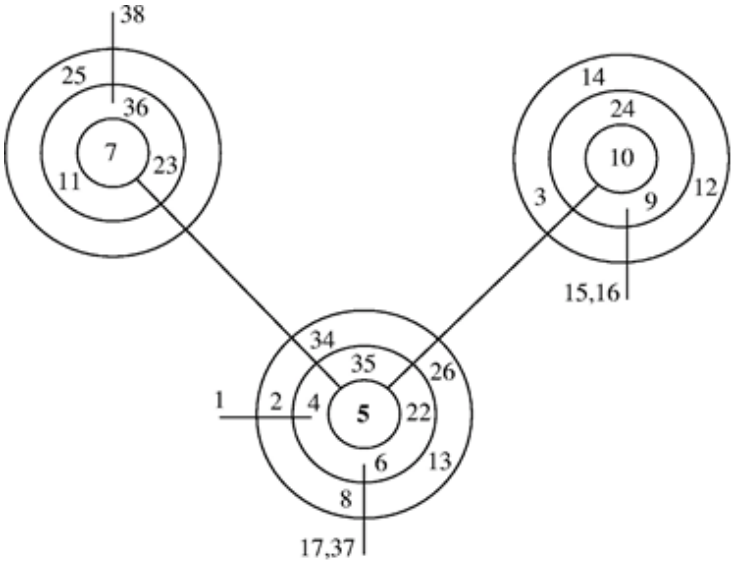



Figure 1.2: BURST analysis of UK isolates showing the relationships between the STs. Central STs represent the founder of the respective group. The inner rings show single locus variants, the outer rings and STS outside the rings are double locus variants, and triple locus variants, respectively. Adapted from Coffey et al. (2006) [38].

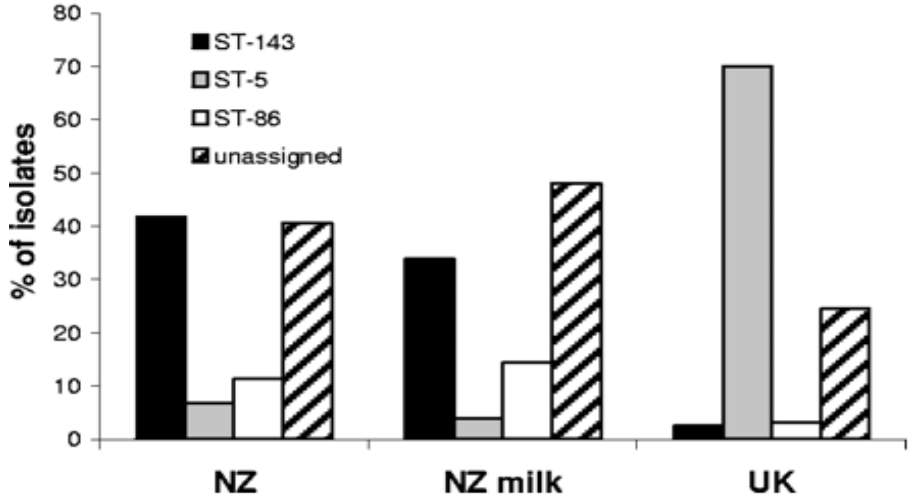



Figure 1.3: Frequency of different clonal complexes in New Zealand and UK. NZ- New Zealand isolates from different sources; NZ milk- New Zealand isolates from milk and UK- UK isolates from milk. Adapted from Pullinger et al. (2006) [168].

potential virulence genes by particular lineages was investigated by the analysis of *hasA*, a gene required for capsule biosynthesis [215]. Highly significant differences were seen in the sequence type (ST) profile obtained from isolates with or without the *hasA* gene [38]. The majority of the UK isolates (16 of 21 [76.2%]), which lacked the *hasA* gene were from sub-clinical infections, whereas, only two isolates were from clinical cases and three isolates from post-treatment cases. This finding was in agreement with several other studies conducted in Denmark and UK [61]. Since the role of capsule as a potential virulence factor is unclear it has been suggested to determine which genes other than *hasA* are also linked to this lineage [38].

## 1.7 *Streptococcus uberis* genome overview

A consortium of researchers completed the sequencing of the 1.85 Mb genome of *S. uberis* strain 0140J in 2009 [216]. Annotation is available via the EMBL/GenBank databases (accession number AM946015). The genome of *S. uberis* 0140J is a single circular chromosome of 1,852,352 bp, which contains 1,825 predicted protein coding sequences (CDS), 62 of which were found to be pseudogenes or gene fragments. Comparisons using representatives of each of the sequenced Streptococcus species showed that ~ 40% of *S. uberis* CDS had identifiable orthologs in the compared streptococcal genomes and the greatest overall conservation with the closely related pathogenic species *S. zooepidemicus* and *S. pyogens* [216]. The phylogenetic relationship among different Streptococci based on combined 16s rRNA and RNase P RNA gene (*rnpB*) is shown in Figure 1.4.

Most of the conserved genomic regions were colinear, with some regions translocated or inverted. Discrete regions of variation including bacteriophage-derived islands, and a putative genomic island were also identified throughout the genome of *S. uberis* 0140J suggestive of different evolutionary origins for this component of the genome. The low number of MGEs (1.7%) in the *S. uberis* genome is unrelated to other related streptococci [164].

A number of CDS involved in central metabolism, transport and gene regulation are conserved between *S. uberis* and *S. zooepidemicus*. On the contrary, they are absent between *S. uberis* and *S. pyogenes*. In comparison to *S. uberis* and *S. zooepidemicus*, the habitat of *S. pyogenes* is highly restricted, this is likely reflected in the lack of shared genes in this specialized bacteria [216]. Although this comparative genomics analysis revealed common evolutionary origins of *S. uberis* along with other related pyogenic streptococci, differences in their host association and pathogenecity suggest that significant divergence from their common ancestor has occurred.

With the advent of next-generation sequencing (NGS) technologies it is possible to sequence whole genomes of pathogens at a lower cost and in a timely manner. Inter and Intra-species comparative genomic studies of a large number of sequenced genomes will possibly reveal a clearer picture of the molecular mechanism of pathogenesis as well as allow identification of potential targets for intervention of this important pathogen.

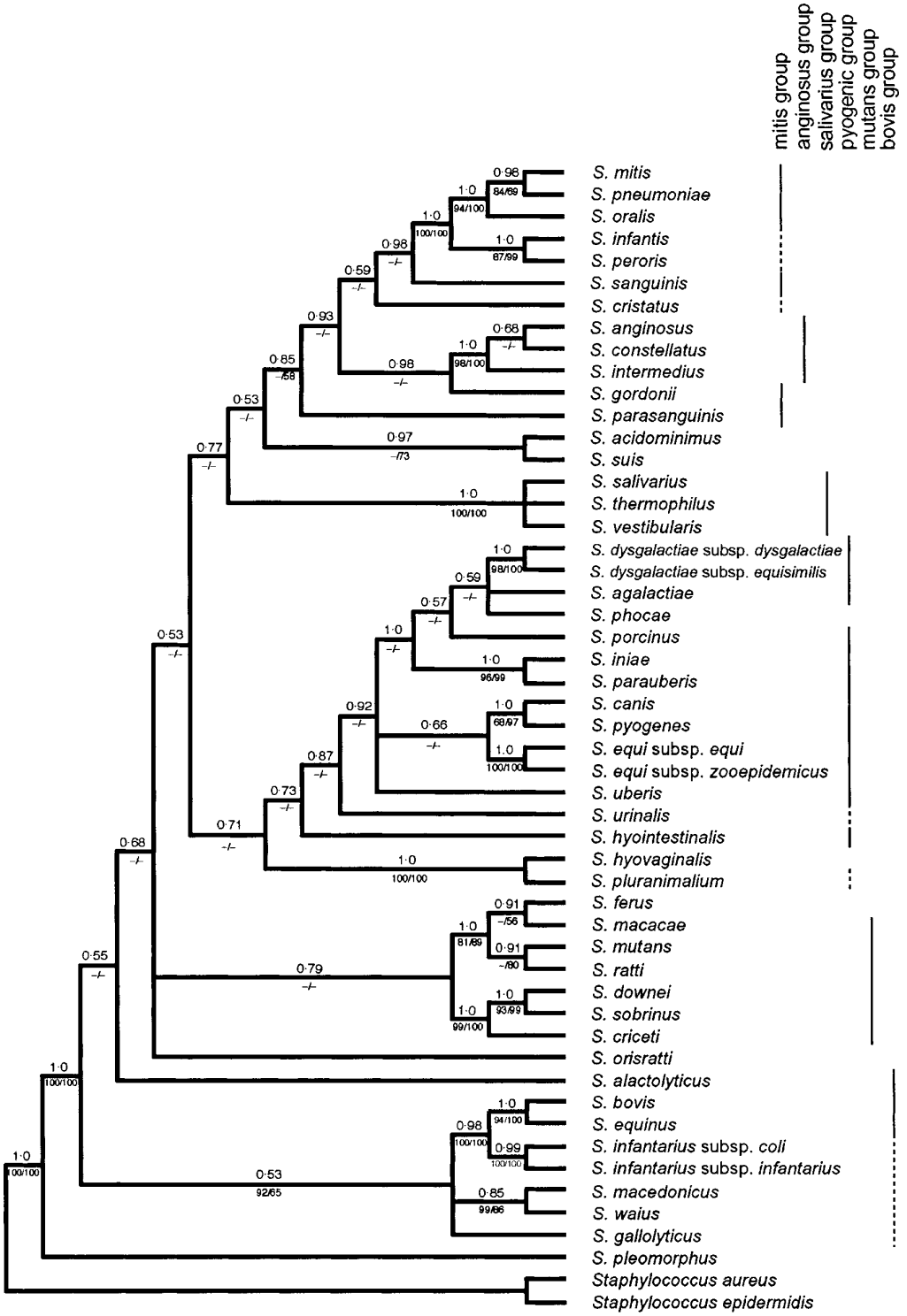



Figure 1.4: Bayesian phylogeny (majority-rule consensus tree ) based on combined 16srRNA and *rnpB*. Posterior probabilities shown in the above branches and values below branches indicate bootstrap percentages for maximum-parsimony (MP) and minimum-evolution (ME) analysis. Bars represents the taxonomic groups within Streptococcus [102]. Solid lines represents original classification and taxa added in subsequent analysis are shown using dashed bars.

## 1.8 Next-generation sequencing technologies

Within the last 5-10 years there has been a dramatic shift from automated Sanger sequencing for genome analysis to NGS technologies that deliver fast, inexpensive and high coverage genome information [141].

Generation of a large number of low-cost and high quality reads have made NGS technology useful of many applications, including:

- Variant discovery by resequencing targeted regions of interest or whole genome.
- *De novo* assembly of bacteria, viral or eukaryotic genomes.
- Gene discovery, population genetics study using metagenomics approaches
- Species classification
- Cataloging the transcriptomes of cells, tissues and organisms (RNA-Seq)
- Genome-wide profiling of epigenetic markers and chromatin structure using other sequence based methods (Chip-seq, methyl-seq, and DNase-Seq)

NGS technologies constitute various strategies based on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. For all approaches, genomic DNA is randomly fragmented to generate fragment template or mate-pair templates. Mate pairs are created by fragmenting the genomic DNA then size-selected inserts are circularized and linked by means of an internal adapter. The mate-pair library is generated by sequencing around the adapter. In paired-end sequencing, after fragmentation of genomic DNA into short (< 300 bp) segments both

ends of the segments are sequenced. In each approach, the template is then immobilized to a solid surface to allow thousands to billions of sequencing reactions to be performed simultaneously.

Different sequencing strategies are exploited by different vendors for example, in Illumina, SOLiD, PGM and 454 systems [144]the initial template molecules are amplified into clusters of "polymerase colonies" or "polonies" since these systems are not sensitive enough to detect single molecules. Alternatively, newer technologies such as Heliscope and PacBio SMRT systems do not require an amplification step due to their high sensitivity in detecting single molecule template extension. Utilization of different strategies also affects the output capacity from different platforms [25].

Currently Illumina is the most widely used sequencing platform. This method adopted the sequencing by synthesis (SBS) approach [144] and uses fluorescently labelled reversible terminator nucleotides on clonally amplified DNA templates immobilized to an acrylamide coating on the surface of a glass flowcell. The initial step of library preparation for Illumina sequencing involves denaturing double strand DNA to produce single strand DNA molecules to which adapters are ligated, these then attach to the flow cell surface (Fig 1.5; step 1). On the surface of the flow cell forward and reverse oligos complementary to the adapter sequences serve as primers to form an initial copy of the individual sequencing template molecule (Fig 1.5; step 2). Fragments attached to the flow cells are used to generate a cluster of identical template molecules by a process known as bridge amplification (Fig 1.5; steps 3-5). Once clusters are generated, nucleotides are sequenced in a repetitive manner during incorporation of flourescently

labeled dNTPs. The four dNTPs contain specific fluorescent labels that serve to identify the base and they prevent multiple extension events by acting in a reversible terminator manner. Fluorescent groups are removed after imaging and the reversible terminator is deactivated to allow for the next incorporation cycle [25].

Sequence by ligation (SBL) differs from SBS in the use of DNA ligase [208] and either

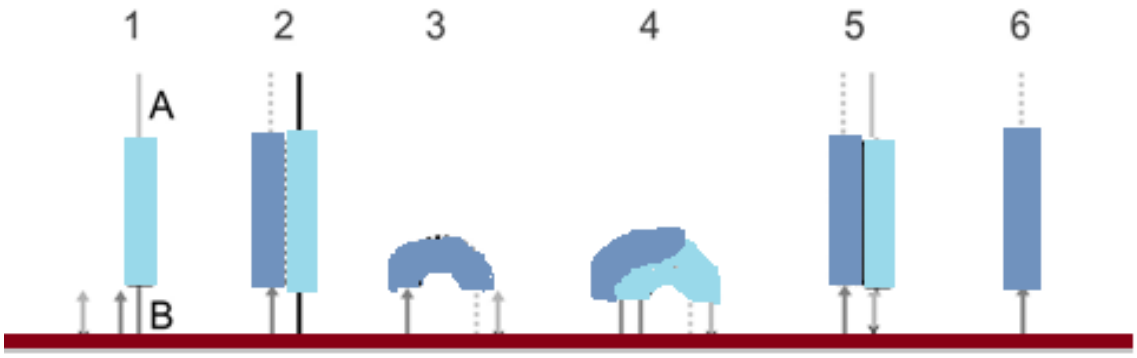



Figure 1.5: Schematic diagram of template generation via bridge amplification [25]. Step 1: random fragmentation of genomic DNA sample and adapter ligation; step 2: bridge amplification of unlabeled nucleotides and enzymes are added to initiate solid-phase amplification; step 3-5: building double-stranded bridges, denaturation of double-stranded DNA, completed amplification results in cluster generation; step 6: incorporation of fluorescently labeled dNTPs and base identification.

one base-encoded probe or two base-encoded probes. The ABI SOLiD sequencer uses the SBL method and has its origins in the system described by Shendure et al. [185] and in work by McKernan et al. [140] at Agencourt Personal Genomics (Beverly, MA, USA) (acquired by Applied Biosystems (Foster City, CA, USA) in 2006). It uses a process of clonal proliferation by emulsion PCR and also uses a di-base sequencing technique in which two nucleotides are read by ligation simultaneously during the sequencing process.

The SOLiD system uses four dyes for all 16 possible pairs of di-bases. The advantage of this technology is the read accuracy since each base in interrogated by two octomer ligations [25]. One of the disadvantages of this technology is that as deciphering the bi-base encoded sequence relies on knowing the first of the bases in the pair, if one of the colours in a read is misidentified for example due to sequencing error this will have an effect on all subsequent bases in the read [131].

One of the newest technologies is nanopore sequencing from Oxford Nanopore technologies and the expected outcome from nanopore sequencers are single particle, amplification free long reads. The basic principle is to pass the molecules through a pore formed by either the protein alpha-hemolysin [37] or manufactured solid state pores [66, 222]. In 2012, Oxford nanopore reported the arrival of their palm sized device MinION which have resulted in various publications which hold great potential in the area of bacterial genomics [7, 100, 128, 129].

## 1.9 Aims and objectives

*S. uberis* is the major bacteria associated with bovine mastitis and it is predicted that vaccination against this pathogen is possible but efforts to achieve this are hampered by a lack of understanding of the pathogen at the genomic level. At the beginning of the project, sequence data generated by the SOLiD platform from 22 strains of *S. uberis* with known differences in virulence were available. The primary objective was to develop pipelines for *de novo* assembly of those sequences and analysis of gain/loss of genes in different strains. In addition an additional 12 strains were sequenced using the Illumina platform allowing comparison of assembly methods between these technologies.

By comparing the sequence of two or more related genomes it is possible to identify parts which are similar and others that are different. Differences in genomes are produced by changing of the DNA sequence or by gain/loss of genes over time. If these changes are useful to the organism they are preserved by the process of natural selection. A greater understanding of adaptive evolution within pathogenic species will link large scale genome studies to the biology of pathogens and identify potential new targets for intervention.

Specific objectives of the study were:

1. To develop pipelines for *de novo* assembly of whole genome sequences available from SOLiD and Illumina MiSeq platforms and comparative analysis of these two platforms

2. To compare assembled genomes of *S. uberis* isolates and analysis of gene gain/loss

to link this information to the virulence properties of *S. uberis*

3. To identify the genetic differences between clinical/sub-clinical and/or virulent/non-virulent isolates that may correlate with and play a causative role in pathogenesis.

4. To functionally annotated proteins using intra- and inter-species protein-protein interaction (PPI) information.

CHAPTER 2

# MATERIALS AND METHODS

## 2.1 Bacterial isolates and genome sequencing

For bacterial whole genome sequencing using the SOLiD platform 22 *S. uberis* isolates (11 from clinical and 11 sub-clinical cases) were used. Their ST profile available in the *S. uberis* mlst database (http://pubmlst.org/suberis/) is provided in Table 2.1. Among these 22 isolates 11 isolates (6 from clinical and 5 from sub-clinical cases) were further sequenced by Illumina MiSeq. Isolates representative of different ST complex and mostly ST-5 complex (most prevalent in UK) were chosen for Illumina sequencing. Bacterial DNA samples were prepared for sequencing by Dr. Sharon Egan University of Nottingham. Library preparation and sequencing of each strain was conducted at DeepSeq, Queens Medical Centre, Nottingham, UK.

Table 2.1: Isolates used in the study for whole genome sequencing. All twenty two isolates were used for SOLiD sequencing and the isolates with an asterisk were chosen for MiSeq Illumina sequencing

| Strain ID | Disease | Year | Clonal complex |
|---|---|---|---|
| 0140J | mastitis (clinical) | 1973 | ST-1 complex |
| *EF20/0145 | mastitis (clinical) | 1970 | ST-86 complex |
| *B362 | mastitis(sub-clinical) | 2000 | ST-86 complex |
| 3448 | mastitis(sub-clinical) | 1999 | N/A |
| 4244 | mastitis(sub-clinical) | 1999 | N/A |
| 4428 | mastitis (clinical) | 1999 | ST-5 complex |
| *C5072 | mastitis(sub-clinical) | 1999 | ST-86 complex |
| 5130 | mastitis (clinical) | 1999 | ST-86 complex |
| 5291 | mastitis (clinical) | 2000 | ST-5 complex |
| *C5388 | mastitis(sub-clinical) | 2002 | ST-5 complex |
| *6736 | mastitis(sub-clinical) | 1999 | ST-143 coplex |
| 6707 | mastitis(sub-clinical) | 2000 | ST-86 complex |
| *6780 | mastitis (clinical) | 2000 | ST-86 complex |
| S7010 | mastitis(sub-clinical) | 2002 | N/A |
| C7131 | mastitis(sub-clinical) | 2002 | ST-5 complex |
| *Ab71 | mastitis (clinical) | 2002 | ST-5 complex |
| Ab1947 | mastitis (clinical) | 2002 | N/A |
| *B190 | mastitis(sub-clinical) | 2000 | ST-5 complex |
| *S6261 | mastitis (clinical) | 2002 | ST-5 complex |
| *C6344 | mastitis (clinical) | 2002 | ST-5 complex |
| *C8329 | mastitis(sub-clinical) | 2002 | ST-5 complex |
| *C9359 | mastitis (clinical) | 2002 | ST-5 complex |

## 2.2 *De novo* assembly of *Streptococcus uberis* genomes from SOLiD short-read data sets

The SOLiD System software "small genomes *de novo* accessory tools for SOLiD 3 plus" (SOLiD *de novo* accessory tools) v.2.0 enables *de novo* reconstruction of small genomes up to 30Mb long from short reads at coverage 50X or higher generated by Applied Biosystems SOLiD 4 system (Biosystem). The software supports different library reads including fragment, mate-paired and paired-end formats. The SOLiD Accuracy Enhancement

Tool (SAET) was used to pre-filter error and improve the average contig length and assembly quality. Before pre-filtering the quality of reads were assessed using the High-Throughput Quality Control (HTQC) tool, which consists of six programs for read quality assessment, read filtration and generation of graphical reports [224]. The SOLiD *de novo* accessory tools use Velvet as the short read assembly tool. Velvet is an open-source package developed at the European Bioinformatics Institute (EBI). Velvet version 1.2.03 [228] was downloaded from source archive http://www.ebi.ac.uk/ zerbino/velvet/. Velvet requires preprocessing of SOLiD reads into colorspace fastq format for initial assembly. An outline of the approach used for genome assembly is shown in Fig. 2.1. Details of individual steps in the pipeline are given in the following sections.

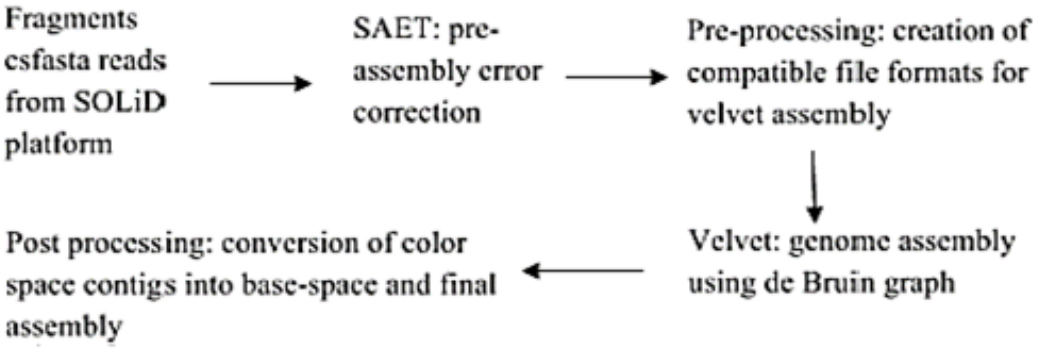



Figure 2.1: Schematic diagram of *de novo* assembly pipelines for SOLiD reads.

### 2.2.1 SOLiD Accuracy Enhancement Tool

SOLiD Accuracy Enhancement Tool (SAET) was used for pre-assembly base call error correction

(http://bcc.bx.psu.edu/download/saet.2.2/SAET.v2.2.pdf). The removal of low quality reads (qv <20) was also conducted to overcome potential memory constraints.

Usage: ./saet_mp <.csfasta> <qual> expectedlength(int) [-options]

By default, SAET takes reads file in csfasta format and a quality value file (when available) in .qual format and an expected length of the reference genome size from which reads were sampled. In our study, the expected genome size was set as 2 Mbp to reflect the genome size of reference strain 0140J (1.85 Mbp).

### 2.2.2 Preprocessor

Velvet is not directly compatible with the colorspace reads available from SOLiD platform so the reads were converted to colorspace fastq format using the perl script preprocessor.pl. The script trims the first base (primer) and the first colour of every read and translates the colours 0123 to the *pseudobases* ACGT

perl preprocessor.pl –f .csfasta –t fragment –o output_directory

-f csfasta file available after SAET correction

-t type of reads; SOLiD fragmenet reads in this study

-o output directory where the preprocessed file will be saved.

### 2.2.3 Choice of short read assemblers

Velvet was used for *de novo* assembly as at the time of analysis this was the only open source tool for the assembly of SOLiD colorspace reads. Velvet uses de Bruijn graph approach [228]. The pipeline uses the read tracking information from Velvet to form an

input file that can be submitted to postprocessing scripts (postprocessor.pl and denovoadp scripts) to extend contigs and convert them to base-space. Velvet was compiled to comply with the color space data, larger K-mer length and 3 categories

> make clean

> make MAXKMERLENGTH=250 CATEGORIES=3 color

#### 2.2.3.1 Running velveth

Velveth helps to construct the dataset for the following program, velvetg:

Usage:

>velveth outputdirectory hashlength [-fileformat][-readtype] filename]

Supported file formats:

fasta (default) , fastq . fasta.gz , fastq.gz ,SAM, BAM, eland, gerald

Read categories :

short (default), shortPaired, short2 (same as short, but for a separate insert-size library)

shortPaired2, long (for Sanger, 454 or even reference sequences)

Multiple k-mers :

Various k-mer lenghts was tested without doing redundant computations:

> velveth outputdirectory m,M,s (.data files.)

m <= k <M with a step of s

#### 2.2.3.2 Running velvetg

velvetg is the core of Velvet where the de Bruijn graph is built and afterwards manipulated to generate final contigs for downstream analysis.

Intitially the following command was run to check the coverage distribution : ./velvetg_de output _directory/ -exp_cov auto

The coverage distribution was optimized using the following R script on the velvet output file, "stats.txt"

data = read.table( "stats.txt", header=TRUE)

hist(data$short1_cov, xlim=range(0,50), breaks=1000000)

library(plotrix)

weighted.hist(data$short1_cov, data$lgth, breaks=0:50)

The histogram generated from the above R script provides the estimation of coverage distribution on the corrected reads and was used for final assembly. For example,

./velvetg_de outputdirectory -exp_cov 180 -cov_cutoff 50 -amos_file yes -read_trkg yes -min_contig_lgth 100

-read_trkg option will cost more memory and calculation time, but produce a more detailed description of the assembly.

-amos_file produce AMOS files and this .afg files have all the assembly information in one data structure and the information contained used for final base-space assembly.

## 2.3 *De novo* assembly of *Streptococcus uberis* genomes from Illumina MiSeq reads

### 2.3.1 Quality control and *de novo* assembly

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for quality control checks on 2x250 bp paired-end raw sequence data coming from Illumina MiSeq pipelines. Due to the high quality reads available from the MiSeq platform reads were not filtered but used directly to generate assemblies using Velvet (version 1.2.10) as described above. Maximum assembly statistics such as N50 (described in section 2.5) was used as the measure to determine optimal K-mer length and a minimum coverage of 10x was selected and the exp_cov option 'auto' was used.

### 2.3.2 Database submission of Illumina assembly

Before submission of genomes, the source and project information was registered to retrieve BioSample and BioProject IDs. Contigs >199bp were submitted as fasta format with the extension .fsa. Required source information was provided such as : [organism= Streptococcus uberis] [strain=EF20] [gcode=11]. Tbl2asn is a command-line program provided by NCBI (ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/) was used to create sequence records for submission to GenBank. The template file (suffix .sbt) containing the source and reference information provided during the project registration and nucleotide sequence in FASTA (suffix .fsa) were used to generate a Sequin file (suffix .sqn) using tbl2asn. Error summary with suffix .val was checked for severity and

type of errors and the errors were fixed during submission. Sequence reads and assembled contigs are available at GenBank under accessions JANW00000000, JATB00000000, JATD00000000, JATK00000000, JATE00000000, JATC00000000, JATI00000000, JATF00000000, JATA00000000, JATG00000000, JATJ00000000, and JATH00000000.

## 2.4 Assembly evaluation

QUAST [84] - a quality assessment tool for evaluating and comparing genome assemblies was used to evaluate assemblies from both SOLiD and Illumina datasets. A number of metrices including the following were used to assess the quality of de novo assembly: N50 contig: the length at which 50% of assembled bases belong to a contig of this length or longer (a higher N50 indicates a better assembly)

- Mean contig length (higher values are better)
- Maximum contig length (higher values are better)
- Number of contigs >200 bases (fewer contigs with larger sizes is preferred) and >1000 bases
- Nx : The largest contig length, L, such that using contigs of length accounts for at least x% of the bases of the assembly.

## 2.5 Genome annotation

Assembled contigs were annotated using the Rapid Annotations using Subsystem Technology (RAST) server [155]. RAST identifies protein-encoding, rRNA and tRNA genes, assigns functions to the genes, and predicts which subsystems are represented in the genome, uses this information to reconstruct the metabolic network and makes the output easily downloadable for the user. PILER-CR [52] and CRISPRs web server [79, 94] was used for rapid identification and classification of clustered regularly interspaced short palindromic Repeats (CRISPRs). In addition to the identification of CRISPRs regions the CRISPRs web server also predicts CRISPR-associated genes within the region. The phage search tool (PHAST) [233] was used to identify, annotate and graphically display prophage sequences within the draft genomes. The prophage regions were classified into three main categories: intact, questionable or incomplete based on the completeness of the identified regions.

CONTIGuator [69] was used to map resulting contigs to the reference genome 0140J for comparative analysis of genomic regions. The program performs a mapping step against the reference genome using the blastn algorithm without mapping contigs to more than one replicon considering many descriptive features like putative overlaps between mapped contigs and syntenic and non-syntenic regions inside and outside contigs. The maps were visualized and analysed using the artemis comparison tool (ACT) [33]. Rapid annotation transfer tool (RATT) [154] was used to transfer the gene annotation of the reference genome to the other genomes and the nucleotide sequences of CDS of orthologs were aligned using Muscle [52]. Assembled contigs are available at GenBank

under the accession numbers provided in Section 2.3.2. RAST annotation are available as Additional file 2.1.

## 2.6 Pan and core-genome analysis

The pan-genome analysis pipeline (PGAP version 1.02) [216, 231] was used for identification of orthologous genes between the twelve sequenced genomes and the reference genome 0140J using the Gene Family (GF) method (50% coverage and with an e-value cut-off 1e-10) (Additional file 4.1). To determine the stability of pan/core genomes, the pan genome (total number of genes identified within a group of samples) and core genome size (those genes shared by a group of samples) was determined when between 2 and 13 genomes were combined in random order. For each combination size (2..13 genomes) 1000 permutations were conducted. Similarity of the samples based in gene presence/absence was conducted using hierarchical clustering. The pvclust package (http://cran.r-project.org/web/packages/pvclust/) using a correlation distance measure and average agglomeration method was used for clustering.

## 2.7 Multi locus sequence typing

MLST profiles of UK *S. uberis* isolates were downloaded from PubMLST database (http://pubmlst.org/suberis/) and analysed using the goeBURST option of PHYLOViZ [68].

## 2.8 Read mapping and variant calling

SOLiD colorspace reads were filtered on the basis of base quality and reads with Phred-like score <20 were excluded from alignment against the reference genome 0140J. Illumina MiSeq, reads were not filtered due to their high quality score. Mapping of SOLiD data was performed using BOWTIE 0.12.7 (9/7/10) [111] using a colorspace index of the reference genome. Illumina MiSeq reads were aligned using BWA [121].

The outputs from both the aligners were in the Sequence alignment/Map (SAM) format. This flexible and generic format stores information about alignments against a reference genome. SAM format has 11 mandatory fields for alignment information such as mapping position, and a number of optional fields [122].

The SAM files were converted to BAM format using SamTools version 0.1.19 [110, 122]. BAM files are binary representation of SAM and are compressed in BGZF format while holding the same information as SAM and so require less disk space. The mpileup function of SamTools was used to call variants for each isolate. The genotype of each variant was called with the 'view' option of bcftools [122]. Genotypes were called for each isolate with a minimum quality for 20.

The Perl script by Richard Emes was used for gene-based SNP analysis. This software requires gff file format of the reference genome and the vcf files (containing genome variation information) of individual strains. Variant calling format (VCF) is a text file format containing meta-information line, header line and data lines containing information about position in the genome. The format can also contain genotype information for each position (https://samtools.github.io/hts-specs/VCFv4.2.pdf).

DnaSP, DNA Sequence Polymorphism [174] software package was used for the analysis of nucleotide polymorphism from aligned DNA sequence data for assembled genomes.

## 2.9 Whole genome phylogeny

The webserver snpTree [114] was used to identify SNP positions from the concatenated 1,377 core genes of 22 isolates using SOLiD data and 13 isolates for MiSeq reads, respectively. PhyML [80] was used for the generation of phylogenetic trees using a GTR model estimated gamma distribution and 4 substitution rate categories. 200 bootstrap replicates were conducted. To determine the possible effect of recombination on phylogeny, the Phi-test [23] was conducted on non-overlapping 500 base windows of the alignment from Illumina assemblies. Following a Bonferroni correction, those windows with significant evidence ($p < 0.05$) of recombination were masked from the alignment and the phylogeny re-determined by PhyML [80].

## 2.10 Heatmap clustering and statistical analysis

Clustering was performed by the hclust function of R and visualised using the Heatmap.2 function of the gplots (http://cran.r-project.org/web/packages/gplots/gplots.pdf). Statistical analysis such as chi-square test, Mann-Whitney U test were carried out using standard R packages [203]. A number of line graphs were generated using MATLAB [136].

## 2.11 Analysis of genes undergoing positive selection

When comparing protein coding genes, two types of substitutions are possible: synonymous and nonsynonymous substitutions. Changes that retain the amino acid encoded by a codon are termed as synonymous substitutions (dS) and, if the amino acid is changed in a codon is termed as nonsynonymous substitutions(dN). The ratio of these two rates (dN/dS) provides the measure of selection pressure in a particular amino acid codon referred to as omega (œâ). Omega has three ranges: $\omega < 1$, $\omega = 1$ and $\omega > 1$ which are referred to as negative (purifying) selection, neutral evolution and positive (adaptive) selection, respectively. In this study we used the PAML suite of tools which incorporates a number of programs such as codon substitution model and estimates parameters (such as œâ) using maximum likelihood to identify codons and lineages under positive selection (Yang and Bielawski, 2000). We used the M0 model to infer a single omega value for each gene alignment. To detect selection by PAML we used the alignment of 1825 homologous gene sets (0140J plus other strains). 209 alignments in which open reading frames were disrupted were excluded from further analysis. For each gene alignment a phylogeny was generated using PhyML [80].

## 2.12 Network visualization and analysis

Network analysis was conducted using cytoscape version 2.8 obtained from http://www.cytoscape.org. Pathway analysis and enrichment was performed using Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com).

## 2.13 Prediction of *Streptococcus uberis* protein-protein interactions

*S. uberis* 0140J intraspecies protein-protein interactions (PPI) data was downloaded from STRING [197] which uses a combination of interaction data providing curated experimental data from different biochemical, biophysical and genetic techniques[103, 125, 175]. STRING also contains computationally predicted interactions from different online resources such as GeneMANIA [147], FunCoup [2], ConsensusPathDB [98] and other resources. It provides some consensus scores based on confidence scoring, coverage (based on number of proteins, organisms and prediction methods [197]. We considered different topological properties such as clustering coefficient, node degree distribution, shortest path length and betweenness centrality (see below for definitions ) using Cytoscape (version 2.8) [190] Analyze Network Plugin.

## 2.14 Homology based host-pathogen protein-protein interaction prediction

We used conventional homology-based approaches to investigate intra-species PPIs to predict host-pathogen PPIs. That is, if a gene encoding protein X exists in the bovine genome and protein Y in the *S. uberis* genome are homologous to a pair of known interacting proteins X' and Y' where X' is Human and Y' another bacterial pathogen, then proteins X and Y are predicted to interact also. Homologous proteins were predicted

between bovine and Human gene and *S. uberis* and pathogens using the blastp algorithm. Identified similar proteins were considered homologs if the region of similarity was ≥ 50% of the length of the query and subject proteins and the e-value was ≤ $1e-10$. The human-bacterial pathogen interaction data was downloaded from the PATRIC database [74]. Cellular Canonical pathway analysis was performed using Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com).

## 2.15 Scripts

Major scripts used and developed in the study are available at github

https://github.com/mdmaqsud/Cov.R

and commandlines used for assembly and mapping are provided in Appendix A2. The function of the major scripts are summarized in Table 2.2.

## 2.16 Additional files

Additional files are available at github https://github.com/mdmaqsud/Cov.R

Table 2.2: Major scripts used in the study are available at https://github.com/mdmaqsud/Cov.R

| Script | Function |
|---|---|
| blast.parser.pl | Parse a NCBI blastx output file and output the top N hits |
| cdna2longest.ORF.6frames.pl | Ouputs longest ORF among by searching 6 reading frames |
| cdna2pep.pl | Converts cdna to amino acid sequence |
| fasta.min.seqlength_check | Extract fasta sequences of specific length provided by the user |
| fasta2match | Extracts fasta sequences based on user defined ID |
| fasta2aln.SNP.count.pl | SNP and Indel calling from multiple sequence alignment |
| glimmer.parser | Parsing glimmer output |
| split_multifasta.pl | Separates multifasta files |
| denovoadp.pl | Converts the colorspace assembly from velvet to fasta format |
| heatmap.R | Produces heatmap custering |
| pvclustering.R | Hierarchial clustering with p-values |

CHAPTER 3

# *De novo* ASSEMBLY AND COMPARATIVE ANALYSIS OF SOLiD AND ILLUMINA SEQUENCING DATASETS

## 3.1 Introduction

With the rapid development of next-generation sequencing (NGS) technologies, genomic sequence information is within reach of most research groups. This has particularly impacted in the area of pathogen genomics where research can improve understanding and potentially improve quality of life. At the time of writing, the leading NGS systems are typically SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina, and GS FLX Titanium/GS Junior from Roche [127]. In this chapter, SOLiD and Illumina MiSeq are reviewed and the advantages and specifics associated with each sequencing system are discussed.

At the outset of this project SOLiD sequence data from twenty two *S. uberis* isolates was available. SOLiD single end data was best suited for whole genome re-sequencing specifically mapping the reads against an available reference genome. However, this approach is limited in that it only allows identification of genes present in the chosen reference genome but will not identify novel genes absent in the reference. This has particular importance when gene gain/loss may impact potential virulence properties of isolates.

The primary objective with these data was to develop pipelines for *de novo* assembly of SOLiD single end reads. Significant improvements in the sequence read length, yield, quality and cost have been observed in recent years due to the advances in next generation sequencing technologies. One such technology is the Illumina MiSeq platform. Given the improvement in sequence quality and availability of improved tools associated with the use of the Illumina platform, twelve of the isolates were re-sequenced using a MiSeq. This chapter introduces the parameter optimization for *de novo* assembly using SOLiD and Illumina platforms to produce high quality *de novo* assemblies to generate as complete draft genomes as possible. The assemblies obtained from both the platforms were analyzed and assessed with respect to the base quality, coverage distribution and K-mer optimization and assessed on the basis of N50 and number of contigs produced. The relationship of assembled draft genomes to known strains are also determined by use of phylogenetic approaches.

### 3.1.1 Technical aspects of NGS data

#### 3.1.1.1 SOLiD sequencing data csfasta

Sequencing data (hierarchical data format, HDF5) from SOLiD V4 is provided in colorspace fasta (CSFASTA) format. This is a text-based format similar to the widely used FASTA format [161], but each read (sometimes also called a tag) is represented as a series of colours (colorspace). Each unique read identifier (readID) consists of four components: the coordinates of the bead (panel, x-pixel and y-pixel) and a tag type, which is F3 for all single-end reads (read starts at the end of the DNA fragment with the adapter using forward (F) ligation chemistry). The colour calls are represented by four digits (0,1,2 and 3) corresponding to the four colours (Fig 3.1). SOLiD 4 system reports unknown colours by a single dot (.). In csfasta reads each fluorescent colour represents a set of two nucleotides rather a single nucleotide as with the other technologies. There are only four fluorophores available for 16 possible di-nucleotide combinations that may result in degenerate codes meaning the same color code for different di-nucleotide base combinations. For example the sequence 'AACTA' will be represented as transitions between nucleotides 0, 1, 2, 3 (also referred to as colours blue=0, green=1, yellow=2 and red=3) (Fig 3.1). The two files available from the colorspace instrument are colorspace fasta file and the corresponding quality file (Phred quality score) with the extension .csfasta and .qual (Appendix A1).

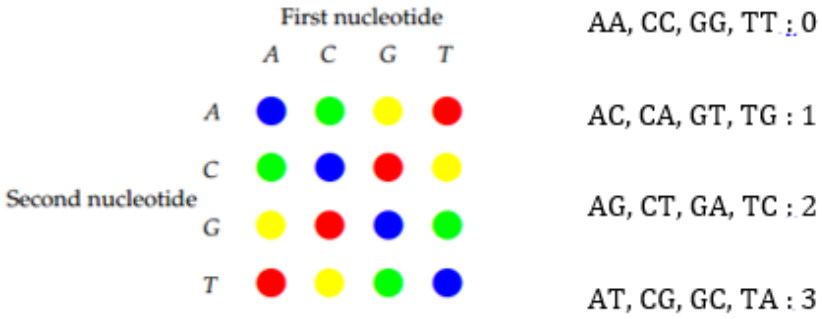



Figure 3.1: Di-nucleotide to colorspace encoding used by SOLiD sequencer (each fluorescent colour represents a set of two nucleotides). Identical transition: 0; complementary transition 3, non-complementary transitions are labelled with 1 and 2.

#### 3.1.1.2 Base quality Scores

Sequencing quality metrics provide useful information about the quality of data. Analysis of error probabilities is based on machine measurements taken during the sequencing run such as image intensity and noise to signal value and angle [178]. The log-transformation of an error probability p was introduced with the base-calling program Phred [57] to obtain a quality value known as Q-score. The Q-score is the most common measurement of base calling accuracy and is defined as a prediction of probability of an error in base calling (Fig 3.2).

QV = -10 $log_{10}$(p).....3.1

The SOLiD sequencer assigns a quality value (QV) to each called colour. The quality value represents an estimate for the accuracy of the colour call, which represents a measure of confidence (or uncertainty). Table 3.1 lists some examples of quality values with corresponding error rates and accuracies. In this work the term quality value, or the abbreviation QV, is used to refer to Phred-like quality scores generated by the

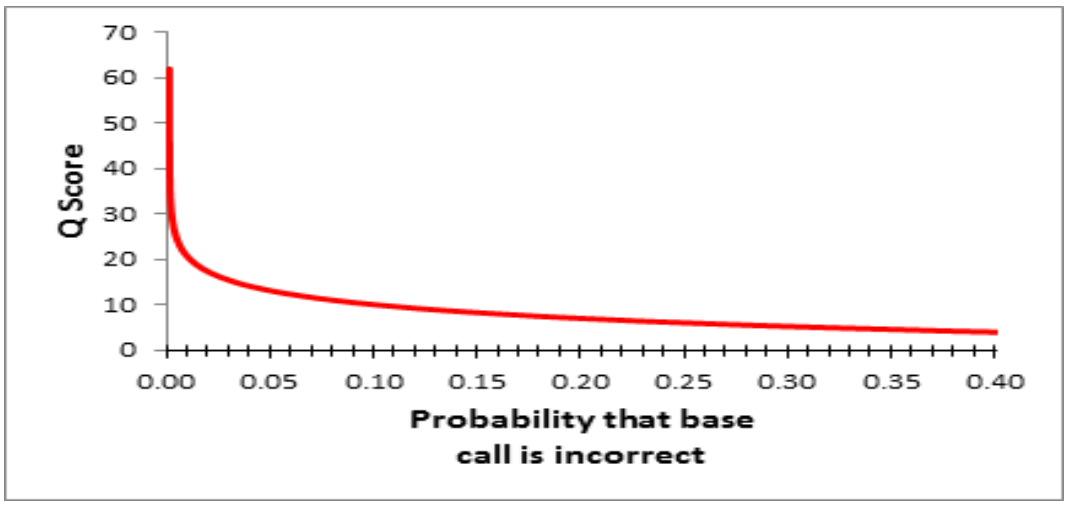



Figure 3.2: Plot of quality of reads (Q-score) as the function of probability of incorrect base call. High Q value indicates higher quality and lower probability of an incorrect base.

SOLiD V4 platform. These quality values are related to individual colours. The Q-score provides important information for how to use the data. Low quality sequencing creates significant hurdles and if the majority of the data is unusable this may result in a waste of time and resources as well as complicating downstream analysis.

Table 3.1: Q-scores, error and accuracy. A call with a quality value of 20 represents a probability of error 1 in 100, in other words the read is 99% accurate

| Q score | Error probability |
|---|---|
| Q40 | 0.00001 (1 in 10,000) |
| Q30 | 0.001 (1 in 1,000) |
| Q20 | 0.01 (1 in 100) |
| Q10 | 0.1 (1 in 10) |

#### 3.1.1.3 Illumina reads

MiSeq which uses Sequencing By Synthesis (SBS) technology integrates cluster generation, SBS, and data analysis in a single instrument. Illumina sequencers provide some initial information in the header of the generated reads (Fig 3.3). It is also flexible allowing read length from 36 bp to 2 x 300 bp [127]. The Illumina MiSeq reads use Phred scores ranging from 0-64. As the reads are available in FASTQ format the score is converted to one of the ASCII symbols. The scores are converted by adding 64 to the Phred score since Illumina version 1.3. Table 3.2 shows some typical quality codes for Illumina. Efficient mapping for millions of short reads to a reference genome and

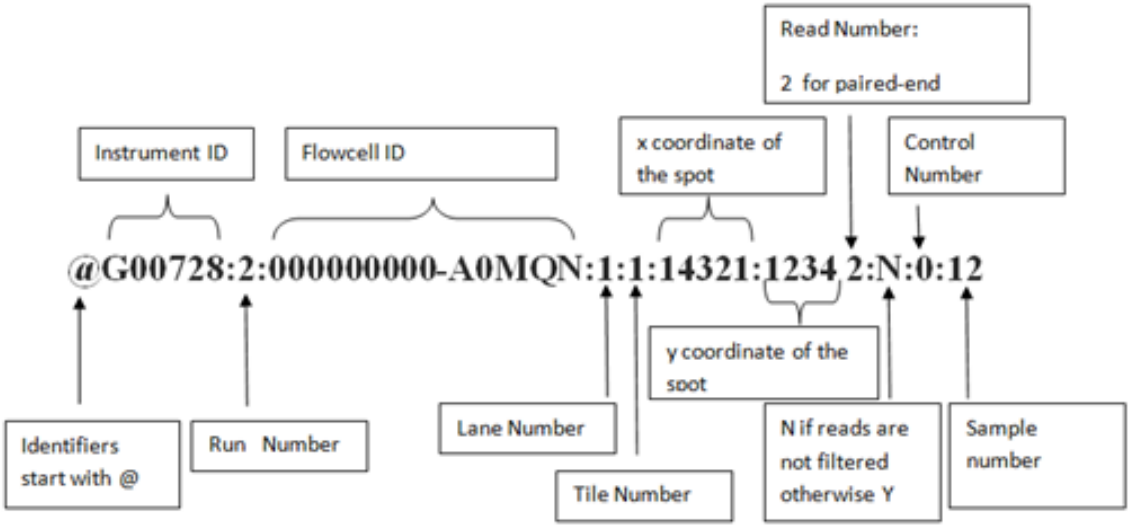



Figure 3.3: Example of a HiSeq FASTQ header. Raw sequence reads are returned in Fastq format. The Fastq header code number provides information on the machine used for sequence run and type of read processing conducted as shown in boxes above.

accurate SNP calling are some of the major challenges in the downstream analysis of NGS data and algorithms for sequence processing and alignment will affect outcome.

Table 3.2: Quality scores of Illumina reads and corresponding ASCII codes.

| Phred score | Quality score | ASCII code |
|---|---|---|
| 25 | 89 | Y |
| 26 | 90 | Z |
| 27 | 91 | [ |
| 28 | 92 | \ |
| 29 | 93 | ] |
| 30 | 94 | ^ |
| 31 | 95 | _ |
| 32 | 96 | ’ |
| 33 | 97 | a |
| 34 | 98 | b |
| 35 | 99 | c |

BWA [122] and BOWTIE [111] have been recommended by Wang et al. [214] as more efficient alignment tools especially for Illumina platform, whereas, NovoalignCS (http://www.novocraft.com/products/novoaligncs/) for SOLiD reads.

With short reads generated from NGS platforms the choice to align to a reference genome or conduct *de novo* assembly is dependent on the intended biological application and other relevant factors such as cost, effort and time. In most cases where a reference genome is not available and a "complete" genome is desirable, *de novo* assembly is required. There has been extensive effort in developing many algorithms for *de novo* assembly especially for bacterial genomes, including AbySS [187], ALLPATHS [27], Edena [86], Velvet [228] and SOAPdenovo [124]. All these programs are based on the de Bruijn graph data structure but differ in how they treat errors and in the use of read-pair information [131].

One of the major disadvantages of many NGS platforms are short read length compared to the conventional Sanger technology, for example, SOLiD sequencing typically yields

36-80 nt (although increased read length can be generated) and cause difficulties for the assembly of genome sequences that contain repetitive sequences, as is the case for many eukaryotes. The most successful *de novo* assembly has included reads from different platforms for example, the genome of the filamentous fungus *Grosmannia clavigera* was assembled from a combination of Sanger, 454, and Solexa sequence data [47]. However, a first draft of the 2.4 Gb Giant Panda genome was assembled from Solexa sequence reads alone [123].

Thus the advancement in this area now makes it feasible to undertake routine *de novo* assembly of whole prokaryote genomes allowing detailed insights into the novel biology of the organism under investigation.

## 3.2 Results

### 3.2.1 Genome sequencing and assembly using SOLiD

We performed SOLiD sequencing of the genomes of twenty two *S. uberis* strains with known clinical status to increase the understanding of the genetic differences between clinical and sub-clinical types (Table 2.1) and to identify sequences potentially associated with virulence. The SOLiD sequencing method was developed for mapping to a reference genome but this approach leaves the possibility of missing regions for which there is no match in the reference genome. To overcome this issue and to identify the new/accessory genes which are not present in the reference genome we attempted to determine the ability of SOLiD sequence reads to be used for a combination of genome mapping and *de*

*novo* assembly. One of the major problems for *de novo* assembly is short sequence reads available from the SOLiD sequencing platform. It was therefore extremely important to optimize different assembly parameters to get high quality long enough contigs for downstream analysis.

#### 3.2.1.1 Filtering error, optimisation of K-mer and coverage distribution improves *de novo* assembly for SOLiD reads

As assembly is highly sensitive to sequencing errors, one of the major challenges when dealing with *de novo* assembly of short reads is the filtering of low quality reads. As the SOLiD platform does not pre-filter low quality reads, a filtering framework SAET, the recommended optional tool from ABI SOLiD was used for filtering error to improve *de novo* assembly. SAET was used with the quality value (QV) cutoff, QV >=20 which equates to error probability of 1%. The assembly quality was assessed on the basis of N50 defined as the length of the contigs which constitute greater than 50% of the total length. It was also noteworthy that filtering using default SAET settings without using any QV threshold could not be assembled as the software (Velvet using default options) ran out of memory (RAM 24GB).

As with many enzymatic reactions, the quality of the reads decreases with time and higher cycle number. This results in a reducing quality score as the read increases in length, Fig 3.4 shows this decrease in quality towards the 3' end of the reads where a greater proportion of the reads are higher quality in the example of good data. Approximately 20 million reads were available for each of the genomes and after filtering

error 50-60% reads could be used for *de novo* assembly using Velvet (Appendix Table A1). The K-mer size is the most important parameter for all de Bruijn graph based

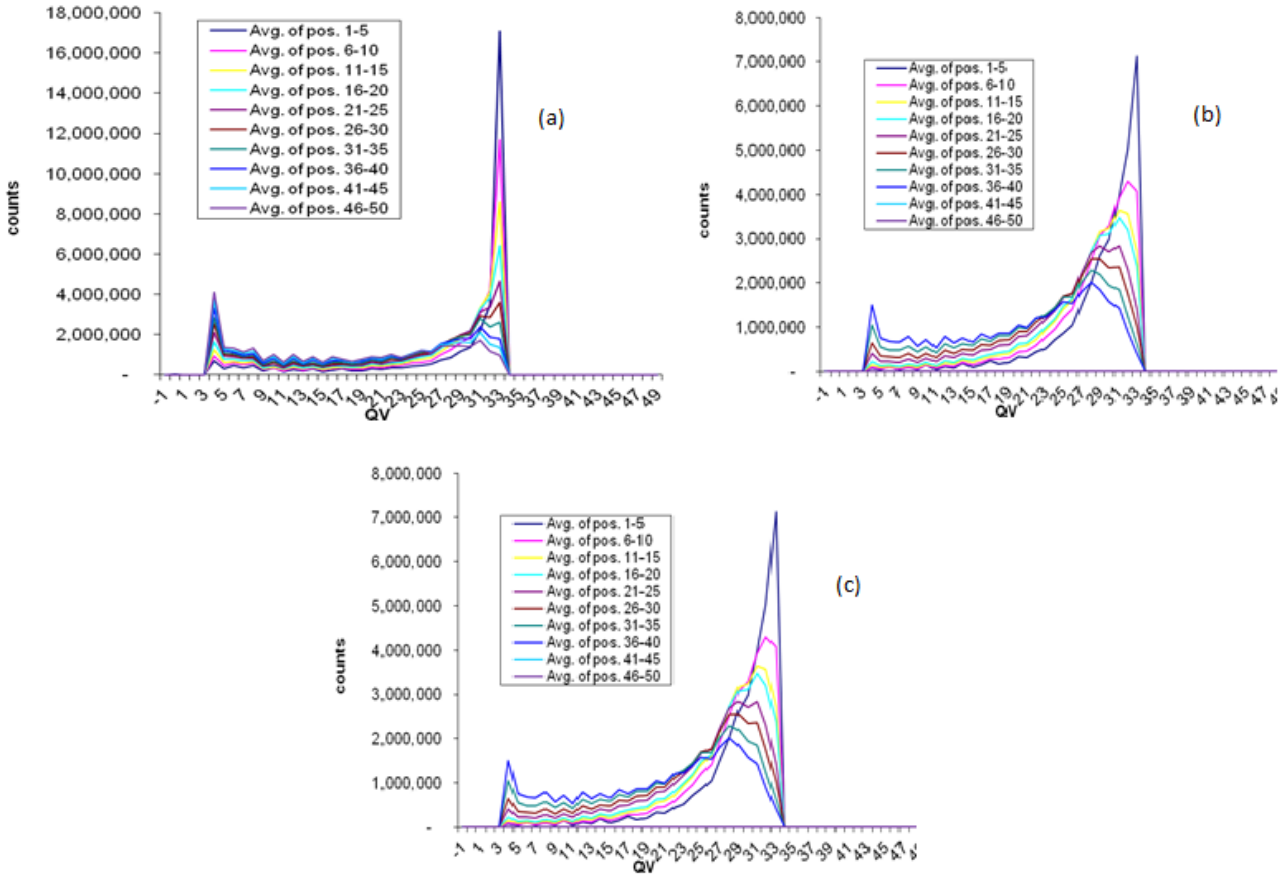



Figure 3.4: Plot showing the relationship between position in sequence read and quality value (isolate 0140J) (a) All reads (b) Reads passing quality value assessment (c) Reads failing quality value assessment.

assembler including Velvet [228]. The optimization of K-mer depends on a number of other factors such as coverage, read length and error rates. The recommended K-mer is usually greater than half of the read length but it is hard to determine the optimum K-mer in advance and so this was determined empirically. The read length available from SOLiD platform was 50 bp and the effect of different K-mer ranging from 25-33

(odd values are accepted by velvet) was initially used to determine the optimum K-mer (Table 3.3). For the rest of the isolates the initial K-mer 29 was used then the K-mer value was increased or decreased one odd value until the N50 value started to decrease. For the majority of genomes the longest N50 was obtained when a K-mer of 29 was used.

Table 3.3: Effect of K-mer on N50 values of twenty isolates using SOLiD reads

| Strain ID | K-mer 25 | K-mer 27 | K-mer 29 | K-mer 31 |
|---|---|---|---|---|
| 0140J | 1546 | 1658 | 1760 | 1639 |
| B362 | 1273 | 1484 | 1562 | 1510 |
| 4244 | 1764 | 1759 | 1867 | 1898 |
| 4428 | 1650 | 1699 | 1754 | 1780 |
| C5072 | 1606 | 1973 | 1910 | 1763 |
| 5130 | 1541 | 1661 | 1633 | 1581 |
| 5291 | 1254 | 1037 | 1096 | 1037 |
| C5388 | 1399 | 1730 | 1679 | 1683 |
| 6736 | 1159 | 1242 | 1207 | 1184 |
| 6707 | 1609 | 1631 | 1757 | 1635 |
| 6780 | 1423 | 1728 | 1678 | 1543 |
| S7010 | 1322 | 1527 | 1726 | 1311 |
| C7131 | 1625 | 1965 | 1870 | 1754 |
| Ab71 | 1693 | 1815 | 1902 | 1868 |
| Ab1947 | 1401 | 1357 | 1254 | 1217 |
| B190 | 1505 | 1808 | 1765 | 1688 |
| S6261 | 1573 | 1809 | 1841 | 1818 |
| C8329 | 1535 | 1938 | 1907 | 1755 |
| C9359 | 1691 | 1937 | 1832 | 1736 |
| 0145/EF20 | 1489 | 1642 | 2045 | 1994 |

Coverage distribution is another important parameter and after filtering high error reads all the strains had coverage of more than 50X (Fig 3.5) (Table 3.4), which was sufficient for *de novo* assembly. *De novo* assembly of short read sequence reads yielded sufficient quality draft genomes for the twenty two *S. uberis* strains. The assembly statistics details are provided in Table 3.5. As one strain (0140J) has previously been sequenced by the Sanger di-deoxy method [216], this could be used to directly evaluate the quality of the

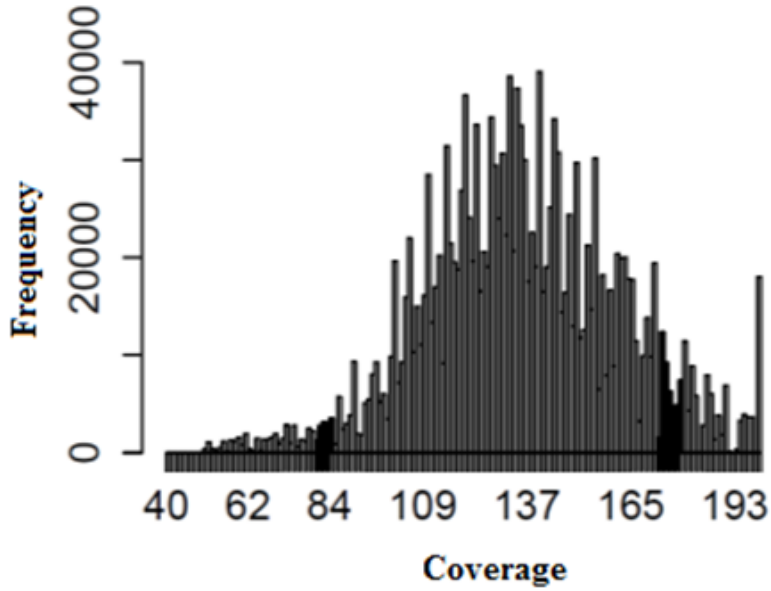



Figure 3.5: Coverage distribution of filtered SOLiD short reads of the genome of 0140J. The figure was generated by R script provided with the Velvet software. Maximum coverage is between 130-140 fold.

0140J genome assembled using short reads. A high degree of agreement (96%) was seen between the assembled and reference genomes size (1.74 Mbp to 1.9 Mbp) and reference genome (1.85 Mbp).

Table 3.4: Optimum coverage distribution for different isolates used during Velvet assembly (SOLiD).

| Strain ID | Expected Coverage | Coverage Cut-off |
|---|---|---|
| 0140J | 200 | 50 |
| 0145/EF20 | 160 | 75 |
| B362 | 165 | 70 |
| 3448 | 150 | 60 |
| 4244 | 210 | 90 |
| 4428 | 180 | 75 |
| C5072 | 140 | 60 |
| 5130 | 150 | 50 |
| 5291 | 113 | 40 |
| C5388 | 189 | 50 |
| 6736 | 60 | 40 |
| 6707 | 165 | 75 |
| 6780 | 227 | 100 |
| S7010 | 169 | 75 |
| C7131 | 150 | 50 |
| Ab71 | 175 | 87 |
| Ab1947 | 155 | 80 |
| B190 | 158 | 70 |
| S6261 | 185 | 70 |
| C6344 | 150 | 70 |
| C8329 | 140 | 67 |
| C9359 | 100 | 40 |

### 3.2.2 Genome sequencing and assembly using Illumina MiSeq

To supplement the SOLiD reads, we performed Illumina sequencing of the genomes of twelve *S. uberis* strains among the isolates already sequenced by the SOLiD V4 platform. The advantage of Illumina over the SOLiD sequencing method includes greater read length and advanced tools for *de novo* assembly which should result in significantly larger contigs and more accurate analysis of genome/gene variations.

#### 3.2.2.1 FASTQC Quality analysis of Illumina reads

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to assess the quality of MiSeq paired end reads. Figure 3.6 and 3.7 show the high quality reads available from Illumina platform.

#### 3.2.2.2 Filtering error of Illumina reads

As discussed, removing low-quality reads can significantly improve the *de novo* assembly. Due to the significantly higher quality reads (Q-score >30) available from Illumina MiSeq (Fig 3.6, Fig 3.7) error filtering was not conducted prior to *de novo* assembly.

#### 3.2.2.3 K-mer and coverage optimization

Read lengths from MiSeq paired-end reads were 250 bp and K-mers ranging from 149-275 were tested with an interval of 10 until the highest N50 and fewest contigs numbers achieved. The interval narrowed down to 5 when it was found that the best assembly was available in the K-mer range between 175-225 (Fig 3.8). To achieve a high-quality

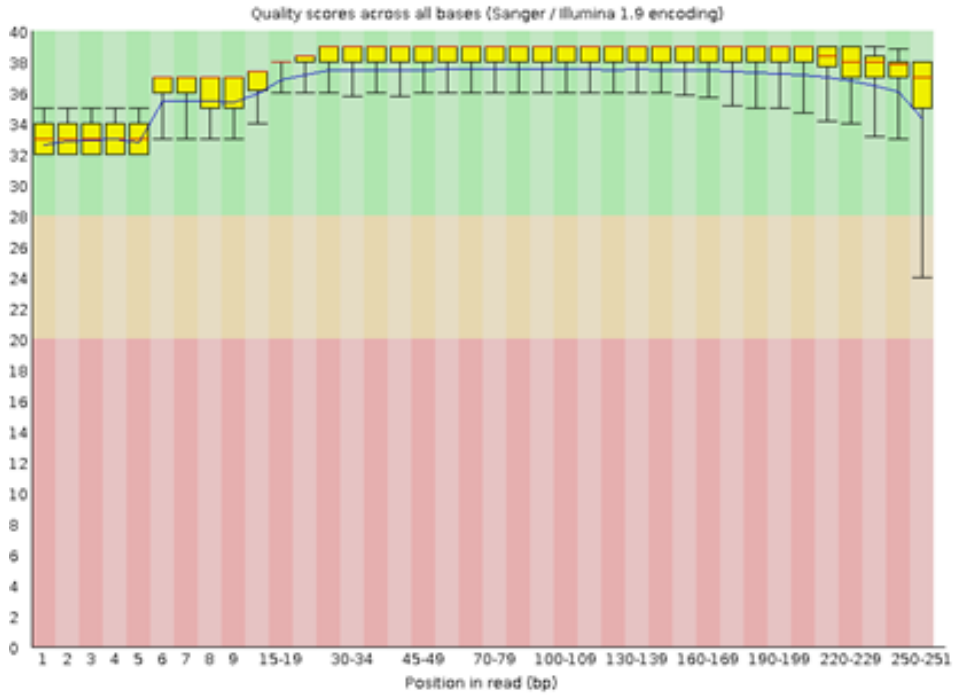



Figure 3.6: Per base quality value of Illumina reads (EF20). A box-and-whisker plot is plotted for each read position. The central red line is the median value. The yellow box represents the inter-quartile range (25-75%). The upper and lower whiskers show minimum and maximum values.

*de novo* assembly with Illumina reads it is important to have a sufficient number of error-free reads covering the whole genome. Due to the improved technology generating longer paired-end MiSeq data, high quality reads with sufficient quality was available for *de novo* assembly. The coverage required depends on a number of factors including the organism, genome size and repeat regions. For the Illumina reads the optimum assembly was available using the expected coverage option auto and the minimum coverage of 10x. As can be seen in the Fig 3.9 coverage of most of the reads are ~x25 unlike SOLiD data which have a wider distribution. A number of basic statistics were calculated for contigs of SOLiD and Illumina assembly, including the N50, N25 and N75 (Table 3.5).

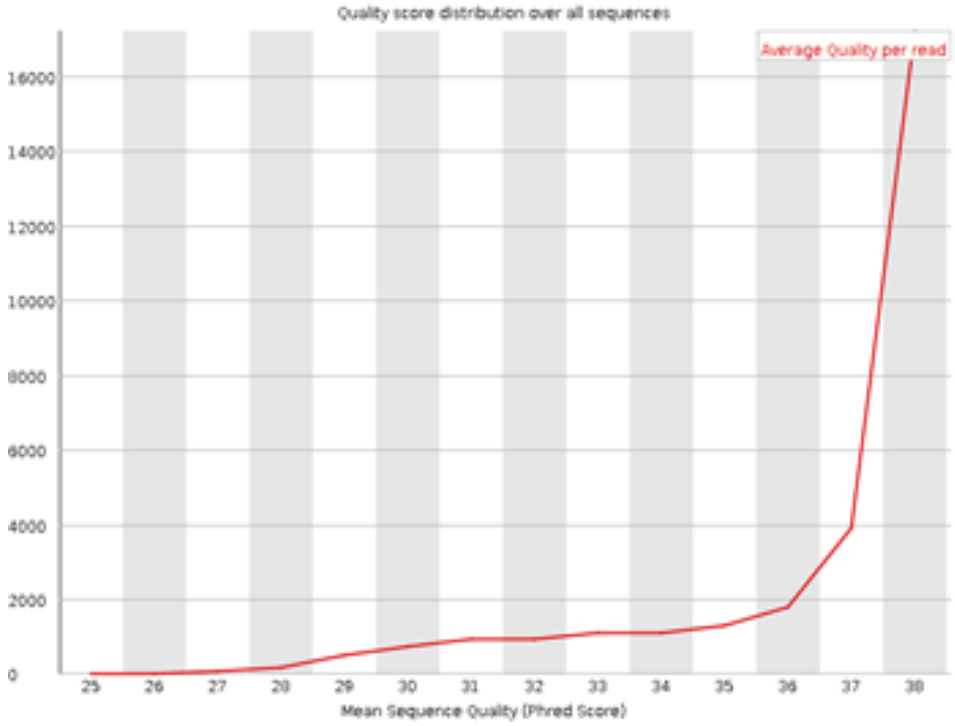



Figure 3.7: The per sequence quality score. This shows the subset of EF20 sequences have universally high quality values with most of the reads with QV > 30.

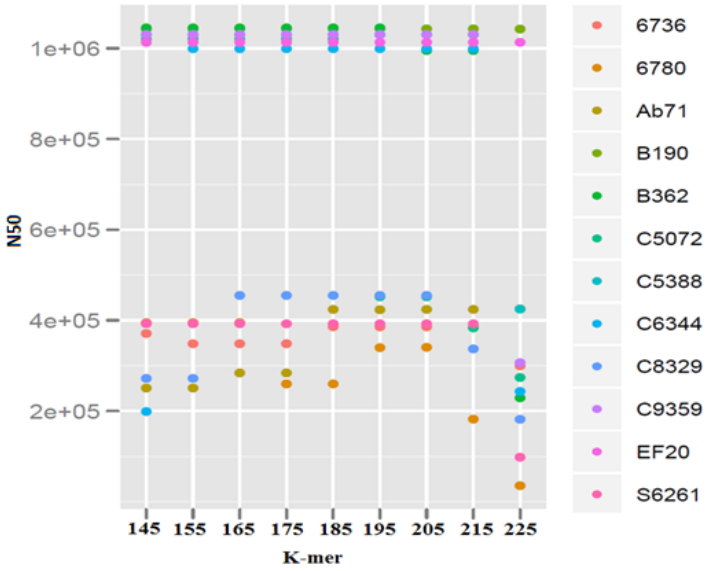



Figure 3.8: Plot showing the affect of K-mer length on N50 of Illumina read assemblies.

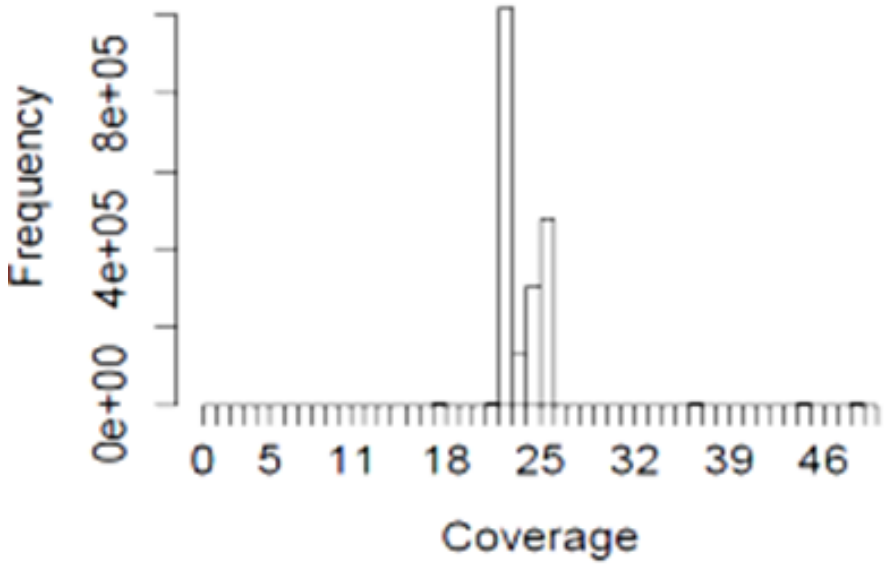



Figure 3.9: Coverage distribution of illumina reads of the genome EF20. Maximum expected coverage is approximately 23.

It can be seen (Table 3.5, 3.6 and Fig 3.10) that the largest genome size available from SOLiD data was 19,07,549 bp with the mean size 1,826,318 bp and minimum size of 1,740,197bp. Conversely, for Illumina the minimum was 1,842,024 with mean genome size of 1,895,485 bp. SOLiD assemblies were highly fragmented with a larger number of contigs compared to the Illumina. For all assemblies, the mean number of contigs >=100 bp for SOLiD reads was 1,098 much higher than the mean number of contigs (21) for the Illumina platform. Mean N50 statistics of Illumina was 713,199 bp and the largest N50 found 102,3474 bp , however, they were 1,808 bp and 2,185 bp obtained from SOLiD data also represent much better assembly obtained from Illumina platform (Fig 3.11). As noted by Yandell et al. [223] an assembly with a 'gene-sized' scaffold N50 length may be a good target for annotation [223]. Using existing genome prediction tools using *ab*

*initio* or *de novo* methods of gene prediction [21, 36] the assemblies available from the SOLiD single end reads even those with significantly lower N50 are useful in predicting the gene repertoire of isolates under study since the average gene size of the reference genome 0140J is 300 amino acids.

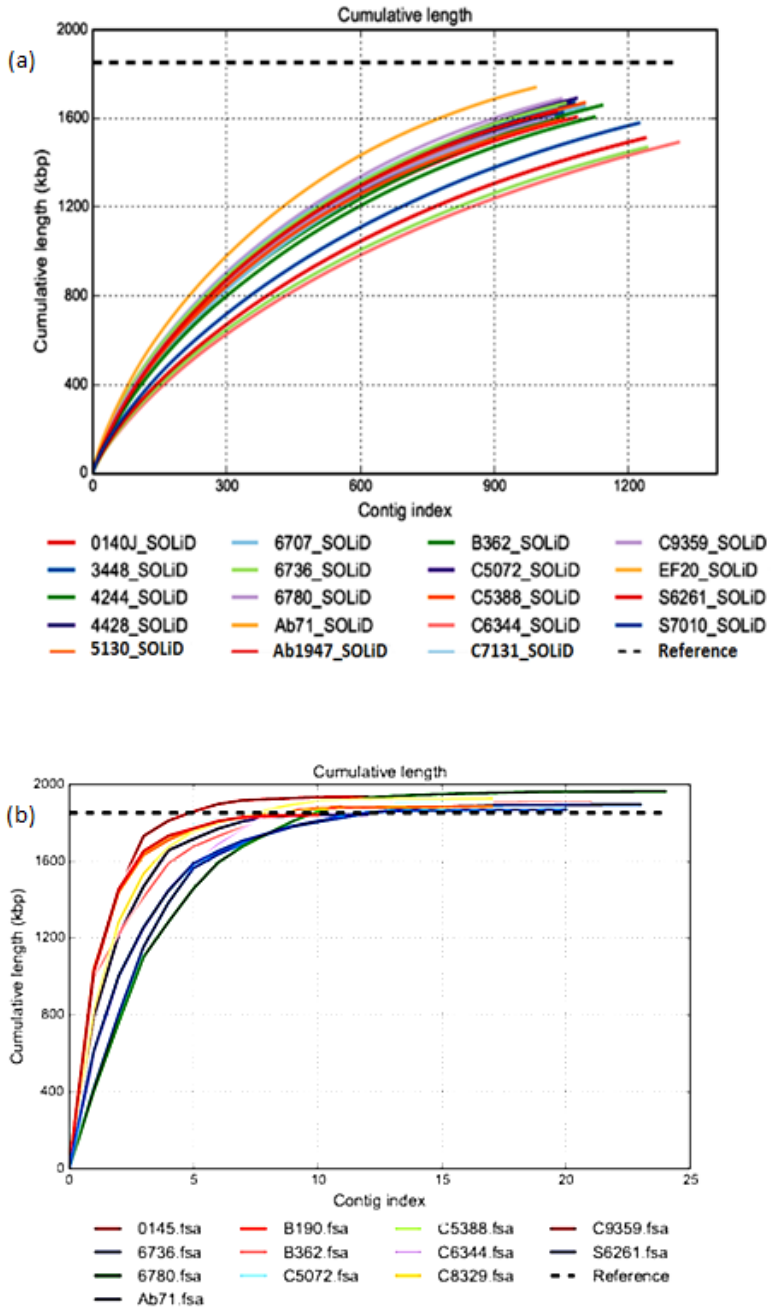



Figure 3.10: Distribution of cumulative lengths of assembled genomes using (a) SOLiD and (b) Illumina reads. The dashed line represents the length of the reference genome 0140J.

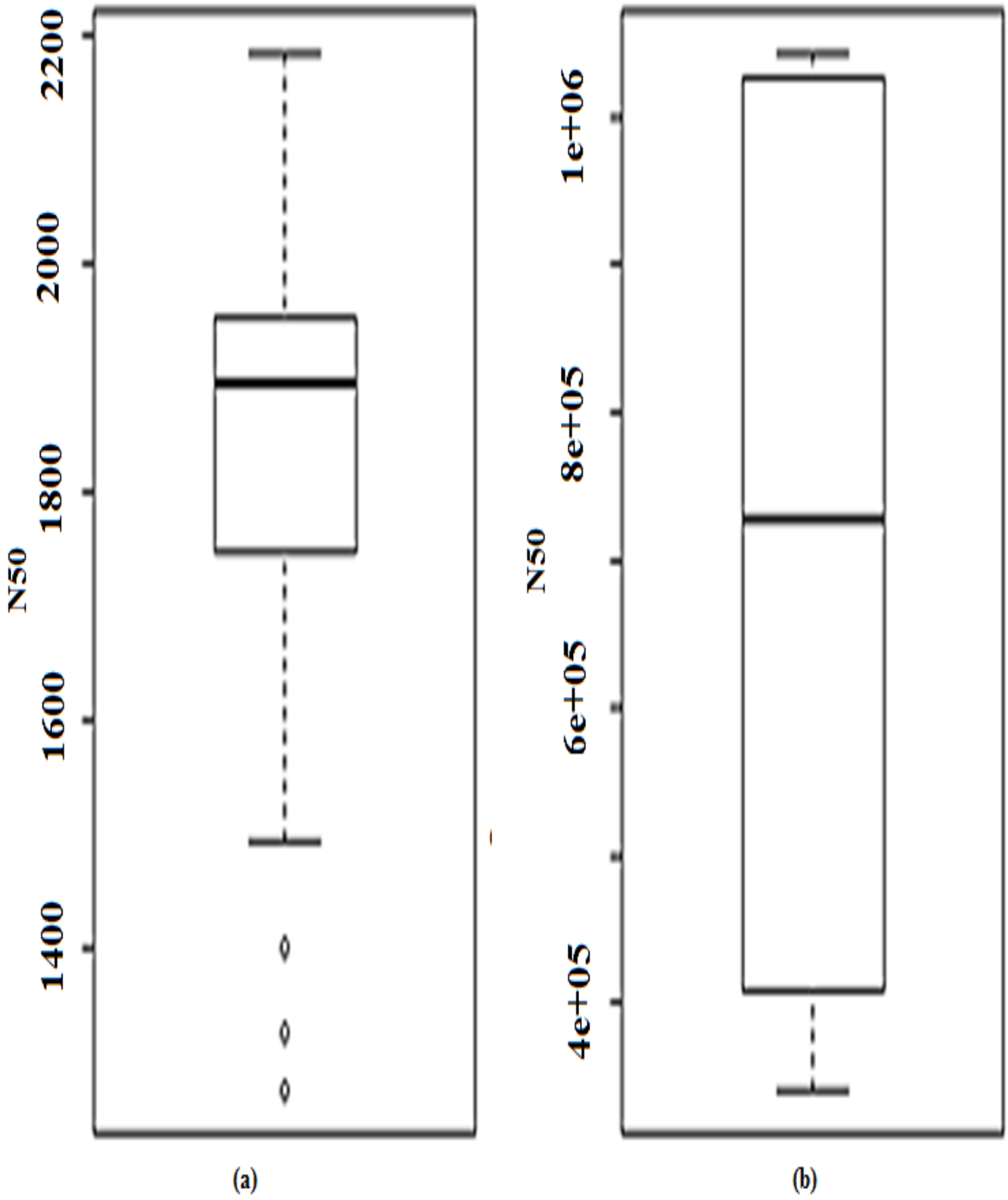



Figure 3.11: Comparison of N50 distribution of (a) SOLiD and (b) Illumia assemblies. For description of N50 see section 2.4.

Table 3.5: *De novo* assembly statistics of twenty *S. uberis* genomes using SOLiD reads. N50, N75- the length at which 50% and 75% of assembled bases belong to a contig of this length or longer, respectively; L50, L75- number of contigs whose combined length is N50 and N75, respectively.

| Strain | Contigs (≥ 0 bp) | Contigs (≥ 1000 bp) | Total length (≥ 0 bp) | Total length (≥ 1000 bp) | contigs | Largest contig | Total length | GC (%) | N50 | N75 | L50 | L75 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0140J | 1815 | 647 | 1781770 | 1289316 | 1084 | 7007 | 1605914 | 37.14 | 1793 | 1116 | 283 | 567 |
| EF20 | 1560 | 650 | 1876586 | 1487353 | 992 | 9642 | 1740291 | 36.87 | 2185 | 1379 | 247 | 497 |
| 6736 | 2240 | 589 | 1742897 | 997082 | 1242 | 5339 | 1471038 | 37.47 | 1327 | 887 | 364 | 702 |
| 6780 | 1643 | 661 | 1846049 | 1402859 | 1050 | 11226 | 1689374 | 37.3 | 1990 | 1215 | 268 | 540 |
| Ab71 | 1552 | 632 | 1740197 | 1334608 | 987 | 8437 | 1592177 | 37.2 | 2005 | 1258 | 257 | 507 |
| B190 | 1731 | 663 | 1792856 | 1338941 | 1053 | 8430 | 1627592 | 37.19 | 1878 | 1170 | 277 | 554 |
| B362 | 2049 | 654 | 1825477 | 1263611 | 1124 | 8721 | 1606451 | 37.23 | 1654 | 1090 | 303 | 598 |
| C5072 | 1747 | 679 | 1847357 | 1391518 | 1084 | 8999 | 1691053 | 37.11 | 1912 | 1189 | 288 | 567 |
| C5388 | 2096 | 633 | 1800147 | 1292624 | 1056 | 8608 | 1597924 | 37.17 | 1895 | 1136 | 272 | 545 |
| C6344 | 1678 | 649 | 1791388 | 1348144 | 1038 | 8109 | 1631647 | 37.25 | 1979 | 1198 | 271 | 536 |
| C8329 | 1719 | 669 | 1837690 | 1384563 | 1059 | 14622 | 1669547 | 37.28 | 1911 | 1185 | 271 | 548 |
| C9359 | 1750 | 643 | 1778212 | 1323273 | 1032 | 8856 | 1607203 | 37.19 | 1898 | 1179 | 267 | 535 |
| S6261 | 1678 | 649 | 1791388 | 1348144 | 1038 | 8109 | 1631647 | 37.25 | 1979 | 1198 | 271 | 536 |
| 3448 | 1910 | 667 | 1863332 | 1364140 | 1079 | 7511 | 1671743 | 37.12 | 1914 | 1168 | 287 | 566 |
| 4244 | 1963 | 683 | 1860533 | 1325767 | 1141 | 6853 | 1659630 | 37.15 | 1747 | 1096 | 302 | 606 |
| 4428 | 1752 | 656 | 1837403 | 1371239 | 1069 | 7475 | 1671740 | 37.04 | 1954 | 1182 | 275 | 549 |
| 5130 | 2226 | 684 | 1907549 | 1366140 | 1102 | 9854 | 1670471 | 37.11 | 1834 | 1183 | 297 | 581 |
| 5291 | 3075 | 586 | 1888484 | 970268 | 1313 | 6747 | 1492964 | 37.3 | 1276 | 850 | 389 | 748 |
| 6707 | 1858 | 685 | 1841426 | 1344130 | 1102 | 7505 | 1655869 | 36.95 | 1794 | 1129 | 302 | 589 |
| Ab1947 | 2362 | 622 | 1805498 | 1069277 | 1238 | 7074 | 1512460 | 37.22 | 1402 | 922 | 359 | 691 |
| C7131 | 2050 | 648 | 1864319 | 1368299 | 1039 | 12629 | 1655291 | 37.15 | 1953 | 1196 | 264 | 533 |
| S7010 | 2557 | 650 | 1858439 | 1161313 | 1224 | 6815 | 1579604 | 37.23 | 1494 | 975 | 346 | 674 |

Table 3.6: *De novo* assembly statistics of twelve *S. uberis* genomes using Illumina MiSeq reads.

| Strain | GenBank Accession No. | Contigs (≥ 0 bp) | Contigs (≥ 1000 bp) | Total length (≥ 1000 bp) | contigs | Largest contig | Total length | GC (%) | N50 | N75 | L50 | L75 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EF20 | JANW00000000 | 19 | 10 | 1934633 | 12 | 1013731 | 1936305 | 36.3 | 1013731 | 280552 | 1 | 3 |
| 6736 | JATB00000000 | 29 | 18 | 1893211 | 23 | 420979 | 1896613 | 36.46 | 348833 | 179842 | 3 | 5 |
| 6780 | JATD00000000 | 27 | 19 | 1961775 | 24 | 404720 | 1965197 | 36.5 | 340886 | 137468 | 3 | 6 |
| Ab71 | JATK00000000 | 13 | 10 | 1841891 | 12 | 792968 | 1843211 | 36.46 | 425644 | 251093 | 2 | 3 |
| B190 | JATE00000000 | 21 | 12 | 1881972 | 17 | 1043356 | 1885693 | 36.47 | 1043356 | 396769 | 1 | 2 |
| B362 | JATC00000000 | 32 | 14 | 1908138 | 21 | 995597 | 1913450 | 36.4 | 995597 | 175855 | 1 | 4 |
| C5072 | JATI00000000 | 20 | 13 | 1908058 | 17 | 1021112 | 1910583 | 36.53 | 1021112 | 452296 | 1 | 2 |
| C5388 | JATF00000000 | 14 | 9 | 1841480 | 10 | 1030559 | 1842024 | 36.52 | 1030559 | 425114 | 1 | 2 |
| C6344 | JATA00000000 | 16 | 12 | 1910975 | 13 | 539809 | 1911734 | 36.51 | 459853 | 215892 | 2 | 4 |
| C8329 | JATG00000000 | 21 | 13 | 1925091 | 17 | 830811 | 1927934 | 36.61 | 455310 | 250943 | 2 | 3 |
| C9359 | JATJ00000000 | 12 | 9 | 1841579 | 10 | 1030658 | 1842338 | 36.52 | 1030658 | 425114 | 1 | 2 |
| S6261 | JATH00000000 | 27 | 16 | 1867900 | 20 | 613122 | 1870741 | 36.51 | 392850 | 189494 | 2 | 4 |

### 3.2.3 Multi locus sequence typing

To determine the relationship among the isolates under study, seven housekeeping genes (*arc, ddl, gki, recP, tdk, tpi, yqiL*) used for MLST profiling of *S. uberis* (http://pubMLST.org/suberis/) were concatenated for each isolate and a phylogenetic tree was constructed using PhyML (Fig 3.12) [81]. Sequence Typing (ST) is a very sensitive method used in MLST profiling and due to the high error rate in the SOLiD assembly the phylogeny or topology of the tree was determined to be unreliable as a dendogram using *de novo* assembled data will not reflect the true pattern of descent. For this reason genes obtained from *de novo* assembly of Illumina reads were used for MLST analysis. Although analysis of the phylogeny of the 13 isolates did not show clustering according to their recorded clinical and sub-clinical status, they resulted in 2 major clusters with isolates B362, 6780 and Ab1947 clustered with the non-virulent isolate EF20. Suggesting a possible independent evolutionary lineage of the less virulent strain.

goEBURST analysis (Fig 3.13) shows the distribution of isolates compared to the other MLST profiles of other UK strains available in *S. uberis* MLST database.

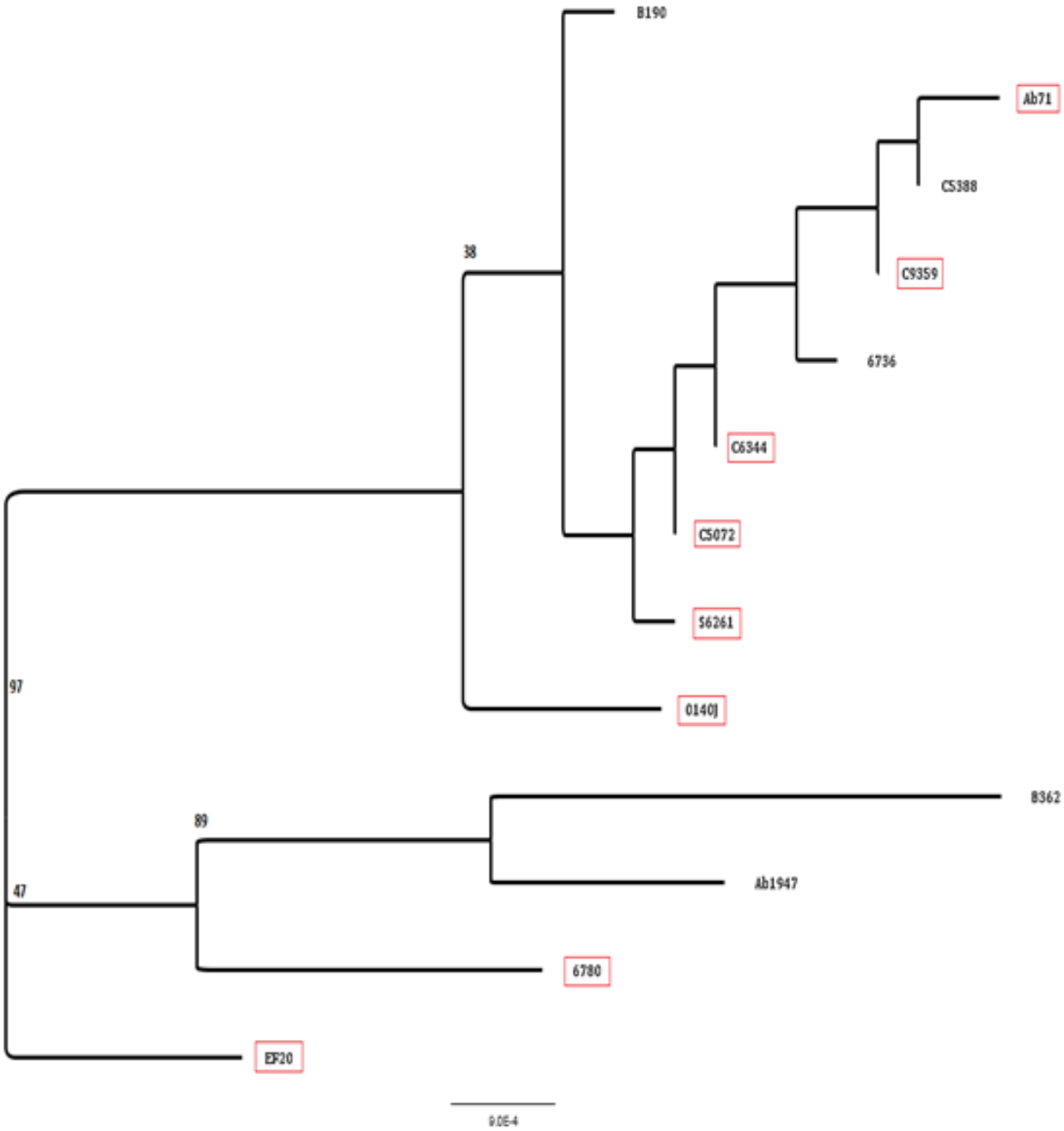



Figure 3.12: Phylogenetic tree based on concatenated MLST genes determined by PhyML [80] using GTR model. 200 bootstraps were conducted. Strains initially recorded as clinical isolates are boxed in red.

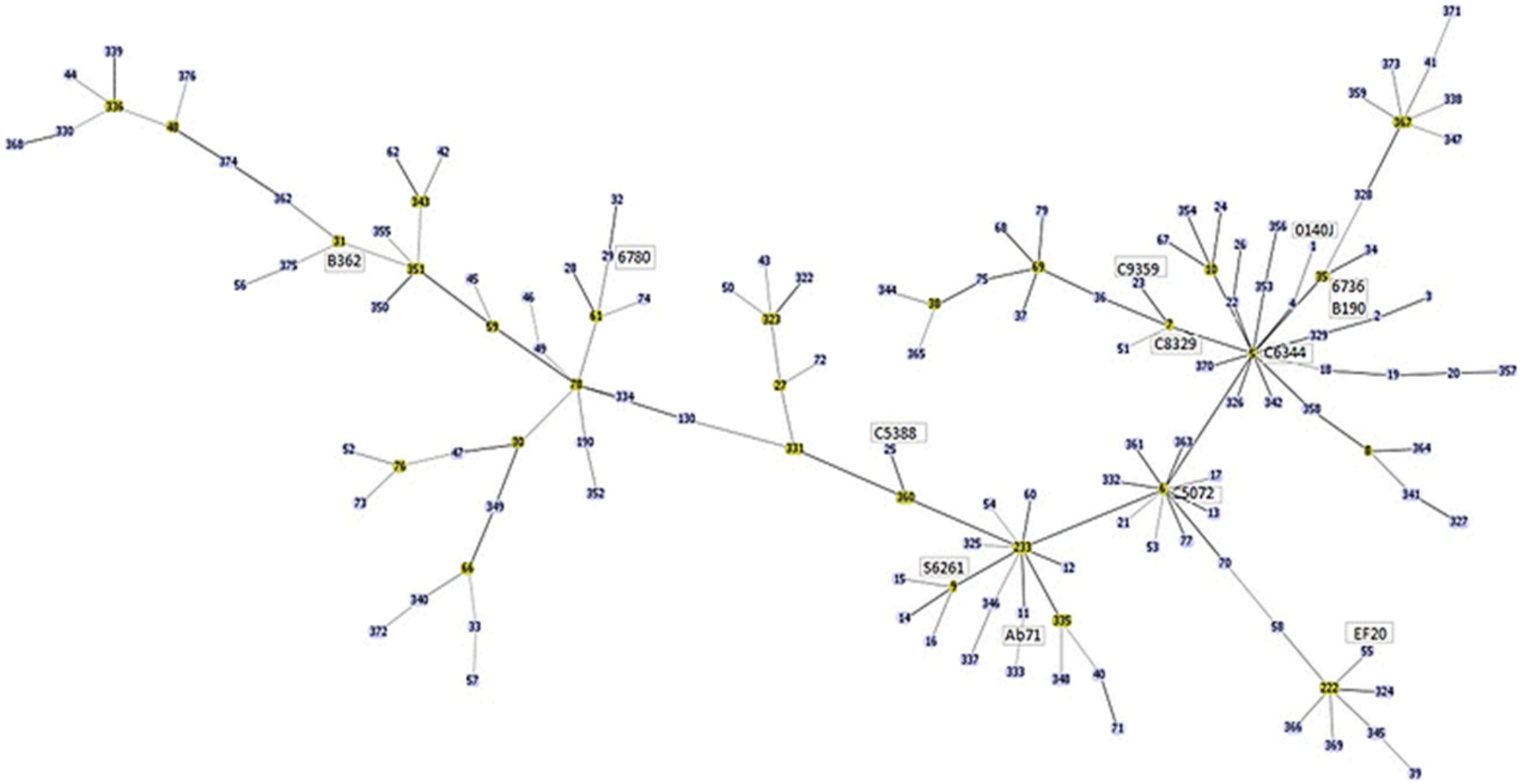



Figure 3.13: MLST distribution of 13 isolates compared to the other UK strains available in the PubMLST database. MLST profiles of UK *S. uberis* isolates were downloaded from PubMLST database (http://pubmlst.org/suberis/) and analysed using the goeBURST option of PHYLOViZ [68]. Each ST is represented as a blue node, edges connect nodes with the greatest number of shared variants. Yellow nodes indicate predicted founder sequence types.

## 3.3 Discussion

At the beginning of the study the *de novo* assembly of bacterial genomes using short reads generated by SOLiD colourspace reads was an intriguing problem. Until we developed the pipelines for *de novo* assembly using SOLiD short reads there was no report of an entire bacterial genome assembly using only colorspace short reads. In fact, the shortness of the reads available from the system was the major problem for *de novo* assembly and this shortcoming was usually mitigated by hybrid assembly using either 454 platform (longer reads) or using traditional Sanger sequencing along with short reads. In this study we developed a strategy that enabled assembly of twenty two draft bacterial genomes isolated from different clinical status of mastitis. We have shown that short reads from the SOLiD platform can facilitate *de novo* assembly of contigs and analyses of bacterial genomes.

Compared to Illumina, the SOLiD platform does produce a larger number of fragmented contigs and hence gaps in the genome. Additional concerns for the use of SOLiD reads are the general lack of tools which could be used to assemble the genomes using color-space reads. Here we used Velvet. Although Velvet is assumed not to consider the sequencing error during assembly, we found the assembly improved significantly when quality filtering was applied prior to assembly. Moreover, it helped to overcome memory constraints and it was possible to assemble the whole genome using only 24 GB RAM. With these improvements and optimization of K-mer length, the quality of the draft genome was only sufficient for annotation of the core and accessory genes.

Illumina MiSeq paired-end reads produced a better assembly compared to SOLiD data.

Assembly using Illumina reads was also less memory greedy and significantly less time consuming compared to the SOLiD data. It took >5 hours to assemble SOLiD reads whereas the assembly was done within 10 minutes using MiSeq data. However, the mean contig length for most of the genomes were ≈1kbp, sufficient to identify gene/CDS using similarity searches or annotation software for SOLiD data. NGS short reads when used with paired-end libraries from Illumina Miseq have been shown to have potential for high quality draft genome assembly; resulted in approximately 1000 fold.

The above results make it clear that assemblies using MiSeq paired reads are substantially superior to what can be accomplished using SOLiD V4 sequencing. Since Velvet is the only known open source tool for colorspace assembly this was also used for comparative analysis of two NGS systems. Illumina reads had substantially lower coverage and coverage requirements could be reduced by using longer reads using other NGS such as paired-end 454 sequencing with deeper coverage Illumina reads. The major objectives of the present study was to find the variation in the genomic content among the clinical and sub-clinical isolates, therefore, the high quality contigs generated using Illumina reads were considered appropriate here.

To determine the relatedness among different isolates we generated phylogeny of isolates assembled using Illumina reads based on concatenated seven genes used for MLST profiling of *S. uberis* (http://pubmlst.org/suberis/). Phylogenetic analysis using three random loci of MLST genes using maximum likelihood showed different topologies when different combination of loci used [38]. Moreover, the bootstrap values demonstrates low resolution in the trees suggesting non-clonal population structure as well as the

evolution of *S. uberis* through recombination as found in other studies [38].

For Illumina sequencing we have chosen isolates from different STs and mostly ST-5 whether we can differentiate between clinical and sub-clinical isolates or to decipher if there is any differences among the ST-5 types according to the clinical status of the diseases. As found in previous studies [38, 168] we could not reveal any clear association between disease status and STs using MLST based phylogeny and this could be explained by the fact that in MLST only seven housekeeping genes are used and these loci are highly conserved among the isolates. However, clustering of three isolates with the known non-virulent isolate EF20 using concatenated MLST phylogey implies that these isolates are likely to show reduced virulence. Phylogeny and comparative analysis based on whole genome repertoire that includes other virulence and pathogen specific variable genes may provide greater insight into the pathogenesis of *S. uberis*.

CHAPTER 4

# COMPARATIVE GENOMICS OF *Streptococcus uberis* ISOLATES OF DIFFERING VIRULENCE

## 4.1 Introduction

Whole genome sequencing (WGS) has enabled the construction of hundreds of complete/draft microbial and eukaryotic genomes as a result of the advancement and affordable cost of NGS technology [127]. Rapid development and tremendous throughput of NGS technology has also greatly influenced the study of bacterial comparative genomics, moving from simple comparison of two or few genomes to hundreds of genomes. Whilst the generation of data has accelerated, genomic comparisons have become more complicated due to the enormous amount of data and the different error generated by different technologies. However, the major goal of bacterial comparative genomics remains to

identify differences among genomes, and to elucidate the differences among the virulent and non-virulent organisms to find the genomic traits responsible for pathogenesis as well as link genotype to the virulence phenotype of the pathogen in genomic scale.

Although sequence differences may alone not provide sufficient information on the specific gene/genomic regions responsible for phenotype or virulence, comparative genomics may still provide useful information on the genomic regions and candidate genes responsible for pathology of *S. uberis*. Whilst over 900 strains of *S. uberis* have been typed using MLST (http://pubMLST.org/suberis/), at the outset of this research, only a single *S. uberis* genome sequence had been reported, strain 0140J (accession number AM946015), selected as a typical virulent UK strain [216].

Moreover, development of a vaccine against *S. uberis* has been hampered by a lack of information on the interaction between pathogen and the host [120]. This lack of knowledge is exemplified by the paucity of information on *S. uberis* strains at the genomic level. The genome of 0140J (1,852,352bp) is one of the smallest Streptococcus genomes that range from 1.8 Mb-2.3Mb [216]. This suggests that through genome reduction, the 0140J genome has become condensed possibly reflecting restricted host-range. It is also possible that the 1,825 protein coding genes of 0140J harbour potential virulence genes which are absent in non-virulent strains, or that loss of accessory genes present in other strains may be associated with the virulence of 0140J.

As an initial attempt to rectify this lack of genomic information and to identify the extent of genome variation between *S. uberis* strains, the genomes of additional strains were determined using high throughput sequencing approaches. As discussed in Chapter 3,

the quality of the genome assembly available from SOLiD datasets is significantly lower compared to the Illumina assembly and for this reason only the twelve isolates sequenced by Illumina MiSeq were considered for comparative genomics study and discussed in this chapter. The strains selected for sequencing are representative of the currently typed UK strains. Comparison of the predicted gene content was performed to identify the core-genome shared by strains and the accessory genome variable between strains.

## 4.2 Results

### 4.2.1 General features of *Streptococcus uberis* genomes

The combined length of the assembled contigs range from 1,837,707 bp to 1,960,858 bp. This range spans the 1,852,352bp of the 0140J reference strain and falls within the 1.8 Mb-2.3Mb predicted by Ward et al. [216]. The median size of the genomes from each group differ (clinical median length 1,887,233.5 bp, sub-clinical median length 1,903,098.5 bp). However, the variation within both groups is large (interquartile range clinical: 74,423.5 bp, sub-clinical: 55,387.5 bp) and as a result there is no significant difference of total genome size (as measured by total assembled contig size) between the two groups. Therefore, this does not support the assumption that all clinical strains would exhibit a reduced size genome like 0140J. This may reflect the relatively loose definition of clinical vs sub-clinical strains.

In a comparative animal model study (Chapter 1 Section 1.3) it was shown that 0140J the well-studied virulent UK isolate was able to cause disease while the isolate EF20

was consistently failing to cause disease in the same model thus substantiating and adding detail to the previous data [88] regarding the nature of virulence of these two strains and confirming their suitability for direct comparison at the genomic level in the elucidation of virulence related features.

The draft genome of EF20 is slightly larger compared to the genome of 0140J. The 1,933,244 bp assembled draft genome of EF20 has a G+C content of 36.5% comparable to the 36.6% G+C content in 0140J. In line with its larger genome, EF20 has an increased number of predicted coding sequences of 1,957 compared to the 1,825 of 0140J. At the gross level, comparative analysis of EF20 and 0140J revealed high-level synteny disrupted by a large number of gene gain/loss and recombination events.

The larger predicted gene count in the EF20 genome is not simply addition within EF20 and/or loss in 0140J. Whilst 1,629 annotated genes are common between the two strains, strain specific genes are present in both 0140J (145 genes present in 0140J not in EF20) and EF20 (222 genes present in EF20 not in 0140J) (Additional file 4.1: orthologous clustering) (Table 4.1 and 4.2).

Table 4.1: Genes present in 0140J and absent from EF20.

| gene id | function |
|---|---|
| *SUB0348* | putative surface-anchored protein (pseudogene) |
| *SUB1059* | hypothetical protein |
| *SUB1060* | hypothetical protein |
| *SUB1660* | hypothetical protein |
| *SUB1835* | hypothetical protein |
| *SUB0301* | membrane protein |
| *SUB0511* | putative membrane protein |
| *SUB1061* | hypothetical protein |
| *SUB1185* | putative DNA |

Continued on next page

Table 4.1 (continued from previous page)

| Gene id | Function |
|---|---|
| *SUB0175* | microcin immunity protein |
| *SUB0498* | response regulator protein |
| *SUB0499* | sensor histidine kinase |
| *SUB0502* | putative bacteriocin (pseudogene) |
| *SUB0505* | bacteriocin |
| *SUB0510A* | conserved hypothetical protein (fragment) |
| *SUB0701* | membrane protein |
| *SU1099* | cation transporting ATPase (fragment) |
| *SUB1180* | putative copper chaperone CopZ |
| *SUB1181* | copper importing ATPase |
| *SUB1182* | hypothetical protein |
| *SUB1183* | Crp family regulatory protein |
| *SUB1186* | putative DNA recombinase (fragment) |
| *SUB1187* | modification methylase |
| *SUB1188* | type III site |
| *SUB1438A* | transposase (pseudogene) |
| *SUB1522* | 3-hydroxybutyryl-CoA dehydrogenase (fragment) |
| *SUB1746* | hypothetical protein |
| *SUB1821* | hypothetical protein |
| *SUB0295* | membrane protein |
| *SUB0296* | ABC transporter ATP binding protein |
| *SUB0297* | membrane protein |
| *SUB0298* | membrane protein |
| *SUB0300* | hypothetical protein |
| *SUB1676* | sensor histidine kinase |
| *SUB1710* | CHY zinc finger protein |
| *SUB0283* | hypothetical protein |
| *SUB0702* | membrane protein |
| *SUB1677* | sensor histidine kinase (pseudogene) |
| *SUB0052* | membrane protein |
| *SUB0279* | type I restriction modification system M protein |
| *SUB0282* | type I restriction modification system R protein |
| *SUB0575* | hypothetical protein |
| *idnO* | Gluconate 5 dehydrogenase |
| *SUB0284* | DeoR family regulatory protein |
| *SUB0285* | ribulose phosphate 3 epimerase |
| *SUB0286* | 6 phospho 3 hexuloisomerase |
| *SUB0287* | Mannitol 1 phosphate 5 dehydrogenase |
| *SUB0288* | PTS system sugar transporter subunit IIA |
| *SUB0293* | acetoin(diacetyl) reductase (pseudogene) |

Continued on next page

Table 4.1 (continued from previous page)

| Gene id | Function |
|---|---|
| *SUB1184* | hypothetical protein |
| *SUB0018* | membrane protein |
| *SUB0051* | hypothetical protein |
| *SUB0058* | membrane protein |
| *SUB0147* | prophage maintenance system killer protein |
| *SUB0164* | putative fusion protein (pseudogene) |
| *SUB0292* | hypothetical protein |
| *SUB0478* | membrane protein |
| *SUB0031* | response regulator protein (pseudogene) |
| *SUB0032* | *ublA*: bacteriocin; Cyclopeptide antibiotic with bacteriolytic activity against most streptococci (except *S.rattus* and *S.mutans*), Listeria spp., enterococci and staphylococci |
| *SUB0033* | ublB:bacteriocin: membrane protein |
| *SUB0034* | *ublC*:bacteriocin:membrane protein |
| *SUB0035* | *ublD*:bacteriocin:ABC transporter ATP binding protein |
| *SUB0278* | hypothetical protein |
| *SUB0504* | putative membrane protein (pseudogene) |
| *SUB0772.1* | hypothetical protein |
| *SUB1029* | mna UDP N acetylgclucosamine epimerase |
| *SUB1063* | haloacid dehalogenase |
| *SUB1598* | hypothetical protein |
| *SUB1600* | putative transposase (fragment) |
| *SUB1601* | keto hydroxyglutarate aldolase |
| *SUB1602* | sugar kinase |
| *SUB1603* | hypothetical protein |
| *SUB1605* | 2 keto 3 deoxygluconate permease |
| *SUB1606* | 5 keto 4 deoxyuronate isomerase |
| *SUB1607* | LacI family regulatory protein |
| *SUB0153* | hypothetical protein |
| *SUB0228* | DNA binding protein |
| *SUB0514* | hypothetical protein |
| *SUB1028* | nucleotide sugar epimerase |
| *SUB1036* | glycosyl transferase |
| *SUB1594* | putative Mg2+/citrate transporter (fragment) |
| *SUB1822* | hypothetical protein |
| *SUB1826* | autolysin regulatory protein |
| *SUB0036* | ublE membrane protein |
| *SUB0037A* | bifunctional purine biosynthesis protein purH |
| *SUB0049* | membrane protein |
| *SUB0064* | hypothetical protein |

Continued on next page

Table 4.1 (continued from previous page)

| Gene id | Function |
|---|---|
| *SUB0091* | 50S ribosomal protein L36 |
| *SUB0096* | AraC family regulatory protein (fragment) |
| *SUB0149* | hypothetical protein |
| *SUB0193* | hypothetical protein |
| *SUB0211* | membrane protein |
| *SUB0217* | hypothetical protein |
| *SUB0218* | hypothetical protein |
| *SUB0219* | hypothetical protein |
| *SUB0229* | DNA binding protein |
| *SUB0230* | phage membrane protein |
| *SUB0280* | type I restriction-modification system S protein (pseudogene) |
| *SUB0292A* | putative dihydrolipoamide dehydrogenase(fragment) |
| *SUB0299* | membrane protein |
| *SUB0327A* | antiholin-like protein (fragment) |
| *SUB0378* | hypothetical protein |
| *SUB0476* | conserved hypothetical protein (pseudogene) |
| *SUB0500* | membrane protein |
| rpmGB | 50S ribosomal protein L33 |
| *SUB0613* | hypothetical protein |
| *SUB0624* | hypothetical protein |
| *SUB0779* | hypothetical protein |
| *SUB0780* | hypothetical protein |
| *SUB0781* | hypothetical protein |
| S*UB0790A* | conserved hypothetical protein (fragment) |
| *SUB0803* | DNA_binding protein |
| *SUB0887* | conserved hypothetical protein (fragment) |
| *SUB1030* | acetyltransferase |
| *SUB1031* | exopolysaccharide biosynthesis protein |
| *SUB1032* | hypothetical protein |
| *SUB1033* | membrane protein |
| *SUB1034* | glycosyl transferase |
| *SUB1035* | putative galacturonic acid acetylase (fragment) |
| *SUB1095* | *sclB* collagen like surface anchored protein |
| *SUB1135* | hypothetical protein |
| *SUB1150* | membrane protein |
| *SUB1170A* | conserved hypothetical protein (pseudogene) |
| *SUB1178A* | putative transposase (fragment) |
| *SUB1340* | conserved hypothetical protein (fragment) |
| *SUB1343* | hypothetical protein |
| *SUB1510* | CsbD like protein |

Continued on next page

Table 4.1 (continued from previous page)

| Gene id | Function |
|---|---|
| *SUB1578* | membrane protein |
| *SUB1597* | hypothetical protein |
| *SUB1607.1* | hypothetical protein |
| *SUB1732A* | 50S ribosomal protein L33 |
| *SUB1748* | dioxygenase |
| *SUB1749* | dioxygenase |
| *SUB1750* | hypothetical protein |
| *SUB1751* | phospholipase/carboxylesterase |
| *SUB1817* | conserved hypothetical protein (fragment) |
| *SUB1818* | Fic protein family phage protein |
| *SUB1820* | hypothetical protein |
| *SUB1823* | hypothetical protein |
| *SUB1824* | phage membrane protein |
| *SUB1825* | hypothetical protein |
| *SUB1827* | hypothetical protein |
| *SUB1828* | hypothetical protein |
| *SUB1830* | hypothetical protein |
| *SUB1839* | DNA binding protein |
| *SUB1865* | ABC transporter ATP binding protein |

Table 4.2: Genes present in EF20 and absent from 0140J.

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.1367 | hypothetical protein |
| fig\|6666666.88081.peg.1338 | 3-demethylubiquinone-9 3-methyltransferase |
| fig\|6666666.88081.peg.1467 | hypothetical protein |
| fig\|6666666.88081.peg.415 | ABC transporter ATP-binding protein |
| fig\|6666666.88081.peg.586 | hypothetical protein |
| fig\|6666666.88081.peg.1889 | hypothetical protein |
| fig\|6666666.88081.peg.1790 | FIG01114110: hypothetical protein |
| fig\|6666666.88081.peg.1665 | FIG01114335: hypothetical protein |
| fig\|6666666.88081.peg.1792 | acetyltransferase, GNAT family |
| fig\|6666666.88081.peg.1893 | hypothetical protein |
| fig\|6666666.88081.peg.1888 | hypothetical protein |
| fig\|6666666.88081.peg.614 | hypothetical protein |
| fig\|6666666.88081.peg.1887 | hypothetical protein |
| fig\|6666666.88081.peg.693 | FIG01119637: hypothetical protein |
| fig\|6666666.88081.peg.140 | [Citrate [pro-3S]-lyase] ligase (EC 6.2.1.22) |

Continued on next page

Table 4.2 (continued from previous page)

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.268 | hypothetical protein |
| fig\|6666666.88081.peg.1468 | hypothetical protein |
| fig\|6666666.88081.peg.691 | hypothetical protein |
| fig\|6666666.88081.peg.580 | Alkaline phosphatase (EC 3.1.3.1) |
| fig\|6666666.88081.peg.121 | hypothetical protein |
| fig\|6666666.88081.peg.694 | hypothetical protein |
| fig\|6666666.88081.peg.1045 | hypothetical protein |
| fig\|6666666.88081.peg.1487 | UDP-glucose 4-epimerase (EC 5.1.3.2) |
| fig\|6666666.88081.peg.136 | Mg2+/citrate complex transporter |
| fig\|6666666.88081.peg.557 | hypothetical protein |
| fig\|6666666.88081.peg.139 | FIG01117820: hypothetical protein |
| fig\|6666666.88081.peg.138 | 2-(5”-triphosphoribosyl)-3’-dephosphocoenzyme-A synthase (EC 2.7.8.25) |
| fig\|6666666.88081.peg.137 | Transcriptional regulator, GntR family |
| fig\|6666666.88081.peg.225 | hypothetical protein |
| fig\|6666666.88081.peg.1493 | Eps11J |
| fig\|6666666.88081.peg.1788 | hypothetical protein |
| fig\|6666666.88081.peg.1521 | Transcriptional regulator, ArsR family |
| fig\|6666666.88081.peg.1524 | Putative pyridine nucleotide-disulphide oxidoreductase |
| fig\|6666666.88081.peg.1356 | hypothetical protein |
| fig\|6666666.88081.peg.1357 | hypothetical protein |
| fig\|6666666.88081.peg.1523 | cationic amino acid transporter |
| fig\|6666666.88081.peg.1522 | YeeE/YedE family protein |
| fig\|6666666.88081.peg.1168 | Lipoteichoic acid synthase LtaS Type IIc |
| fig\|6666666.88081.peg.760 | hypothetical Cytosolic Protein |
| fig\|6666666.88081.peg.141 | Transcriptional regulator, LysR family |
| fig\|6666666.88081.peg.575 | FIG01202080: hypothetical protein |
| fig\|6666666.88081.peg.579 | hypothetical protein |
| fig\|6666666.88081.peg.143 | Uncharacterized protein Bsu YraM |
| fig\|6666666.88081.peg.1116 | hypothetical protein |
| fig\|6666666.88081.peg.80 | hypothetical protein |
| fig\|6666666.88081.peg.975 | Response regulator of the competence regulon ComE |
| fig\|6666666.88081.peg.1556 | cation-transporting ATPase, E1-E2 family |
| fig\|6666666.88081.peg.574 | Glycerol-3-phosphate ABC transporter, periplasmic glycerol-3-phosphate-binding protein (TC 3.A.1.1.3) |
| fig\|6666666.88081.peg.142 | Mg2+/citrate complex transporter |
| fig\|6666666.88081.peg.578 | unknown |
| fig\|6666666.88081.peg.573 | Fibronectin-binding protein |
| fig\|6666666.88081.peg.144 | 2-methylcitrate dehydratase (EC 4.2.1.79) |

Continued on next page

Table 4.2 (continued from previous page)

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.577 | Multiple sugar ABC transporter, ATP-binding protein |
| fig\|6666666.88081.peg.301 | hypothetical protein |
| fig\|6666666.88081.peg.1541 | FIG131328: Predicted ATP-dependent endonuclease of the OLD family |
| fig\|6666666.88081.peg.576 | Hydrolase, haloacid dehalogenase-like family |
| fig\|6666666.88081.peg.759 | hypothetical protein |
| fig\|6666666.88081.peg.1540 | ATP-dependent DNA helicase pcrA (EC 3.6.1.-) |
| fig\|6666666.88081.peg.978 | ABC transporter ATP-binding protein |
| fig\|6666666.88081.peg.634 | hypothetical protein |
| fig\|6666666.88081.peg.645 | FIG01117125: hypothetical protein |
| fig\|6666666.88081.peg.635 | hypothetical protein |
| fig\|6666666.88081.peg.464 | FIG01114860: hypothetical protein |
| fig\|6666666.88081.peg.640 | Sortase A, LPXTG specific |
| fig\|6666666.88081.peg.647 | FIG01116764: hypothetical protein |
| fig\|6666666.88081.peg.1169 | membrane protein |
| fig\|6666666.88081.peg.567 | hypothetical protein |
| fig\|6666666.88081.peg.546 | PTS sysytem, mannitol-specific enzyme II, B component |
| fig\|6666666.88081.peg.549 | Fructose-bisphosphate aldolase class II (EC 4.1.2.13) |
| fig\|6666666.88081.peg.638 | hypothetical protein |
| fig\|6666666.88081.peg.1349 | oxidoreductase, putative |
| fig\|6666666.88081.peg.595 | Triosephosphate isomerase (EC 5.3.1.1) |
| fig\|6666666.88081.peg.1622 | hypothetical protein |
| fig\|6666666.88081.peg.1348 | Thiamin biosynthesis lipoprotein ApbE |
| fig\|6666666.88081.peg.641 | oxidoreductase, aldo/keto reductase family |
| fig\|6666666.88081.peg.639 | Cell wall surface anchor family protein |
| fig\|6666666.88081.peg.979 | hypothetical protein |
| fig\|6666666.88081.peg.643 | transcriptional regulator, Cro/CI family |
| fig\|6666666.88081.peg.147 | Beta-lactamase (EC 3.5.2.6) |
| fig\|6666666.88081.peg.547 | FIG00732228: membrane protein |
| fig\|6666666.88081.peg.1347 | FIG00745083: hypothetical protein |
| fig\|6666666.88081.peg.548 | hypothetical protein |
| fig\|6666666.88081.peg.596 | UPF0047 protein Bsu YugU |
| fig\|6666666.88081.peg.594 | Fructose-bisphosphate aldolase class II (EC 4.1.2.13) |
| fig\|6666666.88081.peg.591 | PTS system, galactitol-specific IIA component (EC 2.7.1.69) |
| fig\|6666666.88081.peg.636 | hypothetical protein |
| fig\|6666666.88081.peg.593 | PTS system, galactitol-specific IIB component (EC 2.7.1.69) |

Continued on next page

Table 4.2 (continued from previous page)

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.592 | PTS system, galactitol-specific IIC component (EC 2.7.1.69) |
| fig\|6666666.88081.peg.148 | N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28) |
| fig\|6666666.88081.peg.590 | Mannitol operon activator, BglG family |
| fig\|6666666.88081.peg.460 | hypothetical protein |
| fig\|6666666.88081.peg.646 | Transcriptional regulator, Cro/CI family |
| fig\|6666666.88081.peg.637 | Signal peptidase I (EC 3.4.21.89) |
| fig\|6666666.88081.peg.648 | Integrase |
| fig\|6666666.88081.peg.550 | FIG00628025: hypothetical protein |
| fig\|6666666.88081.peg.545 | PTS system, IIA component |
| fig\|6666666.88081.peg.702 | hypothetical protein |
| fig\|6666666.88081.peg.825 | hypothetical protein |
| fig\|6666666.88081.peg.669 | hypothetical protein |
| fig\|6666666.88081.peg.680 | FIG01113951: hypothetical protein |
| fig\|6666666.88081.peg.1553 | hypothetical protein |
| fig\|6666666.88081.peg.765 | PTS system, fructose-specific IIB component (EC 2.7.1.69) / PTS system, fructose-specific IIC component (EC 2.7.1.69) |
| fig\|6666666.88081.peg.158 | CTP synthase (EC 6.3.4.2) |
| fig\|6666666.88081.peg.756 | hypothetical protein |
| fig\|6666666.88081.peg.664 | hypothetical protein |
| fig\|6666666.88081.peg.1483 | Glycerol-3-phosphate cytidylyltransferase (EC 2.7.7.39) |
| fig\|6666666.88081.peg.654 | membrane protein, putative |
| fig\|6666666.88081.peg.649 | FIG01117029: hypothetical protein |
| fig\|6666666.88081.peg.671 | FIG01116529: hypothetical protein |
| fig\|6666666.88081.peg.372 | hypothetical protein |
| fig\|6666666.88081.peg.370 | hypothetical protein |
| fig\|6666666.88081.peg.154 | Phenazine biosynthesis protein PhzF |
| fig\|6666666.88081.peg.164 | hypothetical protein |
| fig\|6666666.88081.peg.1629 | Enoyl-CoA hydratase (EC 4.2.1.17) / Enoyl-CoA hydratase [valine degradation] (EC 4.2.1.17) / 3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) |
| fig\|6666666.88081.peg.1489 | Glycosyltransferase |
| fig\|6666666.88081.peg.659 | Transcriptional regulator, XRE family |
| fig\|6666666.88081.peg.657 | unknown |
| fig\|6666666.88081.peg.155 | Protein export cytoplasm protein SecA ATPase RNA helicase (TC 3.A.5.1.1) |
| fig\|6666666.88081.peg.656 | hypothetical protein |
| fig\|6666666.88081.peg.679 | Glycerate kinase (EC 2.7.1.31) |

Continued on next page

Table 4.2 (continued from previous page)

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.1490 | Glycosyltransferase |
| fig\|6666666.88081.peg.763 | Transcriptional antiterminator of lichenan operon, BglG family |
| fig\|6666666.88081.peg.815 | Nucleotide sugar synthetase-like protein |
| fig\|6666666.88081.peg.819 | Accessory secretory protein Asp3 |
| fig\|6666666.88081.peg.817 | Accessory secretory protein Asp1 |
| fig\|6666666.88081.peg.673 | FIG01117874: hypothetical protein |
| fig\|6666666.88081.peg.683 | hypothetical protein |
| fig\|6666666.88081.peg.1484 | CDP-glycerol:poly(glycerophosphate) glycerophosphotransferase (EC 2.7.8.12) |
| fig\|6666666.88081.peg.977 | hypothetical protein |
| fig\|6666666.88081.peg.160 | hypothetical protein |
| fig\|6666666.88081.peg.766 | PTS system, fructose-specific IIBC component (EC 2.7.1.69) |
| fig\|6666666.88081.peg.156 | unknown |
| fig\|6666666.88081.peg.1631 | Long-chain-acyl-CoA dehydrogenase (EC 1.3.99.13) |
| fig\|6666666.88081.peg.768 | hypothetical protein |
| fig\|6666666.88081.peg.151 | ABC transporter, ATP-binding protein |
| fig\|6666666.88081.peg.826 | hypothetical protein |
| fig\|6666666.88081.peg.820 | Protein export cytoplasm protein SecA2 ATPase RNA helicase (TC 3.A.5.1.1) |
| fig\|6666666.88081.peg.157 | Glutamate-rich protein grpB |
| fig\|6666666.88081.peg.651 | transcriptional regulator, Cro/CI family |
| fig\|6666666.88081.peg.767 | Fructose-bisphosphate aldolase class II (EC 4.1.2.13) |
| fig\|6666666.88081.peg.675 | FIG01115107: hypothetical protein |
| fig\|6666666.88081.peg.663 | hypothetical protein |
| fig\|6666666.88081.peg.822 | GftB: Glycosyl transferase, family 8 |
| fig\|6666666.88081.peg.816 | Preprotein translocase secY subunit (TC 3.A.5.1.1) |
| fig\|6666666.88081.peg.764 | PTS system, fructose-specific IIA component (EC 2.7.1.69) |
| fig\|6666666.88081.peg.375 | hypothetical protein |
| fig\|6666666.88081.peg.389 | Phage protein |
| fig\|6666666.88081.peg.1628 | 3-ketoacyl-CoA thiolase (EC 2.3.1.16) @ Acetyl-CoA acetyltransferase (EC 2.3.1.9) |
| fig\|6666666.88081.peg.1627 | Transcriptional regulators, LysR family |
| fig\|6666666.88081.peg.1630 | 4-hydroxybutyrate coenzyme A transferase |
| fig\|6666666.88081.peg.379 | hypothetical protein |
| fig\|6666666.88081.peg.658 | putative transcriptional regulator |
| fig\|6666666.88081.peg.750 | Integrase |
| fig\|6666666.88081.peg.377 | hypothetical protein |

Continued on next page

Table 4.2 (continued from previous page)

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.689 | Cadmium efflux system accessory protein |
| fig\|6666666.88081.peg.368 | hypothetical protein |
| fig\|6666666.88081.peg.769 | hypothetical protein |
| fig\|6666666.88081.peg.369 | hypothetical protein |
| fig\|6666666.88081.peg.150 | putative response regulator; homolog of RumR and ScnR |
| fig\|6666666.88081.peg.373 | hypothetical protein |
| fig\|6666666.88081.peg.653 | ABC transporter, ATP-binding protein |
| fig\|6666666.88081.peg.1632 | Long-chain-fatty-acid–CoA ligase (EC 6.2.1.3) |
| fig\|6666666.88081.peg.644 | hypothetical gene |
| fig\|6666666.88081.peg.1491 | probable glycosyl transferase |
| fig\|6666666.88081.peg.162 | Protein export cytoplasm protein SecA ATPase RNA helicase (TC 3.A.5.1.1) |
| fig\|6666666.88081.peg.165 | hypothetical protein |
| fig\|6666666.88081.peg.667 | Transcriptional regulator, Cro/CI family |
| fig\|6666666.88081.peg.690 | Cadmium resistance protein |
| fig\|6666666.88081.peg.761 | Methylase |
| fig\|6666666.88081.peg.684 | hypothetical protein |
| fig\|6666666.88081.peg.678 | membrane protein, putative |
| fig\|6666666.88081.peg.661 | hypothetical protein |
| fig\|6666666.88081.peg.666 | FtsK/SpoIIIE family protein |
| fig\|6666666.88081.peg.655 | DNA-damage-inducible protein J |
| fig\|6666666.88081.peg.153 | putative transporter, trans-membrane domain bacteriocin immunity protein |
| fig\|6666666.88081.peg.374 | hypothetical protein |
| fig\|6666666.88081.peg.755 | hypothetical protein |
| fig\|6666666.88081.peg.1485 | Polysaccharide biosynthesis protein |
| fig\|6666666.88081.peg.376 | hypothetical protein |
| fig\|6666666.88081.peg.823 | hypothetical protein |
| fig\|6666666.88081.peg.166 | hypothetical protein |
| fig\|6666666.88081.peg.685 | Phage integrase |
| fig\|6666666.88081.peg.670 | FIG01116947: hypothetical protein |
| fig\|6666666.88081.peg.818 | Accessory secretory protein Asp2 |
| fig\|6666666.88081.peg.161 | hypothetical protein |
| fig\|6666666.88081.peg.665 | hypothetical protein |
| fig\|6666666.88081.peg.382 | hypothetical protein |
| fig\|6666666.88081.peg.660 | hypothetical protein |
| fig\|6666666.88081.peg.163 | FIG01117573: hypothetical protein |
| fig\|6666666.88081.peg.676 | FIG01116993: hypothetical protein |
| fig\|6666666.88081.peg.642 | hypothetical protein |

Continued on next page

Table 4.2 (continued from previous page)

| Gene id (RAST) | Function |
|---|---|
| fig\|6666666.88081.peg.762 | Type II restriction endonuclease |
| fig\|6666666.88081.peg.652 | hypothetical protein |
| fig\|6666666.88081.peg.1488 | Wzy |
| fig\|6666666.88081.peg.159 | FIG01117573: hypothetical protein |
| fig\|6666666.88081.peg.662 | hypothetical protein |
| fig\|6666666.88081.peg.672 | hypothetical protein |
| fig\|6666666.88081.peg.1539 | hypothetical protein |
| fig\|6666666.88081.peg.821 | Poly(glycerol-phosphate) alpha-glucosyltransferase GftA (EC 2.4.1.52) |
| fig\|6666666.88081.peg.681 | cell surface protein precursor |
| fig\|6666666.88081.peg.668 | hypothetical protein |
| fig\|6666666.88081.peg.1486 | Acetyltransferase |
| fig\|6666666.88081.peg.674 | FIG01116888: hypothetical protein |
| fig\|6666666.88081.peg.757 | hypothetical protein |
| fig\|6666666.88081.peg.149 | putative histidine kinase, ScnK homolog |
| fig\|6666666.88081.peg.650 | Lysine decarboxylase family |
| fig\|6666666.88081.peg.1492 | Serine acetyltransferase (EC 2.3.1.30) |
| fig\|6666666.88081.peg.152 | putative transporter, trans-membrane domain bacteriocin immunity protein |
| fig\|6666666.88081.peg.677 | conjugal transfer protein, putative |
| fig\|6666666.88081.peg.814 | Glycosyl transferase, family 8 |
| fig\|6666666.88081.peg.758 | Similar to pipB |
| fig\|6666666.88081.peg.974 | hypothetical protein |
| fig\|6666666.88081.peg.54 | hypothetical protein |
| fig\|6666666.88081.peg.829 | hypothetical protein |
| fig\|6666666.88081.peg.1552 | Collagen-like surface protein |
| fig\|6666666.88081.peg.824 | hypothetical protein |
| fig\|6666666.88081.peg.827 | CDP-diacylglycerol–glycerol-3-phosphate 3-phosphatidyltransferase (EC 2.7.8.5) |
| fig\|6666666.88081.peg.858 | hypothetical protein |
| fig\|6666666.88081.peg.1607 | hypothetical protein |
| fig\|6666666.88081.peg.1592 | FIG01115305: hypothetical protein |
| fig\|6666666.88081.peg.682 | hypothetical protein |
| fig\|6666666.88081.peg.828 | PUTATIVE PROLIN-RICH TRANSMEMBRANE PROTEIN |

The copper metabolism operon *SUB1180-SUB1184* was absent in EF20, however, growth comparison studies in restrictive levels of copper does not appear to differen-

tially affect the rate of growth of either strain (results not shown). This may be due to a compensatory gene duplication of *copZ* event of the copper operon located in the region *SUB1462-SUB1464*. Following sequencing of the 0140J strain, a list of suggested virulence genes were proposed for *S. uberis* [216]. Many of these are found in the EF20 genome and appeared to be intact and hence presumably functional, including *SUB1111* (Fibronectin- binding protein), *SUB1273* (Hemolysin like protein), *SUB1154* (C5a peptidase precursor), *SUB0881* (Sortase A), *SUB0145* (*Lbp*), *SUB1095* (Collagen like surface-anchored protein), *SUB1635* (*SUAM* protein), *SUB1785* (*PauA* Streptokinase precursor) suggesting that the simple presence of these genes is insufficient to explain virulence in *S. uberis*. Comparison of the inferred differences between metabolic subsystems highlights multiple differences (Additional file 4.1: Subsystem enrichment) (Section 2.16). Tests for association (2 with Benjamini-Hochberg multiple hypothesis correction $p \leq 0.05$) identifies ten subsystems enriched in either 0140J or EF20. Bacterial checkpoint-control-related cluster, F0F1-type ATP synthase, Fructose and Mannose Inducible PTS, Restriction-Modification System, Formaldehyde assimilation: Ribulose monophosphate pathway and Lysine Biosynthesis DAP Pathway subsystems are over-enriched in 0140J compared to EF20. Whereas, D-Tagatose and Galactitol Utilization, Phage replication and Cadmium resistance subsystems are enriched in EF20 compared to 0140J.

### 4.2.2 Pan and core-genome analysis

Genome comparison using whole genome sequences revealed much higher degree of strain variability compared to using conventional methods such as PFGE. [13, 108, 143]. This led to the development of the distributed genome hypothesis [12] which assumes that bacterial taxa such as species or genera possess a distributed genome in which the variable or unique genes, termed the "pan-genome" exceeds the gene set of any individual by several magnitudes [90, 205]. In general the pan-genome of a group of bacteria is defined as the union of three distinct sets of genes: core genes, accessory genes and unique genes. Each of the three gene sets shows certain characteristics, e.g., the number of genes that are included in it and the distribution of functional classes [28, 90, 112, 205]. Therefore, core, accessory and unique genes can be used for different applications and provide different information about the group of investigated genomes. In the 12 *S. uberis*, strains sequenced here, the total number of genes shared by all strains sequenced and 0140J, was 1,550 genes (Additional file 4.1: core_genome) grouped in 1,509 orthologous clusters (Additional file 4.1: RAST_orthologous_clusters). By repeatedly comparing the shared core-genome clusters obtained from 1000 random combinations of strains we can see that with two genomes the median shared core-genome size is 1,635 genes with a variance (interquartile range) of 44 genes. By comparison of 5 assembled genomes the median is 1,560 with interquartile range of 29 genes. If 10 genomes are compared the core-genome size plateaus at 1,521 genes and the interquartile range is 16 genes (Fig 4.1a).

The identified *S. uberis* core-genome contained mainly housekeeping gene clusters or

operons of genes essential for cellular growth and replication including those involved in cell wall and capsule synthesis, cell division, cell cycle regulation and cell signalling, membrane transport (protein secretion systems), RNA/DNA metabolism, metabolism of cofactors, aromatic compounds, amino acids and derivatives (arginine, lysine, threonine metabolism), phosphorus, fatty acid and lipids and carbohydrate uptake and utilization. Genes that encode mobile genetic elements such as prophage genes, ICE elements and other horizontally acquired genes are underrepresented in the core-genome. Although the core-genome is conserved for a group of bacteria it is not invariable since over time it will be changed by horizontal gene transfer (HGT) and natural selection [115]. Thus, bacterial taxa show significant differences in the number and functional distribution of their pan-genes.

As with the core-genome, with 10-12 *S.uberis* genomes the number of novel genes identified with the addition of an additional genome slows but does not plateau (Fig 4.1 b), suggesting an open pan-genome [204]. Together these comparisons of the core and pan-genomes suggest that sequencing the relatively small number of strains has captured the majority of variation within *S. uberis* genomes. Another group of genes is the pan-genome which consists of the set of accessory genes that are neither core nor unique genes, i.e., genes found in at least two but not all genomes. These are also called "distributed" [90] or "dispensable" [204] genes. Most of the genes found in this group are mobile genetic elements such as prophage genes, CRISPRcas system, bacteriocins and a number of virulent factors described in their respective sections below. They are presumed not to be involved in essential metabolic functions, but, play important roles in

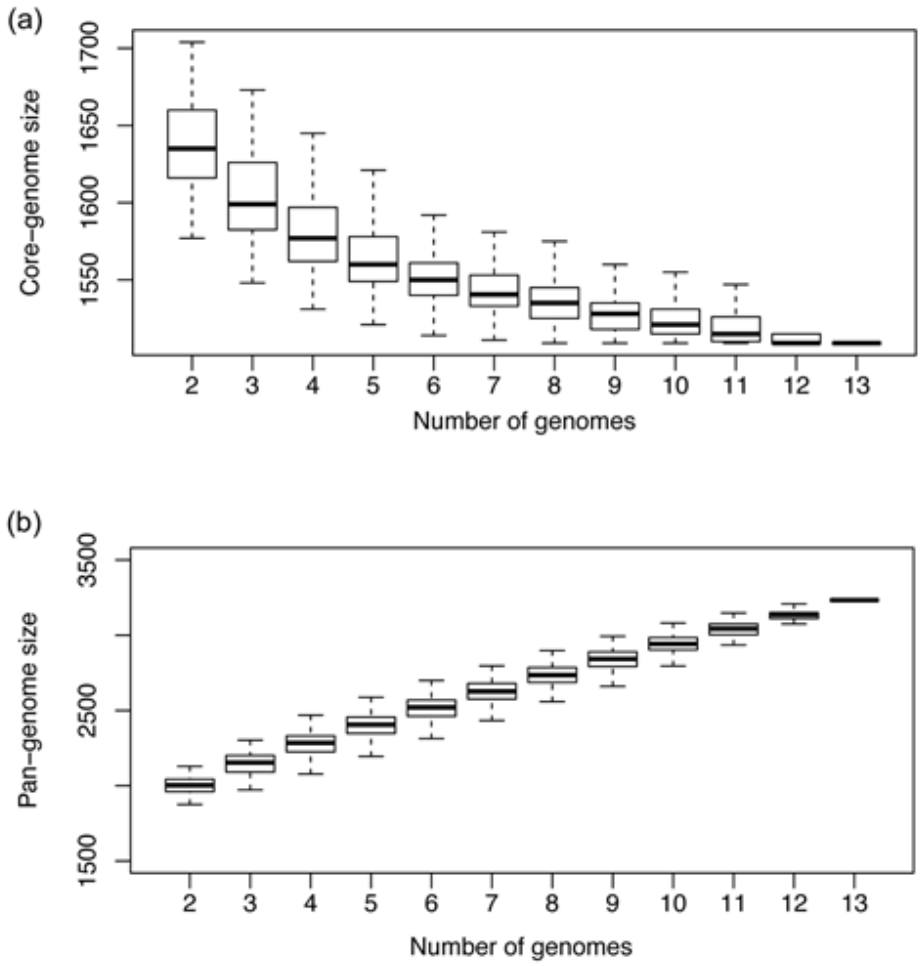



Figure 4.1: Relationship between gene count and genome size (a) Core genome. To determine the stability of pan/core-genomes, the pan-genome size (total number of different genes identified within samples) and core-genome size (number of common genes within samples) was determined when between 2 and 13 genomes were combined in random order. For each combination size (2..13 genomes) 1000 permutations were conducted. The boxplots represent the median and interquartile ranges of 1000 permutations, whiskers extend to the value extremes. (b) Pan-Genome.

genetic variability. They may often be involved in adaptation to a specific niche [108, 116] or manifestation of pathogenicity.

Direct comparison of all newly sequenced genomes with the available 0140J strain identifies between 115-246 accessory genes present in one or more *S. uberis* strains but not present in 0140J (Fig 4.2).

Following the assumption that the genomes of clinical strains are smaller, the clinical isolates were found to have slightly fewer accessory genes (median 180.5) than the sub-clinical isolates (median 193.5), however again these differences were not statistically significant. In addition, hierarchical clustering of the strains based on their shared accessory genome content, does not group strains by clinical status (Fig 4.3).

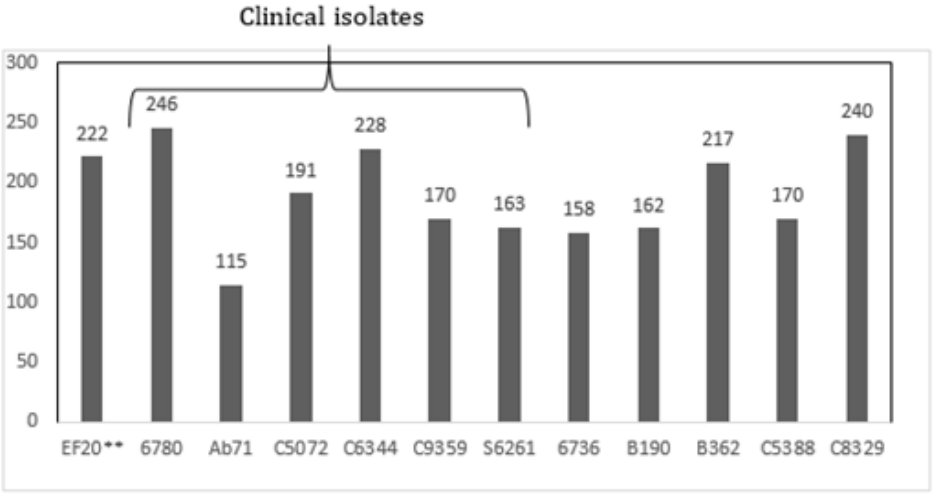



Figure 4.2: Distribution of accessory genes not in 0140J. The mean number of accessory genes in clinical isolates is 186 and sub-clinical isolates 189 with the range of 71 and 86, respectively. The independent t-test showed the p value of 0.8814 indicating no significant association between the number of accessory genes and disease outcome.

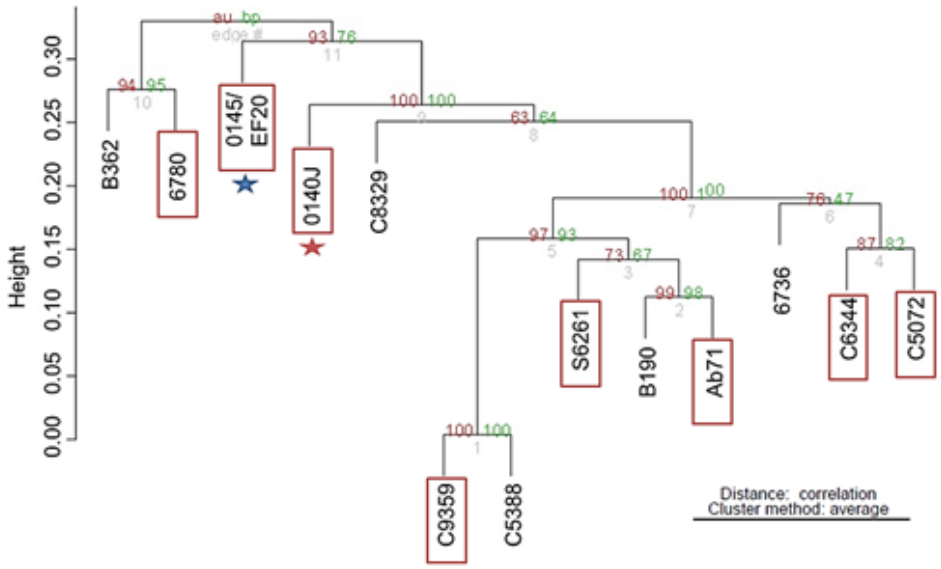



Figure 4.3: Pan-genome clustering of *S. uberis* isolates. Similarity of the samples based in gene presence/absence was conducted using hierarchical clustering (pvclust) (http://cran.rproject.org/web/packages/pvclust/). The red star denotes the virulent strain 0140J, blue star, non-virulent strain EF20. Strains initially recorded as clinical isolates are boxed in red.

### 4.2.3 Comparison of virulence factors

The virulence factors in the *S. uberis* genome are found in the variable component of the genome when compared to other Streptococcal species. It was interesting to note that most of them were found within the core-genome when compared among the *S .uberis* isolates [216].

As with the previous comparison of EF20 and 0140J the reported virulence genes *SUB1111*, hemolysin-like protein *SUB1273, SUB1154, SUB0881, SUB0145, SUB1095, SUB1635* and *SUB1785* were found in the core-genome of all other sequenced strains. Whilst the presence of these genes is conserved in clinical and non-clinical strains, they do exhibit a range of sequence variation between strains described in the Chapter 5.

*PauA* (*SUB1785*) and Hemolysin-like protein (*SUB1273*) are very well conserved between the 12 strains.

Sortase is another important virulence factor, which may be universal in Gram-positive bacteria [51].The sortase-anchored proteins of *S. uberis* are known to contain a conserved amino acid LPxxG or LPxxxD motif [120]. Using the presence of either of these motifs together with a predicted secretory leader motif, 10 genes (*SUB0135, SUB0145, SUB0207, SUB0241, SUB0826, SUB0888, SUB1095, SUB1154, SUB1370* and *SUB1730*) were identified as predicted potential sortase anchored proteins in 0140J. Nine of these have been confirmed to be sortase anchored by comparative proteomic analysis [120], whilst *SUB0241* was not. Sortase-anchored genes *SUB0145, SUB1095* and *SUB1154* have been shown to be important in infection and have been proposed to be virulence candidates [120]. *S. uberis* mutants in which these genes were inactivated were attenuated in their ability to infect cattle.

Using RAST to transfer annotation from the 0140J reference suggested that *SUB1095* was unique to 0140J. Since this is an important virulence factor we manually checked whether this gene is truly absent from all other strains. However, using the Rapid Annotation Transfer Tool (RATT) a *SUB1095* ortholog could be detected in all genomes sequenced suggesting that RAST was unable to annotate this gene due its highly variable sequence.

The CW-PRED, a HMM-based method for cell wall anchored proteins classification system for Gram-positive bacteria was used for prediction and comparative analysis of sortase anchored proteins among the 13 completely sequenced genomes. Using this

method, the number of predicted sortase anchored proteins varied from 11 to 18 (Appendix Table A2). 0140J was found to have the fewest number of sortase substrates while B362 found to have the highest. The sortase C substrate *SUB0867* ($_{391}$LPSTAN$_{396}$) was identified in 0140J and a homolog of this gene was found in all other strains.The role of sortase C in the sporulation of *B. anthracis* has been demonstrated, however, it's role in bacterial virulence is rather unclear. Sortase C substrate may be involved in the adaptation to specific environmental conditions as found with the other GAS [193], however it's role in *S. uberis* has not been investigated. Only one conserved sortase was absent from 0140J, however, a possible homolog (*SUB0241*) with low sequence similarity (22% similar with a coverage of 97%) does exist.

### 4.2.4 Analysis of the has operon

The hyaluronic acid capsule of *S. pyogenes* has been found to play a significant role in the pathogenesis of invasive Group A Streptococcus (GAS) bacteria [135]. *S. uberis* strains isolated from cases of bovine mastitis display variable amounts of hyaluronic acid capsule [53] suggesting that the capsule may be associated with infection. However, Field et al 2003 showed that capsule negative mutants can still cause mastitis and the availability of more capsule in clinical isolates than the environmental isolates [61] may be due to the fact that capsule prevents desiccation in the environment and allows it to persist longer, increasing chances of subsequent infection or even gut colonisation. In *S. uberis* 0140J the arrangement of the hyaluronic acid biosynthetic genes comprising the has operon, differs from the typical "hasABC" arrangement common to GAS [215]. *hasA*(*SUB1697*) encoding

hyaluronan synthase and *hasB*(*SUB1696*) encoding UDP-glucose dehydrogenase are arranged as in other GAS. However the *hasC* homologue (*SUB1691*), encoding UDP-glucose pyrophosphorylase, is encoded in the reverse orientation and separated from hasAB by approximately 3 kb of genome, containing CDS thought to be unrelated to capsule biosynthesis [216]. It is unlikely that this movement of *hasC* affects capsule production, as in GAS bacteria capsule production is only dependent on functional copies of *hasA* and *hasB*, but not *hasC* [6]. All the isolates sequenced here except strain B362 have hasABC in a similar arrangement to that found in 0140J (Fig 4.4). In nine *S. uberis* strains a paralog of hasB (*SUB1027*) was identified. The non-capsular, non-virulent isolate EF20 lacks *SUB1027* and this gene is also missing from isolates B362, 6780 and B190.

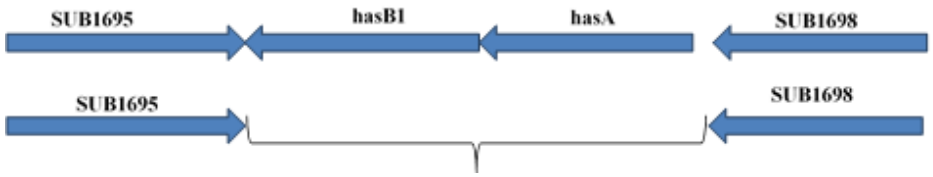



Figure 4.4: Comparison of has operon between 0140J and B362. Deletion of *hasB1* and *hasA* is evident in the isolate B362 which also lacks capsule in it's cell wall (Leigh et al. unpublished data).

### 4.2.5 Analysis of CRISPR-Cas proteins

The CRISPR-Cas (clustered regularly inter- spaced short palindromic repeats - CRISPR - associated proteins) identified in approximately 40% bacteria and most archaea, are genomic regions involved in adaptive immunity against invading genetic elements such

as viruses [46, 70, 92, 99, 192, 211] .

Distinct arrays of short repeats separated by unique spacers have been identified in CRISPR-cas regions and have recently been found to act as nucleases, a helicase, a polymerase and different RNA-binding proteins [94]. CRISPR-cas regions were initially thought to be involved in DNA repair mechanisms, however, the availability of fragments of virus and plasmid genes within the CRISPR-cas region led to the speculation that they might be involved in defense against selfish elements [18, 146, 166]. This mechanism was evident in a number of organisms including the lactic acid bacterium *S. thermophilus* in which integration of phage-specific sequence in CRISPR-Cas system conferred resistant against related phage [11].

The CRISPR-mediated defence processes involve three steps: Adaptation mediated by the incorporation of foreign dna material to the CRISPR-Cas region; transcription of CRISPR RNAs along with spacer repeats and interference with the invasive genetic material leading to their degradation [132].

According to the current classification system there are three major types of CRISPR-Cas system named, type I, type II and type III systems, based on structure and composition CRISPR-Cas region [133, 184]. The maturation of the precursor CRISPR RNA (pre-crRNA) is mediated by specific endonucleases, *Cse3* (type I-E), *Csy4* (type I-F) and *Cas5d* (type I-C) in type I CRISPR-Cas system, whereas in the type III systems *Cas6* is responsible [31, 32, 71, 72, 74, 85, 110, 177, 190, 214, 214]. Trans-activating CRISPR RNA (tracr RNA) binds to the pre-crRNA to form a dual-RNA essential for both crRNA maturation by RNAase III as well as invading DNA cleavage by *Cas9* in type II system

[36, 44, 46, 70, 101, 130].

When the reference strain 0140J was sequenced using Sanger sequencing there was no information available on the presence of CRISPR-cas region in *S. uberis*. Whole Genome Sequencing of isolates revealed CRISPR-Cas regions in all strains except EF20 and 0140J. In strains other than EF20 and 0140J, the type II system [132] which includes the 'HNH'-type (Streptococcus-like) comprising *Cas9/Csn1* (a single, large protein) was located in a conserved region between genes *SUB1084-SUB1085* (Fig 4.5). An additional Type III *cas1-cas6* gene set [132] was found in the isolate B362 inserted between genes *SUB0330-SUB0333* (Fig4.6). In isolates EF20, Ab71 and C5072 an insertion event of two genes, homologous to *S. pneumoniae* integrative and conjugative elements (ICE) are present in the homologous region between *SUB1084-SUB1085*.

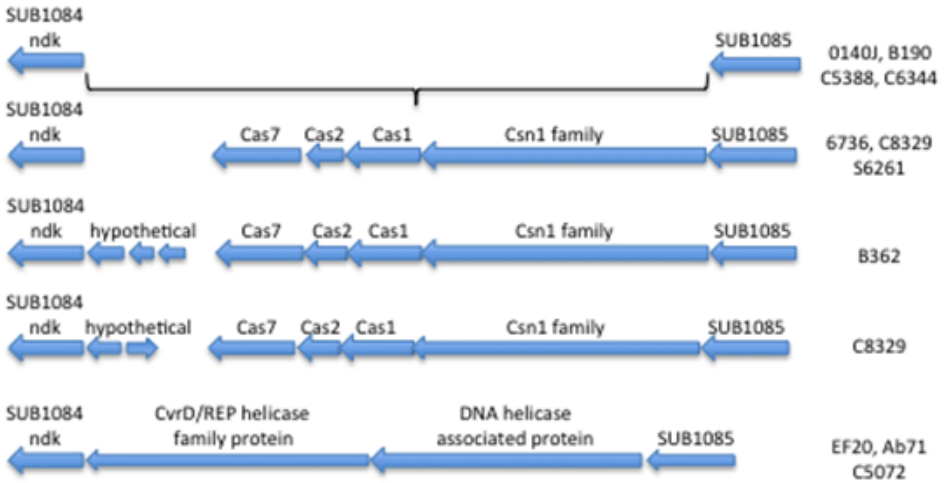



Figure 4.5: Comparative analyses of Type II CRISPRs regions in different isolates. Relative positions of CRISPR-Cas genes in the 13 isolates. The non-virulent strain EF20 along with the isolates Ab71 and C5072 harbour helicase family proteins instead of CRISPRs regions between genes *SUB1084* and *SUB1085*.

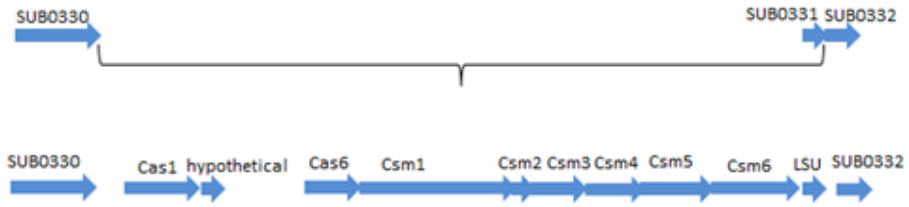



Figure 4.6: Comparative analyses of Type III CRISPRs regions in B362 and 0140J. Relative position of CRISPRs-cas found only in B362 as an insertion between *SUB0330-SUB0332* replacing the CDS *SUB0331*.

### 4.2.6 Analysis of prophage regions

Growing evidence suggests a significant role for prophage regions in the virulence and evolution of many bacteria. For example, lysogeny has been found to contribute to the virulence of a number of organisms including *Vibrio cholerae, S. enterica, E. coli, Clostridium botulinum, Corynebacterium diphtheriae, S. aureus* and *S. pyogenes* [24]. Analysis of prophage regions shows that among the thirteen isolates seven had intact prophage regions (Table 4.3, Fig 4.7). The non-virulent isolate EF20 and sub-clinical isolate B362 did not have any identified prophage regions. An incomplete prophage region is evident in the region *SUB1818-SUB1840* of the isolate 0140J and this region is variable in most of the isolates. The prophage genes were highly diverse and most of them are not conserved between strains making it difficult to use them as genetic markers for epidemiological study as well as phage typing. The diversity of these prophage regions may contribute to the adaptation of lysogens to new hosts as well as environmental conditions by increasing their fitness.

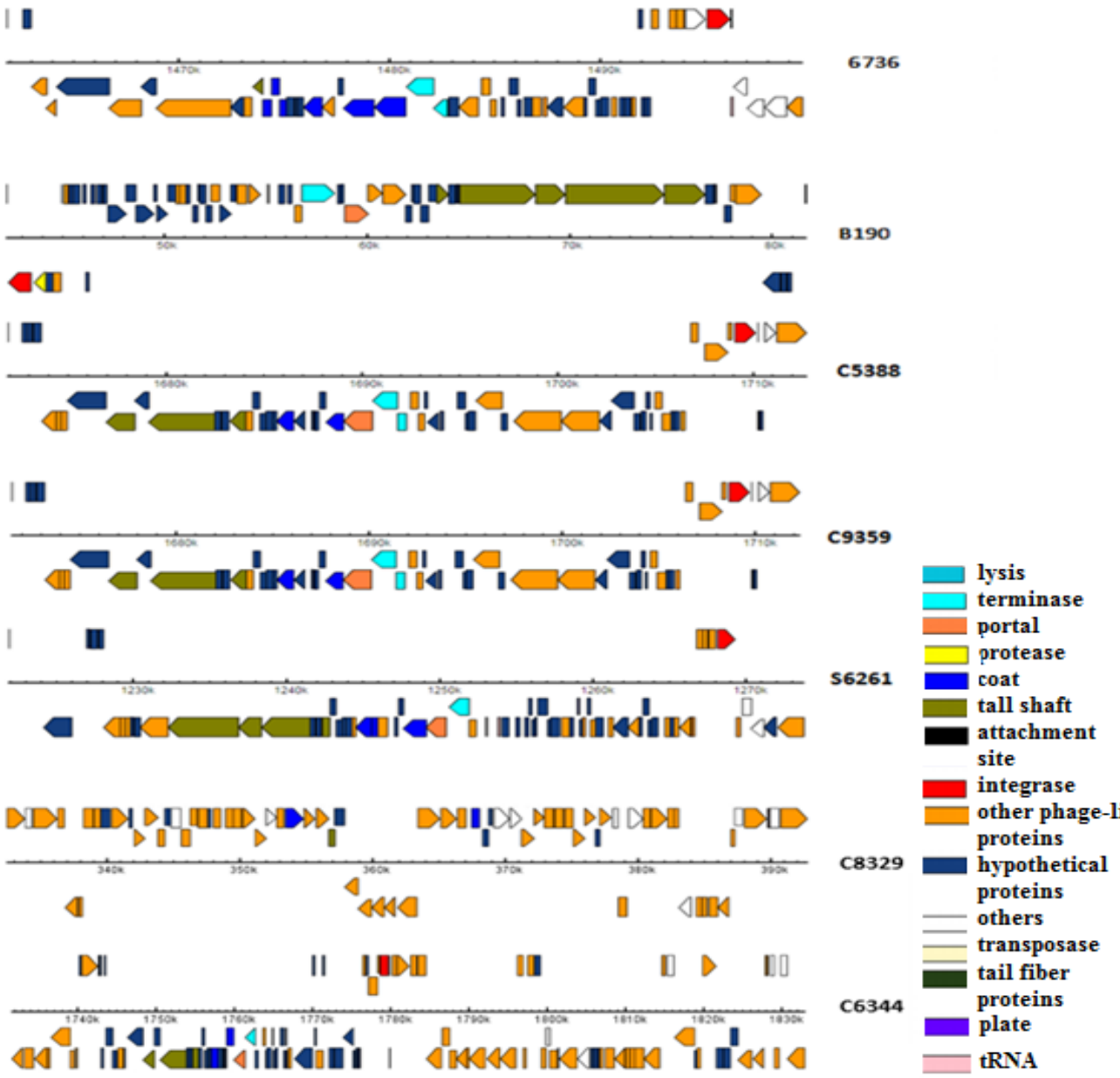



Figure 4.7: Comparisons of the CDS of the seven strains with intact prophage. polygons represent open reading frames of the prophage region, with orientation represented by arrow head. The detailed prophage views of the strains were produced using the online software PHAST (http://phast.wishartlab.com/index.htm). Different colours represent various phage elements and putative CDS.

Table 4.3: Distribution of prophage regions among 13 isolates

| Strain | Region | Length | Status | CDS | Putative phage | GC % | Location compared to 0140J |
|---|---|---|---|---|---|---|---|
| 0140J | 1 | | incomplete | 28 | Lactococcus_phage_bIL311 | 33.6 | SUB1818-SUB1840 |
| EF20 | None | | | | | | |
| 6736 | 1 | 33.5Kb | intact | 51 | PHAGE_Strept_5093_NC_012753 | 35.1 | SUB1470-SUB1471 |
| 6780 | 1 | 35.3Kb | questionable | 23 | PHAGE_Strept_PH10_NC_012756 | 40.6 | SUB1176-SUB1187 |
| Ab71 | 1 | 12.9Kb | incomplete | 15 | PHAGE_Lactoc_bIL311_NC_002670 | 33.1 | SUB181-SUB1840 |
| B190 | 1 | 39.4Kb | intact | 56 | PHAGE_Strept_pyogenes_315_1_NC_004584 | 34.9 | SUB0062-SUB0065 |
| B362 | None | | | | | | |
| C5072 | 1 | 47.7Kb | questionable | 58 | PHAGE_Strept_pyogenes_315_2_NC_004585 | 39.7 | SUB1531-SUB1532 |
| | 2 | 15.8Kb | incomplete | 25 | PHAGE_Lactoc_bIL311_NC_002670 | 32.8 | SUB1818-SUB1840 |
| C5388 | 1 | 40.8Kb | intact | 57 | PHAGE_Strept_P9_NC_009819 | 37.4 | SUB1748-SUB1452 |
| C6344 | 1 | 46.6Kb | questionable | 58 | PHAGE_Strept_pyogenes_315_2_NC_004585 | 38.5 | SUB1531-SUB1532 |
| | 2 | 39.6Kb | intact | 54 | PHAGE_Strept_SMP_NC_008721 | 36.8 | SUB1748-SUB1452 |
| C8329 | 1 | 11.5Kb | incomplete | 15 | PHAGE_Staphy_phi2958PVL_NC_011344 | 40.2 | SUB1183-SUB1190 |
| | 2 | 45.7Kb | intact | 56 | PHAGE_Strept_phi3396_NC_009018 | 36.5 | SUB1531-SUB1532 |
| | 3 | 37.6Kb | questionable | 38 | PHAGE_Lactoc_bIL311_NC_002670 | 35.3 | SUB1818-SUB1840 |
| C9359 | 1 | 40.8Kb | intact | 57 | PHAGE_Strept_P9_NC_009819 | 37.4 | SUB1748-SUB1452 |
| S6261 | 1 | 51.8Kb | intact | 68 | PHAGE_Strept_TP_J34_NC_020197 | 36.4 | SUB1263-SUB1270 |

### 4.2.7 Analysis of bacteriocin production

Bacteriocins are proteinaceous antibiotics produced by bacteria, which kill or inhibit the growth of other bacteria, often providing an advantage in competitive colonization environments. Uberolysin is a novel cyclic bacteriocin produced by *S. uberis* encoded by the operon spanning *SUB0032-SUB0036*. This operon is absent in EF20 and isolate C9359 but is present in all other sequenced strains. Analysis of the 0140J genome identified five genes encoding putative bacteriocin proteins (*SUB0502, SUB0505, SUB0506, SUB0509* and *SUB0512*) [216], of which *SUB502-SUB505* are again absent in EF20 and also isolates B362 and 6780 both of which belong to the ST-86 complex. Whilst bacteriocin production does not define clinical and sub-clinical strains, the absence of almost all bacteriocins in the EF20 genome could put it at a competitive disadvantage with other environmental strains in the dairy cow environment and may reflect (but not explain) it's non-virulent status.

## 4.3 Discussion

The comparison of multiple strains of closely related bacteria provides a valuable resource for the understanding of biological systems. The comparison of 12 newly sequenced strains together with the type 0140J strain of *S. uberis* allows a first comparison of bacteria isolated from clinical and non-clinical infections and the generation of a draft genome of EF20 strain together with the existing 0140J genome, allows for the first time comparison of two naturally occurring strains of *S. uberis* with defined virulence.

The use of comparative genomics approaches enabled the exploration of the genome dynamics of *S. uberis*, allowed the definition of the pan-genome and core-genome, and distribution of virulence factors in different isolates with known virulence. In *S. uberis* the pan-genome size increases with the sequential addition of newly sequenced genomes and the size of the core-genome stabilizes with the addition of new genomes, supporting the prediction of an open-pan genome in *S. uberis*.

The total size of the core-genome is 1,521 genes and most of these are housekeeping genes involved in the maintenance of simple cellular process such as cellular growth and replication, cell cycle regulation, signalling pathways and other major cellular metabolic processes. As found in a number of other Streptococci, most of the dispensable genes were mobile genetic elements (MGE) such as prophage, CRISPR-cas and bacteriocin producing genes [150]. Thus they are involved in shaping the fitness of an isolate in a specific environment.

The genomic variability of *S. uberis* based on the pan-genome is well supported and confirms the MLST-based classification although this did segregate the clinical status

of isolates. The open pan-genome of *S. uberis* also reflects that the species is quickly diversifying by integration of new genes. This suggests that these gains or losses are not fixed in the long term and may change for a specific environmental condition. This scenario has been demonstrated in other bacterial species such in *E. coli* where it was estimated that on average dispensable sequences are lost after 6.7My and 100My in *E. coli* and *S. enterica*, respectively [113].

The ability of *S. uberis* to adapt to different environmental niches [61, 215, 217] suggest that through gene gain and loss the sub-clinical isolates may be capable of becoming clinical, and can lead to clinical mastitis, depending on other environmental factors, such as reduced immunity, poor management or presence of other pathogens [219]. In many bacteria it has been found that in addition to gaining virulence factors, virulent isolates had significantly smaller genomes than non-clinical isolates. Though the size of the virulent isolate 0140J (1.85Mbp) was smaller than the non-virulent isolate EF20 (1.9 Mbp) this finding was not consistent with other clinical isolates. Weinert et al. [219] proposed that the reduced genome size as seen in many pathogenic might be because the presence of 'antivirulence genes', may hinder the full expression of newly acquired virulence factors.

Pathogenic bacteria utilize different cell surface-associated proteins which can interact with host cells and tissues and contribute to pathogenicity. One such mechanism is through covalent linkage by sortase enzymes. Using LPXXD or LPXXXD motifs identified in [53], ten sortase-anchored genes were predicted and all of them were identified within the core genome [215]. A new sortase C substrate (*SUB0867*) was identified using the

tool CW-PRED. All classes of sortase C enzymes so far identified are involved in assembly of pili, thus its role is cell wall attachment for *S. uberis* is questionable, however, they may play role in adhesion and biofilm formation by *S. uberis* [193].

The hyaluronic acid capsule of *S. pyogenes* plays an important role in the pathogenesis of GAS infection and the amount of hyaluronic acid produced by *S. uberis* is variable [215]. The operonic structure of hasABC of *S. uberis* is different from *S. pyogenes* [216]. A homologue (*SUB1027*) of the *S. uberis hasB* (*SUB1696*) was identified and this homologue was absent from non virulent isolate EF20 along with other sub-clinical isolates such as B362 and 6780 that may account for the apparent inability to isolate a *hasB* mutant using phenotypic screening methods [215] and also the inability or reduction of capsule production in these isolates (personal communication Dr. Sharon Egan University of Nottingham).

To the best of our knowledge this is the first report on the presence of CRISPR-cas system in *S. uberis*.Identification of CRISPR-cas genes suggests a defence mechanism against bacteriophage and horizontal gene transfer in *S. uberis*.This also implies the presence of genetic exchange as a result of bacteriophage infection due to evolutionary pressure. Incorporation of beneficial phage in the CRISPR-cas region of *S pyogens* has been suggested to be beneficial for the organism to survive in specific conditions [152]. CRISPR-cas genes may also be involved in the gene regulatory mechanisms for some of the isolates [5]. To the best of our knowledge this is the first report on the presence of CRISPR-cas system in *S. uberis*. Analysis of the 13 *S. uberis* genomes shows the differences in the length of spacer and repeat regions in the CRISPR-cas regions

suggesting the distribution of this region as species dependent. Although CRISPR interference was originally defined as a phage resistance mechanism, CRISPR- Cas systems are now known to play a broader role in limiting horizontal gene transfer [134]. In *S. epidermidis* and *S. pyogenes* CRISPR-Cas systems were shown to prevent new acquisition of plasmids or prophages. This has been proposed to be by blocking entry in similar manner as seen for phage DNA [134, 152]. Similar observations have been made in *Enterococcus faecalis, Enterococcus faecium* and *Campylobacterr jejuni* [51, 157]. Recently in *S. pneumoniae* and *Neisseria meningitides*, CRISPR-Cas systems have been shown to prevent natural transformation [15, 229].

Prophage regions of *S. uberis* are highly diverse and thus it is not plausible to use prophage typing as an epidemiological tool for *S. uberis*. Most of the prophage contents are homologous to other Streptococci indicating the horizontal gene transfer from other Streptococci.

The absence of 0140J bacteriocin operon (*SUB0502-SUB0505*) in the non-virulent isolate EF20 along with other two isolates B362 and 6780 may provide an advantage to be sustainable in an environment by killing other organisms.

The comparison of the strains did not suggest an obvious "smoking gun" gene either present or absent between the virulent or avirulent strains to suggest a previously unknown virulence factor. In addition the genome content did not differentiate between clinical and non-clinical strains. However, it is worth considering that the status as clinical or non-clinical refers to the state of the host animal from which the isolate was obtained, not to the causative agent. For example, the proven non-virulent strain EF20

was isolated from a clinical case and hence is named as a clinical strain but this may have been due to other factors such as a co-infection with another bacterial species/strain and importantly, the genetics of the host. Thus, whilst the data here present a detailed comparison of *S. uberis* bacterial strains, to fully understand virulence and causation of disease, we must take a holistic approach encompassing bacteria, host and environment.

CHAPTER 5

# GENOMIC VARIATION ANALYSIS OF *Streptococcus uberis*

## 5.1 Background

Traditionally, PCR of specific genes followed by sequencing or analysis of a combination of gene sets for example as in MLST, or comparative genome hybridization [232] have been used to study the link between a particular phenotype and association with specific gene/s. However, the advent of the genome-wide association study (GWAS) which involves analysing large numbers of genetic variants such as single nucleotide polymorphisms (SNPs), insertions/deletions (indels) to study the association between phenotype and genotype data [40] provides a promising framework for analysing the genetic basis of bacteria in a much larger scale and in a comprehensive manner [14, 16, 56, 87, 93, 160].

High throughput genotyping using microarray and bead-based assays have been one of the most important tools for investigating common SNPs during the last decade [170]. However, these techniques were associated with high cost and require knowledge of expected SNP variation, hence were limited to bacteria with low genomic diversity [235]. With the availability of NGS with high throughput and lower cost this scenario has recently changed. These technologies made it possible to sequence 100 or more bacterial strains as a routine experiment and opened the door for bacterial GWASs. The use of genome wide SNPs was used to generate a whole genome phylogeny to identify the genes responsible for antituberculosis drug resistance in *Mycobacterium tuberculosis* [59]. Likewise, Laabei et al. used GWAS to identify polymorphisms responsible for the ability of *S. aureus* to invade the human cells [107].

All of these and other GWAS show that this is an unbiased powerful screening approach to discover genomic regions that correlate with a specific phenotype such as bacterial pathogenicity. GWASs can also be linked to the functionality of regulatory pathways and gene expression mechanisms. In this study, we attempted to decipher the phylogenetic relationships between different *S. uberis* isolates with known virulence. Whole genome phylogeny confirmed that isolates clustered into two major clades: one cluster contains the virulent isolate 0140J and the other the non-virulent strain EF20, this led to the analysis of candidate markers contributing the diversification of these two major clades. As noted we did not find any accessory genes that could differentiate between clinical and sub-clinical or virulent and non-virulent isolates and to address this, a GWAS was adopted to determine the feasibility of predicting virulence from the genome sequences

of *S. uberis* isolates. Our findings demonstrate that using whole genome sequence data for large collections of isolates to identify genetic markers associated with a specific trait can be used to infer complex phenotypes from genotype in *S. uberis* [58, 107].

Understanding the forces shaping the evolution of a population is one of the major goals in evolutionary biology. According to the neutral theory, the majority of evolutionary changes are random fixation of mutations causing little or no change in fitness, more rare is positive selection [210]. Positive selection refers to the process of selection of alleles increasing the fitness of an organism. Detection of positive selection is usually performed by comparing protein coding genes and due to their larger effective population size, bacterial genomes tend to be shaped by stronger selection compared to eukaryotes such as humans [34].

Identification of genomic regions under positive selection has been used to infer functionally important regions of genes and genomes. Moreover, identifying regions under varying selection can contribute to the analysis of phenotypic diversity as well as historical forces shaping a population. Several genome-wide studies have been performed for positive selection and in this section, we focus on the detection of selection in protein coding genes.

## 5.2 Results

### 5.2.1 Whole-genome sequencing to identify the relationship among *Streptococcus uberis* isolates

Alignment of 1,377 concatenated core genes (those shared by all strains) comprising 1,294,803 nucleotides identified 43,661 variable sites (SNPs). Phylogenetic analysis revealed two major clades among the isolates one containing the virulent isolate 0140J and one the non-virulent isolate EF20 (Fig 5.1 and 5.2).

Twelve of the isolates were also sequenced using Illumina MiSeq and alignment of concatenated core genes identified fewer (12,982) variable sites among these twelve isolates. In agreement with shared gene content and as seen with analysis of SOLiD data, clustering based on presence or absence of shared SNPs discriminates the genomes based on ST clonal complex rather than by clinical status (Fig 5.2). Analysis of the genomic variations shared between pairs of genomes also showed no obvious association between clinical and subclinicial isolates (Table 5.1). The analysis of tree topology following masking of regions with high probability of recombination shows that the effect of recombination is minimal within in the core genome (Compare Fig 5.2 (a) and (b)).

### 5.2.2 Gene-wide variation analysis

Due to the higher per base quality of the Illumina reads as well as poor assembly statistics and fragmented low quality contigs obtained by SOLiD data, gene wide variation was only conducted using assemblies from Illumina MiSeq data.

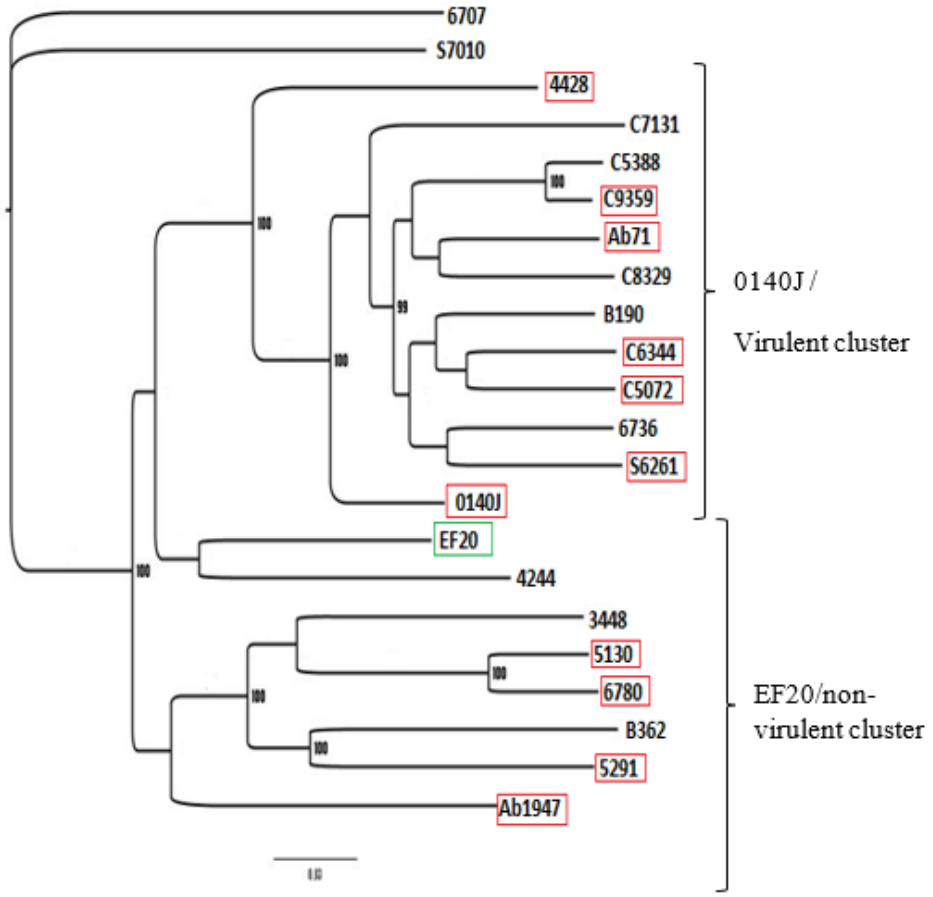



Figure 5.1: Whole genome phylogeny of *S. uberis* isolates using SOLiD reads. A concatenated alignment of 1,377 core genes of 22 isolates was generated by snpTree [114] and the phylogeny determined by PhyML [80] using a GTR model. 200 bootstrap replicates were conducted. Strains initially recorded as clinical isolates are boxed in red. The phylogeny resulted in three major groups one containing the virulent isolate 0140J and one the non-virulent isolate EF20.

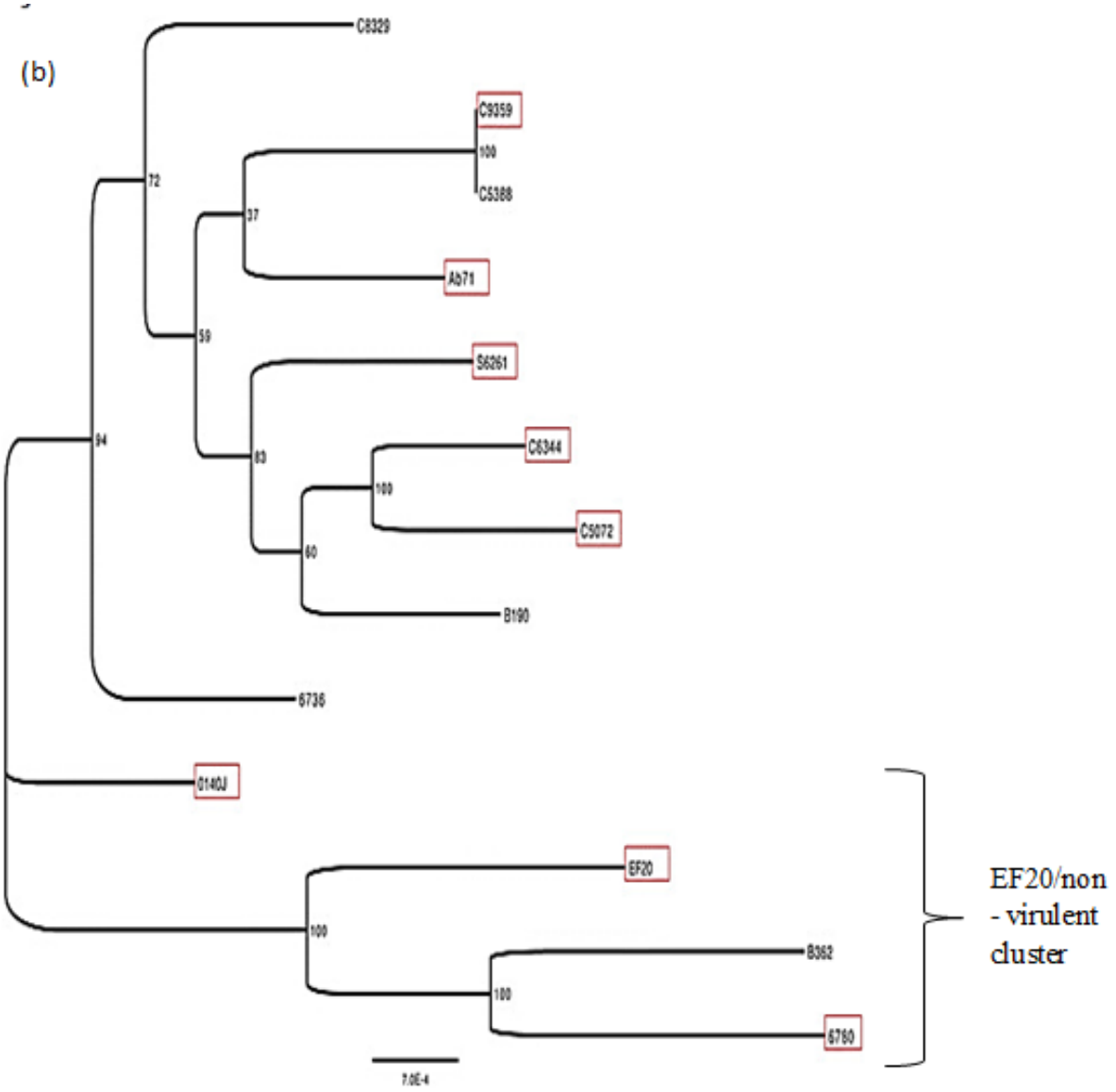



Figure 5.2: Whole genome phylogeny of *S. uberis* isolates using Illumina reads. A concatenated alignment of 1,377 core genes of 13 isolates was generated by snpTree [114]. The Phi-test [23] was conducted on non-overlapping 500 base windows of this alignment. The phylogeny determined by PhyML [80] using a GTR model estimated gamma distribution and 4 substitution rate categories. 200 bootstrap replicates were conducted. Strains initially recorded as clinical isolates are boxed in red. (a) Alignment of complete sequence (b) Alignment of complete sequence where windows with significant evidence of recombination were masked (Phi-test $p < 0.05$ Following a Bonferroni correction). Strains initially recorded as clinical isolates are boxed in red.

Table 5.1: SNP differences among different isolates using Illumina reads (12 isolates)

| Strains | C8329 | B190 | C5072 | 6780 | C5388 | Ab71 | EF20 | 6736 | B362 | C6344 | S6261 | C9359 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C8329 | 0 | 6817 | 7089 | 13541 | 6553 | 6158 | 11620 | 6241 | 13380 | 7185 | 6997 | 6553 |
| B190 | 6817 | 0 | 5425 | 13455 | 6122 | 5911 | 12560 | 6575 | 13588 | 5638 | 5809 | 6117 |
| C5072 | 7089 | 5425 | 0 | 12898 | 6003 | 6142 | 12223 | 6777 | 13475 | 4913 | 6481 | 5964 |
| 6780 | 13541 | 13455 | 12898 | 0 | 12875 | 13528 | 13017 | 13897 | 11104 | 13338 | 13803 | 12817 |
| C5388 | 6553 | 6122 | 6003 | 12875 | 0 | 5850 | 12168 | 6224 | 13129 | 5812 | 6328 | 172 |
| Ab71 | 6158 | 5911 | 6142 | 13528 | 5850 | 0 | 12312 | 6429 | 13404 | 6232 | 5884 | 5841 |
| EF20 | 11620 | 12560 | 12223 | 13017 | 12168 | 12312 | 0 | 12579 | 12317 | 12352 | 12249 | 12146 |
| 6736 | 6241 | 6575 | 6777 | 13897 | 6224 | 6429 | 12579 | 0 | 13975 | 6199 | 5950 | 6197 |
| B362 | 13380 | 13588 | 13475 | 11104 | 13129 | 13404 | 12317 | 13975 | 0 | 13721 | 13687 | 13096 |
| C6344 | 7185 | 5638 | 4913 | 13338 | 5812 | 6232 | 12352 | 6199 | 13721 | 0 | 6332 | 5797 |
| S6261 | 6997 | 5809 | 6481 | 13803 | 6328 | 5884 | 12249 | 5950 | 13687 | 6332 | 0 | 6257 |
| C9359 | 6553 | 6117 | 5964 | 12817 | 172 | 5841 | 12146 | 6197 | 13096 | 5797 | 6257 | 0 |

The mean number of SNPs/bp for core and accessory genes was 0.0043 and 0.0042, respectively, implying slightly more variations in the core-genome repertoire. A frequency distribution plot of the mean number of SNPs/bp for core versus accessory genomes confirms the conservation within the accessory genome (Fig 5.3). There are a number of

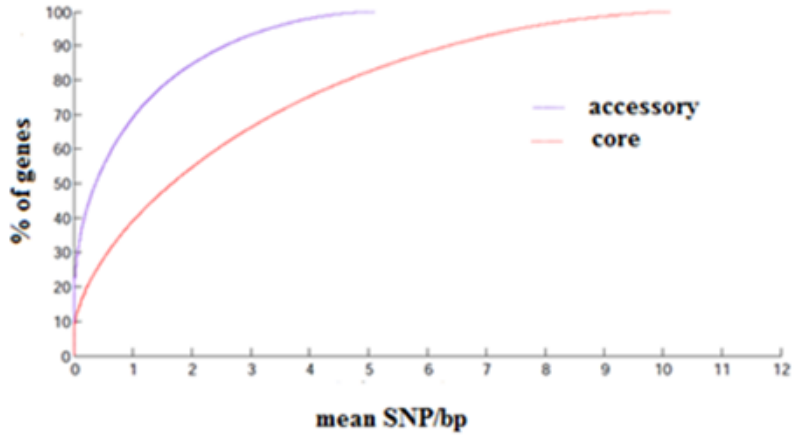



Figure 5.3: Cumulative frequency plot of mean SNP/bp of core and accessory genes. The shift of the distribution to the right shows that the core genes are on average more variable than accessory genes.

core genes with particularly high rate of variant positions (mean SNP/bp >0.05).Since the data were not normally distributed, the Mann-Whitney test was performed which

showed a significant difference ($p < 0.001$) between SNP density of genes of the core and accessory genomes (Fig 5.4). If the variation per base is plotted as a function of genome

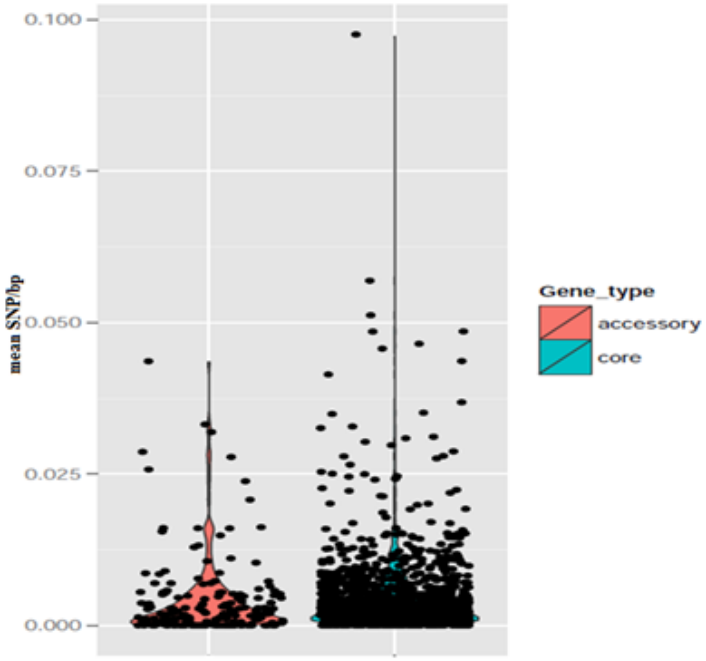



Figure 5.4: SNP distribution between accessory and core genes. As in Figure 5.3 the core genome has on average a higher mean SNP/bp. This is statistically significant (Mann-Whitney U Test, $p < 0.05$).

position, two groups of strains emerge in line with the whole genome phylogeny (Fig 5.5). Isolates B362 and 6780 cluster with the non-virulent isolate EF20 and this prompted us to consider comparative analysis of these groups instead of basing comparative analysis on the clinical and sub-clinical definition of isolates.Whilst the B362, 6780 and EF20 cluster shows generally higher variation, the region around the 60th centile shows a rapid increase in the number of SNPs (Fig 5.5). This region contains genes (*SUB0967-SUB1177*) (Additional file 5.1) (Section 2.16). Gene ontology (GO) (molecular function) analysis of these genes shows that the majority are involved in catalytic

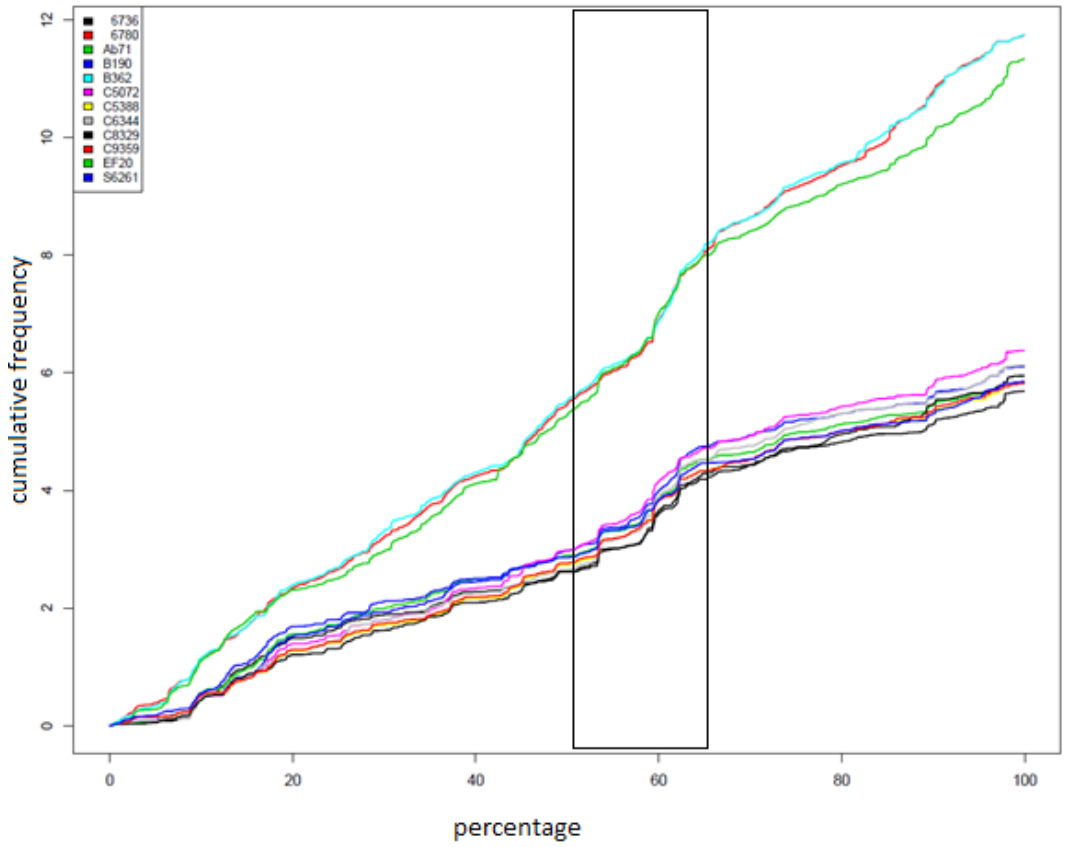



Figure 5.5: Cumulative frequency plot of SNP/bp of each genome as a function of percentage genome position. The rectangular bar highlights the region of rate change indicating the major region of highest variation between genomes. The two clear groups recapitulate the whole genome phylogeny seen in Figure 5.2.

activities such as nucleic acid binding, ATP and GTP binding (Fig 5.6(a)). GO analysis of cellular location shows that most of the genes (135 out of 447) are predicted to be membrane located (~ 30%), cytoplasmic (16%) and intracellular (16%) (Fig 5.6(b)). Comparison of core and accessory genes by GO molecular functional analysis shows that core genes are significantly enriched (chi square test $p < 0.05$) in transporter activity (57), catalytic activity (48), structural constituents of ribosome (39) and hydrolase activity (19) compared to the accessory genome content (Fig 5.6) (Additional file 5.1). Genes

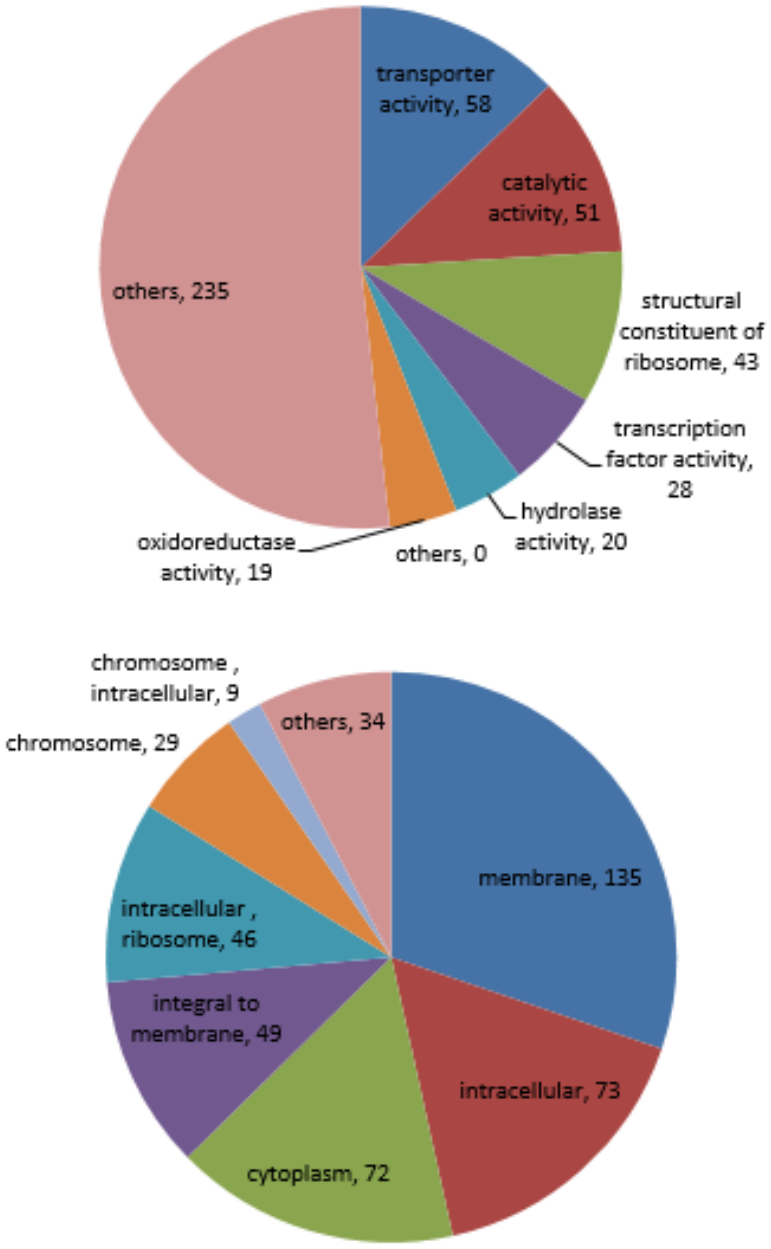
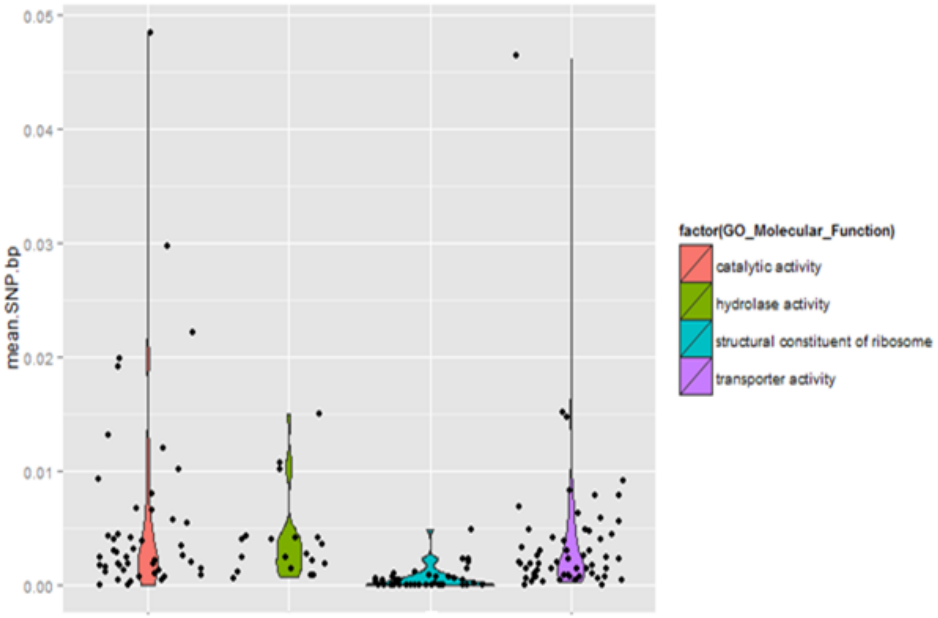


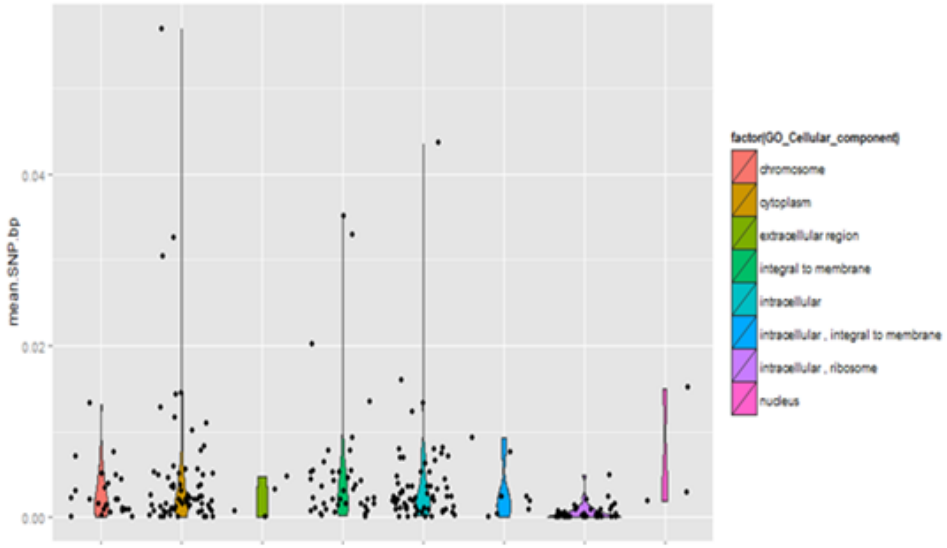

Figure 5.6: Distribution of different functional categories within the core genome using Gene Ontology (GO) terms (a) GO molecular function (454) (b) GO Cellular location (447).

involved in catalytic and transporter activity are more variable compared to the other two enriched categories (Fig 5.7).

Cluster analysis of these categories generally resulted in two major clusters (Fig 5.8-5.9) as found in whole genome phylogeny. We identified a number of genes (Table 5.2) from these cluster analysis such as: *SUB1407, SUB0933, SUB1638, SUB0890, SUB0182, SUB0690, SUB1514, SUB0083, SUB1632, SUB0082, SUB0083, SUB0096, SUB0713* and *SUB1633* which contribute to these two major clusters and they may be used as potential diagnostic markers in epidemiological studies in a relatively larger sample size to see the association of these variations with the pathogenesis of *S. uberis*.

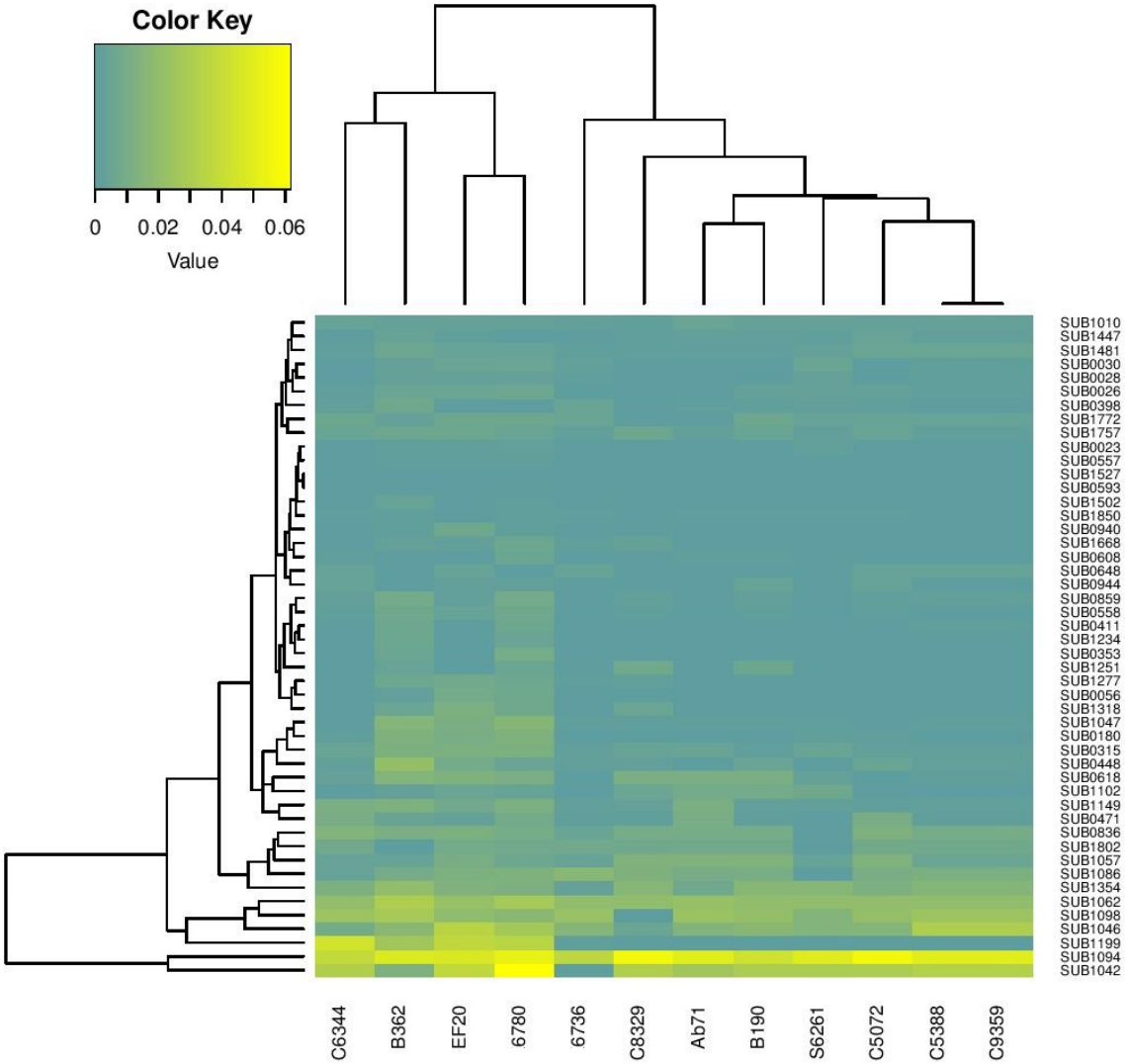



(a)

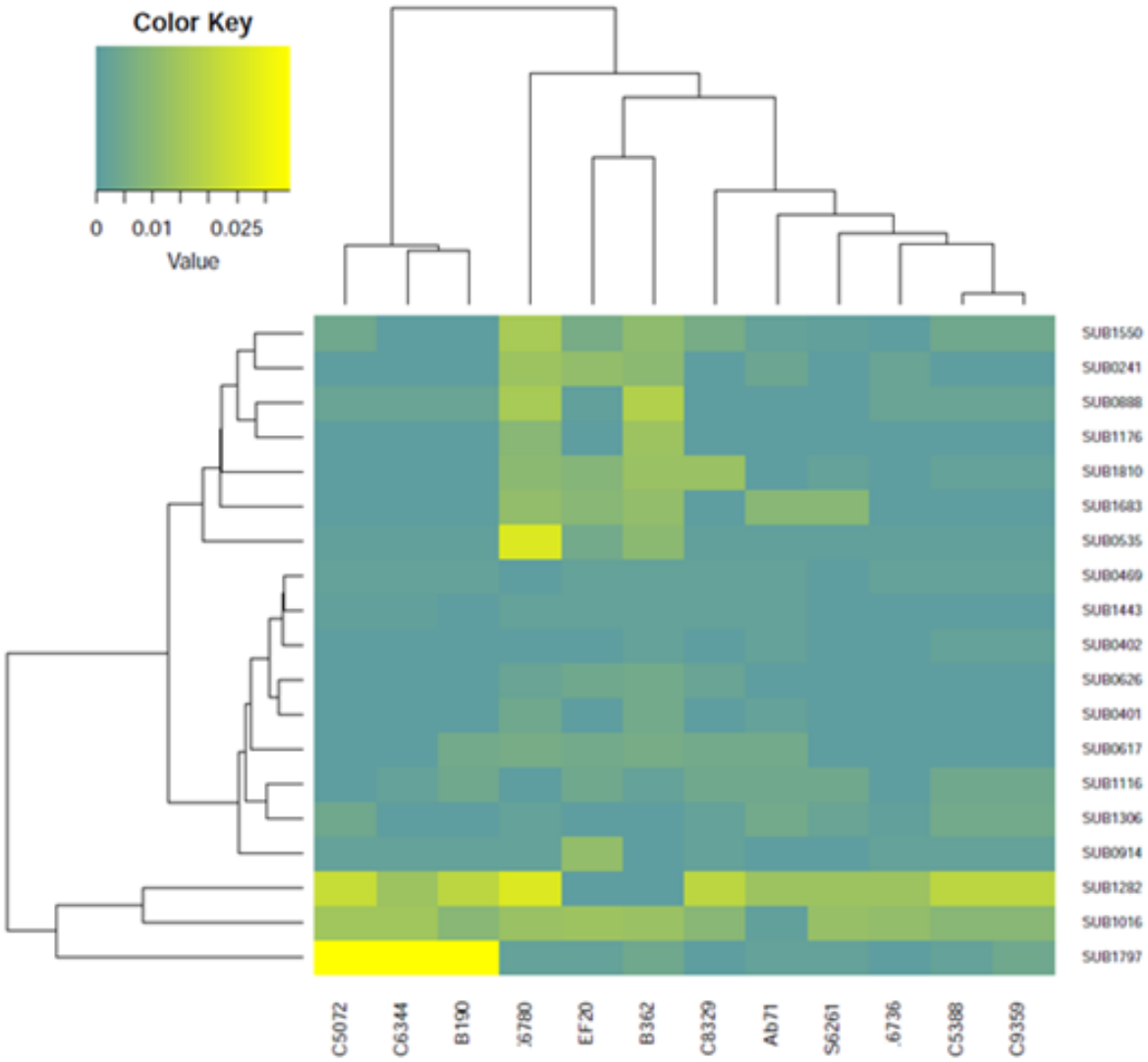



(b)

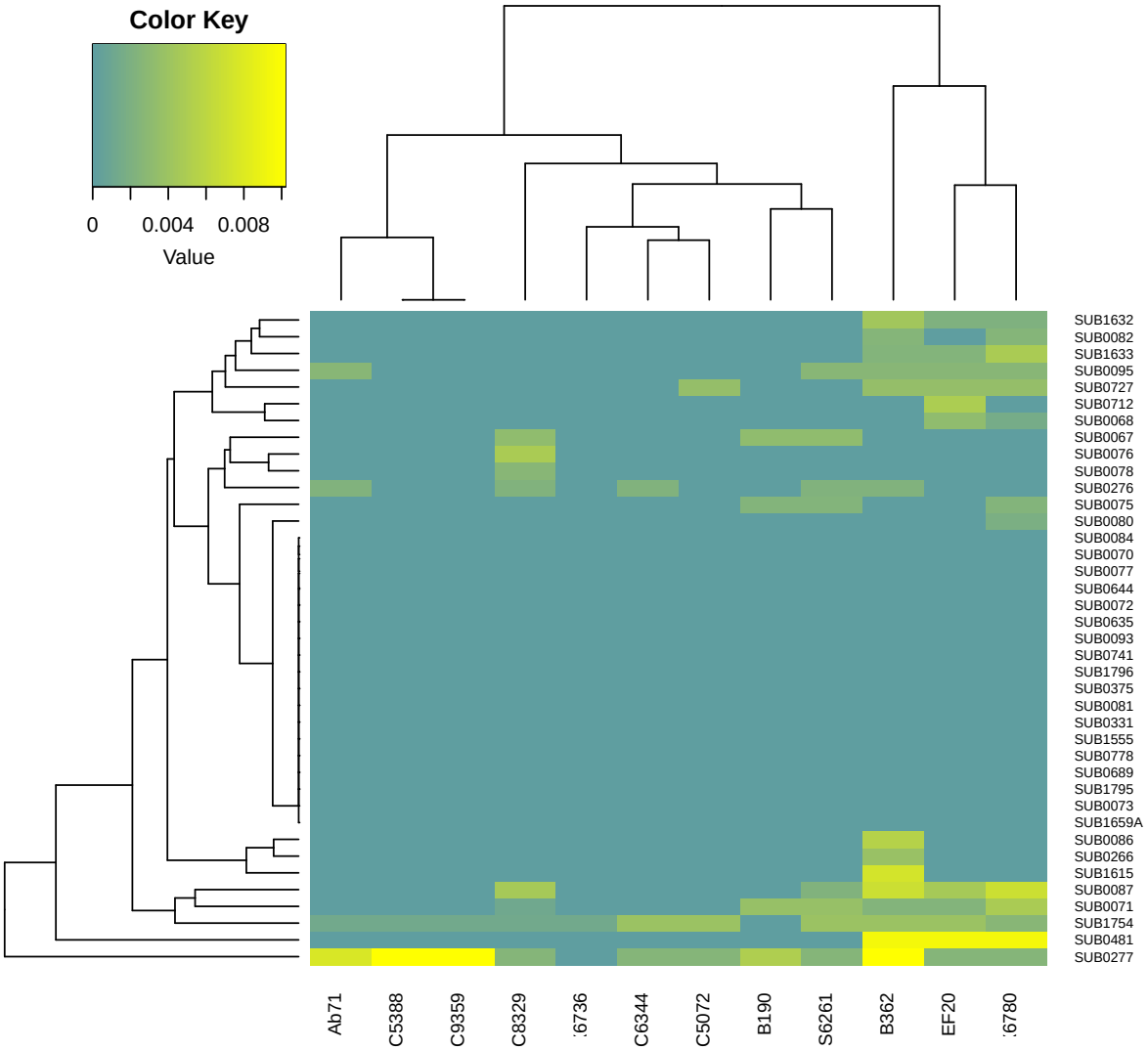
Figure 5.7: Violin plots showing the distribution of mean SNP/bp in the core genome of 12 isolates in enriched GO terms (a) GO Molecular Function (b) GO Cellular location.




5.8 (a) GO: catalytic

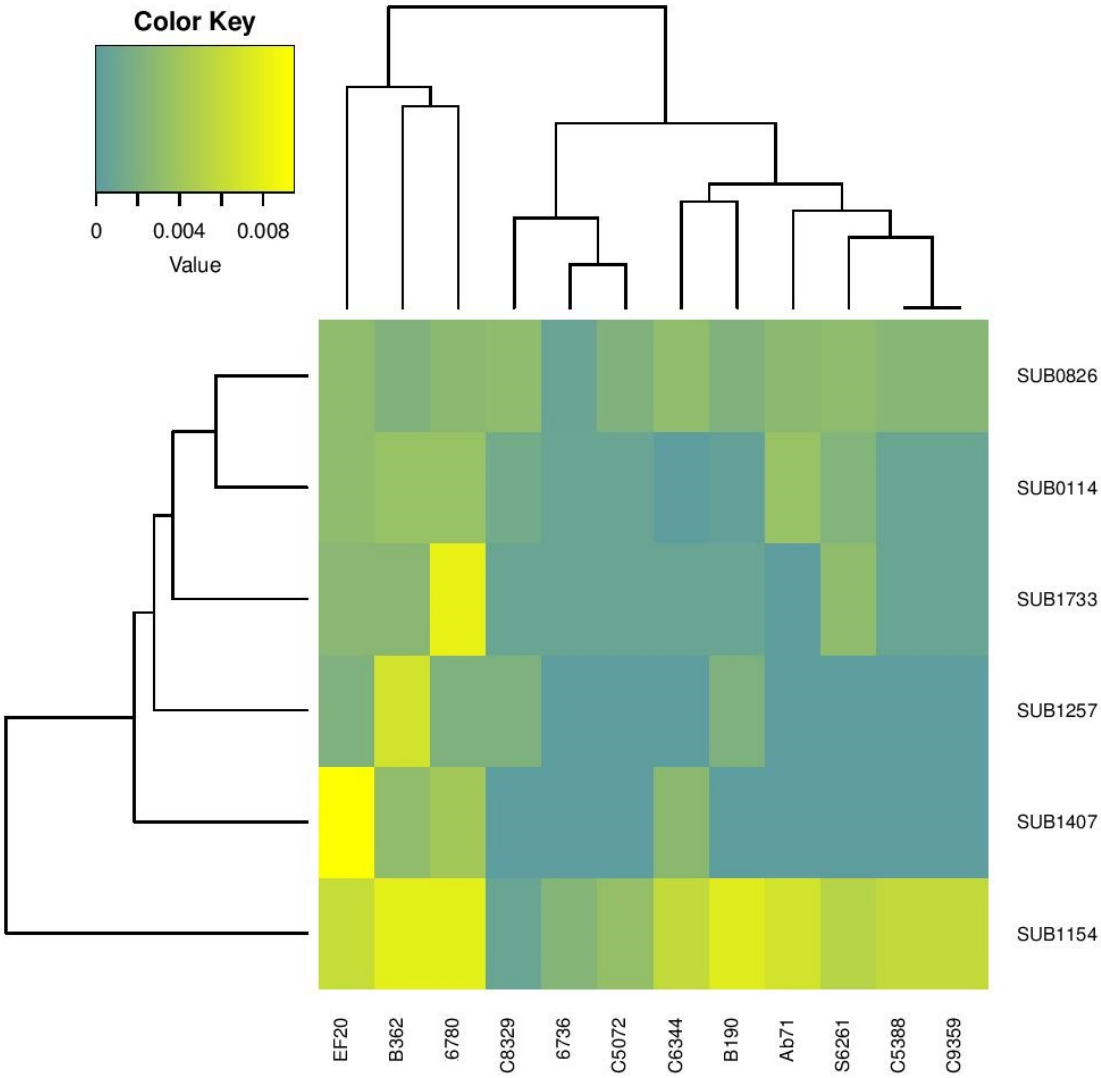



5.8 (b) GO: hydrolase

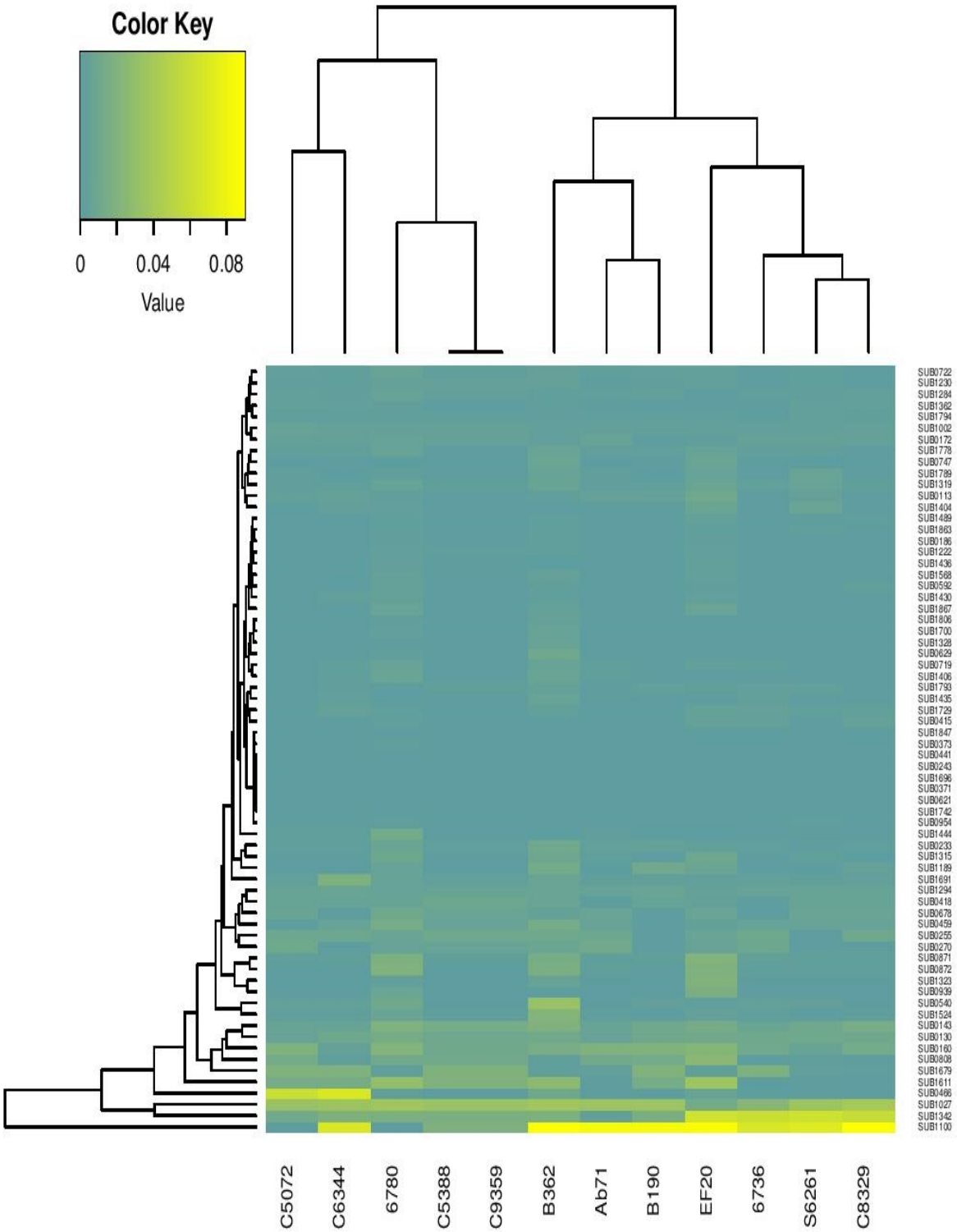



5.8 (c) GO: structural constituents of ribosome

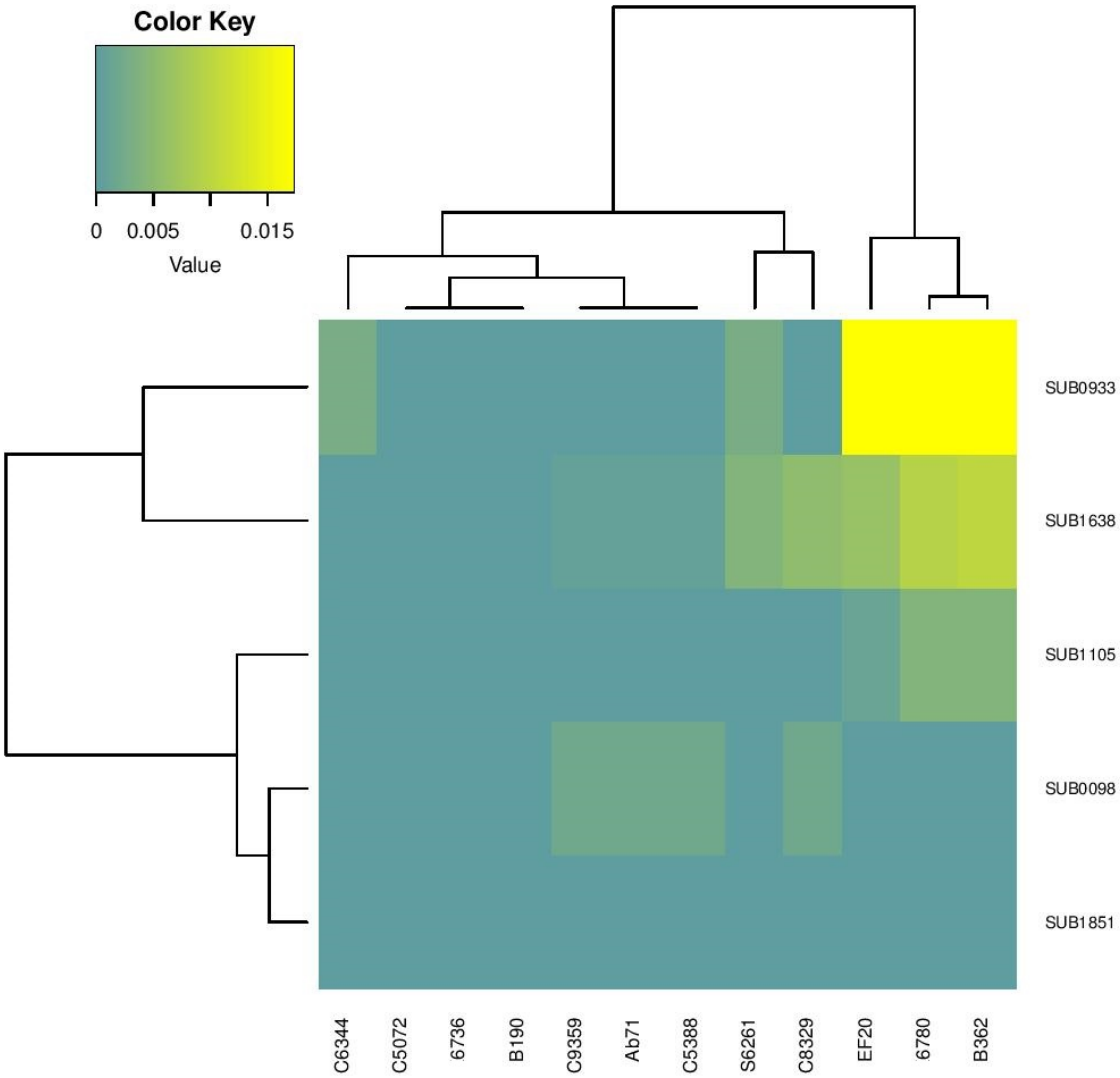
Figure 5.8: SNP density (SNP/bp) of individual genes associated with enriched GO molecular function categories. Heatmap visualization of clustering of genes with similar SNP density between *S. uberis* strains. Colours show SNP density, yellow = higher mean SNP/bp. (a) genes associated with catalytic activity (b) genes associated with hydrolase activity (c) genes associated with structural constituents of ribosome.



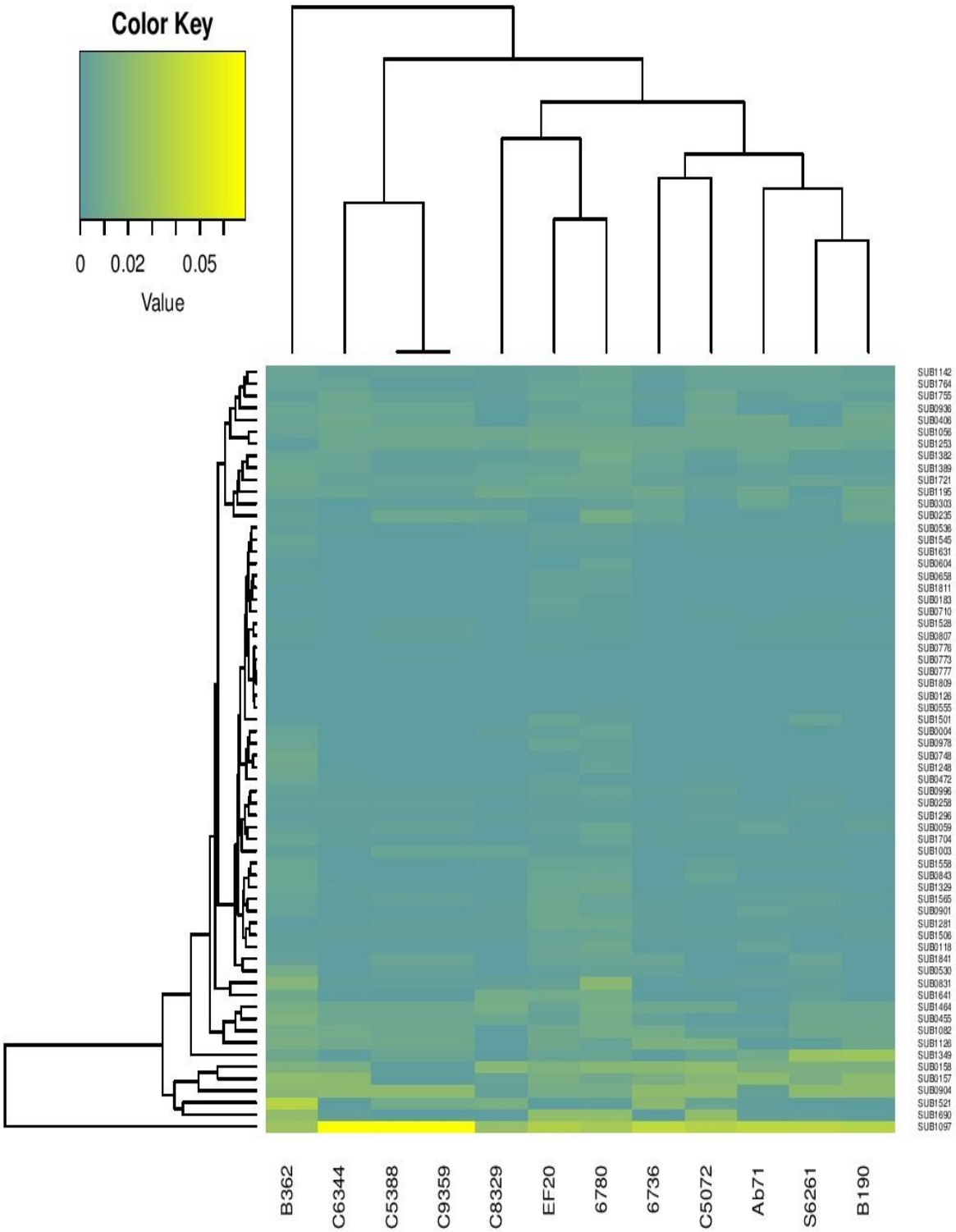

5.9 (a) GO: cell wall



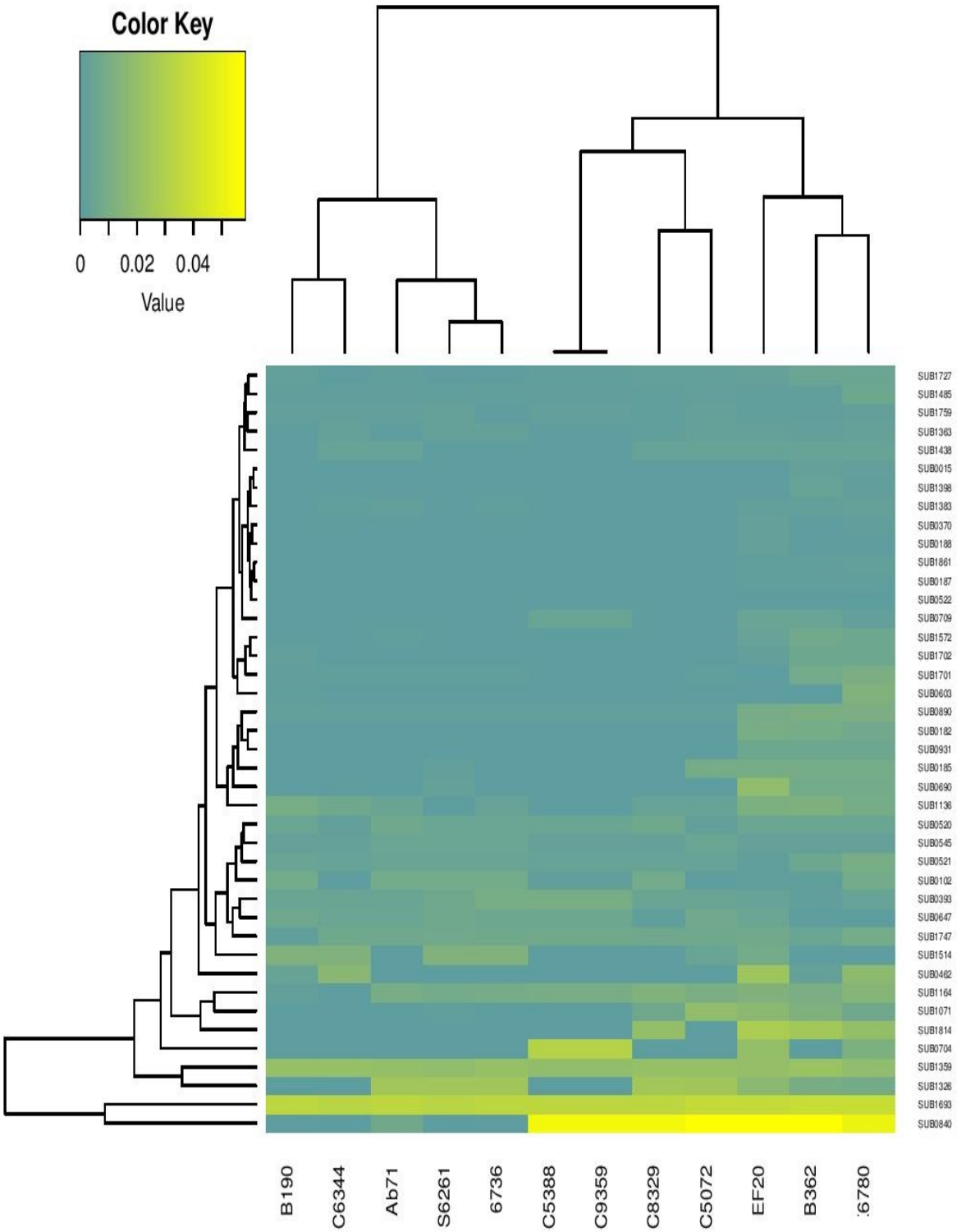

5.9 (b) GO: cytoplasm



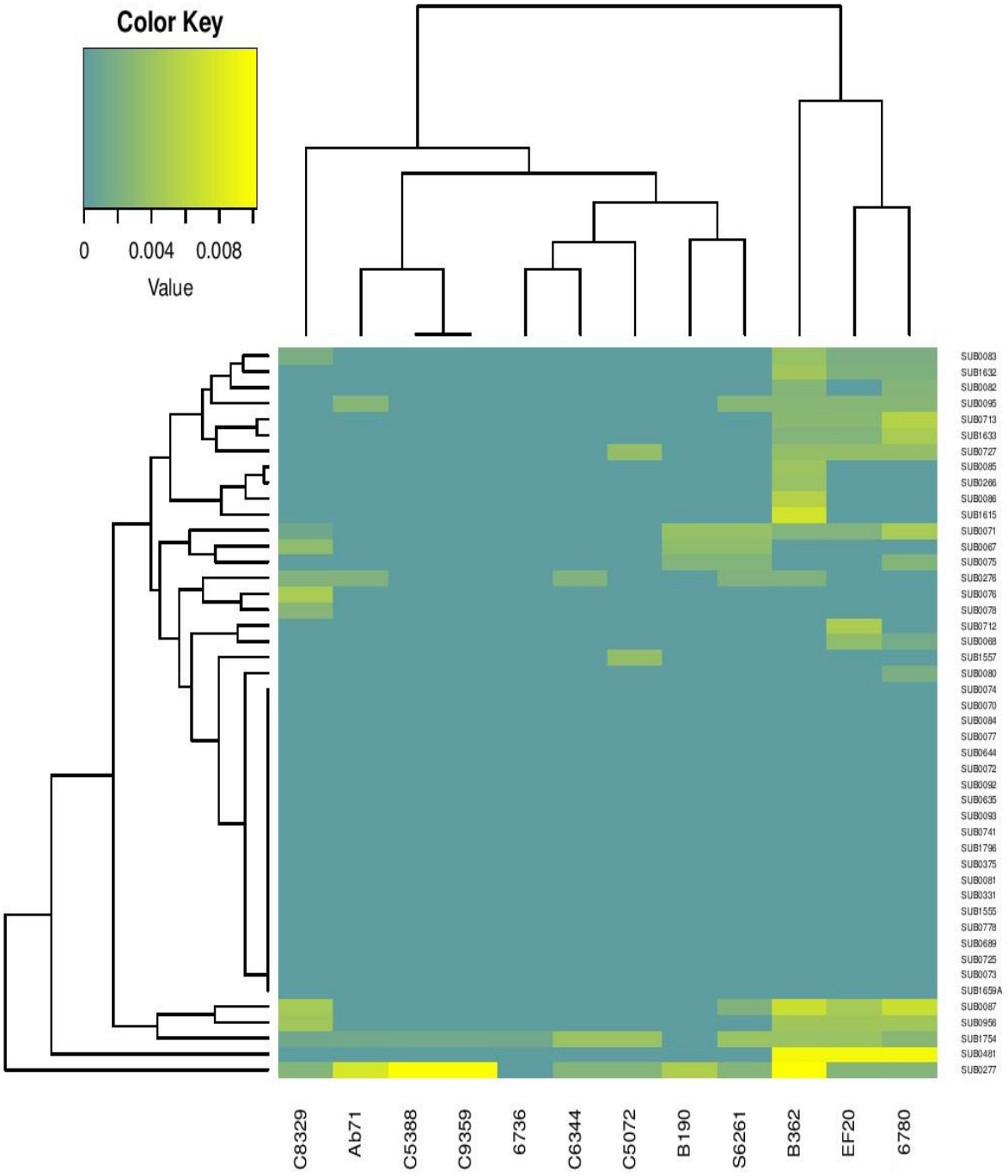

5.9 (c) GO: extracellular




5.9 (d) GO: intracellular




5.9 (e) GO: membrane




5.9 (f) GO: ribosome

Figure 5.9: SNP density (SNP/bp) of individual genes associated with enriched GO cellular location categories. Heatmap visualization of clustering of genes with similar SNP density between *S. uberis* strains. Colours show SNP density, yellow = higher mean SNP/bp. (a) cell wall (b) cytoplasm (c) extracellular (d) intracellular (e) membrane (f) ribosome.

Table 5.2: List of potential genes affecting the grouping of strains into two major categories.

| Gene | Function |
|---|---|
| *SUB1407* | *pbp1A* - penicillin-binding protein 1A |
| *SUB0933* | hypothetical protein |
| *SUB1638* | RmuC family protein |
| *SUB0890* | Na(+)/H(+) antiporter |
| *SUB0182* | sensor histidine kinase |
| *SUB0690* | *mscL* - large-conductance mechanosensitive channel; Channel that opens in response to stretch forces in the membrane lipid bilayer. May participate in the regulation of osmotic pressure changes within the cell (By similarity) |
| *SUB1514* | transglycosylase associated protein |
| *SUB0083* | *rplF* - 50S ribosomal protein L6; This protein binds to the 23S rRNA, and is important in its secondary structure. It is located near the subunit interface in the base of the L7/L12 stalk, and near the tRNA binding site of the peptidyltransferase center (By similarity) |
| *SUB1632* | *rpsG* - 30S ribosomal protein S7; One of the primary rRNA binding proteins, it binds directly to 16S rRNA where it nucleates assembly of the head domain of the 30S subunit. Is located at the subunit interface close to the decoding center, probably blocks exit of the E-site tRNA (By similarity) |
| *SUB0082* | *rpsH* - 30S ribosomal protein S8; One of the primary rRNA binding proteins, it binds directly to 16S rRNA central domain where it helps coordinate assembly of the platform of the 30S subunit (By similarity) |
| *SUB0083* | *rplF* - 50S ribosomal protein L6; This protein binds to the 23S rRNA, and is important in its secondary structure. It is located near the subunit interface in the base of the L7/L12 stalk, and near the tRNA binding site of the peptidyltransferase center (By similarity) |
| *SUB0096* | *araC* - family regulatory protein (fragment) |
| *SUB0713* | *rplT* - 50S ribosomal protein L20; Binds directly to 23S ribosomal RNA and is necessary for the in vitro assembly process of the 50S ribosomal subunit. It is not involved in the protein synthesizing functions of that subunit (By similarity) |
| *SUB1633* | *rpsL* - 30S ribosomal protein S12; With S4 and S5 plays an important role in translational accuracy (By similarity) |

### 5.2.3 Variation of virulence genes

As discussed in Chapter 4, the presence of previously predicted virulent genes [216] are conserved in clinical and non-clinical strains and this led to the assumption that genomic variation rather than gene gain/loss within the isolates might play significant roles in their pathogenesis. We have shown that isolates do exhibit a range of sequence variations between strains. Since these putative virulence genes and sortase-anchored proteins are important for pathogenesis the variation among them were further analysed using DNASP [174].

PauA (*SUB1785*) and Hemolysin-like protein (*SUB1273*) are very well conserved between the 12 strains with only 3 and 5 variable sites, respectively in genes of over 800 bp in length (Table 5.3). In contrast, Collagen like surface-anchored protein (*SUB1095*) and Lactoferrin binding protein (*SUB0145*) are more variable with over 0.4 SNPs/bp. Analysis of variations among 10 potential sortase anchored proteins (*SUB0135, SUB0145, SUB0207, SUB0241, SUB0826, SUB0888, SUB1095, SUB1154, SUB1370* and *SUB1730*) showed that *SUB0145* and *SUB1095* are highly variable between strains (Table 5.3) suggesting that variation between them is maintained by natural selection in turn suggesting that this may be driven by interaction with the host immune system. Conversely, *SUB1154* is more conserved between strains (0.021 SNPs/bp aligned). In nine *S. uberis* strains (Chapter 4) a paralog of *hasB(SUB1027)* was identified. The non-capsular, non-virulent isolate EF20 lacks *SUB1027* and this gene is also missing from isolates B362, 6780 and B190 implying the gene gain/loss related to the genomic variation of isolates as whole genome analysis clusters EF20, 6780 and B362 into a single cluster. *SUB1696* and

*SUB1697* are highly conserved among the isolates and no variation was found within the isolates, however, few genomic variations were found in *SUB1691 (hasC)* and this will be interesting to note if these variations in *hasC* are important in the capsule production of *S. uberis*.

Table 5.3: SNP distribution of (a) known virulence genes (b) sortase-anchored proteins (c) has genes. *denotes genes previously associated with virulence which also contain a predicted sortase-anchor.

| Gene | Gene function | SNP count | Gene size (bp) | Alignment size (bp) | SNP/bp | "SNP/bp |
|---|---|---|---|---|---|---|
| (a) | | | | | | |
| *SUB1785* | PauA (Streptokinase precursor) | 3 | 861 | 861 | 0.003 | 0.01 |
| *SUB1273* | Hemolysin like protein | 5 | 828 | 828 | 0.006 | 0 |
| *SUB0881* | Sortase A (*srtA*) | 8 | 759 | 759 | 0.011 | 0.01 |
| *SUB1154* | C5a peptidase precursor | 73 | 3480 | 3362 | 0.022 | 0 |
| *SUB1635* | *SUAM* protein | 52 | 2637 | 2637 | 0.02 | 0.04 |
| SUB1111 | Fibronectin- binding protein | 42 | 1653 | 1653 | 0.025 | 0.01 |
| (b) | | | | | | |
| *SUB0135* | putative fructan beta-fructosidase precursor | 36 | 3810 | 3810 | 0.009 | NA |
| *SUB0145* * | lactoferrin binding protein | 552 | 1819 | 1352 | 0.408 | 0.06 |
| *SUB0207* | putative surface-anchored protein | 28 | 1500 | 1500 | 0.019 | NA |
| *SUB0241* | putative surface-anchored 2',3'-cyclic-nucleotide 2'-Phosphodiesterase | 48 | 2478 | 2457 | 0.02 | NA |
| *SUB0826* | putative surface-anchored subtilase family protein | 208 | 4492 | 4418 | 0.047 | NA |
| *SUB0888* | putative surface-anchored protein | 20 | 852 | 837 | 0.024 | NA |
| *SUB1095* * | collagen-like surface-anchored protein | 375 | 1452 | 935 | 0.26 | 0.08 |
| *SUB1154* | C5A peptidase precursor | 73 | 3480 | 3435 | 0.021 | NA |
| *SUB1370* | putative zinc carboxypeptidase | 58 | 3225 | 3213 | 0.018 | NA |
| *SUB1730* | putative surface-anchored protein | 128 | 1191 | 1020 | 0.125 | NA |
| (c) | | | | | | |
| *SUB1691* | *hasC*: UDP-glucose pyrophosphorylase | 22 | 1017 | 1017 | 0.022 | 0 |
| *SUB1696* | *hasB*: UDP-glucose dehydrogenase | 0 | 1206 | 828 | 0 | 0 |
| *SUB1697* | *hasA*: hyaluronan synthase | 1 | 1254 | 1254 | 0.008 | 0 |

Variation analysis of these putative virulence genes (Fig 5.10) does not follow the whole genome phylogeny and most of the other clustering in which EF20 clusters with 6780 and B362, implying that other factors identified can also be used as diagnostic markers relating to the pathogenesis of *S. uberis*. *S. uberis SUB0144* is a homologue

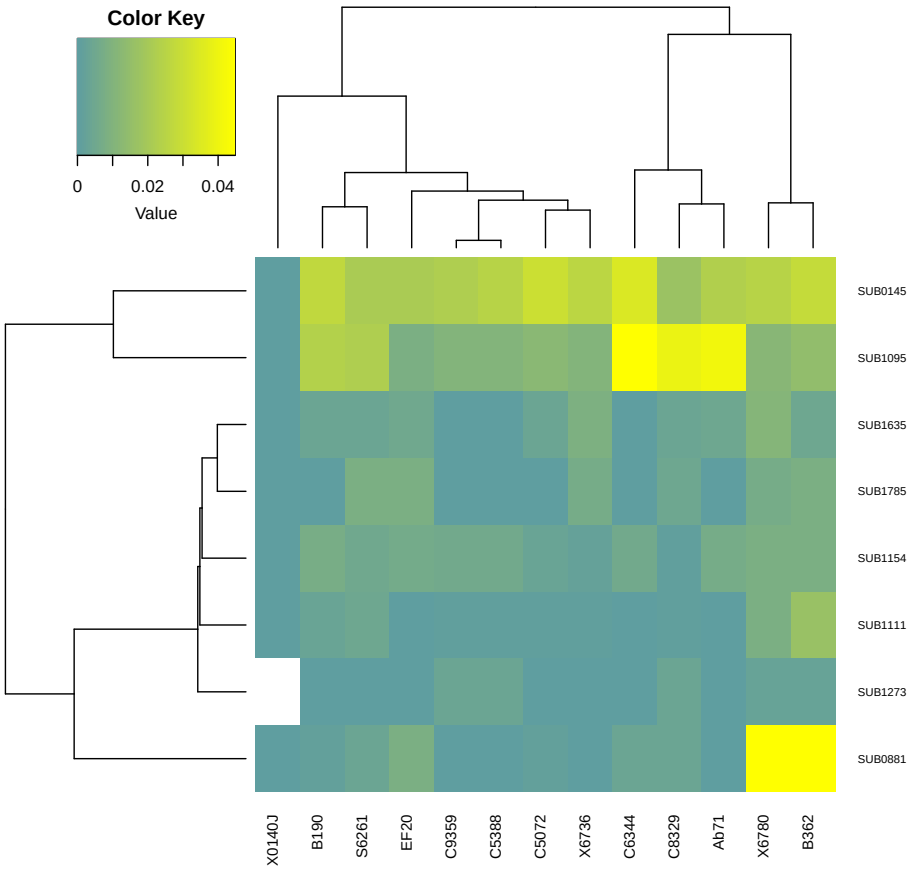



Figure 5.10: SNP density (SNP/bp) of putative virulence genes. Heatmap visualization of clustering of genes with similar SNP density between *S. uberis* strains. Colours show SNP density, yellow = higher mean SNP/bp.

of the *S. pyogenes* virulence regulatory gene *mga*. *SUB0144* (*vru*) of *S. uberis* has been found to regulate a number of virulent genes including *hasA* and *hasB1* (*SUB1696* an *SUB1697*), Lbp (*SUB0145*), SclB (*SUB1095*) and PauA (*SUB1785*) and inactivation

of *vru* resulted in reduced ability to colonize the mammary gland as well as reduced clinical signs of mastitis compared with the wild-type strain [54]. Moreover, Flores et al. showed that a 12-bp deletion in the VNTR region of *mga* promoter at positions -63 to -75 alters GAS virulence, resulting in asymptomatic carrier phenotype [65]. In *S. uberis* we observe a deletion of five bp in a similar region of vru (position -75 to -79). This deletion was seen in several isolates including the non-virulent strain EF20 and 6736. A four bp deletion was seen in positions -76 to -79 in five isolates Ab71, C9359, B362, C5388 and C8329 (Fig 5.11). Whilst not perfectly segregating, this deletion was found in most of the sub-clinical isolates and only in three clinical isolates. Hence, variation in this region may play an important role in the regulation of this regulatory gene and in turn influence the host-pathogen interaction. We manually investigated the polymorphism of 100bp upstream regions of some of the predicted virulent genes (Table 5.3) that may have potential effect on the regulation of respective virulent genes. *SUB0145* (*lbp*), *SUB1095* (surface protein) and *SUB1635* (SUAM protein) showed considerable variation whereas most of them were highly conserved with little or no variation. These data suggests that even within the core virulence genes variations exists in their genomic sequences and these may be important for pathogenesis. Such variations can also be used as markers for the rapid diagnosis of strains with increased propensity to cause mastitis and can shed light to the identification of potential vaccine target.




Figure 5.11: Variation in the *vru* upstream regions. Alignment of upstream region of *vru* gene across 13 strains was generated using Muscle [52]. Boxshade server (http://www.ch.embnet.org/software/BOX_form.html) was used to highlight the high degree of conservation within the aligned region. The position of the initiating methionine codon (ATG) is shown with an arrow. Deletion of TATAA was found in isolates EF20, 6736 and 6780 in position -75 to -79 of *vru* gene. Polymorphism of A to T found in region -79 along with the deletion of four bases TATA in regions -75 to -78 of isolates B362 and C8329. Deletion of four bases ATAA (-76 to -79) was found in 3 isolates Ab71, C5388 and C9359.

### 5.2.4 Genes under positive selection

DNA codon based alignments of orthologous gene families from 1825 genes were generated. Genes annotated as pseudogenes were not alligned and 209 alignments were removed due to suspected alignment errors, frameshift mutations, or because of ambiguous assignment of orthologs. For the remaining 1616 genes, the ratio of dN/dS was determined. Among the 1530 core orthologous genes, 26 showed significant evidence of positive selection (Table 5.4).

Genes under positive selection are enriched in transmembrane transporter activity, oxidoreductase activity, DNA binding, organic cyclic compound binding, catalytic activity which are required for fundamental life processes of bacteria. Analysis of cellular compartment of these genes showed no exceptional enrichment (Fig 5.12). To investigate further, the correlation between the rates of evolution and predicted cellular compartments (intracellular, extracellular and membrane) was analysed. The dN/dS values for the three classes showed that proteins located in the membrane show relaxed selective constraint and tend to have a higher dN/dS ratio (Fig 5.13). Since, only four genes are annotated to be extracellular this may provide misleading information compared to other categories.

Analysis of selected molecular function categories showed that catalytic proteins are evolving faster than transporter encoding genes and hydrolase annotated genes are evolving under strong purifying selection (Fig 5.13).

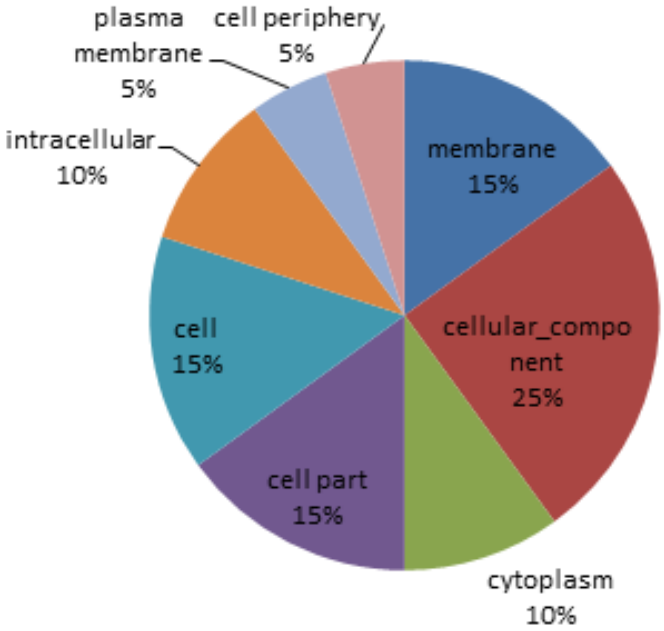



Figure 5.12: Predicted Cellular location of genes under positive selection using Gene Ontology (GO) categories.

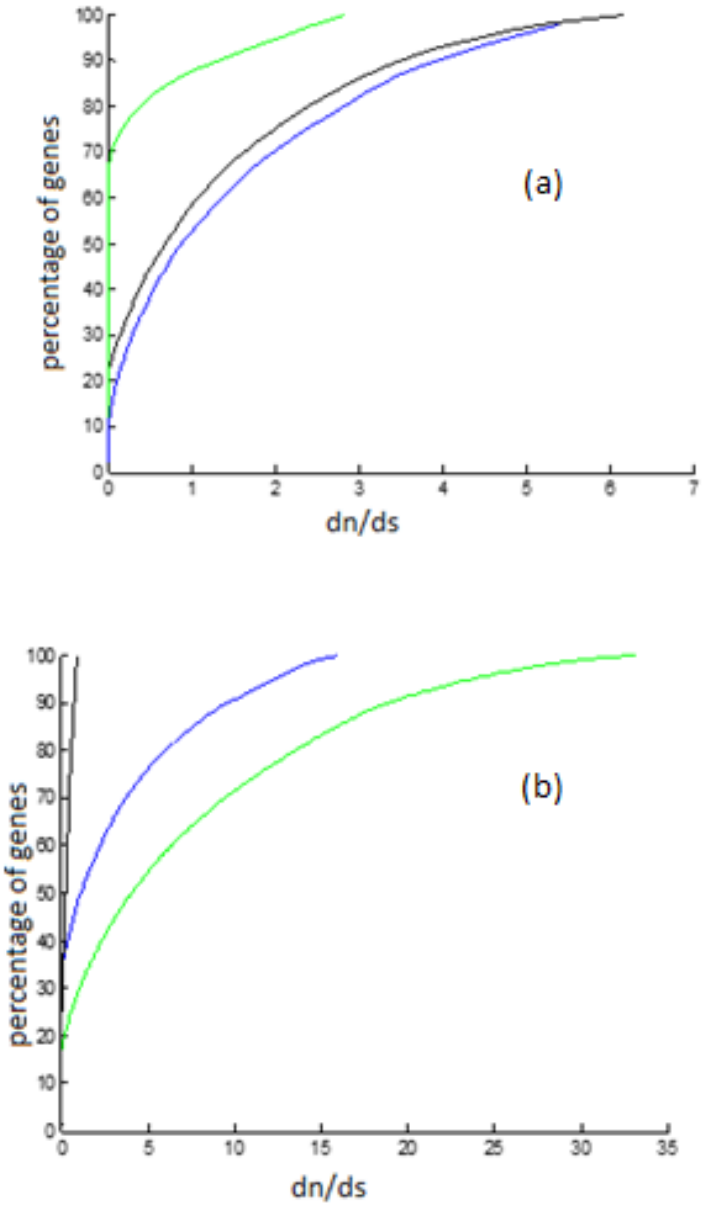



Figure 5.13: Cumulative frequency plot of intensity of evolutionary selection (ratio of non-synonymous/synonymous substitutions (dn/ds) for genes predicted to be evolving under positive selection. (a) Comparison of functional groups : green - hydrolase; black - transporter and blue - catalytic (b) Comparison of different cellular locations: blue - intracellular ; green - membrane ; black - extracellular. The extracellular category contained only 4 genes hence the skewed distribution seen.

Table 5.4: Genes under positive selection.

| Gene | Function |
|---|---|
| *SUB0133* | thioredoxin (104 aa) |
| *SUB0193* | hypothetical protein (56 aa) |
| *SUB0238* | membrane protein (165 aa) |
| *SUB0526* | lactoylglutathione lyase (125 aa) |
| *SUB0597* | membrane protein (182 aa) |
| *SUB0653* | hypothetical protein (190 aa) |
| *SUB0789* | replication initiator protein (274 aa) |
| *SUB0828* | transcription antiterminator |
| *SUB0918* | transporter permease (651 aa) |
| *SUB1007* | lipoprotein (207 aa) |
| *SUB1228* | membrane protein (319 aa) |
| *SUB1239* | glucosamine-6-phosphate isomerase; Catalyzes the reversible isomerization-deamination of glucosamine 6-phosphate (GlcN6P) to form fructose 6-phosphate (Fru6P) and ammonium ion (234 aa) |
| *SUB1339* | membrane protein (118 aa) |
| *SUB1401* | hypothetical protein (77 aa) |
| *SUB1495* | 3-oxoacyl-ACP synthase; Catalyzes the condensation reaction of fatty acid synthesis by the addition to an acyl acceptor of two carbons from malonyl-ACP (410 aa) |
| *SUB1578* | membrane protein (32 aa) |
| *SUB1580* | membrane protein (113 aa) |
| *SUB1581* | membrane protein (114 aa) |
| *SUB1597* | hypothetical protein (35 aa) |
| *SUB1599* | putative transposase (fragment)(113aa) |
| *SUB1695* | MATE family efflux protein (452 aa) |
| *SUB1745* | cold shock protein (67 aa) |
| *SUB1746* | hypothetical protein (89 aa) |
| *SUB1768* | anaerobic ribonucleoside-triphosphate reductase activating protein (204 aa) |
| *SUB1798* | putative DNA-binding membrane protein (54aa) |
| *SUB1840* | phage integrase (388 aa) |

## 5.3 Discussion

It was not possible to differentiate between clinical and sub-clinical isolates simply by analysis of the gene presence/absence rendering the identification of clinically relevant *S. uberis* isolates more difficult. To address this, analysis of genomic variation of the complete genome sequences of *S uberis* was conducted.

Standard microbial genotyping methods such as PFGE and MLST are commonly used during a disease outbreak investigation or in determining evolutionary relationships. These approaches do have some limitations [186]. Comparative and evolutionary analysis using MLST is based on single nucleotide polymorphisms of only seven conserved housekeeping genes. Both of these techniques are unable to identify indels, duplications, or other important genomic events such as gene gain/loss which may affect the virulence properties of pathogens. Several whole genome studies were able to distinguish between pathogenic and non-pathogenic strains [186]. In this study, using unbiased computational approaches we found isolates clustering into two major clades with the virulent and non-virulent isolates. Further analysis using larger sample size may shed light into the variation among the virulent and non-virulent *S. uberis*.

Recombination is a possible confounding factor in bacterial whole genome phylogenetics. However, this did not appear to be a major concern in this study. Posada & Crandall suggested that the effect of recombination on phylogeny depends on the similarity of the sequences and the relative size of regions involved in the recombination event. They found that phylogenetic history was correctly inferred for closely related taxa where the recombination event was ancient and hence shared by multiple taxa. However, for diver-

gent sequences with recent recombination events inference of the incorrect topologies was more common [165].

Strain specific live vaccines have been found to produce some protection, however, antigens responsible for such protection have not been identified [64]. Ward et al. suggested a lack of classical virulence factors in 0140J by analyzing the whole genome sequence. This is confirmed by clustering of these putative virulence factors which did not group isolates into either clinical and sub-clinical or virulent and non-virulent clades. This implies the significant role of other factors in conjunction with these putative virulence factors in the diverse pathogenesis of *S. uberis*. We have identified a number of genes including *SUB1407, SUB0933, SUB1638, SUB0890, SUB0182, SUB0690, SUB1514, SUB0083, SUB1632, SUB0082, SUB0083, SUB0096, SUB0713* and *SUB1633* from different functional, cellular location categories which potentially contribute to the variation of two major clusters. It was interesting to note that among these 14 genes 6 were either 50S or 30S ribosomal proteins with different functional categories.

A number of core genes were more variable than accessory genes and this could be due to the fact that during adaptation to different environmental niches, core genes are changing and hence more variable whilst accessory genes are being acquired or lost during transition from one environment to another. GO enrichment showed that *S. uberis* genes with high variability are enriched in several functional and cellular location categories. Most of the core genes were found to be associated with catalytic, transporter, hydrolase and structural components of ribosomes. Catalytic and transporters genes were more variable compared to the other two categories. Catalytic enzymes are impor-

tant for maintaining different cellular chemical reactions. Variations of such enzymes may enable survival under different environmental conditions by allowing subtle control of chemical reactions in altered environments. Transporter genes are important for nutritional uptake and adaptation in different environmental conditions. The variation of transporter genes might play important roles in the survival of this pathogen in different niches.

Variation in outer membrane proteins have been found to be involved in bacterial pathogenesis by allowing evasion of the host immune system. A number of bacterial membrane proteins such as PorA have been used as the typing method of *Neisseria gonorrhoeae*.In the present study we found several membrane proteins including: *SUB0709, SUB1572, SUB0890, SUB0182, SUB0931, SUB0690* which may be involved in the evolutionary dynamics of two major categories i.e virulent and non-virulent. As such these genes may be used as genotypic markers for these groups.

Adaptation and transmission of *S. uberis* in different environmental conditions and various host niches requires co-ordinated regulation of gene expression in different challenging environmental conditions [216]. The putative gene regulator *vru* has been found to regulate a number of virulent genes [61]. Homologs of *vru* (*mga* and *mgc*) have been shown to regulate virulent genes in other Streptococci [35, 73]. Although changes of expression of putative virulent genes have been observed in *S. uberis*, it was not clear whether *vru* directly or indirectly effected such expression [54]. Egan et al. speculated that the disruption of the *vru* ORF in the vicinity of its start codon was likely to be influencing *SUB0145* expression through altered expression of its gene product, rather

than disruption of a possible shared transcriptional promoter [54]. Evidence for the importance of the *vru* homolog in virulence gene regulation has been suggested by various studies. *mga* the homolog of *vru* has been shown to regulate a number of GAS virulence genes and Flores et al.identified a 12-bp deletion in the promoter of *mga* in isolates from asymptomatic carrier state, but this was absent from a known invasive serotype [65]. This identification led us to investigate ~ 100bp upstream region of *vru* in *S. uberis* and identified a deletion of five base pairs (TATAA) in isolates EF20, 6736 and 6780 in position -75 to -79. A corresponding four bp deletion was found in a number of isolates most of which are sub-clinical including B362. This deletion was absent in the virulent isolate 0140J. This indel might provide advantage for the expression of virulence genes regulated by *vru*. Studies showed that 12-bp deletion in the *mga* promoter contributes to decreased virulence [65].

Genes undergoing positive selection have been found to drive various processes including: adaptation to different environmental niches, evading host immune systems and functional diversification [4]. The 26 genes identified as having evolved under positive selection were connected to a wide variety of functions such as replication, carbohydrate metabolism, transporter, phage integrase and a substantial fraction was annotated as hypothetical and membrane proteins with unknown function. Interestingly none of the putative virulence factors was found to be under positive selection. This highlights the complexity of this system and suggests that consideration of novel genes as possible virulence factors by contributing to escape from the immune recognition or adaptation to the host environment. Genome-wide analysis of *S. uberis* protein-coding genes uncovered a

number of genes that can be connected to virulence and subject to positive selection pressure. Moreover, identification of these genes also implies the role of different virulence mechanisms than previously considered. We argue that the evolution of pathogen specific genes does not provide a complete picture of the evolutionary dynamics of virulence. There is no doubt that the gene gain/loss, genomic variation and diversity in the intergenic region affecting the colonisation and exploitation of novel niches are critical to the pathogenesis of *S. uberis*. The outcome of this study and other Streptococci [77], suggest that the molecular mechanism of virulence are more complex than initially expected.

CHAPTER 6

# COMPUTATIONAL ANALYSIS OF INTER- AND INTRASPECIES PROTEIN-PROTEIN INTERACTION NETWORKS

## 6.1 Introduction

As proteins usually interact with other protein partners to perform their functions, microbial physiology and cellular processes are generally governed by protein-protein interactions(PPIs) [169]. The study of bacterial protein-protein interaction can therefore provide important insights into the molecular mechanisms of gene expression and global gene regulatory mechanisms. In addition, the identification of protein-protein interactions may help with the annotation of hypothetical proteins without known function. In the case of a pathogenic organism, PPI can be within a single species or can be in-

traspecies between host and pathogen. In the case of host-pathogen interaction PPI can, for example, provide an understanding of how pathogenic microorganisms manipulate host mechanisms to escape from host immune responses [169].

Various high-throughput experimental approaches such as yeast two-hybrid (Y2H), mass spectrometry (MS), protein microarrays and tandem affinity purification (TAP) have been performed to predict and analyze PPI networks in many species, such as *Saccharomyces cerevisiae* [181], *Drosophila melanogaster* [75], *C. jejuni* [158], *E. coli* [26], *Helicobacter pylori* [96] and *Leptospira interrogans* [195].

These experimental approaches whilst beneficial do require a large investment of time and money, and also have high false prediction rates [163]. Therefore, they are not routinely conducted for many species. To overcome this, the use of computational approaches to infer possible PPIs are a more pragmatic and feasible way of predicting putative pathogen-host interaction. In this study, we interrogated the PPI of *S. uberis* 0140J (data available from the STRING database) [197].

Host-pathogen interaction data was analyzed by applying homology based approaches to identify known host-pathogen interactions from existing human and other bacterial pathogen PPI data (downloaded from the PATRIC database) [74]. For the homology based host-pathogen interaction study comparative analysis was carried out on the virulent isolate 0140J and non-virulent isolate EF20 strains only.

## 6.2 Results and Discussion

### 6.2.1 *Streptococcus uberis* protein-protein interaction

Following generation of graphs for complete and sub sets of the *S. uberis* proteome (Fig 6.1), simple network statistical parameters such as clustering coefficient, network diameter, network radius, average number of neighbours, node degree distribution etc. were estimated using Cytoscape [190] (Fig 6.2).

**Clustering coefficient:** the degree to which nodes cluster together to form sub networks.

**Node degree:** the number of edges that connect to a specific node. This value can also be called vertex degree.

**Minimum (shortest) path length:** the fewest number of edges that need to be traversed to connect two nodes.

**Betweenness centrality:** the number of shortest paths going through a node.

To determine if genome components or different functional categories of proteins exhibited different parameters, eight categories were compared. These were, the core and accessory genomes, proteins involved in catalytic and hydrolase functions, transporters and proteins located in the ribosome, intracellular or at the bacterial membrane. Albert et al. [1] suggested that biological networks are usually scale-free which means that few nodes have higher than expected connectivity, signifying the importance of so called "hub" proteins in the physiological processes of the cell. Topological properties of predicted gene-gene network in different categories indicates that the network of *S. uberis*

interactions is scale free.

Proteins encoded by the core genome are more connected than the accessory genome due to the much higher number of edges within the core genome, skewing the ratio of these to be much higher for the former (~ 67 and ~ 8 , respectively) (Fig 6.2 e). This highlights the interaction of the core genome and implies that essential genes are mostly within the core genome and involved in the global gene regulation of *S. uberis*. Comparative analysis of core and accessory genomes shows that the core genome has higher scores for several important network parameters due to higher connectivity including clustering coefficient, betweenness centrality and node degree distribution (Fig 6.2-6.3). Since most of the 0140J genes analyzed are within the core genome the complete and core genomes are more similar compared to the accessory genome (Fig 6.2-6.3).

The number of edges were considerably higher (7042) in catalytic groups compared other functional groups such as hydrolase and transporters. Compared to the intracellular group the membrane located subcategory had more nodes but lower clustering coefficient. This indicates more connectivity than proteins in other cellular locations. However, due to the lower number of extracellular GO annotated genes they were not included for comparative gene network analysis to avoid misinterpretation of gene interaction data. As previously discussed (Chapter 5) a number of virulence genes are regulated by *vru* (*SUB0144*) and we have found eight genes (Fig 6.4) interacting with *vru* using systems biology approach. These have all been found to be regulated by *vru* as shown by Sharon et al. [54] confirming the validity of this approach to predict regulation where experimental data is not available.

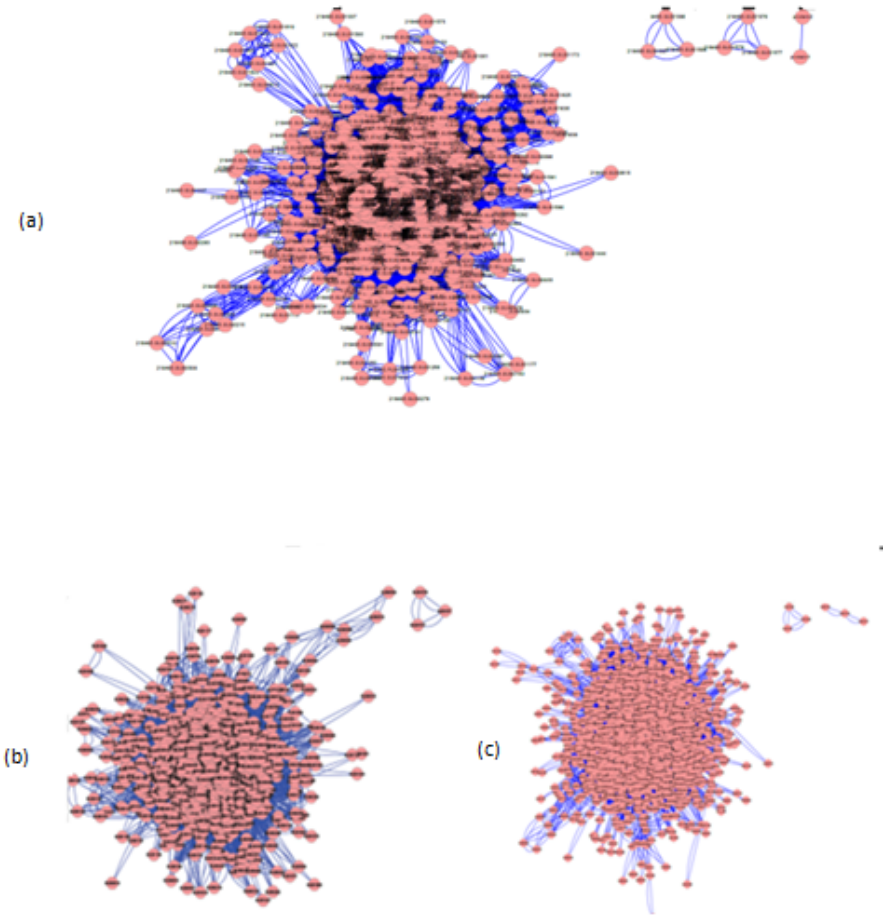



Figure 6.1: Topology of protein-protein interaction networks. Networks visualized using cytoscape (a) 0140J whole genome (b) 0140J core genome (c) 0140J accessory genome.

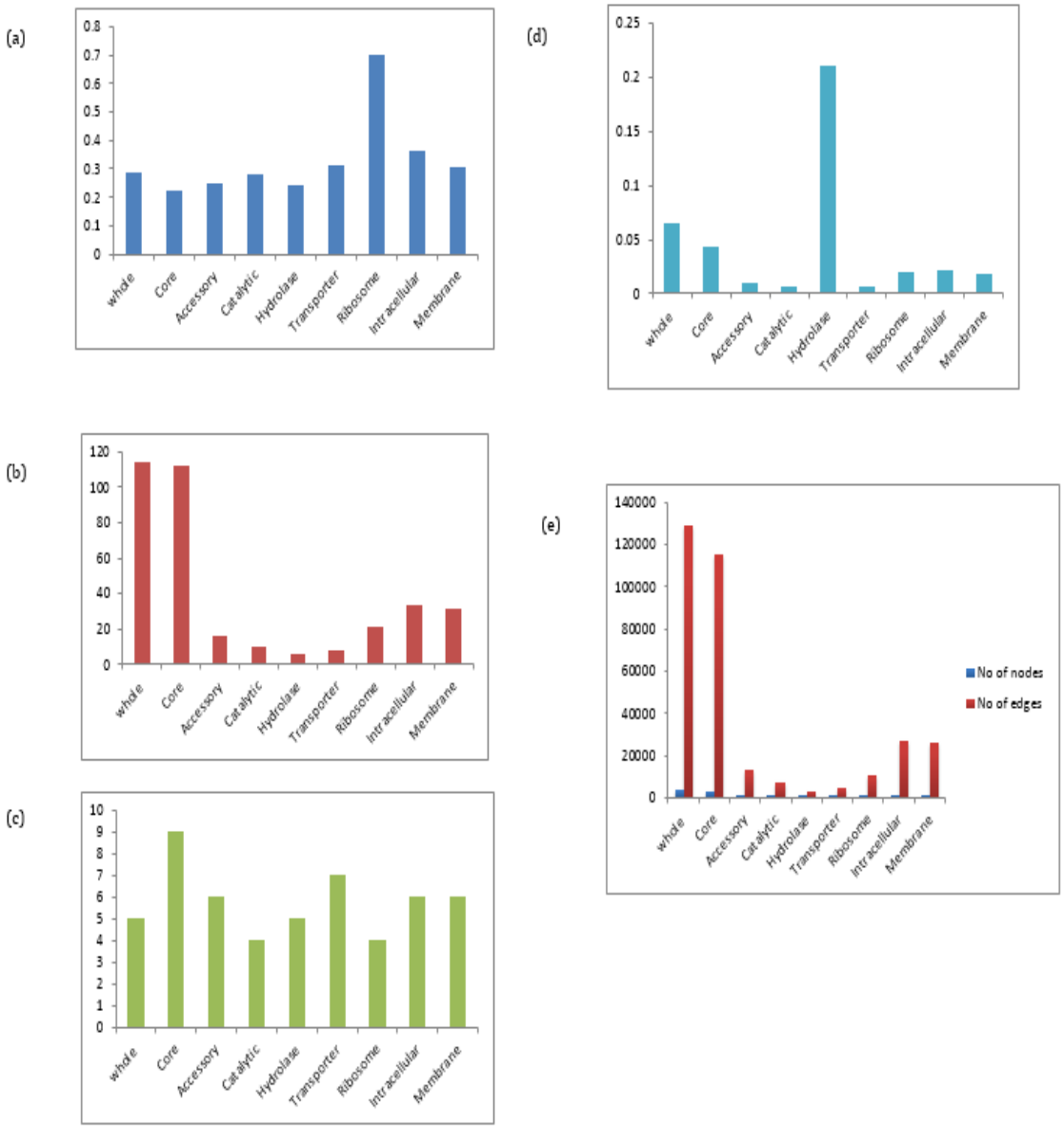



Figure 6.2: Protein-protein interaction network parameters for eight groups: whole - whole genome; Core - core genome; Accessory - Accessory genome; Catalytic - proteins classified as catalytic using GO; Hydrolase - proteins classified as hydrolase using GO; Transporter - proteins classified as a transporter using GO; Ribosome - proteins classified as located in the ribosome using GO; Intracellular, proteins classified as intracellular location using GO; Membrane - proteins classified as membrane located using GO. Network statistics generated using Cytoscape (a) clustering coefficient (b) average number of neighbors (c) network diameter (d) network density (e) number of nodes and edges.

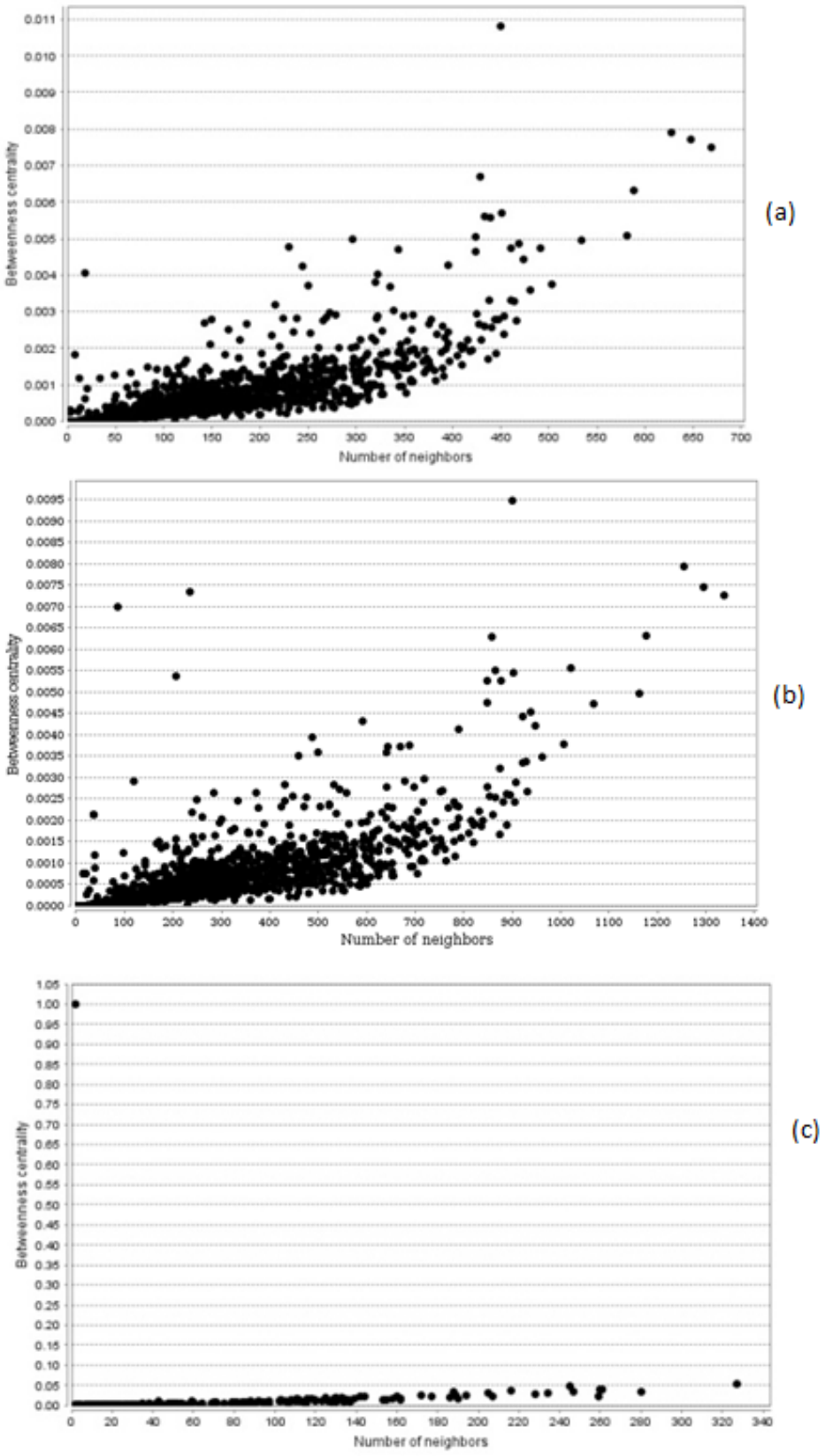



Figure 6.3: Network density of protein-protein interaction networks as measured by betweenness centrality compared to number of interacting proteins (number of neighbours). (a) 0140J whole genome (b) 0140J core genome (c) 0140J accessory genome.

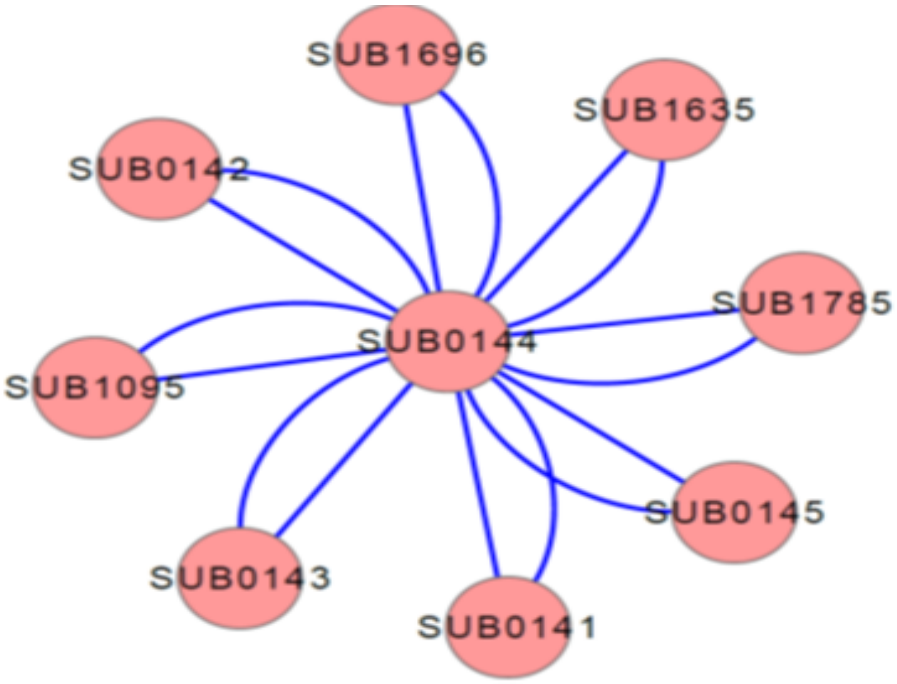



Figure 6.4: Protein-protein interaction network of SUB0144 virulence associated gene *vru*.

### 6.2.2 *Streptococcus uberis*-bovine protein-protein interaction

Tassi et al. [200] explored the pathogenic response to two *S. uberis* strains which differ in their ability to cause disease in cattle. One of the isolates was host-adapted and the other strain isolated from a transient infection in a single animal and was not as contagious as the host-adapted strain. Cows challenged with the putatively host-adapted strain showed clinical signs of mastitis and a specific cytokine response was seen with the increase of IL-6, CXCL8, IL-10, IL-12p40 and IL-17A. Minor elevation of CXCL8 level was also seen with no clinical or immunological response in cows challenged with the nonadapted strain. From our analysis we found that the non-virulent isolate EF20 also potentially interacts with these canonical signalling and immunological pathways. However this is also seen in the virulent isolate 0140J (Table 6.1). Interestingly, some of the canonical pathways were differentially enriched for interacting proteins when EF20 and 0140J predicted networks were compared (Table 6.1). These include several immunological pathways such as dendritic cell maturation and interleukin signalling for IL-6, CXCL8, IL-9, IL-10, IL-12, IL-15, IL-17 and IL-17A (Figure 6.5, 6.6 and 6.7). As Tassi et al. [200] showed, the differential expression of immunological factors at different time points suggests that pathogens may have evolved mechanisms to evade the host immune system by avoiding such interaction or by delaying interaction compared to the non-virulent pathogens. Moyes et al. [149] also studied the bovine gene expression pattern following challenge with *S. uberis* O140J. They identified the most affected gene networks and correlated these findings with other immunological and metabolic responses to mastitis. Their analysis showed 20 significantly enriched canonical pathways (with 20 to 61 genes

in each), The majority of these were signalling pathways. Among the immunological pathways activated wee IL-6 and IL-10 signalling pathways [149].

Human orthologs of predicted *S. uberis*-host interacting proteins are involved in 219 canonical pathways (Table 6.1).

Table 6.1: Canonical pathways interacting with EF20 and 0140J genes predicted by IPA. A multiple-testing corrected p-value are calculated using Benjamini-Hochberg (BH) [109] method and the values shown here are -log(p-value).

| Canonical Pathway | EF20(-log(BH p) | 0140J(-log(BH p) |
|---|---|---|
| Antigen Presentation Pathway | 9.65126129 | 9.486541034 |
| Caveolar-mediated Endocytosis Signaling | 7.054892448 | 7.692836111 |
| Virus Entry via Endocytic Pathways | 6.824457639 | 5.939119573 |
| ILK Signaling | 6.072170263 | 5.844863107 |
| PI3K Signaling in B Lymphocytes | 4.654012194 | 5.626963055 |
| Role of Tissue Factor in Cancer | 6.379529022 | 5.547084201 |
| iNOS Signaling | 5.333676447 | 5.21374608 |
| Dendritic Cell Maturation | 6.479309325 | 5.175285771 |
| Glucocorticoid Receptor Signaling | 5.127509931 | 4.885475663 |
| Interferon Signaling | 4.933451937 | 4.832831874 |
| phagosome maturation | 3.830086006 | 4.76447911 |
| Unfolded protein response | 3.96850951 | 4.599870106 |
| B Cell Receptor Signaling | 4.786498858 | 4.594755517 |
| Activation of IRF by Cytosolic Pattern Recognition Receptors | 4.705378506 | 4.581339718 |
| Cdc42 Signaling | 4.743536892 | 4.558359232 |
| TNFR2 Signaling | 4.587429696 | 4.497262026 |
| Telomerase Signaling | 5.189520576 | 4.414471927 |
| Colorectal Cancer Metastasis Signaling | 4.606726041 | 4.395286021 |
| CD28 Signaling in T Helper Cells | 4.54898394 | 4.391200723 |
| OX40 Signaling Pathway | 4.480982714 | 4.344120233 |
| IL-8 Signaling | 5.022254348 | 4.340096061 |
| Hepatic Fibrosis / Hepatic Stellate Cell Activation | 4.514295667 | 4.320865717 |
| T Cell Receptor Signaling | 4.402893022 | 4.259923796 |
| Role of NFAT in Regulation of the Immune Response | 4.94131762 | 4.258662114 |
| EGF Signaling | 5.005524501 | 4.153863771 |
| Prostate Cancer Signaling | 4.653581339 | 3.894282536 |
| Tec Kinase Signaling | 4.54552239 | 3.880592619 |
| CD40 Signaling | 4.634317207 | 3.826650128 |
| Germ Cell-Sertoli Cell Junction Signaling | 4.467452018 | 3.809522607 |
| iCOS-iCOSL Signaling in T Helper Cells | 3.883136937 | 3.745967737 |
| IL-4 Signaling | 3.829588197 | 3.712848183 |
| Signaling by Rho Family GTPases | 3.896388654 | 3.708288386 |
| Estrogen Receptor Signaling | 4.367644122 | 3.687376345 |
| Complement System | 3.683611443 | 3.597990523 |
| April Mediated Signaling | 2.920708035 | 3.597990523 |
| Chronic Myeloid Leukemia Signaling | 4.263298044 | 3.546519176 |

**Table 6.1 – continued from previous page**

| Canonical Pathway | EF20(-log(BH p)) | 0140J(-log(BH p)) |
|---|---|---|
| Inhibition of Angiogenesis by TSP1 | 3.600842419 | 3.515723208 |
| B Cell Activating Factor Signaling | 2.784422742 | 3.436237983 |
| Pancreatic Adenocarcinoma Signaling | 4.109738387 | 3.427315228 |
| PKC? Signaling in T Lymphocytes | 3.510870226 | 3.378413993 |
| Renal Cell Carcinoma Signaling | 3.476894647 | 3.370447315 |
| VEGF Signaling | 4.009381092 | 3.320858791 |
| Small Cell Lung Cancer Signaling | 4.056015086 | 3.318425052 |
| IGF-1 Signaling | 3.396514484 | 3.27778175 |
| Role of JAK1, JAK2 and TYK2 in Interferon Signaling | 3.31072994 | 3.243959271 |
| Tight Junction Signaling | 2.93781366 | 3.224780038 |
| CTLA4 Signaling in Cytotoxic T Lymphocytes | 3.915264829 | 3.218000603 |
| UVA-Induced MAPK Signaling | 3.864283344 | 3.17305528 |
| MIF Regulation of Innate Immunity | 3.225858226 | 3.143236565 |
| IL-17A Signaling in Gastric Cells | 3.207905231 | 3.141618954 |
| Crosstalk between Dendritic Cells and Natural Killer Cells | 3.239409892 | 3.128854719 |
| IL-2 Signaling | 3.898651203 | 3.119979139 |
| Protein Ubiquitination Pathway | 3.294050477 | 3.118248454 |
| Integrin Signaling | 3.645521579 | 3.071063249 |
| Lymphotoxin ? Receptor Signaling | 3.830505601 | 3.061441594 |
| Lipid Antigen Presentation by CD1 | 3.110308574 | 3.044505666 |
| NF-$\kappa$B Signaling | 3.18837191 | 3.044116506 |
| Death Receptor Signaling | 3.152148541 | 3.042617252 |
| Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes | 2.537856069 | 2.98080679 |
| Ephrin Receptor Signaling | 3.071606665 | 2.929441721 |
| ATM Signaling | 2.982224634 | 2.894293149 |
| RhoGDI Signaling | 2.612675234 | 2.873750691 |
| Systemic Lupus Erythematosus Signaling | 3.821616915 | 2.864410866 |
| JAK/Stat Signaling | 3.530462133 | 2.834063905 |
| IL-10 Signaling | 2.92822809 | 2.834063905 |
| TNFR1 Signaling | 2.904501854 | 2.82435874 |
| Cleavage and Polyadenylation of Pre-mRNA | 2.855662959 | 2.80896908 |
| LPS-stimulated MAPK Signaling | 3.476894647 | 2.787736923 |
| Molecular Mechanisms of Cancer | 2.933214445 | 2.741088433 |
| PPAR/RXR Activation | 2.488091898 | 2.739117066 |
| Role of PKR in Interferon Induction and Antiviral Response | 3.520856209 | 2.712616066 |
| autophagy | 2.784422742 | 2.712616066 |
| Autoimmune Thyroid Disease Signaling | 2.17285597 | 2.709149064 |

**Table 6.1 – continued from previous page**

| Canonical Pathway | EF20(-log(BH p)) | 0140J(-log(BH p)) |
|---|---|---|
| Phospholipase C Signaling | 2.499688273 | 2.688703801 |
| Production of Nitric Oxide and Reactive Oxygen Species in Macrophages | 3.233700197 | 2.687036891 |
| Prolactin Signaling | 3.322272664 | 2.654180199 |
| RANK Signaling in Osteoclasts | 2.742512238 | 2.644188583 |
| Cell Cycle: G1/S Checkpoint Regulation | 2.72641407 | 2.640970718 |
| Clathrin-mediated Endocytosis Signaling | 2.772381679 | 2.63593496 |
| IL-6 Signaling | 3.233188719 | 2.634737058 |
| 4-1BB Signaling in T Lymphocytes | 2.686931294 | 2.623529504 |
| Hypoxia Signaling in the Cardiovascular System | 2.678655641 | 2.593707732 |
| Aryl Hydrocarbon Receptor Signaling | 3.140660166 | 2.577983459 |
| Cyclins and Cell Cycle Regulation | 2.660569324 | 2.569408525 |
| Estrogen-Dependent Breast Cancer Signaling | 3.230234018 | 2.547490606 |
| IL-15 Signaling | 3.177235321 | 2.502285287 |
| Regulation of Actin-based Motility by Rho | 2.588903163 | 2.492586816 |
| Rac Signaling | 3.070063558 | 2.476254152 |
| Mouse Embryonic Stem Cell Pluripotency | 3.026420473 | 2.420511849 |
| RAR Activation | 2.549303985 | 2.417500503 |
| Role of RIG1-like Receptors in Antiviral Innate Immunity | 2.480801257 | 2.411414493 |
| PDGF Signaling | 2.9934695 | 2.371088596 |
| MIF-mediated Glucocorticoid Regulation | 2.408535197 | 2.347034743 |
| IL-17A Signaling in Fibroblasts | 2.408535197 | 2.347034743 |
| Type I Diabetes Mellitus Signaling | 2.445742325 | 2.345829542 |
| Protein Kinase A Signaling | 1.75833655 | 2.34326874 |
| Cardiac Hypertrophy Signaling | 2.143815082 | 2.33629481 |
| GDNF Family Ligand-Receptor Interactions | 2.976007617 | 2.33095552 |
| PEDF Signaling | 2.976007617 | 2.33095552 |
| p53 Signaling | 2.410566634 | 2.316745146 |
| Wnt/Ca+ pathway | 1.834483768 | 2.305241037 |
| MSP-RON Signaling Pathway | 3.027425131 | 2.303635638 |
| PI3K/AKT Signaling | 2.36844877 | 2.264490144 |
| Huntington's Disease Signaling | 2.737009335 | 2.254644 |
| Antiproliferative Role of TOB in T Cell Signaling | 2.302044406 | 2.249004075 |
| Allograft Rejection Signaling | 2.311832097 | 2.225130176 |
| Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis | 2.669051958 | 2.208169089 |
| Ephrin B Signaling | 2.253506987 | 2.173471412 |
| mTOR Signaling | 2.621626668 | 2.135414816 |

**Table 6.1 – continued from previous page**

| Canonical Pathway | EF20(-log(BH p)) | 0140J(-log(BH p)) |
|---|---|---|
| Renin-Angiotensin Signaling | 2.231483151 | 2.135060512 |
| Epithelial Adherens Junction Signaling | 2.233055336 | 2.126600034 |
| HIF1? Signaling | 2.682725259 | 2.12364432 |
| HMGB1 Signaling | 2.612597511 | 2.078602536 |
| Regulation of eIF4 and p70S6K Signaling | 1.812133015 | 2.077020885 |
| Thrombin Signaling | 2.190802303 | 2.0715499 |
| Primary Immunodeficiency Signaling | 2.126484883 | 2.060444795 |
| Communication between Innate and Adaptive Immune Cells | 2.141026817 | 2.056776617 |
| Glioblastoma Multiforme Signaling | 2.555453322 | 2.052667271 |
| UVB-Induced MAPK Signaling | 2.678275206 | 2.015806396 |
| IL-17A Signaling in Airway Cells | 2.631943816 | 2.011987693 |
| IL-1 Signaling | 1.644106653 | 1.993445152 |
| Role of p14/p19ARF in Tumor Suppression | 2.764083028 | 1.97805303 |
| Endometrial Cancer Signaling | 2.625414074 | 1.972374872 |
| RhoA Signaling | 2.063774588 | 1.97031844 |
| Sertoli Cell-Sertoli Cell Junction Signaling | 1.748162384 | 1.968478095 |
| eNOS Signaling | 2.06102644 | 1.958063705 |
| Acute Phase Response Signaling | 2.057630889 | 1.95021702 |
| p70S6K Signaling | 2.457017134 | 1.944173138 |
| DNA Methylation and Transcriptional Repression Signaling | 1.986870485 | 1.94376332 |
| Angiopoietin Signaling | 2.541525411 | 1.936596225 |
| Calcium-induced T Lymphocyte Apoptosis | 1.52709548 | 1.936596225 |
| Remodeling of Epithelial Adherens Junctions | 1.52709548 | 1.936596225 |
| Actin Cytoskeleton Signaling | 2.366643694 | 1.916347877 |
| Role of IL-17A in Arthritis | 2.523718307 | 1.888943998 |
| ERK/MAPK Signaling | 1.976247427 | 1.866410451 |
| Non-Small Cell Lung Cancer Signaling | 2.412946643 | 1.829636032 |
| fMLP Signaling in Neutrophils | 1.898622259 | 1.813155312 |
| Glioma Invasiveness Signaling | 2.427051879 | 1.809801453 |
| IL-17 Signaling | 2.371852663 | 1.795517459 |
| PXR/RXR Activation | 1.864532214 | 1.795517459 |
| Role of Oct4 in Mammalian Embryonic Stem Cell Pluripotency | 1.82448092 | 1.768112999 |
| Neuroprotective Role of THOP1 in Alzheimer's Disease | 1.294533231 | 1.768112999 |
| NRF2-mediated Oxidative Stress Response | 2.587272726 | 1.763203943 |
| IL-9 Signaling | 2.473998795 | 1.751875936 |
| TWEAK Signaling | 1.801269352 | 1.751875936 |
| Ovarian Cancer Signaling | 1.836437706 | 1.747385198 |

**Table 6.1 – continued from previous page**

| Canonical Pathway | EF20(-log(BH p)) | 0140J(-log(BH p)) |
|---|---|---|
| Altered T Cell and B Cell Signaling in Rheumatoid Arthritis | 1.813888906 | 1.740116084 |
| G$\alpha$12/13 Signaling | 2.231483151 | 1.737031066 |
| Induction of Apoptosis by HIV1 | 1.796912421 | 1.734634208 |
| FLT3 Signaling in Hematopoietic Progenitor Cells | 2.292156918 | 1.729446804 |
| Toll-like Receptor Signaling | 1.797517197 | 1.729446804 |
| Apoptosis Signaling | 1.784305064 | 1.711007786 |
| GNRH Signaling | 1.43889946 | 1.701420538 |
| Coagulation System | 1.749925214 | 1.700981131 |
| PTEN Signaling | 1.771169903 | 1.688119525 |
| Regulation of the Epithelial-Mesenchymal Transition Pathway | 2.124115279 | 1.685697852 |
| 3-phosphoinositide Degradation | 1.768623524 | 1.676526642 |
| HER-2 Signaling in Breast Cancer | 2.746339478 | 1.666126622 |
| Graft-versus-Host Disease Signaling | 1.698406928 | 1.64341277 |
| CCR3 Signaling in Eosinophils | 2.118149127 | 1.640616837 |
| PCP pathway | 1.689616902 | 1.628731055 |
| G$\alpha$q Signaling | 1.705284437 | 1.614643319 |
| EIF2 Signaling | 1.71043587 | 1.611414385 |
| HGF Signaling | 2.093540394 | 1.604983144 |
| Reelin Signaling in Neurons | 2.106377398 | 1.575969992 |
| PPAR Signaling | 1.644106653 | 1.57317302 |
| Atherosclerosis Signaling | 1.652550017 | 1.571896286 |
| Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis | 1.661102935 | 1.555792604 |
| Telomere Extension by Telomerase | 1.58374447 | 1.551070147 |
| CD27 Signaling in Lymphocytes | 1.583084673 | 1.529452483 |
| Semaphorin Signaling in Neurons | 1.080500696 | 1.493622069 |
| VEGF Family Ligand-Receptor Interactions | 2.003315209 | 1.491188749 |
| Eicosanoid Signaling | 1.555523012 | 1.491188749 |
| G-Protein Coupled Receptor Signaling | 1.347514953 | 1.490051066 |
| Relaxin Signaling | 1.252650056 | 1.489010451 |
| P2Y Purigenic Receptor Signaling Pathway | 1.563507196 | 1.48475587 |
| Breast Cancer Regulation by Stathmin1 | 1.549956502 | 1.455303804 |
| Mechanisms of Viral Exit from Host Cells | 1.479763253 | 1.433483092 |
| Role of CHK Proteins in Cell Cycle Checkpoint Control | 1.477220923 | 1.424938553 |
| Glioma Signaling | 1.896255015 | 1.423225889 |
| Erythropoietin Signaling | 1.934469355 | 1.409518032 |
| Agrin Interactions at Neuromuscular Junction | 1.467196213 | 1.409518032 |

**Table 6.1 – continued from previous page**

| Canonical Pathway | EF20(-log(BH p)) | 0140J(-log(BH p)) |
|---|---|---|
| Paxillin Signaling | 1.867876177 | 1.399730895 |
| UVC-Induced MAPK Signaling | 1.440164618 | 1.394322952 |
| IL-3 Signaling | 1.899124466 | 1.381026862 |
| Neurotrophin/TRK Signaling | 1.864532214 | 1.353177544 |
| T Helper Cell Differentiation | 1.409950243 | 1.353177544 |
| Wnt/?-catenin Signaling | 1.420211851 | 1.336749277 |
| Melanoma Signaling | 1.915265702 | 1.319952051 |
| Corticotropin Releasing Hormone Signaling | 1.06303311 | 1.314703577 |
| NF-$\kappa$B Activation by Viruses | 1.797517197 | 1.2993287 |
| FAK Signaling | 1.755252391 | 1.288354909 |
| Cholecystokinin/Gastrin-mediated Signaling | 1.353495017 | 1.288133203 |
| IL-12 Signaling and Production in Macrophages | 1.699232696 | 1.287298475 |
| GADD45 Signaling | 1.311745947 | 1.28069706 |
| DNA damage-induced 14-3-3? Signaling | 1.311745947 | 1.28069706 |
| Chemokine Signaling | 1.328717258 | 1.273293276 |
| Insulin Receptor Signaling | 1.34261968 | 1.269027771 |
| Amyotrophic Lateral Sclerosis Signaling | 1.33183691 | 1.26693264 |
| TR/RXR Activation | 1.698690364 | 1.242379338 |
| CXCR4 Signaling | 1.604617787 | 1.217906822 |
| phagosome formation | 2.007512444 | 1.18561216 |
| GM-CSF Signaling | 1.689616902 | 1.184747634 |
| SAPK/JNK Signaling | 1.617528355 | 1.176607335 |
| NGF Signaling | 1.560472236 | 1.146945369 |
| Hereditary Breast Cancer Signaling | 1.520816905 | 1.12770857 |
| Ephrin A Signaling | 1.658840448 | 1.124213855 |
| Fc Epsilon RI Signaling | 1.514615977 | 1.10953911 |
| 3-phosphoinositide Biosynthesis | 1.755489796 | 1.078443647 |
| Endothelin-1 Signaling | 1.42631788 | 1.076280696 |
| Ceramide Signaling | 1.500857408 | 1.062811625 |
| Macropinocytosis Signaling | 1.52709548 | 1.058071403 |
| ErbB Signaling | 1.874703006 | 1.02126155 |
| Type II Diabetes Mellitus Signaling | 1.380346315 | 1.012508745 |
| uperpathway of Inositol Phosphate Compounds | 1.523696628 | 0.939994916 |
| ErbB2-ErbB3 Signaling | 1.41136047 | 0.937473673 |
| Fc$\gamma$RIIB Signaling in B Lymphocytes | 1.329149676 | 0.835966401 |

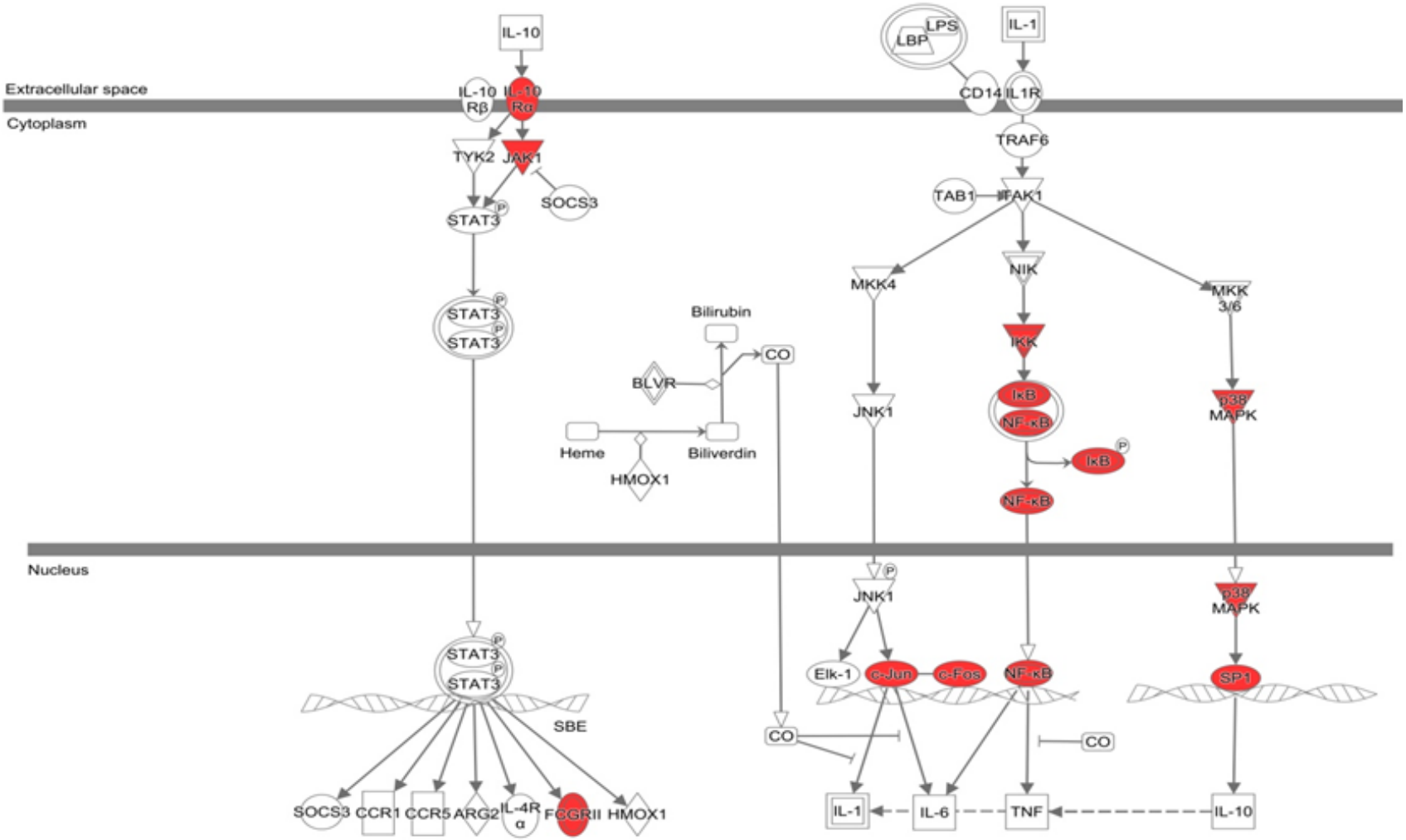
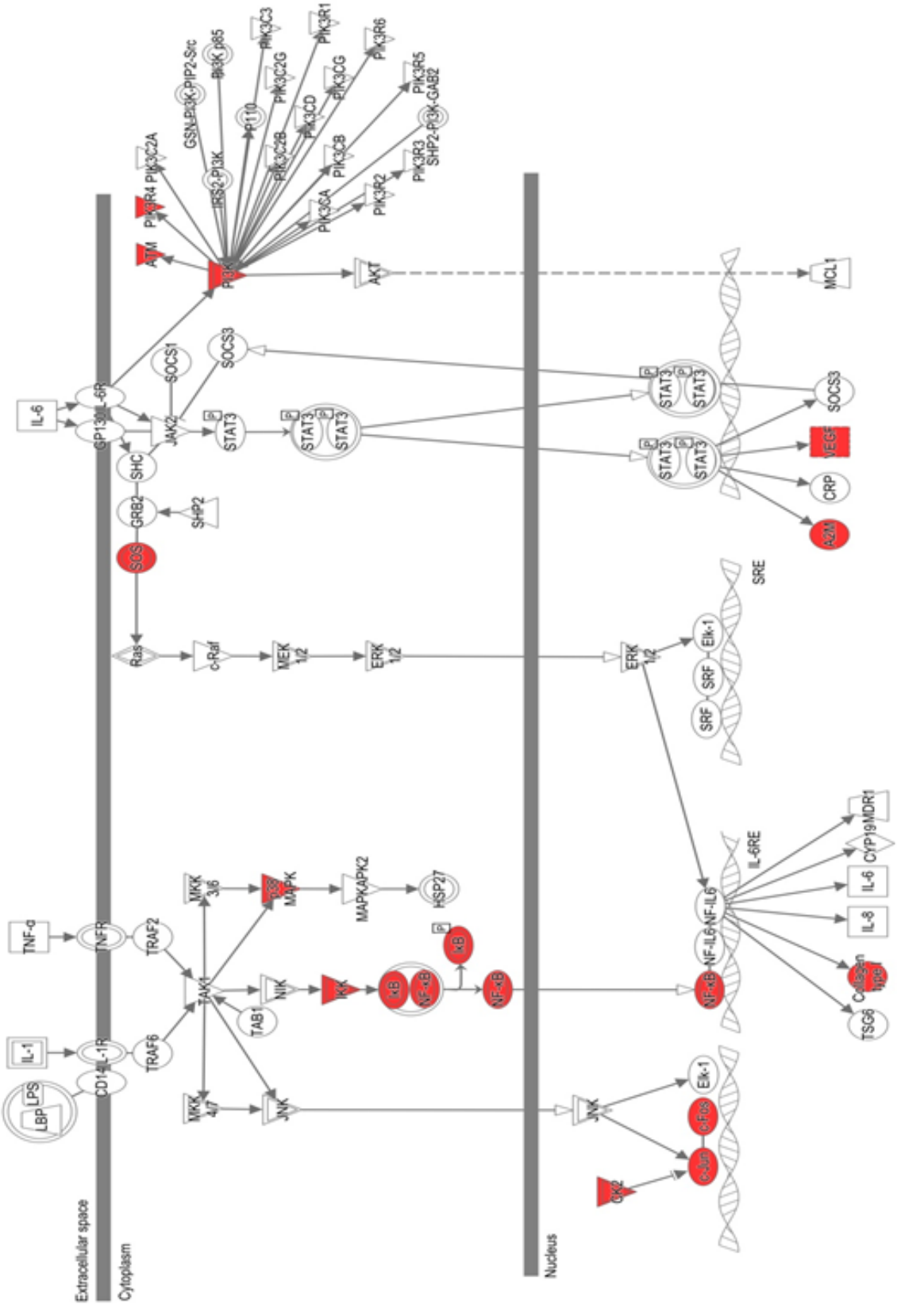



Figure 6.5: IL-10 signaling pathway. Red colour highlights host proteins interacting with pathogen proteins in *S uberis*.

(a) : IL-6 signaling pathway: 0140J



(b) : IL-6 signaling pathway: EF20

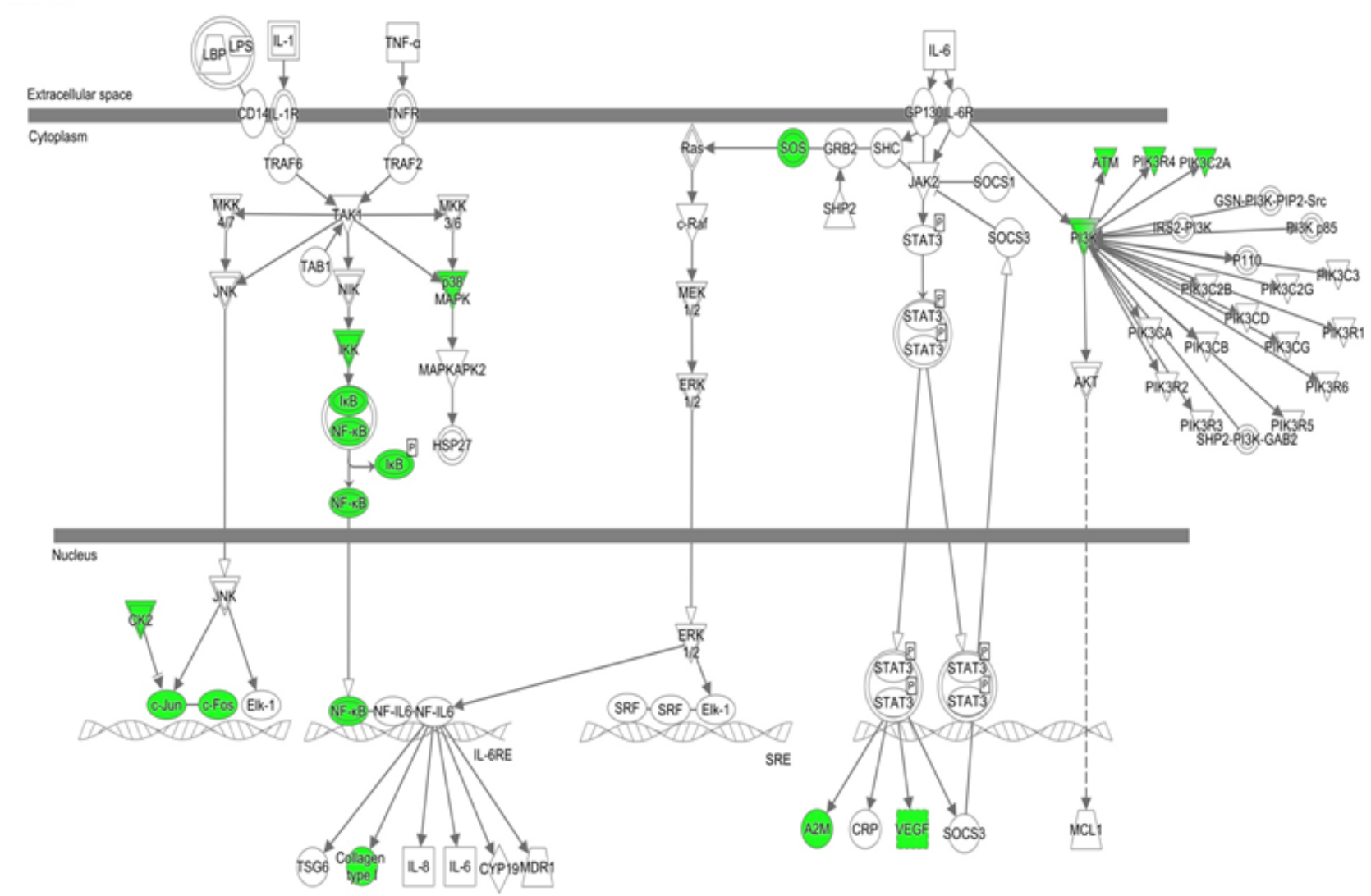



Figure 6.6: IL-6 signaling pathway. Green colour highlights host proteins interacting with pathogen proteins in *S uberis*. (a) *S uberis* 0140J (b) *S uberis* EF20, one of the host proteins PIK3C2 interacts with EF20 specific accessory gene ped.580 not found in 0140J.

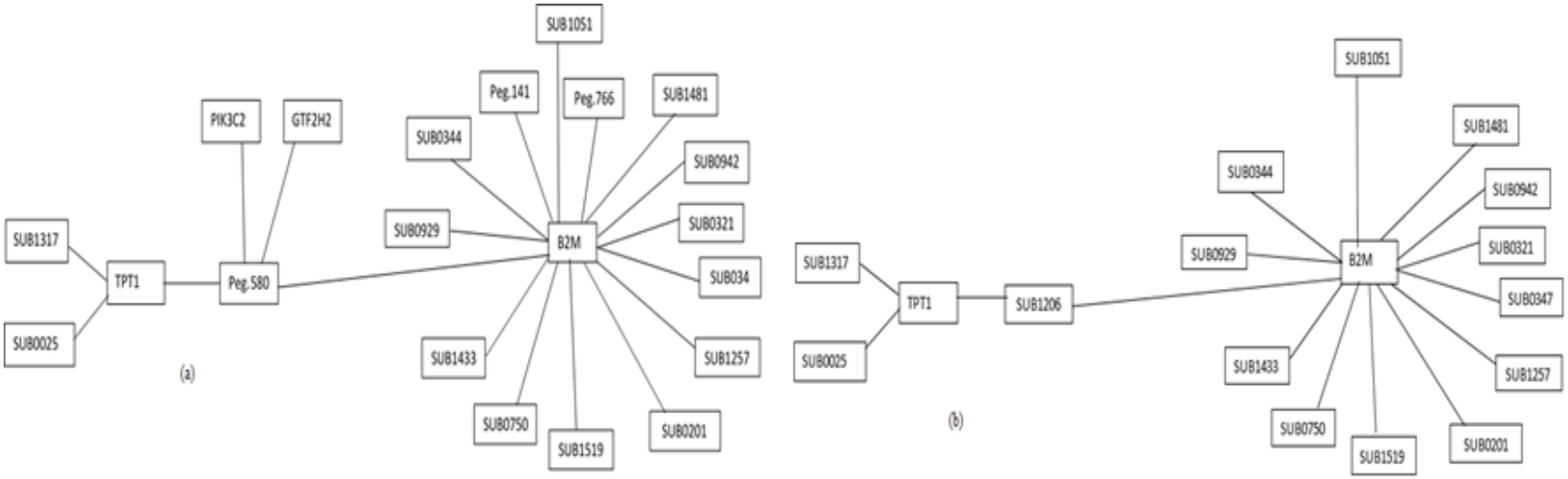



Figure 6.7: Comparative analysis of 0140J and EF20 proteins interacting with host IL-6 signalling pathway. (a) Neighbourhood of host PIK3C2A interactions in EF20, the novel gene peg580 (green circle) is interacting with both PIK3C2A and also GTF2H2,(b) 0140J-host network interaction is absent.

Twenty five pathways were found to have a greater enrichment of interacting proteins from 0140J compared to EF20 (list of proteins interacting with these pathways are provided in Additional File 6.1: PPI_0140J_EF20. These pathways may play significant roles in the pathogenesis or development of mastitis in the host. These 25 signalling pathways should be considered for further comparative analysis to understand the mechanism of varying pathogenesis between *S. uberis* strains in more detail. IL-1 signalling, PI3K/Akt signalling, April Mediated signaling, PI3K signaling, G-Protein Coupled Receptor signaling are some of the major pathways enriched in interactors of 0140J proteins. These have been implicated in many disease processes. IL-1 signaling facilitates the production of leukocytes that attract chemokines KC and MIP-1$\alpha$ which recruit neutrophils to Group B Streptococci (GBS) infection sites, thus assisting these leukocytes to clear the infection [17]. Therefore IL-1 signaling pathway stimulation may be useful for the treatment of GBS infections.

The PI3K/Akt signalling pathway has been found to be involved in several cellular functions and activation of this signalling pathway has been implicated in phagocytosis, pseudopod formation and membrane ruffling one of the key events of GBS invasion of host cells [42]. This pathway also triggers other downstream signalling molecules such as protein kinases for example in *S. suis*, it has been found that PKC inhibitors reduce phagocytosis into J774 macrophage cells [182]. As seen in the host-pathogen interaction data, such pathways may also play important roles in the invasion of *S. uberis* isolates. A proliferation-inducing ligand (APRIL), is involved in immune system regulation. APRIL promotes B-cell proliferation by binding to the B-cell maturation antigen (BCMA) and

calcium modulator cyclophilin ligand interactor (TACI) receptors, which are expressed on resting and activated B-cells. In addition, TACI has been found on a subset of activated T cells. This signalling process releases NF-$\kappa$B from the NF-$\kappa$B/IkB complex and enables NF-$\kappa$B to translocate to the nucleus regulating gene transcription. APRIL has been implicated in promoting growth in several malignant tumor cell lines [227]. Bacterial pathogens take the advantage of effector proteins to propagate infection by diverting host cell signaling and frequently target kinase signaling cascades such as the nuclear factor $\kappa$B (NF-$\kappa$B), mitogen-activated protein kinase (MAPK), phosphatidylinositol 3-kinase (PI3K)/Akt, and p21-activated kinase (PAK) pathways [106].

Some of the 0140J genes have been found to interact with Wnt/Ca+ pathway and in Drosophila the dally-mediated Wnt signalling pathway was found to play important roles in phagocytosis [234]. The Wnt family of signalling proteins are involved in multiple developmental events during embryogenesis, and also effects other signalling pathways such as mitogenic stimulation, cell fate specification, and differentiation [234] making this an interesting signalling pathway requiring further investigation in *S. uberis* infection.

Virulence patterns differ between host-adapted and non-adapted strains [200] and in our study, we also found that comparison of virulent (0140J) and non-virulent (EF20) isolates differ in their potential interaction with different molecular and immunological signalling pathways implying that *S. uberis* pathogenesis is mostly driven by the pathogen type rather than by host factors. Differences in their ability to affect different canonical pathways also supports the idea of variability in their abilities to resist killing

by PMNL or mammary macrophages, however, this finding was contradicted by Leigh et al. [119, 173].

As shown in Table 6.1, CXCL8 signalling is one of the affected immunological pathways and it has been found that initial increase of SCC correlates with the first detectable increase in CXCL8 [10]. The immune response varies in response to different pathogens and even between bacterial strains of same species. The IL-6 signalling pathway affected by IMI is important since this is associated with both proinflammatory and anti-inflammatory activity and plays important role in transitioning from innate to adaptive immunity [62, 200]. One of the accessory proteins (peg.580) of EF20 interacts with the host protein PIK3C2A (phosphatidylinositol-4-phosphate 3-kinase) and this protein is missing in 0140J implying the absence of this interaction. The role of PIK3C2A in host immune response has not been extensively studied and thus it will be interesting to investigate how the interaction makes the differences in host response in the absence of this protein. The increase of TNF-$\alpha$, IL-10, and IL-12p40 was generally associated with the increase in CXCL8 (Figure 6.6) [200] (http://www.biomedcentral.com/1471-2164/10/542/figure/F4). The observed increase in IL-12p40 levels suggested the presence of IL-12 or IL-23, since they share a p40 subunit [153].

These analyses complement data revealed from homology-based PPI (Table 6.1). We also found that these pathways are enriched for host-pathogen interacting proteins. After challenge with *S.uberis* strain O140J, IL-10 and IL-12 levels increased at the same time as CXCL8 levels and in our analysis we again found the enrichment of proteins interacting with the IL-10 signalling pathway complementing experimental studies by

Tassi et al. [200].

Following infection with *S. uberis* the first detectable rise in IL-17A occurs following the influx of CD4+ and CD8+ T. CD4+TH17 cells. These cells produce IL-17A and the differentiation of naive T-cells in CD4+Th17 cells is supported by IL-6 and TGF-$\beta$1. IL-1$\beta$ and TNF-$\alpha$ usually enhance this signalling [212]. From the homology based system biology approaches we can see that though a number of immunological signalling pathway are affected by both virulent and non-virulent isolates there are differences in the enrichment of interacting proteins. Offering a potential avenue for validation by wet lab experiments.

## 6.3 Conclusion

From our analysis and experimental analyses of others it is clear that different *S. uberis* strains target specific signalling networks possibly affecting the downstream event they require to evade the host immune system. Thus computational prediction of protein-protein interaction can play an important role in verifying experimental interaction data and offers potential to infer PPI along with the signalling pathways they may affect. Not all the pathogen interaction data have been extensively studied and thus a lack of experimental data in such cases may hinder understanding the host-pathogen interaction which is required to gain insight into the pathogenesis mechanism of bacterial pathogens. Thus knowledge transfer from computational prediction can be useful in identifying unknown pathways and gene regulatory networks that may relate to the pathogenesis

and not identified by experimental studies due to higher cost and limitations of high throughput study.

CHAPTER 7

# CONCLUSIONS

## 7.1 Summary

The major findings from this research can be summarized as below:

- Development of pipelines which enabled the assembly of draft *S. uberis* genomes solely using data generated by SOLiD single end reads. As the NGS platforms were expensive and the scarcity of assembly tools at that time (~ 2011) the pipelines developed in the present study could have huge impact on the *de novo* assembly of colorspace single end reads.

- At the onset of the project a single reference genome (0140J) was available [216], now an additional 12 quality assemblies from Illumina data are within the public domain.

- Genome variation analysis using unbiased computational approaches could separate virulent and non-virulent lineages and identified several markers that could play significant roles in the variations of these two clusters.

- The deletion of *SUB1027* a homolog of *hasB* (*SUB1696*) in non-virulent isolate EF20 along with other subclinical isolates such as B362 and 6780 may be related to variance in capsule production (personal communication Dr Sharon Egan University of Nottingham). GWAS also identified a conserved deletion in the upstream region of one of the most important regulatory genes *vru* demonstrated the importance of the investigation of noncoding sequences.

- Genome-wide analysis of *S. uberis* protein-coding genes uncovered twenty six genes that could be connected to virulence and which are subject to positive selection pressure which may have been previously overlooked.

- Apart from known signalling pathways we identified a number of host canonical pathways enriched with interacting *S. uberis* 0140J proteins and several of our predictions were complemented by experimental studies [200] thus validating the computational prediction of host-pathogen interaction.

It is important to note the there are some limitations of the studies including:

- One of the limitations of the present study was not considering the gap closure. Moreover, the metrics for assessing genome assembly emphasized contig size, size of scaffolds and especially N50 sizes. This approach is greedy in joining contigs and generating larger contigs may result in mis-assemblies [176].

- Twelve isolates were used for Illumina sequencing and comparative genomics study. This study could have benefited from the inclusion of larger number of samples. Moreover, lllumina reads had substantially lower coverage and with higher coverage the assembly could have been improved and could help in closing a number of gaps and identify a number of undetected genes/regions.

- Annotation information from SOLiD data was not included in pan-genome analysis and this could have provided greater resolution in genome distribution among the isolates. However, this could have also provided misleading information due to higher error rate.

- The present study does not show any functional validation using genotype-phenotype relationships.

- For genewide-variation the variants from SOLiD data was not included. Although they were highly error prone the variants could have been validated using stringent quality control such as coverage information, mapping quality. Variants could have been validated if they are found consistent both in Illumina and SOLiD platforms.

- Deletion in the upstream regions of *vru* indicates the significance of noncoding regions in the genome variation and the study did not provide any information on the indel on the noncoding regions apart from those upstream of putative virulence genes.

- Chapter 6 is extrapolatory in nature due to time limitation. As such, intraspecies

gene network anlaysis was carried out using only 0140J and host-pathogen protein-protein interaction analysis was studied only using 0140J and EF20.

- Although genome-wide analysis based on genomic variation and selection pressure clusters isolates into two major groups, the difference between virulent and non-virulent strains is still not clear since the non-virulent isolate EF20 was also isolated from a clinical case.

## 7.2 Outlook

### 7.2.1 Gap closure

A more general limitation of the study was not considering the gap closure and this could have been mitigated by using hybrid assembly using combination of different tools such as CISA [126] or Post Assembly Genome Improvement Toolkit (PAGIT) (https://www.sanger.ac.uk/resources/software/pagit/).

The assembly from SOLiD and Illumina could also be merged in an attempt to close gaps in assembled genomes. As sequencing technology is rapidly advancing and third generation sequencing such as Oxford Nanopore Technology and PacBio are now available and as these produce significantly longer reads, gap closure especially in repeat and GC rich regions will be much more efficient.

### 7.2.2 Identification of virulence genes

Despite no significant differences found between the virulent and non-virulent isolates with respect to the gene gain and loss of putative virulence genes, this does not rule out differences at the transcriptional level contributing to the differences in their pathology. Analysis of *S. uberis* transcriptome sequencing during an infection as well as analysis of host immune response could be studied in detail and comparative analysis could provide information on the immune responses during clinical and subclinical infections.

Due to the advent of NGS, thousands of potential homologs of sortase substrates are being discovered with poorly understood mechanisms. Thus in future, it will be important to move beyond conventional classification based on sequence similarity, to functional analysis using each structure-function relationship to identify enzyme determinants that control substrate specificity. This will reveal the mechanism of sortase singnalling and attachment to cell surface [193].

It was already known that production of the capsule varies among *S. uberis* strains, the deletion of *SUB1027* a homolog of *hasB* (*SUB1696*) in non-virulent isolate EF20 along with other subclinical isolates such as B362 and 6780 may be related to variance in capsule production (personal communication Dr Sharon Egan University of Nottingham). The role of gene knockouts such as deletion of *SUB1027* a homolog of *hasB* (*SUB1696*) in the process of capsule production needs to be analyzed in detail.

The role of CRIPSR-cas spacer regions have also been implicated in the pathogenesis of some other Streptococci. In *S. pyogens* one of the spacer regions is similar to *trcF* which is a transcriptional repair coupling factor indicating their possible role in the DNA repair

system. Such a mechanism may also be available in *S. uberis*. Further experimental studies will be necessary to investigate the roles of spacer regions.

With the availability of more *S. uberis* genomes the comprehensive picture of the role of prophages in *S. uberis* evolution, adaptation and physiology can be generated. Prophage regions have been found to modulate gene transduction in many isolates such as between *E. faecalis* strains and between enterococcal species and they have also been implicated in killing sensitive competitors and helping bacteria colonize a niche [137]. Such studies could be conducted in *S. uberis* to determine whether prophage integration provides any advantage in adapting to certain environmental conditions. As more sequencing is under way in a related project (Peers Davies and Professor Martin Green Personal communication) it may be possible to investigate if the prophage contents follow any specific dynamics in a specific environment. Their role in the gene regulation can also be investigated to see if they have any role in regulating virulence gene expression. It could be possible that mixed *S. uberis* subpopulations may be formed upon prophage induction and favor survival of one or several of them in an environment [76, 172, 214] .

In this study we analysed the ~ 100 bp upstream of putative virulence genes, however, this could have been complemented by analyzing upstream regions of all coding regions to see the genomic variation in potential promoter sites. Therefore, analysis of intergenic regions using phylogenetic footprinting [216] or mutation rate heterogeneity [213] could provide substantial information on *S. uberis* evolution.

Identification of genes undergoing positive selctions necessitates further study in a larger number of samples to validate putative involvement of specific genes or molecular sys-

tems of *S. uberis* in pathogenesis. Moreover, identification of these genes also implies the role of different virulence mechanisms than previously considered. We argue that the evolution of pathogen specific genes does not provide a complete picture of the evolutionary dynamics of virulence. Another important aspect is identifying novel insertion/deletion (indel) events. This is a complex process and indels sometimes remain undetected using a reference based mapping approach. For this reason we used the alignment of coding genes extracted from *de novo* assembled genomes to infer variation among the protein coding regions. Such analysis could be complemented by using conventional Sanger sequencing to validate our predictions. Third generation sequencing platforms such as Oxford Nanopore [7, 37, 100, 128] and Single Molecule Real Time (SMRT) PacBio sequencing [8, 59] which provide longer reads at high throughput may help to overcome such problems.

As mentioned classical genotyping using MLST and PFGE lacks robust phylogenetic markers [39] thus the genomic variation analysis in this work is the first attempt to distinguish between virulent and non-virulent lineages directly. Further analysis of SNPs in specific positions of the identified genetic markers contributing to the diversity of two major identified lineages of virulent and non-virulent isolates (*SUB1407, SUB0933, SUB1638, SUB0890, SUB0182, SUB0690, SUB1514, SUB0083, SUB1632, SUB0082, SUB0083, SUB0096, SUB0713 and SUB1633*) has the potential to be applied in epidemiological and surveillance settings. These markers can also be used to provide genotype-phenotype association and provide information on the strain-typing and host pathogen interactions during *S. uberis* infection.

### 7.2.3 Analysis of host-pathogen interaction.

Analysis of *S. uberis* intra- species PPI needs further comprehensive analysis with respect to validating predictions using other available sources and one approach could be using homology based bacterial-bacterial interaction data as was conducted for *S. uberis*-bovine PPI network analysis.

Extensive analysis of the predicted PPI network could help to find important pathways and regulatory mechanisms relating to the pathogenesis of *S. uberis*. The novel pathways could be studied in detail and such studies may provide information on prognostic/diagnostic markers of mastitis.

Understanding protein domain-domain interaction could be the next step to understand pathogen biology as they provide information of protein interaction networks as well as prediction of protein function. However, protein-protein interaction is a complex process and prediction/analysis depends on lots of factors such as time and space constraints. The interaction predicted may not be universal or depend on their expression at certain conditions and differs in different cellular developmental stages. As the experimental data only provides a small fraction of actual protein-protein interaction data, computational prediction provides comprehensive information and by integrating other datasets such as gene expression, homology, signalling pathways such predictions can be improved dramatically. The present PPI network analysis could be adopted in a larger scale and this data can be used as reference for future analysis as the comprehensive analysis

of datasets may provide important information both host and pathogen factors and understanding their relationship during disease development stages crucial for providing treatment and prevention of mastitis.

Whilst much is left to understand, the presented body of work represents a comprehensive comparative study of bacterial strains of differing variance. The methods developed and the obstacles overcome have provided pipelines and priortities for further analysis.




# APPENDIX A

## A.1 Colorspace reads

Example of a colorspace fasta read: .csfasta file:

>853_7_463_F3

T3231110.122321002.0011.0012.2213..2

The corresponding color space quality format: .qual file:

>853_7_463_F3

33 32 29 33 33 31 33 -1 33 33 33 31 26 31 32 33 30 -1 27 31 33 33 -1 29 31 31 30 -1 28 29

18 16 -1 -1 28

(https://www.biostars.org/p/43855/).

Table A.1: SOLiD error filtering (data shown for 10 isolates).

| | 3448 | 4428 | Ab1947 | C5072 | 5291 | 6736 | C6344 | C5388 | B190 | 6780 | C7131 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Number of Reads | 19,476,871 | 23,875,614 | 28,615,398 | 21,451,375 | 24,023,754 | 21,161,925 | 22,983,635 | 27,427,561 | 22,300,392 | 28,978,858 | 21,372,178 |
| Number of Reads Sampled | 3,895,374 | 4,775,122 | 5,723,079 | 4,290,275 | 4,804,750 | 4,232,385 | 4,596,727 | 5,485,512 | 4,460,078 | 5,795,771 | 4,274,435 |
| Percentage Sampled | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% | 20.00% |
| Number of Passing Reads Avg QV >20 | 2,172,237 | 2,625,959 | 3,434,164 | 2,344,101 | 2,625,694 | 2,526,400 | 2,583,185 | 2,965,307 | 2,493,636 | 3,532,233 | 2,529,009 |
| Passing Percentage | 55.80% | 55.00% | 60.00% | 54.60% | 54.60% | 59.70% | 56.20% | 54.10% | 55.90% | 60.90% | 59.20% |

Table A.2: Sortase prediction by CW-PRED (Annotation from RAST server from June 2014 was used for prediction).

| ClutserID | 0140J | EF20 | 6736 | 6780 | Ab71 | B190 | B362 | C5072 | C5388 | C6344 | C8329 | C9359 | S6261 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SUB1154 | peg.1618 | peg.817 | .peg.1711 | peg.1100 | peg.1586 | peg.1141 | peg.1633 | peg.1102 | peg.1626 | peg.1105 | peg.1093 | peg.1647 |
| 2 | SUB1370 | peg.1825 | peg.1029 | .peg.126 | peg.1328 | peg.1815 | peg.1367 | peg.1844 | peg.1317 | peg.1831 | peg.1377 | peg.1308 | peg.75 |
| 3 | SUB0888 | peg.1346 | peg.553 | .peg.1432 | peg.834 | peg.1319 | peg.867 | peg.1359 | peg.837 | peg.1360 | peg.827 | peg.828 | peg.1375 |
| 4 | SUB0207 | peg.634 | peg.1735 | .peg.785 | peg.170 | peg.654 | peg.195 | peg.674 | peg.182 | peg.680 | peg.112 | peg.175 | peg.713 |
| 5 | SUB1730 | peg.287 | peg.1430 | .peg.496 | peg.1660 | peg.287 | peg.1719 | peg.340 | peg.1650 | peg.313 | peg.1764 | peg.1641 | peg.412 |
| 6 | SUB0145 | peg.561 | peg.1679 | .peg.709 | peg.113 | peg.595 | peg.138 | peg.614 | peg.124 | peg.620 | peg.53 | peg.117 | peg.654 |
| 7 | SUB0135 | peg.545 | peg.1669 | .peg.699 | peg.103 | peg.585 | peg.116 | peg.604 | peg.114 | peg.610 | peg.43 | peg.107 | peg.644 |
| 8 | SUB0241 | peg.720 | peg.1753 | .peg.821 | peg.193 | peg.671 | peg.229 | peg.692 | peg.200 | peg.707 | peg.129 | peg.193 | peg.731 |
| 9 | SUB0826 | peg.1288 | peg.494 | .peg.1371 | peg.777 | peg.1261 | peg.806 | peg.1302 | peg.780 | peg.1303 | peg.772 | peg.771 | peg.1317 |
| 10 | - | peg.1325 | peg.532 | .peg.1407 | peg.813 | peg.1298 | peg.842 | peg.1338 | peg.816 | peg.1339 | peg.806 | peg.807 | peg.1354 |
| 11 | - | peg.1237 | peg.442 | .peg.1323 | peg.725 | peg.1209 | peg.754 | peg.1254 | peg.728 | peg.1255 | peg.724 | peg.719 | peg.1268 |
| 12 | - | - | peg.1891 | .peg.917 | peg.299 | peg.778 | peg.340 | peg.829 | peg.291 | peg.831 | peg.459 | peg.284 | peg.839 |
| 13 | - | peg.296 | peg.1439 | .peg.505 | peg.1669 | - | peg.1728 | - | peg.1659 | - | peg.1773 | peg.1650 | peg.421 |
| 14 | - | - | - | - | - | peg.612 | - | peg.632 | - | peg.638 | peg.70 | - | - |
| 15 | - | - | peg.1848 | - | - | - | - | peg.787 | - | peg.790 | - | - | - |
| 16 | - | - | peg.1847 | - | - | - | - | peg.786 | - | peg.789 | - | - | - |
| 17 | - | - | peg.1849 | - | - | - | - | peg.788 | - | peg.791 | - | - | - |
| 18 | - | peg.645 | - | - | - | - | peg.214 | - | - | - | - | - | - |
| 19 | - | - | - | .peg.472 | - | - | peg.1691 | - | - | - | - | - | - |
| 20 | - | - | - | - | - | - | - | - | peg.1043 | - | - | peg.1034 | - |
| 21 | - | peg.687 | - | - | - | - | - | - | - | - | - | - | - |
| 22 | - | peg.669 | - | - | - | - | - | - | - | - | - | - | - |
| 23 | - | - | - | .peg.801 | - | - | - | - | - | - | - | - | - |
| 24 | - | - | - | .peg.1651 | - | - | - | - | - | - | - | - | - |
| 25 | - | - | - | - | - | - | peg.1695 | - | - | - | - | - | - |
| 26 | - | - | - | - | - | - | peg.1080 | - | - | - | - | - | - |
| 27 | - | - | - | - | - | - | peg.868 | - | - | - | - | - | - |
| 28 | SUB1095 | - | - | - | - | - | - | - | - | - | - | - | - |
| 29 | - | peg.1559 | - | - | - | - | - | - | - | - | - | - | - |
| 30 | - | - | peg.758 | - | - | - | - | - | - | - | - | - | - |
| 31 | - | - | - | - | peg.1042 | - | - | - | - | - | - | - | - |
| 32 | - | - | - | - | - | peg.1527 | - | - | - | - | - | - | - |
| 33 | - | - | - | - | - | - | - | peg.1573 | - | - | - | - | - |
| 34 | - | - | - | - | - | - | - | - | - | peg.1568 | - | - | - |
| 35 | - | - | - | - | - | - | - | - | - | - | peg.1047 | - | - |
| 36 | - | - | - | - | - | - | - | - | - | - | - | - | peg.1588 |

## A.2 Command lines

**SOLiD error correction :**

./saet_mp <.csfasta> <qual> expectedlength(int) [-options]

Preprocessing for velvet assembly:

perl preprocessor.pl –f .csfasta –t fragment –o output_directory

-f csfasta file available after SAET correction

-t type of reads; SOLiD fragmenet reads in this study

-o output directory where the preprocessed file will be saved.

**Assembly using velveth and velvetg:**

>velveth outputdirectory hashlength [-fileformat][-readtype] filename]

./velvetg_de output _directory/ -exp_cov auto

**Mapping using bwa:**

making index of reference:

bwa index ref

paired-end mapping:

bwa mem ref.fa read1.fq read2.fq > aln-pe.sam

convert sam to bam format and index the bamfile

samtools view -bS -q 20 -F4 aln-pe.sam -o aln-pe.bam

samtools sort aln-pe.bam aln-pe.sorted samtools index aln-pe.sorted.sorted.bam

**Variant calling:**

samtools pileup -f myRef.fasta myBam.bam | java -jar VarScan.jar pileup2s

# BIBLIOGRAPHY


[1] R. ALBERT, *Scale-free networks in cell biology*, Journal of cell science, 118 (2005), pp. 4947–4957.

[2] A. ALEXEYENKO, T. SCHMITT, A. TJÄRNBERG, D. GUALA, O. FRINGS, AND E. L. SONNHAMMER, *Comparative interactomics with funcoup 2.0*, Nucleic acids research, 40 (2012), pp. D821–D828.

[3] R. A. ALMEIDA, D. A. LUTHER, H.-M. PARK, AND S. P. OLIVER, *Identification, isolation, and partial characterization of a novel Streptococcus uberis adhesion molecule (SUAM)*, Veterinary microbiology, 115 (2006), pp. 183–191.

[4] M. ANISIMOVA, J. BIELAWSKI, K. DUNN, AND Z. YANG, *Phylogenomic analysis of natural selection pressure in Streptococcus genomes*, BMC evolutionary biology, 7 (2007), p. 154.

[5] A. ARANAZ, B. ROMERO, N. MONTERO, J. ALVAREZ, J. BEZOS, L. DE JUAN, A. MATEOS, AND L. DOMÍNGUEZ, *Spoligotyping profile change caused by deletion of a direct variable repeat in a Mycobacterium tuberculosis isogenic laboratory strain*, Journal of clinical microbiology, 42 (2004), pp. 5388–5391.

[6] C. D. ASHBAUGH, S. ALBERTÍ, AND M. R. WESSELS, *Molecular analysis of the capsule gene region of group A Streptococcus: the hasAB genes are sufficient for capsule expression*, Journal of bacteriology, 180 (1998), pp. 4955–4959.

[7] N. ATTAR, *Techniques & applications: A first genome assembly for nanopore sequencing*, Nature Reviews Microbiology, 13 (2015), pp. 459–459.

[8] K. F. AU, J. G. UNDERWOOD, L. LEE, AND W. H. WONG, *Improving PacBio long read accuracy by short read alignment*, PLoS One, 7 (2012), p. e46679.

[9] D. BANNERMAN, *Pathogen-dependent induction of cytokines and other soluble inflammatory mediators during intramammary infection of dairy cows*, Journal of animal science, 87 (2009), pp. 10–25.

[10] D. D. BANNERMAN, M. J. PAAPE, J. P. GOFF, K. KIMURA, J. D. LIPPOLIS, AND J. C. HOPE, *Innate immune response to intramammary infection with Serratia marcescens and Streptococcus uberis*, Veterinary research, 35 (2004), pp. 681–700.

[11] R. BARRANGOU, C. FREMAUX, H. DEVEAU, M. RICHARDS, P. BOYAVAL, S. MOINEAU, D. A. ROMERO, AND P. HORVATH, *CRISPR provides acquired resistance against viruses in prokaryotes*, Science, 315 (2007), pp. 1709–1712.

[12] F. BAUMDICKER, W. R. HESS, AND P. PFAFFELHUBER, *The infinitely many genes model for the distributed genome of bacteria*, Genome biology and evolution, 4 (2012), pp. 443–456.

[13] U. BERGTHORSSON AND H. OCHMAN, *Heterogeneity of genome sizes among natural isolates of Escherichia coli.*, Journal of bacteriology, 177 (1995), pp. 5784–5789.

[14] D. E. BESSEN, N. KUMAR, G. S. HALL, D. R. RILEY, F. LUO, S. LIZANO, C. N. FORD, W. M. MCSHAN, S. V. NGUYEN, J. C. D. HOTOPP, ET AL., *Whole-genome association study on tissue tropism phenotypes in group A Streptococcus*, Journal of bacteriology, 193 (2011), pp. 6651–6663.

[15] D. BIKARD, A. HATOUM-ASLAN, D. MUCIDA, AND L. A. MARRAFFINI, *Crispr interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection*, Cell host & microbe, 12 (2012), pp. 177–186.

[16] E. BILLE, J.-R. ZAHAR, A. PERRIN, S. MORELLE, P. KRIZ, K. A. JOLLEY, M. C. MAIDEN, C. DERVIN, X. NASSIF, AND C. R. TINSLEY, *A chromosomally integrated bacteriophage in invasive meningococci*, The Journal of experimental medicine, 201 (2005), pp. 1905–1913.

[17] C. BIONDO, G. MANCUSO, A. MIDIRI, G. SIGNORINO, M. DOMINA, V. L. CARICCIO, M. VENZA, I. VENZA, G. TETI, AND C. BENINATI, *Essential role of interleukin-1 signaling in host defenses against group B streptococcus*, MBio, 5 (2014), pp. e01428–14.

[18] A. BOLOTIN, B. QUINQUIS, A. SOROKIN, AND S. D. EHRLICH, *Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin*, Microbiology, 151 (2005), pp. 2551–2561.

[19] J. BOOTH, *Progress in mastitis control-an evolving problem*, in Proceedings of the British Mastitis Conference, Stoneleigh, 1997, pp. 3–9.

[20] A. J. BRADLEY, *Bovine mastitis: an evolving disease*, The Veterinary Journal, 164 (2002), pp. 116–128.

[21] M. R. BRENT AND R. GUIGO, *Recent advances in gene structure prediction*, Current opinion in structural biology, 14 (2004), pp. 264–272.

[22] A. M. BRITTEN, *The role of diagnostic microbiology in mastitis control programs*, Veterinary Clinics of North America: Food Animal Practice, 28 (2012), pp. 187–202.

[23] T. C. BRUEN, H. PHILIPPE, AND D. BRYANT, *A simple and robust statistical test for detecting the presence of recombination*, Genetics, 172 (2006), pp. 2665–2681.

[24] H. BRÜSSOW, C. CANCHAYA, AND W.-D. HARDT, *Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion*, Microbiology and Molecular Biology Reviews, 68 (2004), pp. 560–602.

[25] H. BUERMANS AND J. DEN DUNNEN, *Next generation sequencing technology: advances and applications*, Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842 (2014), pp. 1932–1941.

[26] G. BUTLAND, J. M. PEREGRÍN-ALVAREZ, J. LI, W. YANG, X. YANG, V. CANADIEN, A. STAROSTINE, D. RICHARDS, B. BEATTIE, N. KROGAN, ET AL., *Interaction*

*network containing conserved and essential protein complexes in Escherichia coli*, Nature, 433 (2005), pp. 531–537.

[27] J. BUTLER, I. MACCALLUM, M. KLEBER, I. A. SHLYAKHTER, M. K. BELMONTE, E. S. LANDER, C. NUSBAUM, AND D. B. JAFFE, *ALLPATHS: de novo assembly of whole-genome shotgun microreads*, Genome research, 18 (2008), pp. 810–820.

[28] S. J. CALLISTER, L. A. MCCUE, J. E. TURSE, M. E. MONROE, K. J. AUBERRY, R. D. SMITH, J. N. ADKINS, AND M. S. LIPTON, *Comparative bacterial proteomics: analysis of the core genome concept*, PloS one, 3 (2008), p. e1542.

[29] A. CAPUCO, R. AKERS, AND J. SMITH, *Mammary growth in Holstein cows during the dry period: Quantification of nucleic acids and histology*, Journal of Dairy Science, 80 (1997), pp. 477–487.

[30] A. CAPUCO, D. WOOD, R. BALDWIN, K. MCLEOD, AND M. PAAPE, *Mammary cell number, proliferation, and apoptosis during a bovine lactation: relation to milk production and effect of bST*, Journal of Dairy Science, 84 (2001), pp. 2177–2187.

[31] J. CARTE, N. T. PFISTER, M. M. COMPTON, R. M. TERNS, AND M. P. TERNS, *Binding and cleavage of CRISPR RNA by Cas6*, Rna, 16 (2010), pp. 2181–2188.

[32] J. CARTE, R. WANG, H. LI, R. M. TERNS, AND M. P. TERNS, *Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes*, Genes & development, 22 (2008), pp. 3489–3496.

[33] T. J. CARVER, K. M. RUTHERFORD, M. BERRIMAN, M.-A. RAJANDREAM, B. G. BARRELL, AND J. PARKHILL, *ACT: the Artemis comparison tool*, Bioinformatics, 21 (2005), pp. 3422–3423.

[34] P. E. CHEN AND B. J. SHAPIRO, *The advent of genome-wide association studies for bacteria*, Current opinion in microbiology, 25 (2015), pp. 17–24.

[35] K. H. CHO AND M. G. CAPARON, *Patterns of virulence gene expression differ between biofilm and tissue communities of streptococcus pyogenes*, Molecular microbiology, 57 (2005), pp. 1545–1556.

[36] K. CHYLINSKI, A. LE RHUN, AND E. CHARPENTIER, *The tracrrna and cas9 families of type ii crispr-cas immunity systems*, RNA biology, 10 (2013), pp. 726–737.

[37] J. CLARKE, H.-C. WU, L. JAYASINGHE, A. PATEL, S. REID, AND H. BAYLEY, *Continuous base identification for single-molecule nanopore DNA sequencing*, Nature nanotechnology, 4 (2009), pp. 265–270.

[38] T. J. COFFEY, G. D. PULLINGER, R. URWIN, K. A. JOLLEY, S. M. WILSON, M. C. MAIDEN, AND J. A. LEIGH, *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 (2006), pp. 1420–1428.

[39] I. COMAS AND S. GAGNEUX, *The past and future of tuberculosis research*, PLoS pathog, 5 (2009), p. e1000600.

[40] A. CORVIN, N. CRADDOCK, AND P. F. SULLIVAN, *Genome-wide association studies: a primer*, Psychological medicine, 40 (2010), p. 1063.

[41] P. COSSART AND R. JONQUIÈRES, *Sortase, a universal target for therapeutic agents against Gram-positive bacteria?*, Proceedings of the National Academy of Sciences, 97 (2000), pp. 5013–5015.

[42] D. COX, C.-C. TSENG, G. BJEKIC, AND S. GREENBERG, *A requirement for phosphatidylinositol 3-kinase in pseudopod extension*, Journal of Biological Chemistry, 274 (1999), pp. 1240–1247.

[43] G. CULLEN AND T. LITTLE, *Isolation of Streptococcus uberis from the rumen of cows and from soil*, Veterinary record, 85 (1969), pp. 115–118.

[44] E. DELTCHEVA, K. CHYLINSKI, C. M. SHARMA, K. GONZALES, Y. CHAO, Z. A. PIRZADA, M. R. ECKERT, J. VOGEL, AND E. CHARPENTIER, *CRISPR RNA maturation by trans-encoded small rna and host factor RNase iii*, Nature, 471 (2011), pp. 602–607.

[45] M. DENIS, N. A. PARLANE, S. J. LACY-HULBERT, E. L. SUMMERS, B. M. BUDDLE, AND D. N. WEDLOCK, *Bactericidal activity of macrophages against Streptococcus uberis is different in mammary gland secretions of lactating and drying off cows*, Veterinary immunology and immunopathology, 114 (2006), pp. 111–120.

[46] H. DEVEAU, J. E. GARNEAU, AND S. MOINEAU, *CRISPR/Cas system and its role in phage-bacteria interactions*, Annual review of microbiology, 64 (2010), pp. 475–493.

[47] S. DIGUISTINI, N. Y. LIAO, D. PLATT, G. ROBERTSON, M. SEIDEL, S. K. CHAN, T. R. DOCKING, I. BIROL, R. A. HOLT, M. HIRST, ET AL., *De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data*, Genome biology, 10 (2009), pp. 1–12.

[48] B. DJABRI, N. BAREILLE, F. BEAUDEAU, AND H. SEEGERS, *Quarter milk somatic cell count in infected dairy cows: a meta-analysis*, Veterinary research, 33 (2002), pp. 335–357.

[49] V. L. DOUGLAS, S. G. FENWICK, D. U. PFEIFFER, N. B. WILLIAMSON, AND C. W. HOLMES, *Genomic typing of Streptococcus uberis isolates from cases of mastitis, in new zealand dairy cows, using pulsed-field gel electrophoresis*, Veterinary microbiology, 75 (2000), pp. 27–41.

[50] M. B. DRENNAN, D. NICOLLE, V. J. QUESNIAUX, M. JACOBS, N. ALLIE, J. MPAGI, C. FRÉMOND, H. WAGNER, C. KIRSCHNING, AND B. RYFFEL, *Toll-like receptor 2-deficient mice succumb to Mycobacterium tuberculosis infection*, The American journal of pathology, 164 (2004), pp. 49–57.

[51] G. DUGAR, A. HERBIG, K. U. FÖRSTNER, N. HEIDRICH, R. REINHARDT, K. NIESELT, AND C. M. SHARMA, *High-resolution transcriptome maps reveal*

*strain-specific regulatory features of multiple Campylobacter jejuni isolates*, PLoS Genet, 9 (2013), p. e1003495.

[52] R. C. EDGAR, *MUSCLE: multiple sequence alignment with high accuracy and high throughput*, Nucleic acids research, 32 (2004), pp. 1792–1797.

[53] S. A. EGAN, D. KURIAN, P. N. WARD, L. HUNT, AND J. A. LEIGH, *Identification of sortase a (Srta) substrates in streptococcus uberis: evidence for an additional hexapeptide (LPXXXD) sorting motif*, Journal of proteome research, 9 (2010), pp. 1088–1095.

[54] S. A. EGAN, P. N. WARD, M. WATSON, T. R. FIELD, AND J. A. LEIGH, *Vru (sub0144) controls expression of proven and putative virulence determinants and alters the ability of Streptococcus uberis to cause disease in dairy cattle*, Microbiology, 158 (2012), pp. 1581–1592.

[55] G. ELSON, I. DUNN-SIEGRIST, B. DAUBEUF, AND J. PUGIN, *Contribution of Toll-like receptors to the innate immune response to Gram-negative and Gram-positive bacteria*, Blood, 109 (2007), pp. 1574–1583.

[56] M. C. ENRIGHT AND B. G. SPRATT, *Multilocus sequence typing*, Trends in microbiology, 7 (1999), pp. 482–487.

[57] B. EWING AND P. GREEN, *Base-calling of automated sequencer traces using phred. ii. error probabilities*, Genome research, 8 (1998), pp. 186–194.

[58] W. J. FAISON, A. ROSTOVTSEV, E. CASTRO-NALLAR, K. A. CRANDALL, K. CHUMAKOV, V. SIMONYAN, AND R. MAZUMDER, *Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes*, Genomics, 104 (2014), pp. 1–7.

[59] M. R. FARHAT, B. J. SHAPIRO, K. J. KIESER, R. SULTANA, K. R. JACOBSON, T. C. VICTOR, R. M. WARREN, E. M. STREICHER, A. CALVER, A. SLOUTSKY, ET AL., *Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis*, Nature genetics, 45 (2013), pp. 1183–1189.

[60] C. FELDMAN, R. COCKERAN, M. JEDRZEJAS, T. MITCHELL, AND R. ANDERSON, *Hyaluronidase augments pneumolysin-mediated injury to human ciliated epithelium*, International journal of infectious diseases, 11 (2007), pp. 11–15.

[61] T. R. FIELD, P. N. WARD, L. H. PEDERSEN, AND J. A. LEIGH, *The hyaluronic acid capsule of Streptococcus uberis is not required for the development of infection and clinical mastitis*, Infection and immunity, 71 (2003), pp. 132–139.

[62] C. A. FIELDING, R. M. MCLOUGHLIN, C. S. COLMONT, M. KOVALEVA, D. A. HARRIS, S. ROSE-JOHN, N. TOPLEY, AND S. A. JONES, *Viral IL-6 blocks neutrophil infiltration during acute inflammation*, The Journal of Immunology, 175 (2005), pp. 4024–4029.

[63] J. FINCH, A. WINTER, A. WALTON, AND J. LEIGH, *Further studies on the efficacy of a live vaccine against mastitis caused by streptococcus uberis*, Vaccine, 15 (1997), pp. 1138–1143.

[64] J. M. FINCH, A. W. HILL, T. R. FIELD, AND J. A. LEIGH, *Local vaccination with killed Streptococcus uberis protects the bovine mammary gland against experimental intramammary challenge with the homologous strain.*, Infection and immunity, 62 (1994), pp. 3599–3603.

[65] A. R. FLORES, R. J. OLSEN, A. WUNSCHE, M. KUMARASWAMI, S. A. SHELBURNE, R. K. CARROLL, AND J. M. MUSSER, *Natural variation in the promoter of the gene encoding the Mga regulator alters host-pathogen interactions in group a Streptococcus carrier strains*, Infection and immunity, 81 (2013), pp. 4128–4138.

[66] D. FOLOGEA, M. GERSHOW, B. LEDDEN, D. S. MCNABB, J. A. GOLOVCHENKO, AND J. LI, *Detecting single stranded DNA with a solid state nanopore*, Nano Letters, 5 (2005), pp. 1905–1909.

[67] M. C. FONTAINE, J. PEREZ-CASAL, X.-M. SONG, J. SHELFORD, P. J. WILLSON, AND A. A. POTTER, *Immunisation of dairy cattle with recombinant Streptococcus uberis GapC or a chimeric CAMP antigen confers protection against heterologous bacterial challenge*, Vaccine, 20 (2002), pp. 2278–2286.

[68] A. P. FRANCISCO, C. VAZ, P. T. MONTEIRO, J. MELO-CRISTINO, M. RAMIREZ, AND J. A. CARRIÇO, *PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods*, BMC bioinformatics, 13 (2012), p. 87.

[69] M. GALARDINI, E. G. BIONDI, M. BAZZICALUPO, AND A. MENGONI, *CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes*, Source code for biology and medicine, 6 (2011), p. 1.

[70] J. E. GARNEAU, M.-È. DUPUIS, M. VILLION, D. A. ROMERO, R. BARRANGOU, P. BOYAVAL, C. FREMAUX, P. HORVATH, A. H. MAGADÁN, AND S. MOINEAU, *The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid dna*, Nature, 468 (2010), pp. 67–71.

[71] E. L. GARSIDE, M. J. SCHELLENBERG, E. M. GESNER, J. B. BONANNO, J. M. SAUDER, S. K. BURLEY, S. C. ALMO, G. MEHTA, AND A. M. MACMILLAN, *Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases*, Rna, 18 (2012), pp. 2020–2028.

[72] E. M. GESNER, M. J. SCHELLENBERG, E. L. GARSIDE, M. M. GEORGE, AND A. M. MACMILLAN, *Recognition and maturation of effector RNAs in a CRISPR interference pathway*, Nature structural & molecular biology, 18 (2011), pp. 688–692.

[73] A. GEYER AND K.-H. SCHMIDT, *Genetic organisation of the M protein region in human isolates of group C and G streptococci: two types of multigene regulator-like (mgrC) regions*, Molecular and General Genetics MGG, 262 (2000), pp. 965–976.

[74] J. J. GILLESPIE, A. R. WATTAM, S. A. CAMMER, J. L. GABBARD, M. P. SHUKLA, O. DALAY, T. DRISCOLL, D. HIX, S. P. MANE, C. MAO, ET AL., *PATRIC:*

*the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species*, Infection and immunity, 79 (2011), pp. 4286–4298.

[75] L. GIOT, J. S. BADER, C. BROUWER, A. CHAUDHURI, B. KUANG, Y. LI, Y. HAO, C. OOI, B. GODWIN, E. VITOLS, ET AL., *A protein interaction map of Drosophila melanogaster*, science, 302 (2003), pp. 1727–1736.

[76] J. GÖDEKE, K. PAUL, J. LASSAK, AND K. M. THORMANN, *Phage-induced lysis enhances biofilm formation in Shewanella oneidensis MR-1*, The ISME journal, 5 (2011), pp. 613–626.

[77] M. R. GRAHAM, L. M. SMOOT, C. A. L. MIGLIACCIO, K. VIRTANEVA, D. E. STURDEVANT, S. F. PORCELLA, M. J. FEDERLE, G. J. ADAMS, J. R. SCOTT, AND J. M. MUSSER, *Virulence control in group a Streptococcus by a two-component gene regulatory system: global expression profiling and in vivo infection modeling*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 13855–13860.

[78] M. GREEN, L. GREEN, A. BRADLEY, P. BURTON, Y. SCHUKKEN, AND G. MEDLEY, *Prevalence and associations between bacterial isolates from dry mammary glands of dairy cows*, Veterinary Record-English Edition, 156 (2005), pp. 71–77.

[79] I. GRISSA, G. VERGNAUD, AND C. POURCEL, *CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats*, Nucleic acids research, 36 (2008), pp. W145–W148.

[80] S. GUINDON, F. DELSUC, J.-F. DUFAYARD, AND O. GASCUEL, *Estimating maximum likelihood phylogenies with PhyML*, Bioinformatics for DNA sequence analysis, (2009), pp. 113–137.

[81] S. GUINDON AND O. GASCUEL, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*, Systematic biology, 52 (2003), pp. 696–704.

[82] J. GÜNTHER, A. CZABANSKA, I. BAUER, J. A. LEIGH, O. HOLST, AND H.-M. SEYFERT, *Streptococcus uberis strains isolated from the bovine mammary gland evade immune recognition by mammary epithelial cells, but not of macrophages*, Veterinary research, 47 (2016), pp. 1–14.

[83] J. GÜNTHER, D. KOCZAN, W. YANG, G. NÜRNBERG, D. REPSILBER, H.-J. SCHUBERTH, Z. PARK, N. MAQBOOL, A. MOLENAAR, AND H.-M. SEYFERT, *Assessment of the immune capacity of mammary epithelial cells: comparison with mammary tissue after challenge with Escherichia coli*, Veterinary research, 40 (2009), pp. 1–14.

[84] A. GUREVICH, V. SAVELIEV, N. VYAHHI, AND G. TESLER, *QUAST: quality assessment tool for genome assemblies*, Bioinformatics, (2013), p. btt086.

[85] R. E. HAURWITZ, M. JINEK, B. WIEDENHEFT, K. ZHOU, AND J. A. DOUDNA, *Sequence-and structure-specific RNA processing by a CRISPR endonuclease*, Science, 329 (2010), pp. 1355–1358.

[86] D. HERNANDEZ, P. FRANÇOIS, L. FARINELLI, M. ØSTERÅS, AND J. SCHRENZEL, *De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer*, Genome research, 18 (2008), pp. 802–809.

[87] L. HERRON-OLSON, J. R. FITZGERALD, J. M. MUSSER, AND V. KAPUR, *Molecular correlates of host specialization in Staphylococcus aureus*, PLoS One, 2 (2007), p. e1120.

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

[89] A. HILL AND C. BRADY, *A note on the isolation and propagation of lytic phages from Streptococcus uberis and their potential for strain typing*, Journal of applied bacteriology, 67 (1989), pp. 425–431.

[90] N. L. HILLER, B. JANTO, J. S. HOGG, R. BOISSY, S. YU, E. POWELL, R. KEEFE, N. E. EHRLICH, K. SHEN, J. HAYES, ET AL., *Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome*, Journal of bacteriology, 189 (2007), pp. 8186–8195.

[91] J. HILLERTON, *Redefining mastitis based on somatic cell count*, Bulletin-FIL-IDF (Belgium), (1999).

[92] P. HORVATH AND R. BARRANGOU, *CRISPR/Cas, the immune system of bacteria and archaea*, Science, 327 (2010), pp. 167–170.

[93] S. L. HOWARD, M. W. GAUNT, J. HINDS, A. A. WITNEY, R. STABLER, AND B. W. WREN, *Application of comparative phylogenomics to study the evolution of Yersinia enterocolitica and to identify genetic differences relating to pathogenicity*, Journal of bacteriology, 188 (2006), pp. 3645–3653.

[94] R. JANSEN, J. EMBDEN, W. GAASTRA, L. SCHOULS, ET AL., *Identification of genes that are associated with DNA repeats in prokaryotes*, Molecular microbiology, 43 (2002), pp. 1565–1575.

[95] B. JAYARAO, B. GILLESPIE, M. LEWIS, H. DOWLEN, AND S. OLIVER, *Epidemiology of Streptococcus uberis intramammary infections in a dairy herd*, Journal of Veterinary Medicine, Series B, 46 (1999), pp. 433–442.

[96] H. JEONG, S. P. MASON, A.-L. BARABÁSI, AND Z. N. OLTVAI, *Lethality and centrality in protein networks*, Nature, 411 (2001), pp. 41–42.

[97] M. S. JIN, S. E. KIM, J. Y. HEO, M. E. LEE, H. M. KIM, S.-G. PAIK, H. LEE, AND J.-O. LEE, *Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide*, Cell, 130 (2007), pp. 1071–1082.

[98] A. KAMBUROV, U. STELZL, H. LEHRACH, AND R. HERWIG, *The ConsensusPathDB interaction database: 2013 update*, Nucleic acids research, (2012), p. gks1055.

[99] F. V. KARGINOV AND G. J. HANNON, *The CRISPR system: small RNA-guided defense in bacteria and archaea*, Molecular cell, 37 (2010), pp. 7–19.

[100] E. KARLSSON, A. LÄRKERYD, A. SJÖDIN, M. FORSMAN, AND P. STENBERG, *Scaffolding of a bacterial genome using MinION nanopore sequencing*, Scientific reports, 5 (2015).

[101] T. KARVELIS, G. GASIUNAS, A. MIKSYS, R. BARRANGOU, P. HORVATH, AND V. SIKSNYS, *crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus*, RNA biology, 10 (2013), pp. 841–851.

[102] Y. KAWAMURA, X.-G. HOU, F. SULTANA, H. MIURA, AND T. EZAKI, *Determination of 16S rRNA sequences of streptococcus mitis and streptococcus gordonii and phylogenetic relationships among members of the genus streptococcus*, International Journal of Systematic and Evolutionary Microbiology, 45 (1995), pp. 406–408.

[103] S. KERRIEN, B. ARANDA, L. BREUZA, A. BRIDGE, F. BROACKES-CARTER, C. CHEN, M. DUESBURY, M. DUMOUSSEAU, M. FEUERMANN, U. HINZ, ET AL., *The intact molecular interaction database in 2012*, Nucleic acids research, (2011), p. gkr1088.

[104] I. KHAN, A. HASSAN, A. ABDULMAWJOOD, C. LAMMLER, W. WOLTER, AND M. ZSCHOCK, *Identification and epidemiological characterization of Streptococcus uberis isolated from bovine mastitis using conventional and molecular methods*, Journal of veterinary science, 4 (2003), pp. 213–224.

[105] M. KOSSAIBATI AND R. ESSLEMONT, *The costs of production diseases in dairy herds in England*, The Veterinary Journal, 154 (1997), pp. 41–51.

[106] A. M. KRACHLER, A. R. WOOLERY, AND K. ORTH, *Manipulation of kinase signaling by bacterial pathogens*, The Journal of cell biology, 195 (2011), pp. 1083–1092.

[107] M. LAABEI, M. RECKER, J. K. RUDKIN, M. ALDELJAWI, Z. GULAY, T. J. SLOAN, P. WILLIAMS, J. L. ENDRES, K. W. BAYLES, P. D. FEY, ET AL., *Predicting the virulence of mrsa from its genome sequence*, Genome research, 24 (2014), pp. 839–849.

[108] C. R. LAING, Y. ZHANG, J. E. THOMAS, AND V. P. GANNON, *Everything at once: comparative analysis of the genomes of bacterial pathogens*, Veterinary microbiology, 153 (2011), pp. 13–26.

[109] C. LANGE, H. LYON, D. DEMEO, B. RABY, E. K. SILVERMAN, AND S. T. WEISS, *A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies*, Human heredity, 56 (2003), pp. 10–17.

[110] B. LANGMEAD, *Aligning short sequencing reads with Bowtie*, Current protocols in bioinformatics, (2010), pp. 11–7.

[111] B. LANGMEAD, C. TRAPNELL, M. POP, S. L. SALZBERG, ET AL., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*, Genome biol, 10 (2009), p. R25.

[112] P. LAPIERRE AND J. P. GOGARTEN, *Estimating the size of the bacterial pan-genome*, Trends in genetics, 25 (2009), pp. 107–110.

[113] J. G. LAWRENCE AND H. OCHMAN, *Molecular archaeology of the Escherichia coli genome*, Proceedings of the National Academy of Sciences, 95 (1998), pp. 9413–9417.

[114] P. LEEKITCHAROENPHON, R. S. KAAS, M. C. F. THOMSEN, C. FRIIS, S. RASMUSSEN, AND F. M. AARESTRUP, *snpTree-a web-server to identify and construct SNP trees from whole genome sequence data*, BMC genomics, 13 (2012), p. S6.

[115] T. LEFÉBURE AND M. J. STANHOPE, *Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition*, Genome Biol, 8 (2007), p. R71.

[116] B. A. LEGAULT, A. LOPEZ-LOPEZ, J. C. ALBA-CASADO, W. F. DOOLITTLE, H. BOLHUIS, F. RODRIGUEZ-VALERA, AND R. T. PAPKE, *Environmental genomics of of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species*, BMC genomics, 7 (2006), p. 171.

[117] J. LEIGH, *Streptococcus uberis: a permanent barrier to the control of bovine mastitis?*, The veterinary journal, 157 (1999), pp. 225–238.

[118] J. LEIGH AND T. FIELD, *Killing of Streptococcus uberis by bovine neutrophils following growth in chemically defined media*, Veterinary research communications, 15 (1991), pp. 1–6.

[119] J. LEIGH, J. FINCH, T. FIELD, N. REAL, A. WINTER, A. WALTON, AND S. HODGKINSON, *Vaccination with the plasminogen activator from Streptococcus uberis induces an inhibitory response and protects against experimental infection in the dairy cow*, Vaccine, 17 (1999), pp. 851–857.

[120] J. A. LEIGH, S. A. EGAN, P. N. WARD, T. R. FIELD, AND T. J. COFFEY, *Sortase anchored proteins of Streptococcus uberis play major roles in the pathogenesis of bovine mastitis in dairy cattle*, Veterinary research, 41 (2010), p. 63.

[121] H. LI AND R. DURBIN, *Fast and accurate long-read alignment with Burrows–Wheeler transform*, Bioinformatics, 26 (2010), pp. 589–595.

[122] H. LI, B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, G. ABECASIS, R. DURBIN, ET AL., *The sequence alignment/map format and SAMtools*, Bioinformatics, 25 (2009), pp. 2078–2079.

[123] R. LI, W. FAN, G. TIAN, H. ZHU, L. HE, J. CAI, Q. HUANG, Q. CAI, B. LI, Y. BAI, ET AL., *The sequence and de novo assembly of the giant panda genome*, Nature, 463 (2010), pp. 311–317.

[124] R. LI, Y. LI, K. KRISTIANSEN, AND J. WANG, *SOAP: short oligonucleotide alignment program*, Bioinformatics, 24 (2008), pp. 713–714.

[125] L. LICATA, L. BRIGANTI, D. PELUSO, L. PERFETTO, M. IANNUCCELLI, E. GALEOTA, F. SACCO, A. PALMA, A. P. NARDOZZA, E. SANTONICO, ET AL., *MINT,*

*the molecular interaction database: 2012 update*, Nucleic acids research, 40 (2012), pp. D857–D861.

[126] S.-H. LIN AND Y.-C. LIAO, *Cisa: contig integrator for sequence assembly of bacterial genomes*, PLoS One, 8 (2013), p. e60843.

[127] L. LIU, Y. LI, S. LI, N. HU, Y. HE, R. PONG, D. LIN, L. LU, AND M. LAW, *Comparison of next-generation sequencing systems*, BioMed Research International, 2012 (2012).

[128] N. J. LOMAN, J. QUICK, AND J. T. SIMPSON, *A complete bacterial genome assembled de novo using only nanopore sequencing data*, Nature methods, (2015).

[129] M.-A. MADOUI, S. ENGELEN, C. CRUAUD, C. BELSER, L. BERTRAND, A. ALBERTI, A. LEMAINQUE, P. WINCKER, AND J.-M. AURY, *Genome assembly using Nanopore-guided long and error-free DNA reads*, BMC genomics, 16 (2015), p. 1.

[130] A. H. MAGADÁN, M.-È. DUPUIS, M. VILLION, AND S. MOINEAU, *Cleavage of phage DNA by the Streptococcus thermophilus CRISPR3-Cas system*, PLoS One, 7 (2012), p. e40913.

[131] A. MAGI, M. BENELLI, A. GOZZINI, F. GIROLAMI, F. TORRICELLI, AND M. L. BRANDI, *Bioinformatics for next generation sequencing data*, Genes, 1 (2010), pp. 294–307.

[132] K. S. MAKAROVA, D. H. HAFT, R. BARRANGOU, S. J. BROUNS, E. CHARPENTIER, P. HORVATH, S. MOINEAU, F. J. MOJICA, Y. I. WOLF, A. F. YAKUNIN, ET AL., *Evolution and classification of the CRISPR–Cas systems*, Nature Reviews Microbiology, 9 (2011), pp. 467–477.

[133] K. S. MAKAROVA, Y. I. WOLF, O. S. ALKHNBASHI, F. COSTA, S. A. SHAH, S. J. SAUNDERS, R. BARRANGOU, S. J. BROUNS, E. CHARPENTIER, D. H. HAFT, ET AL., *An updated evolutionary classification of CRISPR-Cas systems*, Nature Reviews Microbiology, (2015).

[134] L. A. MARRAFFINI AND E. J. SONTHEIMER, *CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA*, science, 322 (2008), pp. 1843–1845.

[135] P. R. MARTIN AND E. A. HØIBY, *Streptococcal serogroup: A Epidemic in Norway 1987–1988*, Scandinavian journal of infectious diseases, 22 (1990), pp. 421–429.

[136] I. MATHWORKS, *MATLAB and Statistics Toolbox release*, 2012.

[137] R. C. MATOS, N. LAPAQUE, L. RIGOTTIER-GOIS, L. DEBARBIEUX, T. MEYLHEUC, B. GONZALEZ-ZORN, F. REPOILA, M. DE FATIMA LOPES, AND P. SERROR, *Enterococcus faecalis prophage dynamics and contributions to pathogenic traits*, PLoS Genet, 9 (2013), p. e1003539.

[138] S. K. MAZMANIAN, G. LIU, H. TON-THAT, AND O. SCHNEEWIND, *Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall*, Science, 285 (1999), pp. 760–763.

[139] T. MCDONALD AND J. MCDONALD, *Streptococci isolated from bovine intramammary infections.*, American journal of veterinary research, 37 (1976), pp. 377–381.

[140] K. J. MCKERNAN, H. E. PECKHAM, G. L. COSTA, S. F. MCLAUGHLIN, Y. FU, E. F. TSUNG, C. R. CLOUSER, C. DUNCAN, J. K. ICHIKAWA, C. C. LEE, ET AL., *Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding*, Genome research, 19 (2009), pp. 1527–1541.

[141] M. L. METZKER, *Sequencing technologies ‚Äî the next generation*, Nature reviews genetics, 11 (2010), pp. 31–46.

[142] M. MILNE, D. BARRETT, J. FITZPATRICK, AND A. BIGGS, *Prevalence and aetiology of clinical mastitis on dairy farms in Devon*, Veterinary record, 151 (2002), pp. 241–242.

[143] A. MIRA, A. B. MARTÍN-CUADRADO, G. D,ÄôAURIA, AND F. RODRÍGUEZ-VALERA, *The bacterial pan-genome: a new paradigm in microbiology*, International Microbiology, 13 (2010), pp. 45–57.

[144] R. D. MITRA AND G. M. CHURCH, *In situ localized amplification and contact replication of many individual DNA molecules*, Nucleic Acids Research, 27 (1999), pp. e34–e39.

[145] S. MITTERHUEMER, W. PETZL, S. KREBS, D. MEHNE, A. KLANNER, E. WOLF, H. ZERBE, AND H. BLUM, *Escherichia coli infection induces distinct local and systemic transcriptome responses in the mammary gland*, BMC genomics, 11 (2010), p. 1.

[146] F. J. MOJICA, J. GARCÍA-MARTÍNEZ, E. SORIA, ET AL., *Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements*, Journal of molecular evolution, 60 (2005), pp. 174–182.

[147] J. MONTOJO, K. ZUBERI, H. RODRIGUEZ, G. D. BADER, AND Q. MORRIS, *Genemania: Fast gene network construction and function prediction for cytoscape*, F1000Research, 3 (2014).

[148] I. MOSHYNSKYY, M. JIANG, M. C. FONTAINE, J. PEREZ-CASAL, L. A. BABIUK, AND A. A. POTTER, *Characterization of a bovine lactoferrin binding protein of Streptococcus uberis*, Microbial pathogenesis, 35 (2003), pp. 203–215.

[149] K. M. MOYES, J. K. DRACKLEY, D. E. MORIN, M. BIONAZ, S. L. RODRIGUEZ-ZAS, R. E. EVERTS, H. A. LEWIN, AND J. J. LOOR, *Gene network and pathway analysis of bovine mammary tissue challenged with Streptococcus uberis reveals induction of cell proliferation and inhibition of ppar$\gamma$ signaling as potential*

*mechanism for the negative relationships between immune response and lipid metabolism*, BMC genomics, 10 (2009), p. 1.

[150] A. MUZZI AND C. DONATI, *Population genetics and evolution of the pan-genome of Streptococcus pneumoniae*, International Journal of Medical Microbiology, 301 (2011), pp. 619–622.

[151] F. NEAVE, F. DODD, R. KINGWILL, AND D. WESTGARTH, *Control of mastitis in the dairy herd by hygiene and management*, Journal of Dairy Science, 52 (1969), pp. 696–707.

[152] T. NOZAWA, N. FURUKAWA, C. AIKAWA, T. WATANABE, B. HAOBAM, K. KUROKAWA, F. MARUYAMA, AND I. NAKAGAWA, *CRISPR inhibition of prophage acquisition in Streptococcus pyogenes*, PLoS One, 6 (2011), p. e19543.

[153] B. OPPMANN, R. LESLEY, B. BLOM, J. C. TIMANS, Y. XU, B. HUNTE, F. VEGA, N. YU, J. WANG, K. SINGH, ET AL., *Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12*, Immunity, 13 (2000), pp. 715–725.

[154] T. D. OTTO, G. P. DILLON, W. S. DEGRAVE, AND M. BERRIMAN, *RATT: rapid annotation transfer tool*, Nucleic acids research, 39 (2011), pp. e57–e57.

[155] R. OVERBEEK, R. OLSON, G. D. PUSCH, G. J. OLSEN, J. J. DAVIS, T. DISZ, R. A. EDWARDS, S. GERDES, B. PARRELLO, M. SHUKLA, ET AL., *The SEED and the*

*Rapid Annotation of microbial genomes using Subsystems Technology (RAST)*, Nucleic acids research, 42 (2014), pp. D206–D214.

[156] M. J. PALLEN, A. C. LAM, M. ANTONIO, AND K. DUNBAR, *An embarrassment of sortases–a richness of substrates?*, Trends in microbiology, 9 (2001), pp. 97–101.

[157] K. L. PALMER AND M. S. GILMORE, *Multidrug-resistant enterococci lack CRISPR-cas*, MBio, 1 (2010), pp. e00227–10.

[158] J. R. PARRISH, J. YU, G. LIU, J. A. HINES, J. E. CHAN, B. A. MANGIOLA, H. ZHANG, S. PACIFICO, F. FOTOUHI, V. J. DIRITA, ET AL., *A proteome-wide protein interaction map for Campylobacter jejuni*, Genome biology, 8 (2007), p. R130.

[159] G. PAULIN-CURLEE, R. SINGER, S. SREEVATSAN, R. ISAACSON, J. RENEAU, D. FOSTER, AND R. BEY, *Genetic diversity of mastitis-associated Klebsiella pneumoniae in dairy cows*, Journal of dairy science, 90 (2007), pp. 3681–3689.

[160] S. J. PEACOCK, C. E. MOORE, A. JUSTICE, M. KANTZANOU, L. STORY, K. MACKIE, G. O'NEILL, AND N. P. DAY, *Virulent combinations of adhesin and toxin genes in natural populations of Staphylococcus aureus*, Infection and immunity, 70 (2002), pp. 4987–4996.

[161] W. R. PEARSON AND D. J. LIPMAN, *Improved tools for biological sequence comparison*, Proceedings of the National Academy of Sciences, 85 (1988), pp. 2444–2448.

[162] W. PETZL, H. ZERBE, J. GÜNTHER, W. YANG, H.-M. SEYFERT, G. NÜRNBERG, AND H.-J. SCHUBERTH, *Escherichia coli, but not Staphylococcus aureus triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow*, Veterinary research, 39 (2008), pp. 1–23.

[163] E. M. PHIZICKY AND S. FIELDS, *Protein-protein interactions: methods for detection and analysis.*, Microbiological reviews, 59 (1995), pp. 94–123.

[164] P. PHUEKTES, P. D. MANSELL, R. S. DYSON, N. D. HOOPER, J. S. DICK, AND G. F. BROWNING, *Molecular epidemiology of Streptococcus uberis isolates from dairy cows with mastitis*, Journal of clinical microbiology, 39 (2001), pp. 1460–1466.

[165] D. POSADA, K. A. CRANDALL, AND E. C. HOLMES, *Recombination in evolutionary genomics*, Annual Review of Genetics, 36 (2002), pp. 75–97.

[166] C. POURCEL, G. SALVIGNOL, AND G. VERGNAUD, *CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies*, Microbiology, 151 (2005), pp. 653–663.

[167] S. M. PRYOR, *Bovine mastitis and ecology of streptococcus uberis*, (2008).

[168] G. D. PULLINGER, M. LÓPEZ-BENAVIDES, T. J. COFFEY, J. H. WILLIAMSON, R. T. CURSONS, E. SUMMERS, J. LACY-HULBERT, M. C. MAIDEN, AND J. A. LEIGH, *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 (2006), pp. 1429–1436.

[169] P. RADIVOJAC, W. T. CLARK, T. R. ORON, A. M. SCHNOES, T. WITTKOP, A. SOKOLOV, K. GRAIM, C. FUNK, K. VERSPOOR, A. BEN-HUR, ET AL., *A large-scale evaluation of computational protein function prediction*, Nature methods, 10 (2013), pp. 221–227.

[170] J. RAGOUSSIS, *Genotyping technologies for genetic research*, Annual review of genomics and human genetics, 10 (2009), pp. 117–133.

[171] P. RAINARD, C. RIOLLET, P. BERTHON, P. CUNHA, A. FROMAGEAU, C. ROSSIGNOL, AND F. B. GILBERT, *The chemokine CXCL3 is responsible for the constitutive chemotactic activity of bovine milk for neutrophils*, Molecular immunology, 45 (2008), pp. 4020–4027.

[172] S. A. RICE, C. H. TAN, P. J. MIKKELSEN, V. KUNG, J. WOO, M. TAY, A. HAUSER, D. MCDOUGALD, J. S. WEBB, AND S. KJELLEBERG, *The biofilm life cycle and virulence of Pseudomonas aeruginosa are dependent on a filamentous prophage*, The ISME journal, 3 (2009), pp. 271–282.

[173] E. ROSEY, R. LINCOLN, P. WARD, R. YANCEY, AND J. LEIGH, *PauA: a novel plasminogen activator from Streptococcus uberis*, FEMS microbiology letters, 178 (1999), pp. 27–33.

[174] J. ROZAS, *Dna sequence polymorphism analysis using DnaSP*, Bioinformatics for DNA sequence analysis, (2009), pp. 337–350.

[175] L. SALWINSKI, C. S. MILLER, A. J. SMITH, F. K. PETTIT, J. U. BOWIE, AND D. EISENBERG, *The database of interacting proteins: 2004 update*, Nucleic acids research, 32 (2004), pp. D449–D451.

[176] S. L. SALZBERG AND J. A. YORKE, *Beware of mis-assembled genomes*, Bioinformatics, 21 (2005), pp. 4320–4321.

[177] D. G. SASHITAL, M. JINEK, AND J. A. DOUDNA, *An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease cse3*, Nature structural & molecular biology, 18 (2011), pp. 680–687.

[178] A. SASSON AND T. P. MICHAEL, *Filtering error from SOLiD output*, Bioinformatics, 26 (2010), pp. 849–850.

[179] Y. SCHUKKEN, F. GROMMERS, D. VAN DE GEER, H. ERB, AND A. BRAND, *Risk factors for clinical mastitis in herds with a low bulk milk somatic cell count. 1. Data and risk factors for all cases*, Journal of Dairy Science, 73 (1990), pp. 3463–3471.

[180] Y. H. SCHUKKEN, J. GÜNTHER, J. FITZPATRICK, M. FONTAINE, L. GOETZE, O. HOLST, J. LEIGH, W. PETZL, H.-J. SCHUBERTH, A. SIPKA, ET AL., *Host-response patterns of intramammary infections in dairy cows*, Veterinary immunology and immunopathology, 144 (2011), pp. 270–289.

[181] B. SCHWIKOWSKI, P. UETZ, AND S. FIELDS, *A network of protein–protein interactions in yeast*, Nature biotechnology, 18 (2000), pp. 1257–1261.

[182] M. SEGURA, M. GOTTSCHALK, AND M. OLIVIER, *Encapsulated Streptococcus suis inhibits activation of signaling pathways involved in phagocytosis*, Infection and immunity, 72 (2004), pp. 5322–5330.

[183] R. K. SELANDER, D. A. CAUGANT, H. OCHMAN, J. M. MUSSER, M. N. GILMOUR, AND T. S. WHITTAM, *Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics.*, Applied and Environmental Microbiology, 51 (1986), p. 873.

[184] M. SERBANESCU, M. CORDOVA, K. KRASTEL, R. FLICK, N. BELOGLAZOVA, A. LATOS, A. YAKUNIN, D. SENADHEERA, AND D. CVITKOVITCH, *Role of the Streptococcus mutans CRISPR-Cas systems in immunity and cell physiology*, Journal of bacteriology, 197 (2015), pp. 749–761.

[185] J. SHENDURE, G. J. PORRECA, N. B. REPPAS, X. LIN, J. P. MCCUTCHEON, A. M. ROSENBAUM, M. D. WANG, K. ZHANG, R. D. MITRA, AND G. M. CHURCH, *Accurate multiplex polony sequencing of an evolved bacterial genome*, Science, 309 (2005), pp. 1728–1732.

[186] S. K. SHUKLA, M. PANTRANG, B. STAHL, A. M. BRISKA, M. E. STEMPER, T. K. WAGNER, E. B. ZENTZ, S. M. CALLISTER, S. D. LOVRICH, J. K. HENKHAUS, ET AL., *Comparative whole-genome mapping to determine Staphylococcus au-*

*reus genome size, virulence motifs, and clonality*, Journal of clinical microbiology, 50 (2012), pp. 3526–3533.

[187] J. T. SIMPSON, K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. JONES, AND I. BIROL, *ABySS: a parallel assembler for short read sequence data*, Genome research, 19 (2009), pp. 1117–1123.

[188] A. J. SMITH, P. N. WARD, T. R. FIELD, C. L. JONES, R. A. LINCOLN, AND J. A. LEIGH, *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 bovine mammary gland*, Infection and immunity, 71 (2003), pp. 4842–4849.

[189] K. SMITH AND J. HOGAN, *Environmental mastitis.*, The Veterinary clinics of North America. Food animal practice, 9 (1993), pp. 489–498.

[190] M. E. SMOOT, K. ONO, J. RUSCHEINSKI, P.-L. WANG, AND T. IDEKER, *Cytoscape 2.8: new features for data integration and network visualization*, Bioinformatics, 27 (2011), pp. 431–432.

[191] L. SORDILLO, K. SHAFER-WEAVER, AND D. DEROSA, *Immunobiology of the mammary gland*, Journal of dairy science, 80 (1997), pp. 1851–1865.

[192] R. SOREK, V. KUNIN, AND P. HUGENHOLTZ, *CRISPR97a widespread system that provides acquired resistance against phages in bacteria and archaea*, Nature Reviews Microbiology, 6 (2008), pp. 181–186.

[193] T. SPIRIG, E. M. WEINER, AND R. T. CLUBB, *Sortase enzymes in Gram-positive bacteria*, Molecular microbiology, 82 (2011), pp. 1044–1059.

[194] C. R. STARR AND N. C. ENGLEBERG, *Role of hyaluronidase in subcutaneous spread and growth of group A streptococcus*, Infection and immunity, 74 (2006), pp. 40–48.

[195] B. SUN, D. LUO, J. SUN, AND J. YAN, *[study on the immunogenicity of major leptospiral genus-specific protein antigens and the distribution of antigens in different serogroups of leptospira interrogans].*, Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi, 27 (2006), pp. 1073–1077.

[196] K. SWANSON, K. STELWAGEN, J. DOBSON, H. HENDERSON, S. DAVIS, V. FARR, AND K. SINGH, *Transcriptome profiling of Streptococcus uberis-induced mastitis reveals fundamental differences between immune gene expression in the mammary gland and in a primary cell culture model*, Journal of dairy science, 92 (2009), pp. 117–129.

[197] D. SZKLARCZYK, A. FRANCESCHINI, M. KUHN, M. SIMONOVIC, A. ROTH, P. MINGUEZ, T. DOERKS, M. STARK, J. MULLER, P. BORK, ET AL., *The string database in 2011: functional interaction networks of proteins, globally integrated and scored*, Nucleic acids research, (2010), p. gkq973.

[198] J. TAGG AND M. RALSTON, *Use of an inhibitor typing scheme to study the epidemiology of Streptococcus uberis mastitis*, New Zealand veterinary journal, 36 (1988), pp. 115–119.

[199] O. TAKEUCHI AND S. AKIRA, *Pattern recognition receptors and inflammation*, Cell, 140 (2010), pp. 805–820.

[200] R. TASSI, T. MCNEILLY, J. FITZPATRICK, M. FONTAINE, D. REDDICK, C. RAMAGE, M. LUTTON, Y. SCHUKKEN, AND R. ZADOKS, *Strain-specific pathogenicity of putative host-adapted and nonadapted strains of Streptococcus uberis in dairy cattle*, Journal of dairy science, 96 (2013), pp. 5129–5145.

[201] R. TASSI, T. N. MCNEILLY, A. SIPKA, AND R. N. ZADOKS, *Correlation of hypothetical virulence traits of two Streptococcus uberis strains with the clinical manifestation of bovine mastitis*, Veterinary research, 46 (2015), pp. 1–12.

[202] L. F. TAYLOR, J. VAN DONKERSGOED, O. M. RADOSTITS, C. W. BOOKER, E. J. DUBOVI, J. VAN DEN HURK, AND E. D. JANZEN, *Investigation of an outbreak of mucosal disease in a beef cattle herd in southwestern Saskatchewan.*, The Canadian Veterinary Journal, 35 (1994), p. 425.

[203] R. C. TEAM, *R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. 2013*, 2014.

[204] H. TETTELIN, V. MASIGNANI, M. J. CIESLEWICZ, C. DONATI, D. MEDINI, N. L. WARD, S. V. ANGIUOLI, J. CRABTREE, A. L. JONES, A. S. DURKIN, ET AL., *Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ‚Äúpan-genome‚Äù*, Proceedings of the National Academy of Sciences of the United States of America, 102 (2005), pp. 13950–13955.

[205] H. TETTELIN, D. RILEY, C. CATTUTO, AND D. MEDINI, *Comparative genomics: the bacterial pan-genome*, Current opinion in microbiology, 11 (2008), pp. 472–477.

[206] L. THOMAS, W. HAIDER, A. HILL, AND R. COOK, *Pathologic findings of experimentally induced Streptococcus uberis infection in the mammary gland of cows.*, American journal of veterinary research, 55 (1994), pp. 1723–1728.

[207] D. TODHUNTER, K. SMITH, AND J. HOGAN, *Environmental streptococcal intramammary infections of the bovine mammary gland*, Journal of dairy science, 78 (1995), pp. 2366–2374.

[208] A. E. TOMKINSON, S. VIJAYAKUMAR, J. M. PASCAL, AND T. ELLENBERGER, *DNA ligases: structure, reaction mechanism, and function*, Chemical reviews, 106 (2006), pp. 687–699.

[209] S. UEMATSU AND S. AKIRA, *Toll-Like receptors (TLRs) and their ligands*, in Toll-like receptors (TLRs) and innate immunity, Springer, 2008, pp. 1–20.

[210] J. VAMATHEVAN, J. D. HOLBROOK, AND R. D. EMES, *The mouse genome as a rodent model in evolutionary studies*, eLS, (2008).

[211] J. VAN DER OOST, M. M. JORE, E. R. WESTRA, M. LUNDGREN, AND S. J. BROUNS, *CRISPR-based adaptive and heritable immunity in prokaryotes*, Trends in biochemical sciences, 34 (2009), pp. 401–407.

[212] M. VELDHOEN, R. J. HOCKING, C. J. ATKINS, R. M. LOCKSLEY, AND B. STOCKINGER, *TGF$\beta$ in the context of an inflammatory cytokine milieu sup-*

*ports de novo differentiation of IL-17-producing T cells*, Immunity, 24 (2006), pp. 179–189.

[213] G. P. WAGNER, C. FRIED, S. J. PROHASKA, AND P. F. STADLER, *Divergence of conserved non-coding sequences: rate estimates and relative rate tests*, Molecular biology and evolution, 21 (2004), pp. 2116–2121.

[214] R. WANG, G. PREAMPLUME, M. P. TERNS, R. M. TERNS, AND H. LI, *Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage*, Structure, 19 (2011), pp. 257–264.

[215] P. N. WARD, T. R. FIELD, W. G. DITCHAM, E. MAGUIN, AND J. A. LEIGH, *Identification and Disruption of Two Discrete Loci Encoding Hyaluronic Acid Capsule Biosynthesis GeneshasA, hasB, and hasC in streptococcus uberis*, Infection and immunity, 69 (2001), pp. 392–399.

[216] P. N. WARD, M. T. HOLDEN, J. A. LEIGH, N. LENNARD, A. BIGNELL, A. BARRON, L. CLARK, M. A. QUAIL, J. WOODWARD, B. G. BARRELL, ET AL., *Evidence for niche adaptation in the genome of the bovine pathogen Streptococcus uberis*, BMC genomics, 10 (2009), p. 54.

[217] P. N. WARD AND J. A. LEIGH, *Characterization of PauB, a novel broad-spectrum plasminogen activator from Streptococcus uberis*, Journal of bacteriology, 184 (2002), pp. 119–125.

[218] J. L. WATTS, *Characterization and identification of streptococci isolated from bovine mammary glands*, Journal of dairy science, 71 (1988), pp. 1616–1624.

[219] L. A. WEINERT, R. R. CHAUDHURI, J. WANG, S. E. PETERS, J. CORANDER, T. JOMBART, A. BAIG, K. J. HOWELL, M. VEHKALA, N. VÄLIMÄKI, ET AL., *Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis*, Nature communications, 6 (2015).

[220] O. WELLNITZ AND R. M. BRUCKMAIER, *The innate immune response of the bovine mammary gland to bacterial infection*, The veterinary journal, 192 (2012), pp. 148–152.

[221] I. W. T. WIBAWAN, C. LÄMMLER, AND F. H. PASARIBU, *Role of hydrophobic surface proteins in mediating adherence of group B streptococci to epithelial cells*, Microbiology, 138 (1992), pp. 1237–1242.

[222] H. YAN AND B. XU, *Towards rapid DNA Sequencing: Detecting Single-Stranded DNA with a Solid-State Nanopore*, Small, 2 (2006), pp. 310–312.

[223] M. YANDELL AND D. ENCE, *A beginner's guide to eukaryotic genome annotation*, Nature Reviews Genetics, 13 (2012), pp. 329–342.

[224] X. YANG, D. LIU, F. LIU, J. WU, J. ZOU, X. XIAO, F. ZHAO, AND B. ZHU, *HTQC: a fast quality control toolkit for illumina sequencing data*, BMC bioinformatics, 14 (2013), p. 1.

[225] R. ZADOKS, *Sources and epidemiology of Streptococcus uberis, with special emphasis on mastitis in dairy cattle*, CAB Rev Perspect Agric Vet Sci Nutr Nat Resour, 2 (2007), p. 15.

[226] R. ZADOKS, H. ALLORE, H. BARKEMA, O. SAMPIMON, G. WELLENBERG, Y. GRÖHN, AND Y. SCHUKKEN, *Cow-and quarter-level risk factors for Streptococcus uberis and Staphylococcus aureus mastitis*, Journal of Dairy Science, 84 (2001), pp. 2649–2663.

[227] B. ZARNEGAR, S. YAMAZAKI, J. Q. HE, AND G. CHENG, *Control of canonical NF-κb activation through the NIK–IKK complex pathway*, Proceedings of the National Academy of Sciences, 105 (2008), pp. 3503–3508.

[228] D. R. ZERBINO AND E. BIRNEY, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs*, Genome research, 18 (2008), pp. 821–829.

[229] Y. ZHANG, N. HEIDRICH, B. J. AMPATTU, C. W. GUNDERSON, H. S. SEIFERT, C. SCHOEN, J. VOGEL, AND E. J. SONTHEIMER, *Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis*, Molecular cell, 50 (2013), pp. 488–503.

[230] X. ZHAO AND P. LACASSE, *Mammary tissue damage during bovine mastitis: causes and control*, Journal of animal science, 86 (2008), pp. 57–65.

[231] Y. ZHAO, J. WU, J. YANG, S. SUN, J. XIAO, AND J. YU, *PGAP: pan-genomes analysis pipeline*, Bioinformatics, 28 (2012), pp. 416–418.

[232] J. ZHOU, *Microarrays for bacterial detection and microbial community analysis*, Current opinion in microbiology, 6 (2003), pp. 288–294.

[233] Y. ZHOU, Y. LIANG, K. H. LYNCH, J. J. DENNIS, AND D. S. WISHART, *PHAST: a fast phage search tool*, Nucleic acids research, (2011), p. gkr485.

[234] F. ZHU AND X. ZHANG, *The Wnt signaling pathway is involved in the regulation of phagocytosis of virus in Drosophila*, Scientific reports, 3 (2013).

[235] M. E. ZWICK, M. K. THOMASON, P. E. CHEN, H. R. JOHNSON, S. SOZHAMANNAN, A. MATECZUN, AND T. D. READ, *Genetic variation and linkage disequilibrium in Bacillus anthracis*, Scientific reports, 1 (2011).